万燕,李毅凡,姚砺. 基于注意力机制的海关报表识别方法研究[J]. 智能计算机与应用,2024,14(4):26-33. DOI:10.20169/j.issn.2095-2163.240404

基于注意力机制的海关报表识别方法研究

万 燕, 李毅凡, 姚 砺
(东华大学 计算机科学与技术学院, 上海 201620)

摘 要: 海关报表作为进出口业务中的重要材料,需要快速识别文本并录入系统,提高业务效率。海关报表图像通常存在字迹模糊粘连、字号过小和噪声污染等问题,增加了报表文本识别的难度。本文针对海关报表图片识别准确率低的问题,提出了基于注意力机制的海关报表识别方法。在 DBNet 模型中引入了注意力机制,提升小字符文本检测能力,使网络更加关注字符相关区域;在视觉模型中引入可变形卷积模块,扩大感受野,并将视觉特征和语义特征增强后通过门控机制实现多模态融合,提升对低质量字符的识别精度。实验结果表明,本文方法在海关报表低质量图像的检测和识别准确率方面领先其他方法。

关键词: 海关报表识别; 注意力机制; 文本识别
中图分类号: TP391.43 **文献标志码**: A **文章编号**: 2095-2163(2024)04-0026-08

Research on the method of customs statement identification based on attention mechanism

WAN Yan, LI Yifan, YAO Li
(College of Computer Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: Customs statements, as important materials in import and export business, need to quickly identify the text and enter it into the system to improve business efficiency. Customs statement images usually have characteristics of blurred and sticky handwriting, too small font size and noise pollution, which increase the difficulty of statement text recognition. In this paper, we propose a customs statement recognition method based on the attention mechanism for the problem of low accuracy of customs statement image recognition. Attention mechanisms are introduced in DBNet model to enhance the small character text detection ability and make the network pay more attention to the character-related regions. The deformable convolution module is introduced in the visual model to expand the perceptual field, and the visual features and semantic features are enhanced to achieve multimodal fusion through the gating mechanism to improve the recognition accuracy of low-quality characters. The experimental results prove that this paper leads other methods in the detection and recognition accuracy of low-quality images in customs statements.

Key words: customs statements recognition; attention mechanisms; text recognition

0 引 言

中国已经成为进出口贸易大国,2022 年中国进出口贸易额突破 40 万亿元,同比增长 7.7%。海关报表和单据是进出口业务中必不可少的材料,快速准确地检测并识别报表中的文字信息并录入系统是海关业务的重要环节。传统的人工录入方法存在成本高、易误录等问题。因此,研究一种快速准确检测识别海关报表的方法具有实际意义和应用价值。

OCR(Optical Character Recognition)技术能够将打印或手写文本转化成计算机可操作的字符,已在票据识别、证件识别等领域有成熟的应用,但缺少识别海关报表和单据的应用研究。海关报表和单据大多采用扫描打印的方式,存在字迹模糊、图像质量差等问题,且报表的字符稠密,字号过小,字符间常存在粘连的情况,样例如图 1 所示,增加了文本检测和识别的难度。

作者简介: 李毅凡(1998-),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。
通讯作者: 万 燕(1970-),女,博士,教授,硕士生导师,主要研究方向:图像处理。Email:winniewan@dhu.edu.cn

收稿日期: 2023-05-03

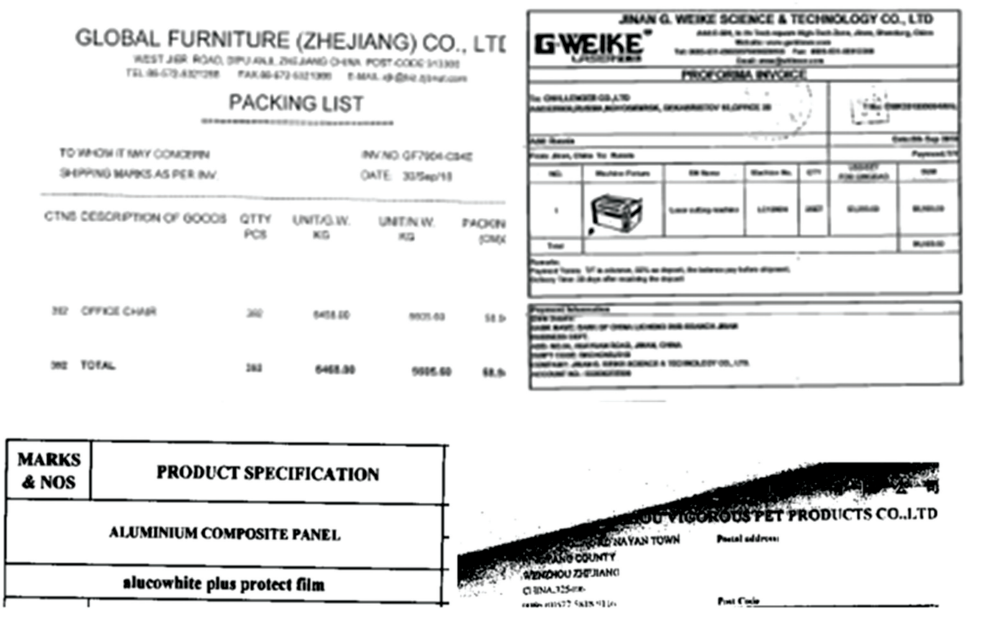


图 1　海关报表样例图

Fig. 1　Customs statement examples

OCR 的核心流程包含文本检测和识别两个阶段。文本检测的任务是定位字符或文本行的位置；文本识别的任务是识别出图像中的字符或文本行内容。

当前，文本检测包含自顶向下和自底向上两类方法。自顶向下的方法是将目标检测任务迁移到文本检测任务中，没有考虑到文本行的形状和文本间粘连密集的特点，导致识别效果较差而被逐渐淘汰。

自底向上的方法是目前文本检测研究的主流方法。Jaderberg[1] 基于 R-CNN 实现了文本区域的定位；王建新等[2] 在此基础上提出了 Deep Text 模型，对文本区域进行了更为精确的回归。为了实现多方向的文本检测，Bai 等[3] 设计了 SegLink 模型，将离散的文本片段连接，形成完整的文本行；在此基础上，Tang 等[4] 又提出了 SegLink++模型，将不同文本子区域相斥，进一步提升了模型性能；Gupta 等[5] 在 YOLO 模型基础上，针对不同大小的图像采用全卷积网络进行文本位置定位，也取得了良好效果；Wang 等[6] 提出 PSENet，采用 FPN 结构（Feature Pyramid Network）将各文本实例划分到像素级别，提出了渐进性尺度扩张算法以获得文本实例分割结果，但模型的后处理比较复杂，造成前向预测效率太低。为解决此问题，Wang 等[7] 提出了 PANNet 模型，在轻量化特征提取与融合网络的基础上附加了一个像素相似向量，用于把文本像素聚集在正确文本核上，从而获得不同实例的检测结果，但对于弯曲文本效果欠佳；Liao 等[8] 提出 DBNet 模型，首次将可微分二值化的算法引入到文本检测任务中，该方法不再需要手工设计后处理步骤，除了输出预测概率图，还预测对应的阈值图，二者结合得到最后的结果，提升了后处理的效率和模型的前向推理速度，对不同方向的文本检测效果良好。

目前，文本识别的主流方法是通过视觉模型提取特征，转化为字符序列送入语言模型提取语义特征，将视觉和语义特征结合实现文本实例的识别。Wojna 等[9] 基于卷积神经网络和循环神经网络提出 Attention-OCR 模型，采用多个卷积神经网络提取特征后输入注意力模块加权；Shi 等[10] 用长短期记忆网络替换了循环神经网络，提高了模型对于不规则文本的识别准确率；Yu 等[11] 提出了一种新的语义推理网络 Spatial Regularization Network（SRN）来替换 RNN 网络，并且引入全局语义推理模块，通过并行传输挖掘语义信息来辅助文本识别；Fang 等[12] 提出 ABINet 模型，设计了一个双向的语言网络，将视觉特征作为辅助信息输入到语言模型迭代；Wang 等[13] 提出 VisionLAN 模型，放弃了复杂的语言模型，采用视觉为主、语义为辅的方法，增强视觉特征，提升模型识别准确率；Zhou 等[14] 提出了 PIMNet 模型，采用并行注意力机制来提高预测文本的速度。

除了上述方法外，端到端的文本检测与识别方法也得到了应用。Liao 等[15] 采用一个无锚点的分割网络替换 Region Proposal Network，实现了多方向

文本的识别。为了解决小尺寸文本图片识别效果差的问题,Liao 等[16]提出了 TextBoxes 模型,能够检测任意方向和小尺寸图片中的文本;Qiao 等[17]提出 TextPerceptron 模型,通过一个形状转换模块,将文本区域校正为规则形状,提升了识别正确率。Liu 等[18]在单阶段目标检测器的基础上提出了 ABCNet 模型,利用贝塞尔曲线建立文本模型,从而提高端到端文本检测和识别的效率。

上述方法在图像质量较好的场景文本和印刷文本的检测任务中取得了较好的效果,但对于图像质量差,小目标文本繁多,文本密集的海关报表图片上表现较差。因此,本文提出了基于注意力机制的海关报表识别方法解决上述问题。

文本检测方法基于 DBNet 模型,该模型引入了可微分二值化算法,实现了阈值的自适应学习,提高了模型的检测精度和文本的检测速度,被 OpenCV、MMOCR 和 Wechat 等多家机构和公司收录作为文本检测模型。原始的 DBNet 模型采用 FPN 作为特征融合模块,本文在 FPN 中引入 SE(Squeeze-and-Excitation)注意力机制,并在特征融合后添加 CBAM(Convolutional Block Attention Module)模块,使网络关注对文本区域影响较大的通道和位置,降低不重要的通道和位置的权重,从而提升模型的表现能力和泛化能力。在此基础上优化了损失函数,降低了背景像素损失的权重,更好地评估模型表现。

文本识别网络采用视觉模型和语言模型融合的方式,ResNet 和 Transformer。为了解决 ResNet 因网络深度增加导致的退化问题,本文在其中引入可变形卷积模块扩大感受野,使采样区域更贴合字符的尺寸和形状。提取到的视觉特征转换成字符序列,输入语言模型提取语义特征,得到的视觉特征和语义特征经过 Transformer 增强,通过一个门控机制进行融合,融合后的结果再输入语言模型进行迭代,使网络更好地利用视觉和语义信息做出预测,提高模型对低质量文本的识别能力。

1 基于注意力机制改进的 DBNet 文本检测网络

由于现有的文本检测和识别技术在噪声繁多的海关报表上效果不佳,难以完成自动化识别报表的任务。因此,本文设计了针对海关报表检测识别任务的网络模型,其识别过程如图 2 所示,包括文本检测和文本识别两个模块。

文本检测模块在 DBNet 模型基础上进行改进,DBNet 模型虽然在场景文本检测领域表现优异,但是在海关报表检测任务表现较差,原始 DBNet 模型在海关报表的检测效果如图 3 所示,可以看出原始DBNet 模型漏检较多,对于字符较小或长度较短的文本实例检测效果差,造成诸多关键数据的遗漏。

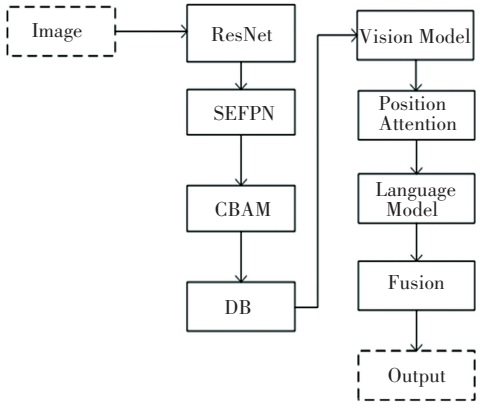


图 2 海关报表检测识别过程图

Fig. 2 Customs statement detection and recognition process chart

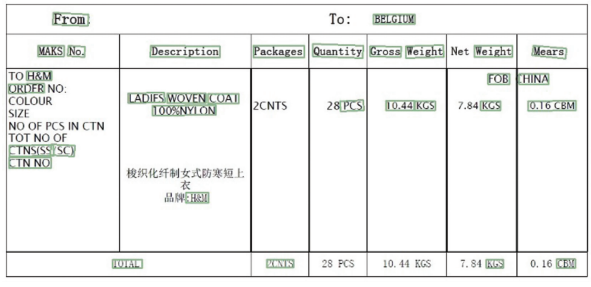


图 3 DBNet 检测海关报表效果图

Fig. 3 Customs statements in DBNet detection result image

原始的 DBNet 模型采用轻量级 ResNet18 作为骨干网络提取特征,然后输入 FPN 模型进行多次下采样,得到不同尺度的特征图,在下采样过程中会丢失部分细节信息导致网络无法提取到鲁棒的特征,对小目标文本区域检测效果差。

为解决上述问题,本文提出基于注意力机制改进的 DBNet 文本检测模型,进行了三项改进,其网络结构如图 4 所示。首先,在 FPN 中融合 SE 注意力机制组成 SEFPN,SE 模块能够获取全局信息,补充因 FPN 下采样导致的细节信息丢失;其次,引入 CBAM 模块,CBAM 包含通道注意力和空间注意力机制,能够进一步补充细节信息,增强网络对文本所在位置和通道的感知能力;最后,对损失函数进行优化,修正了背景损失函数的权重,使损失函数更贴合海关报表文本检测的特点,改善模型表现。

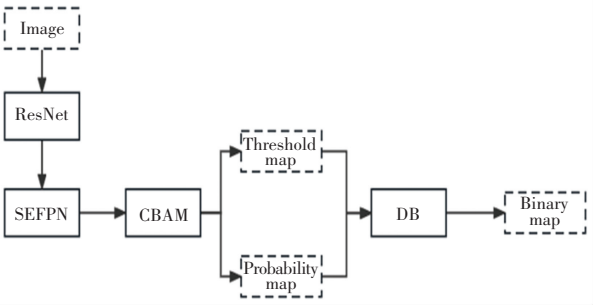


图 4　基于注意力机制的 DBNet 网络结构

Fig. 4　Structure diagram of DBNet based on attention mechanism

引入 SE 注意力机制。DBNet 模型采用 FPN 作为特征融合模块，对于小目标的物体检测效果欠佳。本文在 FPN 中引入 SE 模块组成 SEFPN，SE 模块是一种关注通道特征的注意力机制，网络结构如图 5 所示。首先，全局平均池化输入特征图 F，把各通道特征图压缩成一常量，获得通道级全局特征；其次，利用全连接网络 FC（Fully Connected Network）产生各通道权重，也就是学习各通道之间的相互关系，最后对各通道特征图进行权重归一化，从而获得 F'。SE 模块能够让模型关注更为重要的通道，操作前后不改变特征图大小且参数量小，具有即插即用的特点，可以嵌入到任何网络结构中。

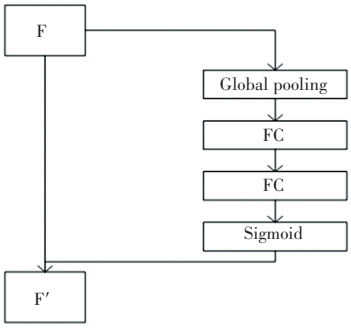


图 5　SE 结构

Fig. 5　Squeeze-and-Excitation structure diagram

引入 CBAM 模块。CBAM 包括通道注意力与空间注意力两部分，其结构如图 6 所示。输入特征图先进入通道注意力模块中，再通过平均池化层与最大池化层获取两个不同空间上下文信息；将这两个特征输入多层感知机进行融合并输出最终通道注意力特征图。

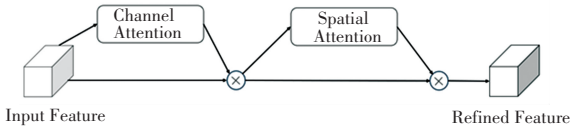


图 6　CBAM 结构

Fig. 6　CBAM structure diagram

通道注意力和空间注意力机制的级联组合能够使模型更加关注和任务相关的通道与位置，增强了模型的表现能力。

优化损失函数。在文本检测任务中，文字所占像素点为前景像素，其余像素点为背景像素，基于分割的检测网络需要对每个像素点分别计算损失。由于海关报表文本图片的特殊性，前景像素所占的区域比例很小，背景像素所占的区域比例很大，二者极端的差异导致原有的损失函数不能很好地评估模型性能。

本文在原有损失函数中添加一个手工设置的超参数，减小背景损失所占权重。优化后的损失函数由 3 部分构成，如式（1）所示：

$$L = L_s + \alpha * L_b + \beta * L_t \tag{1}$$

其中，L_s 代表概率图的损失；L_b 代表二值图的损失；L_t 代表阈值图的损失；α 和 β 是可训练的超参数。

概率图的损失包含前后景像素点损失，二者可通过交叉熵损失计算得到，超参数 *Weight* 实现对背景像素损失加权，*Weight* 设置在 0 和 1 之间时，得到的背景像素损失值将显著减小，如式（2）所示：

$$L_s = Weight * negLoss + posLoss \tag{2}$$

其中，*negLoss* 代表背景像素点的损失和，*posLoss* 代表前景像素点的损失和。

2　基于可变形卷积的多模态融合文本识别网络

当前，采用视觉模型和语言模型结合是文本识别任务的主流方法，通常需要将二者得到的特征进行融合再预测，常见的视觉特征和语义特征融合方式有以下 4 种：简单组合视觉特征和语义特征，如 SRN 模型；视觉特征作为辅助增强语义特征，如 PIMNet 模型；语义特征作为辅助增强视觉特征，如 VisionLAN；视觉特征和语义特征同时增强再融合输出，即本文方法。

对于文本识别任务，基于图像的视觉特征和基于上下文的语义特征都是不可或缺的，有机地结合二者是提高模型预测精度的关键。因此，本文提出基于可变形卷积的多模态融合文本识别网络，在视觉模型中引入可变形卷积，并在融合模块中添加 Transformer 增强视觉和语义特征，充分提取视觉和语义两个模态的信息，识别网络结构如图 7 所示。

引入可变形卷积。视觉模型采用 ResNet45 和 Transformer 作为骨干网络提取特征，为了解决网络深度增加导致的退化和普通卷积感受野固定的问

题,本文将 ResNet 中 3×3 卷积层替换为可变形卷积,扩大感受野从而提取更广泛的特征。

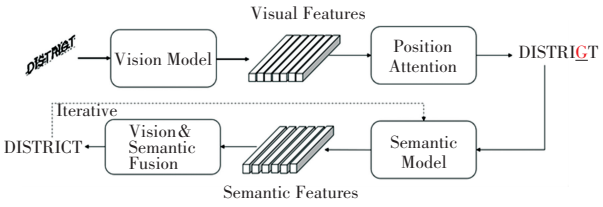


图 7 文本识别网络结构图

Fig. 7 Text recognition network structure diagram

可变形卷积结构如图 8 所示,与普通卷积采样大小固定不同,可变形卷积额外添加一个卷积层来计算 x 和 y 方向的偏移量,再采用双线性插值将偏移量转换为整型,得到对应的像素值。

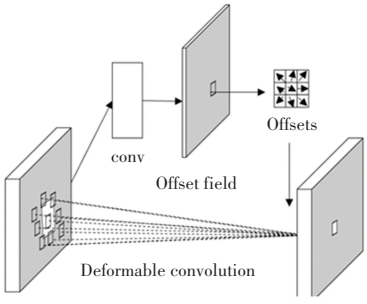


图 8 可变形卷积结构图

Fig. 8 Deformable convolution structure diagram

视觉模型得到的字符序列经过位置编码后作为语言模型的首次输入,语言模型由多个 Transformer 块构成,默认为 3 个。本文设计的文本识别模型和 VisionLAN 相同,引入了带有掩码的 Transformer,在训练过程中会随机遮盖某个位置的字符像素,再通过语义特征学习恢复,从而增强语义模型的鲁棒性。融合模块由多个 Transformer 组成,视觉和语义特征经过 Transformer 增强后,通过一个门控机制将视觉特征和语义特征融合,传入全连接层得到预测结果。G 是可训练参数,如公式(3)所示:

$$G = Sigmoid([F_V, F_L] W_{gate}) \tag{3}$$

其中,F_V 和 F_L 分别是增强后的视觉特征和语义特征。

Fusion 是最终融合后输出的结果,如公式(4)所示:

$$Fusion = G \times F_V + (1 - G) \times F_L \tag{4}$$

3 实验与分析

3.1 实验环境和数据集

所有实验基于 Ubuntu20.04 操作系统,系统内存 32 GB,使用两张 Nvidia RTX 3090 显卡,软件版本 Pytorch1.7.1,cuda11.0。

文本检测任务使用 ICDAR2015 数据集,共包含 1 500 张图像,7 548 个标注文本实例,其中 1 000 张作为训练集,500 张作为测试集,每组实验训练 500 轮。

文本识别任务使用 MJSynth 和 SynthText 训练集,二者都是用于场景文本识别的合成数据集。MJSynth 包含超过 900 万张字符级标注的合成文本图像;SynthText 包含超过 800 万张合成文本图像。

为了验证本文提出文本识别方法的有效性,首先在公开测试集进行实验,然后在海关报表数据集进行测试。公开数据集分为两组:规则文本测试集和不规则文本测试集。规则文本测试集的图片主要是水平文本行图片,不规则文本测试集包含多种形状的文本实例,图片质量更差,噪声干扰较多。

在公开数据集测试后,本文构建了一个包含 700 张真实海关报表文本图片和 300 张添加不同程度高斯噪声的报表图片的测试数据集,用于验证模型在海关报表上的表现。

3.2 实验结果比较和分析

3.2.1 文本检测网络的实验结果和分析

实验包含 4 组,第一组实验分别对原始 DBNet 模型,引入 SE 后的模型,以及本文引入 SE 和 CBAM 后的模型进行训练并在测试集上评估,实验结果见表 1。评价指标主要包括精度(*Precision*)、召回率(*Recall*)、*F*1 分数和 *FPS*(Frames Per Second)。

表 1 文本检测实验结果

Table 1 Text detection experiment results

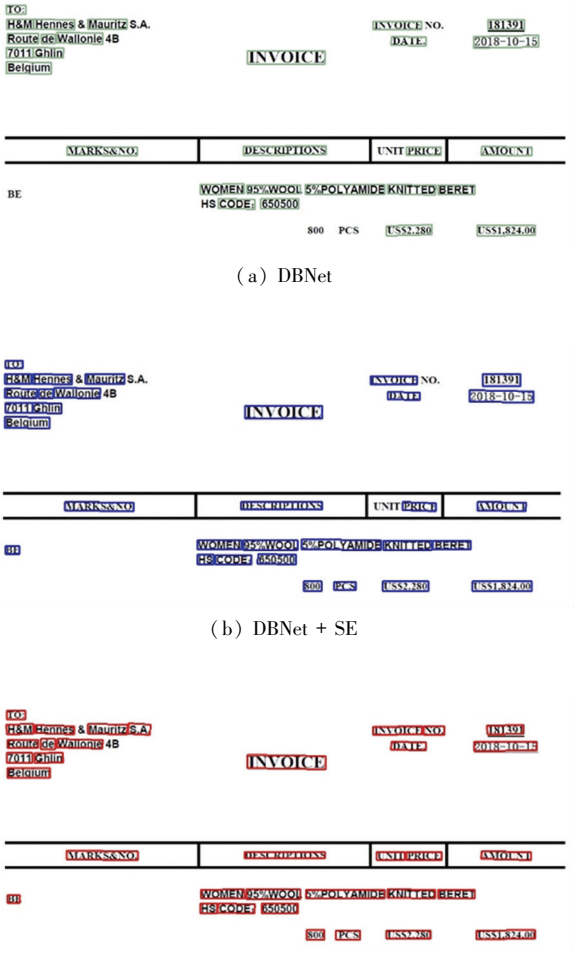
模型	*Precision*/ %	*Recall*/ %	*F*1/ %	*FPS*
原始 DBNet 模型	90.6	62.5	76.0	40.4
DBNet+SE 模型	88.5	67.5	77.5	30.3
DBNet+SE+CBAM 模型(本文)	89.3	72.5	**80.0**	24.8

实验结果表明在引入 SE 和 CBAM 模块后,*F*1 分数相较于原始模型分别上升了 1.5%和 4.0%,检测能力得到了显著提升。

各模型在海关报表上检测效果如图 9 所示,明显看出添加 SE 和 CBAM 模块的模型报表文字检测效果更好,几乎不存在漏检情况。

第二组实验基于添加注意力机制后的检测网络,并引入优化后的损失函数,采用不同的背景损失权重进行实验并在测试集评估,实验结果见表 2,可见添加背景损失权重能够提高模型的评估指标分数,当权重参数设置为 0.01 时,模型在测试集上表

现最好。



（a）DBNet

（b）DBNet + SE

（c）DBNet + SE + CBAM

图 9　各模型在海关报表上检测效果图

Fig. 9　Comparison of the text detection effect of each model customs statement

表 2　背景损失权重对比实验

Table 2　Background loss weight comparison experiment

权重	*Precision*/ %	*Recall*/ %	*F*1/ %	*FPS*
0.100	89.6	76.4	82.3	30.5
0.010	90.6	76.8	**83.9**	28.2
0.001	89.7	74.8	81.2	30.8

为了比较改进后的模型和其他主流文本检测模型的差异，本文在 ICDAR2015 测试集上进行了对比实验，最终结果见表 3。实验结果表明本文提出的基于注意力机制的 DBNet 模型相比其他主流文本检测算法 *F*1 指标更高。

表 3　主流文本检测模型对比实验结果

Table 3　Comparison experiment results of mainstream text detection models

模型	*Precision*/ %	*Recall*/ %	*F*1/ %	Year
TextSnake	84.9	80.4	82.6	2018
DBNet	88.1	76.2	81.8	2019
PAN	86.3	81.9	82.9	2019
FAST-T[19]	86.0	77.9	81.6	2021
DBNet++[20]	90.1	77.2	83.1	2022
本文模型	90.6	76.8	**83.9**	

3.2.2　文本识别网络的实验结果比较和分析

文本识别网络负责识别图片中的文本信息。实验共有 4 组，第一组对原始视觉模型和添加可变形卷积的视觉模型分别进行 8 轮训练，并在测试集上评估准确率，实验结果见表 4。实验结果表明在视觉模型中加入了可变形卷积模块后，识别准确率有明显提升，为语义模型和融合模型的训练提供了良好的先验知识。

表 4　视觉模型对比实验结果

Table 4　Vision model comparison experiment results　　%

是否添加可变形卷积	规则数据集				不规则数据集			
	IIIT	SVT	$IC13_S$	$IC13_L$	$IC15_S$	$IC15_L$	SVTP	CUTE
×	92.6	85.9	91.9	90.1	76.0	72.3	77.2	78.8
√	**93.7**	**87.2**	**94.9**	**93.8**	**81.9**	**78.1**	**83.3**	**83.0**

第二组实验选取了不同方式融合视觉和语义特征的网络进行对比，验证不同融合方式对识别准确率的影响，实验结果见表 5，可见采用视觉语义同时增强再融合的方法更具优势。

表 5　不同融合方式模型对比实验结果

Table 5　Experiment results of comparing models of different fusion methods　　%

模型	规则数据集			不规则数据集		
	IIIT	SVT	$IC13_S$	$IC15_S$	SVTP	CUTE
SRN	94.8	91.5	95.5	82.7	79.5	84.7
PIMNet	95.8	90.7	93.5	80.6	85.7	86.5
VisionLAN	95.8	91.7	95.7	83.7	86.0	88.5
本文模型	**95.9**	**93.4**	**98.1**	**85.6**	**90.5**	**89.2**

第三组实验目的是验证本文改进的识别网络与其他主流文本识别网络的差异,本文选取了 SEED 模型、MaskOCR 模型和 TrOCR 模型进行对比实验,实验结果见表 6,可见本文提出的文本识别网络模型在多个公开测试集上具有明显优势。

表 6 主流文本识别模型对比实验结果

Table 6 Comparison experiment results of mainstream text recognition models

模型	年份	规则数据集				不规则数据集				耗时/ms
		IIIT	SVT	$IC13_S$	$IC13_L$	$IC15_S$	$IC15_L$	SVTP	CUTE	
SEED	2020	93.8	89.6		92.8	80.0		81.4	83.6	53.3
MaskOCR	2022	95.8	94.7	98.0		87.3		89.9	89.1	
TrOCR	2022	90.1	91.0	97.3	96.3	81.1	75.0	90.4	86.8	22.9
本文模型		**95.9**	93.4	**98.1**	**96.3**	85.6	**81.7**	**90.5**	**89.2**	32.6

为了验证本文模型在海关报表文本识别优势,进行了第四组对比实验,真实噪声报表数据集包含 700 张从报表图片截取的带有噪声的文本行图片,模拟噪声报表数据集包含 300 张随机添加高斯噪声的报表文本行图片,实验结果见表 7,可见本文模型在两个海关报表数据集分别取得了 82.0%和 68.7%的准确率,均高于其他主流模型。

表 7 不同模型在海关报表数据集对比实验结果

Table 7 Experiment results of comparing different models in customs statement dataset %

模型	真实噪声报表数据集	模拟噪声报表数据集
SEED	62.8	59.8
TrOCR	70.8	60.6
MaskOCR	75.6	64.5
本文模型	82.0	68.7

不同模型在报表图片的识别结果如图 10 所示,可见其他模型存在误检漏检等问题,识别效果较差,本文提出的基于可变形卷积的多模态融合文本识别模型在海关报表数据集上均领先其他模型。

图 10 各模型在海关报表识别效果图

Fig. 10 Effect of each model in the recognition of customs reports

4 结束语

针对海关报表中图像质量差、文本字号小、存在笔画缺失、墨迹干扰等情况,本文提出了基于注意力机制的海关报表识别方法解决上述问题。在 DBNet 模型的 FPN 模块中引入 SE 注意力机制,得到通道级别的全局特征,并在该模块后引入 CBAM 注意力机制,提高网络在重要通道和重要位置的权重,以此应对单字符、小字符和噪声繁多的文本区域漏检问题;在视觉模型中引入可变形卷积,扩大卷积感受野,更加贴合文本的尺寸和形状,并在融合模块中引入 Transformer 增强视觉和语义信息,使二者更加有机地结合,提高文本识别精度。本文模型在多个公共数据集上准确率超过目前先进的其他模型,同时在两个海关报表数据集上识别准确率分别达到 82.0%和 68.7%,均领先其他模型。

本文针对海关报表文本自动化识别提出了一种可行的方法,满足实时检测识别的速度和精度,对于海关业务自动化智能化转型具有重要意义和应用价值。

参考文献

[1] JADERBERG M, SIMONYAN K, VEDALDI A, et al. Reading text in the wild with convolutional neural networks [J]. International Journal of Computer Vision, 2016, 116(1): 1-20.

[2] 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): 32.

[3] 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.

[4] 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.

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

[6] 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.
[7] WANG W，XIE E，SONG X，et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision.2019：8440-8449.
[8] 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. 2020：11474-11481.
[9] WOJNA Z，GORBAN A N，LEE D S，et al. Attention-based extraction of structured information from street view imagery [C]//Proceedings of 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE，2017：844-850.
[10] SHI B，YANG M，WANG X，et al. Aster：An attentional scene text recognizer with flexible rectification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2018，41(9)：2035-2048.
[11] 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. IEEE，2020：12113-12122.
[12] 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.
[13] WANG Y，XIE H，FANG S，et al. From two to one：A new scene text recognizer with visual language modeling network[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：14194-14203.
[14] ZHOU Y，QIAO Z，WEI J，et al. Pimnet：A parallel，iterative and mimicking network for scene text recognition [C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021：2046-2055.
[15] LIAO M，LYU P，YAO C，et al. Mask textspotter：An end-to-end trainable neural network for spotting text with arbitrary shapes [C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018：67-83.
[16] LIAO M，SHI B，BAI X. TextBoxes++：A Single-shot oriented scene text detector[J]. IEEE Transactions on Image Processing，2018，27(8)：3676-3690.
[17] QIAO L，TANG S，CHENG Z，et al. Text perceptron：Towards end-to-end arbitrary-shaped text spotting[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020：11899-11907.
[18] LIU Y，CHEN H，SHEN C，et al. Abcnet：Real-time scene text spotting with adaptive bezier-curve network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.2020：9809-9818.
[19] CHEN Z，WANG W，XIE E，et al. FAST：Searching for a faster arbitrarily-shaped text detector with minimalist kernel representation[J]. arXiv preprint arXiv：2111.02394，2021.
[20] LIAO M，ZOU Z，WAN Z，et al. Real-time scene text detection with differentiable binarization and adaptive scale fusion[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2022，45(1)：919-931.