Hans Journal of Data Mining 数据挖掘, 2023, 13(2), 154-164
Published Online April 2023 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2023.132015

Hans汉斯

基于端到端的复杂场景中文文字识别方法研究

帅梓涵，胡金蓉，郎子鑫，罗月梅，李桂钢

成都信息工程大学计算机学院，四川 成都

收稿日期：2023年3月13日；录用日期：2023年4月13日；发布日期：2023年4月21日

摘　要

近年来，由于成功挖掘了场景文本检测和识别的内在协同作用，端到端场景文本识别引起了人们的极大关注。然而，最近最先进的方法通常仅通过共享主干来结合检测和识别，这些方法由于其尺度和纵横比的极端变化不能很好地处理场景文本。在本文中，我们提出了一种新的端到端场景文本识别框架，称为ES-Transformer。与以往以整体方式学习场景文本的方法不同，我们的方法基于几个代表性特征来执行场景文本识别，这避免了背景干扰并降低了计算成本。具体来说，使用基本特征金字塔网络进行特征提取，然后，我们采用Swin-Transformer来建模采样特征之间的关系，从而有效地将它们划分为合理的组。在提升识别精度的同时降低了计算复杂度，不再依赖于繁杂的后处理模块。对中文数据集的定性和定量实验表明，ES-Transformer优于现有方法。

关键词

端到端，文字识别，Transformer，深度学习

Research on End-to-End Chinese Text Recognition Method in Complex Scenes

Zihan Shuai, Jinrong Hu, Zixin Lang, Yuemei Luo, Guigang Li

School of Computer Science, Chengdu University of Information Technology, Chengdu Sichuan

Received: Mar. 13th, 2023; accepted: Apr. 13th, 2023; published: Apr. 21st, 2023

Abstract

In recent years, due to the successful exploration of the inherent synergistic effect of scene text detection and recognition, end-to-end scene text recognition has attracted great attention. However, the most recent state-of-the-art methods usually only combine detection and recognition by sharing backbones, and these methods cannot handle scene text well due to extreme variations in scale and aspect ratio. In this paper, we propose a new end-to-end scene text recognition frame-

文章引用: 帅梓涵, 胡金蓉, 郎子鑫, 罗月梅, 李桂钢. 基于端到端的复杂场景中文文字识别方法研究[J]. 数据挖掘, 2023, 13(2): 154-164. DOI: 10.12677/hjdm.2023.132015

work called ES-Transformer. Unlike previous methods that learn scene text in a holistic way, our approach performs scene text recognition based on several representative features, which avoids background interference and reduces computational cost. Specifically, we use a basic feature pyramid network for feature extraction, and then we employ Swin-Transformer to model the relationships between the sampled features, effectively partitioning them into reasonable groups. By improving recognition accuracy and reducing computational complexity, ES-Transformer no longer relies on complex post-processing modules. Qualitative and quantitative experiments on Chinese datasets show that ES-Transformer outperforms existing methods.

Keywords

End-to-End, Text Recognition, Transformer, Deep Learning

1. 引言

文字作为一种传播与联系的文明产物，遍布在人们的生活和工作中，尤其是互联网时代开启后，人们通过文字在网络上高效且多样地获取各种信息。随着物联网时代的到来，人工智能领域不断发展，其中计算机视觉研究领域中的文字检测与识别也受到广泛关注。长期以来，场景文本识别一直是一个活跃的研究领域，由于其在自动驾驶、智能导航和关键实体识别等领域的广泛应用，引起了大量关注。场景文本检测作为场景文本读取的关键先验组件，旨在精确定位场景图像中的文本。尽管现有方法取得了显著的改进[1] [2]，但由于场景文本的多样性，例如不同比例、复杂照明、透视失真、多方位和复杂形状，这仍然是一项具有挑战性的任务。此外，传统的场景文本定位方法将检测和识别视为两个独立的任务，首先定位和裁剪输入图像上的文本区域，然后通过将裁剪区域输入文本识别器来预测文本序列。这种流水线可能具有一些限制，例如 1) 这两个任务之间的错误累积，例如，不精确的检测结果可能严重阻碍文本识别的性能；2) 内存消耗大，推理效率低。因此对复杂场景下的精确且运行速度快的文字识别有了迫切需要。

近年来随着人工智能的发展，受 Transformer [3]在自然语言处理(NLP)中的优势的启发，许多工作[4] [5] [6]将其引入视觉任务中，以提取全局范围特征并对图像中的长距离依赖性建模，同时显示出良好的性能。特别是在对象检测中，基于 DETR 的方法[6] [7]成功地使用 Transformer 从以前的对象检测框架中移除了复杂的手工设计过程[8]。

尽管 Transformer 在基于 DETR 的框架中为全局范围特征建模带来了优势[4]，但它们可能会遇到处理小对象和高计算复杂性的问题。例如，最近基于 DETR 的场景文本检测器[9]无法在 ICDAR2015 数据集[10]和 ICDAR2017-MLT 数据集[11]上实现令人满意的检测精度，因为这两个数据集中的文本实例具有更大的尺度和纵横比方差。Transformer 在小尺度下捕获特征图上的小文本通常是不够的，而使用多尺度特征图的基于 DETR 的方法的时间成本是不可预测的。并且，如果检测仅基于输入特征中的视觉信息，则检测器容易被背景噪声分散注意力，并提出不一致的检测。同一图像中文本之间的交互是消除背景噪声影响的关键因素，因为同一单词的不同字符可能包含很强的相似性，例如背景和文本样式。使用 Transformer 可以学习文本实例之间的丰富交互。例如，Yu 等人[12]使用 Transformer 使文本在语义层面上相互交互。Fang 等人[13]和 Wang 等人[14]进一步采用 Transformer 来模拟文本之间的视觉关系。尽管

最近的一些工作[6]通过优化注意操作提高了基于 Transformer 的对象检测器的效率，但它们在场景文本检测中未能获得更好结果。

在本文中，我们提出了 ES-Transformer，这是一个端到端可训练的基于 Transformer 的框架，旨在实现文本检测和识别之间更好的协同。我们认为，由于前景文本实例仅占据场景图像中的几个小而窄的区域，因此不需要使用所有像素的关系进行特征学习。直观地，我们首先采样并收集与场景文本高度相关的特征。然后，我们采用 Transformer 对采样特征之间的关系进行建模，以便对它们进行适当分组。受益于 Transformer 的强大注意力机制，每个特征组将对应于文本实例，这对于预测其边界框非常方便。

因此，本文所使用的 ES-Transformer 与以往的场景文本检测方法[15] [16] [17]不同，我们的检测方法使用 Efficient-Net 以整体的方式学习场景文本图像的深度表示，我们仅基于几个代表性特征的检测方法具有三个突出的优点：

1) 它可以显著消除冗余背景信息，这有利于提高检测过程的有效性和效率；

2) 使用 Transformer 对采样特征进行分组，我们可以获得更精确的分组结果和边界框，而无需任何后处理操作；

3) 由于特征采样和分组以端到端的方式实现，这两个阶段可以共同提高最终检测性能。

此外，与最近的模型[18] [19] [20]的比较也证明了我们方法的有效性。

2. 数据

ICDAR’19-ReCTs 数据集[21]包含 25 k 个带注释的招牌图像，其中 20 k 个图像用于训练集，其余用于测试集。所有图片均来自美团点评集团，由美团商家在不受控制的条件下使用手机摄像头收集。与其他数据集不同，该数据集主要关注招牌上的中文文本阅读。为了美观或突出某些元素，招牌中汉字的布局和排列要复杂得多。与英文文本相比，中文文本的类数通常要多得多，常用字符 6 k 多，排版复杂，字体繁多。该数据集主要包含商店标志的文本，并为每个字符提供注释。图 1 显示了一些示例图像。

Figure 1. The main recognition objects of this paper (25,000 signboard images)
图 1. 本文的主要识别对象(25,000 张招牌图像)

3. 本文方法

3.1. 模型方案

在本节中，我们首先介绍所提出的场景文本识别方法的总体架构。包括四个组件：1) 基于 Efficient-Net 的特征提取模块 2) 基于查询的文本检测器；3) 识别转换模块，用于桥接文本检测器和识别器；和 4) 基于注意力的识别器。

如图所示，我们提出的基于 Transformer 的架构由主干网络、特征采样网络和特征分组网络组成。主干是基本特征金字塔网络(FPN)，配备 EfficientNet-B3 [22]。我们使用 EfficientNet-B3 作为特征提取模块，该模块由七个移动反向瓶颈(MBConv 块)组成，如图 2 中的“Conv 块#1”至“Conv 块#7”所示。为了利用多尺度特征，将特征映射 F3、F5 和 F7 (输入图像尺度为 1\8、1\16 和 1\32)。在我们的特征采样网络中，首先通过 Coord 卷积层[23]和受约束的可变形池层将三个特征映射下采样到更小的尺度(即 1\8、1\16、1\32)。然后，采用多个卷积层来生成置信评分图，以区分代表性文本区域。之后，我们只选择每个标度层 k 中具有最高分数的特征，并将它们聚集成具有形状 $\left(\sum_k N_K, C\right)$ 的序列形式，其中 C 是信道号。在我们的特征分组网络中，采样特征首先与位置嵌入连接。然后，我们采用 Transformer-encoder 层来建模它们之间的关系，并隐式聚合来自同一文本实例的特征。最后，通过识别器对识别的结果进行输出。

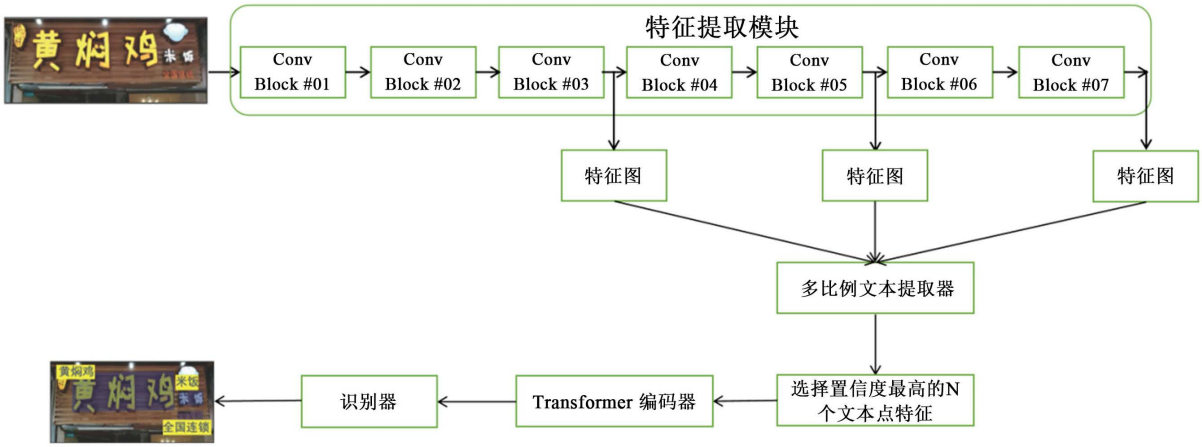

Figure 2. Framework of character recognition model based on ES-Transformer
图 2. 基于 ES-Transformer 的文字识别模型框架

3.1.1. 特征提取模块

尽管在目标检测中具有新颖的结构和良好的性能，但基于 Transformer 的方法由于尺度和纵横比的极端变化，不能很好地进行场景文本检测。继之前的文本检测器之后，我们使用 FPN 的多尺度特征来提高检测性能。然而，这样的方案产生了难以承受的计算成本和 Transformer 更长的收敛时间。我们观察到前景文本实例只占据小而窄的区域，用于定位文本的有用信息相对稀疏。因此，我们提出了一种特征采样网络，以减少多尺度特征所涉及的冗余背景噪声，降低计算复杂度并促进 Transformer 的特征学习。

多尺度文本提取器：为了对前景文本中的代表性特征进行采样，我们应用了一个简单的多尺度文本提取器来预测像素级文本区域的置信度。按照 CoordConv [23]，我们首先将每个特征图与两个额外的归一化坐标通道连接起来，以引入位置信息。设 F 表示不同比例(即 1\4、1\8、1\16)的 FPN 特征图，

$$F = \left\{ f_k \in R^{H_k \times W_k \times C} \middle| k = 0,1,2 \right\} \tag{1}$$

然后，通过以下方式注入位置信息：

$$\hat{f}_k = Conv\left(f_k \oplus C_k \right) \tag{2}$$

$\oplus$ 表示串联操作，$C_k \in R^{H_k \times W_K \times 2}$ 表示归一化坐标。

最后，我们构造了一个由卷积层和 Sigmoid 函数组成的简单评分网络，以生成所有尺度上代表性文本区域的置信评分图。为了更好地区分每个文本实例中不同位置处像素的重要性，使用位置上的不同分

数进行监督。为了生成分数图，我们调整了一般对象检测中的高斯热图生成，用于单词级别的文本实例。具体而言，实现二维高斯分布以生成 S 的基本真值 $S^t=\left\{S_k^t \middle| k=0,1,2\right\}$，确保每个文本实例的中心部分具有最高的重要性分数，并且分数从中心到轮廓逐渐减少。

3.1.2. 特征采样模块

为了减少冗余背景噪声，我们设计了一种选择与前景文本高度相关的代表性特征的策略。这些特征包含前景文本的丰富几何和上下文信息，足以进行文本定位。让我们表示预测得分图，并且

$$S=\left\{S_k \in R^{H_k' \times W_k'} \middle| S_k=S\left(\hat{f}_k\right), k=0,1,2\right\} \tag{3}$$

为了减少冗余背景噪声，我们设计了一种选择与前景文本高度相关的代表性特征的策略。这些特征包含前景文本的丰富几何和上下文信息，足以进行文本定位。让我们表示预测得分图，并且

$$\bar{F}=\left[\hat{f}_n \in R^C \middle| n=0,1,\cdots,N\right] \tag{4}$$

其中 $N=\sum_{k=0}^{2} N_k$，且 N_k 是不同尺度下的选定特征的数量。

因此，可以显著减少所有尺度下的巨大特征的数量。主要选择的特征可能来自前景文本区域，该区域将包含足够的几何和上下文信息，用于文本检测。

3.1.3. 特征分组模块

通过特征选择，只有少数与前景文本高度相关的代表性特征被连接用于输入转换器建模。为了保留采样特征的位置信息，我们将位置嵌入添加到 F 中。然后，我们采用 Transformer 结构，通过注意力机制聚合来自同一文本实例的特征。基本形式是一个有四个 Transformer 编码器层的堆叠网络，它们由自我注意模块、前馈层和层规范化组成。我们构建了我们的自我注意模块，即

$$Attn\left(\hat{F}\right)=\text{softmax}\left(\frac{Q\left(\hat{F}\right)K\left(\hat{F}\right)^{\mathrm{T}}}{\sqrt{C'}}\right)V\left(\hat{F}\right) \tag{5}$$

其中 $\hat{F} \in R^{N \times C'}$ 表示带位置嵌入的采样特征，C' 是通道数。Q、K 和 V 表示不同的线性层。对于以前的方法，在特征地图上应用注意操作的核心问题 $x \in R^{H \times W \times C'}$ 是所有空间位置的计算复杂性。在原 DETR [5]编码器中，注意操作的复杂度为 $\mathrm{O}\left((\mathrm{HW})^2 C'\right)$，与空间大小呈二次方关系。然而，在我们的方法中，它只与所选特征的数目 N 有关，复杂度变为 $O\left(N^2 C'\right)$。在我们的实现中，选定的数字 $N^2 \ll (HW)^2$，因此可以显著降低 Transformer 的复杂性。最后，输出的文本特征被送入两个预测头进行分类和文本检测。文本检测头由完全连接的层和 Sigmoid 函数组成(图 3)。

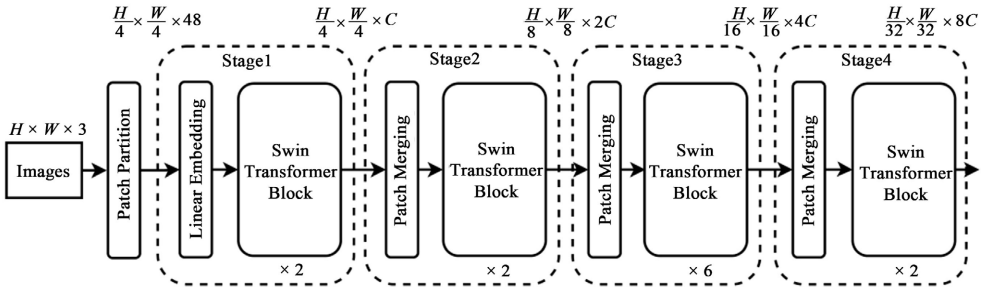

Figure 3. Model framework based on Swin Transformer
图 3. 基于 Swin Transformer 的模型框架

3.1.4. 注意力机制

在 Transformer 的基于窗口计算自注意力的方式虽然很好的解决了内存和计算量的问题，但是窗口与窗口之间没有了通信，没能达到全局建模的效果，这就限制了模型的能力。移动窗口[24]被提出后，先进性一次窗口的自注意力计算，再进行一次移动窗口后的自注意力计算，这样就实现了窗口与窗口之间的通信，从而达到了全局建模的效果。

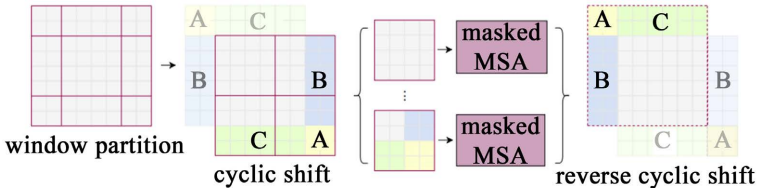

Figure 4. A framework for calculating self-attention based on moving windows
图 4. 基于移动窗口计算自注意力的框架

通过循环移位，由图 4 中的 window partition 图变成图 4 的 cyslic shift 图，经过循环移位后，图片重新分成 4 个窗口，移动窗口前是 4 个窗口，经过循环移位后仍是 4 个窗口，这就使得窗口数量还是 4 个，窗口的数量就固定了，这就使得计算难度降低了。

循环移位后仍然有一些问题存在，A，B，C 所在的三个窗口中不仅包含了原本就在这个地方的元素，同时也包含了 A，B，C 三个从很远的地方移位过来的元素，每个窗口中原本存在的元素和移位过来的元素之间是没有什么较大的关系的，因为二者离的比较远，所以我们不需要对二者进行注意力的计算。针对这个问题 Swin-Transformer 原作者团队提出了掩码操作，从而能够让一个窗口中不同的区域之间能用一次前向过程就能把自注意力就散出来，而相互之间都不干扰。

3.2. 试验参数设置

对于数据集，我们采用 LSVT [25]、ArT [26]、ReCTS [21]和合成预处理数据来训练模型。

我们提取了输入图像分辨率为 1/8、1/16、1/32 的 4 个特征地图，用于文本检测和识别。

我们使用图像批量大小为 8 的图像对模型进行训练。使用以下数据增强策略：1) 随机缩放；2) 随机旋转；3) 随机裁剪。其他策略，如随机亮度、对比度和饱和度也会在训练期间应用。

此外，我们还采用了数据增强策略，例如随机尺度训练，短尺寸从 640 到 896 (间隔为 32)唯一选择，长尺寸小于 1600；和随机裁剪，我们确保裁剪图像不会剪切文本实例(对于一些难以满足条件的特殊情况，我们不应用随机裁剪)。我们使用 4 个 2080 ti 训练我们的模型，图像批量大小为 8。最大迭代次数为 20 epoch；初始化学习率为 0.01，在第 10 次迭代时降至 0.001。

3.3. 损失函数

识别损失

在特征地图上应用识别转换后，背景噪声得到有效抑制，从而可以更精确地界定文本区域。这使我们能够仅使用顺序识别网络来获得有希望的识别结果，而无需校正模块为了增强细粒度特征提取和序列建模，我们采用了受[24]启发的基于移动窗口计算自注意机制作为识别编码器。因此，它可以有效地提取细粒度特征，同时保持全局建模能力。至于解码器，我们只需使用空间注意模块来跟踪。确认损失如下：

$$L = -\frac{1}{T}\sum_{k=1}^{T}\log p\left(y_i\right) \tag{6}$$

其中 T 是序列的最大长度， $p(y_i)$ 是序列的概率。

3.4. 试验评估

文字识别数据集使用字级精度、召回率和 f1 分数进行评估。如果重复的单词出现在基本事实中，那么它们也应该出现在预测中。精确率表示返回的图片中正确的有多少。召回率代表有多少张应该返回的图片没有找到。F1 值代表 Precision 与 Recall 的调和平均。精确性、召回率和 f1 分数描述如下：

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

其中，*TP* (真阳性)：被模型预测为正的正样本；将正类预测为正类；

FP (假阳性)：被模型预测为正的负样本；将负类预测为正类；

FN (假阴性)：被模型预测为负的正样本；将正类预测为负类；

TN (真阴性)：被模型预测为负的负样本；将负类预测为负类。

4. 结果

4.1. 与其它注意力机制的比较

为了模型的合理性，本研究分别使用精度、召回率、F1 三个指标作为验证模型优劣的指标来进行比较。为了进行公平的比较，在训练过程中模型的其它设置不变，仅对模型进行修改。

Table 1. The evaluation results of different attention mechanisms (precision, recall rate, F1) on the ICDAR'19-ReCTs data set. The bold value in each column represents the best value (the same below).
表 1. 在 ICDAR'19-ReCTs 数据集上不同注意力机制的评估结果(精度、召回率、F1)其中每一列中加粗的值代表最佳值(下同)

Attention mechanism	precision	Recall	F1
Attention	73.7	74.3	74.0
基于窗口计算自注意力	83.9	87.3	85.6
基于移动窗口计算自注意力	88.3	90.1	89.7

表 1 结果表明，当为常规注意力机制时，训练的模型的识别最差，其精度和召回率最低，F1 值最低，识别效果最差。当使用基于窗口计算自注意力作为注意力机制时，训练的文字识别模型表现中等，优于常规注意力机制作为注意力机制时的结果。而当基于移动窗口计算自注意力作为注意力机制时，模型在测试集上的结果(PRESION = 88.3, RECALL = 90.1, F1 = 89.7)均优于基于窗口计算自注意力作为注意力机制时的识别效果。具体而言，基于移动窗口计算自注意力作为注意力机制时，识别的精度，召回率，F1 评分效果均高于常规注意力机制和基于窗口计算自注意力。所以使用基于移动窗口计算自注意力时，模型具有最佳的文字识别效果。

4.2. 与其它模型比较

Table 2. Evaluation results of different models (precision, recall rate, F1) on the ICDAR'19-ReCTs data set
表 2. 在 ICDAR'19-ReCTs 数据集上不同模型的评估结果(精度、召回率、F1)

模型	precision	Recall	F1
EAST	73.7	74.3	74.0
PSENet-1s	83.9	87.3	85.6
Mask TextSpotter	88.9	89.3	89.0
ES-Transformer	88.3	90.1	89.7

本节将之前取得较好结果的模型 EAST [18]，PSENET-1S [19]，MASK TEXTPOTTER [20]模型与本文所提出的 ES-TRANSFORMER 模型作对比，以此验证和评估本文提出的复杂场景下的文字识别模型的有效性。其中，三个模型的试验和参数设置与本文模型一致。从不同杂场景下的文字识别模型在中文测试集上的评估价格(表 2)可得，本文提出的中午文字识别模型对文字有着最佳效果，具有最高的 RECALL 和 F1。MASK TEXTPOTTER 预测效果仅次于本文的模型。而 EAST 的识别最差，其 PRECISION/RECALL 仅为 73.7/74.3，在文字识别过程中存在较大误差。

4.3. 消融实验

Table 3. Evaluation results of ablation experiment on ICDAR'19-ReCTs data set (precision, recall rate, F1)
表 3. 在 ICDAR'19-ReCTs 数据集上消融实验的评估结果(精度、召回率、F1)

Model	Resnet 50	Efficient-Net	CoordConv	Top N Furture	precision	Recall	F1
Model 1	√				81.4	82.1	81.7
Model 2		√			83.2	84.1	83.6
Model 3		√	√		83.6	84.3	84.0
Ours		√	√	√	88.3	90.1	89.7

为了评估所提出的组件的有效性，我们对 ES-TRANSFORMER 进行了消融研究。首先对 resnet50 与 Efficient-net 的主干网络效果进行对比，通过使用 Efficient-net 使得 PRECISON 和 RECALL 都得到了提高，

如表 3 所示，使用 CoordConv 后，场景文本识别的结果均有了一定量的提升。CoordConv 极大地提高了检测器和识别器的性能。这主要是因为 CoordConv 给卷积加上坐标，从而使其具备了空间感知能力。

我们还比较 top-nfurture 的效果，使用了 top-nfurture，在 PRECIOSON 与 RECALL 上提升了 4.7 和 5.8，F1 提高了 5.1，有了更好的识别能力，效果上有了较大提升。

4.4. 个例分析

本文分别给出一组个例来对模型的识别性能进行具体的比较和说明，在图 5 中，四张图片分别代表四种不同的错误。在第一张图片中，类似于印章的“炜”由于字的少见与不同于常规字体的繁体，使得模型将这个字错误的识别成“津”。在第二张图片中，在红色背景的“牛麒麟”中，由于距离的远和字体的小，只识别其中的“牛”字。在第三张图片中，“超市出入口”被“出入平安”所遮挡，使得模型只识别出来其中的“超”和“口”两个字，其他的都没有识别出来。在四张图片中，由于广告的年久失

Figure 5. Text recognition results on ICDAR'19-ReCTs data set
图 5. 在 ICDAR'19-ReCTs 数据集上文字识别结果

修，其中的一排字体很多都是缺失一半或者全部缺失。所以在识别时候给模型造成困难，检测和模型效果不佳。这些可能是因为多种原因导致的：1) 复杂场景下的中文文字识别一直以来都是一个难题，文字信息很容易被周围环境因素遮挡，所以考虑图片的全局信息将给文字识别带来很大助力。2) 本文所使用的训练数据集有限，而虽然基本包含中文常用文字，但是对于一些不常用的文字和繁体或者不同书法类型字体缺乏训练，导致对部分文字识别能力有限。因此，在未来的工作将一步考虑以上问题，改进算法并考虑更加注重图片全局信息。

5. 结论

本文基于卷积神经网络网络框架建立了一个复杂场景下的中文文字识别模型，得到以下主要结论：

1) 为了对前景文本中的代表性特征进行采样，我们应用了一个简单的多尺度文本提取器来预测像素级文本区域的置信度。可以很好降低模型在多尺度特征上的计算复杂度。

2) 使用卷积网络，可以很好消除背景的冗余信息，利于提高 Transformer 的效率。

3) 在特征选择的时候，只有少数与前景文本高度相关的代表性特征被连接用于输入转换器建模。为了保留采样特征的位置信息，我们将位置嵌入添加到模型中。解决文本具有更极端的尺度和长宽比导致的 Transformer 通常无法在小尺度上充分地获取小文本特征问题。

尽管本文提出的复杂场景下文字识别模型具有较好的识别能力，但也存在一些可以改进的空间：因为中文文字的含义是连贯的，如果可以结合语义信息进入模型，将具有更好的识别性能。

基金项目

四川省科技计划项目(编号：2023YFQ0072，22QYCX0082)。

参考文献

[1] He, M.H., *et al*. (2021) MOST: A Multi-Oriented Scene Text Detector with Localization Refinement. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 8809-8818. https://doi.org/10.1109/CVPR46437.2021.00870

[2] Li, X., *et al*. (2018) Shape Robust Text Detection with Progressive Scale Expansion Network. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 16-17 June 2019, 9328-9337.

[3] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2007) Attention Is All You Need. *Proceedings NIPS*, Vancouver, 3-6 December 2007, 5998-6008.

[4] Carion, N., Massa, F., Synnaeve, G., *et al*. (2020) End-to-End Object Detection with Transformers. *Computer Vi-*

sion-ECCV 2020: 16*th European Conference*, Glasgow, 23-28 August 2020, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

[5] Dai, X., Chen, Y., Yang, J., *et al*. (2021) Dynamic detr: End-to-End Object Detection with Dynamic Attention. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 11-17 October 2021, 2988-2997. https://doi.org/10.1109/ICCV48922.2021.00298

[6] Meng, D., Chen, X., Fan, Z., Zeng, G., *et al*. (2021) Conditional detr for Fast Training Convergence. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 11-17 October 2021, 3651-3660. https://doi.org/10.1109/ICCV48922.2021.00363

[7] Zhu, X., Su, W., Lu, L., *et al*. (2020) Deformable DETR: Deformable Transformers for End-to-End Object Detection. *ICLR* 2021, 3-7 May 2021, 1-16.

[8] Liu, W., Anguelov, D., Erhan, D., *et al*. (2016) Ssd: Single Shot Multibox Detector. *Computer Vision-ECCV* 2016: 14*th European Conference*, Amsterdam, 11-14 October 2016, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[9] Raisi, Z., Naiel, M.A., Younes, G., *et al*. (2021) Transformer-Based Text Detection in the Wild. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 3162-3171. https://doi.org/10.1109/CVPRW53098.2021.00353

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

[11] Nayef, N., Yin, F., Bizid, I., *et al*. (2017) Icdar2017 Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification-rrc-mlt. 2017 14*th IAPR International Conference on Document Analysis and Recognition* (*ICDAR*) *IEEE*, Vol. 1, 1454-1459. https://doi.org/10.1109/ICDAR.2017.237

[12] Yu, D., Li, X., Zhang, C., *et al*. (2020) Towards Accurate Scene Text Recognition with Semantic Reasoning Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 14-19 June 2020, 12113-12122. https://doi.org/10.1109/CVPR42600.2020.01213

[13] Fang, S., Xie, H., Wang, Y., *et al*. (2021) Read like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 7098-7107. https://doi.org/10.1109/CVPR46437.2021.00702

[14] Wang, Y., Xie, H., Fang, S., *et al*. (2021) From Two to One: A New Scene Text Recognizer with Visual Language Modeling Network. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 11-17 October 2021, 14194-14203. https://doi.org/10.1109/ICCV48922.2021.01393

[15] Baek, Y., Lee, B., Han, D., Yun, S., *et al*. (2019) Character Region Awareness for Text Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 9365-9374. https://doi.org/10.1109/CVPR.2019.00959

[16] 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

[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*, **34**, 11474-11481. https://doi.org/10.1609/aaai.v34i07.6812

[18] 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

[19] Li, Y., Wu, Z., Zhao, S., *et al*. (2020) PSENet: Psoriasis Severity Evaluation Network. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 800-807. https://doi.org/10.1609/aaai.v34i01.5424

[20] Lyu, P., Liao, M., Yao, C., *et al*. (2018) Mask Textspotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 67-83. https://doi.org/10.1007/978-3-030-01264-9_5

[21] Zhang, R., Zhou, Y., Jiang, Q., *et al*. (2019) Icdar 2019 Robust Reading Challenge on Reading Chinese Text on Signboard. 2019 *International Conference on Document Analysis and Recognition* (*ICDAR*) *IEEE*, Sydney, 20-25 September 2019, 1577-1581. https://doi.org/10.1109/ICDAR.2019.00253

[22] Tan, M. and Le, Q. (2019) Efficientnet: Rethinking model Scaling for Convolutional Neural Networks. *International Conference on Machine Learning*. *PMLR*, Long Beach, 9-15 June 2019, 6105-6114.

[23] Liu, R., Lehman, J., Molino, P., *et al*. (2018) An Intriguing Failing of Convolutional Neural Networks and the Coordconv Solution. *Proceedings of the* 32*nd International Conference on Neural Information Processing Systems*, Montréal, 2-8 December 2018, 9628-9639.

[24] Liu, Z., Lin, Y., Cao, Y., Hu, H., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted

Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 11-17 October 2021, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

[25] Sun, Y., Ni, Z., Chng, C.K., Liu, Y., *et al*. (2019) ICDAR 2019 Competition on Large-Scale Street View Text with Partial Labeling-rrc-lsvt. 2019 *IEEE International Conference on Document Analysis and Recognition* (*ICDAR*), Sydney, 20-25 September 2019, 1557-1562. https://doi.org/10.1109/ICDAR.2019.00250

[26] Chng, C.K., Liu, Y.L., Sun, Y.P., *et al*. (2019) Icdar2019 Robust Reading Challenge on Arbitrary-Shaped Text-rrc-art. 2019 *IEEE International Conference on Document Analysis and Recognition* (*ICDAR*), Sydney, 20-25 September 2019, 1571-1576. https://doi.org/10.1109/ICDAR.2019.00252