文章编号:1671-4598(2024)09-0256-06 DOI:10.16526/j.cnki.11-4762/tp.2024.09.036 中图分类号:TP391.41; TP18 文献标识码:A

基于改进 MTSv2 的场景文本检测和识别算法研究

王艳媛，茅正冲，杨雨涵
(江南大学 物联网工程学院，江苏 无锡 214000)

摘要：在自然场景图像中，丰富的文本内容对于全面理解场景非常重要。针对自然场景文本图像存在背景复杂、文本粘连、文本多角度等问题，提出一种基于改进 MTSv2 的文本检测和识别算法；检测算法以 MTSv2 为基础网络，采用 CBAM 注意力机制增大特征图中的小型文本的权重，更好捕捉图像中的关键特征；融合 CE-FPN 结构，减轻多尺度融合产生的特征混叠问题；引入 focal loss 函数，减少正负样本分布不均衡对识别准确率的影响，使网络更加关注难以分类的样本，改善模型的泛化能力；通过多个文本数据集进行训练，并在 ICDAR2015 数据集上进行验证，改进后模型对场景文本检测和识别的准确率达到了 89.3%，召回率达到了 87.6%，F_1 值达到了 88.5%，相比于原模型都有一定程度的提高。

关键词：场景文本；文本检测；文本识别；CBAM；CE-FPN；注意力机制

Research on Scene Text Detection and Recognition Algorithm Based on Improved MTSv2

WANG Yanyuan, MAO Zhengchong, YANG Yuhan
(School of Interest of things, Jiangnan University, Wuxi 214000, China)

Abstract: In natural scene images, rich text content is very important for a comprehensive understanding of the scene. Aimed at the problems of complex background, sticky text, and multi-angle text in natural scene text images, a text detection and recognition algorithm based on improved MTSv2 is proposed. The detection algorithm takes MTSv2 as the base network, adopts the convolutional block attention module (CBAM) attention mechanism to increase the weight of small text in the feature map, so as to better capture the key features in the image; the channel enhancement-feature pyramid network (CE-FPN) structure is fused to alleviate the feature aliasing problem generated by multi-scale fusion; The focal loss function is introduced to reduce the influences of the positive and negative sample distribution imbalance on the recognition accuracy, making the network focused on difficult to classify the samples, and improving the generalization ability of the model. Through training on multiple text datasets and validation on the ICDAR2015 data set, the accuracy of the improved model on the scene text detection and recognition reaches 89.3%, the recall rate reaches 87.6%, and the F_1 value reaches 88.5%, this model improves the above indicators to a certain extent compared with the original model.

Keywords: scene text; text detection; text recognition; CBAM; CE-FPN; attention mechanism

0 引言

文本是保存和交流信息的基本工具之一，随着计算机视觉领域的发展，许多新兴应用场景都需要提取图像文本中的信息[1]。光学字符识别（OCR，optical character recognition）作为传统的文本检测算法非常适用于基于数字平面的扫描文档的信息[2]，但场景文本图片的定位和识别技术对于提高工业自动化水平和场景理解能力有着非常重大的意义，相比于 OCR 也具有更大的挑战，例如：图像模糊、背景复杂，文本多角度、字体风格多变等。自然场景文本的检测和识别在实践中有丰富的研究背景和广泛的应用价值，可以应用于图像搜索、自动驾驶、实时翻译和机器人导航等场景[3]。

自然场景中的文本存在模糊、阴影、光照变化等噪声和干扰的影响，以及自然场景中的文本字体具有多样性且文本会以不同方向出现的特点[4]。为了解决各种因素对于场景文本的检测和识别带来的影响，提出了许多的场景文本检测和识别的方法。HE 等在 EAST[5]模型的检测模块简

收稿日期:2023-09-05； **修回日期**:2023-10-16。
基金项目:国家自然科学基金(61901206);国家自然科学基金青年项目(6170185)。
作者简介:王艳媛(1999-),女,硕士研究生。
通讯作者:茅正冲(1964-),男,硕士,副教授。
引用格式:王艳媛,茅正冲,杨雨涵.基于改进 MTSv2 的场景文本检测和识别算法研究[J].计算机测量与控制,2024,32(9):256-261.

化了中间过程，在文本识别模块中，提出了两种策略来提高编码字符空间信息的准确性。第一种策略是引入注意力对齐学习，通过学习对文本中不同字符的关注程度来提高准确性；第二种策略是引入额外的聚焦损失，让网络在关键位置更加关注字符边界来提高准确性。但该方法对于弯曲、旋转等不规则文本的效果不是很理想。BAEK 等[6]则以 CRAFT[7]文本检测框架为基础提出了 CRAFTS 方法，使用薄板样条变换纠正文本纠正模块和字符区域注意力模块中的任意形状的文本区域，同样取得了较好的识别性能。WANG 等[8]提出基于文本中心区域的任意形状文本的表示方法，并成功区分了密集相邻文本。同时，他们还采用了轻量级注意力识别头来改进识别模型，并提出了高效的端到端识别框架 PAN＋＋[8]。LIU 等[9]则提出了一种基于贝塞尔曲线的场景文本检测和识别方法 ABCNet，该方法用三阶贝塞尔曲线对不规则形状文本自适应拟合，设计贝塞尔对齐层来精确提取任意形状文本实例的卷积特征，在效率和精度上都具有很大优势。

上述提出的研究方法都能较好的检测和识别场景文本，然而对于难以分类的小型文本的检测和识别效果并不理想，存在漏检、误检的现象。这是因为小型文本的字符高度和宽度较小，导致小型文本字符之间的间距较小，更容易受到相邻文本和背景的干扰，使得网络更难分辨和提取。且小型文本在图像上的分辨率像素通常较低，字符更易失去清晰度，边缘模糊不清，导致字符细节更难被捕捉。

如何检测和识别场景文本中的小型文本内容，并提高文本检测的准确率，是一个具有挑战性的问题。本文基于 MaskTextspotterv2 网络改进，首先在骨干网络 ResNet50 中加入 CBAM 注意力机制，增加小型文本的权重，强化小型文本的特征；其次融合 CE-FPN 网络，减少跨尺度融合造成的特征混叠，避免不必要的混杂的综合特征导致的混淆定位和识别问题；最后在网络中采用 focal loss 函数，使模型更关注相对难分类的文字区域，提高文本检测和识别鲁棒性。实验结果表明，文本改进的算法能够检测出一定的场景文本中的小型文本，有效提高了检测和识别的准确性。

1　MaskTextspotter 模型结构及其原理

MaskTextspotter 系列基于经典的目标检测网络 Mask R-CNN[10]，总体架构主要包括两个部分：一个基于实例分割的检测部分和一个基于字符分割的识别部分[11]。本文选择的 MTSv2[12]相比于原来的 MaskTextspotter 结构的不同之处主要在于：在识别分支增加了空间注意模块（SAM，spatial attentional module），即将空间注意力应用到识别部分。SAM 能直接解码二位特征映射，以更好地表示各种形状。先通过双线性插值将给定的特征映射调整为固定形状，这个特征包括 MaskTextspotter 中的 ROI 特征，也包括独立识别模型中的主干特征图，然后依次通过一个卷积层、一个最大池化层和一个卷积层，将此时的输出特征与位置编码特征拼接起来一起再通过基于空间注意力的循环神经网络（RNN，recurrent neural network），最后在文本识别过程中输出文本序列。

2　改进的 MTSv2 的模型结构

2.1　改进后系统架构图

场景文本的检测和识别算法需要考虑要场景文本背景复杂，图片模糊，存在光照变化等影响，同时要求能够检测和识别场景文本图片中的小型文本。但是由于各种噪声和干扰的存在，小型文本的检测和识别较为困难。为此，提出改进 MTSv2 算法，其系统架构图如图 1 所示，在骨干网络的残差网络部分引入 CBAM 注意力机制，使网络增强小型文本的特征，更容易识别出小型文本，提高检测的准确率；同时，融合了 CE-FPN 网络，减少跨尺度融合造成的特征混叠，减少混叠特征对检测和识别的影响；使用 focal loss 函数，减少正负样本不均衡对网络造成的影响，提高网络鲁棒性。

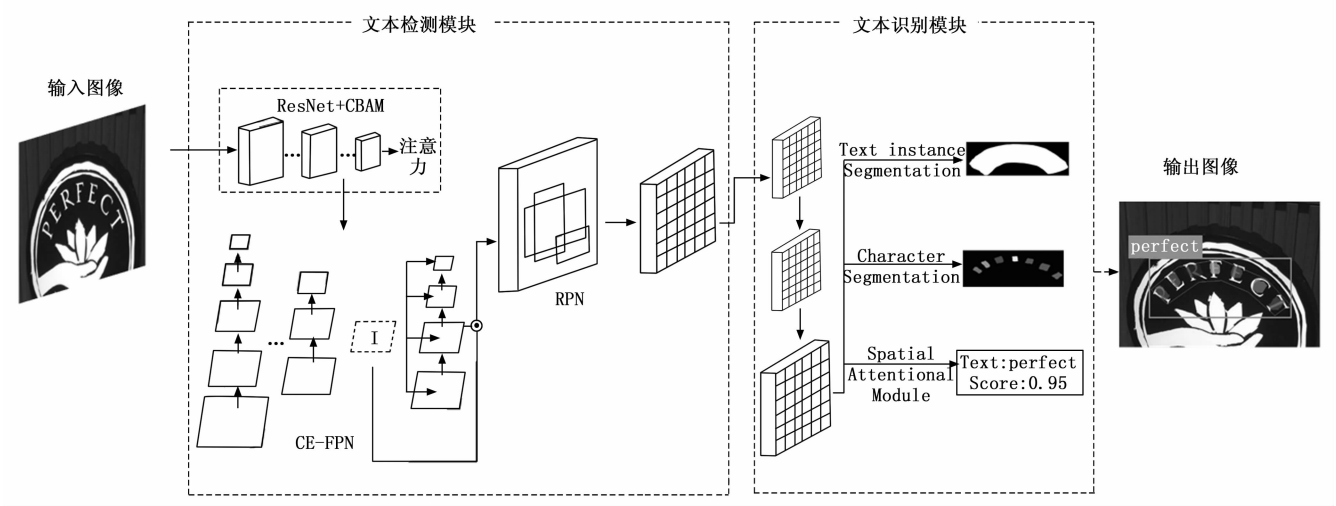

图 1　改进 MTSv2 的架构图

2.2 CBAM

卷积注意力模块（CBAM，convolutional block attention module）[14]是一种融合了空间和通道注意力机制的模块。它在集成到任何卷积神经网络（CNN，convolutional neural network）架构中时具有优越性，能无缝地忽略其开销。该模块通过在通道和空间两个维度上生成注意力特征图信息，并与输入的特征图相乘以进行自适应特征优化，生成最终的特征图。如图 2 所示，在 ResNet 网络的每个卷积阶段之后引入 CBAM 注意力机制，自适应地调整特征图的通道和空间信息，帮助网络更好地理解和表示输入数据。

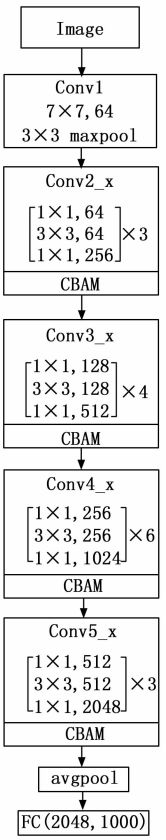

图 2 引入 CBAM 的 ResNet 网络

通道注意力模块（CAM，channel attention module）从输入数据中提取特征，这些特征通常是以多个通道的形式表示，每个通道对应不同的特征或信息。接着对每个通道的特征使用平均池化和最大池化来聚合特征映射的空间信息，其中全局平局池化是对每个通道的二维图像求均值，而全局最大值池化是选取局部接受域中值最大的点。接下来，将特征图输入共享网络，压缩输入特征图的空间维度，并进行逐元素求和。最后，通过 sigmoid 激活操作产生通道注意力图。通道注意力机制可以表达为：

$$M_c(F)=\sigma\{MLP[AvgPool(F)]+MLP[MaxPool(F)]\} \quad (1)$$

空间注意力模块（SAM，Spatial Attention Module）先对 CAM 输出的特征图进行基于通道的压缩，并进行平均值池化和最大值池化操作。接着将得到的特征图通道拼接后进行 7×7 卷积，并降维成 1 个通道，最后通过 sigmoid 激活后生成空间注意力特征。空间注意力机制可表达为：

$$M_s(F)=\sigma\{f^{7\times7}[AvgPool(F);MaxPool(F)]\} \quad (2)$$

CBAM 注意力机制能够自动学习图像中的关键信息，并将更多的注意力放在这些重要区域上，这样可以提高模型的表达能力，帮助模型更好地捕捉图像中的细节和上下文信息，从而提高模型在各种视觉任务中的性能。在 CAM 中，它能自动地选择和调整特征图中不同通道的重要程度。它可以帮助模型突出局部的关键的通道信息，这样能够提升检测小型文本的能力。而在 SAM 中，它可以自动地调整特征图中不同位置的重要程度。它可以帮助模型更加关注小型文本所在的区域，增强对小型文本的敏感度，从而提高小型文本的检测和识别准确率。

2.3 通道增强特征金字塔网络

传统的特征金字塔网络（FPN，feature pyramid network）[15]采用 1×1 卷积来减少输出特征的通道维度，间接损失了通道信息。不同层的语义信息之间可能存在区别，采用差值的方法进行跨尺度融合可能会导致混叠问题，即综合特征的混杂可能会干扰定位和识别任务。

根据传统 FPN 所产生的一些问题，将原先骨干网络 ResNet50-FPN 中的 FPN 网络结构结合通道增强特征金字塔网络 CE-FPN（Channel Enhancement-Feature Pyramid Network）[16]，其基本架构如图 3 所示。CE-FPN 通过卷积得到四层特征，表示为 $\{C_2, C_3, C_4, C_5\}$，分别相较于原图的 $\{4, 8, 16, 32\}$ 缩放，其中 $\{F_2, F_3, F_4\}$ 为通过 1×1 卷积获得 256 维通道的特征，特征金字塔由 $\{P_2, P_3, P_4\}$ 自上而下的方法获得。重复特征融合不仅会造成明显的混叠效应，还会带来额外的计算负担，降低检测效率。故不加入 FPN 语义中的特征最高级别的 F_5 和 P_4 进行差值计算和最大值池化计算，而是通过 $\{P_2, P_3, P_4\}$ 进行差值和最大值池化计算获得整合图 I，最后的预测在模型生成的 $\{R_2, R_3, R_4, R_5\}$ 上分别执行，这与原始的特征金字塔相互对应，代替原 FPN 的输出结果。

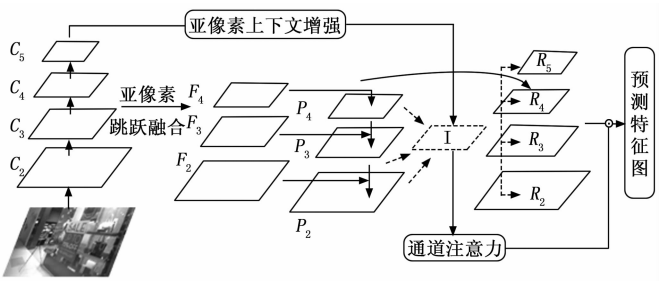

图 3 通道增强特征金字塔网络的总体架构图

CE-FPN 由 3 个主要部分构成，分别是：亚像素跳跃融合 SSF（SSF，sub-pixel skip fusion）、亚像素上下文增强 SCE（SCE，sub-pixel content enhancement）以及通道注意

力指导模块 CAG（CAG，channel attention guided module）。在 SSF[16]部分，没有采用传统的线性插值来进行上采样，而是采用残差融合亚像素卷积的方法对将特征进行上采样到目标大小，并对其进行 1×1 的卷积操作，然后将该结果与亚像素卷积后的特征进行融合，最终生成目标结果 F_4 和 F_3。通过同时进行的通道增强和上采样操作，取代了原始的 1×1 卷积和线性上采样方法，从而减少了由于信道减少而导致的信息丢失。SCE 模块融合大视野局部信息和全局上下文信息，用于提取更多的特征表示。在亚像素上下文增强部分，主要是对 C_5 的特征进行 3 种处理，先应用 3×3 卷积进行特征提取，同时将通道维数变为亚像素上采样；然后通过使用 3×3 最大池化操作，将输入特征进行下采样，将其尺寸调整为 $w\times h$，再经过 1×1 卷积层扩展通道尺寸；接着在 C_5 执行全局平均池化，以获取全局上下文信息；将得到的特征图与 P_4、P_3、P_2 进行特征融合，得到特征图 I。SCE 通过扩展 3 个不通尺度的表征特征，增大了 C_5 的感受野，提升了 I 的表征能力。引入 CAG 模块优化每个级别上的最终特征集成。它还用来指导 FPN 消除重叠影响，它只通过整合图 I 提取通道权重，然后乘以不同的输出特征。整合图 I 经过注意力指导模块得到权重 α，再通过线性插值和最大值池化得到 $\{R_5, R_4, R_3, R_2\}$，最后将 R_i（$i=2$，3，4，5）乘以权重 α，得到预测特征图，用于后续的场景文本的检测和识别。

CE-FPN 能够推广到各种基于 FPN 的检测器，只增加了少量的计算量。为了解决 FPN 构建过程中跨尺度融合造成的信息混叠问题，本文提出融合该模块，减少信息损失对文本检测和识别的影响，并优化复杂集成化的层的最终特征。

2.4　focal loss

focal loss[17]是一种用于解决类别不平衡问题的损失函数，它通过减小容易分类样本的权重，增大难分类样本的权重，从而更加关注难以分类的样本。将 Fast R-RCNN 使用的交叉熵损失函数修改为 focal loss，交叉熵损失函数公式定义为：

$$\text{Loss}=L(y,\hat{p})=-y\log(\hat{p})-(1-y)\log(1-\hat{p}) \quad (4)$$

其中：$\hat{p}$为预测概率大小，y 为标签值，在二分类中对应 0，1。而 focal loss 的公式定义为：

$$L_{fl}=-(1-p_t)^{\gamma}\log(p_t) \quad (5)$$

其中：$\gamma>0$ 为可调节因子，本文选取 γ 值为 2，当 γ 取值为 0 时，focal loss 函数与交叉熵损失函数等价。γ 取值越大，难分类样本的损失就越大。p_t 表示分类器预测 t 类样本的概率值，反映了与类别 y 的接近程度。p_t 的表示形式如下：

$$p_t=\begin{cases}\hat{p}\rightarrow y=1\\ 1-\hat{p}\rightarrow \text{otherwise}\end{cases} \quad (6)$$

在自然场景文本检测和识别中，文字通常是较小的目标，而且通常数据集中的正负样本比例不均衡。在这样的条件下使用传统的交叉熵损失容易导致模型过度关注易分类的背景区域，而忽略了难以分类的文字区域。而 focal loss 将易分类样本的权重降低，使得模型更加关注难以分类的文字区域，提高文本检测的准确性和鲁棒性。

3　实验结果与分析

3.1　实验细节

本文的实验平台为 Ubuntu20.04.3 系统，框架选用 Pytorch，所选取的数据集为英文场景文本的数据集，包括（STR，synthetic text renderer）团队发布的合成文本数据集 SynthText[18]，该数据集通过合成技术生成了大量带有真实背景的合成文本图像；国际字符识别大会 ICDR 提供的自然场景文本检测和识别的数据集：ICDAR 2013 和 ICDAR 2015[19]；以及弯曲文本数据集 total-text[20]。模型采用 SGD 梯度下降算法进行优化，初始学习率为 0.01，权重衰减为 0.001，动量为 0.9，学习率在第 10 万次和第 20 万次迭代时分别衰减到原始值的十分之一，而在第 27 万次结束迭代。在微调阶段，初始学习率为 0.001，学习率在第 10 万次迭代时衰减到原始值的十分之一，并在第 30 万次结束迭代。

预训练和微调结束之后将 ICDAR2015 的 500 张图像用于测试。通过非极大值抑制（NMS，non-maximum suppression）后，RPN 生成的候选框被输入 Fast R-CNN，将 Fast R-CNN 和 NMS 过滤后的错误的和多余的候选框去除，保留下来的候选框被输入到掩码分支，以生成文本实例、字符和文本序列相关的映射。最后根据网络预测的结果生成文本实例边界框，检验网络对于文本的检测能力。

3.2　评价指标

本文选择精确率（Precision）、召回率（Recall）和 F_1 值[21]作为衡量模型检测和识别的能力的指标。精确率衡量了模型正确预测为正的占全部预测为正的比例。它的计算公式为：

$$p=\frac{TP}{TP+FP} \quad (8)$$

其中：TP 是真正例个数，FP 是假正例个数。精确率越高，说明模型将负例预测为正例的能力越弱。召回率衡量了模型正确预测为正例的样本占真实正例的比例。它的计算公式为：

$$R=\frac{TP}{TP+FN} \quad (9)$$

其中：TP 是真正例个数，FN 是假负例个数。召回率越高，说明模型发现正例的能力越强。F_1 值是精确率和召回率的综合评估指标，它是精确率和召回率的调和均值，计算公式为：

$$F_1=\frac{2PR}{P+R} \quad (10)$$

F_1 值综合考虑了模型的准确性和全面性，是一个常用的综合指标。

3.3 实验结果及分析

3.3.1 消融实验

为了验证本研究对 MTSv2 提出的 3 种改进策略的有效性，本文进行实验证明。通过在原有模型基础上逐步添加改进措施，使用相同的参数配置进行训练，并选择 ICDAR2015 数据集进行消融试验。我们使用 F_1 值作为模型评价指标。表 1 展示了各个改进点的有效性，√ 表示使用了相应的改进，空表示未使用改进。

表 1 消融实验结果

改进措施	第一组	第二组	第三组	第四组
CBAM		√	√	√
CE-FPN			√	√
focal loss				√
F_1 值/%	83.4	85.0	86.7	88.4

消融实验以原生 MTSv2 模型作为基准，通过表 1 可以看出，文本对于文本检测和识别网络的改进效果得到了有效验证。通过表格的第一组和第二组数据对比可以得出，在原来的网络基础上添加 CBAM 模块，模型的 F_1 值的从 83.4%提高到 85.0%，表明 CBAM 模块能够有效的使网络关注特征图中的小型文本的权重，提高文本检测的准确率。通过表格的第二组和第三组数据对比可以得出，在模型中增加 CE-FPN 结构，模型的 F_1 值的从 85.0% 提高到 86.7%，说明网络在一定程度上解决了跨尺度融合产生的特征混叠问题。通过表格的第三组和第四组数据对比可以得出，模型引入 focal loss 函数，模型的 F_1 值的从 86.7% 提高到 88.4%，说明网络更加关注难以分类的样本，能够一定程度上改善网络的性能。

3.3.2 文本性能测试

为了证明改进后的 MTSv2 模型在文本检测方面的优势，本文将其与一些经典的文本检测算法进行了对比。本文使用 ICDAR2015 数据集检验各类方法模型的性能，该数据集具有背景复杂，光照不均、图片模糊等干扰因素。将本模型检测端与其他模型的检测段进行比较，结果如表 2 所示。

表 2 ICDAR2015 数据集上的对比结果

算法	精确率	召回率	F_1 值
CTPN[22]	0.740	0.520	0.610
Seglink[23]	0.731	0.768	0.750
EAST[5]	0.805	0.728	0.744
TextBoxes++[24]	0.872	0.767	0.817
CRAFT[7]	0.854	0.895	0.874
MTSv2[10]	0.866	0.873	0.870
改进后的 MTSv2	0.893	0.876	0.885

从表 2 可以看出，本文改进后的模型在 ICDAR2015 数据集进行性能测试，检测的精确率、召回率以及 F_1 值都有着一定程度的提高。改进后的 MTSv2 由于加入了针对小目标权重的 CBAM 注意力机制以及其他改进措施，相比于原生 MTSv2 模型在检测的精确率上提高了 2.7%，网络复杂度的相对提高对网络检测的召回率有一定影响，所以召回率的提升不是很大，相较于原模型只提高了 0.3%，但改进后的 MTSv2 检测的 F_1 值相比于原模型提高了 1.5%。故从综合层面来讲，本文的算法在一些经典的文本检测和识别模型中取得了一定的优势，算法的整体表现是优于其他模型的。

总之，本文提出的基于 MTSv2 的改进模型在准确性和全面性的综合考虑下取得了最佳效果，在没有引入复杂结构的前提下，提高了原网络的性能，不会过多增加网络的计算负担，且提高了英文场景文本检测的准确率。

图 4 为本文改进后的模型在 total-text 数据集训练后选取的测试结果图。从图中可以看出该模型能够比较准确的检测和识别出自然场景中的英文文本，误检漏检的情况较少。不论是复杂背景还是不同光照的情况下，都能识别出图片中的英文文本，同时对于图片中的小型文本也能够准确的检测和识别出来。

图 4 改进后的 MTSv2 的检测和识别效果图

4 结束语

为了解决自然场景文本的畸变、遮挡、多角度以及背景复杂等造成检测和识别难度较大的问题，以及为了进一步提高检测和识别的准确率，本文提出了基于 MTSv2 的改进模型，在骨干网络中 ResNet50 中加入 CBAM 注意力机制，增大特征图中的小型文本的权重，提高检测的准确率；通过融合 CE-FPN 网络，减少跨尺度融合可能会造成混叠的问题；最后引入 focal loss 损失函数，对于容易分类的样

本，降低其权重，而对于难分类的样本，增加其权重，使网络更关注难以分类的样本。

实验表明，本文提出的基于 MTSv2 的改进模型提高了自然场景文本检测和识别的精度，对于自然场景文本的检测和识别的准确率达到了 89.3%，召回率达到了 87.6%，F_1 值达到了 88.5%，证明了算法的有效性。目前所做工作主要是针对英文场景文本的检测和识别。在后续工作中，将尝试减少模型参数，实现轻量化的模型结构；以及增加多语种数据，实现识别多语种。

参考文献：

[1] 刘崇宇，陈晓雪，罗灿杰，等. 自然场景文本检测与识别的深度学习方法 [J]. 中国图象图形学报，2021，26 (6)：1330 - 1367.

[2] 谢开宇. 基于分割和编解码的 OCR 技术研究 [D]. 哈尔滨：哈尔滨工业大学，2021.

[3] 童朝娣. 卷积神经网络在车牌字符识别中的应用 [J]. 电子技术与软件工程，2019 (1)：68.

[4] 周　燕，韦勤彬，廖俊玮，等. 自然场景文本检测与端到端识别：深度学习方法 [J]. 计算机科学与探索，2023，17 (3)：577 - 594.

[5] ZHOU X Y，et al. EAST：An efficient and accurate scene text detector [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2017：5551 - 5560.

[6] GREENHALGH J，MIRMEHDI M. Recognizing text-based traffic signs [J]. IEEE Transactions on Intelligent Transportation Systems，2014，16 (3)：1360 - 1369.

[7] BAEK，YOUNGMIN，et al. Character region awareness for text detection [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2019：9365 - 9374.

[8] WANG W，XIE E，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，2021，44 (9)：5349 - 5367.

[9] 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.

[10] HE K M，et al. Mask R-CNN [C] //Proceedings of the IEEE International Conference on Computer Vision，2017：2961 - 2969.

[11] LYU P，LIAO M，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.

[12] LIAO M，LYU P，HE M，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.

[13] ZHU X，CHENG D，ZHANG Z，et al. An empirical study of spatial attention mechanisms in deep networks [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision，2019：6688 - 6697.

[14] WOO，S et al. CBAM：Convolutional block attention module [C] //Proceedings of the European Conference on Computer Vision (ECCV)，2018：3 - 19.

[15] LIN T Y，DOLLAR P，GIRSHICK R，et al. Feature pyramid networks for object detection [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2017：2117 - 2125.

[16] LUO Y H，et al. CE-FPN：Enhancing channel information for object detection [J]. Multimedia Tools and Applications，2022，21：30685 - 30704.

[17] LIN T Y，et al. Focal loss for dense object detection [C] //Proceedings of the IEEE International Conference on Computer Vision，2017：2980 - 2988.

[18] BAEK，JEONGHUN，et al. What is wrong with scene text recognition model comparisons dataset and model analysis [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision，2019：4715 - 4723.

[19] KARATZAS，DIMOSTHENIS，et al. ICDAR 2015 competition on robust reading [C] //2015 13th International Conference on Document Analysis and Recognition (ICDAR)，IEEE，2015：1156 - 1160.

[20] CHNG，CHEE KHENG，CHAN C S. Total-Text：A comprehensive dataset for scene text detection and recognition [C] //2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)，IEEE，2017：935 - 942.

[21] 胡高丽，文成玉. 自然场景下交通标识文本检测与识别算法研究 [J]. 成都信息工程大学学报，2022，37 (2)：171 - 176.

[22] TIAN Z，et al. Detecting text in natural image with connectionist text proposal network [C] //Computer Vision-ECCV 2016：14th European Conference，Amsterdam，The Netherlands，October 11 - 14，2016，Proceedings，Part VIII 14. Springer International Publishing，2016：56 - 72.

[23] SHI B G，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.

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