DOI:10.19651/j.cnki.emt.2106770

基于注意力机制的自然场景文本检测*

宋彭彭 曾祥进 郑安义 米 勇
(武汉工程大学 计算机科学与工程学院 武汉 430205)

摘 要: 针对自然场景文本检测中没有明确全局特征的重要性,导致文本检测过程中存在文本的误检、漏检问题,提出了基于注意力机制的自然场景文本检测方法。该方法在 CTPN 网络的基础上,利用 ResNet 网络及特征融合技术提取更深层次的多层网络文本特征;同时将注意力机制引入改进后的特征提取网络中,通过从所有位置聚集的相同特征来增强原始特征,并获取注意力权重,对全局注意力进行汇集,明确需要关注的特征。其次,针对自然场景下文本定位精度不高的问题,使用 GIoU 损失代替坐标损失,同时引入 Focal Loss 损失函数对原有损失函数进行改进。实验表明,该方法在自然场景文本图片检测中获得了 83%的召回率、87%的准确率和 85%的 F 值,保证了文本检测过程中文本信息的完整性。

关键词: 文本检测;CTPN;ResNet;注意力机制;GIoU;Focal Loss
中图分类号: TP391 **文献标识码:** A **国家标准学科分类代码:** 520.6040

Text detection in natural scenes based on attention mechanism

Song Pengpeng Zeng Xiangjin Zheng Anyi Mi Yong
(School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China)

Abstract: Aiming at the fact that the importance of global features is not clear in the text detection of natural scenes, which leads to the misdetection and missed detection of text in the text detection process, a natural scene text detection method based on attention mechanism is proposed. Based on the CTPN network, this method uses the ResNet network and feature fusion technology to extract deeper multi-layer network text features; at the same time, the attention mechanism is introduced into the improved feature extraction network, which is enhanced by the same features gathered from all positions the original features, and the attention weight is obtained, the global attention is collected, and the features that need attention are clarified. Secondly, for the problem of low text positioning accuracy in natural scenes, GIoU loss is used instead of coordinate loss, and the Focal Loss loss function is introduced to improve the original loss function. Experiments show that this method obtains a recall rate of 83%, a precision rate of 87% and an F value of 85% in the text image detection of natural scenes, which ensures the integrity of the text information in the text detection process.

Keywords: text detection;CTPN;ResNet;attention mechanism;GIoU;Focal Loss

0 引 言

在自然场景中准确获取文本信息有很多实际应用,如文本分析、信息检索、语言翻译等,因此它成为了计算机视觉领域的研究热点之一[1-2]。这项工作由文本检测与识别两部分组成,相对于识别而言文本检测任务更具有挑战性,检测的准确性将直接影响后期的识别,因此本文将研究重点放在文本检测上。准确检测文本位置面临的主要问题在于复杂的背景和文本模式的差异[3],自然场景文本由于自身特点(如:文字方向多样、尺度大小不一、排列方式杂乱等)的复杂,传统的自上而下的文本检测方法[4]已经难以满足当前的需求。近年来,目标检测在深度学习领域快速发展,受其影响文本检测方法也进入了新的研究阶段。但是直接将目标检测用于文本检测得到的结果很不理想,主要的原因在于:1)文本方向不定和文字自身特征使得通用的目标检测框很难应用于文本检测;2)文本尺度大小不一造

收稿日期:2021-05-24
*基金项目:国家自然科学基金项目(61502354)、湖北省教育厅重点研究项目(D20171503)、武汉工程大学研究生教育创新基金项目(CX2020214)资助

成文本行检测难度提升;3)文本排列类别不同造成文本检测区域区分的特征难以得到。

针对这些问题,提出了大量的深度学习解决方法用于对已有网络进行改进。其中,主要的方法有:1)基于回归文本框的文本检测。如:CTPN[5]网络模型的提出,解决了文本尺度频繁变化的问题,主要采用拆分大文本行为多个小文本行最后使用文本先构造算法将其连接的方法,特征提取方面是在 Faster R-CNN[6]的基础上为了使网络学习文本的序列特征加入了 BiLSTM 使得对于文本序列的检测效果显著提升。2)基于语义分割的文本检测。如:EAST[7]网络模型的提出,解决了基于回归文本框方法中候选框与真正框匹配过程中的耗时缺点,实现了文本检测在速度和准确率上的提高。CTPN 网络模型中采用 VGG16 作为基础的网络来提取文本特征,在特征提取时对特征重要性没有明确的表示,而且提取特征耗费大量时间,最终在检测上会产生许多的误检和漏检情况。另外由于网络深度的限制,对抽象特征的提取得不到提高,获得的语义信息的性能也比较差,使得文本检测的判别和预测能力难以加强。

针对 CTPN 网络中存在的不足,本文设计了新的网络模型结构,使用 ResNet50 作为基础特征提取网络,获得语义信息更强的深层次文本特征,利用特征融合得到不同网络层次的特征,引入注意力机制明确特征提取过程中的重点特征。其次为增加自然环境中的定位精度,在损失函数中加入边框回归的广义交并比损失函数(generalized-IoU,GIoU)[8],同时使用 Focal Loss[9]损失函数代替原有损失函数。本文利用改进后的网络在 CTPN 的基础上研究,针对特征提取网络进行改进提升了获取文本检测框的准确性,对损失函数进行改进提高了文本定位精度。与 CTPN 相比,本文提升了文本检测信息的完整性。

1 CTPN 算法

CTPN 算法进行文本检测过程如下:利用 VGG16 获取输入图像的 Conv5 特征图,然后使用 3×3 的滑动窗口在输出的特征图中密集移动得到 256 维的特征向量。为了适应各种尺度的文本框,设计 10 个固定宽度为 16 pixel,高度从 11~273 pixel(每次除 0.7)变化的锚框。检测中单独考虑每个独立的文本框可能会导致在非文本对象上的误检,为了提高定位的精度,将文本行分割成一系列小的文本框,文本框具有很强的序列特征,利用其序列上下文信息极大地方便了文本的检测。因此使用双向的 LSTM 对两个方向的文本进行循环上下文编码。最后将其输入到全连接层产生输出。CTPN 的网络模型结构如图 1 所示。

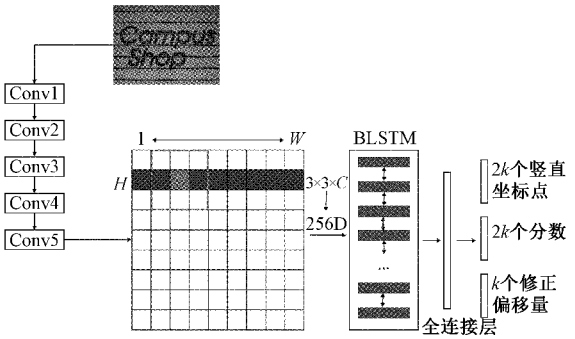


图 1 网络模型结构

2 基于注意力机制的文本检测方法

2.1 注意力模块

计算机视觉中全局上下文特征的建模发挥着重要的作用,为了明确图像的重点特征需要对其全局特征进行有效的建模,SENet[10]、GENet[11]和 PSANet[12]通过重新缩放不同的信道,获取全局特征与信道的关系。注意力模块 CBAM[13]则通过对不同的通道和空间实行重新校准,使网络明白需要提取哪些特征。但是这些方法实现特征融合均采用的是重新缩放,这样的全局特征建模不能满足当前文本检测的需求。

GCNet[14]利用 NLNet[15]对全局特征的有效建模,使其可以明确需要关注的特征;利用 SENet 的轻量级属性,使其可以方便的插入用于处理计算机视觉问题的各种网络中。而且对于各种检测任务,如目标检测、文本检测等,GCNet 可以获得比 NLNet 和 SENet 更好的效果。

全局上下文块的具体框架如图 2 所示,公式如下:

$$z_i = x_i + \boldsymbol{W}_{v2} ReLU\left(LN\left(\boldsymbol{W}_{v1}\sum_{j=1}^{N_p}\frac{e^{\boldsymbol{W}_k x_j}}{\sum_{m=1}^{N_p} e^{\boldsymbol{W}_k x_m}} x_j\right)\right) \quad (1)$$

其中,$\alpha_j = \dfrac{e^{\boldsymbol{W}_k x_j}}{\sum_m e^{\boldsymbol{W}_k x_m}}$ 表示汇集全局注意力的权重,$\delta(\cdot) = \boldsymbol{W}_{v2} ReLU(LN(\boldsymbol{W}_{v1}(\cdot)))$ 表示瓶颈变换。由框架图可知,GCNet 块主要包含:1)汇集全局注意力对上下文特征进行建模;2)通过瓶颈变换获取通道的依赖;3)采用广播添加方法实现特征融合。

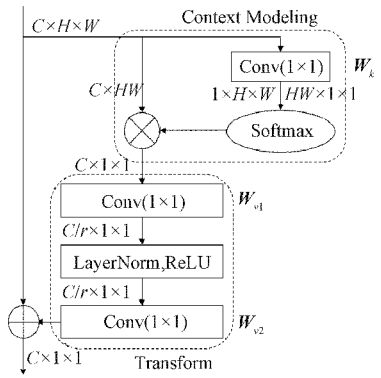


图 2 全局上下文块

为了获得 SENet 的轻量级属性,GCNet 利用瓶颈变换

模块代替 1×1 卷积，使得模块从原本拥有 $C \cdot C$ 数量的参数变为 $2 \cdot C \cdot C/r$，其中 r 是瓶颈比率，C/r 表示瓶颈的隐藏维度。然而，随着瓶颈变换的增加，对模块的优化变得困难，因此，在 ReLU 之前加入了层规范化用以简化优化，并将其作为一个正则化器进行推广。这对于对象检测和实例分割起着至关重要的作用。由此得到的 GCNet 模块既可以对全局特征建模，又拥有轻量级属性，使得该模块可以应用到每一个卷积层，获取更好的依赖关系。

2.2　特征提取网络的改进

为了使 CTPN 文本检测网络在特征提取过程中能够得到更深层次的语义信息以及明确重点文本特征，本文在 CTPN 算法的基础上改进了特征提取网络并引入了 GCNet 注意力机制，形成新的文本检测网络，其核心框架如图 3 所示。

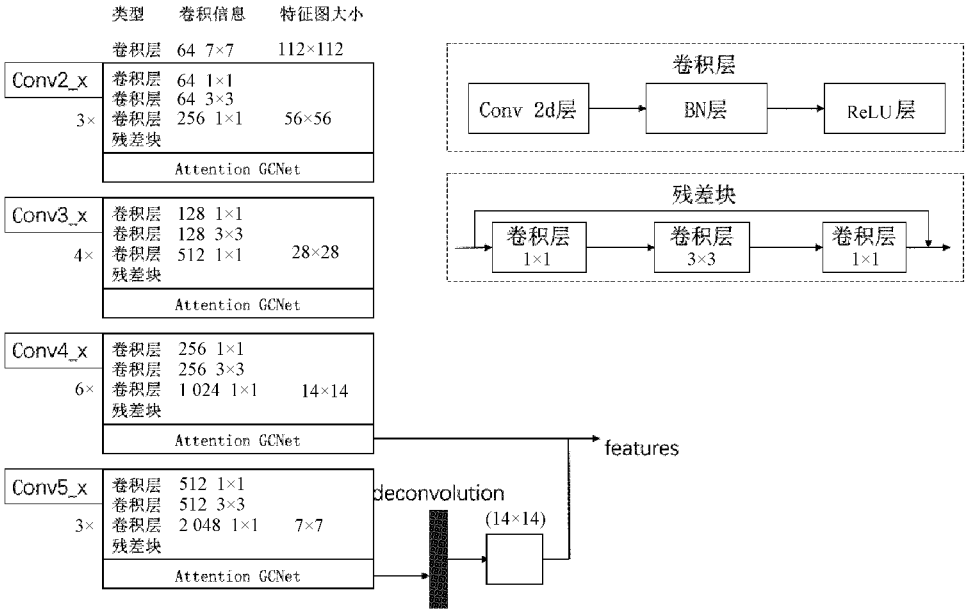


图 3　改进后的算法结构

在利用 ResNet[16] 网络进行文本下采样过程中，为了得到不同网络层次之间的特征，对 ResNet 网络进行了多层特征融合，采用特征融合技术对 ResNet50 中的 Conv5_x 使用反卷积，将得到的特征与 Conv4_x 的特征进行拼接，获得含有更丰富信息的特征，提高了文本检测的准确率。为了明确网络中的特征信息，在网络中引入注意力模块 GCNet，对每一层输出的文本特征使用 1×1 卷积和 softmax 函数获取注意力权重，然后进行注意力汇集，获取全局上下文特征，之后再使用 1×1 卷积进行特征变换，最后使用加法将全局上下文特征聚合到每个位置的特征，实现更有效地全局上下文建模。

2.3　损失函数改进

在神经网络中，损失函数是评价网络预测值与真实值误差的重要依据，通过损失函数反向传播来更新网络参数，损失函数的改进，有助于获得更好的预测效果。CTPN 主要使用了两种损失函数，在预测框坐标回归中使用的是 SmoothL1 损失函数，在类别概率上使用了交叉熵损失函数。公式如下：

$$L(\boldsymbol{s}_i,\boldsymbol{v}_j,\boldsymbol{o}_k)=\frac{1}{N_s}\sum_i L_s^{cl}(\boldsymbol{s}_i,\boldsymbol{s}_i^*)+\frac{\lambda_1}{N_v}\sum_j L_v^{re}(\boldsymbol{v}_j,\boldsymbol{v}_j^*)+\frac{\lambda_2}{N_o}\sum_k L_o^{re}(\boldsymbol{o}_k,\boldsymbol{o}_k^*) \tag{2}$$

其中，$\boldsymbol{s}_i$、$\boldsymbol{v}_j$、$\boldsymbol{o}_k$ 表示网络预测输出；$\boldsymbol{s}_i^*=\{0,1\}$、$\boldsymbol{v}_j^*$、$\boldsymbol{o}_k^*$ 表示真值标签；λ_1 和 λ_2 表示平衡不同任务的损失权重，根据实验设置为 1.0 和 2.0；N_s、N_v、N_o 为标准化参数，表示对应任务使用的样本数量。

CTPN 使用 SmoothL1 损失函数作为回归预测损失会造成两个检测框的损失值一致，但是效果却区别很大。如图 4 所示，其中虚线框为真实物体目标框，实线框为预测框，可以看到 3 个目标框 SmoothL1 损失相同时，目标框与真实框的 IoU 区别很大。在目标检测中，IoU 可以作为一种距离测量标准，如果直接使用 IoU 计算损失函数，会出现两个问题：1）当目标框和真实框没有相交时，IoU 为 0，此时不能反映两框之间的距离，同时损失函数为 0，没有梯度回传，无法进行参数更新。2）当目标框和真实框存在平移旋转时，IoU 相同，但两框重合度不同。针对以上问题，提出使用 $GIoU$ 来计算预测框坐标回归损失函数，公式如下：

$$GIoU=IoU-\frac{C-U}{C} \tag{3}$$

$GIoU$ 是一种距离测量标准。其中，C 为目标框和真实框最小外包面积，U 为目标框和真实框覆盖的总面积，当 IoU 值为 0 时，$GIoU$ 的值依然存在，且 C 会根据目标框和真实框的变化而变化，从而很好解决直接使用 IoU 计

算损失函数的问题。$GIoU$ 来计算损失函数的公式如下:

$$L_{GIoU} = 1 - GIoU \tag{4}$$

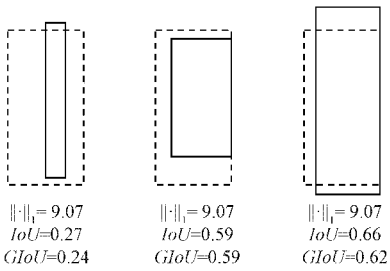


图 4 3 种情况示意图

在文本检测过程中受到复杂背景的影响,会存在大量的难分样本,使用交叉熵损失函数会使各个样本的权重一样,占总的损失值中多的是容易分的样本,因此模型优化的方向并不是本文所希望的那样。为了提高模型检测的准确率,使用 Focal Loss 损失函数代替 CTPN 中的类别损失函数。Focal Loss 损失是在交叉熵损失的基础上修改而来,公式如下:

$$FL(p) = \begin{cases} -\alpha(1-p)^{\gamma}\log(p), y = 1 \\ -(1-\alpha)p^{\gamma}\log(1-p), y = 0 \end{cases} \tag{5}$$

其中,α 为控制正负样本参数;$p \in [0,1]$ 为对应标签的模型的估计概率;y 为实际标签值;$y \in \{0,1\}$,γ 为控制难易样本参数,始终大于 0。当 p 越大时,$(1-p)^{\gamma}$ 越小,从而减少了大概率目标的损失贡献,加强了网络对难分目标的学习。改进后的网络损失函数公式如下:

$$Loss = \sum L_{GIoU} + \sum FL(p) \tag{6}$$

改进后的损失函数使用 $GIoU$ 损失作为预测框坐标回归损失,使用 Focal Loss 损失函数代替 CTPN 中的类别损失函数,解决了文本定位精度不高和背景干扰问题,提高了网络检测的准确率。

3 实验与分析

3.1 实验数据集及评估指标

本次实验使用文本检测基准数据集 ICDAR2013[17] 对本文方法进行有效评估,ICDAR2013 共有 462 张真实场景图像,分为 229 张训练集和 233 张测试集。图像中的文字能很好地聚焦且多数是水平的。

对数据集 ICDAR2013 本文采用 DetEval 的评估方法,主要通过 3 种情况来考虑标记框与检测框是否匹配,分别是 1 对 1、1 对多和多对 1。最后计算精确度、召回率和 F 值来验证模型的有效性。精确度、召回率和 F 值分别定义为:

$$Precision(\overline{G},\overline{D},t_r,t_p) = \frac{\sum_k \sum_j Match_D(D_j^k, G^k, t_r, t_p)}{\sum_k |D^k|} \tag{7}$$

$$Recall(\overline{G},\overline{D},t_r,t_p) = \frac{\sum_k \sum_i Match_G(G_i^k, D^k, t_r, t_p)}{\sum_k |G^k|} \tag{8}$$

$$F-measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

其中,$G^k \in \overline{G}, k = 1,\cdots,N$ 表示真实文本框;$D^k \in \overline{D}$,$k = 1,\cdots,N$ 表示检测文本框;$t_r \in [0,1]$ 表示对区域召回率的约束;$t_p \in [0,1]$ 表示对区域精度的约束;文献中 $t_r = 0.8, t_p = 0.4$;$Match_G$ 和 $Match_D$ 表示对真实文本框和检测文本框的不同类型的匹配函数。

3.2 改进方法结果对比

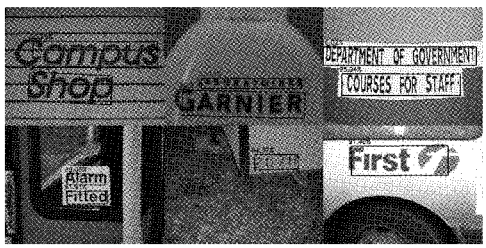
在网络训练过程中,相同的环境下采用不同的改进方法进行对比实验,其实验结果如表 1 所示,在不同的方法中,打"√"的地方表示使用了该方法。在改进算法 1 中利用 ResNet50 进行文本特征提取,使得特征提取网络获得更深层次的文本特征,将算法的准确率提升 0.03,召回率提升 0.02;为了获得不同层次的网络之间的特征,在改进算法 2 中对 ResNet50 的 Conv4_x 与 Conv5_x 进行特征融合,改进后的算法的准确率提升 0.04,召回率提升 0.04;为了明确特征提取过程中需要的特征,改进算法 3 引入了注意力机制,将全局上下文特征聚合到每个位置的特征,实现更有效地全局上下文建模,改进后的算法的准确率提升 0.08,召回率提升 0.05;为了提高目标的定位精度,在改进算法 4 中对损失函数进行改进,使用改进后的损失函数进行训练,改进后的算法的准确率提升 0.09,召回率提升 0.07。

表 1 不同改进方法对算法性能的提升

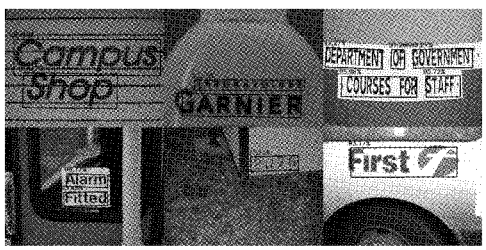
方法	ResNet50	特征融合	注意力机制	改进损失函数	召回率	准确率	F 值
CTPN					0.76	0.78	0.77
改进算法 1	√				0.78	0.81	0.79
改进算法 2	√	√			0.80	0.82	0.81
改进算法 3	√	√	√		0.81	0.86	0.83
改进算法 4	√	√	√	√	0.83	0.87	0.85

在相同环境下，对不同特征网络及不同注意力机制进行对比实验。

实验 1：使用 ResNet50 替换 VGG16，并对 Conv5_x 进行反卷积与 Conv4_x 特征融合作为文本特征提取网络与 CTPN 进行对比，实验结果如图 5 所示。通过检测结果可以发现 ResNet50 检测的结果更好，表明了 ResNet50 在特征提取上强于 VGG16，而且它的参数量相较于 VGG16 更少，降低了计算时间。



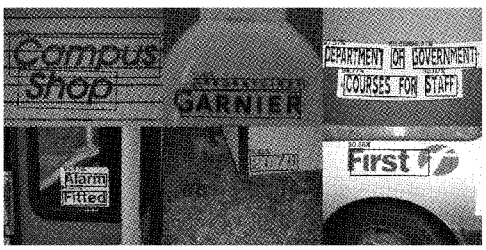
(a) VGG16



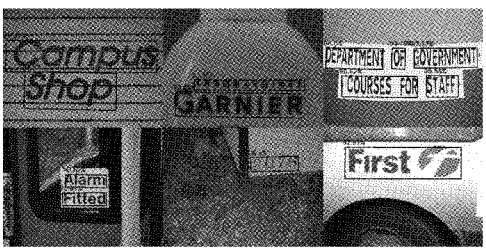
(b) ResNet50

图 5　不同网络实验结果对比图

实验 2：在替换特征提取网络的基础上添加注意力模块(GCNet)与添加注意力模块(CBAM)进行对比，实验结果如图 6 所示。通过检测结果可以发现，CBAM 仍然存在少量的漏检问题，且检测出来的文本区域较大，而 GCNet 注意力模块可以有效对全局上下文建模，减少文本检测的漏检问题同时也完整地检测出文本区域。



(a) 使用CBAM注意力机制



(b) 本文方法

图 6　不同注意力机制实验结果对比图

3.3　实验结果评估

通过对比实验发现，本文提出的在特征提取网络上的改进对于自然场景文本检测取得了较好的结果。为了验证本文方法与其他算法的检测性能，本次实验使用 ICDAR2013 数据集进行检测，实验结果如表 2 所示，本文方法的召回率为 0.83，准确率为 0.87，F 值为 0.85，与其他算法相比在召回率、准确率和 F 值上都得到了提升。

表 2　ICDAR2013 数据集检测对比结果

方法	Recall	Precision	F 值
文献[18]	0.73	0.81	0.77
文献[19]	0.79	0.83	0.81
文献[20]	0.80	0.83	0.81
文献[21]	0.79	0.85	0.82
本文方法	0.83	0.87	0.85

4　结　　论

本文提出了一种基于注意力机制的自然场景文本检测方法，该方法可以进行端到端的文本训练，有利于与文本识别一起训练，实现端到端的文本检测与识别。而且在加入 GCNet 模块后实现了对远程依赖的有效建模和轻量级计算，使得提取到的图片语义信息更丰富；引入 Focal Loss 损失函数，同时使用 GIoU 损失代替坐标回归损失，提高了自然场景下文本定位精度，解决了前景和背景不平衡问题。通过在不同的实验中发现，本文提出的方法都可以获得较好的结果，然而对于多方向的文本检测效果相对较差，因此后面将考虑进行多方向的文本检测。

参考文献

[1]　孙婧婧，张青林. 基于轻量级网络的自然场景下的文本检测[J]. 电子测量技术，2020，43(8)：101-107.

[2]　李英杰，全太锋，刘武启. 基于 MSER 的自适应学习自然场景文本检测[J]. 小型微型计算机系统，2020，41(9)：1966-1971.

[3]　王瑞，龙华，邵玉斌，等. 基于 Labeled-LDA 模型的文本特征提取方法[J]. 电子测量技术，2020，43(1)：141-146.

[4]　YIN X C，PEI W Y，ZHANG J，et al. Multi-orientation scene text detection with adaptive clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2015，37(9)：1930-1937.

[5]　TIAN Z，HUANG W L，HE T，et al. Detecting text in natural image with connectionist text proposal network [C]. European Conference on Computer Vision，Springer，Cham，2016：56-72.

[6]　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.

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

[9] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2980-2988.

[10] 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.

[11] HU J, SHEN L, ALBANIE S, et al. Gather-excite: Exploiting feature context in convolutional neural networks[J]. ArXiv Preprint,2018, ArXiv:1810.12348.

[12] ZHAO H S, ZHANG Y, LIU S, et al. Psanet: Point-wise spatial attention network for scene parsing [C]. Proceedings of the European Conference on Computer Vision(ECCV), 2018: 267-283.

[13] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3-19.

[14] CAO Y, XU J R, LIN S, et al. Gcnet: Non-local networks meet squeeze-excitation networks and beyond[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[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] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[17] KARATZAS D, SHAFAIT F, UCHIDA S, et al. ICDAR 2013 robust reading competition[C]. 2013 12th International Conference on Document Analysis and Recognition, IEEE, 2013: 1484-1493.

[18] CHNG C K, CHAN C S, LIU C L. Total-text: Toward orientation robustness in scene text detection [J]. International Journal on Document Analysis and Recognition(IJDAR), 2020, 23(1): 31-52.

[19] WANG X B, JIANG Y Y, LUO Z B, et al. Arbitrary shape scene text detection with adaptive text region representation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 6449-6458.

[20] XIE E, ZANG Y H, 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.

[21] YE J, CHEN Z, LIU J H, et al. TextFuseNet: Scene text detection with richer fused features[C]. IJCAI, 2020:516-522.

作者简介

宋彭彭,硕士研究生,主要研究方向为图像处理、深度学习、文本检测与识别。

E-mail:1170552818@qq.com

曾祥进,副教授,博士,主要研究方向为智能机器人控制、机器视觉、嵌入式系统设计、运动控制等。

E-mail:xjzeng21@163.com

郑安义,硕士研究生,主要研究方向为图像处理、计算机视觉、深度学习。

E-mail:313430434@qq.com

米勇,硕士研究生,主要研究方向为图像处理、机器视觉、口标检测。

E-mail:2532740174@qq.com