文章编号:1671-4598(2023)07-0021-07 DOI:10.16526/j.cnki.11—4762/tp.2023.07.004 中图分类号:TP391.43 文献标识码:A

基于像素分配的文本检测方法研究

吉训生[1],喻 智[1],徐晓祥[2]
(1.江南大学 物联网工程学院,江苏 无锡 214122;
2.无锡市创凯电气控制设备有限公司,江苏 无锡 214400)

摘要:针对现有方法在场景文本检测上的不足,提出一种基于像素分配方的场景文本检测方法,并采用了交叉注意力模块和多尺度特征自适应模块来分别在空间和和通道上优化特征提取;为了丰富不同尺度的特征表示,采用多尺度特征自适应模块进行自动分配不同尺度特征的权重;为了有效获取上下文信息,将特征网络提取到的特征送入交叉注意力模块;对每个像素,在其所在的水平路径和垂直路径上收集上下文信息;再通过循环操作,每一个像素便可以在全图范围内获取上下文信息;通过全卷积网络方法,使用多任务学习框架学习文本实例的几何特征,结合多任务学习的结果完成像素到文本框的分配,经过简单处理后重建文本实例的多边形边界框;在任意形状公开数据集 Total-text 上进行测试,文章方法的召回率、精确率、F 值分别为 75.71%、89.15%、81.89%,在多方向公开数据集 ICDAR2015 上也表现良好,经实验得召回率、精确率、F 值分别为 79.06%、89.24%、83.84%,证明了文章方法的有效性。

关键词:图像处理;文本检测;交叉注意力;像素分配

Research on Text Detection Method of Pixel-based Allocation

JI Xunsheng[1], YU Zhi[1], XU Xiaoxiang[2]
(1. School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China;
2. Wuxi CK Electric Control Equipment Co., Ltd., Wuxi 214400, China)

Abstract: Aiming at the shortcomings of existing methods in scene text detection, a scene text detection method based on the pixel allocation is proposed, and the cross-attention module and multi-scale feature self-adaptive module are used to optimize the feature extraction in the space and channel respectively. In order to enrich the feature representations of different scales, the multi-scale feature self-adaptive module is used to automatically assign the feature weights of different scales. In order to effectively obtain the contextual information, the extracted features of the feature network are fed into the cross-attention module. The contextual information of each pixel is collected on its horizontal path and vertical path. Then through the loop operation, the context information of the whole image can be obtained in each pixel. The multi-task learning framework is used to learn the geometric features of the text instance by the full convolutional network method, and the results of the multi-task learning are combined to complete the allocation of pixels to the text box, and the polygonal bounding box of the text instance is reconstructed after simple processing. Tested on the public dataset Total-text with any shape, the recall rate, precision rate, and F value of the proposed method are 75.71%, 89.15%, and 81.89%, respectively, which also performs well on the multi-directional public dataset ICDAR2015. The recall rate, precision rate, and F value are 79.06%, 89.24%, and 83.84%, respectively, which proves the effectiveness of the method in this paper.

Keywords: image processing; text detection; cross-attention; pixel allocation

0 引言

场景文本检测在产品搜索、在线教育、即时翻译和电路板字符检测等有着广泛的应用,受到学术界和工业界的广泛关注。由于前景文本和背景物体差异很大,并且文本在形状、颜色、字体、方向和比例等方面也存在变化的多样性,文本检测仍然是一项具有挑战性的任务。传统方法中的特征都是人为设计的[1-2],需要大量的先验知识,且模型的鲁棒性差。基于回归[4-9]和分割[10-13]的深度学习的文本检测方法可以学习有效的特征来检测文本。文献[4]扩展了 SSD(single shot MultiBox detector)算法[3]以解决不规则文本框呈现的不同长宽比的问题,通过修改卷积内核和锚框的大小来有效地捕获各种文本形状。文献[5]在 Faster R-CNN[6](Faster Regions with CNN)引入 RoI-Pooling(regions of interest pooling),以检测任意方向的场景文本。文献[7]使用全卷积网络 FCN(fully convolutional networks)[8]直接预测像素级的四边形框,而不需要文本候选框和预设锚点。文献[9]提出了一种基于 FCN 的注意机制,从本质上抑制了特征图中的背景干扰,实现文本的精

收稿日期:2023-02-08; **修回日期**:2023-03-07。
基金项目:国家自然科学基金(61771223);江苏省重点研发计划项目(BE2018334)。
作者简介:吉训生(1969-),男,江苏南通人,博士,教授,硕士生导师,主要从事微弱信号处理方向的研究。
通讯作者:喻 智(1998-),男,江西南昌人,大学本科,硕士研究生,主要从事场景文本检测方向的研究。
引用格式:吉训生,喻 智,徐晓祥.基于像素分配的文本检测方法研究[J].计算机测量与控制,2023,31(7):21-27.

确检测。文献［10］先将文本区域当作若干组件，再将文本组件连接为文本区域以实现文本检测。文献［11］将文本实例与像素之间的连通性进行文本分割，再根据分割结果生成边界框。文献［12］设计文本中心线，通过文本中心线上多个圆环来预测任意形状文本区域。李敏等[14]先根据文本像素颜色进行聚类，再对文本检测。

为了提升网络性能，各种注意力模块广泛地应用于深度学习的中，文献［15］计算特征图中各空间点之间的相关矩阵，利用非局部模块生成注意图，然后引导上下文信息聚合。文献［16］通过叠加两个交叉注意力模块，更有效地从所有像素中获取上下文信息，有效地增强了特征表示。文献［17］等提出 SE（squeeze-and-excitation）模块，SE 通过建模通道之间的相互依赖关系，利用网络的全局损失函数自适应地重新矫正通道之间的特征相应强度，实现各个通道的权重自动分配。文献［18］等为了解决通道注意和空间注意只能有效的捕获了局部信息，但不能捕获通道之间的长依赖关系的问题，提出一种有效利用不同尺度特征图中的空间信息的方法，在更细粒度水平上提取多尺度的空间信息。

在文本实例中两个字符间距很大或很小时，像素的预测比较模糊，容易产生误判。对于基于锚框的检测方法而言，待检测的文本长度未知，且文本行的长宽比也未知，无法确定锚框的尺寸，这一定程度上加大了检测的难度。不仅如此，由于缺乏全局上下文信息，在分割两个紧密相连的文本实例时，难以通过语义分割的方法来分离，对于长文本的检测又容易被切分成不同的文本实例。

为了解决上述问题，本文提出一种像素分配的场景文本检测方法，采用循环交叉注意力模块可以有效聚合上下文信息，像素到文本框的分配可以完成文本实例的检测。交叉注意力模块（CCAB，criss-cross attention block）收集每个像素所在的水平路径和垂直路径上的信息，进一步通过循环操作整合全图范围内的上下文信息。由于不同尺度的特征感受野不一样，进而侧重描述的信息也不同，为了得到更丰富的多尺度特征，对多尺度特征进行自动分配权重。本文使用 FCN 模型和多任务学习机制，将高级对象信息和低级像素信息进行整合，完成任意形状文本的检测。其中多任务学习包括文本中心区域得分（Score），像素到文本框的 4 个顶点的偏移（Quad），像素到文本框的分配（PBA，pixel to box assignment），像素到文本上下边界的偏移（TBO，text border offset）。结合 Score 的二值化图和 Quad 确定文本候选框，通过 PBA 可以有效地解决长文本被分割成不同文本实例片段以及相邻较近的文本实例无法区分的问题。在像素分配过程中，为了更有效地区分不同的文本实例所属的文本框，针对候选框外的像素添加一个惩罚来抑制其分配到文本候选框中，最后通过 TBO 细化多边形文本框，输出预测的文本框的多边形框。

1 交叉注意力和像素分配的文本检测方法

1.1 整体结构

本文所使用的深度学习的网络结构结构如图 1 所示，采用全卷积网络的多任务学习框架来重建文本区域的各种几何特性，通过循环 CCAB 模块聚合上下文信息，采用像素到文本框分配的方法完成文本实例分割。输入图像经过主干神经网络提取特征，本文的主干神经网络采用 ResNet50-FPN 结构，并采用多尺度自适应模块 MFA（multi-scale feature adaptive module）来融合不同尺度的特征。融合特征 F 通过循环 CCAB 来整合像素的全局依赖性，以获取更具代表性的特征 F''。在特征 F'' 上多任务学习，获取 Score、Quad、像素到 PBA 以及 TBO，结合 4 个任务的结果后并经过后处理得到多边形文本框。通过 Score 和 Quad

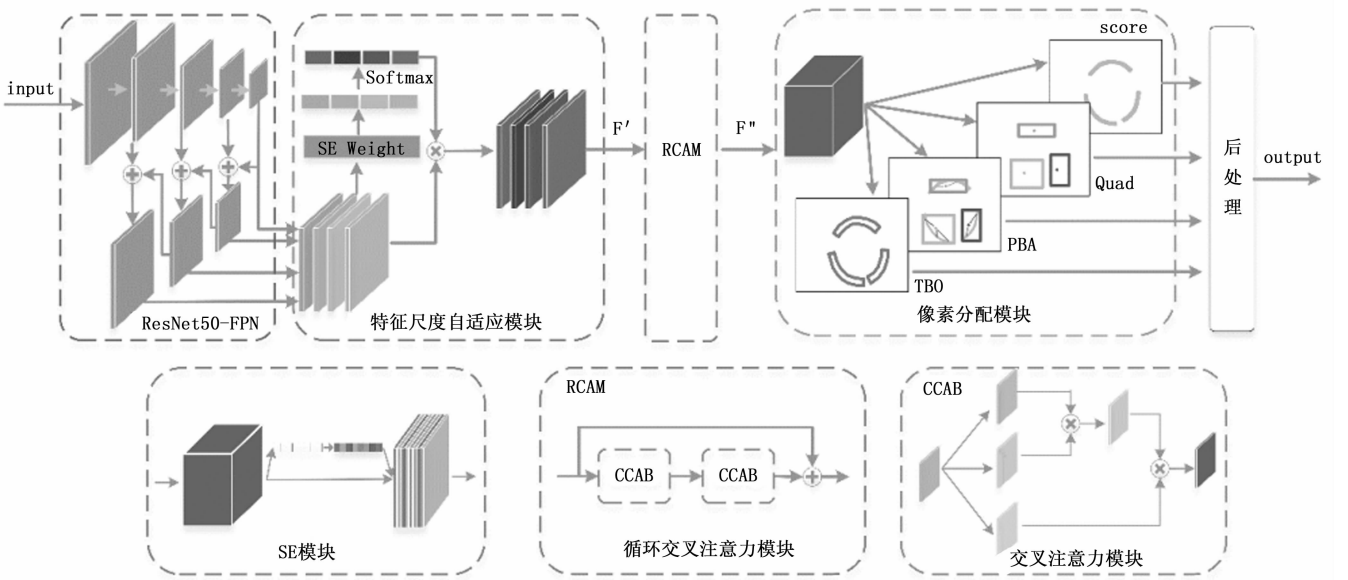

图 1 网络结构图

先确定粗矩形框，PBA 将相邻的矩形框内的像素分配到其对应的矩形框中，最后通过 TBO 对 Score 图中的像素进行上下边框的预测，根据每一个像素点到文本框的上下边界的距离来推导文本框。本文的方法是一种端到端的场景文本检测方法，输入场景文本图像，直接输出文本区域的文本框预测。

1.2　特征尺度自适应模块

不同尺度的特征具有不一样的感受野，它们所侧重描述了不同尺度的信息。对于视觉任务而言，有效地提取多尺度特征能够大大提升特征的表示能力。通道注意力机制自提出以来，被广泛应用到深度学习网络中，在视觉任务中其对网络性能的提升起到了不可忽视的作用。通道注意力机制允许网络有选择的对各个通道的重要性进行加权，从而有更多的信息输出。多数方法中采用了 SE 模块来对特征图的各个通道进行权重自动分配，但是 SE 模块只引入了通道注意力而忽略了全局的空间信息。为了丰富多尺度特征，本文采用多尺度特征自适应模块 MFA 来加强特征表示。如图 1 中主干神经网络 ResNet50-FPN 所示，ResNet50-FPN 输出有 4 个尺度的特征，MFA 模块将对这 4 个不同尺度的特征进行自适应融合。如图 1 中多尺度特征自适应模块所示，对每个尺度的特征采用 SE 模块提取不同尺度特征图的注意力，获取通道方向的注意力向量，通过 softmax 重新校准通道方向的注意力向量，获得多尺度通道的权重分配。将获取到的多尺度特征注意力向量逐个对多尺度特征进行加权。利用 MFA 模块获取到的多尺度特征信息更加细化。

1.3　循环交叉注意力模块

注意力模块广泛应用于各种任务中，本文采用循环交叉注意力模块更有效地获取上下文信息。将通过特征提取模块获取到的特征 F 输入交叉注意力模块中，生成新的特征图 F'，将纵横路径中每个像素的上下文信息结合。如图 2 (a) 中灰色路径所示，特征图 F' 仅在水平方向和垂直方向获取上下文信息，难以获得更加丰富和密集的上下文信息。因此，将特征图 F' 再次输入交错注意力模块，对于图 2 (a) 中灰色路径上的所有像素的水平方向和垂直方向再次获取上下文信息，输出特征图 F''。本文中的 2 个交叉注意力模块共享参数，只增加了一小部分计算开销，使得网络的性能更高。

交叉注意力模块如图 3 所示，特征图 $F \in \mathbb{R}^{C\times H\times W}$ 经过 2 个 $1*1$ 的卷积运算分别得到 $Q \in \mathbb{R}^{C'\times H\times W}$ 和 $K \in \mathbb{R}^{C'\times H\times W}$，为了减少运算，本文中取 $C' = C/8$。计算 Q 和 K 的相似性并获取到注意力图。

$A \in \mathbb{R}^{(H+W-1)\times H\times W}$。计算过程如式 (1) 所示：

$$d_{i,u} = Q_u K_{i,u}^{T} \tag{1}$$

其中：$Q_u \in \mathbb{R}^{C'}$ 是 Q 在空间维度上位置为 u 的像素向量，$K_u \in \mathbb{R}^{(H+W-1)\times C'}$ 是 K 在空间维度上位置为 u 的像素所在水平方向和竖直方向的像素向量，$K_{i,u} \in \mathbb{R}^{C'}$ 是 K_u 的第 i 个元素，$d_{i,u}$ 是 $D \in \mathbb{R}^{(H+W-1)\times H\times W}$ 在空间维度上位置为 u 的第 i 个元素，对 $D \in \mathbb{R}^{(H+W-1)\times H\times W}$ 进行 softmax 运算得到注意力图 A。

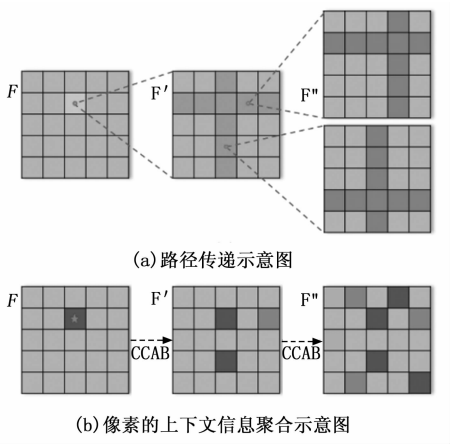

图 2　交叉注意力模块的信息示意图

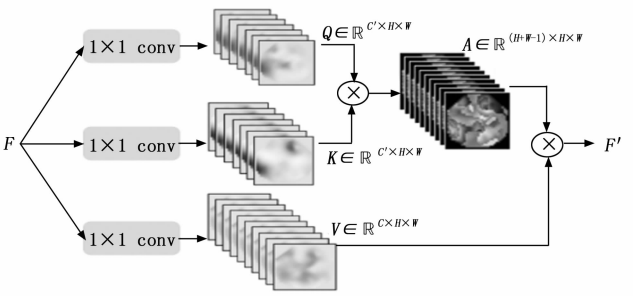

图 3　交叉注意力模块细节图

特征图 $F \in \mathbb{R}^{C\times H\times W}$ 经过 1 个 $1*1$ 的卷积运算得到 $V \in \mathbb{R}^{C\times H\times W}$，将特征图 V 在注意力图 A 上聚集得到新的特征图 F，计算过程如式 (2) 所示：

$$F'_u = \sum_{i=1}^{(H+W-1)} A_{i,u} V_{i,u} + F_u \tag{2}$$

其中：$A_{i,u}$ 是注意力图 A 在空间维度上位置为 u 的第 i 个通道上的标量值，$V_u \in \mathbb{R}^{(H+W-1)\times C}$ 是特征图 V 在空间维度上位置为 u 的像素所在水平方向和垂直方向的像素向量，$V_{i,u} \in \mathbb{R}^{C}$ 是 V_u 中的第 i 个元素，$F_u \in \mathbb{R}^{C}$ 是 F 在空间维度上位置为 u 的像素向量，$F'_u \in \mathbb{R}^{C}$ 是 F' 在空间维度上位置为 u 的像素向量。将获取到的上下文信息添加到原始的特征图 F 中，来聚合上下文信息以及增强特征的表示。交叉注意力模块的运算通过矩阵的运算来完成运算的加速。

1.4　像素分配

在本文中，通过在增强的特征 F'' 上多任务训练来获取 Score、QUAD、PBA 和 TBO，Score 如图 4 所示，

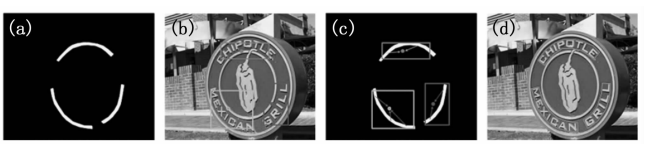

图 4　多任务学习结果

再经过后处理得到文本的实例分割。将图像送入 FCN 网络中，先获取 Score 和 QUAD，其中 Score 是像素在［0，1］内的得分图，分数代表了该像素预测的几何形状的置信度，而 Quad 表示的是文本区域的像素到 4 个顶点的偏移量。Score 只有 1 个通道，Quad 中的矩形框有 4 个顶点以及每个顶点的偏移量包含水平偏移量和竖直偏移量，因此，Quad 的输出包含 8 个通道。根据文本中心区域得分图和给定阈值进行二值化，以及几何输出 Quad 来获取四边形文本候选框，用 NMS 来抑制重叠的文本候选框，文本候选框如图 4（b）所示。通过 PAB 将文本中心区域内的像素分配到对应的文本框中来区分不同的文本实例，如图 4（c）所示。先计算四边形文本候选框的中心点位置，再根据像素到文本中心点预测的偏移量来分配到对应的文本框，像素的文本中心预测到候选框的中心有水平偏移量和垂直偏移量。

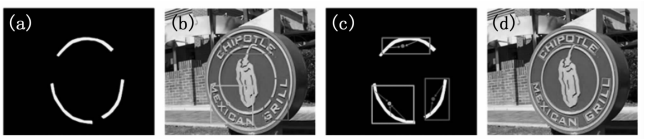

图 4 多任务学习结果

像素到不同候选框的分配如图 5 所示，C_1 和 C_2 分别表示不同文本候选框的中心点，像素 P_1 和像素 P_2 分别属于 C_1 和 C_2 的候选框，用一个向量来表示像素 P_2 到文本框中心 C_1 的预测，其中 d_{off} 表示预测的中心到文本框中心 C_1 的偏移量。为了更加有效地区别不是该文本框中的像素，针对文本框以外的像素添加一个惩罚，P_2 属于文本框 C_1 外的像素，由于 P_2 对文本框中心点 C_1 的预测向量和文本框的边界有相交，定义 P_2 到交点的距离 d_{out} 作为惩罚以抑制像素 P_2 对文本中心点 C_1 的预测。本文方法结合高级的对象信息和底层的像素信息，可以有效地将文本中心区域中的像素分配到对应的文本候选框中，并且能够解决长文本实例在分割时被切割成多个不同文本实例的问题。惩罚 d_{out} 可以很好地区分相邻较近的文本实例。

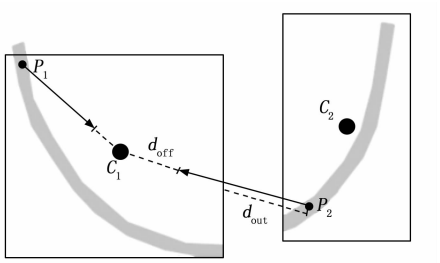

图 5 像素分配示意图

通过 TBO 细化每一个文本实例的文本边界框，对文本中心区域的每个像素预测上下边界的偏移量重建文本实例的边界框。如图 6 所示，像素 P 到文本候选框的边界有水平偏移量和垂直偏移量，而边界有上下 2 个边界，因此 TBO 的预测输出有 4 个通道。

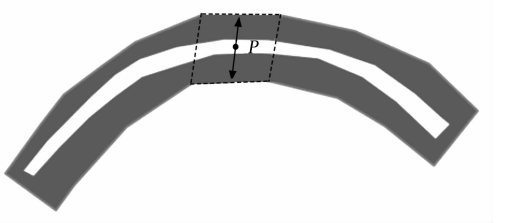

图 6 像素到上下边界偏移量示意图

1.5 损失函数

本文的损失函数分为 4 个部分，分别是得分图损失 L_{score}，四边形回归损失 L_{quad}，像素到文本中心偏移损失 L_{pba}，文本到上下边界便宜损失 L_{tbo}。多任务学习总损失 L 的计算如式（3）所示，

$$L = \lambda_1 L_{\text{score}} + \lambda_2 L_{\text{quad}} + \lambda_3 L_{\text{pba}} + \lambda_4 L_{\text{tbo}} \tag{3}$$

其中：λ_1，λ_2，λ_3，λ_4 用于权衡 4 个损失之间的重要性，本文中分别设置为 1，1.5，1.5，1 。在多任务训练中 L_{score} 为二分类损失，其他 3 个损失都是回归损失。

为了简化计算过程，文本使用 EAST[7]（efficient and accurate scene text）中介绍的得分图损失 L_{score} 和四边形回归损失 L_{quad}。采用类平衡交叉熵作为 L_{score} 损失，L_{score} 计算如式（4）和（5）所示，

$$L_{\text{score}} = -\beta Y^* \log \hat{Y} - (1-\beta)(1-Y^*)\log(1-\hat{Y}) \tag{4}$$

$$\beta = 1 - \frac{\sum y^*}{|Y^*|} \tag{5}$$

其中：Y^* 是真值，$\hat{Y}$ 是预测输出，β 是用于权衡正负样本的一个系数，$y^* \in Y^*$。

本文的回归损失函数采用 $smoothed_{L_1}$ 和 L_{quad}，计算如式（6）和式（7）所示，

$$L_{\text{quad}} = \min_{\tilde{Q} \in P_Q} \sum_{\substack{c_i \in C_Q \\ \tilde{c}_i \in \tilde{C}_Q}} \frac{smoothed_{L_1}(c_i - \tilde{c}_i)}{8 \times N_{Q^*}} \tag{6}$$

$$smoothed_{L_1} = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{7}$$

其中：$C_Q = \{x_1, y_1, \cdots, x_4, y_4\}$ 是 4 个顶点的坐标值，N_{Q^*} 为四边形中的最小边长，P_{Q^*} 为不同顶点顺序的几何。本文规定点的顺序为左上角为第一个顶点，左下角为最后一个顶点，顶点按顺时针方向排列。而 L_{pba} 损失和 L_{tbo} 损失直接使用 $smoothed_{L_1}$ 损失计算。

2 实验结果与分析

2.1 数据集介绍

本文主要采用 Total-text [19] 和 ICDAR2015[20] 两个公共数据集作为训练与测试数据集。Total-text 是一个具有挑战性的任意形状文本检测数据集，它由 1 255 张训练图片和 300 张测试图片组成，这些图像具有多种不同的文本形态：水平、任意方向和弯曲形状，其文本实例的标签由字符级别的标注组成。ICDAR2015 是针对多方向文本检测的数据集，该数据集包括 1 000 张训练图片和 500 张测试图片，而且图像中的文本以英文和数字为主，这些图像的文本实例

标签也是字符级别的标签。

2.2　标签的生成

数据集的标签采用 14 点标注，左上角的坐标为第一个点，按顺时针方向依次标注，左下角的坐标点为最后一个点。如图 7（a）所示，文本区域的原始标签为外围橙色虚线框，本文中心区域为红色实线框，按 0.5 的缩减比例在原始签上缩减得到文本中心区域标签。如图 7（b）所示，QUAD 的标签为文本中心区域的像素点到四边形文本框的 4 个顶点的偏移，如图 7（c）所示，在文本框内的像素点的 PBA 的标签为像素到四边形文本框中心的偏移，若像素在文本框外部时，需要额外加上惩罚 d_{out}。如图 7（d）所示，TBO 的标签为文本中心区域的像素到边界框的偏移，其中线段 K_1K_2 的斜率为线段 P_1P_2 斜率与线段 P_4P_3 斜率的平均值。通过线段 K_1P_0 与线段 K_1K_2 的长度之比，在线段 P_1P_2 上确定像素 P_0 的上边界偏移点，同理可确定下边界偏移点。

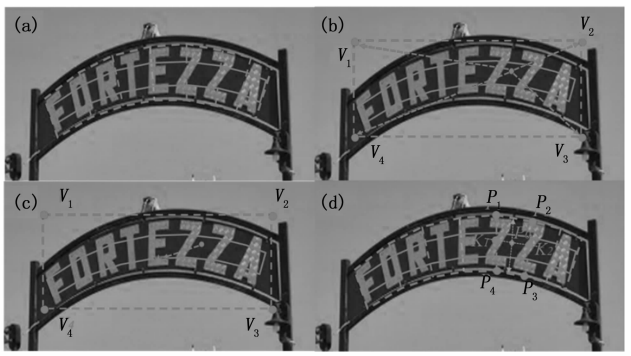

图 7　标签生成细节图

2.3　评价指标

本文实验中的评价指标主要是精确率 Precision、召回率 Recall 以及二者的综合评价 F-measure，计算公式如式（8）～（10）所示，

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F\text{-}measure = \frac{2Precision \times Recall}{Precision + Recall} \tag{10}$$

其中：TP、FP、FN 分别为混淆矩阵中的真阳值、假阳值、假阴值。

2.4　消融实验

本次所有实验的训练和测试都是在同一配置下完成，CPU：Intel（R）Xeon（R）Gold 6 278C CPU @ 2.60 GHz * 16，GPU：NVIDIA Tesla V100 16 GB，RAM：32G。本文实验的优化器使用 Adam 优化器，其中 $\beta_1 = 0.9$，$\beta_2 = 0.999$，学习率为 0.001，每个批次设置为 12。为了提高模型在数据集上的性能，对数据进行随机缩放和随机旋转进行数据增强，最后将数据调整和填充到 512 * 512 的大小。

为了验证交叉注意力模块在网络中发挥的作用，在 Total-text 数据集上进行消融实验对比，为了保持单一控制变量原则，上述的参数设置均相同。在消融实验中 $R=0$ 表示没有交叉注意力模块，$R=1$ 表示只有一个交叉注意力模块，$R=2$ 表示对交叉注意力模块进行 2 次循环操作，$R=3$ 表示对交叉注意力模块三次循环操作，实验结果如表 1 所示。

表 1　Total-text 数据集上交叉注意力模块的消融实验结果

Method	R(%)	P(%)	F(%)	FPS
$R=0$	**78.25**	82.92	80.52	**21.60**
$R=1$	76.66	86.52	81.29	20.31
$R=2$	75.71	89.15	81.89	19.27
$R=3$	75.69	**89.24**	**81.91**	15.54

根据表 1 可以看出，在加入交叉注意力模块后，精确率 Precision 和综合指标 F-measure 都有所提升。当 $R=1$ 时的精确率 Precision 比没有交叉注意力模块的结果提高 3.6%，F-measure 指标提升了 0.77%，表明了交叉注意力模块在横纵路径上聚合上下文信息。$R=2$ 时的精确率 Precision 提高 6.23%，F-measure 指标提升了 1.37%。证明 2 个交叉注意力模块可以提取全图的上下文信息。$R=3$ 时的精确率 Precision 只提高了 6.32%，F-measure 指标提升了 1.39%。同时，随着注意力模块的增加，网络的实时性有小幅度下降，且随着 R 的增加，FPS 逐步减少。

为进一步验证交叉注意力的有效性在 ICADR2015 数据集上进行相同的消融实验对比，上述的参数设置同 Total-text 数据集保持一致。实验结果如表 2 所示。

表 2　ICADR2015 数据集上交叉注意力模块的消融实验结果

Method	R(%)	P(%)	F(%)	FPS
$R=0$	80.16	83.26	81.68	**26.87**
$R=1$	78.93	85.31	81.99	22.51
$R=2$	79.09	89.21	83.85	18.09
$R=3$	79.42	**89.57**	**84.19**	16.24

在 ICADR2015 数据集上的交叉注意力模块的消融实验结果中精确率 Precision 和综合指标 F-measure 都有所提升。当 $R=1$ 时的精确率 Precision 比没有交叉注意力模块的结果提高 2.05%，F-measure 指标提升了 0.31%，表明了交叉注意力模块在横纵路径上聚合上下文信息的有效性。$R=2$ 时的精确率 Precision 提高 5.95%，F-measure 指标提升了 2.17%。$R=3$ 时的精确率 Precision 提高了 5.95%，F-measure 指标提升了 2.17%。将 $R=2$ 与 $R=3$ 进行对比，$R=3$ 时的精确率 Precision 只提高了 0.36%，F-measure 只提升了 0.34%，但是网络每秒处理的帧数却减少了 10.63 FPS。

在 ICADR2015 数据集和 Total-text 数据集上的交叉注意力模块消融实验可知，交叉注意力模块能够显著提升网络性能。当只有一个交叉注意力模块时只能聚合单一路径上的上下文信息，对网络的提升效果还能进一步加强。通过循环操作，将单一路径扩增到了横纵 2 个路径，根据实

验结果显示，对交叉注意力模块进行 2 次循环操作，更进一步提升了网络性能。继续加入循环操作，虽然网络性能提升了，但是提升效果不明显，且随着循环操作的堆叠，网络的实时性下降严重，因此进行 2 次循环操作能够兼顾网络性能与网络的实时性。

2.5 对比实验

为了进一步验证本文方法的有效性，分别在 Total-text 和 ICDAR2015 数据集上对本文方法和近年出现的其他方法进行对比。实验结果如表 3 和 4 所示。从表 3、表 4 中的实验结果表明，本文方法在弯曲文本数据集 Total-text 上取得了 75.71% 的召回率、89.15% 的精确率和 81.89% 的 F-measure 值。

表 3 不同算法在 Total-text 数据集上的比较

Method	R(%)	P(%)	F(%)
SegLink[10]	23.80	30.30	26.70
EAST[7]	36.20	50.00	42.00
Textsnake[12]	74.50	82.70	78.40
SAST [21]	76.86	83.77	80.17
ATRR [22]	76.20	80.90	78.50
CRAFT [23]	79.90	87.60	83.60
Textfield [24]	79.90	81.20	80.60
Lomo [25]	75.70	88.60	81.60
SpcNet[26]	82.80	83.00	82.90
Pan++ [27]	80.50	88.40	84.20
Ours	75.71	89.15	81.89

表 4 不同算法在 ICADR2015 数据集上的比较

Method	R(%)	P(%)	F(%)
EAST[7]	73.50	83.60	78.20
SSTD[9]	73.9	80.2	76.9
SegLink[10]	76.8	73.1	75
DeepReg [28]	80	82	81
Textsnak[12]	80.4	84.9	82.6
PAN [29]	81.9	84	82.9
PixelLink[11]	81.7	82.9	82.3
SATS[21]	87.09	86.72	86.91
Ours	79.06	89.24	83.84

其中精确率相较其他算法提升较大。在多方向文本数据集 ICADR2015 上也取得了较好的结果，其中召回率，精确率，以及 F_1 值分别为 79.06%、89.24%、83.84%。ICADR2015 数据集与 Total-text 数据集的实验结果中精确率均在所有的对比实验中取得了最好的成绩。部分任意文本检测的实验结果可视化如图 8 所示，可以看出本文方法对包含密集文本、长文本以及相邻较近的文本的图像有较好的检测结果。

图 8 部分实验结果图

3 结束语

本文提出一种基于像素分配的场景文本检测方法。该方法采用多任务学习机制对文本中心区域的分割，像素到文本框顶点的偏移，像素到文本框的分配和像素到上下边界的偏移 4 个任务进行训练，通过像素到文本框的分配，能够有效解决长文本被分割成不同文本实例和相邻很近的文本无法区分的问题。通过多尺度特征自适应分配权重，丰富了多尺度特征，结合交叉注意力模块，在空间和通道上细化了特征的表示，在 ICADR2015 数据集和 Total-text 数据集上进行实验，实验结果证明了本文方法的有效性。

参考文献：

[1] EPSHTEIN B, OFEK E, WEXLER Y. Detecting text in natural scenes with stroke width transform [C] //2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 2963 - 2970.

[2] YIN X C, YIN X, HUANG K, et al. Robust text detection in natural scene images [J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36 (5): 970 - 983.

[3] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector [C] //European conference on computer vision. Springer, Cham, 2016: 21 - 37.

[4] LIAO M, SHI B, BAI X, et al. Textboxes: A fast text detector with a single deep neural network [C] //Thirty-first AAAI conference on artificial intelligence, 2017.

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

[6] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks [J]. Advances in neural information processing systems, 2015, 28: 91 - 99.

[7] ZHOU X, YAO C, WEN H, et al. East: an efficient and accurate scene text detector [C] //Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017: 5551 - 5560.

[8] MILLETARI F, NAVAB N, AHMADI S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation [C] //2016 fourth international conference on 3D vision (3DV). IEEE, 2016: 565 - 571.

[9] HE P, HUANG W, HE T, et al. Single shot text detector with

regional attention [C] //Proceedings of the IEEE international conference on computer vision, 2017: 3047 - 3055.

[10] SHI B, BAI X, BELONGIS S. Detecting oriented text in natural images by linking segments [C] //Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 2550 - 2558.

[11] DENG D, LIU H, LI X, et al. Pixellink: Detecting scene text via instance segmentation [C] //Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32 (1).

[12] LONG S, RUAN J, ZHANG W, et al. Textsnake: A flexible representation for detecting text of arbitrary shapes [C] //Proceedings of the European conference on computer vision (ECCV), 2018: 20 - 36.

[13] 程　琦，王国栋，赵　毅. 基于分散注意力与路径增强特征金字塔的文本检测 [J]. 激光与光电子学进展，2020，57 (24)：249 - 257.

[14] 李　敏，郑建彬，詹恩奇，等. 基于文本像素颜色聚类的场景文本检测算法 [J]. 激光与光电子学进展，2019，56 (7)：139 - 146.

[15] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks [C] //Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794 - 7803.

[16] HUANG Z, WANG X, HUANG L, et al. Ccnet: Criss-cross attention for semantic segmentation [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 603 - 612.

[17] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C] //Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 7132 - 7141.

[18] ZHANG H, ZU K, LU J, et al. EPSANet: An efficient pyramid squeeze attention block on convolutional neural network [C] //Proceedings of the Asian Conference on Computer Vision, 2022: 1161 - 1177.

[19] CH′NG C K, CHEN C S. Total-text: A comprehensive dataset for scene text detection and recognition [C] //2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2017, 1: 935 - 942.

[20] KARATZAS D, GOMEZ-BIGORDA L, NICOLAOU A, et al. ICDAR 2015 competition on robust reading [C] //2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2015: 1156 - 1160.

[21] WANG P, ZHANG C, QI F, et al. A single-shot arbitrarily-shaped text detector based on context attended multi-task learning [C] //Proceedings of the 27th ACM international conference on multimedia, 2019: 1277 - 1285.

[22] JIANG X, XU S, ZHANG S, et al. Arbitrary-shaped text detection with adaptive text region representation [J]. IEEE Access, 2020, 8: 102106 - 102118.

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

[24] XU Y, WANG Y, ZHOU W, et al. Textfield: Learning a deep direction field for irregular scene text detection [J]. IEEE Transactions on Image Processing, 2019, 28 (11): 5566 - 5579.

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

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

[27] WANG W, XIE E, LI X, et al. PAN++: Towards Efficient and Accurate End-to-End Spotting of Arbitrarily-Shaped Text [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[28] HE W, ZHANG X Y, YIN F, et al. Deep direct regression for multi-oriented scene text detection [C] //Proceedings of the IEEE International Conference on Computer Vision, 2017: 745 - 753.

[29] WANG W, XIE E, SONG X, et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 8440 - 8449.

(上接第 20 页)

[30] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-tuned salient region detection [C] //2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009: 1597 - 1604.

[31] TERRIER. The Addon Repository for the 3D Space Simulator Celestia [EB/OL]. https://www.celestiamotherlode.net/catalog/spacecraft.php, 2017.

[32] AUTODESK. 3ds Max: Create massive worlds and high-quality designs [EB/OL]. 2022 [2022-10-13]. https://www.autodesk.com/.

[33] ZHANG W, CONG M Y, WANG L P. Algorithm for optical weak small targets detection and tracking: review [C] //IEEE International Conference on Signal and Image Processing, Maui, USA, Aug. 23 - 25, 2003: 643 - 647.

[34] STOKES G, VO C, SRIDHARAN R, et al. The space—based visible program [C] // Space Conference & Exposition, 2013.

[35] PAYKARI P, STARCK J L, FADILI M J. True cosmic microwave background power spectrum estimation [J]. Astronomy and Astrophysics, 2012, 541.

[36] LENDERMANN M, TAN J S Q, KOH J M, et al. Computational Imaging Prediction of Starburst-Effect Diffraction Spikes [J]. Scientific reports, 2018, 8 (1): 16919.