文章编号：2095-2163(2023)12-0032-06 中图分类号：TP311.5 文献标志码：A

低质量海关报表字符识别模型研究

万 燕，范艺环，姚 砺，朱彦锦
（东华大学 计算机科学与技术学院，上海 201620）

摘 要：海关报表单据图像质量差，其中字符往往有模糊、笔画缺失、笔画粘连和噪声污染等特点。本文针对海关报表单据中低质量字符识别准确率低的问题，提出了 Enhanced-DBNet 文本检测模型并改进 ABINet 文本识别模型。基于 DBNet 模型重新设计其主干网络，引入可变形卷积模块（DCN）扩大感受野，提高长文本识别能力；采用双向特征金字塔增强模块（FPEM），使网络具有更强的表征能力；引入特征融合模块（FFM）将图像高层次语义特征和低层次位置特征充分融合。针对形近字符难区分的问题，在 ABINet 模型中引入可变形注意力模块，使注意力集中在字符相关区域，捕获到更多的字符特征。对比实验结果表明，本文的模型在海关报表低质量字符上的检测和识别准确率优于当前其他模型。

关键词：海关报表；字符识别；DBNet；ABINet

Research on character recognition model of low quality customs statements

WAN Yan, FAN Yihuan, YAO Li, ZHU Yanjin
（College of Computer Science and Technology, Donghua University, Shanghai 201620, China）

Abstract: The image quality of customs report documents is poor, and the characters are often characterized by blur, missing strokes, adhesion of strokes, and noise pollution. This paper proposes the Enhanced-DBNet text detection model and improves the ABINet text recognition model to solve the problem of low accuracy of low-quality character recognition in customs report documents. This paper redesigns the backbone network based on the DBNet model, introduces the deformable convolution module (DCN) to expand the receptive field and improves long text recognition capabilities, and uses the bidirectional feature pyramid enhancement module (FPEM) to make the network stronger representation capabilities and introduce feature fusion. The module (FFM) fully integrates high-level semantic features and low-level positional features of the image. To solve the problem of difficulty in distinguishing characters with similar shapes, a deformable attention module is introduced in the ABINet model to focus attention on character-related areas and capture more character features. Through comparative experiments, the model in this article has better detection and recognition accuracy on low-quality characters in customs reports than other current models.

Key words: customs statements; character recognition; DBNet; ABINet

0 引 言

海关报表单据中的报关信息尤为重要，快速且准确地识别海关报表单据中的文字成为海关供应链高效开展工作极为重要的一环。当前，采用人工识别的方式手动将信息输入到系统中，存在效率低下、成本高、漏识别、误识别等问题。

目前，光学字符识别（Optical Character Recognition，OCR）技术已有广泛的应用，能够准确地识别常见票据中的字符，如发票、火车票等，但是针对海关报表单据识别的公开研究成果非常少。由于海关报表单据采用机器扫描打印的方式，往往具有字迹不清晰、笔画粘连、单据图像模糊、噪声多等特点，且在海关报表单据中字符数量多、字体较小、文本行长宽比变化较大、文本密集度高，导致海关报表单据中字符的检测和识别难度更大。

OCR 识别过程包括文本检测和文本识别两个阶段。文本检测目的在于定位每一个文本实例的边界框；文本识别目的在于识别文本框中的文字的内容。

在文本检测方面，EAST（Efficient and Accuracy Scene Text）算法采用全卷积神经网络（FCN）直接对

作者简介：万 燕（1970-），女，博士，教授，硕士生导师，主要研究方向：图像处理、纤维的自动识别；范艺环（1996-），男，硕士研究生，主要研究方向：图像处理；姚 砺（1967-），男，博士，副教授，硕士生导师，主要研究方向：软件测试技术；朱彦锦（2000-），女，硕士研究生，主要研究方向：行为识别、可解释人工智能。

收稿日期：2022-12-15

文本行的边框进行检测，但在检测长文本时存在截断现象。渐进式扩展网络（PSENet）提出了一种渐进尺度扩展的像素分割算法，区分相邻的文本实例[1]；SegLink 模型回归文本片段的边界框，并通过预测边界框之间的链接关系将局部合并为一个整体，以处理长文本实例[2]；SegLink++模型在 SegLink 模型基础上，进一步提出实例感知组件分组算法，用来更好的分离文本实例，但是其模型收敛速度慢[3]；DBNet 模型中提出可微分二值化，将二值化这一过程融入模型中训练，得到自适应阈值，将分割网络和二值化过程联合优化[4]。但是上述模型在海关报表单据中，对字符不清晰的文本区域会出现检测不到的情况，或者因为噪声的干扰错误地将噪声区域归类为文本实例。

在文本识别方面，CRNN（Convolutional Recurrent Neural Network）算法可以将文本识别转换为时序依赖的序列问题，用双向长短时记忆网络（BiLSTM）和 CTC（Connectionist Temporal Classification）算法建模字符之间的上下文关系[5]；SRN（Semantic Reasoning Networks）模型中提出并行注意力模块，利用全局语义推理网络结合视觉特征，提高文本识别准确率[6]；SVTR（Scene Text Recognition with a Single Visual Model）模型中采用单视觉模型完成文字识别任务[7]；ABINet 模型中提出一种双向完形填空（BCN）语言模型，采用多模态融合的方式完成低质量图像文本识别任务[8]。

上述方法在印刷质量较好的字符上能取得很好的效果，但是针对海关报表中低质量的字符，识别准确率降低。

本文针对海关报表单据中检测和识别遇到的难题，提出了 Enhanced-DBNet 文本检测模型并对文本识别模型 ABINet 进行改进。在 Enhanced-DBNet 模型中采用可变形卷积神经网络（DCN）以适应不同长度的文本，引入双向特征金字塔增强模块（FPEM）增强不同尺度的特征，并使用特征融合模块（FFM）将不同层级的特征进行充分融合；在改进的 ABINet 模型中引入可变形注意力机制，使注意力模块能够聚焦于感兴趣区域并捕获更多的特征，提升视觉模型的识别效果。对海关报表中 1 200 张低质量字符图像进行识别，将本文提出的模型与 SRN 模型、SVTR 模型、ABINet 模型做对比实验，结果表明：本文所提出的模型在 1 200 张海关报表低质量字符图像上达到了 80.2%的识别准确率，明显高于其他模型。

1　基于 Enhanced-DBNet 和改进 ABINet 的模型结构

目前优秀的场景文本检测和识别模型都难以准确地识别海关报表单据中的低质量字符。文本检测阶段，DBNet 模型采用特征金字塔（FPN），表征能力弱、感受野较小，会出现低层位置特征信息丢失等问题；在文字识别阶段，外形相近的字符，如字母 i 和数字 1、字母 o 和数字 0 难以辨别，虽然 ABINet 模型提出多模态融合模型，但是语言模型依赖于视觉模型的输出结果，视觉模型预测的结果影响到最终语言模型修正的结果。

本文针对海关报表中低质量字符的识别提出了 Enhanced-DBNet 和改进的 ABINet 相结合的模型，模型结构图如图 1 所示。

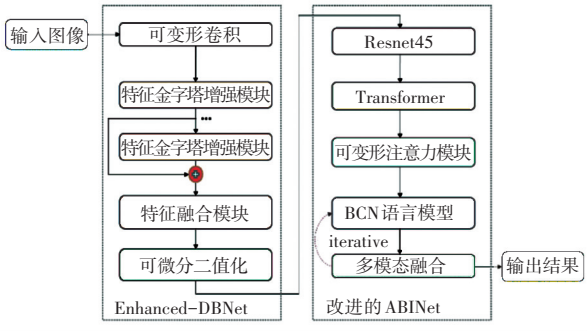


图 1　Enhanced-DBNet 和改进 ABINet 的模型结构图

Fig. 1　Model structure diagram based on Enhanced-DBNet and improved ABINet

用两阶段方法对海关报表中的字符进行识别。文本检测阶段，首先引入可变形卷积模块（DCN）使网络根据图像内容自主学习采样点的形状，适应不同长宽比的文本实例；其次，采用双向特征金字塔增强模块（FPEM），通过较小的计算开销增强不同尺度的特征，扩大感受野；最后，引入特征融合模块（FFM）将低级和高级特征信息充分融合，使模型具有更强的表征能力。在文本识别阶段，针对形近字符难区分及不清晰字符难以准确识别的问题，引入可变形注意力模块，使注意力聚焦字符附近感兴趣区域并捕获更多信息特征，有效提升视觉模型对不清晰文本和形近文字的识别能力。

本文模型的整体流程：首先，将文本图片输入到文本检测网络中，使用轻量 Resnet18 网络进行特征提取，经过 FPEM 增强特征后，由 FFM 将特征图进行融合，再采用可微分二值化（DB）方法将文本实例分割出来；其次，将分割出的文本实例送入文本识别模型中，使用 Resnet34 网络提取特征，使用

Transformer 模型序列化建模,再通过可变形注意力模块输出视觉模型的预测结果,再由双向完形填空语言模型(BCN)通过迭代校正视觉模型中预测错误的字符,提升不清晰字符的预测准确率,得到最终的识别结果。

1.1 Enhanced-DBNet 模型

海关报表图像经过卷积操作,语义信息不够充分,会包含较多的噪声,而高层次特征则包含有更多的语义信息,但是对文本的细节感知能力较差。因此,本文提出 Enhanced-DBNet 模型,引入特征金字塔增强模块(FPEM)、特征融合模块(FFM)对特征进行加强和融合,扩大感受野,使特征图同时包含文本的语义信息和位置细节信息。Enhanced-DBNet 模型结构图如图 2 所示。

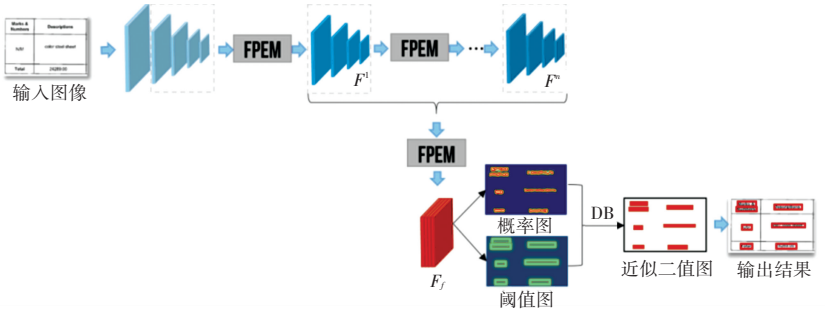


图 2 Enhanced-DBNet 模型结构

Fig. 2 Enhanced-DBNet overall model structure

本文提出的 Enhanced-DBNet 模型中,首先将图像输入到主干网络,主干网络中引入可变形卷积(DCN)模块,使网络可以自主学习可变形采样点的形状,自底向上进行卷积操作,获取到原图大小的 1/2、1/4、1/8、1/16、1/32 的特征图,由 1/4、1/8、1/16、1/32 特征图组成一个特征金字塔;其次,通过级联多个 FPEM 模块增强特征金字塔,经过 FFM 模块再将多个金字塔融合为特征图 F_f,通过特征图 F_f 生成对应的阈值图和概率图,根据两者生成二值图像,得到最终的文本识别结果。

FPEM 模块不仅包含上采样增强还包含下采样增强,双向增强不同尺度的特征。FPEM 模块中特征图连接部分使用的是可分离卷积,减少计算量。通过级联多个 FPEM 模块可以产生多个增强的金字塔,在扩大感受野的同时加深网络,使不同尺度的特征表达的更加充分。FPEM 模块细节如图 3 所示。

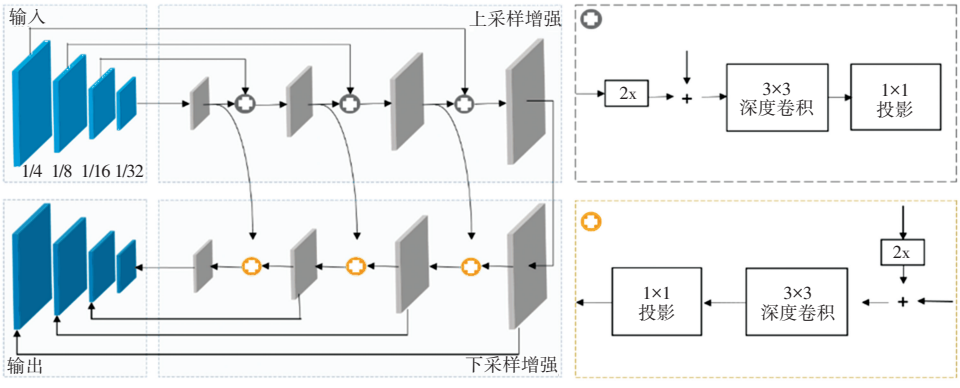


图 3 FPEM 模块细节

Fig. 3 Detail drawing of FPEM module

FFM 模块通过逐元素相加方式将相同层级对应位置的元素进行组合,将组合后的特征图上采样,获取到原图 1/4 大小的特征图用于预测。FFM 模块细节如图 4 所示。

1.2 改进的 ABINet 模型

在文本识别阶段,本文改进了 ABINet 网络结构中的视觉模型,采用可变形注意力机制,增强模型的灵活性,捕获更多的信息特征,提高视觉模型的预测准确率进而提高最终文字识别结果的准确率。

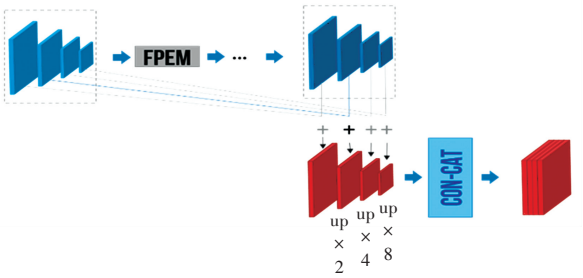


图 4　FFM 模块细节

Fig. 4　Detail drawing of FFM module

引入可变形注意力模块。可变形注意力模块结构图如图 5 所示。输入一张图像，图像中的字符为 2，字符轮廓与 z 相似，如果采用位置注意力机制，其采样点只在固定的位置，采样的特征分散。序列建模后，经过注意力计算很有可能将其识别为 z，但是可变形注意力模块可以将采样的位置偏移，计算权重矩阵时对“2”这一字符笔画所在区域分配更高的权重。

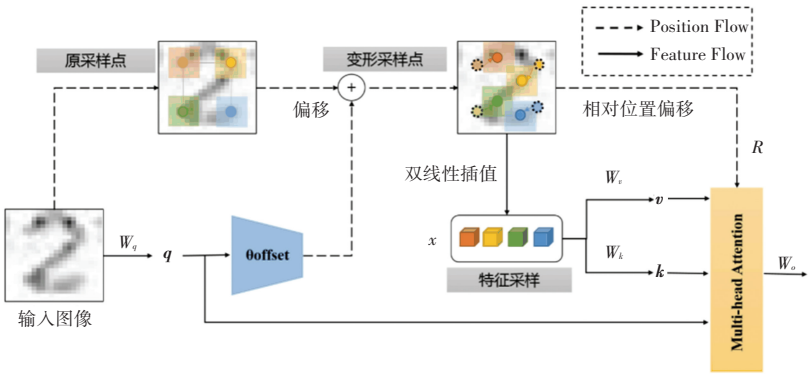


图 5　可变形注意力模块结构

Fig. 5　Deformable attention module structure

本文在 ABINet 模型中采用可变形注意力模块，通过查询向量 $\boldsymbol{q}$ 和每一个键向量 $\boldsymbol{k}$ 做交互时，给对应的 $\boldsymbol{k}$ 的位置加一个偏移量，候选的 $\boldsymbol{k}$ 和值向量 $\boldsymbol{v}$ 被转移到重要区域，聚焦在图像中文字附近的位置，在偏移后的位置用双线性插值进行特征采样，作为 $\boldsymbol{k}$ 和 $\boldsymbol{v}$，式(1)：

$$\begin{cases} \boldsymbol{q} = x\ W_q \\ \tilde{\boldsymbol{k}} = \tilde{x}\ W_k \\ \tilde{\boldsymbol{v}} = \tilde{x}\ W_v \end{cases} \tag{1}$$

其中，$\tilde{\boldsymbol{k}}$ 和 $\tilde{\boldsymbol{v}}$ 表示偏移之后的键向量和值向量。

将 $\boldsymbol{q}$ 输入到轻量偏移网络 θ_{offset} 中，生成偏移量 Δp，然后得到变形点 $\tilde{x}$，式(2)：

$$\begin{cases} \Delta p = \theta_{\text{offset}}(\boldsymbol{q}) \\ \tilde{\boldsymbol{x}} = \varphi(x; p + \Delta p) \end{cases} \tag{2}$$

其中，采样函数 $\varphi(\cdot;\cdot)$ 使用双线性插值方法，具备可微性，插值过程如式(3)所示：

$$\varphi(m;(p_x,p_y)) = \sum_{(r_x,r_y)} f(p_x,r_x) f(p_y,r_y)\, m[r_y,r_x,:] \tag{3}$$

其中，$f(a,b) = \max(0, 1 - |1 - b|)$。

由于只在最接近 (p_x,p_y) 的 4 个积分点上不为 0，因此双线性插值法简化了式(3)到 4 个点的加权平均值。

本文对 $\boldsymbol{q}$、$\boldsymbol{k}$、$\boldsymbol{v}$ 进行注意力机制计算，并采用相对位置偏移 r 为多头注意力模块提供更强大的相对位置偏移信息。可变形注意力头的输出如式(4)所示：

$$z = \text{softmax}\left(\frac{\boldsymbol{q}\,\tilde{\boldsymbol{k}}^T}{\sqrt{c}} + \varphi(\hat{B};R)\right)\tilde{\boldsymbol{v}} \tag{4}$$

其中，$\hat{B}$ 对应的是相对位置的偏置表，R 表示相对位置的偏移。

通过在笛卡尔坐标系的 x 和 y 方向上进行索引，得到相对位置偏置 B，由于可变形注意力模块具有连续的键位置，首先将偏移量做归一化，计算在 $[-1,1]$ 范围内的相对位移，然后做插值 $\varphi(\hat{B};R)$，得到最终的采样点。

2　实验与分析

2.1　实验环境和数据集

本文实验基于 Ubuntu20.04 LTS 操作系统，系统内存为 32 G，使用两块 RTX 2080Ti 显卡进行训练。

实验主要分为两个部分，即文本检测算法和文本识别算法的训练和验证。

文本检测算法：基础实验环境为 python = 3.6，

pytorch=1.5, numpy=1.17.4, 设置学习率为 0.001, 初始化阈值为 0.7, 每个实验均训练 1 200 轮。由于海关报表中文本都是水平文本, 所以本文采用的数据集为 ICDAR 2015, 共包含 1 500 张图片, 1 000 张作为训练集, 500 张作为测试集。

文本识别算法: 基础实验环境为 python=3.7, pytorch=1.1.0, 初始学习率设置为 0.000 1, 经过 6 轮之后衰减为 0.000 01, Batch-size 设置为 150, 使用 Adam 优化器, 实验一共训练 8 轮, 用于训练的数据集是两个合成数据集 MJSynth(MJ)和 SynthText(ST), 将 ICDAR 2013(IC13)、ICDAR 2015(IC15)、IIIT 5KWords(IIIT)、街景文本(SVT)、街景文字透视(SVTP)和 CUTE80(CUTE)作为测试数据集。本文从现有的海关报表图像中截取 700 张图片, 并进行数据增强, 将图像数量扩充至 1 200 张, 用于验证本文改进的模型在海关报表低质量字符上识别的效果。

2.2 实验结果比较与分析

2.2.1 DBNet 模型改进前后的实验结果和分析

在 DBNet 模型引入 DCN 模块、引入 FPEM 模块、FFM 模块后对模型性能的影响结果做对比。使用训练好的模型对 ICDAR-2015 测试集进行验证, 文本检测实验结果见表 1。

表 1 文本检测实验结果表

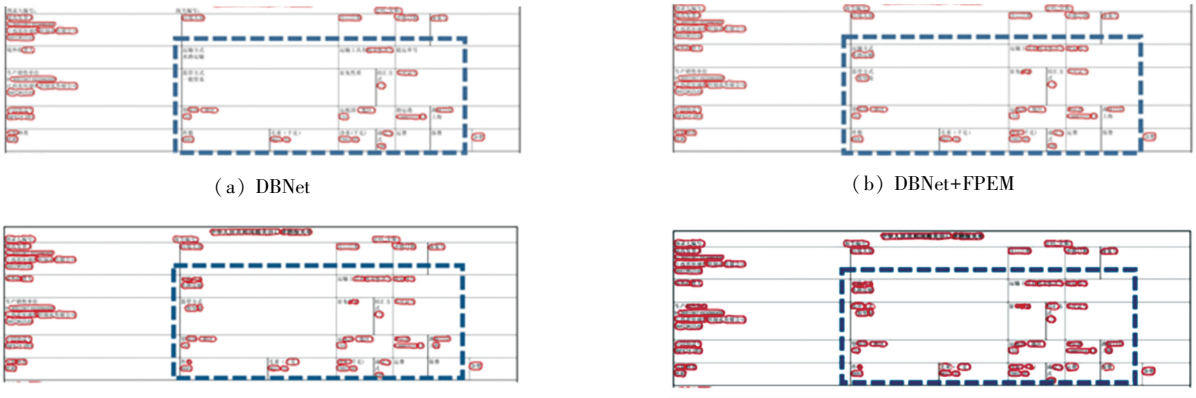
Table 1 Text detection experiment result table

模型名	可变形卷积	特征增强模块	精确率	召回率	*F* 值	每秒帧数
DBNET+DCN	有	FPN	88.6	73.8	80.5	21.8
DBNET	无	FPN	88.1	76.2	81.8	19.5
DBNET+FPEM+FFM	无	FPEM-FFM	87.4	75.4	81.0	24.4
DBNET+DCN+FPEM+FFM	有	FPEM-FFM	**90.0**	**76.6**	**82.7**	22.7

本文将海关报表图像分别输入到 DBNet 模型、DBNet + DCN 模型、DBNet + FPEM + FPN 模型、Enhanced-DBnet 模型中检验分割效果如图 6 所示, 文本中红色框代表检测出的文本区域, 图片中蓝色虚线框区域是存在差异的主要部分, 从图 6 中看出, 本文提出的 Enhanced-DBNet 模型对海关报表中文字区域检测取得了最好的检测效果。

2.2.2 ABINet 模型改进前后的实验结果和分析

将分割好的文本实例送入文本识别网络进行文字内容的识别。对改进的视觉模型(ABINet)经过 8 轮的训练, 将训练好的模型用 6 个测试集来验证效果, 对比本文采用可变形注意力模块与原文采用的位置注意力模块, 同时还对比平行注意力模块对模型的识别效果, 文本识别实验结果见表 2。



(a) DBNet (b) DBNet+FPEM

(c) DBNet+FPEM+FFM (d) Enhanced-DBNet

图 6 各模型对海关报表单据的检测结果图

Fig. 6 Diagram of the detection results of each model on the customs report documents

表 2 文本识别实验结果表

Table 2 Text recognition experiment result table

注意力模块	Transformer 层数	IC13	IC15	SVT	SVTP	IIIT	CUTE
平行注意力	3	94.5	81.1	89.5	83.7	94.3	**86.8**
位置注意力	3	94.9	81.7	**90.4**	**84.2**	94.6	86.5
可变形注意力	3	**95.0**	**82.7**	90.1	83.9	**94.7**	86.5

通过表 2 可以看出，本文改进后的模型在通用数据集上识别准确率提高，取得了较好的结果。为了进一步验证本文提出的模型在海关报表上的识别效果，在现有的海关报表单据中截取出的 700 张低质量的图片，包含字迹模糊、噪声干扰、字迹粘连、笔画缺失等情况，本文通过加入噪声或运动模糊等方式将图像数量扩充至 1 200 张。用本文改进的模型对这 1 200 张图片进行识别，并与当前先进的文本识别模型即 SRN 模型、SVTR 模型、ABINet 模型进行比较，1 200 张低质量海关报表单据图像上的识别准确率，见表 3。

表 3　1 200 张低质量海关报表图像的识别准确率

Table 3　Recognition accuracy of 1 200 low-quality customs report images

模型名称	识别正确个数	识别准确率/%
SRN	759	63.3
SVTR	878	73.1
ABINet	928	77.3
改进的 ABINet(本文)	963	**80.2**

用本文改进的 ABINet 模型与目前一些优秀的文字识别模型对海关报表中的部分文本实例识别对比的结果如图 7 所示，可以看出本文对海关报表中一些低质量文本字符能够准确的识别，优于其他模型的识别效果。

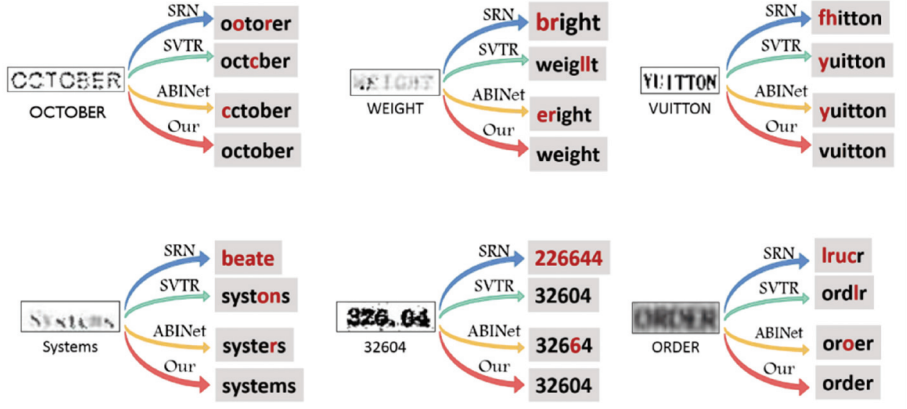


图 7　识别结果对比图

Fig. 7　Comparison chart of recognition results

3　结束语

针对海关报表单据中字符不清晰，关键信息字迹模糊、字迹连续、笔画缺失，噪声干扰等情况，本文基于场景文本检测模型 DBNet 和场景文本识别模型 ABINet 进行改进。首先，在 DBNet 模型中引入 FPEM 和 FFM 模块，并加入 DCN 模块，充分提取文字特征；其次，在 ABINet 模型中引入可变形注意力模块，使特征聚焦在文字关键区域。对比目前先进的模型，在公共数据集上取得不错的结果，同时对 1 200张海关报表图片验证，本文提出的模型能够有效的提升海关报表中低质量字符的识别结果，识别准确率达 80.2%，高于其它模型。

参考文献

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

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

[3] TANG J，YANG Z，WANG Y，et al. Seglink++：Detecting dense and arbitrary-shaped scene text by instance-aware component grouping[J]. Pattern Recognition，2019，96：106954.

[4] LIAO M，WAN Z，YAO C，et al. Real-time scene text detection with differentiable binarization [C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York：AAAI，2020：11474-11481.

[5] SHI B，BAI X，YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2016，39(11)：2298-2304.

[6] YU D，LI X，ZHANG C，et al. Towards accurate scene text recognition with semantic reasoning networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：12113-12122.

[7] DU Y，CHEN Z，JIA C，et al. Svtr：Scene text recognition with a single visual model[J]. arXiv preprint arXiv：2205.00159，2022.

[8] FANG S，XIE H，WANG Y，et al. Read like humans：Autonomous，bidirectional and iterative language modeling for scene text recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：7098-7107.