文章编号:1671-4598(2026)06-0062-09 DOI:10.16526/j.cnki.11-4762/tp.2026.06.009 中图分类号:TP391.4 文献标识码:A

基于注意力机制和多尺度特征的场景文本检测算法

李 琼[1],邢景松[1],廖海斌[2],翁 灿[1]
(1.武汉工程大学 电气信息学院,武汉 430205;
2.武汉纺织大学 电子与电气工程学院,武汉 430200)

摘要:针对定位信息传播路径过长,不能充分挖掘不同尺度特征的语义信息,导致难以利用底层定位信息来预测文本边界框的问题,提出了一种基于注意力机制和多尺度特征的场景文本检测方法;改进特征提取模块,构建内嵌十字交叉注意力机制的特征提取器提取多尺度特征,获取上下文感知信息;引入特征融合模块路径聚合网络(PANet)对不同尺度的特征映射进行融合,并提供多尺度的上下文语义信息,使分割网络生成更精细的边界分割结果,并在预测阶段重建损失函数;为了验证该方法的有效性,在公开数据集 ICDAR2015,CTW1500 和 Total-Text 三个公开数据集上进行实验,其综合指标 F_1 值分别达到了 87.4%,82.3%,83.4%,基于注意力机制和多尺度特征的场景文本检测 CM-STD 的性能优于经典的 EAST 方法,且可与当前的 LOMO 方法相媲美。

关键词:场景文本检测;注意力机制;多尺度特征融合;损失函数;任意形状文本

Scene Text Detection Algorithm Based on Attention Mechanism and Multi-scale Features

LI Qiong[1], XING Jingsong[1], LIAO Haibin[2], WENG Can[1]
(1. School of Electrical and Information Engineering, Wuhan Institute of Technology, Wuhan 430205, China;
2. School of Electronic and Electrical Engineering, Wuhan Textile University, Wuhan 430200, China)

Abstract: The propagation path of localization information is too long and the semantic information of different scale features can not be fully mined, which makes it difficult to predict the text bounding box using the underlying localization information. For these issues, a scene text detection method based on attention mechanism and multi-scale features is proposed; Improve the feature extraction module, construct a feature extractor with an embedded cross-crossing attention mechanism, extract the multi-scale features, and obtain the context-aware information; Introduce the feature fusion module path aggregation network (PANet), fuse feature mappings of different scales, provide multi-scale contextual semantic information, make the segmentation network generate more fine boundary segmentation results, and rebuild the loss function in the prediction stage; To validate the effectiveness of the method, experiments are conducted on the three public datasets of ICDAR2015, CTW1500 and Total-Text, which shows that the comprehensive F_1 reach 87.4%, 82.3% and 83.4%, respectively. The CM-STD method is superior in performances to the classical EAST method, with equivalent to the current LOMO method.

Keywords: scene text detection; attention mechanisms; multi-scale feature fusion; loss function; arbitrary shape texts

0 引言

近年来,场景文本检测和识别受到了广泛关注,并已被用于许多基于视觉的应用。例如智慧交通、AR 交互、智能文档分析等等。与传统扫描文本图像不同,场景文本通常具有各种比例和形状,并且字体多样,尺度

收稿日期:2025-07-18; **修回日期**:2025-08-25。
基金项目:江西省主要学科学术和技术带头人培养计划——领军人才项目(20204BCJ22014)。
作者简介:李 琼(1974-),女,硕士,副教授。
引用格式:李 琼,邢景松,廖海斌,等.基于注意力机制和多尺度特征的场景文本检测算法[J].计算机测量与控制,2026,34(6):62-70.

不一，背景复杂，例如复杂背景下文本易与场景元素重叠，导致误检相似纹理或漏检遮挡文本；文本尺度差异达数十倍，固定特征提取易失小文本特征或大文本边界；弯曲、不规则文本难被传统方法完整建模。当前研究热点又围绕这些挑战展开：1）注意力机制与上下文建模结合，强化文本特征、抑制背景，扩大感受野；2）优化多尺度特征融合，减少浅层信息丢失，实现多尺度文本协同检测；3）引入 Transformer 突破卷积局限，提升任意形状文本建模能力，但存在推理慢、小样本泛化差问题。而这些都需要深度学习检测技术。

随着深度学习技术的飞速发展，文本检测领域呈现出以深度学习方法主导的趋势。目前，场景文本检测的主流方法大致可以归为两类：一个是基于回归的方法，另一个是基于分割的方法。

基于回归的方法从目标检测算法发展而来，根据预置锚框或像素点预测的回归框来生成文本检测框。文献［1］提出利用四边形边界框进行检测，修改锚点比率和卷积大小，调整目标框使之适应文本实例纵横比的显著变化。文献［2］中区域建议网络（RRPN，region proposal net-work）通过旋转候选网络区域生成带有倾斜的文本区域，最后通过旋转感兴趣区域层（RROI，Rotated Region of Interest Layer）将候选区域映射到特征图上得到文本检测区域。文献［3］描述了一种高效且精确的场景文本检测器（EAST，efficient and accurate scene text detector），该方法通过全卷积网络输出生成像素级检测结果，然后通过目标检测非极大值抑制算法（NMS，non-maximum suppression）生成文本区域。文献［4］提出了一种字符级文本检测方法，文献［5］和文献［6］也检测到了字符级别的文本实例。他们最初预测了所有潜在的字符框。然后，该方法增加了字符框关联概率。最后，他们通过相应的链接将方框连接起来，以获得文本轮廓。文献［7］通过嵌入深度形态学来捕捉文本的规律性，从而检测任意形状的文本。

基于分割的方法从语义分割发展而来，语义分割通过分割网络从图像中分割文本后进行像素分类，根据分割结果构建文本行，且能通过局部预测来描述各种形状的文本。实例分割网络[8]（PSENet，progressive scale expansion network）中通过渐进尺度扩展的算法分割相近的文本实例，通过缩小文本掩码以生成更小的掩码，可以有效地分离粘性文本并大致定位它们。文献［9］提出像素嵌入，通过计算像素之间的特征距离对像素进行聚类，以检测包含不同纵横比和不精确边界的文本实例。文献［10］使用一种有效的金字塔纵向和横向序列建模方法来检测任意形状的文本实例。文献［11］通过傅里叶轮廓嵌入方法表示文本轮廓，并且运用傅里叶逆变换和 NMS 来生成最后的检测结果。文献［12］首先预测了文本中心区域（TCR，text central region）、扩展率和文本区域（TR，text region）。然后，作者利用展开比和多边形展开算法将 TCR 展开为 TR。文献［13］提出了一种简单有效的基于 Transformer 的场景文本检测框架，将 Transformer 应用到文本检测领域。

基于分割的方法从像素层面进行预测，因此在处理不规则文本图像方面具有显著优势。其中一些方法可以被视为自上而下的方法，基于掩码区域卷积神经网络[14]（Mask RCNN，mask region-based convolutional neural network）的方法会生成类似掩码的文本分割图作为最终结果。在收缩掩码的优势下，现有的实时文本检测方法能够以更有效的方式表示文本实例，极大地提高了文本检测速度的优势。文献［15］通过收缩掩码、文本掩码和类似向量对文本实例进行建模。他们通过将缩小掩码扩展到基于相似向量的文本掩码来获得最终的文本区域。文献［16］通过固定的收缩距离向外扩展收缩掩模轮廓，以重建文本轮廓。这些方法具有更少的预处理过程和后处理步骤，加快了检测的推理过程。文献［17］通过对象级后处理重建文本轮廓，并提出了一种基于同心掩模和多视图特征模块的新文本表示方法。较少的预处理和后处理步骤，可以加快检测的推理过程，但检测精度会受到收缩掩码识别结果的影响。算法通过像素级分类任务训练网络来学习收缩掩模特征会导致模型的感受野较小，并可能在一些背景区域中错误检测到与文本实例相似的纹理特征。

为了解决这一问题，减少漏检和误检现象，提高检测准确率，本文基于收缩掩码方法的优势，设计了一种基于注意力机制和多尺度特征的场景文本检测算法，主要贡献如下：

1）为了得到复杂背景下文本实例的多尺度特征信息，使用基于注意力机制的骨干网络作为特征提取器，扩大感受野，增强算法对上下文信息的感知能力和文本轮廓特征提取能力。

2）为了获得多尺度上下文语义信息，特征融合阶段使用特征路径聚合网络，以增加自下而上的传播路径，解决图像因多尺度变化和网络传递丢失浅层位置信息的问题。

3）预测阶段重建损失函数，加快模型收敛速度，缓解前景目标的混淆问题。

4）对常用开源数据集进行了广泛的实验，以证明该方法实现了优于或可与现有方法相媲美的性能。

1　本文算法

在本节中介绍了所提出算法的整体管道和网络架构，基于注意力机制和多尺度特征的场景文本检测

(CM-STD, cross-attention and multi-scale feature based scene text detector) 的整体网络框架如图 1 所示。首先，采用具有嵌入交叉注意力机制与残差网络结合 (CCA-ResNet, criss-cross attention-ResNet) 作为特征提取器，实现多级别多尺度的特征提取过程。输入图像被馈送到特征提取网络中以获得特征图，记为 $\{C_2, C_3, C_4, C_5\}$，从 C_2 到 C_5 的分辨率依次降低为前者的 1/2，通过在空间维度跨通道交互，增强多尺度特征的语义关联，再经残差连接融合原始特征与注意力增强特征，输出融合后的多级特征。然后将输入特征图馈送到特征路径聚合网络 (PAFPN, path aggregation feature pyramid network) 中得到特征 $\{N_2, N_3, N_4, N_5\}$，并将金字塔特征图上 2 倍上得到 $\{P_2, P_3, P_4, P_5\}$，逐元素相加融合同层级特征，融合后 2 倍下采样得 $\{N_2, N_3, N_4, N_5\}$，实现多尺度跨层交互，再按通道拼接 $\{N_2, N_3, N_4, N_5\}$，生成整合多尺度信息的特征 F。之后将特征 F 作为预测阶段的输入，用于预测概率图 P 和阈值图 T，由 P 和 T 计算出近似二值图 B，通过式 (1) 推理即可得到最终的文本边界框：

$$B = \frac{1}{1 + e^{[-(P-T)]}} \tag{1}$$

1.1 骨干网络

CM-STD 算法使用 CCA-ResNet 骨干网络作为特征提取器，CCA-ResNet 网络在 ResNet50 网络的基础上进行改进，并且嵌入 CCA，以此增大算法对上下文信息的感知能力和文本轮廓特征的提取能力，骨干网络框架如图 2 所示，中间的残差层分别由不同个数的残差结构组成。

1) 卷积核数量。卷积核数量是算法考虑网络计算量的要素之一，大的 7×7 卷积核虽然能够扩大感受野但也会增加网络计算量，多个 3×3 卷积核能够在实现大卷积核效果的同时捕获特征图像素四周的上下文信息，增加网络深度，提高网络的特征提取能力。如式

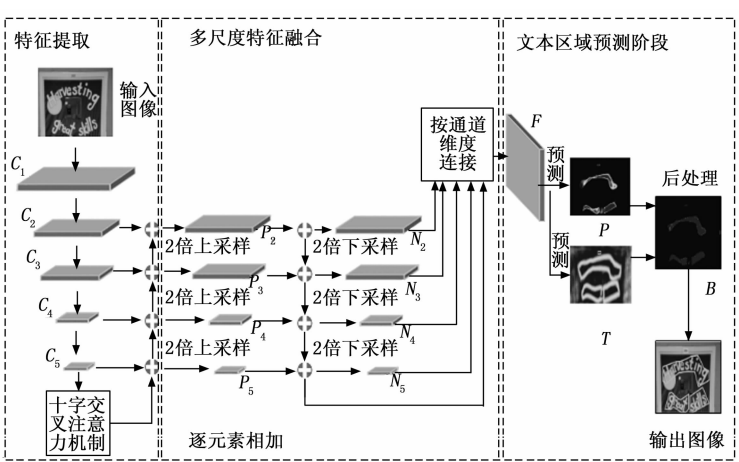


图 1 算法整体框架图

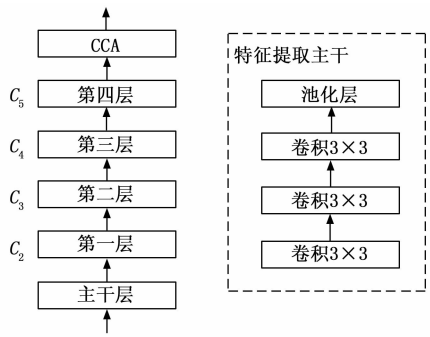


图 2 骨干网络框架

(2) 所示，在通道数不变的假设下，设卷积层输入与输出通道数均为 C。使用 3 个 3×3 卷积替代一个 7×7 卷积，参数量减少至原参数量的约 55%，实现了效率与性能的平衡：

$$\alpha = \frac{3 \times 3 \times 3 \times C^2}{7 \times 7 \times C^2} \tag{2}$$

2) 下采样改进。下采样操作能够降低特征图的维度，在保留图像特征的同时减少参数量，避免过拟合现象的出现。在 ResNet 网络中通常按照卷积步长为 2 来完成下采样操作。池化层通过降低特征面的分辨率来获得具有空间不变性的特征，起到二次提取特征的作用。常用的池化方法有最大池化即提取特征图局部区域内的最大像素值、均值池化即提取特征图的平均像素值、随机池化等。最大池化能够最大程度地降低特征图的背景信息，为了提高网络对图像前景特征和纹理特征的提取能力，因此在残差层中使用最大池化层替代下采样操作。而平均池化层能够保留完整的特征图信息，因此在 stem 层使用二维平均池化，通过平均池化将特征图下采样后再与全连接层相连接。

3) 可变形卷积。骨干网络使用可变形卷积改进卷积层，解决因文本形状未知性而带来的特征表示不足的问题。可变形卷积通过一个标准卷积单元计算得到尺度不变，通道数为 $2N$ 的偏移域，分别代表每层像素点在 x 轴，y 轴的偏移量。标准卷积层加上偏移量后，卷积核的大小和位置可以根据输入特征图的内容进行自适应调整，降低网络复杂度，以此提高具有极端纵横比和形状不规则文本实例的检测结果。

在 CCA-ResNet 骨干网络中的阶段 conv3、conv4 和 conv5 中，在所有 3×3 卷积层中应用调制可变形卷积。

1.2 十字交叉注意力机制

注意力机制源于人眼的视觉感知研究，在训练之后自动调整配置参数，能够像人类视觉一样重视有用信息，减少无效信息的干扰，提高模型

的效率与准确率。为了使用轻量级计算和内存对局部特征表示上的全图像依赖性进行建模，本文所提算法在特征提取器中嵌入十字交叉注意力机制即 CCA 模块，使网络计算像素相关性的同时在水平和垂直方向收集上下文信息，增强网络捕获场景上下文信息的能力。

十字交叉注意力机制如图 3 所示，给定局部特征图 H，该模块首先在 H 上应用两个具有 1×1 滤波器的卷积层，将通道降维并生成查询矩阵 $\boldsymbol{Q}$，键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$，然后通过查询矩阵和键矩阵的类同操作，得到相关矩阵 $\boldsymbol{A}$。取 $\boldsymbol{Q}$ 矩阵中位置 i 处的一点，记为 $\boldsymbol{Q}_i$，同时在 $\boldsymbol{K}$ 矩阵中，可从位置 $\boldsymbol{i}$ 以及同列同行的其余位置提取到 $H+W-1$ 个向量，构成集合 K_i，$K_{i,j}$ 定义为 K_i 中的第 j 个元素，图中类同操作定义为：

$$d_{i,j}=\boldsymbol{Q}_i\boldsymbol{K}_{i,j}^{T} \tag{3}$$

其中：B 为批处理数，C 为通道数，H 和 W 分别为特征图的高和宽，$d_{i,j}$ 表示 $\boldsymbol{Q}_i$ 和 $\boldsymbol{K}_{i,j}^{T}$ 矩阵乘法得到的相关结果，然后在 A 的 $(H+W-1)$ 维度上使用 Softmax 函数，得到注意力图 A^{*}，之后将得到的注意力图 A^{*} 和值矩阵 $\boldsymbol{V}$ 通过聚合操作重新加权特征。值矩阵 $\boldsymbol{V}$ 在 i 点空间位置按行，列方向获得值向量 V_i，Softmax 函数和聚合操作计算公式如式（4）和（5）所示：

$$\mathrm{Softmax}(x)_i=\frac{e^{xi}}{\sum_{j=1}^{n}e^{xj}} \tag{4}$$

$$\boldsymbol{X}_i^{h}=\boldsymbol{V}_i^{T}A_i \tag{5}$$

其中：$\boldsymbol{X}_i^{h'}$ 表示输出子特征图 $\boldsymbol{X}^{h'}$ 在 i 点的特征向量，A_i 表示在 i 点对应行列范围内的注意力权值。将上下文信息添加到局部特征 $\boldsymbol{X}$ 以增强逐像素表示。因此，它具有广泛的上下文视角，并根据空间注意力图选择性地聚合上下文，这些特征表示实现了相互增益，并且对于语义分割更加鲁棒。

最后把不同特征子空间的输出级联后与可以自学习的标量 α 相乘，与输入相加得到最终输出，使网络通过学习特征残差的方式来降低学习难度：

$$Y=\alpha\sum_{i}^{H+W-1}X_i^{h}+X \tag{6}$$

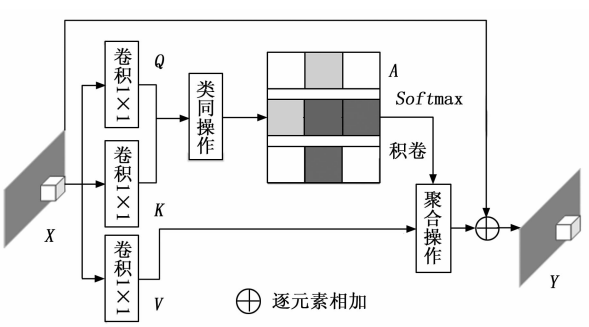


图 3　十字交叉注意力机制示意图

其中：Y 为最终输出特征图，X 为输入特征图。

1.3　多尺度特征融合

在神经网络中底层特征有着精确的位置信息，高层特征拥有较强的语义信息。由于网络层数的加深，底层特征向高层传播时会越来越难以获取精确的位置信息，丢失较多的浅层特征，最终导致无法精确地检测出小目标的文本实例。常见的场景文本检测通常采用特征金字塔结构[18]（FPN，feature pyramid network）进行特征提取和融合，兼顾大目标和小目标文本实例。为了缩短底层向高层传播的路径和充分利用底层位置信息，在 FPN 的基础上增加自下而上的路径聚合增强模块（PA-Net，path aggregation network）构造特征路径聚合网络（PAFPN，feature path aggregation network）进行特征融合，捕获远程浅层特征，如图 4 所示。其网络路径聚合增强模块提高了浅层特征层位置信息的利用率，在保留横向连接的同时增加路径聚合，缩短了底层特征向高层传播的路径，以此保留了底层特征传递时损失的精确位置信息。传统 FPN 是单一自上而下长路径传播特征，易丢底层位置信息，横向连接对特征交互支撑弱，而 PAFPN 构建“自顶向下＋自底向上”双向路径，缩短底层特征传播距离，减少位置信息丢失。通过这种方式落地到检测效果，既提升小目标文本检测精度，降低漏检、误检率，适配不同大小文本检测，又增强面对模糊、干扰场景时的鲁棒性和精度，使网络在复杂环境中更稳定。

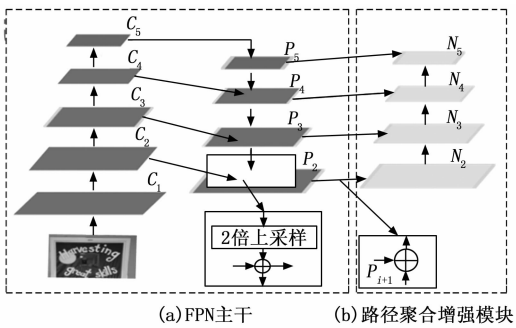


图 4　PAFPN 网络结构

路径聚合增强模块通过上采样得到具有更高分辨率的特征图，特征层通过将输入的长和宽扩大至原来的两倍后，然后与上一个特征层进行堆叠，实现浅层特征与深层特征的促进融合。随后，深层语义信息回传，并再次进行下采样与下一特征层堆叠。通过横向连接将自下而上生成的特征图与上采样结果融合，最终得到特征集 $\{P_2, P_3, P_4, P_5\}$。增强路径从最底层的 P_2 开始，逐层传递至 P_5，过程中逐级完成下采样。最终，路径聚合后生成的新特征映射用 $\{N_2, N_3, N_4, N_5\}$ 表示，对应原特征集 $\{P_2, P_3, P_4, P_5\}$。

如图 5 所示，特征融合结构的具体过程如下：每一特征层通过横向连接，将更粗糙的特征映射 P_{i+1} 与相应更高分辨率的特征映射 N_i 结合，然后生成新的特征映射 N_{i+1}。首先，特征映射 N_i 通过一个步长为 2 且为 3×3 卷积层，目的是为了缩小空间尺寸；接下来，通过横向连接的方式，将其特征映射 P_{i+1} 的每个元素与其下采样映射进行相加，完成融合。最终，融合后的特征映射再经过 3×3 卷积层完成进一步处理，生成连续网络的 N_{i+1}。这一迭代过程从低层开始，并逐步接近 P_5 结束。

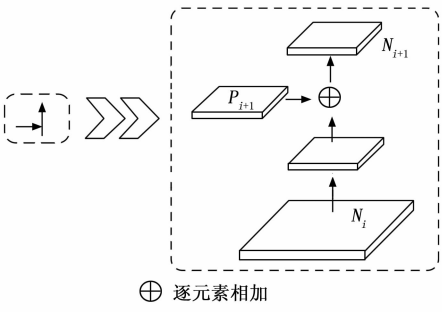


图 5 自下向上路径聚合增强模块示意图

1.4 可微二值化

检测过程中最后一步是通过后处理算法预测得到文字区域的检测结果。通常在基于分割的文本检测网络中，由二值化预测图推理文本边界，最终的二值化都会使用固定阈值获取，而且取不同的阈值会对性能产生不同的影响。标准二值化过程描述如式（7）所示：

$$B_{i,j}=\begin{cases}1 & \text{if} \quad P_{i,j}\geqslant t\\ 0 & \text{otherwise}\end{cases} \tag{7}$$

其中：t 为预先定义的阈值，$(i,\ j)$ 为图中的坐标点，像素值为 1 的认为是文本区域。

然而标准二值化是不可微的，因此在训练期间无法与分割网络一起进行优化。为了解决这个问题，预测阶段使用近似阶跃函数对每个像素点进行自适应二值化，由网络学习得到二值化的阈值，彻底将二值化过程加入网络中一起训练。利用分割网络得到的特征图预测概率图 P 和阈值图 T，二值化公式如式（8）所示：

$$B'_{i,j}=\frac{1}{1+e^{-k(P_{i,j}-T_{i,j})}} \tag{8}$$

其中：B 是近似二值图，T 是从网络中学习得来的自适应阈值图，k 是放大系数，在复杂场景中为了平衡精度与泛化能力，一般设置 k 为 50[16]。

1.5 标签生成

在可微分二值化的后处理过程中，预测完整的文本区域图需要生成与概率图和阈值图相对应的标签再进行网络模型的训练。对于完整的文本实例标签，本文将输入图像的文本区域看作是一个多边形，然后将多边形的区域设置为 1，其他区域设置为 0，并使用一组线段来描述这个多边形，可以有效地表示文本的连通区域：

$$G=\{S_k\}_{k=1}^{n} \tag{9}$$

其中：G 是一个包含多个子图的集合，n 是顶点的数量，数据集不同 n 的取值也可能不同。k 是一个索引，用于标识每个子图 S_k 对应的特定尺度。S_k 是在特定尺度 k 下，从输入图像中提取的特征图。

通过二值图得到文本连通区域后，再利用偏移量 D 扩大连通区域得到文本区域，收缩后的偏移量 D 根据原始多边形的周长 L 和面积 A 计算：

$$D=\frac{A(1-r^2)}{L} \tag{10}$$

其中：r 为收缩率，根据经验设置[8]为 0.4。

使用 Vatti 裁剪算法将文本区域 G 收缩偏移量 D 后变成收缩区域 G_S，如式（11）所示：

$$G_S=\{x\in G\mid dist(x,\partial G\geqslant D\} \tag{11}$$

其中：∂G 代表文本区域 G 的边界，$dist(x,\partial G)$ 计算点 x 到边界的距离。阈值图生成标签的过程和概率图生成标签的过程类似，将文本区域 G 以相同的偏移量 D 扩大到 G_d，如式（12）所示：

$$G_d=\{x\in \boldsymbol{R}^2\mid dist(x,G\geqslant D\} \tag{12}$$

R^2 是二维平面空间，$dist(x,G)$ 计算点 x 到文本区域 G 的最短距离。考虑将 G_S 和 G_d 之间区域的距离作为文本边界，然后计算其中到 G 中最接近的线段的距离来生成阈值图的标签，即阈值图距离标签 y^*，如式（13）所示：

$$y*=\min_{i=1..n}\frac{\|(x-p_i)\times(q_i-p_i)\|}{\|q_i-p_i\|} \tag{13}$$

p_i 和 q_i 均为 G 内的线段端点，i 为边界线段。阈值图需反映“边界过渡区”与文本核心的关联度，因此计算 G_S 与 G_d 之间的区域到 G 边界线段的距离的最小值作为标签。收缩掩模的可视化如图 6 所示。

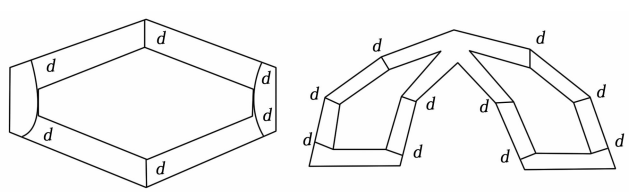


图 6 收缩掩模的可视化

在训练期间，对概率图、阈值图和近似二值化图进行监督，其中概率图和近似二值化图将共享相同监督。为了提高效率使用概率图，以便可以去除阈值分支。盒子的形成过程包括 3 个步骤：

1）首先用恒定阈值（0.2）对概率图/近似二值图进行二值化，得到二值图；

2）从二进制映射中获得连接区域（收缩文本区域）；

3）通过裁剪算法以偏移量 D 进行扩张得到扩张后的文本区域，以此生成文本边界框。

偏移量 D' 公式如式（14）所示：

$$D' = \frac{A' \times r'}{L'} \tag{14}$$

其中：A' 为多边形的面积，L' 是收缩后的多边形的面积，收缩率 r' 在本文中设置为 1.5。D 和 D' 是相互关联的，前者定义文本的内边界标签，后者定义文本的外边界标签，二者共同作用于文本区域的检测与提取。

1.6　损失函数

本文损失函数由预测阶段 3 个分支的组成，分别是概率图的损失 L_S，二值图的损失 L_b 和阈值图的损失 L_t 的加权：

$$L = L_S + \alpha \times L_b + \beta \times L_t \tag{15}$$

根据相对应损失的数值，α 和 β 的默认值为 5 和 10。概率图损失 L_S 使用 DiceLoss 损失函数，DiceLoss 损失函数可以缓解样本中前景和背景面积不平衡带来的消极影响，适用于场景文本任务，计算公式如式（16）和（17）所示：

$$Dice = \frac{2 \times (Pred \cup True)}{Pred \cup True} \tag{16}$$

$$DiceLoss = 1 - Dice \tag{17}$$

其中：$Pred$ 为网络输出的预测结果，$True$ 为真实标签结果，$Dice$ 系数作为一种评估两种样本相似性的度量函数，其值越大，意味着预测样本和真实样本之间的相似性越高。

二值图损失函数 L_b 使用二元交叉熵函数（BCE，binary cross entropy），并且为了解决正负数不平衡的问题，在 BCE 中采用 hard negative mining 的采样方法：

$$L_b = \sum_{i \in S_l} y_i \log x_i + (1 - y_i) \log(1 - x_i) \tag{18}$$

其中：S_l 是采样集，其中的正负样本的比例为 1∶3。

L_t 是扩大后的文本多边形 G_d 内预测值和标签值之间的距离之和，其具体表达式如式（19）所示：

$$L_t = \sum_{i \in R_d} | y_i^* - x_i^* | \tag{19}$$

其中：y_i^* 是每个阈值图的标签，R_d 扩大后的文本多边形 G_d 内像素的一组索引。

2　实验结果与分析

2.1　数据集与评价指标

本文实验在多方向文本数据集 ICDAR2015[26]，弯曲文本数据集 Total-Text[27] 和 CTW1500[28] 数据集上开展。

ICDAR2015 数据集是一个常用的多方向文本检测数据集，其中包括很多小的和低分辨率的文本实例，由 1 000 张训练图像和 500 张测试图像组成。CTW1500 数据集是一个已经被广泛使用的曲线文本检测数据集，文本尺度变化很大，其中有 1 000 张训练集图像和 500 张测试集图像。Total-Text 数据集由 1 255 张训练图像和 300 张测试图像组成，包含了多种文本方向，水平、竖直以及弯曲文本数据集。

本文检测性能的评估标准，主要考虑 4 个性能参数，分别是准确率（P，precision），召回率（R，recall），综合评价指标（F_1）和检测速度（FPS，frames per second）。

1）准确率 P 表示被判定为正样本的数量占真实正样本数量的比例。

2）召回率 R 表示真实正样本的数量占被判定为正样本的比例。

3）F_1 是准确率 P 和召回率 R 的加权调和平均值，其计算公式为：

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{20}$$

4）检测速度 FPS，单位为帧/s。

2.2　实验设置

实验所用到的软硬件环境：使用 Pytorch 深度学习网络框架实现，一块 3080Ti 显卡进行实验。训练过程中采用 SGD 优化器，训练批量大小为 8，初始学习率为 0.007，权重衰减为 0.000 1，梯度冲量设为 0.9，为避免训练后期出现过拟合与梯度震荡，学习率会随着训练轮次（Epochs）的增大采用阶梯式衰减策略（每迭代一定 Epochs 后按固定比例下调，如每 300 epochs 衰减至原学习率的 1/10）进行动态梯度调节，整体训练迭代次数设定为 1 200 次。

另外，为提升模型的泛化能力与鲁棒性，采用多维度数据增强策略：包括角度范围为（−10°，10°）的随机旋转、随机翻转及随机剪裁等多种数据增强方法来提高模型的泛化性和鲁棒性。所有处理后的图像都重新调整为 640×640 像素来提高训练的效率。

2.3　消融实验

为了验证 CM-STD 算法中改进后的骨干网络 CCA-ResNet50，注意力机制 CCA，多尺度特征融合模块 PAFPN，重构损失函数的有效性，在 ICDAR2015 数据集上进行各个模块的消融实验，与其他方法不同，本文算法的实验均没有经过 SynthText 等大规模数据集的预训练。

2.3.1　重构损失函数

在预测阶段优化损失函数，对预测模块进行改进，在骨干网络之后加入重构损失函数与骨干网络对比，实验结果如表 1 所示。准确率提高了 0.4%，召回率提高

了 0.5%，F_1 值提高了 0.4%，证明预测阶段中重建损失函数对文本检测结果具有一定的积极作用。

表 1 重构损失函数的性能影响 %

重构损失函数	P	R	F_1
—	87.4	82.7	85.0
√	87.8	83.2	85.4

2.3.2 特征路径聚合网络

在同时叠加重构损失函数的基础上，引入 FPN 的基础增加自下而上的路径聚合增强模块，构造特征路径聚合网络（PAFPN）进行特征融合，捕获远程浅层特征。实验以未引入 PAFPN 为参照，另一组则启用 PAFPN 对比，如表 2 所示。F_1 值比改进预测模块的方法提高了 1%，准确率提高到了 2.2%。

表 2 PAFPN 的性能影响 %

PAFPN	P	R	F_1
—	87.8	83.2	85.4
√	90.0	83.2	86.4

路径聚合增强模块减少了底层位置信息的损失，增强了高层特征表示，提高了文本像素的分类能力，减少了漏检和误检现象的发生。

2.3.3 十字交叉注意力机制

基线模型的骨干网络默认使用为 ResNet50，对 ResNet50 的特征提取进行改进。在同时叠加重构损失函数以及 PAFPN 的基础上，在特征提取阶段嵌入 CCA 机制。实验以未引入 CCA 为参照，另一组则启用 CCA 对比，实验结果如表 3 所示，召回率虽有所下降，但准确率提升了 1.4%，F_1 值提高了 0.2%。

表 3 CCA 的性能影响 %

CCA	P	R	F_1
—	90.0	83.2	86.4
√	91.4	82.7	86.8

使用改进后的特征提取器，算法性能够得到更优结果，CCA 模块能够有效收集水平和垂直方向上的上下文信息，将上下文信息添加到局部特征表示中，以此增强了网络的特征提取能力。

2.3.4 CCA-ResNet 骨干网络

基线模型 CM-STD 的骨干网络默认使用 ResNet50，使用改进后的 CCA-ResNet50 进行替换。在同时叠加重构损失函数、PAFPN 和 CCA 的基础上，对比两个不同的骨干网络，实验结果如表 4 所示。召回率提升了 1.4%，F_1 值提高了 1.4%。说明改进后的 CCA-ResNet 模块能有效提升模型的检测性能。

表 4 CCA-ResNet 的性能影响 %

CCA-ResNet	P	R	F_1
—	91.4	82.7	86.8
√	91.0	84.1	87.4

2.4 对比实验

为了验证本文提出的算法的有效性，将本文算法在多方向文本，弯曲文本，任意形状文本的公开数据集上与其它文本检测算法进行对比实验。

表 5 是不同算法在 ICDAR2015 多方向文本数据集上的文本检测性能对比结果。本文方法在 ICDAR2015 数据集上的召回率为 84.1%，准确率为 91.0%，F_1 值为 87.4%，检测速度为 5.9 帧/s，F_1 值在所有算法中取得了最高值。与准确率最高的 LOMO 相比，F_1 值提高了 0.2%。与召回率最高的方法 SAE 相比，召回率仅相差 0.4%，F_1 值提高了 3.4%。与检测速度最快的方法 PAN 相比，尽管运行速度较慢，但 F_1 值提高了 4.5%，召回率提高了 2.2%，准确率提高了 7%。实验结果证明，本文算法在多方向长文本数据集上具有较好的检测性能。

表 5 ICDAR2015 实验结果

模型	P/%	R/%	F_1/%	FPS/(f/s)
EAST[3]	83.6	73.5	78.2	13.2
Textboxes++[1]	87.2	76.7	81.7	11.6
RRPN[2]	84.0	77.0	80.0	—
PSENet[8]	81.5	79.7	80.6	1.6
Textsnake[19]	84.9	80.4	82.6	1.1
PAN[15]	84.0	81.9	82.9	**26.1**
PixelLink[20]	85.5	82.0	83.7	1.1
SAE[9]	85.1	**84.5**	84.8	3.0
FCENet[11]	85.1	84.2	84.6	—
LOMO[21]	**91.3**	83.5	87.2	3.4
CM-STD	91.0	84.1	**87.4**	5.9

为了验证算法在粘性和弯曲文本实例上的检测能力，本文在弯曲数据集上将 CM-STD 与其他方法进行比较，实验结果如表 6 和表 7 所示。

表 6 是不同算法在 CTW1500 弯曲文本数据集上的文本检测性能对比结果。本文方法在 CTW1500 数据集上的检测速度为 6.3 帧/s，召回率为 79.9%，准确率为 84.9%，F_1 值为 82.3%。与准确率最高的方法 LOMO 相比，F_1 值提高 1.5%，召回率提高了 3.4%，与召回率最高的方法 Testsnake 相比，F_1 值提高 6.7%，准确率提高了 17%，与 F_1 值最高的 PCR 方法，F_1 值仅相差 0.1%，与检测速度最快的 PAN 方法相比，召回率提高了 2.2%，准确率提高了 0.3%，F_1 值提高了 1.2%，同时检测速度达到 6.3 帧/s，具有不慢的运行速度。

表 6　CTW1500 实验结果

模型	P/%	R/%	F_1/%	FPS
EAST[3]	78.7	49.1	60.4	21.2
Textsnake[19]	67.9	**85.3**	75.6	—
PSENet[8]	80.6	75.6	78.0	3.9
LOMO[21]	**85.7**	76.5	80.8	—
PAN[15]	84.6	77.7	81.0	**39.8**
ABCNet[23]	83.8	79.1	81.4	9.5
TextRay[24]	80.4	82.8	81.6	—
PCR[25]	85.3	79.8	**82.4**	—
CM-STD	84.9	79.9	82.3	6.3

表 7　Total-Text 实验结果　　%

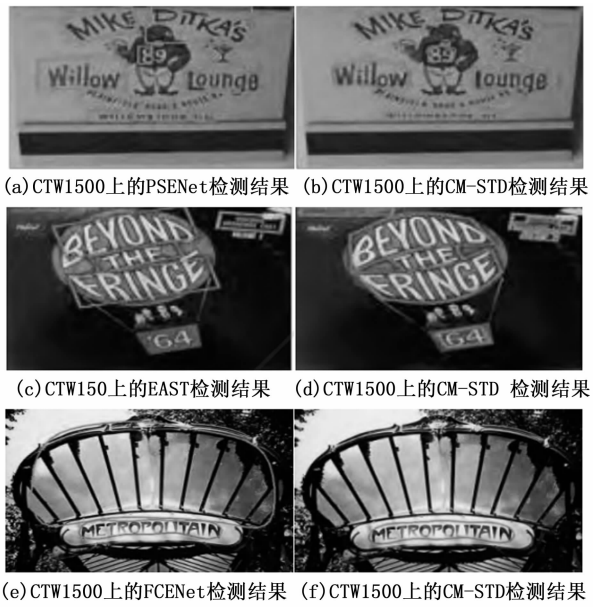
模型	P	R	F_1
EAST[3]	50.0	36.2	42.0
PSENet[8]	84.0	78.0	80.9
Textsnake[19]	82.7	74.5	78.4
TextField[22]	81.2	79.9	80.6
PAN[15]	88.0	79.4	83.5
FCENet[11]	87.4	79.8	83.4
LOMO[21]	**87.6**	79.3	**83.6**
CM-STD	86.3	**80.6**	83.4

表 7 是不同算法在 Total-Text 弯曲文本数据集上的文本检测性能对比结果。本文方法在 Total-Text 数据集上的 F_1 得分达到了 83.4%，召回率为 80.6%，准确率为 86.3%。召回率在所有算法中取得了最高值。与准确率最高的 PAN 方法相比，F_1 值的差异仅为 0.1%，召回率提高了 1.2%。与 F_1 值最高的 LOMO 方法相比，F_1 值仅相差 0.2%，召回率提高了 1.3%。CTW1500 和 Total-Text 数据集的结果表明，CM-STD 在处理单词和行级别的弯曲文本时具有优越性。

2.5　可视化结果

为了将网络 CM-STD 与现有方法进行比较，在广泛使用的文本检测数据集（即 ICDAR2015、CTW1500 和 Total-Text）上进行了实验。CTW1500 数据集上文本检测的可视化如图 7 所示。

从图 7（a）和图 7（b）可以看出，CM-STD 可以准确地定位更多的文本示例，这源于网络的 PAFPN 模块，通过自上而下特征传递与横向连接，整合深浅层特征，解决传统单尺度网络对小尺寸文本的检测；从图 7（c）和图 7（d）可以看出，CM-STD 方法可以准确地检测到弯曲幅度较大的弯曲文本，这关键在于 CCA 和可微二值化的协同作用；从图 7（e）和图 7（f）可以看出，它可以有效地减少复杂背景干扰引起的误检测，这得益于 CCA-ResNet 骨干网络的注意力机制，聚焦文本核心特征。上述 3 组图像中分别存在不同尺度和弯曲文本，CM-STD 网络可以准确地检测到它们，这证明了网络在检测不同尺度和弯曲文本方面的优势。



(a)CTW1500上的PSENet检测结果　(b)CTW1500上的CM-STD检测结果

(c)CTW150上的EAST检测结果　(d)CTW1500上的CM-STD 检测结果

(e)CTW1500上的FCENet检测结果　(f)CTW1500上的CM-STD检测结果

图 7　CTW1500 可视化对比结果分析

3　结束语

针对复杂场景背景和可变文本任意形状的问题，研究提出了一种用于检测自然场景中任意形状文本的网络 CM-STD。CM-STD 的框架由特征提取器、特征融合网络和文本区域预测组成。为了获得更具鉴别性和丰富性的上下文特征信息，使用 CCA-ResNet 骨干网作为特征提取器来提取文本实例的多尺度特征信息。为了解决语义传播过程中潜在信息丢失的问题，特征融合网络使用特征路径聚合网络来融合不同信息尺度的特征。文本区域预测通过使用可区分的二值化模块和优化损失函数提高了所提出方法的文本检测效率。在无预训练的条件下，CTW1500、Total-Text 和 ICDAR2015 等几个任意形状文本的数据集上的实验结果证明了网络 CM-STD 的有效性，未来，希望能扩展到端到端文本检测方法。

参考文献：

[1] LIAO M，SHI B，BAI X，et al. TextBoxes＋＋：a single-shot oriented scene text detector［J］. IEEE Transactions on Image Processing，2018，27：3676－3690.

[2] SHI B，BAI X，BELONGIE S. Detecting oriented text in natural images by linking segments［C］//Proc. of the IEEE Conference on Computer Vision and Pattern Recognition，2017：2550－2558.

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

[4] TANG J，YANG Z，WANG Y，et al. Seglink＋＋：De-

tecting dense and arbitrary-shaped scene text by instance-aware component grouping [J]. Pattern Recognition, 2019, 96: 1-35.

[5] ZHANG S, ZHU X, HOU J, et al. Deep relational reasoning graph network for arbitrary shape text detection [C] //Proc. of theIEEE Conference on Computer Vision and Pattern Recognition, 2020: 9699-9708.

[6] MA C, SUN L, ZHONG Z, et al. Relatext: exploiting visual relationships for arbitrary-shaped scene text detection with graph convolutional networks [J]. Pattern Recognition, 2021, 111: 1-15.

[7] XU C, JIA W, WANG R, et al. MorphText: deep morphology regularized accurate arbitrary-shape scene text detection [J]. IEEE Transactions on Multimedia, 2023, 25: 4199-4212.

[8] WANG W, XIE E, LI X, et al. Shape robust text detection with progressive scale expansion network [C] //Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 9336-9345.

[9] TIAN Z, SHU M, LÜ P, et al. Learning shape-aware embeding for scene text detection [C] //Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 4234-4243.

[10] ZHANG S, LIU Y, JIN L, et al. Opmp: an omnidirectional pyramid mask proposal network for arbitrary-shape scene text detection [J]. IEEE Transactions on Multimedia, 2020, 23: 454-467.

[11] ZHU Y, CHEN J, LIANG L, et al. Fourier contour embedding for arbitrary-shaped text detection [C] //Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2021: 3123-3131.

[12] JIANG X, XU S, ZHANG S, et al. Arbitrary-shaped text detection with adaptive text region representation [J]. IEEE Access, 2020, 8 (102): 106-118.

[13] TANG J Q, ZHANG W Q, LIU G Y. Few could be better than all: feature sampling and grouping for scene text detection [C] //Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2022: 4563-4572.

[14] HE K, GKIOXARI G, DOLLáR P, et al. Mask R-CNN [C] //Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2961-2969.

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

[16] LIAO M, WAN Z, YAO C, et al. Real-time scene text detectionwith differentiable binarization [C] //Proc. of the AAAI Conference on Artificial Intelligence, 2020: 11474-11481.

[17] YANG C, CHEN M, XIONG Z, et al. CM-Net: Concentric mask based arbitrary-shaped text detection [J]. IEEE Transactions on Image Processing, 2022, 31: 2864-2877.

[18] LIN T Y, DOLLAR P, GIRSHCKR B, et al. Feature pyramid net-works for object detection [C] //Proc. of the IEEE International Conference on Computer Vision, 2017: 2117-2125.

[19] LONG S B, RUAN J Q, ZHANG W J, et al. TextSnake: a flexible representation for detecting text of arbitrary shapes [C] //Proc. of the European Conference on Computer Vision, 2018: 19-35.

[20] DENG D, LIU H F, LI X L, et al, PixelLink: detecting scene text via instance segmentation [C] //Proc. of the AAAI Conference on Artificial Intelligence, 2018: 6773-6780.

[21] ZHANG C Q, LIANG B R, HUANGZ M, et al, Look more thanonce: an accurate detector for text of arbitrary shapes [C] //Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 10544-10553.

[22] 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.

[23] LIU Y, CHEN H, SHEN C, et al, ABCNet: real-time scene text spotting with adaptive Bezier-curve network [C] //Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 9809-9818.

[24] WANG W W, CHEN Y F, WU F, et al, TextRay: Contour-based geometric modeling for arbitrary-shaped scene text detection [C] //Proc. of the 28th ACM International Conference on Multimedia, 2020: 111-119.

[25] DAI P, ZHANG S, ZHANG H, et al, Progressive contour regress-ion for arbitrary-shape scene text detection [C] //Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2021: 7393-7402.

[26] KARATZAS D, GOMEZ-BIGORRA L, NICOLAS L, et al. ICDAR 2015 competition on robust reading [C] //Proc. of the 13th International Conference on Document Analysis and Recognition, 2015: 1156-1160.

[27] CHNG C K, CHAN C S. Total-Text: a comprehensive dataset forscene text detection and recognition [C] //Proc. of the 14th International Conference on Document Analysis and Recognition, 2017: 935-942.

[28] LIU Y L, JIN L W, ZHANG S T, et al. Detectingcurve text in the wild: newdataset and new solution [J]. ArXivPreprintArXiv: 1712. 02170, 2017.