计算机科学
COMPUTER SCIENCE

基于改进区域候选网络的场景文本检测

李俊林, 欧阳智, 杜逆索

引用本文

李俊林, 欧阳智, 杜逆索.基于改进区域候选网络的场景文本检测[J]. 计算机科学, 2023, 50(2): 201-208.

LI Junlin, OUYANG Zhi, DU Nisuo. Scene Text Detection with Improved Region Proposal Network[J]. Computer Science, 2023, 50(2): 201-208.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211000191

基于改进区域候选网络的场景文本检测

李俊林[1] 欧阳智[2] 杜逆索[1,2]
1 贵州大学计算机科学与技术学院 贵阳 550025
2 贵州大学贵州省大数据产业发展应用研究院 贵阳 550025
(729741445@qq.com)

摘 要 自然场景中的文本图像具有十分复杂多变的特征,使用区域候选网络(Region Proposal Network,RPN)提取文本矩形位置候选框是不可或缺的一个步骤,能够极大地提升文本检测的精度。然而最近的研究表明,通过最小化平滑的 L_1 损失函数来回归矩形候选框中心点、宽和高的方式容易产生边界信息缺失、回归不准确等问题。针对这一问题,提出了一种基于改进区域候选网络的场景文本检测模型。首先,使用残差网络和特征金字塔网络组成的骨干网络生成共享特征图。然后,使用改进的回归取点方式和基于顶点的 VIOU 损失函数(Vertex-IOU)在共享特征图上生成系列文本矩形候选框。接着,使用 ROI Align 将这些候选框转化为固定大小的特征图在全连接层进行边界框预测。最后,在 ICDAR2015 数据集上进行对比实验,结果表明,与其他模型相比,所提模型可以提升检测精度,证明了所提模型的有效性。

关键词:深度学习;场景文本检测;区域候选网络;回归方式;损失函数
中图法分类号 TP391

Scene Text Detection with Improved Region Proposal Network

LI Junlin[1],OUYANG Zhi[2] and DU Nisuo[1,2]
1 College of Computer Science and Technology,Guizhou University,Guiyang 550025,China
2 Guizhou Big Data Academy,Guizhou University,Guiyang 550025,China

Abstract Scene text images have very complex and changeable features. Using region proposal network(RPN) to extract text rectangle position candidate boxes is an indispensable step,which can greatly improve the accuracy of text detection. However,recent studies show that the methods of regressing the center point,width and height of the text rectangular candidate boxes by minimizing the smooth L_1 loss function would easily cause problems such as missing boundary information and inaccurate regression. Therefore,this paper proposes a scene text detection model based on improved region proposal network. First,the backbone network composed of the residual network and the feature pyramid network is used to generate a shared feature map. Then,an improved regression method and vertex-based loss function(Vertex-IOU) are used to generate a series of text rectangular candidate boxes on the shared feature map. Finally,ROI Align is used to convert these candidate boxes into fixed-size feature maps for bounding box regression in the fully connected layer. Through comparative experiments on ICDAR2015 dataset,the results show that the test effect is improved compared with other models,which proves the effectiveness of our model.

Keywords Keywords deep learning,Scene text detection,Region proposal network,Regression method,Loss function

1 引言

场景文本检测(Scene Text Detection)是检测自然场景中的文本区域并用包围框标记文本区域的任务,准确的检测结果将有利于在文本识别任务中从文本框内提取文字。文本检测与识别对于理解场景内容具有重要的意义,因而具有广泛的现实应用,如智能检测系统、基于内容的图像搜索系统、场景理解、产品搜索及无人驾驶系统等。目前,场景文本检测已成为计算机视觉、模式识别、文档分析与识别等领域的一个热点研究方向[1-3]。

传统的场景文本检测方法主要使用滑动窗口方法,依赖手工设计的特征生成文本候选区域,并使用随机森林(Random Forest)和支持向量机(Support Vector Machine,SVM)等机器学习方法对文本框位置进行定位。这些方法依赖于手工设计的特征,无法有效提取到文本图像的关键边界信息,因此定位准确率较低。

到稿日期:2021-10-26 返修日期:2022-03-18
基金项目:贵州省科学技术厅重大科技计划项目(黔科合重大专项字[2018]3002);贵州大学培育项目(贵大培育[2020]41号)
This work was supported by the Major Scientific and Technological Special Project of Guizhou Province China([2018]3002) and Cultivation Project of Guizhou University([2020]-41).
通信作者:杜逆索(nsdu@gzu.edu.cn)

近些年，以 VGGNet[4]，ResNet[5] 和 FCN[6] 等网络结构结合特征金字塔网络(Feature Pyramid Network，FPN)为基础的深度学习模型相继被提出。如 Xuan 等[7]将高层先验语义和低级特征相结合，提出了一种新颖的基于高层先验语义的显著目标检测算法模型，其能够高效地识别图像的显著目标区域。相比于传统方法，这些模型的一个巨大优势在于，通过组合低层特征形成了更加抽象的高层特征。这不仅能有效地提取到文本区域的语义信息，还能将高层信息中缺失的位置信息在底层特征中提取出来，避免了手工提取特征过程中的繁琐低效，大大提高了检测效率。此外，相比于传统方法中采用滑动窗口生成文本矩形候选框的方式，目前广泛使用的 RPN[8] 能获取到更加精准的文本候选区域，且能节省大量的成本。

这些方法主要可以分为两类：基于图像分割的场景文本检测方法和基于回归的场景文本检测方法。目前，由于大多数据集并不存在像素级别的标注，且基于图像分割的方法后续处理较为复杂，而基于回归的方法通常能节省大量的成本，因此本文将围绕基于回归的场景文本检测方法展开。

针对文本检测图像中的极端长文本、弯曲文本，基于回归方法中的 RPN 网络通常以 3 种宽高比(2∶1，1∶1，1∶2)搭配 3 种大小尺度(128，256，512)在共享特征图中的每个特征点上生成 9 个 anchor box。然后在两个全连接层上分别进行分类和回归训练，分类的目的在于划分文本与背景。回归的目标则是优化以中心点、宽、高表示的文本候选区域。传统 RPN 网络通过预定义大量的候选框基本能完整地覆盖文本区域，但在回归过程中采取回归中心点、宽、高的方式容易缺失单个文本实例部分关键的边界点信息，压缩矩形区域候选框，使获取到的矩形区域候选框不能完整地包含单个文本实例(如图 1 所示)。因此，传统 RPN 网络回归时的取点方式亟待改进。

图 1 传统 RPN 网络检测效果图

Fig. 1 Detection effect of traditional RPN network

另一方面，目前 RPN 网络广泛使用的平滑的 L_1 损失函数并非根据场景文本检测评估规则定制，这使得具有相同 IOU 值的几对边界框可能具有不同的 L_1 范数损失，因此最小化平滑的 L_1 损失函数与提升预测边界框与真实边界框的 IOU 值并没有强大的正相关性。在优化过程中，一个损失函数的选择决定了文本区域候选框能否更好、更快地与目标真实框拟合。虽然 Yu 等[9]提出的 IOU 和 Rezatofighi 等[10]提出的 GIOU 都稍微改善了这个问题，但仍存在回归不准确的问题。

针对这两个问题，本文提出了基于改进区域候选网络的场景文本检测模型。首先，构建一个残差网络 ResNet50 和特征金字塔网络组成的骨干网络，在解码层中增加一个横向连接来提取多尺度目标的强语义特征。然后，在区域候选网络中使用多点(本文实验中采用了 13 个点)回归的方式优化基于顶点的 VIOU 损失函数(Vertex-IOU)生成矩形候选框。具体地，边界框上的 12 个点用于文本矩形位置候选框边界上的局部区域信息优化，而中心点则用于文本矩形位置候选框的总体位置信息定位。最后，使用 ROI Align 将这些候选框转化为固定大小的特征图，在全连接层使用不同的 ROI head 进行边界框预测。

本文的主要贡献有以下两点：

(1)在区域候选网络中提出多点回归的检测方式，与传统 RPN 网络相比，其能更好地解决卷积神经网络感受视野较小的问题，能提取到更加丰富的边界信息。实验证明，其能有效扩大单个文本实例的检测范围。

(2)提出了基于顶点的 VIOU 损失函数(Vertex-IOU)，通过直接优化预测边界框与真实边界框顶点之间的归一化距离来更好、更快地提升预测边界框与真实边界框之间的 IOU 值。相比于传统的平滑的 L_1 损失函数，其实现了更好的性能，并且可以广泛用于一般场景文本检测方法中。

2 相关工作

目前大多数场景文本检测的研究方法主要是基于深度学习网络来开展的，这些方法可以分为两类：基于图像分割的场景文本检测方法和基于回归的场景文本检测方法。基于回归的场景文本检测方法将文本视为一般的目标对象，通过调整目标检测框架来进行文本框检测，最后通过回归的方式来得到文本位置信息，这些方法通常采用两阶段的方式来回归文本区域的位置信息。第一阶段训练 RPN 网络生成一系列的文本区域候选框，第二阶段训练目标区域检测网络来对文本区域候选框进行分类和边界细化。例如，Tian 等[11]提出的 CTPN(Connectionist Text Proposal Network)采用垂直候选框划分的方法，通过自顶向下的原则对场景文本进行预测。它主要借鉴了 Faster R-CNN 中 RPN 的思想，将文本检测问题转换成定位细粒度文本提议，使精度得以提升。Ma 等[12]提出了旋转区域建议网络(Rotation Region Proposal Network，RRPN)，通过生成旋转提议框来适应多方向文本检测，并加入旋转感兴趣区域(Rotation Region-of-Interest，RRoI)池化层来更好地拟合文本检测区域，减少非文本区域冗余，确保文本检测的计算效率得到提升。这些方法在应对尺度差异较小的文本时效果不错，但对于极端长文本，这些方法不能有效处理文本尺度差异的问题。

与一般目标检测的对象相比，场景文本尺度变化大得多，有些是极端长文本，有些是弯曲文本，因此针对文本检测图像中的场景文本尺度差异问题，Zhang 等[13]提出了一种可以逐步调整文本检测框的网络模型。该模型采用了类似于 EAST(An Efficient and Accurate Scene Text Detector)[14]的直接回归网络，以每像素的方式预测字符或文本框，但其感受野依旧很小，无法包含到完整文本区域，因此在迭代优化模块中该模型

引入了角落注意力机制来回归每个角落的坐标偏移，在同一感受野内通过不断迭代来感知更准确的边界信息。而 Baek 等[15]提出了一种基于单字符(Character)的文本检测网络。该模型不以整个文本检测框为目标，而是通过检测单个字符以及字符间的联系来确定最终的文本框，通过解决卷积神经网络感受视野较小的问题来实现长文本检测。综上，这些方法都只考虑了文本图像的单一特征，并没有考虑文本图像的多尺度特征，因而无法提取到文本图像完整的边界信息。

另一方面，目前在 RPN 网络中普遍使用的平滑的 L_1 损失函数存在收敛慢、回归不准确等问题。为了使候选框能更好地与真实边界框拟合，Yu 等[9]提出在目标检测框架中使用 IOU 损失函数来取代普遍使用的平滑的 L_1 损失函数。但 IOU loss 在预测边界框与真实边界框不相交的情况下并不会提供任何移动梯度，对此 Rezatofighi 等[10]提出了一个惩罚项 $\frac{|C|-A\cup B}{|C|}$(其中 A 表示预测边界框，B 表示真实边界框，C 表示两框的最小外接矩形，$|C|$ 表示该最小外接矩形面积)用于优化两框不相交的情况。然而，当预测框位于真实框中时，这个惩罚项将不起作用，该损失函数将退化为 IOU，并不能直接优化两框之间的位置信息。

综上所述，本文主要针对深度学习网络中传统回归方式无法完整、准确地包含单个文本实例的问题，提出了基于改进区域候选网络的场景文本检测模型。与大多数模型一样，本文采用两阶段的体系结构，使用区域候选网络来生成矩形候选框。但有别于传统 RPN 网络，一方面，本文使用了多点(本文实验中采用了 13 个点)回归的方式来取代传统方法中采用中心点、宽和高的回归方式，能有效改善卷积神经网络感受视野较小的问题；另一方面，本文提出了一种基于顶点的 VIOU 损失函数(Vertex-IOU)，通过直接优化预测边界框与真实边界框顶点之间的归一化距离，来提升预测边界框与真实边界框之间的 IOU 值。实验表明，相比于其他模型，本文提出的模型能给出更加精确的文本候选区域，有效提升了文本检测的精度。

3 基于 IRPN 的场景文本检测模型

3.1 主要框架

本文提出的模型框架如图 2 所示，包括改进的区域候选网络(Improved Region Proposal Network，IRPN)和 Vertex-IOU。首先，采用 ResNet50 和 FPN 组成的骨干网络生成共享特征图，其中既包含底层特征中的位置信息，又包含高层特征中的语义信息。然后，通过 IRPN 在共享特征图上生成若干初始候选区域。接着，采用多点的边界回归方式和 VIOU 损失函数对初始候选区域进行分类和回归训练，得到文本矩形候选框。最后，使用 Roi Align 将文本矩形候选框在共享特征图上转化为固定大小的特征图，并根据不同数据特性，使用不同的 ROI head 来对边界框进行细化。

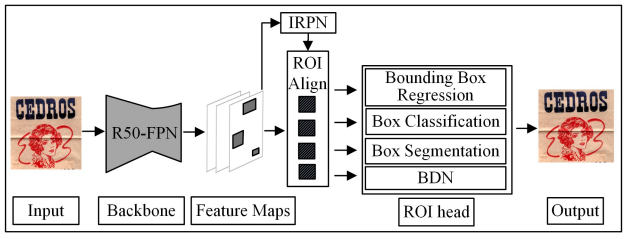

图 2 所提模型整体框架图

Fig. 2 Overall framework of the proposed model

3.2 IRPN 模块

RPN 网络(如图 3 所示)在现有的场景文本检测方法中得到了广泛的应用。在传统方法里，它首先在共享特征图上使用滑动窗口以特征图的每一个点进行滑动，每个点根据 3 种宽高比(2∶1，1∶1，1∶2)搭配 3 种大小的尺度(128，256，512)生成 9 个 anchor box $\{X_{\min}^t, Y_{\min}^t, X_{\max}^t, Y_{\max}^t\}$(即由该矩形框左上角和右下角坐标点来表示)，并将其编码为 $\{X_t, Y_t, \omega_t, h_t\}$(即由该矩形框中心点坐标($x$，$y$)、宽($\omega$)和高($h$)来表示)格式进行回归训练，训练的目标是优化 *Smooth* L_1 *Loss*(式(1))。然后，预测一个 4 维的回归向量 $\{\boldsymbol{\Delta x}, \boldsymbol{\Delta y}, \boldsymbol{\Delta \omega}, \boldsymbol{\Delta h}\}$，通过式(2)来优化当前边界框 $\{X_C, Y_C, \omega_C, h_C\}$，得到一个预测的边界框 $B_{\text{pred}}=\{X_{\text{pred}}, Y_{\text{pred}}, \omega_{\text{pred}}, h_{\text{pred}}\}$。

$$Smooth_{L_1}(x)=\begin{cases}0.5\,x^2, & |x|<1\\ |x|-0.5, & |x|\geqslant 1\end{cases} \tag{1}$$

$$\begin{cases}X_{\text{pred}}=X_c+\omega\boldsymbol{\Delta x}\\ Y_{\text{pred}}=Y_c+h\boldsymbol{\Delta y}\\ \omega_{\text{pred}}=\omega_c\,e^{\boldsymbol{\Delta x}}\\ h_{\text{pred}}=h_c\,e^{\boldsymbol{\Delta y}}\end{cases} \tag{2}$$

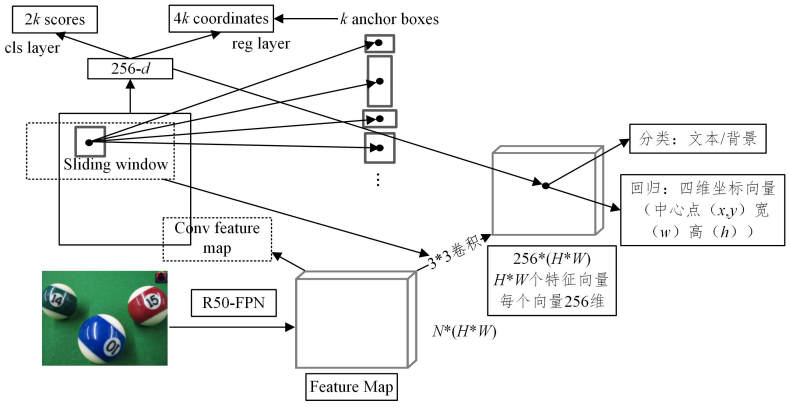

图 3 传统 RPN 回归示意图

Fig. 3 Schematic diagram of traditional RPN regression

实验证明，采取回归中心点、宽、高的方式容易缺失部分关键的边界点信息，压缩矩形区域候选框，使获取到的矩形区域候选框不能完整地包含单个文本实例，从而影响检测精度。因此，本文首先根据文本图像相比于其他一般目标检测图像所具有的纵横比普遍较大或较小的特点，在编码阶段对中心点赋予更高的权重，试图降低因纵横比差异过大带来的检测不准确的不良效果。另外，本文提出了一种新的回归取点方式，使 RPN 网络在训练时可以参考到语义更加重要的区域，从而回归到更完整的矩形候选框。即先定义一组初始点 $Z=\{X_t,Y_t\}_{t=0}^{12}$（共计 13 点），如图 4 所示，这 13 个点分别对应矩形框的中心点、4 个边界点以及边界上的三等分点。每个点的优化方法如式(3)所示：

$$Z=\{X_t,Y_t\}_{t=0}^{12}=\{(X_t+\omega_c\boldsymbol{\Delta X}_t,Y_t+h_c\boldsymbol{\Delta Y}_t)\}_{t=0}^{12} \tag{3}$$

其中，$\{X_t,Y_t\}_{t=1}^{12}$ 表示从矩形边界框左上角顶点开始，以逆时针顺序进行编号的点，特别地，$\{X_0,Y_0\}$ 表示中心点。$\{\boldsymbol{\Delta X}_t,\boldsymbol{\Delta Y}_t\}_{t=0}^{12}$ 表示对当前框的预测偏移量，ω_c 和 h_c 表示当前框的宽和高。这 13 个点都采用式(3)进行优化，其中边界框上的 12 个点用于文本矩形位置候选框边界上的局部区域信息优化，而中心点则用于文本矩形位置候选框的总体位置信息定位。然后使用式(4)对这些点进行处理，得到文本矩形候选框 $Proposal=\{X_{\text{pred}}^{\min},Y_{\text{pred}}^{\min},X_{\text{pred}}^{\max},Y_{\text{pred}}^{\max}\}$。

$$\begin{cases}X_{\min}=\min(\{X_t\}_{t=1}^{12}),X_{\max}=\max(\{X_t\}_{t=1}^{12})\\Y_{\min}=\min(\{Y_t\}_{t=1}^{12}),Y_{\max}=\max(\{Y_t\}_{t=1}^{12})\\\text{if}((X_{\max}-X_{\text{ctr}})>(X_{\text{ctr}}-X_{\min}))\\X_{\min}=2*X_{\text{ctr}}-X_{\max}\\\text{else:}\\X_{\max}=2*X_{\text{ctr}}-X_{\min}\\\text{if}((Y_{\max}-Y_{\text{ctr}})>(Y_{\text{ctr}}-Y_{\min}))\\Y_{\min}=2*Y_{\text{ctr}}-Y_{\max}\\\text{else:}\\Y_{\max}=2*Y_{\text{ctr}}-Y_{\min}\\Proposal=\{X_{\text{pred}}^{\min},Y_{\text{pred}}^{\min},X_{\text{pred}}^{\max},Y_{\text{pred}}^{\max}\}\\\quad=\{X_{\min},Y_{\min},X_{\max},Y_{\max}\}\end{cases} \tag{4}$$

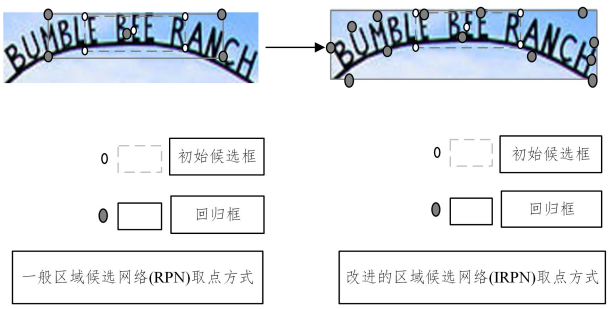

图 4 传统 RPN 与 IRPN 取点方式的比较

Fig. 4 Comparison of point taking methods of traditional RPN and IRPN

与传统的 RPN 网络相比，IRPN 网络中采用的新的取点方式能更加周全地考虑文本矩形位置候选框边界上的局部区域信息，在处理极端长文本、弯曲文本时可以对边界点实现更加精细的定位，从而使生成的文本矩形位置候选框能更完整、准确地包围单个文本实例。

3.3 Vertex-IOU 损失函数

在传统 RPN 中，L_n（$n=1$ 或 2）范式的损失函数被广泛应用为 Bounding Box 回归，该方法不是根据场景文本检测评估规则定制的，因此不是提升预测盒和目标盒之间 IOU 值的最佳选择；同时，采用 L_n 范数损失优化的 RPN 网络对尺度变化十分敏感，具有相同 IOU 值的几对边界框可能具有不同的 L_n 范数损失，如图 5 所示，使用平滑的 L_1 损失函数计算得出 3 组框的损失值均为 8.41，然而 3 组框回归的情况并不一致，最右侧一组 L_{IOU} 值最小，回归效果也最好，这说明使用平滑的 L_1 损失函数与提升回归框质量并没有强大的正相关性。

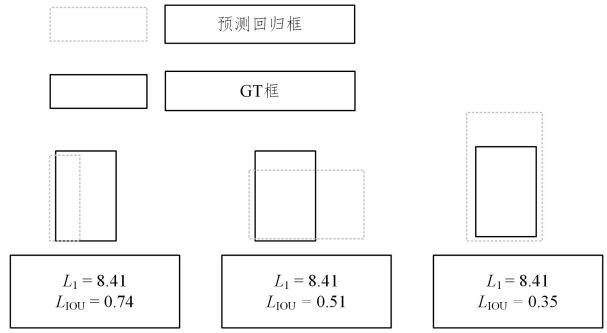

图 5 L_1 范数损失与 IOU 值变化示意图

Fig. 5 Schematic diagram of smooth L_1 loss and IOU change

为了解决这一问题，Yu 等[9]将 IOU loss 用作提升 IOU 度量的损失函数，如式(5)所示：

$$L_{\text{IOU}}\ loss=1-\frac{B\cap B^{gt}}{B\cup B^{gt}} \tag{5}$$

其中，B 为预测边界框，B^{gt} 为真实边界框。然而，IOU loss 在预测边界框与真实边界框不重叠时为 0，此时并不会提供任何移动梯度，IOU loss 将不会收敛。针对以上问题，本文提出一个新的损失函数 VIOU(Vertex-IOU)来优化预测边界框与真实边界框的 IOU 值。VIOU 定义如式(6)所示：

$$L_{\text{VIOU}}\ loss=1-IOU+R_B \tag{6}$$

其中，R_B 为一惩罚项，用于直接优化预测边界框与真实边界框之间的位置信息。R_B 定义如式(7)所示：

$$\begin{cases}V_L^2=(X_L^{GT}-X_L^t)^2+(Y_L^{GT}-Y_L^t)^2\\V_R^2=(X_R^{GT}-X_R^t)^2+(Y_R^{GT}-Y_R^t)^2\\\mathcal{R}_B=\dfrac{V_L^2}{C^2}+\dfrac{V_R^2}{C^2}\end{cases} \tag{7}$$

其中，$V_L{}^2$ 表示预测边界框与真实边界框左上角顶点的欧氏距离，$V_R{}^2$ 表示预测边界框与真实边界框右下角顶点的欧氏距离，C^2 表示预测边界框与真实边界框最小外接矩形对角线的距离。因此，$\frac{V_L^2}{C^2}$ 和 $\frac{V_R^2}{C^2}$ 分别表示预测边界框与真实边界框左上角、右下角顶点之间的归一化距离。VIOU 示意图如图 6 所示，在两框不相交时，IOU 值为 0，此时并不提供任何移动梯度，而 VIOU 损失函数的惩罚项则可以通过优化两顶点间的归一化距离来推动两框进行相交。当两框相交甚至处于包含关系时，该惩罚项能对两顶点间的归一化距离进行优化；此外，由于预测框与真实框纵横比不一致时也会影响两框之间的顶点距离，该惩罚项还能对预测框的纵横比进行调整，使之更加匹配真实边界框。

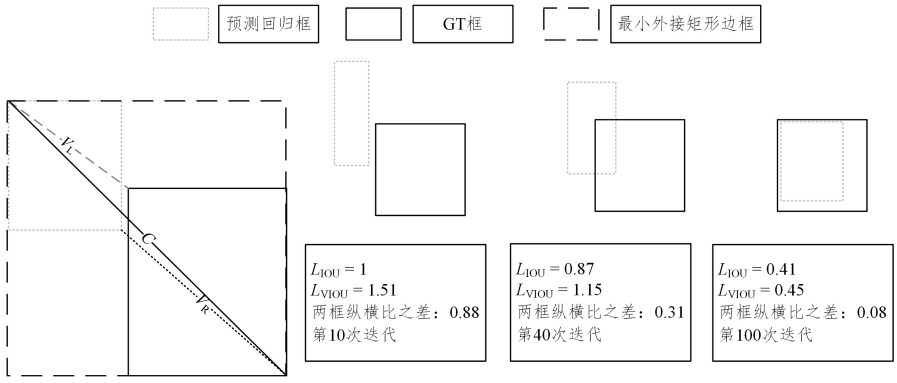

图 6 VIOU 示意图

Fig. 6 Schematic diagram of vertex-based loss function

3.4 后处理

在使用 IRPN 网络生成系列文本矩形候选框后，本文首先使用 Roi Align 在共享特征图上将这些文本矩形候选框转化为固定大小的特征图，然后使用不同的 ROI head 来对边界框进行细化。针对具有一定方向的文本，本文使用 Bounding Box Regression 和 Box Classification 两个分支进行分类和回归训练，并使用 Box Segmentation 分支生成图像的 mask，通过 NMS 算法提取其中具有较强响应的边界点生成最终的四边形预测框。另外，为了缓解本文模型可能出现标注顺序敏感的问题，提升对关键边界的学习能力，本文最终模型在生成系列文本矩形候选框后引入了 Liu 等[16]提出的 BDN 方法来生成更加准确、有序的四边形文本预测框。

4 实验与结果分析

4.1 数据集及评价指标

本文在 ICDAR2015 数据集上进行评估，以测试所提方法在多方向文本中的性能。ICDAR2015 数据集是最受欢迎的面向场景文本检测的基准数据集之一，其中的图片主要从街道和购物中心拍摄，因此这个数据集的挑战依赖于定向、小和低分辨率的文本。ICDAR2015 数据集包含 1 000 个训练样本和 500 个测试样本，其中有大约 2000 个可识别的四边形字级边界框。

ICDAR2015 数据集使用精确率（Precision，P）、召回率（Recall，R）和调和平均值（H-means）3 种评估指标来评估模型在数据集上的检测效果。具体地，当检测结果矩形框与 Ground-truth 矩形框之间的 IOU 值大于 0.5 时，则认为该检测结果正确，反之则认为是错误的。精确率（Precision，P）和召回率（Recall，R）分别定义为：

$$P=\frac{|T_P|}{E} \tag{8}$$

$$R=\frac{|T_P|}{T} \tag{9}$$

其中，T_P，E，T 分别表示正确的检测结果集合、检测结果集合以及 Ground-truth 集合。调和平均值（H-means）则定义为：

$$H\text{-}means=\frac{2PR}{P+R} \tag{10}$$

4.2 实现环境与参数

本文提出的模型在 Mask R-CNN[17] Pytorch 版本基础上实现，由于 ICDAR2015 数据集标注样本较少，直接对该数据集训练容易导致过拟合。Huang 等[18]提出在具有大量标注信息的基类数据集上训练后的网络应用于只有少量标注样本的新类数据集得到的模型，将具有更好的泛化能力，本文采用了该训练方法。首先在 SynthText[19] 数据集中使用 Adam 优化器进行预训练，然后在 Nayef 等[20] 提出的 ICDAR2017MLT 数据集上进行 fine-tune，最后在 ICDAR2015 数据集上训练，后面两阶段使用 SGD 优化器进行训练。另外，本文只使用了官方提供的训练图像在 4 块 NVIDIA TITANXP GPU 上开展训练，数据增强方式使用了随机旋转、随机水平翻转和随机裁剪。训练时总迭代次数不超过 20 000 次，初始学习率设置为 1×10^{-2}，在迭代次数为 10 000 次和 15 000 次时分别下降到 1×10^{-3} 和 1×10^{-4}，动量（Momentum）设置为 0.9，重量衰减设置为 0.000 1。训练图像的短边设置为（680，720，760，800，840，880，920，960，1 000），最大尺寸限制为 1 480。在训练阶段，被标记为“DO NOT CARE”的模糊文本将被自动忽略，不参与训练。在测试阶段，本文采用单尺度方法进行测试（便于与其他方法进行公平比较），尺度和最大尺寸设置为（1 200，1 600），并通过官方的评估协议来评估实验结果。此外，本文也使用了 Liu 等[21]提出的多边形非极大值抑制（Polygon Non-Maximum Suppression，PNMS）来抑制冗余检测，本文中这一阈值设置为 0.25。

4.3 消融实验

为了验证 IRPN 网络和基于顶点的 VIOU 损失函数的有效性，本节在 ICDAR2015 数据集中开展了消融实验，所有的模型均只采用官方图像进行训练和测试。在 ICDAR2015 数据集中的消融实验的结果如表 1 所列。

表 1 ICDAR2015 数据集上的消融实验

Table 1 Ablation experiments on ICDAR2015 dataset

（单位：%）

Method	*Recall*	*Precision*	*H-mean*
传统 RPN	80.6	86.8	83.5
IRPN	81.8	87.6	84.6
RPN+VIOU	80.5	88.7	84.4
IRPN+VIOU	82.8	88.8	85.4
OURS(IRPN+VIOU+BDN)	85.6	89.8	87.6

从表 1 中可以看到，相较于原始 RPN 网络，IRPN 的 *H-means* 提升了 1.1%。本文分析其中一个原因在于 IRPN

网络通过多点回归的方式能更加有效地学习到文本矩形位置候选框边界上的局部区域信息，并且由于给中心点分配了足够高的权重，因此能够对文本矩形位置候选框实现更加精准的定位，使生成的文本矩形位置候选框能更完整、准确地包围单个文本实例，从而提升检测精度。另外，传统 RPN 网络在应用了 VIOU 损失函数后，*H-means* 提升了 0.9%，该结果表明 VIOU 相较于其他两种损失函数能更好地提升两框之间的 IOU 值，不管两框能否相交都能推动预测框向目标框移动。当使用 IRPN 结合 VIOU 损失函数时 *H-means* 提升了 1.9%，在使用 SynthText 数据集进行预训练并使用 BDN 方法提升模型对关键边的学习能力时 *H-means* 则能提升 4.1%。实验结果表明所提模型具有强大的泛化能力，即使是在处理低分辨率、噪声多的 ICDAR2015 数据集时也能实现良好的检测效果。

4.4 与其他模型的结果比较分析

为了验证本文模型的效果，将其与基于图像分割的方法（EAST[14]、Lyu 等[22]、PixelLink[23]、PAN[24]、TextDragon[25]、TextField[26]、PSENet[27]、Richard 等[28]、Zhang 等[29]、DB[30]、EFPN[31]）以及基于回归的方法（MSR[32]、CRAFT[15]、SegLink[33]、Liao 等[34]、CTPN[11]、BDN[16]、ContourNet[35]、Xie 等[36]、Mask R-CNN[17]）进行比较，实验结果如表 2 所列。

表 2 在 ICDAR2015 数据集上的实验结果

Table 2 Results on ICDAR2015 dataset

（单位：%）

类别	方法	第一阶段是否对 RPN 改进	第二阶段（采用方法）	*Recall*	*Precision*	*H-mean*
基于图像分割的方法	EAST[14]	—	—	73.5	83.6	78.2
	Lyu 等[22]	—	—	70.7	94.1	80.7
	PixelLink[23]	—	—	81.7	82.9	82.3
	PAN[24]	—	—	77.8	82.9	80.3
	TextDragon[25]	—	—	81.8	84.8	83.1
	TextField[26]	—	—	83.9	84.3	84.1
	PSENet[27]	—	—	84.5	86.9	85.7
	Richard 等[28]	—	—	85.4	83.1	84.2
	Zhang 等[29]	—	—	83.2	87.7	85.4
	DB[30]	—	—	88.2	82.7	85.4
	EFPN[31]	—	—	89.2	82.0	85.5
基于回归的方法	MSR[32]	—	—	78.4	86.6	82.3
	CRAFT[15]	—	—	84.3	89.8	86.9
	SegLink[33]	—	—	73.1	76.8	75.0
	Liao 等[34]	无	RRD	79.0	85.6	82.2
	CTPN[11]	无	双向 LSTM	74.0	52.1	61.0
	BDN[16]	无	BDN	83.8	89.4	86.5
	ContourNet[35]	无	LOTM	86.1	87.6	86.9
	Xie 等[36]	无	LACC	86.0	87.0	87.0
	Mask R-CNN[17]	无	MASK	80.6	86.8	83.5
	OURS	有	BDN	**85.6**	**89.8**	**87.6**

注：为了公平比较，本表仅列出了单尺度实验结果，并未列出涉及识别的方法

从表中的 *Precision* 来看，本文模型仅低于 Lyu 等[22]的模型，但该模型的 *H-means*（80.7%）显著低于本文模型（87.6%），说明本文所提的 VIOU 损失函数相比其他模型使用方法能更好地优化文本矩形候选框与 Ground-truth 之间的距离，从而提升检测精度。对表中的 *Recall* 进行比较分析，可以看到本文模型的 *Recall* 为 85.6%，低于 Xie 等[36]（86.0%）、ContourNet[35]（86.1%）、DB[30]（88.2%）和 EFPN[31]（89.2%）模型，但这几种方法在取得良好 *Recall* 值的同时 *Precision* 表现不佳，在 *Precision* 上分别低于本文模型 2.8%，2.2%，7.1%，7.8%。综合来看，本文模型的 *Precision* 为次优，*Recall* 表现一般，但 *H-means* 为最优。进一步分析，与同类型基于回归的模型相比，本文模型是在针对文本检测两阶段方法中第一阶段文本区域候选框进行改进。如果只进行第一阶段的比较，本文模型取得的 85.4% 的 *H-means* 优于目前其他的方法（BDN[16] 为 83.5%，ContourNet[35] 为83.9%），在第二阶段引入 BDN[16] 模型后同样优于其他方法，证明了本文方法对提取文本候选框的有效性，以及在处理低分辨率数据集时的强大泛化能力。该数据集的可视化结果如图 7 所示。

图 7 在 ICDAR2015 数据集上的检测结果

Fig. 7 Detection results on the ICDAR2015 dataset

结束语 针对传统 RPN 网络回归矩形候选框时容易产生边界信息缺失、回归不准确等问题，本文提出了一种改进的区域候选网络模型。首先，提出了新的回归取点方式，通过多点回归试图降低因纵横比差异过大带来的检测不准确的不良效果，并对文本矩形位置候选框边界上的局部区域信息进行有效学习来解决文本尺度差异问题。然后，提出了一种新的

基于顶点的 VIOU 损失函数（Vertex-IOU），通过直接优化预测边界框与真实边界框顶点之间的距离来提升两框之间的 IOU 值，从而提升整体的检测精度。最后，在 ICDAR2015 数据集上，与其他模型的实验结果进行比较分析，结果表明所提模型针对传统 RPN 网络进行的改进能更充分地学习文本框边界上的局部区域信息，从而实现更好的检测效果。因此，本文模型不仅适用于具有传统 RPN 网络的场景文本检测模型，而且所提出的 VIOU 损失函数也可以用于一般场景文本检测边界盒回归方法，体现了本模型具有较强的泛化性。

参 考 文 献

[1] WANG R M，SANG N，DING D，et al. Text Detection in Natural Scene Image：A Survey [J]. Acta Automatica Sinaca，2018，44(12)：2113-2141.

[2] MIAO Y Q，LIU S Q，ZHANG W Z，et al. Chinese text detection algorithm in natural sceneimages[J]. Computer Engineering and Design，2018，39(3)：804-807，818.

[3] JIANG W，ZHANG C S，YIN X C. Deep Learning Based Scene Text Detection：ASurvey [J]. Acta Electronica Sinica，2019，47(5)：1152-1161.

[4] SIMONYAN K，ZISSERMAN A. Very DeepConvolutional Networks for Large-Scale Image Recognition[C]// Proceedings of the International Conference on Learning Representations. San Diego：2015.

[5] HE K M，ZHANG X Y，REN S Q，et al. Deep Residual Learning for Image Recognition[C]// Proceedings of the 2016 IEEE Conference onComputer Vision and Pattern Recognition. Las Vegas：IEEE，2016：770-778.

[6] LONG J，SHELHAMER E，DARRELL T. Fully Convolutional Networks for Semantic Segmentation[C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston，Massachusetts：IEEE，2015：3431-3440.

[7] XUAN D D，WANG J，WANG Z. Salient target detection based on high-level priori semantics[J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition)，2020，32(2)：304-312.

[8] ROSS G. Fast R-CNN[C]// Proceedings of the 2015 IEEE International Conference on ComputerVision. Santiago，Chile：IEEE，2015：1440-1448.

[9] YU J H，JIANG Y N，WANG Z Y，et al. UnitBox：An Advanced Object Detection Network[C]// Proceedings of the 2016 ACM Multimedia Conference. Amsterdam：2016：516-520.

[10] REZATOFIGHI H，TSOI N，GWAK J，et al. Generalized intersection over union：A metric and a loss for bounding box regression[C]// Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach，CA：IEEE，2019：658-666.

[11] TIAN Z，HUANG W，HE T，et al. Detecting Text in Natural Image with Connectionist Text Proposal Network[C]// Proceedings of the 14th European Conference on Computer Vision. Amsterdam，2016：56-72.

[12] MA J Q，SHAO W Y，YE H，et al. Arbitrary-Oriented Scene Text Detection via RotationProposals[J]. arXiv：1703. 01086，2017.

[13] ZHANG C Q，LIANG B R，HUANG Z M，et al. Look More Than Once：An Accurate Detector forText of Arbitrary Shapes [C]// Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach，CA：IEEE，2019：10552-10651.

[14] ZHOU X，YAO C，WEN H，et al. EAST：An Efficient and Accurate Scene Text Detector[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，Hawaii：IEEE，2017：2642-2651.

[15] BEAK Y，LEE B，HAN D，et al. Character Region Awareness for Text Detection[C]// Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach，CA：IEEE，2019：9365-9374.

[16] LIU Y L，ZHANG S，JUN L W，et al. Omnidirectional scene text detection with sequential-free box discretization[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao：2019：3052-3058.

[17] HE K M，GEORGIA G，PIOTR D，et al. Mask R-CNN[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice：IEEE，2017：2980-2988.

[18] HUANG D，CHEN Z，FENG X. Object detection method based on graph convolution net under limitedsamples[J]. Journal of Chongqing University of Technology(Natural Science)，2022，36(6)：172-180.

[19] ANKUSH G，ANDREA V，ANDREW Z. Synthetic Data for Text Localisation in Natural Images[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas：IEEE，2016：2315-2324.

[20] NIBAL N，FEI Y，IMEN B，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：2017：1454-1459.

[21] LIU Y L，JIN L W，ZHANG S T，et al. Detecting Curve Text in the Wild：New Dataset and NewSolution[J]. arXiv：1712. 02170，2017.

[22] LYU P Y，YAO C，WU W H，et al. Multi-Oriented Scene Text Detection via Corner Localization and Region Segmentation [C]// Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition，Salt Lake City. Utah：IEEE，2018：7553-7563.

[23] DENG D，LIU H F，LI X L，et al. PixelLink：Detecting Scene Text via Instance Segmentation[C]// Proceedings of the 32th AAAI Conference on Artificial Intelligence. New Orleans，Louisiana：2017：6773-6780.

[24] WANG W H，XIE E Z，SONG X G，et al. Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network[C]// Proceedings of the 2019 IEEE International Conference on Computer Vision. Seoul：IEEE，2019：8439-8448.

[25] FENG W，HE W H，YIN F，et al. TextDragon：An End-to-End Framework for Arbitrary Shaped Text Spotting[C]// Proceedings of the 2019 IEEE International Conference on Computer Vision. Seoul：IEEE，2019：9075-9084.

[26] XU Y C, WANG Y K, ZHOU W, et al. TextField: Learning a Deep Direction Field for Irregular Scene Text Detection[J]. arXiv: 1812.01393, 2018.

[27] WANG W H, XIE E Z, LI X, et al. Shape Robust Text Detection With Progressive Scale Expansion Network[C]// Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA: IEEE, 2019: 9336-9345.

[28] RICHARDSON E, AZAR Y, AVIOZ O, et al. It's All About The Scale-Efficient Text Detection Using Adaptive Scaling [C]// Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision. Aspen, Colorado: IEEE, 2020: 1844-1853.

[29] ZHANG L, LIU Y, XIAO H, et al. Efficient Scene Text Detection with Textual Attention Tower[C]// ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020: 4272-4276.

[30] LIAO M, WAN Z, YAO C, et al. Real-Time Scene Text Detection with Differentiable Binarization[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: 2020: 11474-11481.

[31] SHAO H L, JI Y, LIU C P, et al. Scene Text Detection Algorithm Based on Enhanced Feature Pyramid Network[J]. Computer Science, 2022, 49(2): 248-255.

[32] XUE C H, LU S J, ZHANG W. MSR: Multi-Scale Shape Regression for Scene Text Detection[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao: 2019: 989-995.

[33] SHI B G, BAI X, SERGE J B. Detecting Oriented Text in Natural Images by Linking Segments[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii: IEEE, 2017: 3482-3490.

[34] LIAO M H, ZHU Z, SHI B G, et al. Rotation-Sensitive Regression for Oriented Scene Text Detection[C]// Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah: IEEE, 2018: 5905-5918.

[35] WANG Y X, XIE H T, ZHA Z J, et al. ContourNet: Taking a Further Step toward Accurate Arbitrary-shaped Scene Text Detection[C]// Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 11750-11759.

[36] XIE B H, QIN Y L, ZHANG Y J. Scene Text Detection Based on Learning Active Center ContourModel[J]. Computer Engineering, 2022, 48(3): 244-252, 262.

LI Junlin, born in 1997, postgraduate. His main research interests include computer vision and scene text detection.

DU Nisuo, born in 1986, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include simulation and data science.

(责任编辑:柯颖)