DOI:10.16136/j.joel.2022.12.0130

具有自校正与注意力机制相结合的场景文本检测

孙 鹏[1],刘 粤[1],强观臣[1],熊 炜[1,2,3*],付 尧[1],李利荣[1,2]

(1.湖北工业大学 电气与电子工程学院,湖北 武汉 430068;2.襄阳湖北工业大学产业研究院,湖北 襄阳 441003;3.美国南卡罗来纳大学 计算机科学与工程系,南卡罗来纳 哥伦比亚 29201)

摘要:在日常生活中,存在着丰富的文本信息,对这些信息的提取,能够极大地提高人们的生活品质。但自然场景中文本信息表达形式丰富多样,文本形状各异,在检测过程中存在误检、文本区域定位不准问题。针对以上不足,本文提出了一种具有自校正与注意力机制相结合的文本检测方法。首先,在ResNet50骨干网络中嵌入自校正卷积(self-calibrated convolution,SConv)及高效通道注意力(efficient channel attention,ECA),使网络能够校正全局无关信息的干扰,并集中关注于文本区域,提取更加丰富的语义信息;其次,在特征融合后加入协调注意力(coordinate attention,CA),纠正不同尺度的特征图在融合过程中产生的位置偏差。最后,通过修正后的特征图预测得到多个不同尺度的文本实例,采用渐进尺度扩展算法,求出最终检测到的文本实例。实验结果表明,在任意方向数据集ICDAR2015以及弯曲文本数据集Total-Text、SCUT-CTW1500上,相比于改进前的ResNet50综合指标F值分别提升了1.0%、5.2%、5.4%,证明了本方法具有良好的检测能力。

关键词:自校正卷积(SConv);高效通道注意力(ECA);协调注意力(CA);渐进尺度扩展算法

中图分类号:TP183 **文献标识码**:A **文章编号**:1005-0086(2022)12-1287-09

Scene text detection with self-calibration and attention mechanism

SUN Peng[1], LIU Yue[1], QIANG Guanchen[1], XIONG Wei[1,2,3*], FU Yao[1], LI Lirong[1,2]

(1. School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan, Hubei 430068, China; 2. Xiangyang Industrial Research Institute, Hubei University of Technology, Xiangyang, Hubei 441003, China; 3. Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29201, USA)

Abstract: In daily life, there are rich text information, the extraction of such information can greatly improve people's quality of life. However, there are various forms of text information expression and different text shapes in natural scenes, which result in false detection and inaccurate location of text regions. In order to solve these problems, this paper proposes a text detection method with self-calibration and attention mechanism. Firstly, the self-calibrated convolution (SConv) and efficient channel attention (ECA) are embedded in the backbone of ResNet50 to correct the interference of irrelevant global information and concentrate on the text area to extract more abundant semantic information Secondly, coordinated attention (CA) is added after feature fusion to correct the position deviation of feature map in different scale. Finally, several text instances of different scales are predicted by the modified feature map, and the final detected text instances are obtained by using the progressive scale expansion algorithm. The experimental results show that the comprehensive index F-measure is increased by 1.0%, 5.2% and 5.4% respectively compared with the unmodified ResNet50 on the arbitrary direction data set ICDAR2015 and the curved text data set Total-Text and SCUT-CTW1500. It is proved that this method has good detection ability.

Key words: self-calibrated convolutions (SConv); efficient channel attention (ECA); coordinate attention (CA); progressive scale expansion algorithm

* **E-mail**: xw@mail.hbut.edu.cn

收稿日期:2022-03-04 **修订日期**:2022-04-11

基金项目:国家自然科学基金(61571182,61601177)、国家留学基金(201808420418)、湖北省自然科学基金(2019CFB530)、湖北省科技厅重大专项(2019ZYYD020)和襄阳湖北工业大学产业研究院科研项目(XYYJ2022C05)资助项目

1 引　言

场景文本检测与识别技术不仅能迅速确定图中文本的具体位置、还可以从中提取包含的文本信息，为进一步获取更有价值的内容进行分析、理解提供有力支撑。该技术广泛应用于文本即时翻译、票据数据识别、文字图像检索与分类、智能机器人、无人驾驶、工业制造等场景。通常文本检测结果很大程度上决定着文本识别中是否能正确识别文字，因此将提升文本检测的准确性和精确性作为本文研究重点。

近几年，越来越多的科研人员将深度学习应用于自然场景文本检测，借鉴目标检测、语义分割、实例分割的方法，获得了较高的检测率、识别率以及泛化能力。目前，可以将这些采用深度学习的方法分为两类：基于回归和基于分割的文本检测方法。基于回归的方法是将图中文本看作同一类检测目标，在图像中预测多个文本候选框，然后进行分类和回归，直接检测出整个文本实例。采用该文本检测方法有：LOMO[1]（look more than once）、多尺度回归网络[2]（multi-scale regression，MSR）。它们都以目标检测的方式通过回归水平矩形框或不同方向的多边形实现文本检测，但检测结果往往包含较多无关的背景信息，尤其是在密集文本图像中难以分隔文本实例。而基于分割的文本检测方法则是先通过卷积神经网络得到每个像素为文本的概率，根据设定的阈值分割得到文本的基本组件，之后进行后处理，将组件逐步组成一个完整的文本实例，该方法能够很好解决密集文本问题。其主要方法有渐进尺度扩展网络[3]（progressive scale expansion network，PSENet）、像素聚集网络[4]（pixel aggregation network，PAN）、可微分二值化[5]（differentiable binarization，DB）。

由于自然场景中的文本在语种、颜色、字体、方向、尺寸大小上文本样式多种多样，有的文本图像背景极其复杂，甚至类似文本，加之在获取图像时受到光照、拍照技术影响，图像出现低对比度、低分辨率或有遮挡、伪影等现象，导致自然场景文本检测技术研究有很大的难度。因此，本文提出了一种具有自校正与注意力机制相结合的文本检测方法。首先在 ResNet50 网络结构上将原卷积拆分为 4 个小卷积、两个分支（校正分支、上下文语义继承分支）；在不引入额外的学习参数下，使得网络能够获得不同尺度的空间信息，通过校正分支对提取到的特征进行校正；同时在网络的每一层后引入高效通道注意力（efficient channel attention，ECA）机制，使网络检测的焦点集中于文本区域，提取更加丰富的语义信息；其次在特征融合后加入协调注意力（coordinate attention，CA）机制，修正不同尺度的特征图在自顶向下的融合过程中产生的位置偏差。最后通过卷积预测得到不同大小的文本实例核，并使用渐进尺度扩展算法，从最小核逐渐扩大到最大核，得到最终的文本实例。在模型训练过程中使用修改后的平方 *Dice* 系数损失函数，该损失可加大模型误检、漏检的惩罚力度，从而提升模型检测的准确率和召回率。

2 文本检测方法

本文提出的文本检测网络结构如图 1 所示，主要有 3 个部分：文本特征提取模块（feature extraction module）、文本特征增强模块（feature enhancement module）以及文本后处理模块（post processing module）。第一部分是提取特征，在原 ResNet50[6] 骨干网络的基础上，于 Bottleneck 中将自校正卷积[7]（self-calibrated convolutions，SConv）替换原网络中 3×3Conv 卷积，它仅仅考虑每个空间位置周围的上下文信息，避免全局上下文信息中无关区域的干扰；其次在每一层的Bottleneck末尾引入ECA[8]，在不

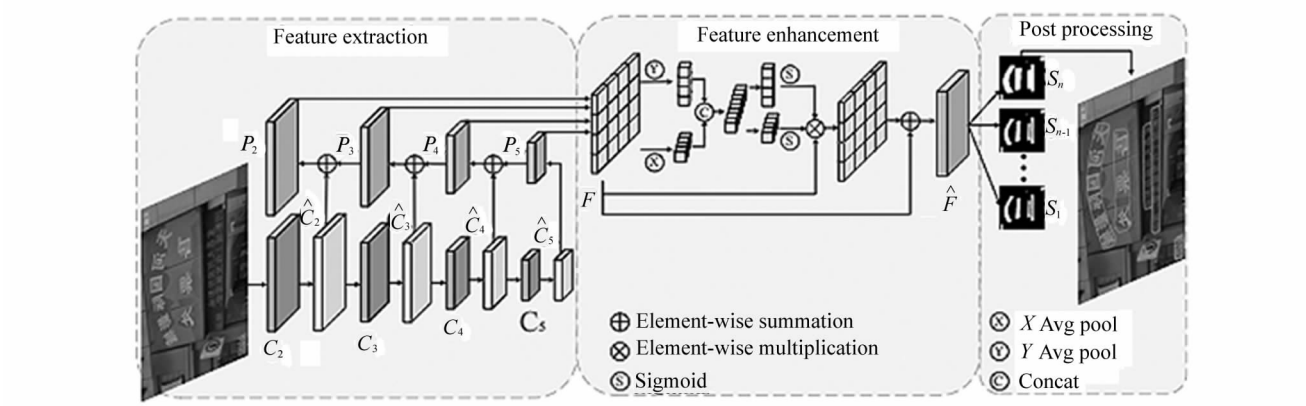

图 1　文本检测网络结构图

Fig. 1　Diagram of text detection network structure

降低特征图维度的情况下可交互通道信息,增强网络的特征提取能力。第二部分是特征增强,从骨干网络的每一层抽取 4 张不同尺度的特征图,采用特征金字塔网络[9](feature pyramid networks,FPN)的方式融合特征得到特征图 F;接着使用 CA[10],对原特征图进行增强得到新特征图,来增强文本边界信息,以扩大文本与非文本区域的区别度。第三部分则是求取文本检测结果,首先使用 1×1Conv 对特征图 $\hat{F}$ 进行卷积操作得到不同尺度的文本实例 S_1—S_n;然后通过渐进尺度扩展算法从最小尺寸文本实例 S_1 扩展到最大尺寸文本实例 S_n,得到最终的文本预测结果。

2.1　文本特征提取

在网络模型输入场景文本图像前,将图像进行随机亮度变换、随机旋转、缩放、裁剪等数据增强操作,得到 640×640×3 固定大小的图像信息。在将骨干网络中卷积替换为 SConv 后,于卷积层 C_2—C_5 后面嵌入 ECA,嵌入方式如图 2 所示,之后从每个 ECA 层提取不同尺度的特征图 $\hat{C}_2$、$\hat{C}_3$、$\hat{C}_4$、$\hat{C}_5$,其大小分别为原输入图像的 1/4、1/8、1/16、1/32。接着以 FPN 的方式自顶向下进行特征融合操作(见图 1):使用 1×1Conv 对顶层特征图 $\hat{C}_5$ 卷积得到特征图 P_5;而 P_2、P_3、P_4 由相同层级的 $\hat{C}_2$、$\hat{C}_3$、$\hat{C}_4$ 经过 1×1Conv 卷积操作后与上一层级的特征图 P_3、P_4、P_5 经过 2 倍上采样相加融合得到;最后将 P_2、P_3、P_4、P_5 进行 Concat 操作得到融合后的特征图 F。

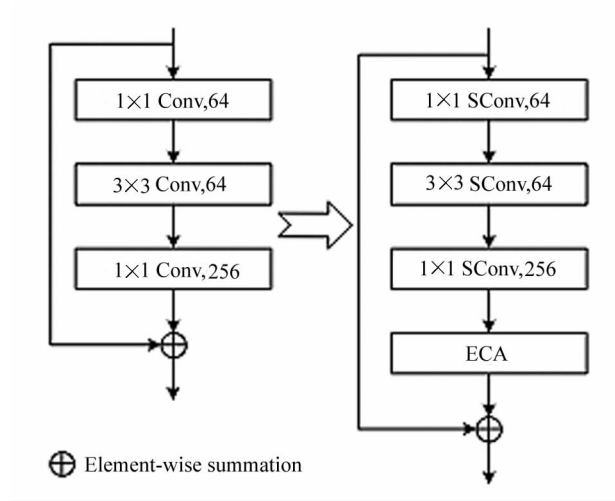

图 2　Bottleneck 结构图

Fig. 2　Bottleneck structure diagram

P_2、P_3、P_4 计算式如式(1)所示:

$$P_n = \mathrm{Conv}_{1\times1}(C_2) + \mathrm{Up}_{r2}(P_{n+1}), \tag{1}$$

式中,n 取 2,3,4,$\mathrm{Conv}_{1\times1}$ 表示 1×1 卷积,Up_{r2} 表示用双向线性插值法实现的 2 倍上采样。

2.1.1　SConv

本文提出使用 SConv 代替传统的卷积方式提高模型的文本特征提取能力,该卷积在两个不同尺度空间中进行卷积特征变换,于每个空间位置周围构建空间和通道之间的相关性,使每个卷积层的感受野变大,从而丰富语义信息。图 3 是骨干网络中使用的SConv模块,其操作过程可划分为以下3个步

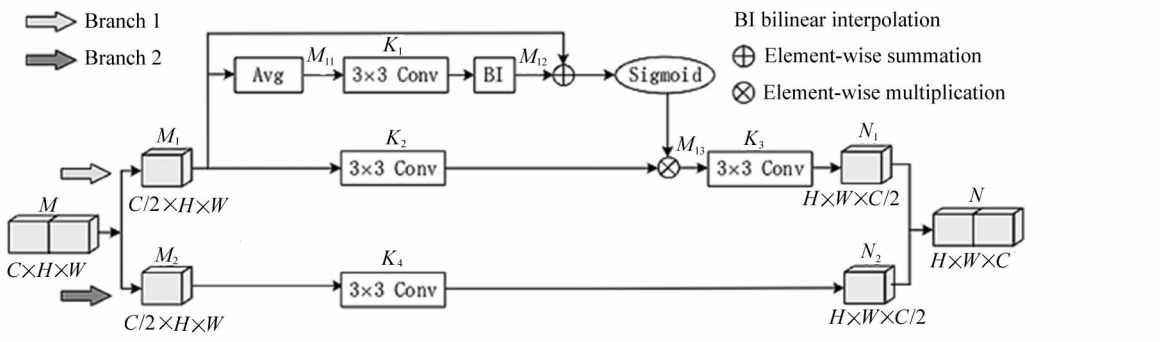

图 3　SConv 模块

Fig. 3　Self-calibrated convolutions module

骤:1) 将输入大小为 $C\times H\times W$ 的特征 M,分成 M_1、M_2,其大小均为 $C/2\times H\times W$;2) 将大小为(C,C,3,3)的原卷积核 K 分成 4 个小卷积核 K_1、K_2、K_3、K_4,其大小均为($C/2$,$C/2$,3,3),用来收集不同类型的上下文信息;卷积核 K_1、K_2、K_3 在分支 1 中对 M_1 进行自校正操作,得到 N_1;而卷积核 K_4 于分支 2 中对原尺度特征 M_2 进行卷积操作,以保留原始的空间背景信息,得到 N_2;3) 拼接原尺度空间输出特征 N_1、N_2,得到与输入 M 大小一致特征 N。在分支 1 的自校正处理中,首先使用大小为 4×4 平均池化层对特征 M_1 进行下采样,得 M_{11},如式(2)所示,再经过卷积核 K_1 提取特征,并进行上采样得 M_{12},将尺寸恢复到

M_1 大小，如式(3)所示，上采样方式为双线性插值法；后与 M_1 相加，使用 Sigmoid 激活函数映射计算出校正权重，并与卷积核 K_2 提取特征后相乘进行校正，得 M_{13}，如式(4)所示，最后通过 K_3 卷积操作得到该分支校正后的特征 N_1，如式(5)所示。

$$M_{11} = \mathrm{AvgPool}(M_1), \tag{2}$$

$$M_{12} = \mathrm{Up}_{r2}(F_1(M_{11})) = \mathrm{Up}_{r2}(M_{11} * K_1), \tag{3}$$

$$M_{13} = F_2(M_1) * \sigma(M_1 + M_{12}) = M_1 * K_2 * \sigma(M_1 + M_{12}), \tag{4}$$

$$N_1 = F_3(M_{13}) = M_{13} * K_3, \tag{5}$$

式中，F_i 表示卷积核 K_i 的卷积操作，σ 为 Sigmoid 函数。

传统卷积感受野往往受到限制，忽略了上下文的语义信息，通过校正操作，使得特征图中每个位置能够考虑其周围的上下文信息，将其得到的上下文信息标量嵌入到原尺度空间。其不仅可以模拟通道之间的依赖性，有效地扩大 SConv 层的感受野，还在一定程度上避免了不相关区域的信息干扰。

2.1.2　ECA

在特征提取过程中使用卷积操作虽能融合其感受野内的空间信息，但往往忽视了它们各通道之间存在的相关性，因此通过通道注意力机制来增强骨干网络的特征提取能力。而特征提取中的降维操作会无差别地丢失特征提取的重要信息，对通道注意力机制的预测有一定的负面影响，并且所需要的特征并非与图像的所有通道相关，所以没有必要建立特征与所有通道之间的联系，并且建立这种联系将耗费模型更多的计算资源。本文通过一维卷积在实现局部通道信息交互的基础上可有效避免降维操作，降低模型复杂度的同时让模型保持良好的性能。

该通道注意力机制实现原理如图 4 所示，X 为输入的原始图像信息，经过全局平均池化(global average pooling，GAP)得到未降维的信息，然后进行尺寸为 k 的一维卷积操作，在相邻通道的小部分范围内实现跨通道信息交互，而 k 的值与通道数成正比，其计算公式如式(6)所示，然后经过 Sigmoid 函数得到各通道的权重值，再与原始输入信息相乘得到含有通道注意力的图像信息 $\hat{X}$，如式(7)所示。

$$k = \Psi(C) = \left| \frac{\log_2 C}{2} + 0.5 \right|_{\mathrm{odd}}, \tag{6}$$

$$\hat{X} = \sigma(\mathrm{C1D}_k(X)) * X, \tag{7}$$

式中，$|t|_{\mathrm{odd}}$ 表示距离 t 最近的奇数，$\mathrm{C1D}_k$ 表示核数为 k 的一维卷积操作。

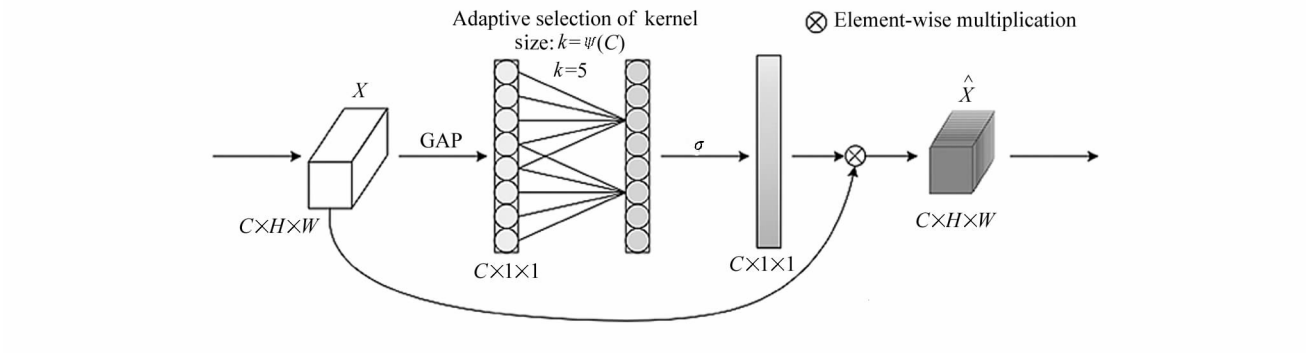

图 4　ECA 模块

Fig. 4　Efficient channel attention module

2.2　文本特征增强

使用 FPN 融合特征时存在两个缺陷，首先，在特征融合前，不同层的特征图需使用 1×1Conv 进行降维，降维后特征图由于感受野大小不一，不同尺度的特征所含有的语义信息存在差异，融合后弱化了多尺度特征的信息表达能力。其次，由于特征融合过程是自顶向下，且金字塔中较高层的特征却因通道减少缺失细节信息，其融合过程会将不同层特征图的语义差异不断放大。因此通过 FPN 融合得到的特征图，文本区域边界定义不清晰，甚至存在一定的偏差，对文本检测的后处理产生影响，对于以上影响可以通过通道、位置注意力相结合的方式对特征图进行修正。因此，本文采用 CA 对 FPN 融合后的特征图进行处理，在训练的过程中不断调整各通道、各位置的权重值，从而得到更为准确、可靠的特征图。

CA 结构如图 5 所示，对于给定输入 X，使用大小为 $(h,1)$、$(1,w)$ 池化层，分别沿着横轴和纵轴进行平均池化，对每个通道在这两方向上进行编码。在横轴上通道 c 的输出为 $Z_c^w(w)$，同理在纵轴上的输

出为 $Z_c^h(h)$,如式(8)、(9)所示。接着连接这两特征映射,并使用 1×1Conv 进行操作以及非线性映射,得到特征图 f,如式(10)所示;然后将特征图沿着空间维度拆分成两个张量 g^h、g^w,如式(11)、(12)所示,并与原输入相乘得到输出 $\hat{X}$,如式(13)所示。

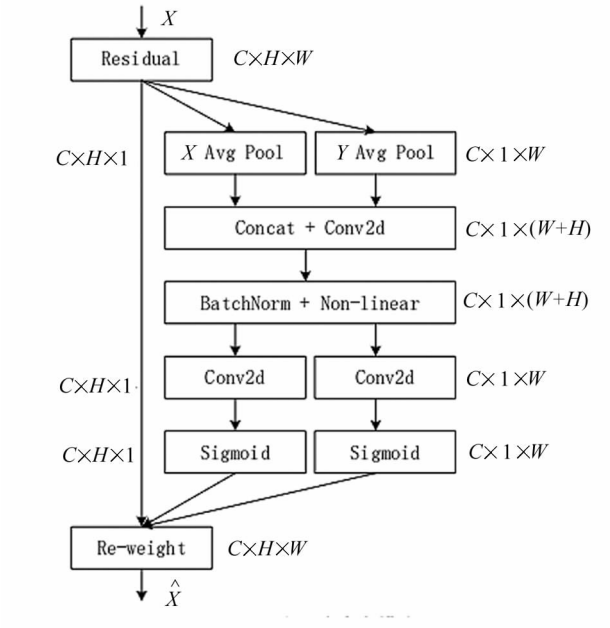

图 5　CA 模块

Fig. 5　Coordinate attention module

$$Z_c^w(w)=\frac{1}{H}\sum_{0\le i\le H}X(j,w), \tag{8}$$

$$Z_c^h(w)=\frac{1}{W}\sum_{0\le i\le W}X(h,i), \tag{9}$$

$$f=\delta(\mathrm{Conv}_{1\times1}([z^h,z^w])), \tag{10}$$

$$g^h=\sigma(\mathrm{Conv}_{1\times1}(f^h)), \tag{11}$$

$$g^w=\sigma(\mathrm{Conv}_{1\times1}(f^w)), \tag{12}$$

$$\hat{X}(i,j)=X(i,j)\times g^h(i)\times g^w(j), \tag{13}$$

式中,[·,·]表示两个张量的拼接操作,δ 表示非线性映射。

相比于全局通道注意,CA 得到的特征向量并不是使用二维全局池化将整个特征张量变进行转换,而是在水平方向和垂直方向上分别汇聚输入的特征,然后经过 Sigmoid 激活函数分别编码成两个不同方向的一维注意力图。通过这种方式处理后,输出不仅具有长程依赖关系,还保存着准确的位置信息。因此,将得到的特征图转换成注意力图,与输入相乘,获得的额外信息可以弥补 FPN 融合过程产生的偏差。

2.3　渐进尺度扩展算法

通过修正后的特征图 $\hat{F}$ 得到不同尺度的文本实例核 S_1-S_n,采用渐进尺度扩展算法对其进行后处理,从最小文本实例核 S_1 逐渐扩展到最大文本实例核 S_n,其扩展过程如图 6 所示(图中 0 表示 S_i 中的像素,1 表示 S_{i+1} 中的像素,不同颜色表示不同的文本实例),在扩展的过程中 S_{i+1} 中某一位置属于 S_i 中同一文本实例时,将该位置进行合并(即图中该位置的颜色变为 S_i 中该文本实例的颜色)。对于存在不能确定像素的归属情况(如图中的 X),解决的原则是先到先得(如在算法中左下角的文本实例先扩展到冲突像素 X,则该像素属于该文本实例)。

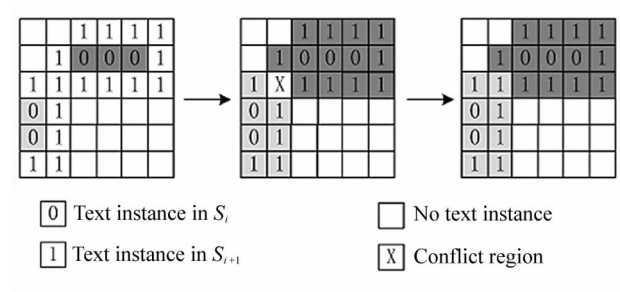

图 6　渐进尺度扩展算法

Fig. 6　Progressive scale expansion algorithm

2.4　损失函数的改进

本文在训练模型时使用的损失函数 L 由两部分构成:L_c 和 L_s,两者按一定的权重求和作为整个模型的损失函数。L_c 是用来衡量未缩放时预测和真实标注之间文本实例的匹配度,L_s 则是用来衡量缩放后的匹配度;损失函数 L 的计算方法如式(14)所示:

$$L=\lambda L_c+(1-\lambda)L_s, \tag{14}$$

式中,λ 是 L_c 和 L_s 的权重系数,λ 的值取 PSENet[3] 实验中的最优值 0.7。

本文采用 *Dice* 系数[11] 损失函数表示 L_c、L_s,*Dice* 系数的计算如式(15)所示:

$$D(S_i,G_i)=\frac{2\sum_{x,y}(S_{i,x,y}\times G_{i,x,y})}{\sum_{x,y}S_{i,x,y}^2+\sum_{x,y}G_{i,x,y}^2}, \tag{15}$$

式中,$S_{i,x,y}$和$G_{i,x,y}$分别表示预测的最终结果 S_i 和训练样本的真实标注 G_i 在图中位置(x,y)处的像素值。考虑到文本图像中难免会有类似文字笔画的背景信息,容易产生误检,降低检测准确率,为了更好区别文本和非文本区域,在训练模型时借鉴了在线难例挖掘(online hard example mining,OHEM)[12] 训练方法,将正、负样本的比例设为 1∶3。

损失函数 L_c 的计算方法如式(16)所示:

$$L_c=1-D^2(S_n*M,G_n*M_{\mathrm{mask}}), \tag{16}$$

式中,M_{mask}表示 OHEM 得到的训练掩码。

损失函数 L_s 的计算方法如式(17)所示:

$$L_s = 1 - \frac{\sum_{i=1}^{n-1} D^2(S_n * \hat{M}, G_n * \hat{M})}{n-1}$$

$$\hat{M}_{x,y} = \begin{cases} 1, \text{if } S_{n,x,y} \geq 0.5 \\ 0, \text{otherwise} \end{cases}, \tag{17}$$

式中，$\hat{M}$ 表示 S_n 中文本区域的掩码，$S_{n,x,y}$ 表示 S_n 中位置(x,y)处的像素值。

因实验中发现损失函数 L_c、L_s 的值收敛于接近于零的极小值，本文使用的损失函数不同于文献[3]中的 L_c、L_s，如式(18)所示，本文将 *Dice* 系数进行了平方操作，使得损失函数的值相比于修改前增大了，加大了模型对误检、漏检的惩罚力度，使得模型预测结果更加接近真实标注，从而改善模型性能。

$$\begin{cases} L_c = 1 - D(S_n * M_{\text{mask}}, G_n * M_{\text{mask}}) \\ L_s = 1 - \dfrac{\sum_{i=1}^{n-1} D(S_n * W, G_n * W)}{n-1} \end{cases}。 \tag{18}$$

3 实验与分析

3.1 实验数据集

本文在实验中选择 ICDAR2015、Total-Text 和 SCUT-CTW1500 3 个常用的文本检测数据集进行训练和测试。ICDAR2015 是个多方向文本检测数据集，图中文字较小，主要有英文和数字；文本区域的标注形式采用 4 个顶点表示的矩形框，共有 1 500 张图片，训练集和测试集的图片数量比为 2∶1。Total-Text 和 SCUT-CTW1500 都是多方向且图中文本弯曲的数据集，包含不同风格、颜色、光照度等情况下的文本，其中 Total-Text 数据集中图片共有 1525 张，其中 1 225 张为训练集，其余为测试集；SCUT-CTW1500 使用 14 个顶点表示多边形来标注文本区域，其含有 1 000 张用于训练的图片，以及 500 张用于测试的图片。

3.2 实验参数

本文实验的所使用的软件平台为 Ubuntu18.04 以及 Pytorch 深度学习框架，使用的显卡为含有 8 G 内存的 NVIDIA RTX 3070。在训练过程中忽略那些模糊的文本实例，并对训练数据集进行数据增强：改变图片的长宽比、随机旋转图像、随机裁剪。选择随机梯度下降(stochastic gradient descent，SGD)算法作为网络模型的优化器，权重衰减率为 5×10^{-4}，动量为 0.99。训练时参数 Batch-size 的值为 2，模型的初始学习率为 1×10^{-4}，共训练 600 个 Epoch，训练到第 100、300、500 个 Epoch 时，学习率下降至前一阶段的 10%。在 Total-Text、SCUT-CTW1500 数据集中将缩放的最小文本标注缩小为原文本框的 70%，而在 ICDAR2015 数据集中因文字较小，则将其设置为 50%。

3.3 评估指标

为了评价模型的检测性能，本文使用 3 个常用的文本检测评价指标：准确率 P(precision)、召回率 R(recall)以及 F 值(F-measure)，其中 F 值综合了准确率和召回率，三者的计算如式(19)－(21)所示。

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

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

$$F = \frac{2 \times P \times R}{P + R}, \tag{21}$$

式中，TP 表示模型能正确预测出文本区域的总个数，FP 表示文本区域预测错误的总个数，FN 表示未能预测到的文本区域的总个数。

3.4 消融实验

为了验证本文方法的有效性，在弯曲文本数据集 SCUT-CTW1500 进行了消融实验，实验结果如表 1 所示，表中，SCNet50 表示将 ResNet50 中卷积替换成SConv的方法，D^2 Loss表示本文修改后的损失函数。

表 1　在 SCUT-CTW1500 数据集上的消融实验

Tab. 1　Ablation experiment on SCUT-CTW1500 dataset

Method	P/%	R/%	F/%
ResNet50	81.74	74.70	78.06
SCNet50	83.45	78.89	81.11
SCNet50＋ECA	84.69	78.44	81.44
SCNet50＋ECA＋CA	87.69	77.11	82.06
SCNet50＋ECA＋CA＋D^2Loss	**88.41**	**78.89**	**83.38**

从表 1 中可以看出，相比于 ResNet50，在骨干网络中使用 SConv，模型检测结果的准确率提升了1.71%，召回率提升了4.19%，F 值提升了3.05%；此外，也表明模型中嵌入 ECA、CA 注意力机制以及改进损失函均对模型性能有改善，最终，改进后模型的检测结果在准确率、召回率、F 值上分别提高了 6.67%，4.19%，5.32%。

图 7 为本文消融实验中的可视化过程，图(a)为检测的原图。由图(b)、(c)可见，改进后模型不会将天空、树叶等背景信息误检为文字，表明在骨干网络中采用 SConv 极大地避免了周围环境信息的干扰；由图(c)－(f)可见，对于字体颜色与背景及其相似的文字“SHOPPING GENT”，模型对其检测的完整度越来越高，边界越来越清晰，说明了引入 ECA、CA、

修改损失函数的有效性。

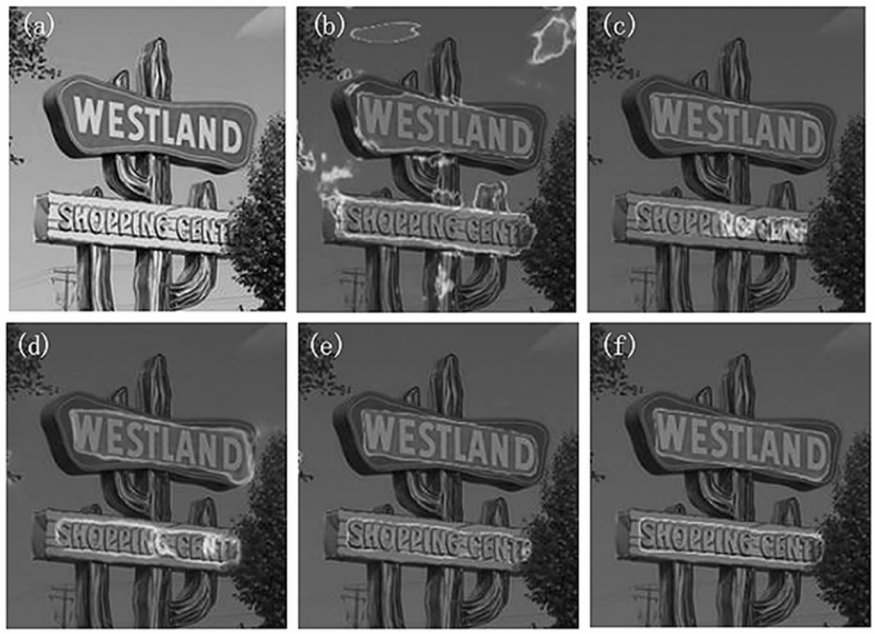

图 7　消融实验可视化:(a) 原图;(b) ResNet50;(c) SCNet50;(d) 高效通道注意力;(e) 协调注意力;(f) D^2 损失

Fig. 7　Visualization of ablation experiment:(a) Original picture;(b) ResNet50;(c) SCNet50;(d) ECA;(e) CA;(f) D^2 Loss

3.5　对比实验

为了进一步验证本文提出的检测方法的有效性,分别在 ICDAR2015、Total-Text、SCUT- CTW 1500 数据集上与近些年主流方法进行对比。结果如表 2—表 4 所示,表中,Ext 栏"√"表示该方法使用了大量额外数据进行了预训练,"-"表示未进行预训练。

表 2　不同方法在 ICDAR2015 上的性能比较

Tab. 2　Performance comparison of different methods on ICDAR2015

Method	Paper	Ext.	P/%	R/%	F/%
PSENet[3]	CVPR′19	—	81.5	79.7	80.6
TextSnake[13]	ECCV′18	√	84.9	80.4	82.6
SegLink++[14]	PR′19	√	83.7	80.3	82.0
SAEmbed[15]	CVPR′19	√	88.3	85.0	86.6
PAN[4]	ICCV′19	—	82.9	77.8	80.3
LOMO[1]	CVPR′19	√	91.3	83.5	87.2
DB[5]	AAAI′20	√	**91.8**	83.2	**87.3**
Boundary[16]	AAAI′20	√	82.2	**88.1**	85.0
Ours		—	84.8	78.6	81.6

由表 2—表 4 可知,对于 ICDAR2015 数据集,本文所提出的检测方法在综合指标 F 值上达到了 81.6%,相比同样无额外数据预训练方法 PAN、PSENet,提高了 1%。在弯曲文本数据 Total-Text 上,准确率相比于其他算法均有所提升,综合指标 F 值仅次于 PAN、ABCNet,F 值比 PSENet 提升了 5.2%;在 SCUT-CTW1500 数据集上,相比于表中其他方法,本文检测方法的准确率及综合指标 F 值达到最优,比 PSENet 分别高出了 7.8%、5.4%。总体上,本文方法对弯曲文本的检测效果较好,其水平达到了近两年使用了大量数据进行预训练的方法。

表 3　不同方法在 Total-Text 上的性能比较

Tab. 3　Performance comparison of different methods on Total-Text

Method	Paper	Ext.	P/%	R/%	F/%
PSENet[3]	CVPR′19	—	81.8	75.1	78.3
TextSnake[13]	ECCV′18	√	82.7	74.5	78.4
SegLink++[14]	PR′19	√	82.1	80.9	81.5
PAN[4]	ICCV′19	—	88.0	79.4	83.5
LOMO[1]	CVPR′19	√	87.6	79.3	83.3
TextRay[17]	MM′20	√	83.5	77.9	80.6
ABCNet[18]	CVPR′20	√	87.9	**81.3**	84.5
DB[5]	AAAI′20	√	87.1	82.5	**84.7**
Ours		—	**89.3**	78.4	83.5

表4 不同方法在CTW1500上的性能比较

Tab. 4 Performance comparison of different methods on CTW1500

Method	Paper	Ext.	P/%	R/%	F/%
PSENet[3]	CVPR'19	—	80.6	75.6	78.0
TextSnake[13]	ECCV'18	√	67.9	**85.3**	75.5
SegLink++[14]	PR'19	√	82.8	79.8	80.9
SAEmbed[15]	CVPR'19	√	82.7	77.8	80.1
PAN[4]	ICCV'19	—	84.6	77.7	81.0
LOMO[1]	CVPR19	√	85.7	76.5	80.8
TextRay[17]	MM'20	√	82.8	80.4	81.6
ABCNet[18]	CVPR'20	√	84.4	78.5	81.4
DB[5]	AAAI'20	√	86.9	80.2	83.4
Ours		—	**88.4**	79.0	**83.4**

4 结 论

本文提出了一种新的场景文本检测方法，在ResNet50的基础上，将普通的卷积替换为SConv，在扩大网络感受空间的同时使其具有更准确的鉴别区域，避免了后续降维操作中丢失更多的语义信息；此外在网络的每一个Bottleneck后引入ECA，适当地跨通道交互，使得特征图具有更丰富的语义信息；在使用特征金字塔融合特征后嵌入CA，于通道和位置注意力中调整权重，能够解决对不同尺度的特征图自顶向下融合过程中导致的特征图失真。最后通过修改损失函数，加大模型预测错误的惩罚力度，从而提升模型的预测精度。实验表明：本文所提出的场景文本检测方法在弯曲文本和多方向文本中的检测效果表现优异，对未来的文本检测技术研究也有重要意义。在将来的研究中，笔者将进一步探索新的文本检测方法，在保证模型预测准确率的同时，设计模型结构更精简、预测精度更高、检测速度更快的算法；进一步探索文本识别算法，解决文本识别问题；进一步探索一种端到端的文本检测与识别算法，实现文本检测与文本识别的结合。

参考文献：

[1] ZHANG C Q, LIANG B R, HUANG Z M, et al. Look more than once: an accurate detector for text of arbitrary shapes[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 16-20, 2019, Long Beach, CA, United States. New York: IEEE, 2019: 10544-10553.

[2] LU Y F, ZHANG A X, LI Y, et al. Multi-scale scene text detection based on convolutional neural network[C]//Proceedings-2019 Chinese Automation Congress, CAC, 2019, November 22-24, 2019, Hangzhou, China. New York: IEEE, 2019: 583-587.

[3] WANG W H, XIE E Z, LI X, et al. Shape robust text detection with progressive scale expansion network[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 16-20, 2019, Long Beach, CA, United States. New York: IEEE, 2019: 9328-9337.

[4] WANG W H, XIE E Z, SONG X G, et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network[C]//IEEE International Conference on Computer Vision, October 27-November 02, 2019, Seoul, Korea (South). New York: IEEE, 2019: 8439-8448.

[5] LIAO M H, WAN Z Y, YAO C, et al. Real-time scene text detection with differentiable binarization[C]//AAAI 2020 - 34th AAAI Conference on Artificial Intelligence, February 7-12, 2020, New York, NY, United States. Menlo Park: AAAI, 2020: 11474-11481.

[6] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, United States. New York: IEEE, 2016: 770-778.

[7] LIU J J, HOU Q B, CHENG M M, et al. Improving convolutional networks with self-calibrated convolutions[C]//IEEE/CVF Computer Society Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, Virtual, Online, United States. New York: IEEE, 2020: 10093-10102.

[8] WANG Q L, WU B G, ZHU P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, Virtual, Online, United States. New York: IEEE, 2020: 11531-11539.

[9] YI L, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings-30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, July 21-26, 2017, Honolulu, HI, United States. New York: IEEE, 2017: 936-944.

[10] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 20-25, 2021, Nashville, TN, Virtual, Online, United States. New York: IEEE, 2021: 13708-13717.

[11] FAUSTO M, NAVAB N, AHMADI S. V-Net: fully convolutional neural networks for volumetric medical image segmentation[C]//Proceedings-2016 4th International Conference on 3D Vision, 3DV 2016, October 25-28, 2016, Stanford, CA, United States. New York: IEEE, 2016: 565-

571.

[12] CHU J, GUO Z X, LENG L. Object detection based on multi-layer convolution feature fusion and online hard example mining[J]. IEEE Access, 2018, 6: 19959-19967.

[13] LONG S B, RUAN J Q, ZHANG W J, et al. TextSnake: a flexible representation for detecting text of arbitrary shapes[C]//The 15th European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. Berlin: Springer Verlag, 2018: 19-35.

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

[15] TIAN Z T, SHU M, YU P Y, et al. Learning shape-aware embedding for scene text detection[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 16-20, 2019, Long Beach, CA, United States. New York: IEEE, 2019: 4229-4238.

[16] WANG H, LU P, ZHANG H, et al. All you need is boundary: toward arbitrary-shaped text spotting[C]//AAAI 2020-34th AAAI Conference on Artificial Intelligence, February 7-12, 2020, New York, NY, United States. Menlo Park: AAAI, 2020: 12160-12167.

[17] WANG F F, CHEN Y F, WU F, et al. TextRay: contour-based geometric modeling for arbitrary-shaped scene text detection[C]//MM 2020-Proceedings of the 28th ACM International Conference on Multimedia, October 12-16, 2020, Seattle, WA, Virtual, Online, United States. New York: Association for Computing Machinery, Inc, 2020: 111-119.

[18] LIU Y L, CHEN H, SHEN C H, et al. ABCNet: real-time scene text spotting with adaptive bezier-curve network[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, Virtual, Online, United States. New York: IEEE, 2020: 9806-9815.

作者简介:

熊　炜　(1976—),男,博士,副教授,硕士生导师,主要从事数字图像处理和计算机视觉方面的研究.