Computer Science and Application 计算机科学与应用, 2022, 12(11), 2608-2618
Published Online November 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1211265

Hans汉斯

基于注意力机制的自然场景文本检测算法

王宪伟，洪智勇*，余文华，王惠吾，吴卓霖

五邑大学智能制造学部，广东 江门

收稿日期：2022年10月25日；录用日期：2022年11月23日；发布日期：2022年11月29日

摘　要

针对目前主流场景文本检测算法在进行多尺度特征融合时不能够充分利用高、低层信息造成的文本漏检，以及长文本边界检测错误的问题，本文提出一种应用注意力机制的多尺度特征融合与残差坐标注意力的场景文本检测算法。该算法将注意力特征融合模块嵌入到金字塔中，通过纠正不同尺度特征的不一致性来提取更多的细节信息，以改善文本的漏检；在融合之后，使用残差坐标注意力模块在纵、横两个方向上捕获方向感知和位置敏感信息，细化边界信息，以优化长文本检测的效果。通过在公开数据集ICDAR 2015和Total-Text上的实验结果表明，该算法在F分数上分别达到了85.5%和83.6%，在推理速度上分别达到了22.4 FPS和40 FPS，相较于DBNet网络，在推理速度上略有下降，但在F分数上分别提高3.2%和0.8%。

关键词

场景文本检测，深度学习，多尺度特征，注意力特征融合模块，残差坐标注意力模块

Natural Scene Text Detection Algorithm Based on Attention Mechanism

Xianwei Wang, Zhiyong Hong*, Wenhua Yu, Huiwu Wang, Zhuolin Wu

Department of Intelligent Manufacturing, Wuyi University, Jiangmen Guangdong

Received: Oct. 25th, 2022; accepted: Nov. 23rd, 2022; published: Nov. 29th, 2022

Abstract

Aiming at the problems of text omission caused by the failure of the mainstream scene text detection algorithm to make full use of the high and low-level information in the multi-scale feature fusion, and the error of long text boundary detection, this paper proposes a scene text detection algorithm which applies the multi-scale feature fusion of attention mechanism and the residual

*通讯作者。

文章引用: 王宪伟, 洪智勇, 余文华, 王惠吾, 吴卓霖. 基于注意力机制的自然场景文本检测算法[J]. 计算机科学与应用, 2022, 12(11): 2608-2618. DOI: 10.12677/csa.2022.1211265

coordinate attention. The model embedded the attention feature fusion module into the pyramid. It extracts more detailed information by correcting the inconsistency of features at different scales to improve the missed detection of text; after feature fusion, the residual coordinate attention module is used to capture orientation-aware and position-sensitive information in vertical and horizontal directions, refine boundary information to optimize the effect of long text detection. The experimental results on the public datasets ICDAR 2015 and Total-Text show that the model achieves 85.5% and 83.6% in F-measure, respectively, and 22.4 FPS and 40 FPS in inference speed. Compared with the DBNet network, this network has a slight decrease in inference speed, but 3.2% and 0.8% improvement in F-measure, respectively.

Keywords

Scene Text Detection, Deep Learning, Multi-Scale Feature, Attention Feature Fusion (AFF) Model, Residual Coordinate Attention (RCA) Model

1. 引言

近年来，对于自然场景下的文本阅读，因其在很多场景都有着广泛的实际应用，受到越来越多研究人员的关注，如信息检索、视觉问答、实时翻译和自动驾驶等。而文本检测作为文本阅读的关键性组件，目的是将视觉捕捉到的文本进行定位，其文本边界框的精确度对于后续的文本识别至关重要。另外，由于自然场景下的文本在尺度、方向、形状多样性和背景等因素的干扰下，使文本检测任务仍具有很大的挑战[1]。

早期的文本检测主要是应用机器学习方法来学习人工设计的特征，虽然这类方法具有一定的可解释性，但特征的设计往往难度大、成本高。近年来，深度学习的出现极大地推动了场景文本检测技术的提升，能够通过模型学习得到深度特征，避免了人工设计特征的繁琐工作，准确率明显高于传统方法，同时也能够很好地应对复杂场景，大致可分为两大类：基于回归的算法和基于分割的算法。

基于回归的算法是通过直接回归文本边界框的点坐标来编码文本实例。文本不同于常规目标，对于长矩形的文本，TextBoxes [2]设计具有不同纵横比的锚框，同时使用不规则的 1*5 卷积以避免正方形卷积带来的噪音；CTPN [3]将长文本分为一系列宽度固定的锚框，使用 BLSTM 来进行序列建模；对于多方向的文本，RRPN [4]引入了具有方向的锚框来生成带有方向角度的倾斜提议框，同时提出旋转 RoI 池化来调整它的方向；TextBoxes++ [5]采用四边形表示代替传统的矩形框，直接回归四个顶点来检测多方向文本；RRD [6]使用具有不同设计的回归与分类分支来分别提取旋转感知特征和旋转不变性特征；对于不规则的文本，CTD [7]使用 14 个点来表示文本区域，提出横向和纵向偏移拼接(TLOC)来学习各点之间的相关性；文献[8]提出自适应点数的文字区域表示，使用 LSTM 来细化文本区域，每个时间步都会预测一对边界点，直到找不到新点为止。基于回归的方法通常只需简单的后处理，但并受限于边界框的表示，使得检测任意形状的效果不容乐观。

基于分割的方法是在像素级别进行特征表示，结合后处理算法生成文本实例。鉴于 FCN [9]网络能够同时考虑局部和全局上下文信息，已被广泛用于生成文本分割图，文献[10]采用两个 FCN 分别生成文本区域显著图与每个字符的中心，文献[11]采用 FCN 生成三种分数图：文本/非文本、字符类别与相邻字符链接方向，文献[12]提出 TextSnake，采用 FCN 生成带有半径和方向信息的文本中心线分数图和文本区域；针对语义分割方法对于相邻文字难区分的问题，文献[13]提出渐进式尺度扩展网络

(PSENet)，对于每个文本实例生成不同尺度的内核，并逐渐扩展内核至实例分割图，但该算法速度较慢、效率低，文献[14]提出可学习的像素聚合(PA)后处理策略，同时提出可级联特征金字塔增强模块(FPEM)和特征融合模块(FFM)。此外，受实例分割思想的鼓舞，PixelLink [15]在像素级别使用 8 个方向信息来编码边界框，预测像素间的连接关系，TextField [16]从最近的文本边界指向每个文本点的方向场，由二维向量的图像表示。基于分割的方法更适用于检测任意形状的场景文本，但后处理相对复杂，实时性较差。

场景文本不同于文档上的文本，其尺度变化非常大，几个像素到几百像素不等，对于深度模型对图像提取的特征，低层的特征图包含更多的位置信息和纹理信息，语义信息较少，而高层特征图包含更多的语义信息和抽象信息，空间信息较少。对于场景文本检测算法，为能够充分利用高、低层的特征信息，DBNet [17]、TextSnake [12]、PSENet [13]、PAN [14]及文献[18]等均采用了不同的特征融合方法，然而，它们集中在构建复杂的路径且通常采用简单的加或拼接的方式来整合不同层的特征信息，仅仅提供特征图的固定线性组合，这样的融合方法容易将浅层特征埋没在背景噪声中，另外，由于文本没有明显的边界，且一般比较长，使得最终的特征表示并不能很好地适应长文本，导致检测到错误的边界。

总之，为实现更准确的文本检测，不仅依赖于高、低层特征信息的充分利用，还在于最后得到的特征表示能否自适应文本的特点。针对以上问题，本文基于 DBNet [17]算法从以下两方面进行改进，提出一个更高效的场景文本检测算法。首先，注意力特征融合模块(Attention Feature Fusion, AFF) [19]用于特征金字塔中，以提升多尺度特征的融合效果；其次，残差坐标注意力模块(Residual Coordinate Attention, RCA) [20]用于融合后的特征，以捕捉远距离特征的相关性，细化边界信息，具体如下：

AFF 能够同时关注相邻特征图的特征信息，相互引导，其中，多尺度通道注意力模块(Multi-Scale Channel Attention Model, MS-CAM)能够在通道维度上聚合局部和全局多尺度上下文信息，可以同时强调分布更全局的大文本和分布更局部的小文本，纠正不同尺度特征的不一致性，保留更多的细节信息，避免引入额外的噪声而造成文本的漏检。

RCA 是一种将位置信息与通道信息相结合的注意力，它将通道注意力沿水平和垂直方向分解为两个具有方位感知的注意力图，在一个更长的范围上捕捉文本的边界信息，有利于长文本的边界检测，同时，残差结构也能避免重要信息的丢失，且计算量不大。

2. 网络结构

2.1. 概述

本文所提出的算法是基于可微分二值化场景文本检测 DBNet 网络进行优化，其网络结构如图 1 所示。首先，使用带有可变形卷积(DCN) [21]的 ResNet-18 [22]作为骨干网络来提取图像的基本特征，DCN 通过增加偏移量来使采样点发生偏移，能够对不同尺度或感受野自适应的定位，更适用于文本的多尺度特征；接着，相邻两层的特征图通过 AFF 模块进行初步融合，然后上采样到相同的尺度并沿通道维度进行拼接；之后，通过 RCA 模块得到最终的特征图；最后，使用反卷积操作分别得到同输入图像同样尺度的概率图(P)和阈值图(T)，之后由公式(1)计算得到二值图(B)，k 在实验中设置为 50，最后根据 B 得到文字检测结果。

$$B_{i,j}=\frac{1}{1+\mathrm{e}^{-k\left(P_{i,j}-T_{i,j}\right)}} \tag{1}$$

2.2. AFF 模块

本 AFF 模块不同于普通 FPN [23]中相邻特征图进行简单的相加操作，而是通过注意力机制解决尺度的不一致性来提升融合效果，其结构如图 2(左)所示。

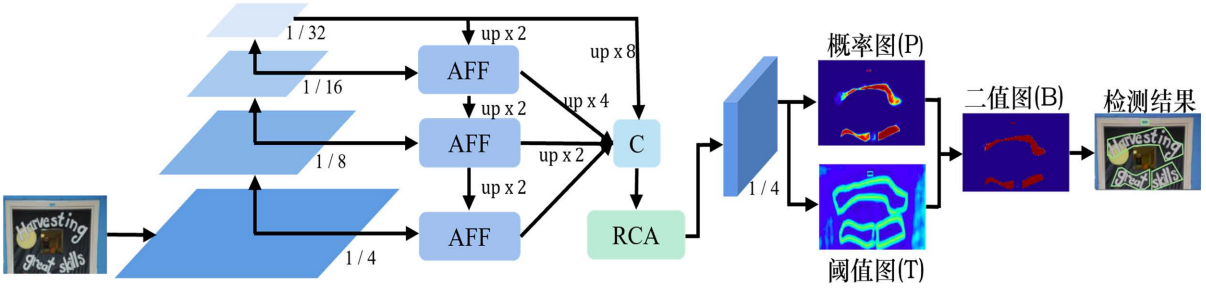

Figure 1. Structure diagram of scene text detection based on attention mechanism
图 1. 基于注意力机制的场景文本检测结构图

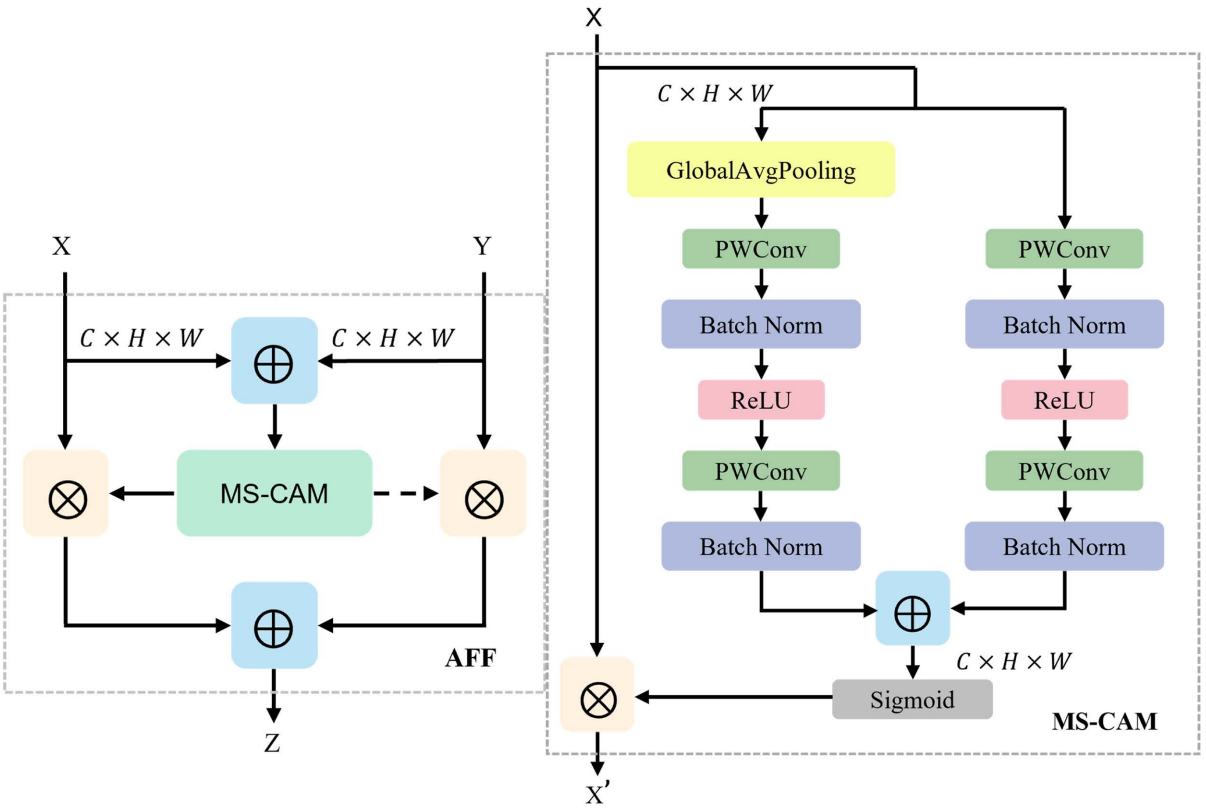

Figure 2. AFF structure diagram (left), MS-CAM structure diagram (right)
图 2. AFF 结构图(左)，MS-CAM 结构图(右)

如图所示，该模块同时关注相邻不同尺度的特征图，使其相互引导，纠正不同尺度特征图之间的不一致性。通常低层特征图拥有更大的分辨率，能够保留更多的文本细节，对于尺度较小的文本更加重要，而高层特征图则拥有更多的语义信息，能够更好地区分文本与背景干扰。图中，$X,Y\in \mathrm{R}^{C\times H\times W}$ 分别定义为特征金字塔中高、低等级语义特征图，C 表示通道数量，$H\times W$ 表示特征图的大小，在这里，为降低内存的开销与计算量，使用 1×1 卷积已将通道数减少至 256。对于高层特征图使用最上采样方法以确保与低层特征图具有相同的大小，随后采用逐像素相加对其进行初步融合，并将将融合后的结果送入 MS-CAM 得到融合权重，之后对不同尺度的特征进行重矫正以增强特征，最后再进行逐像素相加进行二次融合，AFF 可以被表示为公式(2)：

$$Z=\mathrm{M}\left(X\oplus Y\right)\otimes X+\left(1-\mathrm{M}\left(X+Y\right)\right)\otimes Y \tag{2}$$

M 为 MS-CAM 的简写，⊕表示逐元素相加，⊗表示逐元素相乘。

MS-CAM 的关键思想是通过改变空间池大小，可以在多个尺度上实现通道注意力。其结构如图 2(右)所示。它将 AFF 模块初步融合的结果 $X' \in \mathrm{R}^{C\times H\times W}$ 使用两个不同的分支去获得通道注意权重，X' 可由公式(3)得到。其中，一个分支使用全局平均池化去整合全局特征，以强调分布更全局的大文本，由公式(4)表示，全局平均池化由公式(5)表示；另一个分支直接使用逐点卷积去提取局部通道上下文信息，强调分布更局部的小文本，由公式(6)表示：

$$X' = X \otimes M(X) = X \otimes \sigma\left(L(X) \oplus G(X)\right) \tag{3}$$

$$G(X) = B\left(\mathrm{PWConv}_2\left(\delta\left(B\left(\mathrm{PWConv}_1\left(g(X)\right)\right)\right)\right)\right) \tag{4}$$

$$g(X) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X[:,i,j] \tag{5}$$

$$L(X) = B\left(\mathrm{PWConv}_2\left(\delta\left(B\left(\mathrm{PWConv}_1\left(X\right)\right)\right)\right)\right) \tag{6}$$

$M(X)$表示由 MS-CAM 提取得到的注意力权重，B 表示批量归一化，σ 表示 Sigmoid 激活函数 PWConv1 和 PWConv2 表示为逐点卷积，核大小分别为 $C/r\times C\times 1\times 1$ 和 $C\times C/r\times 1\times 1$，其中，$L(X)$与输入特征图具有相同的形状，可保留低层特征图中的细微细节。

2.3. RCA 模块

残差坐标注意模块是将位置信息嵌入到通道注意中，从水平和垂直空间方向编码远程依赖和通道关系，然后聚合特征。其结构如图 3 所示，主要分为两步：坐标信息编码和坐标注意生成。

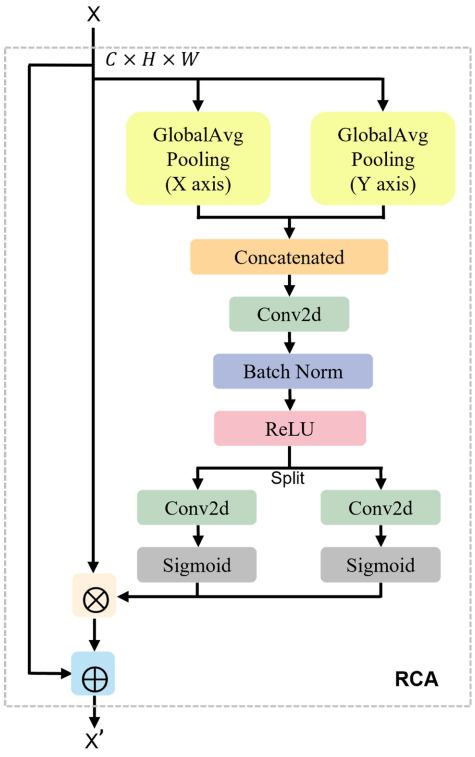

Figure 3. RCA structure diagram
图 3. RCA 结构图

坐标信息编码：全局池化是将全局空间信息压缩到通道描述符中，而这会使文本的位置信息难以保留，因此，这里使用两个空间范围的池化内核$(H, 1)$和$(1, W)$沿水平和垂直方向对每个通道进行编码，第 c 个通道在高为 h 和宽为 w 的输出可分别表示为公式(7)和(8)。

$$Z_c^h(h)=\frac{1}{W}\sum_{0\le i<W}X_c(h,i) \tag{7}$$

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

上述两个变换分别沿两个空间方向聚合特征，产生一对方向感知特征图，使模块能够沿一个空间方向捕获长范围的文本边界信息，以适应长文本的特点，沿另一个空间方向保留精确位置信息，来更准确的定位到图像中的文本。

坐标注意生成：将上述两个变换在空间维度上进行拼接，并使用 1 × 1 卷积来压缩通道，之后进行归一化和非线性变化，可由公式(9)表示：

$$f=\delta\left(B\left(\mathrm{Conv}_{1\times1}\left(\left[z^h,z^w\right]\right)\right)\right) \tag{9}$$

这里，$[\cdot,\cdot]$表示沿空间维度进行拼接，$f\in\mathrm{R}^{C/r\times(H+W)}$为中间特征图，$r$ 为缩减率，实验中设置为 8，之后沿空间维度将 f 差分为两个单独的张量 $f^h\in\mathrm{R}^{C/r\times H}$ 和 $f^w\in\mathrm{R}^{C/r\times W}$，接着使用 1 × 1 卷积调整注意力图的通道数，使其等于输入特征图的通道数。最后，使用 Sigmoid 函数进行归一化得到权重，最终的输出可以由公式(10)表示：

$$y_c(i,j)=x_c(i,j)\times g_c^h(i)\times g_c^w(j)+x_c(i,j) \tag{10}$$

$x_c(i,j)$表示输入特征图，$g_c^h(i)$和$g_c^w(j)$分别表示在水平和垂直方向上的注意力权重。同时，残差结构的使用也能避免重要信息的丢失。

3. 实验结果与分析

3.1. 数据集

为了验证所提算法的有效性，我们在公开数据集 ICDAR 2015 [24]和 Total-Text [25]上进行实验，通过准确率 P (Precision)、召回率 R (Recall)和 F 均值(F-measure)来验证。

ICDAR 2015 数据集包含 1000 张训练图像和 500 张验证图像，文本方向各异，区域由四边形的 4 个顶点注释。

Total-Text 数据集包含 1255 张训练图像和 300 验证图像，包含水平、多方向和弯曲文本。

3.2. 实验细节

针对训练数据，我们首先忽视了文本标签为“##”的文本区域，随后采用随机裁剪，随机旋转和随机翻转以增强数据，增加模型的泛化能力，最后采用 EAST 中提到的裁剪方法将被增强图像重裁剪到大小为 640 × 640 以使网络训练更高效。

我们将模型放在单 GTX 3090 GPU 上训练 1200 轮，Batch size 设置为 16，采用 Adam 作为优化器，初始学习率设置为 0.001，使用 warmup 预热学习率 5 轮，使用 cosine 学习率更新策略，且没有进行预训练，在推理阶段，将 batch size 设置为 1，保持图像的纵横比，并为每个数据集设置合适的高度来调整输入图像大小。

3.3. 对比实验

我们采用 DCN-ResNet-18 作为骨干网络，使用 AFF 用于多尺度特征融合，引入 RCA 来对融合后的特征图进行矫正，使用 ICDAR 2015 数据集来评估所提模型在多方向文本上的检测性能，在推理期间，保持图像的纵横比，调整短边到 736 与 1152，同原 DBNet 及其他经典场景文本检测算法进行比较，结果见表 1。

当我们将图像短边调整到 736 时，可以看到，在准确度和 F 分数上分别达到了 89%和 83.6%，相较于 DBNet 在准确度上提高 2.2%，在 F 分数上提高 1.3%，相较于其它方法，我们所提出的方法在 F 分数上同样具有很强的竞争力，且在推理速度上也具有很大的优势。当我们将图像的短边调整到 1152 时，在召回率和 F 分数上分别提升至 82.5%和 85.5%，较之前分别提高了 3.7%和 1.9%。

Table 1. Comparison of results on the ICDAR 2015 dataset
表 1. 在 ICDAR 2015 数据集上的结果比较

方法	P (%)	R (%)	F (%)	FPS
CTPN [3]	74.2	51.6	60.8	7.1
SegLink [26]	73.1	76.8	75.0	-
EAST [27]	83.6	73.5	78.2	13.2
TextBoxes++ [5]	87.8	78.8	82.9	2.3
PixelLink [15]	82.9	81.7	82.3	7.3
TextSnake [12]	84.9	80.4	82.6	1.1
PSENet [13]	81.5	79.7	80.6	1.6
PAN [14]	84.0	81.9	82.9	26.1
DBNe (736) [17]	86.8	78.4	82.3	48
Ours (736)	89	78.8	83.6	46.6
Ours (1152)	88.7	82.5	85.5	22.4

为验证模型在弯曲文本上的检测性能，本文选用 Total-Text 数据集进行实验，网络结构同上，将短边调整到800，其结果如表2所示，在准确率、召回率及F分数上分别达到了89.3%、78.5%和83.6%，较TextSnake相比，在 F 分数上提高 5.2%，较 DBNet 相比，同样有 0.8%的提升，且推理速度同样具有竞争力。

由此可见，本文方法不论是在检测多方向文本还是弯曲文本，在准确率和速度上均有一定的竞争力。

Table 2. Comparison of results on the Total-Text dataset
表 2. 在 Total-Text 数据集上的结果比较

方法	P (%)	R (%)	F (%)	FPS
TextSnake [12]	82.7	74.5	78.4	-
ATRR [8]	80.9	76.2	78.5	-
TextField [16]	81.2	79.9	80.6	-
PSENet [13]	84.0	78.0	80.9	3.9
PAN [14]	88.0	78.4	83.5	39.6
LOMO [28]	87.6	79.3	83.3	-
DBNet [17]	88.3	77.9	82.8	50
Ours	89.3	78.5	83.6	46

3.4. 消融实验

本文在 ICDAR 2015 上进行消融实验，保持图像的纵横比，将短边调整到 1152，结果如表 3 所示，其中，Baseline 为原 DBNet 在本实验设备及参数配置上的复现结果，当我们仅将特征融合模块替换为 AFF 时，在召回率和 F 分数上分别有 3.06%和 0.88%的提升，当仅使用 RCA 模块时，在 F 分数上略有下降，但结合 AFF 模块使用时，F 分数上又有所激增，达到了 85.47%，相较于 Baseline 提高 1.59%，验证了 AFF 模块和 RCA 模块的有效性。

Table 3. Comparison of ablation experimental results of each module on ICDAR 2015
表 3. 各模块在 ICDAR 2015 上的消融实验结果比较

方法	P (%)	R (%)	F (%)	FPS
Baseline	88.84	79.45	83.88	24.7
+ RCA	89.06	78.98	83.41	23.9
+ AFF	87.11	82.51	84.76	22.1
+ AFF + RCA	88.70	82.47	85.47	20.4

我们对本文所使用的 AFF 模块同其他多尺度特征融合算法 FPN、FPEM_FFM 和 CAM 进行了比较，将它们嵌入到本文所使用的方法中以进行公平的比较，对测试图片的短边调整到 1152，实验结果如表 4 所示，可以看到，我们所使用的 AFF 模块在检测速度上略有减少，但在 F 分数上超越了其它方法，这得益于 AFF 模块能够在尺度融合的时候，同时考虑高、低层的特征信息，相互引导，纠正尺度的不一致性，避免引入额外的噪声，从而使得检测性能更高效。

Table 4. Comparison of the results of different feature fusions on ICDAR 2015
表 4. 不同特征融合在 ICDAR 2015 上的结果比较

特征融合模块	P (%)	R (%)	F (%)	FPS
FPN [23]	88.84	79.45	83.88	24.7
FPEM_FFM [14]	88.55	79.8	83.84	13.6
CAM [29]	85.82	83.24	84.51	22.9
AFF	88.70	82.47	85.47	20.4

与此同时，我们对本文所使用的 RCA 模块同其它注意力 SE 和 CBAM 进行了比较，实验策略同上，结果如表 5 所示，相较于通道注意力(SE)和通道、空间注意力(CBAM)，我们所使用的残差坐标注意力(RCA)在准确率、召回率和 F 分数均超越了其他方法，且对检测速度影响不大，这是由于 RCA 能够在避免重要信息丢失的同时，保留文本位置信息并在更长的范围矫正长文本的边界信息。

Table 5. Comparison of results of different attentions on ICDAR 2015
表 5. 不同注意力在 ICDAR 2015 上的结果比较

注意力模块	P (%)	R (%)	F (%)	FPS
SE [30]	85.26	81.98	83.59	22.0
CBAM [31]	86.6	81.79	84.13	18.4
RCA	88.70	82.47	85.47	20.4

3.5. 可视化结果

图 4 为本文方法与可微分二值化 DBNet 网络在 ICDAR 2015 数据集上的可视化结果，从图中可以看出，DBNet 网络在某些小尺度文本实例上存在漏检，以及对于长文本实例生成不准确的边界框；而本文提出的模型使用 AFF 模块与 RCA 模块相结合，不仅能保留更多的细节特征，也能捕捉远距离的特征相关性，使得对于小尺度文本以及长文本检测效果更好，进一步验证了所提模型的有效性。

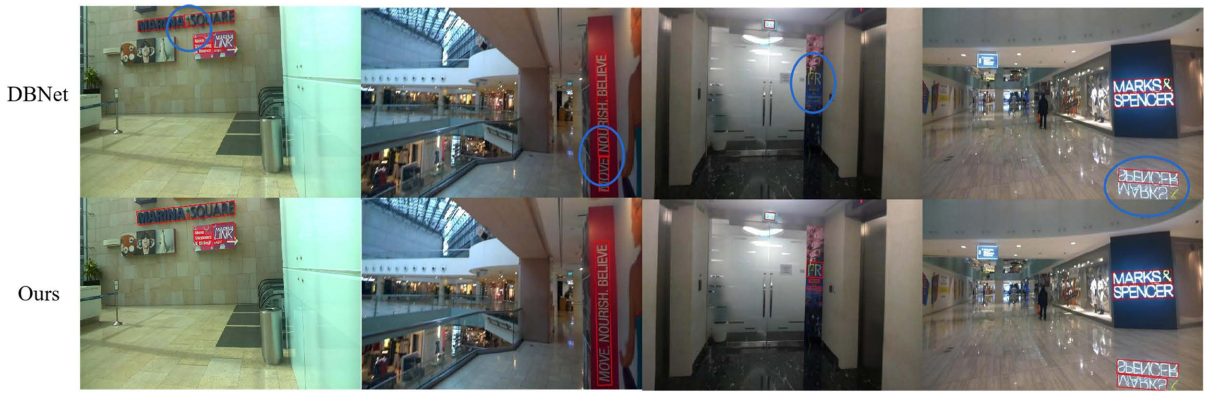

Figure 4. Visualization results on ICDAR 2015
图 4. 在 ICDAR 2015 上的可视化结果

4. 总结

在这篇文章中，我们提出基于注意力机制的场景文本检测算法，在对多尺度进行特征融合时，使用注意力特征融合模块(AFF)来纠正尺度的不一致性，充分利用高、低层信息，抑制背景噪声，以改善文本的漏检。之后，对融合的特征使用残差坐标注意力模块(RCA)在纵、横方向上捕捉长范围文本的位置信息，细化边界信息，使得算法对具有极端长宽比的文本检测效果更好。实验表明，我们所提出的算法在准确性和实时性上有着较好的表现。在接下来的工作中，我们将结合文本识别算法，去设计一个实时的端到端文本提取器。

基金项目

五邑大学港澳联合研发基金(2019WGALH21)；广东省基础与应用基础研究基金(2020A1515011468)；广东省普通高校特色创新类项目(2019KTSCX189)。

参考文献

[1] 刘崇宇, 陈晓雪, 罗灿杰, 金连文, 薛洋, 刘禹良. 自然场景文本检测与识别的深度学习方法[J]. 中国图象图形学报, 2021, 26(6): 1330-1367.

[2] Liao, M., Shi, B., Bai, X., *et al.* (2017) Textboxes: A Fast Text Detector with a Single Deep Neural Network. *Proceedings of the* 31*st AAAI Conference on Artificial Intelligence*, San Francisco, 4-9 February 2017, 4161-4167. https://doi.org/10.1609/aaai.v31i1.11196

[3] Tian, Z., Huang, W., He, T., *et al.* (2016) Detecting Text in Natural Image with Connectionist Text Proposal Network. In: Leibe, B., Matas, J., Sebe, N., Welling, M., Eds., *European Conference on Computer Vision*, Vol. 9912, 56-72. https://doi.org/10.1007/978-3-319-46484-8_4

[4] Ma, J., Shao, W., Ye, H., *et al.* (2018) Arbitrary-Oriented Scene Text Detection via Rotation Proposals. *IEEE Transactions on Multimedia*, **20**, 3111-3122. https://doi.org/10.1109/TMM.2018.2818020

[5] Liao, M., Shi, B. and Bai, X. (2018) Textboxes++: A Single-Shot Oriented Scene Text Detector. *IEEE Transactions on Image Processing*, **27**, 3676-3690. https://doi.org/10.1109/TIP.2018.2825107

[6] Liao, M., Zhu, Z., Shi, B., *et al*. (2018) Rotation-Sensitive Regression for Oriented Scene Text Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 5909-5918. https://doi.org/10.1109/CVPR.2018.00619

[7] Liu, Y.L., *et al*. (2017) Detecting Curve Text in the Wild: New Dataset and New Solution. *ArXiv*, 1712.02170.

[8] Wang, X., Jiang, Y., Luo, Z., *et al*. (2019) Arbitrary Shape Scene Text Detection with Adaptive Text Region Representation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 16-20 June 2019, 6449-6458. https://doi.org/10.1109/CVPR.2019.00661

[9] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

[10] Zhang, Z., Zhang, C., Shen, W., *et al*. (2016) Multi-Oriented Text Detection with Fully Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 4159-4167. https://doi.org/10.1109/CVPR.2016.451

[11] Yao, C., Bai, X., Sang, N., *et al*. (2016) Scene Text Detection via Holistic, Multi-Channel Prediction. *ArXiv*, 1606.09002.

[12] Long, S., Ruan, J., Zhang, W., *et al*. (2018) Textsnake: A Flexible Representation for Detecting Text of Arbitrary Shapes. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 20-36. https://doi.org/10.1007/978-3-030-01216-8_2

[13] Wang, W., Xie, E., Li, X., *et al*. (2019) Shape Robust Text Detection with Progressive Scale Expansion Network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 16-20 June 2019, 9336-9345. https://doi.org/10.1109/CVPR.2019.00956

[14] Wang, W., Xie, E., Song, X., *et al*. (2019) Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, 27-28 October 2019, 8440-8449. https://doi.org/10.1109/ICCV.2019.00853

[15] Deng, D., Liu, H., Li, X., *et al*. (2018) Pixellink: Detecting Scene Text via Instance Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, 2-7 February 2018, 6773-6780. https://doi.org/10.1609/aaai.v32i1.12269

[16] Xu, Y., Wang, Y., Zhou, W., *et al*. (2019) Textfield: Learning a Deep Direction Field for Irregular Scene Text Detection. *IEEE Transactions on Image Processing*, **28**, 5566-5579. https://doi.org/10.1109/TIP.2019.2900589

[17] Liao, M., Wan, Z., Yao, C., *et al*. (2020) Real-Time Scene Text Detection with Differentiable Binarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 7, 11474-11481. https://doi.org/10.1609/aaai.v34i07.6812

[18] 谢斌红, 秦耀龙, 张英俊. 基于学习主动中心轮廓模型的场景文本检测[J]. 计算机工程, 2022, 48(3): 224-252+262.

[19] Dai, Y., Gieseke, F., Oehmcke, S., *et al*. (2021) Attentional Feature Fusion. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, 3-8 January 2021, 3560-3569. https://doi.org/10.1109/WACV48630.2021.00360

[20] Hou, Q., Zhou, D. and Feng, J. (2021) Coordinate Attention for Efficient Mobile Network Design. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 19-25 June 2021, 13713-13722. https://doi.org/10.1109/CVPR46437.2021.01350

[21] Dai, J., Qi, H., Xiong, Y., *et al*. (2017) Deformable Convolutional Networks. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 764-773. https://doi.org/10.1109/ICCV.2017.89

[22] He, K., Zhang, X., Ren, S., *et al*. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[23] Lin, T.Y., Dollár, P., Girshick, R., *et al*. (2017) Feature Pyramid Networks for Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 2117-2125. https://doi.org/10.1109/CVPR.2017.106

[24] Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., *et al*. (2015) ICDAR 2015 Competition on Robust Reading. 2015 *IEEE* 13*th International Conference on Document Analysis and Recognition*, Tunis, 23-26 August 2015, 1156-1160. https://doi.org/10.1109/ICDAR.2015.7333942

[25] Chng, C.K. and Chan, C.S. (2017) Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition. In: *Proceedings of the* 14*th IAPR International Conference on Document Analysis and Recognition*, IAPR Press, New York, 935-942. https://doi.org/10.1109/ICDAR.2017.157

[26] Shi, B., Bai, X. and Belongie, S. (2017) Detecting Oriented Text in Natural Images by Linking Segments. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 2550-2558.

https://doi.org/10.1109/CVPR.2017.371

[27] Zhou, X., Yao, C., Wen, H., *et al*. (2017) East: An Efficient and Accurate Scene Text Detector. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 5551-5560. https://doi.org/10.1109/CVPR.2017.283

[28] Zhang, C., Liang, B., Huang, Z., *et al*. (2019) Look More than Once: An Accurate Detector for Text of Arbitrary Shapes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 16-20 June 2019, 10552-10561. https://doi.org/10.1109/CVPR.2019.01080

[29] Wei, J., Wang, Q., Li, Z., *et al*. (2021) Shallow Feature Matters for Weakly Supervised Object Localization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 5993-6001. https://doi.org/10.1109/CVPR46437.2021.00593

[30] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[31] Woo, S., Park, J., Lee, J.Y., *et al*. (2018) CBAM: Convolutional Block Attention Module. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1