双域感知下多方显式信息协同的场景端到端文本识别

陈平平,林 虎*,陈宏辉,谢肇鹏
(福州大学物理与信息工程学院,福建福州 350108)

摘 要: 在复杂自然场景的端到端文本识别中,由于文本和背景难以区分,文本检测的位置信息和识别的语义信息不匹配,无法有效利用检测和识别之间的相关性. 针对该问题,本文提出双域感知下多方显式信息协同的自然场景端到端文本识别方法(Multi-party Synergetic explicit Information with Dual-domain Awareness text spotting,MSIDA),通过强化文本区域特征和边缘纹理,利用文本检测和识别特征之间的协同作用提高端到端文本识别性能. 首先,设计融合文本空间和方向信息的双域感知模块(Dual-Domain Awareness,DDA),增强文本实例的视觉特征信息;其次,提出多方显式信息协同模块(Multi-party Explicit Information Synergy,MEIS)提取编码特征中的显式信息,通过匹配对齐用于检测和识别的位置、分类和字符多方信息生成候选文本实例;最后,协同特征通过解码器引导可学习的查询序列获得文本检测和识别的结果. 相比最新的 DeepSolo(Decoder with explicit points Solo)方法,在 Total-Text、ICDAR 2015 和 CTW1500 数据集上,MSIDA 模型的准确率分别提升 0.8%、0.8% 和 0.4%. 代码和数据集在 https://github.com/msida2024/MSIDA.git 可以获取.

关键词: 计算机视觉;场景文本图像;文本检测;端到端文本识别;特征信息关联

基金项目: 国家自然科学基金(No.62171135);福建省杰青项目(No.2022J06010);福建省教育厅重点攻关项目(No.2023XQ004);福州科技局项目(No.2023-P-001)

中图分类号: TP391.1;TP183 **文献标识码:** A **文章编号:** 0372-2112(2025)03-0974-12

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240919

End-to-End Scene Text Spotting Under Dual Domain Awareness Based on Multi-Party Synergetic Explicit Information

CHEN Ping-ping, LIN Hu*, CHEN Hong-hui, XIE Zhao-peng
(*College of Physics and Information Engineering*, *Fuzhou University*, *Fuzhou*, *Fujian* 350108, *China*)

Abstract: In the end-to-end text recognition of complex natural scenes, because text and background are difficult to distinguish, the location information detected by text and the semantic information recognized do not match, and the correlation between detection and recognition cannot be effectively utilized. In response to this problem, this paper proposes a multi-party synergetic information with dual-domain awareness text spotting (MSIDA). By enhancing text region features and edge textures, the synergies between text detection and recognition features are utilized to improve end-to-end text recognition performance. Firstly, a dual-domain awareness (DDA) module integrating text space and direction information is designed to enhance the visual feature information of text instances. Secondly, a multi-party explicit information synergy (MEIS) is proposed to extract explicit information from coding features and generate candidate text instances by matching and allocating the position, classification and character multi-party information used for detection and recognition. Finally, cooperative features guide learnable query sequences through decoders to obtain text detection and recognition results. Compared to the latest decoder with explicit points solo (DeepSolo) method, on the Total-Text, ICDAR 2015 and CTW1500 datasets, the accuracy of MSIDA improved respectively by 0.8%, 0.8% and 0.4%. The code and datasets are available at https://github.com/msida2024/MSIDA.git.

Key words: computer vision; scene text images; text detection; text spotting; feature information synergy

收稿日期:2024-10-16;修回日期:2025-02-25;责任编辑:覃怀银

*通讯作者:林虎

Foundation Item(s): National Natural Science Foundation of China (No.62171135); Fujian Talent Project (No.2022J06010); Project of Education Department (No.2023XQ004); Fuzhou Science and Technology Planning (No.2023-P-001)

1 引言

图像中的文本包含大量有效信息,可以根据不同目的对其进行检测和提取. 场景端到端文本识别将文本检测和文本识别整合到统一框架中,有效地提取视觉文本信息,广泛应用于自动驾驶[1]、智能导航[2]和生物医学[3]等领域. 然而,自然场景图像背景复杂,文本因色彩、尺寸、遮挡、方向等多种因素难以和背景区分,影响了视觉和语义信息的提取,增加文本识别的难度. 同时,文本检测和识别作为两个不同的子任务,如何处理好检测和识别之间的关系是端到端文本识别的难题. 近年端到端文本识别方法主要分为三类:基于感兴趣区域(Region-of-Interest, RoI)的方法、基于分割的方法和基于Transformer[4]的方法.

基于感兴趣区域的方法通常遵循先检测后识别的顺序,首先检测文本实例,然后利用基于感兴趣区域的连接器提取检测特征,最后输入到识别器中,相关流程如图1(*a*)所示. Li等人[5]研究了基于RoI的端到端文本检测和识别系统. TextDragon[6]、FOTS[7]利用连接器在两个任务共享计算和视觉信息,提取文本特征进行文本识别. Yao等人[8]则实现了对不同方向文本的检测,但仅局限于规则文本. ABCNet[9]在文本框生成中引入贝塞尔曲线实现了对弯曲文本的检测. 然而,尽管这些方法有一定的效果,但额外的连接器融合文本检测和识别导致位置信息和语义信息难以对齐,缺少检测和识别模块之间的协同作用[10],并且需复杂的后处理操作,如非极大值抑制(Non-Maximum Suppression, NMS).

基于分割的方法在具有共享主干网络的框架中,并行完成文本检测和识别两个子任务,如图1(*b*)所示. Lyu等人[11,12]引入字符分割模块,利用字符级注释来解决任意形状端到端文本识别问题,且不涉及NMS和RoI操作,保证了检测和识别的高效率. Xing等人[13]提出了卷积字符网络,可以有效处理多方向和弯曲文本. 然而,这些方法采用并行分支来分割字符和实例,容易受到噪声的干扰,且缺少各个子任务之间的交互.

基于Transformer[14]的方法是目前主流的端到端文本识别方法. TESTR[15]采用双解码器架构,共享主干和编码器的特征来增强检测和识别的交互,利用文本框位置显式信息建模可学习的查询序列,如图1(*c*)所示. 但是,两个子任务分别利用不同的可学习查询序列,以及不同的解码器并行完成检测和识别,缺少检测和识别的交互,并引入异质性[16]. Yair等人[17]将检测和识别任务统一到单解码器架构中,但需要额外的RNN模块. Ye等人[16]实现了更简洁统一的框架DeepSolo(Decoder with explicit points Solo),如图1(*d*)所示,使用单编码器和单解码器架构,利用具有显式信息的显式点引导检测和识别,但没有考虑到文本检测和识别之间不同的特征模式,且检测中的位置信息难以和识别中的语义信息匹配,导致识别区域和检测区域无法准确对齐. 同时,仅通过图像特征等隐式信息或单一任务的显式信息,文本检测无法利用文本识别的分类信息引导定位.

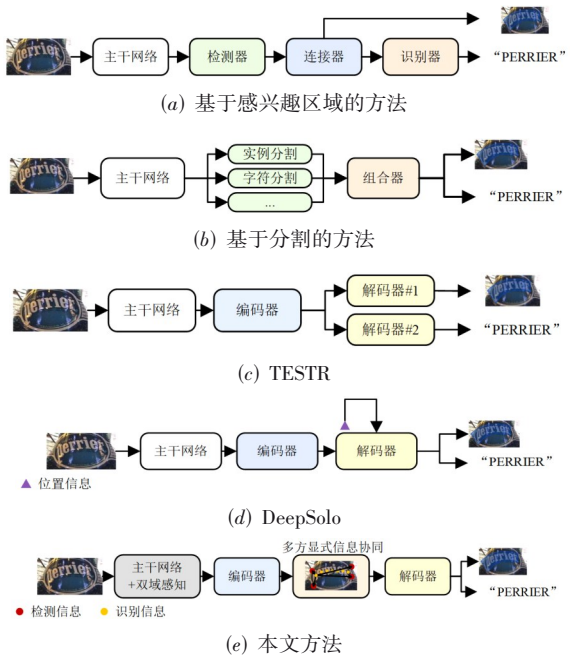


图1　相关端到端文本识别方法流程比较

针对以上问题,本文提出一种双域感知下多方显式信息协同的自然场景端到端文本识别方法(Multi-party Synergetic explicit Information with Dual-domain Awareness text spotting, MSIDA). 首先,采用双域感知模块(Dual-Domain Awareness, DDA)增强文本实例的空间感知和方向感知,关注图像中的文本内容,提高图像文本区域边缘纹理细节;其次,多方显式信息协同模块(Multi-party Explicit Information Synergy, MEIS)对编码特征提取文本检测和识别的显式信息,对齐两个任务不同形式的特征,得到包含位置信息和语义信息的协同特征. 利用文本检测的位置和形状信息,可以帮助提取识别文本内容. 同时,识别的位置和分类信息也可以引导检测区分不同的文本实例和背景. 最后,协同特征通过解码器引导可学习的查询序列,获得文本检测

和识别的结果，提高文本识别和端到端识别的准确性.

综上所述，本文所提MSIDA方法主要贡献可总结为以下三点：

(1)设计多方显式信息协同MEIS模块，通过编码特征中的文本位置、文本分类和字符分类信息，生成包含位置和语义信息的协同特征.

(2)设计双域感知DDA模块，使得本文方法更好地区分文本和背景，增强文本区域特征，突出文本实例边界纹理和细节.

(3)在公共数据集Total-Text、ICDAR2015和CTW1500的实验结果表明，MSIDA在检测准确率、端到端识别准确率方面都优于已有的代表性方法. 相比于DeepSolo方法，MSIDA在公开数据集Total-Text准确率提升了0.8%，在ICDAR 2015提升了0.8%，在CTW1500提升了0.4%.

2 本文方法

文本提出的MSIDA方法整体网络结构如图2所示. 首先，采用ResNet-50[18]主干网络对图像提取特征，在双域感知DDA模块中通过空间和方向感知增强文本区域的视觉特征，突出文本实例位置. 其次，编码器利用增强后的图像特征，预测一组文本实例的贝塞尔中心曲线，并对其进行N个点的均匀采样. 这些点经过协同MEIS模块生成文本的位置、分类以及字符信息，并编码为多方显式信息协同特征. 再次，解码器接收到协同特征后，可学习查询序列收集有效的文本特征. 最后，通过四个并行的多层感知机(Multi-Layer Perceptron，MLP)解决实例分类、字符分类、中心线采样点预测、边界点预测四个任务.

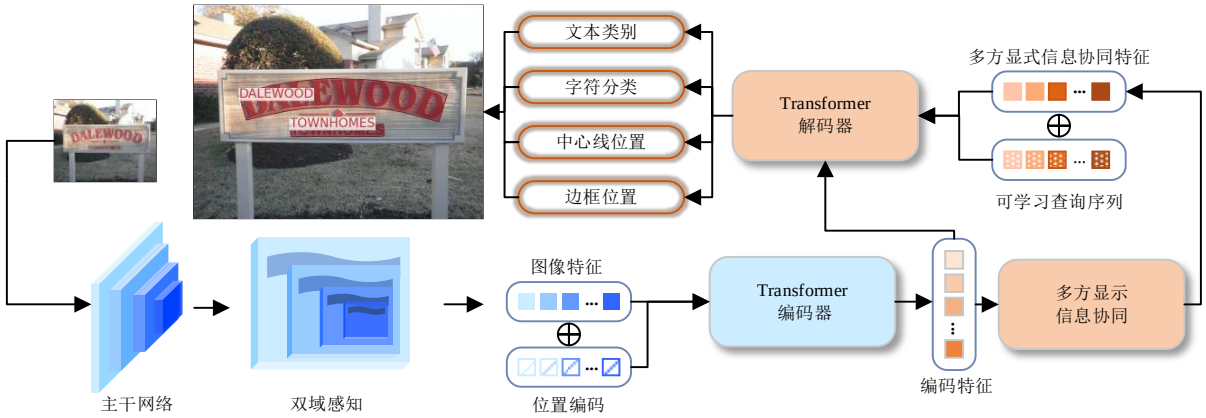


图2 整体网络结构

2.1 双域感知

双域感知通过融合空间感知和方向感知两个并行子网络，以输入特征$\boldsymbol{F}\in\mathbb{R}^{C\times H\times W}$，$C$为特征图通道数，$H$和$W$为高度和宽度，生成文本区域增强特征$\boldsymbol{F}_{\mathrm{TRE}}\in\mathbb{R}^{C\times H\times W}$，网络框架如图3所示.

为学习不同的语义，DDA将输入特征图划分为跨通道维度的G个子特征，分组可以表示为$\boldsymbol{F}=[\boldsymbol{F}_0,\boldsymbol{F}_i,\cdots,\boldsymbol{F}_{G-1}]$，$\boldsymbol{F}_i\in\mathbb{R}^{C//G\times H\times W}$，其中$G\ll C$.

在空间感知中，分组特征$\boldsymbol{F}_i$经由卷积核大小为3的卷积层提取语义，突出文本区域在全局空间中的位置. 为更好拟合线性变换，突出全局像素上下文关系，分别通过SoftMax函数和全局平均池化(Global Average Pooling，GAP)来获取图像特征$\boldsymbol{F}_{\mathrm{S1}}$和$\boldsymbol{F}_{\mathrm{A1}}$，计算过程如式(1)和式(2)所示.

$$\boldsymbol{F}_{\mathrm{S1}}=\delta\left(\mathrm{Conv}_{3\times3}(\boldsymbol{F}_i)\right) \tag{1}$$

$$\boldsymbol{F}_{\mathrm{A1}}=\mathrm{GAP}\left(\mathrm{Conv}_{3\times3}(\boldsymbol{F}_i)\right) \tag{2}$$

其中，$\boldsymbol{F}_{\mathrm{S1}}\in\mathbb{R}^{C//G\times H\times W}$为空间感知在通道维度$C$进行SoftMax的跨通道特征，$\boldsymbol{F}_{\mathrm{A1}}\in\mathbb{R}^{C//G\times1\times1}$为空间维度$H$和$W$进行全局平均池化的跨空间特征，$\delta$表示SoftMax函数.

在方向感知中，为突出字体边界，提升文本边界纹理细节，分组特征$\boldsymbol{F}_i$由水平方向和垂直方向的平均池化分解为两个一维的特征向量，经过拼接后再由卷积核大小为1的卷积层得到跨方向特征向量$\boldsymbol{F}_X,\boldsymbol{F}_Y$，计算过程如式(3)所示.

$$\boldsymbol{F}_X=\boldsymbol{F}_Y=\mathrm{Conv}_{1\times1}\left(\mathrm{Cat}\left(\mathrm{Xavg}(\boldsymbol{F}_i),\mathrm{Yavg}(\boldsymbol{F}_i)\right)\right) \tag{3}$$

其中，$\boldsymbol{F}_X\in\mathbb{R}^{C//G\times1\times W}$为水平方向的特征向量，$\boldsymbol{F}_Y\in\mathbb{R}^{C//G\times H\times1}$为垂直方向的特征向量，Cat表示拼接操作，Xavg表示沿水平方向的平均池化，Yavg表示沿垂直方向的平均池化.

跨方向特征向量$\boldsymbol{F}_X,\boldsymbol{F}_Y$经过Sigmoid函数加权到分组特征$\boldsymbol{F}_i$中，组间归一化后，同样分别通过SoftMax和GAP来获取视觉特征$\boldsymbol{F}_{\mathrm{S2}}$和$\boldsymbol{F}_{\mathrm{A2}}$，并与空间维度分支的特征相乘，最后得到加权后的文本区域增强特征$\boldsymbol{F}_{\mathrm{TRE}}$，

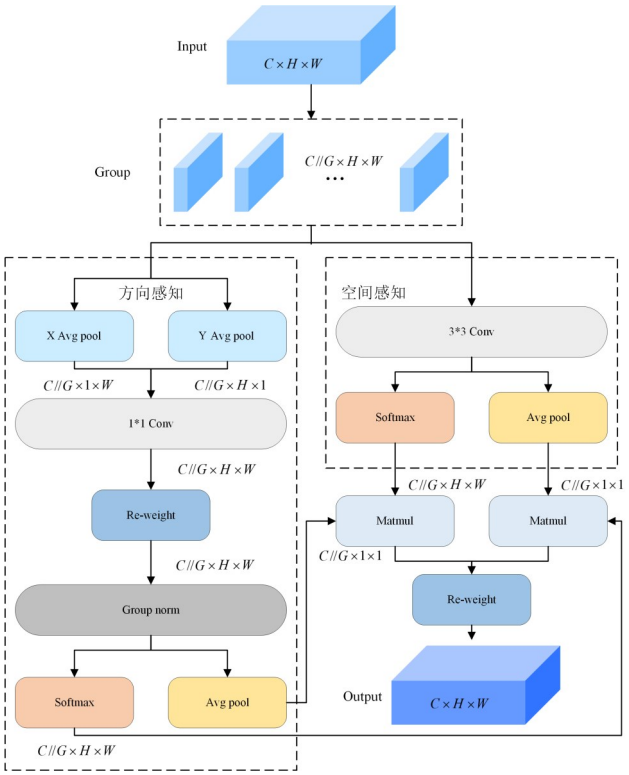


图3　双域感知模块网络框架

计算过程如式(4)~式(6)所示:

$$\boldsymbol{F}_{\mathrm{S2}}=\delta\left(\mathrm{Weight}\left(\boldsymbol{F}_i,\sigma(\boldsymbol{F}_X,\boldsymbol{F}_Y)\right)\right) \tag{4}$$

$$\boldsymbol{F}_{\mathrm{A2}}=\mathrm{GAP}\left(\mathrm{Weight}\left(\boldsymbol{F}_i,\sigma(\boldsymbol{F}_X,\boldsymbol{F}_Y)\right)\right) \tag{5}$$

$$\boldsymbol{F}_{\mathrm{TRE}}=\mathrm{Weight}\left(\mathrm{Mat}(\boldsymbol{F}_{\mathrm{S1}},\boldsymbol{F}_{\mathrm{A2}}),\mathrm{Mat}(\boldsymbol{F}_{\mathrm{S2}},\boldsymbol{F}_{\mathrm{A1}})\right) \tag{6}$$

其中,$\boldsymbol{F}_{\mathrm{S2}}\in\mathbb{R}^{C//G\times H\times W}$ 为方向感知中的跨通道特征,$\boldsymbol{F}_{\mathrm{A2}}\in\mathbb{R}^{C//G\times 1\times 1}$ 为方向感知中的跨空间特征,σ 表示 Sigmoid 函数,Weight 表示加权操作,Mat 表示矩阵相乘.

2.2　候选文本实例生成

与先前的端到端文本识别框架[19,20]不同,本文利用文本检测和识别的不同显式信息,协同引导预测基于贝塞尔曲线[21]的文本中心线[16],生成候选文本实例. 具体来说,为建立全局图像的长距离关系,捕捉文本与文本、文本与背景之间的关系,不同特征层经过文本区域增强得到图像特征 $\boldsymbol{F}_{\mathrm{TRE}}$,由可变形注意力[22]的 Transformer 编码器得到编码特征 $\boldsymbol{F}_{\mathrm{EN}}$:

$$\boldsymbol{F}_{\mathrm{EN}}=\mathrm{Encoder}\left(\mathrm{PE}\left(\mathrm{Flattened}(\boldsymbol{F}_{\mathrm{TRE}})\right)\right) \tag{7}$$

其中,$\boldsymbol{F}_{\mathrm{EN}}\in\mathbb{R}^{256\times HW}$,$HW$ 是特征图高和宽的乘积,Flattened 表示将特征在通道维度展开,PE 表示正弦位置编码函数,Encoder 表示编码器.

MEIS 模块结构如图 4 所示,对编码特征 $\boldsymbol{F}_{\mathrm{EN}}$ 的每个像素 i,提取用于检测的文本位置信息 T_{pos} 和文本分类信息 T_{class},用于识别的字符分类信息 T_{char},并通过两个子任务的信息协同交互生成 K 个候选文本实例,确定代表的中心线,得到多方显式信息协同特征.

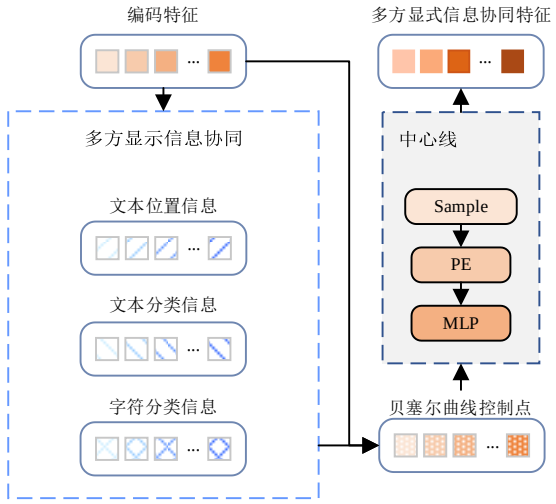


图4　多方显式信息协同结构图

文本位置信息 $T_{\mathrm{pos}}\in\mathbb{R}^{8\times HW}$ 通过由 MLP 预测 $\boldsymbol{F}_{\mathrm{EN}}$ 的每个像素 i 得到,包含像素 i 到四个贝塞尔曲线控制点 $P_i=\{p_{i_0},p_{i_1},p_{i_2},p_{i_3}\}$ 的偏移量信息 $\Delta P_i=\{\Delta p_{i_0},\Delta p_{i_1},\Delta p_{i_2},\Delta p_{i_3}\}$,从而确定一条代表文本实例边界的贝塞尔曲线. 本文采用了与 ABCNet[21] 相同的贝塞尔曲线控制点计算方法,并通过偏移量信息矫正坐标,计算如下:

$$p_{i_j}=\left(\sigma\left(\Delta p_{ix_j}+\sigma^{-1}(\hat{p}_{ix})\right),\sigma\left(\Delta p_{iy_j}+\sigma^{-1}(\hat{p}_{iy})\right)\right) \tag{8}$$

其中,$\hat{p}=\left(\hat{p}_{ix},\hat{p}_{iy}\right)\in[0,1]^2$ 表示像素的 i 二维归一化坐标,x,y 表示横纵坐标,$j\in\{0,1,2,3\}$ 表示四个贝塞尔曲线控制点的索引.

文本分类信息 T_{class} 包含像素 i 是否为文本区域的信息,通过使用线性层对 $\boldsymbol{F}_{\mathrm{EN}}$ 进行文本和非文本的分类,并计算置信度为

$$T_{\mathrm{class}}=\sigma\left(\mathrm{Linear}(\boldsymbol{F}_{\mathrm{EN}})\right) \tag{9}$$

其中,$T_{\mathrm{class}}\in\mathbb{R}^{1\times HW}$,1 是像素 i 为文本区域的概率,Linear 表示线性层.

字符分类信息 $T_{\mathrm{char}}\in\mathbb{R}^{38\times HW}$ 是由大小为 $|A|$ 的概率向量组成的长度为 HW 的序列,A 表示由‘0’到‘9’、‘A’到‘z’、‘unknow’和一个文本控制位(共 38 种情况)组成的分类集,包含候选文本中的字符分类信息和文本控制位信息. 利用检测信息和识别信息对文本分类的相关性,通过字符分类信息 T_{class} 的文本控制位重新对文本分类信息中的权重赋值,生成 K 个候选文本实例 T_k,其中 $k\in\{0,1,\cdots,K-1\}$,且 $K\ll i$,计算为

$$T_K=\mathrm{Sel}(P_i,T_{\mathrm{char}}\times T_{\mathrm{class}}) \tag{10}$$

其中,$T_{\mathrm{char}}=\mathrm{MLP}(\boldsymbol{F}_{\mathrm{EN}})$,$P_i$ 表示预测的所有贝塞尔曲线

控制点，Sel表示从所有控制点中选出置信度排名前K个的函数.

为检测任意形状和方向的文本，利用候选文本实例T_k生成中心线，并使用伯恩斯坦多项式[23]对中心线进行均匀采样N个点，得到中心线归一化坐标$T_{\text{coord}} \in \mathbb{R}^{K\times N\times 2}$. 通过进一步对该信息$T_{\text{coord}}$建模得到多方显式信息协同特征$M_q$，表示为

$$M_q = \text{ReLU}\left(\text{MLP}\left(\text{PE}(T_{\text{coord}})\right)\right) \tag{11}$$

其中，$M_q \in \mathbb{R}^{K\times N\times 256}$，PE表示正弦位置编码函数，ReLU表示激活函数.

2.3 可学习查询序列

为捕获不同文本实例之间的关系，通过将多方显式信息协同特征M_q与可学习查询序列L_q合并为复合查询序列$Q_q \in \mathbb{R}^{K\times N\times 256}$，在$K$个候选文本实例之间进行自注意力. 与Polynomials等人的研究[23,24]相同，首先使用跨维度N的自注意力机制，挖掘文本实例与可学习查询序列之间关系，其中，Keys和Queries相同，Values仅包含可学习信息，即$K_q = Q_q, V_q = L_q$. 其次，采用可变形交互注意力机制来融合更新后的复合查询序列Q_q和编码特征$\boldsymbol{F}_{\text{EN}}$中的多尺度文本特征，采样点$T_{\text{coord}}$引导可变形注意力机制生成参考点，得到解码特征$\boldsymbol{F}_{\text{DE}} \in \mathbb{R}^{K\times N\times 256}$.

2.4 端到端文本识别预测

$\boldsymbol{F}_{\text{DE}} \in \mathbb{R}^{K\times N\times 256}$已经具有多方显式协同信息，通过4个预测头从中提取对应子任务的信息：

(1)实例分类. 通过文本中心线上N个点的平均置信度作为文本实例的分类依据，利用线性层进行文本和非文本的分类.

(2)字符分类. 在文本中心线上均匀采样的N个点覆盖每个字符，对每个点利用线性层进行字符分类，排除没有覆盖字符的点.

(3)中心线采样点. 通过$\text{MLP}_{\text{coord}}$预测中心线采样点$T_{\text{coord}}$到中心线地面真实值的坐标偏移量.

(4)边界点. 通过$\text{MLP}_{\text{boundry}}$预测$T_{\text{coord}}$到文本实例上下边界线的地面真实值的坐标偏移量.

2.5 成本函数及损失函数

在二分匹配中，本文采用匈牙利算法[25]得到地面真实值Y和预测集$\hat{Y}$的最优内射函数φ:$[Y]\to[\hat{Y}]$，使得匹配成本C最小化，计算为

$$\arg\min_{\varphi} \sum_{g=0}^{G-1} C\left(Y^{(g)}, \hat{Y}^{(\varphi(g))}\right) \tag{12}$$

整体成本函数为

$$\begin{aligned} C = {} & \lambda_{\text{cls}}\text{FL}\left(\hat{b}^{(\varphi(g))}\right) + \lambda_{\text{char}}\text{CTC}\left(c^{(g)}, \hat{c}^{(\varphi(g))}\right) \\ & + \lambda_{\text{cd}} \sum_{n=0}^{N-1} \left\| p_n^{(g)} - \hat{p}_n^{(\varphi(g))} \right\| \end{aligned} \tag{13}$$

其中，G表示每张图像地面真实值的文本实例个数，λ_{cls}、λ_{char}和λ_{cd}为平衡任务的超参数，在第一项中采用与Lin等人研究[26,27]中类似的焦点损失FL(Focal Loss)，$\hat{b}^{(\varphi(g))}$表示纯文本实例的概率，第二项采用时间分类损失CTC(Connectionist Temporal Classification loss)[28]计算预测和真实字符之间的损失，第三项利用L_1范数计算中心线上预测点和地面真实值之间的差值.

整体损失函数L包括预测损失函数L_{pred}和多方显式信息辅助损失函数L_{aux}，如式(14)所示：

$$L = L_{\text{pred}} + L_{\text{aux}} \tag{14}$$

预测损失函数L_{pred}包括实例分类损失L_{cls}、字符分类损失L_{char}、中心线采样点损失L_{cd}和边界损失L_{bd}，计算过程如式(15)所示：

$$L_p = \sum_k \left(\lambda_{\text{cls}} L_{\text{cls}}^{(k)} + \lambda_{\text{char}} L_{\text{char}}^{(k)} + \lambda_{\text{cd}} L_{\text{cd}}^{(k)} + \lambda_{\text{bd}} L_{\text{bd}}^{(k)}\right) \tag{15}$$

其中，λ_{cls}、λ_{char}、λ_{cd}和λ_{bd}为4个超参数，表示对应损失的权重，k表示候选文本实例T_k的索引.

实例分类损失L_{cls}采用焦点损失FL来权衡文本实例保真度和识别性能，计算为

$$\begin{aligned} L_{\text{cls}}^{(k)} = {} & -\text{II}_{\{k\in \text{Im}(\varphi)\}} \alpha \left(1-\hat{b}^{(k)}\right)^{\gamma} \log\left(\hat{b}^{(k)}\right) \\ & -\text{II}_{\{k\notin \text{Im}(\varphi)\}} (1-\alpha) \left(\hat{b}^{(k)}\right)^{\gamma} \log\left(1-\hat{b}^{(k)}\right) \end{aligned} \tag{16}$$

其中，II表示指示函数，Im(φ)是映射φ所代表的图像.

字符分类损失L_{char}采用CTC损失来解决文本地面真实值和预测结果之间的长度不一致的问题，计算过程如式(17)所示：

$$L_{\text{char}}^{(k)} = \text{II}_{\{k\in \text{Im}(\varphi)\}} \text{CTC}\left(t^{(\varphi^{-1}(k))}, \hat{t}^{(k)}\right) \tag{17}$$

中心线采样点损失L_{cd}和边界损失L_{bd}采用L_1范数计算中心线，以及上下边界线采样点到真实文本框采样点的差值，计算如下：

$$L_{\text{cd}}^{(k)} = \text{II}_{\{k\in \text{Im}(\varphi)\}} \sum_{n=0}^{N-1} \left\| p_n^{(\varphi^{-1}(k))} - \hat{p}_n^{(k)} \right\| \tag{18}$$

$$\begin{aligned} L_{\text{bd}}^{(k)} = {} & \text{II}_{\{k\in \text{Im}(\varphi)\}} \sum_{n=0}^{N-1} \left(\left\| \text{top}_n^{(\varphi^{-1}(k))} - \hat{\text{top}}_n^{(k)} \right\| \right. \\ & \left. + \left\| \text{bot}_n^{(\varphi^{-1}(k))} - \hat{\text{bot}}_n^{(k)} \right\| \right) \end{aligned} \tag{19}$$

为加强多方显式信息的作用，增加文本检测和识别之间的交互，并使文本实例的贝塞尔曲线预测得更加准确，在信息协同中加入辅助监督的多方显式信息损失函数L_{aux}：

$$L_{\text{aux}} = \sum_i \left(\lambda_{\text{cls}} L_{\text{cls}}^{(i)} + \lambda_{\text{cd}} L_{\text{cd}}^{(i)}\right) \tag{20}$$

3 实验结果与分析

在本节中，首先介绍了实验中使用的数据集、评价指标和实现细节. 其次，将本文所提出的方法与其他方法进行比较. 最后，通过消融研究来验证该方法的有效性.

3.1 数据集

本文基于Total-Text[29]、ICDAR 2015(IC15)[30]和CTW1500[31]场景文本识别数据集来评估算法的性能. Total-Text是任意形状场景文本数据集,包含水平、多向和弯曲文本实例,由1 255张训练图像和300张测试图像组成. IC15是文档分析与识别国际会议(International Conference on Document Analysis and Recognition, ICDAR)公开的四边形场景文本数据集,包含1 000张训练图像和500张测试图像. CTW1500是任意形状长文本场景文本数据集,有1 000张训练图像和500张测试图像. 此外,本文额外采用以下数据集预训练:(1)Synth150K[9]合成数据集,包含94 723张多方向文本图像和54 327张带有弯曲文本的图像;(2)ICDAR 2017 MLT(MLT17)[32]多语言场景文本数据集;ICDAR 2013(IC13)[33],包含229张带有水平文本的训练图像. 本文实验采用的所有对比模型都使用相同的预训练方式.

3.2 评价指标

本文基于上述数据集中使用的标准评价指标评估该方法的场景文本检测和端到端场景文本识别性能,包括召回率(Recall,R)、准确率(Precision,P)和F得分(F-score,F).

3.3 实验细节

本文算法模型基于PyTorch 1.8实现. 所有实验均在4块NVIDIA GeForce RTX 3090 GPU上进行,显存为96 GB,图像批量大小为8. 该模型使用ResNet-50[18]作为主干网络,多尺度特征来自主干网络的最后三个阶段,并未使用FPN[34]. Transformer的参数与先前的研究类似[22],可变形注意力机制的头部$H=8$,采样点$K=4$,使用6层编码器和解码器. 没有特殊说明的情况下,文本实例中心线采样点$N=25$. 模型采用AdamW[35]优化器进行训练,β_1和β_2分别设置为0.9和0.999,并将整体损失函数C以及预测损失函数L_{pred}中的损失权重λ_{cls}、λ_{char}、λ_{cd}和λ_{bd}分别设置为1.0、1.0、0.5和0.5,将实例分类损失L_{cls}采用的焦点损失FL中的α和γ分别设置为0.25和2.0.

3.4 主流模型性能对比分析

为验证MSIDA在任意形状场景文本、多方向场景文本和长文本上的有效性,本文在Total-Text[29]、ICDAR 2015(IC15)[30]和CTW1500[31]文本识别数据集上评估算法的性能,与现有的主流端到端文本识别模型进行比较,包括SwinTextSpotter[10]、TESTR[15]和DeepSolo[16]等. 此外,为对比本文方法与现有主流方法的模型复杂度,本文将MSIDA与其他方法在模型参数量方面进行评估.

在Total-Text数据集上与现有主流方法的对比结果如表1所示,其中,加粗字体表示为最佳值,下划线表示为次优值,None表示在无词典的端到端文本识别F得分,Full表示在有词典的F得分,"*"表示该模型基于Transformer架构.

表1　与现有主流方法在Total-Text数据集上的对比

模型	预训练数据集	文本检测			None	Full
		P	R	F		
TextDragon[6]	Synth800K	85.6	75.7	80.3	48.8	74.8
CharNet[13]	Synth800K	88.6	81.0	84.6	63.6	—
CRAFTS[36]	Synth800K+IC13	89.5	**85.4**	87.4	78.7	—
Mask TextSpotter v3[12]	Synth800K+IC13+IC15+SCUT	—	—	—	71.2	78.4
PGNet[37]	Synth800K+IC15	85.5	86.8	86.1	63.1	—
PAN++[38]	Synth800K+COCO-Text+MLT17+IC15	—	—	—	68.6	78.6
ABCNet v2[21]	Synth150K+MLT17	90.2	<u>84.1</u>	87.0	70.4	78.1
TESTR[15]*	Synth150K+MLT17	<u>93.4</u>	81.4	86.9	73.3	83.9
SwinTextSpotter[10]*	Synth150K+MLT17+IC13+IC15	—	—	87.2	72.4	83.0
SPTS[39]*	Synth150K+MLT17+IC13+IC15	—	—	—	74.2	82.4
TTS[17]*	Synth800K+COCO-Text+IC13+IC15+SCUT	—	—	—	78.2	86.3
DeepSolo[16]*	Synth150K+MLT17+IC13+IC15	93.1	82.1	87.3	<u>79.7</u>	<u>87.0</u>
LATextSpotter[40]	Synth150K+MLT17	**94.1**	81.9	<u>87.6</u>	76.6	84.8
MSIDA(ours)*	Synth150K+MLT17+IC13+IC15	93.0	83.0	**87.7**	**80.5**	**87.5**

在无词典和有词典的情况下,MSIDA的端到端文本识别F得分均优于现有主流模型. 在无词典时,虽然缺少了词典对文本识别的引导,但本文方法可以利用MEIS模块中的位置信息,其端到端文本识别F得分比TESTR、SwinTextSpotter和SPTS模型分别提升了7.2%、6.2%和6.3%. 对比最新模型DeepSolo和LATextSpotter[40]分别提升了0.8%和3.9%,并且MSIDA在使用更少训练数据的情况下仍然优于TTS模型2.3%. 相比于其他基于Transformer的方法,在有词典的情况下,MSIDA能突出图片的文本特征,准确率达到最高,相较

DeepSolo 和 LATextSpotter 分别提高了 0.5% 以及 2.7%. 在文本检测中，由于 DDA 模块增强文本区域边界，MSIDA 的 F 得分也达到 87.7% 的最佳值. 实验结果表明，所提方法有效提高文本检测和端到端文本识别的准确性，并适用于任意形状场景文本识别. MSIDA 在 Total-Text 数据集上的可视化结果和如图 5 所示. 其中，第一行为原始场景文本图像，第二行为本文方法可视化识别结果.

此外，图 6 展示了 MSIDA 与现有主流模型 DeepSolo 的端到端文本识别结果对比. 其中，第一张图片为 MSIDA 方法的识别结果，第二张图片为对比方法 DeepSolo 的识别结果. 可以看到，本文方法的 DDA 能突出文本区域特征，识别出 DeepSolo 未能识别的“BARKS”文本. 在 MEIS 协同检测与识别的作用下，能正确识别出 DeepSolo 识别错误的“YOUK”词汇，验证了本文方法的有效性.

图 5 MSIDA 在 Total-Text 数据集上的可视化结果

图 6 MSIDA 与 DeepSolo 端到端识别可视化结果对比

在 ICDAR 2015 数据集上与现有主流方法的对比结果如表 2 所示，S 表示每张图像使用独立词典的端到端文本识别 F 得分，词典包含 100 个对应图像单词和少量其他图像干扰词. W 表示使用包含数据集所有单词词典的 F 得分，G 表示使用包含大约 90 000 个单词的通用词典的 F 得分.

表 2 与现有主流方法在 ICDAR 2015 数据集上的对比

模型	预训练数据集	文本检测			端到端文本识别		
		P	R	F	S	W	G
TextDragon[6]	Synth800K	92.5	83.8	87.9	82.5	78.3	65.2
CharNet[13]	Synth800K	91.2	88.3	89.7	80.1	74.5	62.2
CRAFTS[36]	Synth800K+IC13	89.0	85.3	87.1	83.1	82.1	74.9
Mask TextSpotter v3[12]	Synth800K+IC13+IC15+SCUT	—	—	—	83.3	78.1	74.2
PGNet[37]	Synth800K+IC15	91.8	84.8	88.2	83.3	78.3	63.5
PAN++[38]	Synth800K+COCO-Text+MLT17+IC15	—	—	—	82.7	78.2	69.2
ABCNet v2[21]	Synth150K+MLT17	90.4	86.0	88.1	82.7	78.5	73.0
TESTR[15]*	Synth150K+MLT17	90.3	**89.7**	90.0	85.2	79.4	73.6
SwinTextSpotter[10]*	Synth150K+MLT17+IC13+IC15	—	—	—	83.9	77.3	70.5
SPTS[39]*	Synth150K+MLT17+IC13+IC15	—	—	—	77.5	70.2	65.8
TTS[17]*	Synth800K+COCO-Text+IC13+IC15+SCUT	—	—	—	85.2	81.7	77.4
DeepSolo[16]*	Synth150K+MLT17+IC13+IC15	92.8	87.4	90.0	86.8	81.9	76.9
LATextSpotter[40]	Synth150K+MLT17	**93.6**	85.9	89.6	85.6	79.5	75.1
MSIDA(Ours)*	Synth150K+MLT17+IC13+IC15	91.5	89.1	**90.3**	**87.6**	**82.6**	**78.0**

结果表明，MSIDA 与主流模型相比，在主要指标文本检测以及端到端文本识别 F 得分上均达到最佳. 在端到端文本识别任务中，与基于 Transformer 框架并且使用相同训练数据集的方法相比，本文方法在使用独立词典(S)、数据集词典(W)和通用词典(G)的情况下，相较于 DeepSolo 分别提高了 0.8%、0.7% 和 1.1%. 虽然在通用词典(G)条件下，词典词汇量扩大，有效先验识别信息占比下降，但 MSIDA 可以利用文本检测任务中的显式信息，协同识别信息引导端到端文本识别，性能仍保持优势. 在文本检测任务中，虽然 LATextSpotter 的

准确率为最优值,但MSIDA方法中DDA模块可以引导模型关注文本区域,因此综合指标F得分最高.

在CTW1500数据集上与现有主流方法的对比结果如表3所示. 结果表明,即使在长文本数据集中,MSIDA依然超过现有主流方法. 在文本检测任务中,本文方法准确率相较ABCNet v2和TESTR分别提升了4.8%和2.4%. 在无词典的情况下,相较同为Transformer架构的次优DeepSolo提高了0.4%. 进一步验证本文方法准确性和在长文本检测识别中的有效性.

表3　与现有主流方法在CTW1500数据集上的对比

模型	文本检测			None	Full
	P	*R*	*F*		
TextDragon[6]	84.5	82.8	83.6	39.7	72.4
ABCNet v2[21]	85.6	83.8	84.7	57.5	77.2
TESTR[15]*	92.0	82.6	87.1	56.0	**81.5**
DeepSolo[16]*	—	—	—	64.2	81.4
MSIDA(ours)*	**92.2**	**87.0**	**89.5**	**64.6**	**81.5**

表4给出了本文方法与最近的DeepSolo[16]等方法的模型参数量对比. 在使用主干网络ResNet-50时,MSIDA模型参数量仅超过SPTS方法. 结合表1的实验结果,所提方法的模型参数量与对比方法保持同一数量级,但性能上依旧具有优势.

表4　与现有主流方法模型参数量的对比

模型	Backbone	Parameters/M
Mask TextSpotter[11]	ResNet-50-FPN	45.5
SPTS[39]	ResNet-50	36.5
SwinTextSpotter[10]	Swin Transformer	113.7
DeepSolo[16]	ResNet-50	42.5
MSIDA(Ours)	ResNet-50	40.9

3.5　消融实验

本节在Total-Text数据集上进行了多组实验,调制预测损失函数L_{pred}中影响检测和识别性能的字符分类损失权重λ_{char}和边界损失权重λ_{bd}. 同时,以浮点运算数(FLOating Point operations,FLOPs)为指标,评估模型在不同输入情况下的计算复杂度. 为了证明所提出的算法的有效性与可行性,本节还设计了四组消融实验评估DDA模块和MEIS模块在端到端文本识别的影响,以及可视化DDA模块在视觉和文本边界突出中所起的引导作用.

3.5.1　损失函数优化

本小节研究了不同λ_{char}和λ_{bd}取值对模型性能的影响. 在Total-Text数据集上的实验如表5和表6所示. 可以看到λ_{char}和λ_{bd}对检测和识别性能有直接的影响. 其中,加粗字体表示最佳.

表5　不同λ_{char}取值在Total-Text数据集上的实验结果

λ_{char}	文本检测			None	Full
	P	*R*	*F*		
0.25	92.9	82.8	87.5	79.8	87.3
0.50	**93.0**	**83.0**	**87.7**	**80.5**	**87.5**
0.75	93.5	81.1	86.9	79.7	86.8
1.00	93.3	81.3	86.9	80.0	86.9

表6　不同λ_{bd}取值在Total-Text数据集上的实验结果

λ_{bd}	文本检测			None	Full
	P	*R*	*F*		
0.25	93.0	82.0	87.2	79.8	87.1
0.50	93.0	**83.0**	**87.7**	**80.5**	**87.5**
0.75	**93.4**	82.2	87.4	80.1	87.3
1.00	93.5	81.7	87.2	**80.5**	87.2

3.5.2　双域感知模块的有效性

在本小节,为评估DDA模块在端到端场景文本识别结果的影响,使用不同的感知方法进行对比分析. 在Total-Text数据集上的消融实验结果如表7所示. 其中,空间感知表示仅使用空间特征突出文本实例位置,方向感知表示仅使用方向特征增强文本边缘纹理.

表7　不同感知方法在Total-Text数据集上的实验结果

增强方法	文本检测			None	Full
	P	*R*	*F*		
空间感知	92.9	82.7	87.5	79.9	87.1
方向感知	92.6	82.3	87.2	80.0	86.9
DDA	**93.0**	**83.0**	**87.7**	**80.5**	**87.5**

根据表7的实验结果显示,无词典情况下,由于方向感知能够有效区分文本区域和背景,使用DDA模块比仅利用空间感知模型的准确率提升0.6%. 此外,与仅利用方向感知相比,DDA模块能够利用空间信息准确定位文本区域,识别准确率提升了0.5%. 实验证明,DDA模块可以有效利用空间和方向特征,感知图像中的文本区域,提高场景文本边缘的细节纹理.

3.5.3　多方显式信息协同的有效性

本小节评估了由不同信息引导端到端文本识别对实验结果的影响. 在Total-Text数据集上设置了三组对比实验,即使用识别语义信息、检测位置信息和多方协同显式信息,其他设置均相同. 消融实验结果如表8所示.

表8　不同信息引导识别在Total-Text数据集上的实验结果

引导信息	文本检测			None	Full
	P	*R*	*F*		
识别语义信息	92.4	80.8	86.2	78.4	86.4
检测位置信息	**93.4**	78.1	85.1	78.1	85.2
MEIS	93.2	**82.4**	**87.5**	**79.2**	**86.7**

由实验结果可知，仅采用检测位置信息比采用MEIS的文本检测的召回率降低4.3%，F得分降低2.4%. 这是由于仅利用位置信息的模型缺少语义信息，在识别过程中容易出现误判和漏判，从而导致召回率明显降低. 对比仅使用识别语义信息的实验组，所提协同MEIS模块在无词典条件下准确率提升了0.8%. 结果表明，MEIS模块可以有效利用检测和识别之间的特征相关性，提升文本检测和端到端文本识别的准确率.

3.5.4 双域感知和多方显式信息协同消融实验

本节评估了所提DDA模块和MEIS模块对实验结果的影响. 在Total-Text和ICDAR2015数据集上设置不同的消融实验组，并将其与完整网络的准确率和浮点运算数进行比较. 消融实验结果如表9和表10所示，“√”和“×”分别表示包含和不包含该模块.

表9 DDA和MEIS在Total-Text数据集上的消融实验结果

实验组	DDA	MEIS	文本检测			None	Full	FLOPs/G
			P	R	F			
1	×	×	93.4	78.1	85.1	78.1	85.2	370.8
2	√	×	**94.0**	80.8	86.9	78.9	86.5	371.1
3	×	√	93.2	82.4	87.5	79.2	86.7	386.6
MSIDA(Ours)	√	√	93.0	**83.0**	**87.7**	**80.5**	**87.5**	387.0

表10 DDA和MEIS在ICDAR2015数据集上的消融实验结果

实验组	DDA	MEIS	文本检测			S	W	G
			P	R	F			
1	×	×	93.4	86.1	89.6	86.3	81.6	76.4
2	√	×	**94.0**	86.7	90.2	86.4	82.2	77.9
3	×	√	92.3	87.3	90.1	87.1	82.1	77.5
MSIDA(Ours)	√	√	91.5	**89.1**	**90.3**	**87.6**	**82.6**	**78.0**

在实验组1中，本文同时删除了DDA模块和MEIS模块，降低了视觉特征提取过程中对文本实例的视觉特征提取能力，减弱检测与识别任务之间的交互. 实验选用单一的位置显式信息进行端到端文本识别过程的引导. 在Total-Text数据集上的结果表明，无词典的情况下，实验组1与所提出的MSIDA相比准确率下降了2.4%. 其次，在实验组2中删除了MEIS模块，保留DDA模块增强文本区域的特征信息. 可以看到，在有词典情况下，相较实验组1提升了1.3%，且仅增加了少量FLOPs. 最后，在实验组3中删除了DDA模块，直接使用主干网络图像特征输入编码器，保留MEIS模块利用检测和识别任务的不同信息. 在IC15数据集的结果表明，独立词典(S)情况下，本组实验由于缺少对文本区域的感知增强，准确率对比本文方法下降了0.5%. 实验结果证明了DDA模块和MEIS模块的有效性和必要性.

3.6 可视化结果

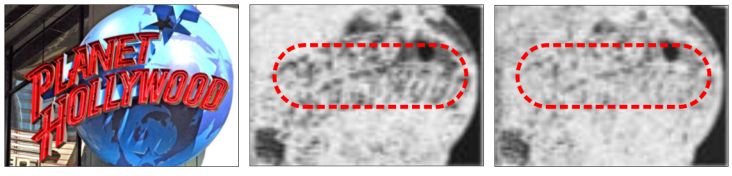
本节将MSIDA中的DDA模块和MEIS模块进行可视化，并给出了本文方法下的错分样本. 将DDA输出的增强特征 $\boldsymbol{F}_{\mathrm{TRE}} \in \mathbb{R}^{C\times H\times W}$ 通过热力图的形式渲染，得到突出文本区域的可视化结果.

图7(a)为选自Total-Text数据集的原始图片，图7(b)为经过DDA模块的特征图，图7(c)为ResNet-50直接输出的特征图，高亮部分表示该区域特征权重占比大. 由可视化结果验证，特征图经过DDA模块后有效提取文本实例特征，突出场景中的文本区域.

其次，本节还通过灰度图的形式，验证DDA模块提高文本边缘细节纹理的有效性，可视化结果如图8所示.

(a) Total-Text原始图片 (b) DDA热力图 (c) ResNet-50热力图

图7 DDA突出文本区域的可视化效果



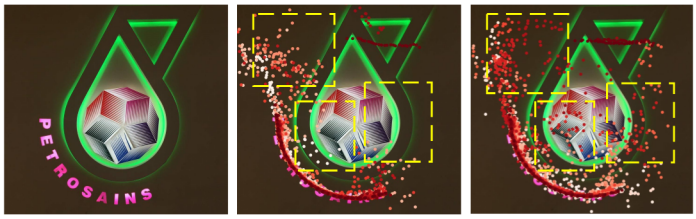
(a) Total-Text原始图片 (b) DDA灰度图 (c) ResNet-50灰度图

图8 DDA提高文本边缘细节的可视化效果

由灰度图可视化对比结果可知，特征图经过DDA模块的方向感知后，可以有效提高文本边缘的细节纹理，将文本和背景区域有效分离，有利于进行下游的识别任务.

此外,本节通过可视化MEIS模块验证其有效性,通过颜色变化可视化候选文本实例中心线采样点的权重,可视化结果如图9所示.其中,颜色越深的点表示权重越大.

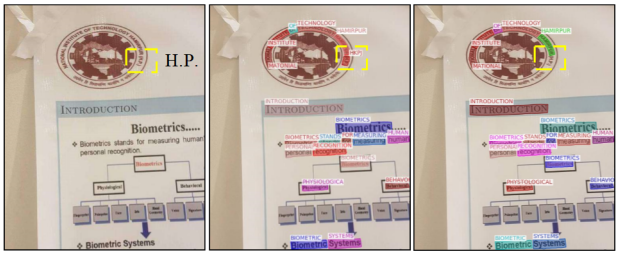
图9(*a*)为选自Total-Text数据集的原始图片,图9(*b*)为经过MEIS模块选取的候选文本实例中心线采样点,图9(*c*)为经过Transformer输出的候选文本实例中心线采样点.由图9可知,MEIS同时利用检测和识别信息,更好地锁定文本实例位置,使中心线采样点向待检测文本聚拢,减少无意义的候选文本实例,使得文本区域内的候选点权重更大,提高端到端文本识别准确率.



(*a*) Total-Text原始图片　(*b*) MEIS采样点　(*c*) Transformer采样点

图9　多方显式信息协同模块可视化效果

最后,在Total-Text上的错分样本如图10所示.其中,图10(*a*)为数据集原图,图10(*b*)为MSIDA识别结果,图10(*c*)为DeepSolo识别结果.相较于DeepSolo,本文MSIDA能够更好检测出目标文本实例.但在其方向垂直并且有轻微翻转角度的情况下,本文方法将"H.P."误识别为"HKPJ".因此,针对垂直和翻转文本还有研究和提高的空间.



(*a*) Total-Text原始图片　(*b*) MSIDA　(*c*) DeepSolo

图10　MSIDA在Total-Text数据集上的错分样本

4　结论

本文提出了一种MSIDA场景端到端文本识别网络,利用显式信息增强端到端网络架构中检测和识别交互.在视觉特征提取方面,设计DDA模块,融合空间分支和方向分支的图片特征,突出场景文本区域并提高边缘细节.通过MEIS模块提取文本实例位置、分类、字符等显式信息,匹配对齐检测任务中的位置信息和识别任务中的语义信息,引导可学习序列,形成检测和识别之间的协同交互特征.在Total-Text、ICDAR 2015和CTW1500数据集上的实验结果表明,本文提出的MSIDA有效提高文本检测和端到端识别的准确性.协调检测和识别任务之间的关系是端到端文本识别任务中的主要挑战,本文方法还有研究空间,以适应更加复杂的场景文本,如垂直文本和翻转文本.因此,未来工作将进一步研究该类文本数据的检测识别.

参考文献

[1] ZHANG C S, TAO Y F, DU K, et al. Character-level street view text spotting based on deep multisegmentation network for smarter autonomous driving[J]. IEEE Transactions on Artificial Intelligence, 2022, 3(2): 297-308.

[2] DESOUZA G N, KAK A C. Vision for mobile robot navigation: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(2): 237-267.

[3] 孟伟伦, 郭景峰, 邢珂萱, 等. 基于字形特征的中文医学命名实体识别方法[J]. 电子学报, 2024, 52(6): 1945-1954.
MENG W L, GUO J F, XING K X, et al. A Chinese medical named entity recognition method based on glyph features[J]. Acta Electronica Sinica, 2024, 52(6): 1945-1954. (in Chinese)

[4] 黄俊炀, 陈宏辉, 王嘉宝, 等. 多域字符距离感知的场景文本图像超分辨率重建[J]. 电子学报, 2024, 52(7): 2262-2270.
HUANG J Y, CHEN H H, WANG J B, et al. Scene text image super-resolution reconstruction based on perceiving multi-domain character distance[J]. Acta Electronica Sinica, 2024, 52(7): 2262-2270. (in Chinese)

[5] LI H, WANG P, SHEN C H. Towards end-to-end text spotting with convolutional recurrent neural networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 5248-5256.

[6] FENG W, HE W H, YIN F, et al. TextDragon: An end-to-end framework for arbitrary shaped text spotting[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 9075-9084.

[7] LIU X B, LIANG D, YAN S, et al. FOTS: Fast oriented text spotting with a unified network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Piscataway: IEEE, 2018: 5676-5685.

[8] YAO C, BAI X, LIU W Y, et al. Detecting texts of arbitrary orientations in natural images[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2012: 1083-1090.

[9] LIU Y L, CHEN H, SHEN C H, et al. ABCNet: Real-time scene text spotting with adaptive bezier-curve network[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 9809-9818.

[10] HUANG M X, LIU Y L, PENG Z H, et al. SwinTextSpotter: Scene text spotting *via* better synergy between text detection and text recognition[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 4583-4593.

[11] LIAO M H, LYU P Y, HE M H, et al. Mask TextSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2): 532-548.

[12] LIAO M H, PANG G, HUANG J, et al. Mask TextSpotter V3: Segmentation proposal network for robust scene text spotting[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 706-722.

[13] XING L J, TIAN Z, HUANG W L, et al. Convolutional character networks[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Neural Information Processing Systems, 2017, 30: 1-9.

[15] ZHANG X, SU Y W, TRIPATHI S, et al. Text spotting transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 9509-9518.

[16] YE M Y, ZHANG J, ZHAO S S, et al. DeepSolo: Let transformer decoder with explicit points solo for text spotting[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 19348-19357.

[17] YAIR KITTENPLON, INBAL LAVI, SHARON FOGEL, et al. Towards weakly-supervised text spotting using a multi-task transformer[EB/OL]. (2022-02-14)[2025-03-11]. https://arxiv.org/abs/2202.05508.

[18] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[19] JIA D, YUAN Y H, HE H D, et al. DETRs with hybrid matching[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 19702-19712.

[20] 邹北骥, 郭建京, 朱承璋, 等. 基于自适应色彩聚类和上下文信息的自然场景文本检测[J]. 电子学报, 2018, 46(6): 1436-1444.
ZOU B J, GUO J J, ZHU C Z, et al. Natural scene text detection based on adaptive color clustering and context information[J]. Acta Electronica Sinica, 2018, 46(6): 1436-1444. (in Chinese)

[21] LIU Y L, SHEN C H, JIN L W, et al. ABCNet v2: Adaptive bezier-curve network for real-time end-to-end text spotting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(11): 8048-8064.

[22] ZHU X, SU W, LU L, et al. Deformable DETR: Deformable transformers for end-to-end object detection[EB/OL]. (2021-03-18)[2025-03-11]. https://arxiv.org/abs/2010.04159.

[23] POLYNOMIALS B. Introduction to the Mathematics of Computer Graphics[M]. RhodeIsland: American Mathematical Society, 2016.

[24] YE M Y, ZHANG J, ZHAO S S, et al. DPText-DETR: Towards better scene text detection with dynamic points in transformer[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(3): 3241-3249.

[25] KUHN H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics Quarterly, 1955, 2(1/2): 83-97.

[26] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2999-3007.

[27] 师硕, 覃嘉俊, 于洋, 等. 基于改进ConvMixer和动态焦点损失的视听情感识别[J]. 电子学报, 2024, 52(8): 2824-2835.
SHI S, QIN J J, YU Y, et al. Improved ConvMixer and focal loss with dynamic weight for audio-visual emotion recognition[J]. Acta Electronica Sinica, 2024, 52(8): 2824-2835. (in Chinese)

[28] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification[C]//Proceedings of the 23rd International Conference on Machine Learning. New York: ACM, 2006: 369-376.

[29] CHENG C K, CHAN C S, LIU C L. Total-Text: Toward orientation robustness in scene text detection[J]. Interna-

tional Journal on Document Analysis and Recognition (IJDAR), 2020, 23(1): 31-52.

[30] KARATZAS D, GOMEZ-BIGORDA L, NICOLAOU A, et al. ICDAR 2015 competition on robust reading[C]//2015 13th International Conference on Document Analysis and Recognition (ICDAR). Piscataway: IEEE, 2015: 1156-1160.

[31] LIU Y L, JIN L W, ZHANG S T, et al. Curved scene text detection via transverse and longitudinal sequence connection[J]. Pattern Recognition, 2019, 90: 337-345.

[32] NAYEF N, YIN F, BIZID I, et al. ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification - RRC-MLT[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Piscataway: IEEE, 2017: 1454-1459.

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

[34] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 936-944.

[35] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[EB/OL]. (2019-01-04) [2025-3-11]. https://arxiv.org/abs/1711.05101.

[36] BAEK Y, SHIN S, BAEK J, et al. Character region attention for text spotting[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 504-521.

[37] WANG P F, ZHANG C Q, QI F, et al. PGNet: Real-time arbitrarily-shaped text spotting with point gathering network[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(4): 2782-2790.

[38] WANG W H, XIE E Z, LI X, et al. PAN++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 5349-5367.

[39] PENG D Z, WANG X Y, LIU Y L, et al. SPTS: Single-point text spotting[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 4272-4281.

[40] LI Z C, QU Y D, XIE H T, et al. LATextSpotter: Empowering transformer decoder with length perception ability[C]//2024 IEEE International Symposium on Circuits and Systems (ISCAS). Piscataway: IEEE, 2024: 1-5.

作者简介

陈平平　男,1986年出生于福建省泉州市.现为福州大学电子信息工程系教授,博士生导师.主要研究方向为信息处理、人工智能与计算机视觉.中国电子学会会员编号:E190021215M.
E-mail: ppchen.xm@gmail.com

林　虎　男,2001年出生于福建省三明市.现为福州大学物理与信息工程学院研究生.主要研究方向为计算机视觉、场景文本检测、场景文本端到端识别.
E-mail: linhu_noah@outlook.com

陈宏辉　男,1998年出生于福建省南平市.现为福州大学物理与信息工程学院研究生.主要研究方向为计算机视觉、场景文本检测、场景文本端到端检测识别.
E-mail: 726673517@qq.com

谢肇鹏　男,1995年出生,现为福州大学先进制造学院讲师.主要研究方向为强化学习,信道编码与无线通信等.
E-mail: xzp_fzu@163.com