基于 Transformer 解码的端到端场景文本检测与识别算法

郑金志[1,2]，汲如意[1,2]，张立波[1,3]，赵琛[1,3]

（1. 中国科学院软件研究所智能软件研究中心，北京 100190；2. 中国科学院大学，北京 100190；
3. 中国科学院软件研究所计算机科学国家重点实验室，北京 100190）

摘 要：针对任意形状的场景文本检测与识别，提出一种新的端到端场景文本检测与识别算法。首先，引入了文本感知模块基于分割思想的检测分支从卷积网络提取的视觉特征中完成场景文本的检测；然后，由基于 Transformer 视觉模块和 Transformer 语言模块组成的识别分支对检测结果进行文本特征的编码；最后，由识别分支中的融合门融合编码的文本特征，输出场景文本。在 Total-Text、ICDAR2013 和 ICDAR2015 基准数据集上进行的实验结果表明，所提算法在召回率、准确率和 F 值上均表现出了优秀的性能，且时间效率具有一定的优势。

关键词：文本检测；文本识别；端到端；Transformer

中图分类号：TP391

文献标志码：A

DOI: 10.11959/j.issn.1000–436x.2023070

End-to-end scene text detection and recognition algorithm based on Transformer decoders

ZHENG Jinzhi[1,2], JI Ruyi[1,2], ZHANG Libo[1,3], ZHAO Chen[1,3]

1. Intelligent Software Research Center, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
2. University of Chinese Academy of Sciences, Beijing 100190, China
3. State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China

Abstract: Aiming at the detection and recognition task of arbitrary shape text in scene, a novelty scene text detection and recognition algorithm which could be trained by end-to-end algorithm was proposed. Firstly, the detection branch of text aware module based on segmentation idea was introduced to detect scene text from visual features extracted by convolutional network. Then, a recognition branch based on Transformer vision module and Transformer language module encoded the text features of the detection results. Finally, the text features encoded by the fusion gate in the recognition branch were fused to output the scene text. The experimental results on the three benchmark datasets of Total-Text, ICDAR2013 and ICDAR2015 show that the proposed algorithm has excellent performance in recall, precision, F-score, and has certain advantages in efficiency.

Keywords: text detection, text recognition, end-to-end, Transformer

0 引言

文字作为人类知识保存与传播的主要手段，是人类最具有影响力的创造之一，是人类文明的基石[1]。如何准确、高效地阅读与理解场景图像中的文本，成为计算机视觉领域一个重要的研究课题。采用手工设计文本视觉特征的传统算法在处理自然场景文本检测与识别时，视觉特征的提取依赖研究人员的经验。因此，这类算法具有较大的局限性，且缺乏鲁棒性[2]。一方面，场景文本检测与识别作为图像视觉理解任务中的重要研究内容[3]，在视觉理解、盲人辅助、自动驾驶、图像检索、文本图像描述、文本视觉问答、视觉导航等领域都能得到广泛的应用；另一方面，随着计算

收稿日期：2022–10–29；**修回日期**：2023–01–31
通信作者：汲如意，jrylovezd@gmail.com

机科学的发展，计算机的性能得到了较大提升，这使基于神经网络的深度学习成为可能[4]。而深度学习的迅速发展，为场景文本检测与识别任务性能提升提供了可观的前景。因此，研究基于深度学习的场景文本检测与识别算法具有一定的理论与现实意义。

从场景图像中读取文本作为图像视觉理解任务中的一项重要研究课题，其可以划分为 3 个子任务，即场景文本检测、场景文本识别以及端到端的场景文本检测与识别[1-5]。场景文本检测是定位出场景图像中的文本区域，提供给后续视觉任务[6]。场景文本识别是识别出文本图像区域或图像块中的文本内容[7-11]。在场景文本检测与识别过程中，可以先使用场景文本检测方法定位包含文本的区域，然后切割文本区域的图像块，将该图像块输入场景文本识别模型识别出文本。但是这种方式在文本检测与识别阶段都需要对场景图进行视觉提取，难以做到视觉特征的共享，存在重复计算的问题，而且需要对检测和识别模型进行分别训练，训练过程也较复杂。为了缓解这些问题，端到端的场景文本检测与识别任务便引起了广泛关注[12-14]。通常，端到端的场景文本检测与识别模型是将场景文本检测与识别分支结合起来，通过端到端的方式完成整个模型的训练。端到端的场景文本检测与识别是在一个模型中完成自然场景中的文本检测与识别任务，在检测与识别过程中实现视觉特征共享，不必进行重复提取。端到端的场景文本检测与识别是本文研究的主要内容。

端到端的场景文本检测与识别分为基于像素级别的分割与基于序列到序列的文本生成 2 种思路。基于像素分割的场景文本检测与识别方法需要对每个像素的文本类别进行预测，可以并行识别场景图像中的多个字符或文字[6-7]。但是，这种方法通常识别精度较低，效果较差。因此，有些研究人员将场景文本检测与识别任务视为视觉特征到文本字符序列的生成任务，通过循环神经网络（RNN, recurrent neural network）对视觉特征进行解码的方式完成文本的识别[12-15]。基于循环神经网络的方法对场景文本依次解码，存在误差累积、识别速度受限等问题。

尽管场景文本的检测与识别任务已经发展了很多年，取得了较大进步，但是仍然面临众多挑战，主要原因在于：自然场景中的文本本身存在字体大小不一、形状任意多变、位置随机、文本方向不定等情况；场景图像存在遮挡、畸变、弯曲、颜色失真、光照不均、分辨率低、背景复杂等问题[16]。

为了缓解速度与精度之间的矛盾，本文提出一种能够进行并行解码的端到端的场景文本检测与识别模型，该模型能够处理任意形状文本的检测与识别。本文的主要贡献概括如下。

1) 设计了包含文本感知模块的检测分支，该模块能够在文本视觉提取过程中增强文本前景特征、抑制背景噪声、提高文本视觉特征的表达能力。

2) 提出了由 Transformer 视觉模块（TVM, Transformer vision module）、Transformer 语言模块（TLM, Transformer language module）和融合门组成的识别分支，在充分提取视觉特征的基础上进一步挖掘了语义信息。视觉模块与语言模块通过位置编码进行并行识别，具有解码速度快的优点。

3) 设计了新的端到端的场景文本检测与识别模型，由基于分割思想的检测分支和基于 Transformer 与融合门的识别分支组成。基于分割思想的检测分支能够实现像素级别的文本检测精度，而基于 Transformer 的识别分支改变了 RNN 循环解码方式，提高了解码速度。

4) 实验结果表明，本文算法在任意形状文本数据集 Total-Text 上实现了具有竞争性的识别性能。在无字典约束和全字典约束情况下，文本识别的 F 值分别达到了 70.9%和 78.1%。此外，该算法具有较高的时间效率。

1　相关工作

本节将从场景文本检测、场景文本识别以及端到端的场景文本检测与识别 3 个方面对当前的发展进行简要介绍。

1.1　场景文本检测

场景文本检测的难点主要在于场景文本的尺寸多变、角度不定以及背景复杂等。文献[17]通过检索字符以及字符之间的依附关系有效地检测文本区域，取得了不错的效果，该方法能够检测具有一定弯曲程度的文本。文献[18]提出的深度关系推理图（DRRG, deep relational reasoning graph）网络将每个文本实例划分为一系列的矩形组件；然后通过定义文本组件对象的长、宽、角度等几何属性建立关系图，引入图神经网络进行关系推理，从而完

成任意形状文本的检测。

基于分割的方法能够精确到像素级别预测，因此其在任意形状的文本检测任务中得到了较广泛的关注。文献[19]提出了基于分割的框架通过嵌入聚类对任意形状的文本进行分割预测。在分割预测过程中，首先对文本的前景区域进行掩码分割，然后在文本前景掩码范围内部预测文本的中心区域。每个文本中心区域代表一个文本实例，最后对每个文本区域进行全图分割，全图是对整个文本区域进行完整的预测。文献[20]采用了与文献[19]类似的方法，不同的是文献[20]用文本实例的边框预测替代了文献[19]中的文本全图预测。文献[21]针对任意形状场景文本的检测任务提出了基于分割的实时上下文感知（RSCA, real-time segmentation-based context-aware）模型。文献[3]提出了增强特征金字塔网络（EFPN, enhanced feature pyramid network）模型，该模型设置了语义传递比率不变的特征增强模块和改善边界位置的重建空间分辨率模块。文献[22]设计了可微二值化模块，使模型在分割推荐过程中的二值化阈值具有更强的鲁棒性。

1.2 场景文本识别

场景文本识别从场景文本检测的文本图像块中提取视觉特征，然后通过解码识别出文本内容，这与机器翻译等自然语言处理任务比较类似。因此，可以将其看作一个特征序列到文本序列的编码解码生成任务[8-9]。最初，常见方法的主要思想是使用 RNN 从文本图像区域的视觉特征中解码出文本内容。例如，文献[8]中提出的模型由提取视觉特征的残差网络（ResNet, residual neural network）编码器和基于 2D 注意力的长短时记忆（LSTM）解码器模型组成。在近阶段的进展中，Transformer 逐渐得到了广泛关注。例如，文献[23]使用卷积网络和 Transformer 作为编码器对视觉特征进行编码，然后使用 Transformer 作为解码器对编码特征进行解码；文献[24]直接使用 Transformer 对卷积网络提取的视觉特征进行解码输出识别文本。

有些研究人员认为场景文本中蕴含着文本语义知识，因此文本识别过程中既要考虑视觉信息，又要提取隐藏的文本语义信息。基于这一思想，文献[9-11]提出的文本识别模型将识别过程分为 2 个阶段：首先，基于视觉特征初步识别文本；然后，考虑文本上下文对初步识别的文本进行二次修正，修正后的结果作为模型的最终识别结果。例如，文献[9]提出的语义推理网络（SRN, semantic reasoning network）设置了并行视觉注意力模块，基于视觉信息进行初始识别；然后，通过全局语义推理模块在视觉识别的文本之间挖掘语义信息进行推理。文献[10]提出的 RobustScanner 由卷积神经网络（CNN, convolutional neural network）编码器和解码器组成，其中解码器中设置了位置增强分支与混合分支进行两阶段的场景文本识别。文献[11]提出的自主双向迭代网络（ABINet, autonomous, bidirectional and iterative network）设置了视觉模块和语言模块，可以进行视觉和语言两阶段识别。

1.3 端到端的场景文本检测与识别

端到端的场景文本检测与识别算法主要分为基于像素分割预测和基于 RNN 序列生成两类。基于像素分割预测的算法通过对场景文本的前景分类预测完成文本的检测与识别。例如，文献[25]提出的 Text perceptron 首先基于分割的思想进行文本检测，然后通过设置的形状转化模型将不规则文本转化为规则文本，最后通过识别网络进行识别。文献[26]提出的点集网络（PGNet, point gathering network）设计了对文本中心线、文本边界偏移、文本方向偏移、文本字符分类的多目标任务，规避了感兴趣区域（ROI, region of interest）和非极大值抑制（NMS, non-maximum suppression）操作。这类算法能够适应任意形状的文本，但是对文本字符关系的编码有限，识别精度较低。

与基于像素分割预测的算法相比，基于 RNN 的算法能够更好地编码文本字符之间的关系，从而提升文本的识别精度。例如，文献[5]提出的模型首次将 CNN 和 RNN 应用到端到端的场景文本检测与识别任务。文献[27]引入感兴趣区域旋转提取文本区域的特征，然后输入由 CNN 和 LSTM 组成的文本识别分支，识别场景文本。为了能够对自然场景图中的不规则文本进行端到端的检测与识别，有些工作通过引入具有尺度感知能力的注意力机制，提取多尺度的图像视觉特征，然后使用 RoI 进行特征对齐，并使用 RNN 对对齐后的特征进行解码生成识别文本[28-30]。例如，文献[28]在 RNN 分支中设置了文本对齐层和具有字符注意力机制的 LSTM 循环模块。文献[29]通过分割掩码的方式完成文本检测，通过基于 LSTM 的识别器完成对文本的识别。文献[30]提出的掩码注意力引导一阶段（MANGO, mask attention guided one-stage）文本检测与识别框架设

计了位置感知掩码注意力（PMA, position-aware mask attention）模块，将图像中不同文本实例映射到不同的特征通道中，省去了 RoI 操作。

此外，还存在另一类将分割与循环解码生成文本进行组合的方法[13,15]。例如，文献[13]在字符分割的基础上并行设置了空间注意力模块，使用基于空间注意力的门控循环单元（GRU, gated recurrent unit）进行文本解码能够在一定程度上挖掘文本字符之间的语义信息，提升场景文本的识别精度。

在上述端到端的场景文本检测与识别算法中，基于像素分割预测的算法能够精确到像素级别检测，但是由于缺乏对文本语义信息的挖掘，识别精度具有一定的局限性；而基于 RNN 解码的序列生成算法在循环中对序列进行解码，时间效率较低。为了能够在保持算法识别性能的同时提高识别效率，本文提出了一种基于 Transformer 编码的场景文本检测与识别算法。该算法使用并行 Transformer 代替 RNN 挖掘文本语义信息，提升了识别精度。

2　算法设计

本文算法框架结构如图 1 所示，主要由主干网络、检测分支与识别分支三部分组成。主干网络提取视觉特征，检测分支从视觉特征中定位出文本区域，识别分支识别出文本区域包含的文本内容。图 1 中虚线箭头只在训练过程中存在。检测分支包含分割推荐网络、检测监督器两部分。识别分支包含 TVM、TLM 和融合门三部分。该模型的检测与识别过程概括为给定待识别的场景图像，首先，由主干网络提取视觉特征；然后，由检测分支基于分割的思想生成文本区域的推荐；最后，根据检测分支的分割推荐模块和文本感知模块（TAM, text aware module）输出视觉特征，由基于 Transformer 解码器的识别分支完成文本的识别。

2.1　主干网络和检测分支

主干网络的主要功能是提取场景图像的视觉特征。检测分支中设置了分割推荐网络和检测监督器。分割推荐网络中基于视觉特征在 U-Net 各层进行融合操作后，设置了文本感知模块和分割推荐模块，目的是更好地提取文本视觉特征、完成文本检测任务；检测监督器的作用主要是训练过程中监督分割推荐网络的学习，而前向推理过程中监督器不参与计算。

2.1.1　主干网络

本文使用 ResNet50[31]作为主干网络提取视觉特征。给定场景图像 I，可通过主干网络提取视觉特征

$$R_v = R(I) \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C} \tag{1}$$

其中，H 和 W 是原始输入图像 I 的高和宽，C 是特征通道的维度，R 是 ResNet50。

2.1.2　分割推荐网络

1) 特征融合

如图 1 所示，检测分支使用 U-Net 结构的卷积模块对主干网络提取的视觉特征进行融合。在不同尺度上进行特征图的融合，从而使融合后的特征具有较强的尺度鲁棒性。在融合特征的过程中，4 个不同尺度的特征分别由一个 3×3 的卷积层和一个上采样层进行尺度归一化，然后对归一化的特征进行拼接（concat），获得融合特征。融合特征定义为

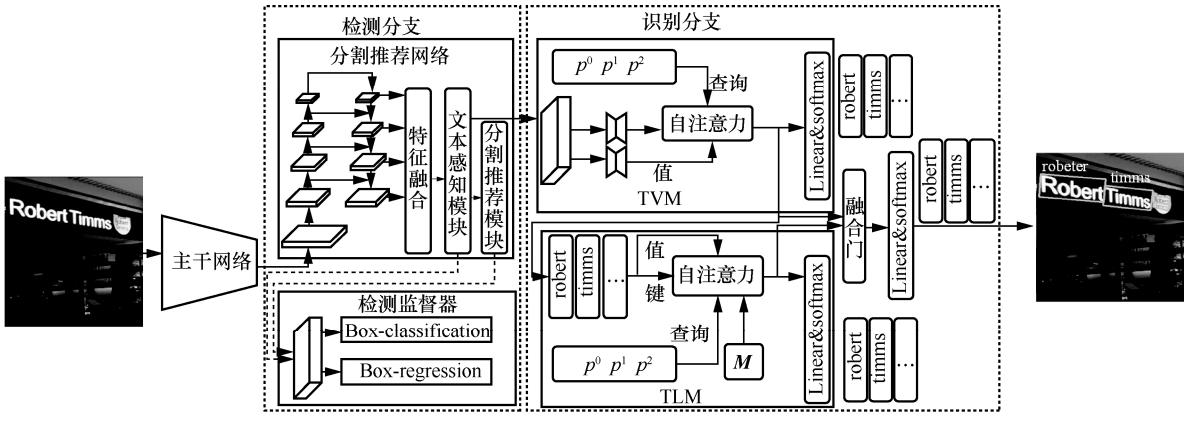


图 1　本文算法框架结构

$$F = \text{U-Net}(R_v) \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C} \tag{2}$$

2) 文本感知模块

自然场景图像中的文本通常具有任意形状，包括具有一定的方向、弯曲、形变等情况。受到文献[21]的启发，为了在视觉特征提取过程中能更好地感知文本区域，提高文本视觉特征的提取质量，本文在视觉特征融合模块后设置了如图 2 所示的文本感知模块。图 2 中，每个操作旁边标注的是对应操作后视觉特征的维度信息，δ 为第一个卷积后特征图的缩小比例，本文在实验中设置 δ=4。

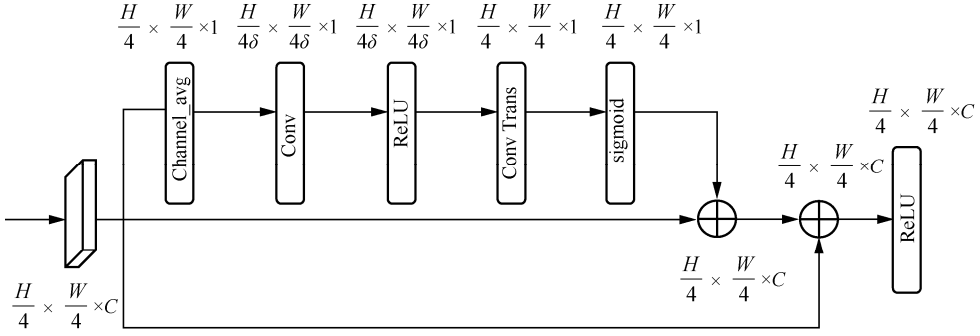


图 2　文本感知模块

文本感知模块的处理过程可以分为三步：感知权重计算、文本空间感知和残差链接增强。感知权重计算过程如下：对融合特征进行通道方向上的池化操作；在池化后的单通道特征图上进行卷积操作和 ReLU 激活，主要目的是提取特征空间之间的非线性关系；通过反卷积操作将特征图恢复到原始特征图的尺寸，提取每个空间位置上的文本注意力；将恢复后的特征图通过 sigmoid 激活函数提取感知权重。文本空间感知过程是将感知权重与融合特征进行广播乘操作，然后将乘积结果与融合特征进行残差链接，进行 ReLU 激活增强，输出文本感知特征。融合特征通过文本感知模块获得文本感知特征，可定义为

$$F_{\text{tam}} = T(F) \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C} \tag{3}$$

3) 分割推荐模块

对于弯曲、形变等不规则文本，传统区域推荐网络（RPN, region proposal network）[5-6]生成的文本区域会存在重叠、相互干扰等情况，从而影响文本识别的精度。文献[15]中提出的基于分割的文本推荐网络能够更加精细地生成相邻的文本区域，缓解相邻文本之间的干扰，因此，在文本感知模块之后，本文使用了与文献[15]类似的结构。在文本感知特征的基础上，定义文本分割特征图为

$$S = S(F_{\text{tam}}) \in \mathbb{R}^{H\times W\times 1} \tag{4}$$

其中，分割模块最后一层为 sigmoid 层，即 S 的取值范围为[0,1]。从文本感知特征 F_{tam} 中获得分割特征图 S 的推荐结构如图 3 所示。

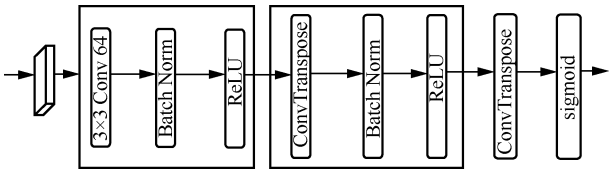


图 3　分割特征图 S 的推荐结构

对分割特征图 S 进行二值化，输出场景图二值化的分割结果为

$$B_{i,j} = \begin{cases} 1, & S_{i,j} > t \\ 0, & \text{其他} \end{cases} \tag{5}$$

其中，$B_{i,j}$ 为分割图在 (i,j) 位置上二值化的结果，$S_{i,j}$ 为分割图在 (i,j) 位置上的取值，t 为二值化的阈值，本文在实验中设置 t=0.5。B 中的连通域为文本区域。

场景图中相邻文本可能存在相互连通的情况[32-33]，不利于文本分割推荐的划分，因此在生成分割特征图 S 的过程中，需要对文本的训练标签进行压缩，从而使不同的文本处于相互隔离状态。基本思想是每个文本按一定比例向中心区域压缩，从而使不同的文本区域分离。本文采用了与文献[15,33-34]相同的策略，使用文献[35]中提出的 Vatti 裁剪算法，通过在文本区域周围裁剪数量为 v 的像素实现提取文本中心区域的目的。裁剪像素 v 由裁剪系数 r、多边形的面积 s 和周长 p 决定，即 $v = \dfrac{s(1-r^2)}{p}$，实

验中 r 设置为 0.4。

分割特征图 S 的二值化 B 对应文本中心区域。因此，获得 B 后，需要再次使用 Vatti 裁剪算法对文本中心区域进行扩大处理，从而恢复出完整的文本区域。Vatti 裁剪算法扩大文本区域时的像素偏移量 $v'=s'\frac{r'}{p'}$。其中，s'、p'、r' 分别是 B 中文本连通域的面积、周长、膨胀系数，本文实验中膨胀系数设置为 3.0。

如图 4 所示，分割推荐网络训练时文本区域真实标签的生成以及前向推理时的文本区域推荐过程可概括为：首先，通过 Vatti 裁剪算法缩小文本区域，获得分割推荐的真实标签，如图 4(b)所示，对分割推荐网络进行训练；其次，对分割特征图 S 进行二值化得到 B，如图 4(c)所示；最后，使用 Vatti 裁剪算法膨胀文本中心区域，输出多边形文本区域推荐，如图 4(d)所示。

这种基于分割的推荐方法更适合形状多变的不规则文本，尤其是场景中文本较为稠密的情况。获得文本多边形区域后，为了提取文本的视觉特征，同时抑制背景以及相邻文本的干扰，需要对文本的 ROI 特征进行掩码处理。掩码生成的过程如下：多边形的分割推荐最小水平外接矩形中，多边形内的像素视为文本前景设置为 1，多边形外的像素视为背景或噪声设置为 0，连通域输出为文本推荐掩码 R_m。

2.1.3　检测监督器

由卷积层、池化层和全连接层组成的 Fast R-CNN[36]是一种较高效的视觉目标检测模型，其中全连接层以池化特征作为输入可以完成对应的分类和回归任务。如图 1 所示，受到文献[15]的启发，本文在模型训练过程中使用 Fast R-CNN 作为文本检测模块的检测监督器，对分割推荐网络进行检测监督。

2.2　识别分支

本文将端到端的场景文本检测与识别任务看成视觉特征到字符序列的文本生成任务。如图 1 所示，本文模型在完成文本检测之后，将文本分割推荐结果与文本感知模块的视觉特征输入文本识别分支，识别出场景文本。识别分支由基于 Transformer 的视觉模块、语言模块和融合门三部分组成。

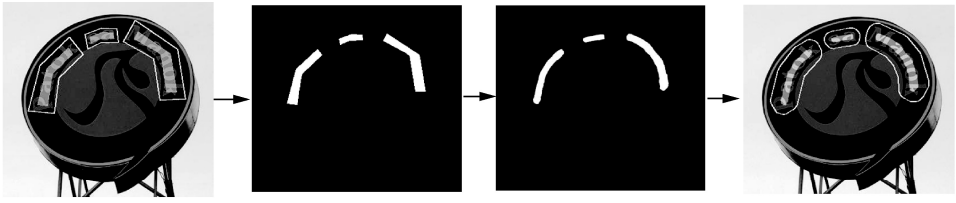
基于 Transformer 的序列解码器能够不依赖历史解码信息进行并行训练，相比于 RNN 解码器具有解码速度快、并行能力强等优点。因此，本文在识别分支设置了 Transformer 进行解码。Transformer 视觉模块基于视觉特征对文本特征进行解码，完成文本初始识别；Transformer 语言模块从 Transformer 视觉模块的文本初始识别中挖掘语义信息，使最终的识别能够在文本序列中感知上下文信息，提升文本的识别精度。融合门的主要作用是对视觉模块和语言模块提取的文本特征进行融合，从而在识别过程中能充分而全面地考虑视觉特征与语义特征。

2.2.1　Transformer 视觉模块

文本由字符按一定顺序排列组成，文本中字符序列的位置信息对文本的识别具有重要作用，包含语义信息。因此，解码过程中使用字符序列的一维位置编码信息对提升识别精度具有重要意义。本文采用 cos 编码方式对位置信息进行编码[10-11]，位置编码可表示为

$$\mathrm{PE}_{(\mathrm{pos},2i)}=\sin\left(\frac{\mathrm{pos}}{10\,000^{\frac{2i}{\dim}}}\right) \tag{6}$$

$$\mathrm{PE}_{(\mathrm{pos},2i+1)}=\cos\left(\frac{\mathrm{pos}}{10\,000^{\frac{2i}{\dim}}}\right) \tag{7}$$



(a) 文本区域真实标签　(b) 文本中心区域的标签掩码　(c) 文本中心区域的推荐掩码　(d) 分割推荐

图 4　真实标签与前向生成分割推荐的过程示意

其中，sin 表示正弦函数，cos 表示余弦函数，pos 表示字符在文本中序列的位置索引，dim 表示位置向量的维度（本文设置为 512），$2i+1$表示奇数维的索引，$2i$ 表示偶数维的索引。

如图 1 所示，识别分支中的 Transformer 视觉模块由式(6)和式(7)的位置编码结果作为 Transformer 的查询 Query，文本感知模块的文本视觉特征 F_{tam} 通过 2 个 U-Net 网络作为键 Key、值 Value。Transformer 视觉模块解码的文本字符特征可表示为

$$g_v = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{C}}\right)V \tag{8}$$

其中，Q、K 和 V 分别为查询 Query、键 Key 和值 Value 的缩写，K^{T} 是 K 的转置。这里的 Transformer 模块设置了一层 Transformer 单元，该层有 8 个注意力头。

然后，将解码的字符特征通过线性变换与 softmax 激活函数，输出 Transformer 视觉模块的文本识别结果

$$P_v = \text{softmax}(T_v) \tag{9}$$

其中，P_v 为视觉模块的识别结果，$T_v = W_v(g_v)$，W_v 为可训练的超参数。

Transformer 视觉模块的主题思想是将位置编码信息作为查询，从视觉特征中解码文本内容。但是字符视觉特征较差，比如出现遮挡或者模糊等情况时，识别效果会受到限制。而文本通常包含一定的语义信息，当字符的视觉特征不足以正确识别出该字符时，可以根据字符的上下文挖掘语义信息对当前字符进行修正识别。为此，本文在视觉模块后面设置了语言模块。

2.2.2 Transformer 语言模块

Transformer 语言模块的主要目的是从基于视觉识别的结果中挖掘语义信息，进一步优化识别结果。受到文献[11]工作的启发，Transformer 语言模块将语义信息的挖掘视为双向填空问题，使用自掩码 Transformer[11]挖掘语义概念。

与视觉模块类似，将位置编码作为查询 Query，视觉模块的文本解码特征 g_v 作为键 Key、值 Value。通过位置编码信息从视觉模块编码的文本特征 g_v 中挖掘语义信息解码场景文本。如图 1 所示，在 Transformer 语言模块的自注意力编码中设置了掩码矩阵 $\boldsymbol{M}$，该矩阵对角线上的元素为负无穷，非对角线上的元素为 0。掩码矩阵的主要作用是模拟填空网络，在编码过程中屏蔽当前字符，而只考虑上下文信息。编码过程可以表示为

$$g_L = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{C}} + \boldsymbol{M}\right)V \tag{10}$$

其中，Q 是字符在序列中的位置编码，K、V 是键 Key、值 Value。这里的 Transformer 模块本文设置了 4 层 Transformer 单元，每层 8 个注意力头。Transformer 语言模块的识别可表示为

$$P_L = \text{softmax}(F_L) \tag{11}$$

其中，P_L 为语言模块的识别结果，$F_L = W_l(g_L)$，W_l 为可训练的超参数。

2.2.3 融合门

为了充分融合视觉特征与语义特征进行识别，模型中设置了融合门[9-11]。如图 1 所示，将 Transformer 视觉模块和语言模块编码的文本特征输入融合门，将融合门的输出特征进行线性变换，然后由激活函数输出识别结果。特征融合过程可表示为

$$\begin{cases} W_f = \sigma([g_v, g_L]W_g) \\ F_g = W_f g_v + (1 - W_f) g_L \end{cases} \tag{12}$$

其中，σ 表示 sigmoid 激活函数，g_v 和 g_L 分别由式(8)和式(10)可得，$[g_v, g_L]$ 表示特征 g_v 和 g_L 的拼接，W_g 和 W_f 分别表示可训练的超参数和融合权重。

识别分支的识别结果可表示为

$$P_F = \text{softmax}(F_f) \tag{13}$$

其中，P_F 为识别分支的识别结果，$F_f = W_r(F_g)$，W_r 为可训练的超参数。

2.3 多目标损失函数

为了训练端到端的模型，本文设计了多目标损失函数

$$L = \alpha L_{\text{spn}} + \beta L_{\text{det}} + \gamma L_{\text{rec}} \tag{14}$$

其中，L_{spn} 是基于分割的文本推荐损失函数，L_{det} 是检测监督器的损失函数，L_{rec} 是文本识别分支的损失函数，α、β、γ 是平衡因子，取值分别为 1、0.1、1。

对于基于分割推荐的任务，本文采用 dice 损失[37]，该损失定义为

$$L_{\text{spn}} = 1 - \frac{2\sum(SG)}{\sum S + \sum G} \tag{15}$$

其中，S 为分割网络输出的分割特征图，G 为分割目标图，$\sum(SG)$ 为分割图与目标图的交，$\sum S + \sum G$ 为分割图与目标图的并。

检测监督器是一个 Fast R-CNN[28]，因此检测监督器的损失 L_{det} 的定义同文献[28]。识别分支的损失主要由 Transformer 视觉模块、Transformer 语言模块和融合门三部分的损失组成。识别分支的损失可表示为

$$L_{\text{rec}} = \gamma_v L_v + \gamma_L L_L + \gamma_F L_F \tag{16}$$

其中，L_v、L_L 和 L_F 分别是 g_v、g_L 和 F_g 对应的交叉熵损失函数，γ_v、γ_L 和 γ_F 是平衡因子。

3　实验

为了验证本文算法的有效性，本节在几个基准数据集上进行了实验验证，并对实验结果进行了分析。此外，本节还对本文的实验设置进行了说明，并对实验结果进行了展示和分析。

3.1　实验设置

本节对实验用到的主要数据集、实验参数、训练策略以及评价指标等进行介绍。

3.1.1　数据集

SynthText[38]数据集收集了约 80 万张场景图像，场景图像中的文本由渲染合成而来，并不是真实的场景文本。合成数据集在训练阶段应用。使用合成数据集进行训练的原因是真实场景中样本的注释较困难，工作量大，无法获得足够的真实场景样本进行训练，而通过渲染获得的合成数据集可以降低人工标注的成本，一定程度上缓解训练样本不足的问题。

ICDAR2013[39]数据集包含 299 张图像的训练集和 233 张图像的测试集。数据集中包含英文文本，文本以水平方向为主，包含文本和字符 2 种级别的标注。为了验证本文算法在多方向文本上的性能，实验中对 ICDAR2013 数据集进行了不同角度的旋转。

Total-Text[40]数据集中训练集含有图像 1 255 张，测试集含有图像 300 张。该数据集的场景图像包含水平文本、多方向文本和弯曲文本。

ICDAR2015[41]数据集中训练集含有图像 1 000 张，测试集含有图像 500 张。该数据集的样本收集于谷歌眼镜，图像中的文本存在畸变、模糊、分辨率低、文本较小等情况，场景更具偶发性。

SCUT（scut-eng-char）数据集[42]仅用于模型训练，主要是为了增加训练样本的多样性和数量，包含自然场景下室内文本图像。为了尽可能地保证公平比较，实验中使用 Mask TextSpotter v2[13]官网提供的 SCUT 数据集，该数据集包含 1 162 张场景文本图像。

3.1.2　实验参数与训练策略

为了进行公平比较，本文采用了与文献[15]相同的实验设置和训练策略。端到端的检测与识别模型将检测与识别作为一个整体进行训练。对不同数据集进行验证时不需要在各个数据集上进行单独训练。训练分为预训练和微调 2 个阶段。

在 SynthText 数据集上进行预训练，然后进行微调。微调训练时 mini-batch 大小设置为 8，每个 batch 中的样本按照 2:2:2:1:1 的比例从 SynthText、ICDAR2013、ICDAR2015、Total-Text 和 SCUT（scut-eng-char）[42]这 5 个数据集中随机抽取。此外，微调阶段采用了数据增强和多尺度训练策略。

为了增强数据的多样性，提升训练模型的泛化能力，同时也为了尽可能公平地与之前的算法进行对比，本文实验采用了深度学习中常用的数据预处理操作[15]。例如，数据增强过程中对输入图像在[–90°,90°]范围内随机旋转，还使用了诸如随机调整饱和度、亮度、对比度等数据增强策略。多尺度训练策略中输入图像的短边被随机调整为 5 个大小，分别是 600、800、1 000、1 200、1 400。

本文实验中将文本序列最大长度设置为 32，识别字符类别设置为 37，包括 10 个数字、26 个字母和一个标记位，识别解码器中 Transformer 设置为一层。实验由 Pytorch 深度学习框架实现，在两块 NVIDIA TITAN RTX 上进行训练，每块显存为 24 GB。预训练阶段设置默认参数优化器为 SGD，初始学习率为 0.02，权重衰减为 0.001，动量衰减为 0.9，训练迭代 30 万次，当迭代到第 10 万次和第 20 万次时学习率衰减十分之一。微调阶段，初始学习率为 0.001，共迭代 30 万次，当迭代到第 10 次和第 20 万次时学习率衰减十分之一。在没有明确说明的情况下，旋转 ICDAR2013 和 Total-Text 数据集上输入图像的短边重置为 1 000，ICDAR2015 数据集上输入图像的短边重置为 1 440。

3.1.3　评价指标

为了验证算法的有效性，本文采用的评价指标

主要包括准确率（Precision）、F 值（F-score）和召回率（Recall），计算式分别为

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \tag{17}$$

$$\text{F-score}=\frac{2\text{TP}}{2\text{TP+FP+FN}} \tag{18}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{19}$$

其中，TP 表示真样本中被正确预测的文本实例数目，FP 表示假样本中被错误预测为真样本的文本实例数目，FN 表示真样本中没有被正确预测的实例数目。在检测任务中，当文本预测区域与真实标签的区域重叠度（IoU, intersection over union）大于给定的阈值时，该文本视为被准确检测到，实验中 IoU 阈值设置为 0.5。

3.2　水平场景文本及其旋转文本的对比实验

为了验证本文算法对水平场景文本及其旋转情况下的识别性能，本节在 ICDAR2013 及其旋转数据集（Rotation_ICDAR2013 数据集）上进行了实验。

在 Rotation_ICDAR2013 数据集上进行实验，分析平衡因子 γ_v、γ_L 和 γ_F 的不同取值对算法性能的影响。不同平衡因子下获得的 F 值如表 1 所示。从表 1 中可以看出，当 γ_v、γ_L 和 γ_F 都设置为 1 时，算法性能达到最优；更高或者更低比例的平衡因子设置并没有带来明显的性能提升。这说明训练过程中 3 个模块的目标优化具有相当或相近的重要性。在之后的实验中，默认平衡因子 γ_v、γ_L 和 γ_F 都设置为 1。

表 1　　不同平衡因子下获得的 F 值

平衡因子			旋转角度	
γ_v	γ_L	γ_F	45°	60°
2	2	1	73.0%	72.3%
1	1	1	74.4%	74.2%
2	1	1	72.3%	73.3%
1	2	1	73.6%	74.5%
1	1	2	73.7%	73.9%

2 种算法在 ICDAR2013 数据集上的可视化如图 5 所示。从图 5 可以看出，相对于 CharNet 算法随着旋转角度的增加检测与识别性能逐渐下降，本文算法对水平方向的文本以及旋转 45°和 60°的文本都能进行相对准确的检测和识别，这证明了本文算法在识别水平与旋转场景文本时的优越性。

表 2 列出了本文算法与其他算法[7,13]在旋转 ICDAR2013 数据集上的端到端识别性能。从表 2 中可以看出，当水平文本旋转 45°时，本文算法端到端的识别精度在召回率、准确率以及 F 值上分别达到了 63.7%、89.4%和 74.4%；当水平文本旋转 60°时，识别精度在召回率、准确率以及 F 值上分别达到了 63.7%、88.9%和 74.2%。

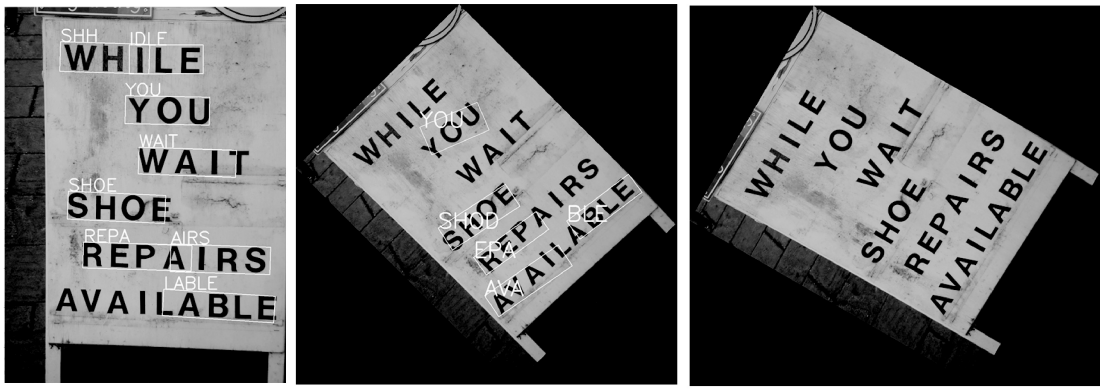
3.3　方向文本的对比实验

为了验证本文算法在方向不确定的场景文本上的性能，本节在 ICDAR2015 数据集上进行了实验，可视化结果如图 6 所示，其中，多边形为算法检测出的文本区域，多边形旁的白色字体为算法识别出的文本内容。由图 6 可知，与 CharNet 算法相比，本文算法能够更加精确地检测和识别出场景文本。本文算法与其他算法[5,7,12,25-26,30,43-48]在 ICDAR2015 数据集上的 F 值如表 3 所示。其中，G、W、S 分别表示 3 种字典约束类型，G 表示一般字典，也被称为全字典，包含的文本内容较多，约 9 万个词；S 表示强字典，包含的词最少；W 表示弱字典，词的量介于 G 和 S 之间。当使用字典约束时，只有字典范围内的文本才会考虑进行识别。本文算法在 G、W、S 这 3 种字典约束类型下的端到端的场景文本检测与识别分别达到了 73.7%、76.8%和 80.5%。与 SPTS v2 相比，在 S 约束下降低了 1.2%，但在 G 约束和 W 约束下分别提升了 3.4%和 1.2%。在所有算法中，本文算法在 G 约束的情况下 F 值最高，证明本文算法对字典约束的需求最小。

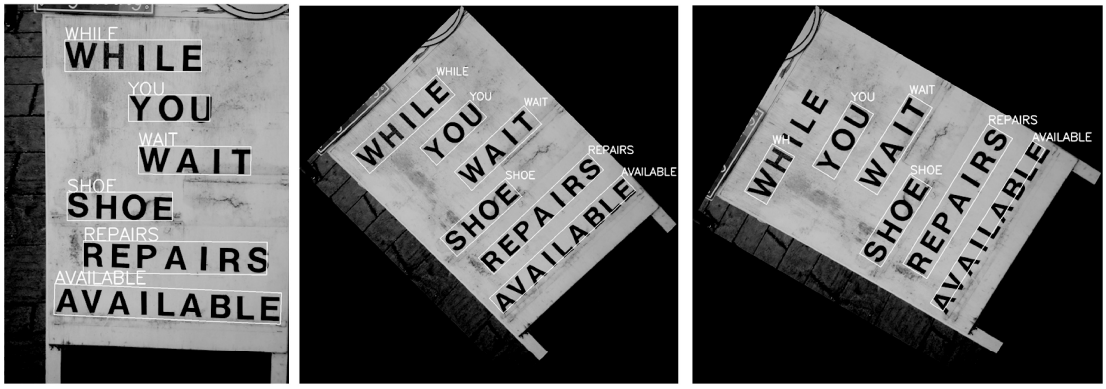
在无字典约束情况下，本文算法的 F 值达到了 69.2%，在所有算法中是最高的，这表明本文算法更加适合应用在现实世界中文本约束字典包含的文本内容较广或者无法约束字典的场景，且本文算法的时间效率最高。

3.4　任意形状文本的对比实验

在 Total-Text 数据集上进行实验的目的是验证算法在任意形状文本数据集上的性能。2 种算法在 Total-Text 数据集上的可视化结果如图 7 所示，其中，多边形为算法检测出的文本区域，多边形旁的白色字体为算法识别出的文本内容。从图 7 可以看出，对于包含了水平方向、弯曲文本的 Total-Text 数据集图像，与 CharNet 算法相比，本文算法具有更优秀的识别性能。



(a1) CharNet算法的识别效果（水平方向）　(a2) CharNet算法的识别效果（旋转45°）　(a3) CharNet算法的识别效果（旋转60°）



(b1) 本文算法的识别效果（水平方向）　(b2) 本文算法的识别效果（旋转45°）　(b3) 本文算法的识别效果（旋转60°）

图 5　2 种算法在 ICDAR2013 数据集上的可视化

表 2　　本文算法与其他算法在旋转 ICDAR2013 数据集上的端到端识别性能

算法	旋转 45°				旋转 60°			
	召回率	准确率	F 值	帧率/(frame·s^{-1})	召回率	准确率	F 值	帧率/(frame·s^{-1})
CharNet	35.5%	34.2%	33.9%	0.3	8.40%	10.30%	9.30%	0.4
Mask TextSpotter	45.8%	66.4%	54.2%	2.9	48.3%	68.2%	56.6%	2.8
本文算法	63.7%	89.4%	74.4%	9.3	63.7%	88.9%	74.2%	8.9

(a) CharNet算法的结果

(b) 本文算法的结果

图 6　2 种算法在 ICDAR2015 数据集上的可视化

表 3　　本文算法与其他算法在 ICDAR2015 数据集上的 F 值

算法	端到端的场景文本检测与识别				帧率/(frame·s^{-1})
	G 约束	W 约束	S 约束	无字典约束	
Mask TextSpotter v1	62.4%	73.0%	79.3%	—	—
CharNet R-50	62.2%	74.5%	80.2%	60.72%	0.8
TextBoxes++	51.9%	65.9%	73.3%	—	—
TextDragon	65.2%	78.3%	82.5%	—	—
Text Perceptron	65.1%	76.6%	80.5%	—	—
Boundary TextSpotter	64.1%	75.2%	79.7%	—	—
PGNet	63.5%	78.3%	83.3%	—	—
MANGO	67.3%	78.9%	81.8%	—	—
ABCNet v2	73.0%	78.5%	82.7%	—	—
TOSS	52.4%	59.6%	65.9%	—	—
SPTS	65.8%	70.2%	77.5%	—	1.5
SPTS v2	70.3%	75.6%	81.7%	—	—
本文算法	73.7%	76.8%	80.5%	69.2%	4.4

(a) CharNet算法的文本检测与识别结果

(b) 本文算法的检测与识别结果

图 7　2 种算法在 Total-Text 数据集上的可视化

为了进行定量分析，本文对具有代表性的优秀算法[5,7,12-14,26-27,29,44-46]在 Total-Text 数据集上的 F 值进行了统计，结果如表 4 所示。从表 4 可以看出，在全字典约束和无字典约束情况下，本文算法都达到了最优性能。其中，在无字典约束的情况下，本文算法的性能优于最高的 ABCNet v2，F 值达到了 70.9%，实现了新的最优性能；在全字典约束的情况下，本文算法达到了 78.1%，与 ABCNet v2 持平。

通过对 Total-Text 数据集上的实验分析可知，本文算法对任意形状的文本检测与识别也具有较

好的性能，证明了本文算法的优越性。

3.5　时间效率分析

为了衡量本文算法的时间效率并与当前算法进行公平比较，在本文硬件环境下对部分算法进行复现，并在表 2～表 4 中统计了对应的帧率。从表 2～表 4 中可以看出，本文算法具有最优的时间效率，与第二名相比，本文算法的时间效率提升了 2～3 倍。在旋转 45° 的 ICDAR2013、旋转 60°的 ICDAR2013、ICDAR2015 和 Total-Text 数据集上分别达到 9.3 frame/s、8.9 frame/s、4.4 frame/s 和 6.4 frame/s 的帧率。

3.6　消融研究

为了验证本文算法中 TAM、TVM 和 TLM 的性能，本文在 Total-Text 数据集上设置了消融研究。消融研究中的统计数据如表 5 所示。设置了 TAM+TVM 的模型与仅设置了 TVM 的模型相比，在无字典约束与全字典约束情况下的 F 值分别提升了 0.6%和 0.8%。设置了 TVM+TLM 的模型在无字典约束情况下，3 个指标均低于设置了 TAM+TVM+TLM 的模型，其中，准确率低了 1.2%；在全字典约束情况下，准确率和 F 值均低于设置了 TAM+TVM+TLM 的模型。这证明本文的 TAM 具有一定的文本感知能力，是有效的。

设置了 TVM+TLM 的模型与仅设置了 TVM 的模型相比，在无字典约束与全字典约束情况下，F 值分别提升了 3.6%和 0.9%。设置了 3 个模块的性能高于仅设置了 TAM+TVM 的性能，在全字典约束时召回率提升了 0.9%；在无字典约束下，召回率、准确率和 F 值分别提升了 3.8%、3.1%和 3.6%。这表明了 TLM 的设置提升了文本识别的精度，是有效的。

表 4　　**各算法 Total-Text 数据集上的 F 值**

算法	端到端的场景文本检测与识别		帧率/(frame·s^{-1})
	无字典约束	全字典约束	
Mask TextSpotter v1	52.9%	71.8%	—
FOTS	32.2%	—	—
CharNet H-88	66.6%	—	0.5
TextDragon	48.8%	74.8%	—
Mask TextSpotter v2	65.3%	77.4%	3.1
Unconstrained	67.8%	—	—
ABCNet	64.2%	75.7%	—
Boundary TextSpotter	65.0%	76.1%	—
PGNet	63.1%	—	—
ABCNet v2	70.4%	78.1%	3.5
TOSS	65.1%	74.8%	—
本文算法	70.9%	78.1%	6.4

表 5　　**消融研究中的统计数据**

模块	无字典约束			全字典约束		
	召回率	准确率	F 值	召回率	准确率	F 值
TVM	61.9%	79.6%	66.7%	68.7%	87.4%	76.9%
TAM+TVM	58.3%	79.4%	67.3%	69.4%	88.4%	77.7%
TVM+TLM	61.9%	81.3%	70.3%	70.3%	87.3%	77.8%
TAM+TVM+TLM	62.1%	82.5%	70.9%	70.3%	87.7%	78.1%

4 结束语

本文提出了一种基于 Transformer 并行解码的场景文本检测与识别模型，可对场景图像中任意形状的文本进行端到端的检测与识别。场景文本通常存在形状复杂、视觉特征质量差、文本识别效率低的问题，该模型通过设计的文本感知模块能够感知文本形态，更加鲁棒性地提取文本视觉特征；基于 Transformer 的视觉模块和语言模块的识别分支，通过并行解码的方式提升模型的识别效率。此外，融合门的设计使模型在识别过程中综合考虑了文本的视觉特征和语义信息，提升了任意形状场景文本的识别精度。实验结果表明，本文提出的模型在端到端的场景文本检测与识别任务中具有更强的实用性，不仅能够端到端地检测与识别任意形状的场景文本，还具有一定的时间效率。未来的工作中，将研究如何基于文本字符之间的视觉关系、语义关系以及几何关系构建关系图，设计具有图推理能力的模型，进一步提升端到端的场景文本检测与识别性能。

参考文献：

[1] LONG S B, HE X, YAO C. Scene text detection and recognition: the deep learning era[J]. International Journal of Computer Vision, 2021, 129(1): 161-184.

[2] 陈卓, 王国胤, 刘群. 结合多粒度特征融合的自然场景文本检测方法[J]. 计算机科学, 2021, 48(12): 243-248.
CHEN Z, WANG G Y, LIU Q. Natural scene text detection algorithm combining multi-granularity feature fusion[J]. Computer Science, 2021, 48(12): 243-248.

[3] 邵海琳, 季怡, 刘纯平, 等. 基于增强特征金字塔网络的场景文本检测算法[J]. 计算机科学, 2022, 49(2): 248-255.
SHAO H L, JI Y, LIU C P, et al. Scene text detection algorithm based on enhanced feature pyramid network[J]. Computer Science, 2022, 49(2): 248-255.

[4] 丁明宇, 牛玉磊, 卢志武, 等. 基于深度学习的图片中商品参数识别方法[J]. 软件学报, 2018, 29(4): 1039-1048.
DING M Y, NIU Y L, LU Z W, et al. Deep learning for parameter recognition in commodity images[J]. Journal of Software, 2018, 29(4): 1039-1048.

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

[6] LYU P Y, LIAO M H, YAO C, et al. Mask TextSpotter: an end-to-end trainable neural network for spotting text with arbitrary shapes[C]// European Conference on Computer Vision. Berlin: Springer, 2018: 71-88.

[7] XING L J, TIAN Z, HUANG W L, et al. Convolutional character networks[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2020: 9125-9135.

[8] LI H, WANG P, SHEN C H, et al. Show, attend and read: a simple and strong baseline for irregular text recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019: 8610-8617.

[9] YU D L, LI X, ZHANG C Q, et al. Towards accurate scene text recognition with semantic reasoning networks[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 12110-12119.

[10] YUE X Y, KUANG Z H, LIN C H, et al. RobustScanner: dynamically enhancing positional clues for robust text recognition[C]//European Conference on Computer Vision. Berlin: Springer, 2020: 135-151.

[11] FANG S C, XIE H T, WANG Y X, et al. Read like humans: autonomous, bidirectional and iterative language modeling for scene text recognition[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 7094-7103.

[12] FENG W, HE W H, YIN F, et al. TextDragon: an end-to-end framework for arbitrary shaped text spotting[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2020: 9075-9084.

[13] 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.

[14] LIU Y L, CHEN H, SHEN C H, et al. ABCNet: real-time scene text spotting with adaptive bezier-curve network[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 9806-9815.

[15] LIAO M H, PANG G, HUANG J, et al. Mask TextSpotter v3: segmentation proposal network for robust scene text spotting[C]//European Conference on Computer Vision. Berlin: Springer, 2020: 706-722.

[16] 王建新, 王子亚, 田萱. 基于深度学习的自然场景文本检测与识别综述[J]. 软件学报, 2020, 31(5): 1465-1496.
WANG J X, WANG Z Y, TIAN X. Review of natural scene text detection and recognition based on deep learning[J]. Journal of Software, 2020, 31(5): 1465-1496.

[17] BAEK Y, LEE B, HAN D, et al. Character region awareness for text detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 9357-9366.

[18] ZHANG S X, ZHU X B, HOU J B, et al. Deep relational reasoning graph network for arbitrary shape text detection[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 9696-9705.

[19] TIAN Z T, SHU M, LYU P Y, et al. Learning shape-aware embedding for scene text detection[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 4229-4238.

[20] 李煌, 王晓莉, 项欣光. 基于文本三区域分割的场景文本检测方法[J]. 计算机科学, 2020, 47(11): 142-147.
LI H, WANG X L, XIANG X G. Scene text detection based on triple segmentation[J]. Computer Science, 2020, 47(11): 142-147.

[21] LI J C, LIN Y, LIU R R, et al. RSCA: real-time segmentation-based context-aware scene text detection[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2021: 2349-2358.

[22] LIAO M H, ZOU Z S, WAN Z Y, et al. Real-time scene text detection with differentiable binarization and adaptive scale fusion[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 919-931.

[23] SHENG F F, CHEN Z N, XU B. NRTR: a no-recurrence sequence-to-sequence model for scene text recognition[C]//Proceedings of International Conference on Document Analysis and Recognition (ICDAR). Piscataway: IEEE Press, 2020: 781-786.

[24] YANG L, DANG F, WANG P, et al. A holistic representation guided attention network for scene text recognition[J]. arXiv Preprint, arXiv: 1904.01375v3, 2019.

[25] QIAO L, TANG S L, CHENG Z Z, et al. Text perceptron: towards end-to-end arbitrary-shaped text spotting[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 11899-11907.

[26] WANG P F, ZHANG C Q, QI F, et al. PGNet: real-time arbitrarily-shaped text spotting with point gathering network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 2782-2790.

[27] LIU X B, LIANG D, YAN S, et al. FOTS: fast oriented text spotting with a unified network[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 5676-5685.

[28] HE T, TIAN Z, HUANG W L, et al. An end-to-end TextSpotter with explicit alignment and attention[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 5020-5029.

[29] QIN S Y, BISSACO A, RAPTIS M, et al. Towards unconstrained end-to-end text spotting[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2020: 4703-4713.

[30] QIAO L, CHEN Y, CHENG Z Z, et al. MANGO: a mask attention guided one-stage scene text spotter[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021, 35(3): 2467-2476.

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

[32] ZHOU X Y, YAO C, WEN H, et al. EAST: an efficient and accurate scene text detector[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 2642-2651.

[33] LIAO M H, WAN Z Y, YAO C, et al. Real-time scene text detection with differentiable binarization[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 11474-11481.

[34] WANG W H, XIE E Z, LI X, et al. Shape robust text detection with progressive scale expansion network[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 9328-9337.

[35] VATTI B R. A generic solution to polygon clipping[J]. Communications of the ACM, 1992, 35(7): 56-63.

[36] GIRSHICK R. Fast R-CNN[C]//Proceedings of IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2016: 1440-1448.

[37] MILLETARI F, NAVAB N, AHMADI S A. V-net: fully convolutional neural networks for volumetric medical image segmentation[C]//Proceedings of 2016 Fourth International Conference on 3D Vision (3DV). Piscataway: IEEE Press, 2016: 565-571.

[38] GUPTA A, VEDALDI A, ZISSERMAN A. Synthetic data for text localisation in natural images[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 2315-2324.

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

[40] CH'NG C K, CHAN C S. Total-text: a comprehensive dataset for scene text detection and recognition[C]//Proceedings of 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Piscataway: IEEE Press, 2018: 935-942.

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

[42] ZHONG Z, JIN L, ZHANG S, et al. DeepText: a unified framework for text proposal generation and text detection in natural images[J]. arXiv Preprint, arXiv: 1605.07314v1, 2016.

[43] LIAO M H, SHI B G, BAI X. TextBoxes++: a single-shot oriented scene text detector[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2018, 27(8): 3676-3690.

[44] WANG H, LU P, ZHANG H, et al. All You need is boundary: toward arbitrary-shaped text spotting[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020:

12160-12167.

[45] 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.

[46] TANG J Q, QIAO S, CUI B L, et al. You can even annotate text with voice: transcription-only-supervised text spotting[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM Press, 2022: 4154-4163.

[47] PENG D, WANG X, LIU Y, et al. SPTS: single-point text spotting[J]. arXiv Preprint, arXiv: 2112.07917, 2021.

[48] LIU Y, ZHANG J, PENG D, et al. SPTS v2: single-point scene text spotting[J]. arXiv Preprint, arXiv: 2301.01635v1, 2023,

[作者简介]

郑金志（1989– ），男，河南周口人，中国科学院大学博士生，主要研究方向为机器视觉、自然语言处理等。

汲如意（1988– ），男，山东日照人，博士，中国科学院软件研究所助理研究员，主要研究方向为机器学习、计算机视觉、图像处理、模式识别等。

张立波（1989– ），男，安徽阜阳人，博士，中国科学院软件研究所副研究员、硕士生导师，主要研究方向为图像处理、模式识别等。

赵琛（1967– ），男，云南普洱人，博士，中国科学院软件研究所研究员、博士生导师，主要研究方向为编译技术、操作系统、网络软件等。