中图法分类号：TP391.4　**文献标识码**：A　**文章编号**:1006-8961(年) -

论文引用格式：

像素聚合和特征增强的任意形状场景文本检测

师广琛[1]，巫义锐[1,*]

1. 河海大学计算机与信息学院, 南京 211100

摘　要：**目的** 获取场景图像中的文本信息对理解场景内容具有重要意义，而文本检测是对文本识别、理解的基础。目前，场景文本检测是最具挑战性的任务之一，正受到越来越多的研究关注。**方法** 本文提出了一种高效的任意形状文本检测器：非局部像素聚合网络，该方法使用特征金字塔增强模块和特征融合模块进行轻量级特征提取，保证了速度优势；同时引入非局部操作以增强骨干网络的提取特征的能力，使其检测的准确性得到提高。非局部操作是一种注意力机制，它能捕捉到文本像素之间的内在关系。此外，本文设计了一种特征向量融合模块，用于融合不同尺度的特征图，使尺度多变的场景文本实例的特征表达得到增强。**结果** 本方法在3个场景文本数据集上与其他方法进行了比较，在速度和准确度上均表现突出。在ICDAR 2015数据集上，本方法比最优方法的F值提高了1.5%，检测速度达到了23.1FPS；在CTW 1500数据集上，本方法比最优方法的F值提高了1.8%，检测速度达到了71.8FPS；在Total-Text数据集上，本方法比最优方法的F值提高了0.8%，检测速度达到了64.3FPS，远远超出其它方法。**结论** 本文所提出的方法兼顾了准确性和实时性，在准确度和速度方面均处于领先水平。

关键词：目标检测；场景文本检测；神经网络；非局部模块；像素聚合；实时检测；任意形状

Arbitrary shape scene text detection based on pixel aggregation and feature enhancement

Shi Guangchen[1]　Wu Yirui[1,*]

1. Department of School of Computer and Information, Hohai University, Nanjing 211100, China

Abstract: Objective Text can be seen everywhere in real life, such as street signs, billboards, newspapers and other items. The text on these items expresses the information they want to convey. The ability of text detection determines the level of text recognition and understanding of the scene. With the rapid development of modern technologies such as computer vision and Internet of Things, many emerging application scenarios need to extract text information from images. In recent years, some new methods for detecting scene text have been proposed. However, many of these methods are slow in detection due to the complexity of the huge post-processing methods of the model, which limits their deployment in reality. On the other hand, the previous high-efficiency text detectors mainly used quadrilateral bounding boxes for prediction, and it is difficult to accurately predict arbitrary-shaped scene. **Method** In this paper, an efficient arbitrary shape text detector is proposed named non-local pixel aggregation network (Non-local PAN). Non-local PAN follows a segmentation-based method to detect scene text instances. In order to increase the detection speed, the backbone network must be a lightweight network. However, the presentation capabilities of lightweight backbone

收稿日期：2020 年 8 月 27 日　；　**修回日期**：2021 年 2 月 7 日

基金项目:本课题得到国家重点研发计划(No.2018YFC0407901);国家自然科学基金(No.B200202177);中央高校基本科研业务费专项资金资助(No.B200202177);江苏省自然科学基金(No.BK20170892)资助.

Supported by:National Key R&D Program of China(Grant 2018YFC0407901);the Fundamental Research Funds for the Central Universities(Grant B200202177); the Natural Science Foundation of China(Grant 61702160); the Natural Science Foundation of Jiangsu Province(Grant BK20170892).

networks are usually weak. Therefore, in this paper, a non-local module is added to the backbone network to enhance its ability to extract features. Resnet-18 is used as the backbone network of Non-local PAN, and non-local modules are embedded before the last residual block of the third layer. In addition, in this paper a feature vector fusion module is designed to fuse feature vectors of different levels to enhance the feature expression of scene texts of different scales. The feature vector fusion module is formed by concatenating multiple feature vector fusion blocks. Causal convolution is the core component of the feature vector fusion block. After training, it can predict the fused feature vector based on the previously input feature vector. This paper also uses a lightweight segmentation head, which can effectively process features with a small computational cost. The segmentation head contains two key modules, namely the feature pyramid enhancement module (FPEM) and the feature fusion module (FFM). FPEM is cascadable and has a low computational cost. It can be attached behind the backbone network to deepen its characteristics of different scales and make it more expressive. After that, FFM merges the features generated by FPEM of different depths into the final features for segmentation. Non-local PAN uses the predicted text area to describe the complete shape of the text instance, and predicts the core of the text to distinguish different text instances. The network also predicts the similarity vector of each text pixel to guide each pixel to the correct core. **Result** This method is compared with other methods on three scene text datasets, and it has outstanding performance in speed and accuracy. On the ICDAR 2015 dataset, the F value of this method is 1.5% higher than the best method, and the detection speed reaches 23.1FPS; on the CTW 1500 dataset, the F value of this method is 1.8% higher than the best method, and the detection speed has reached 71.8FPS; on the Total-Text dataset, the F value of this method is 0.8% higher than the best method, and the detection speed has reached 64.3FPS, which is far beyond other methods. In addition, we design parameter setting experiments to explore the best location for non-local module embedding. Experiments have proved that the effect of embedded the non-local module is better than non-embedding, indicating that non-local modules play an active role in the detection process. According to the detection accuracy, the effect of embedding non-local blocks into the second, the third, and the fourth layers of Resnet-18 is significant, while the effect of embedding the fifth layer is not obvious. Among them, embedding non-local blocks in the third layer has the best effect. We designed ablation experiments on the ICDAR 2015 dataset for the non-local module and the feature vector fusion module. The experimental results prove that the superiority of the non-local module does not come from deepening the network, but from its own structural characteristics. The feature vector fusion module also plays an active role in the scene text detection process. It combines feature maps of different scales to enhance the feature expression of scene texts with variable scales. **Conclusion** In this paper, an efficient text detection method for arbitrary shape scene is proposed, which takes into account accuracy and real-time. The experimental results show that the performance of our model is better than the previous methods, and our model is in the leading level in accuracy and speed.

Key words： object detection; scene text detection; neural network; non-local module; pixel aggregation; real-time detection; arbitrary shape

0 引 言

文本在现实生活中处处可见，物品上的文字表达了它们想传递的信息。对文本的检测能力则决定了对文本的识别和对场景的理解水平。随着计算机视觉和物联网等现代技术的高速发展，许多新兴的应用场景都需要提取图像中的文本信息，比如获取路牌中的指路信息为自动驾驶的汽车指引方向，门牌号识别实现无人送货等。近年来提出了一些检测场景文本的新方法，例如 TextSnake（Long 等，2018）达到较高的检测水平，但其后处理方法过于庞大复杂，因而检测速度较慢，这从根本上限制了它在现实中的应用。而 Gliding vertex（Xu 等，2020）等高效文本检测方法主要是以四边形边界框进行预测，这在检测弯曲的文本时会产生偏斜。

为了解决上述问题，本文提出了一种高效的任意形状场景文本检测器：非局部像素聚合网络(Non-local PAN)，可用于检测多方向、任意形状的场景文本，本方法可以在速度和性能之间取得良好的平衡。本方法使用特征金字塔增强模块和特征融合模块进行轻量级特征提取，为了弥补轻量级网络提取特征能力不足的缺陷，为骨干网络嵌入非局部模块以增强其提取特征的能力；此外，本文提出了特征向量融合模块，用于增强多尺度场景文本的特征表达，使其检测的准确性得到提高。如图 1 所示，与近年提出的其它方法相比，本方法的性能与速度均处于领先水平。

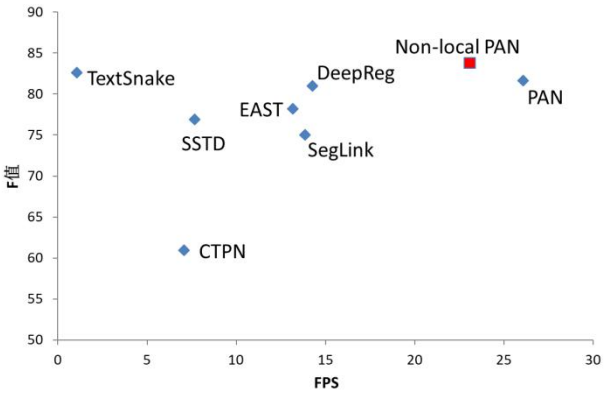

图 1　各模型检测速度和准确度的对比图

Fig.1　Comparison chart of detection speed and accuracy of each model

1 相关工作

1.1 实时场景文本检测

实时场景文本检测需要使用一种快速的方法来生成高质量的文本预测结果。EAST（Zhou 等，2017）直接使用 FCN（Long 等，2015）来预测得分图和相应的坐标，然后使用非极大值抑制得到输出结果。EAST 的整个流程非常简洁，因此可以做到实时性检测。 MCN（Liu 等，2018）将文本检测问题表达为基于图的聚类问题，并在不使用非极大值抑制的情况下生成边界框，使得 MCN 可以在 GPU 上完全并行化。但是，这些方法是专为四边形文本检测而设计的，对任意形状场景文本的预测结果非常不理想。

1.2 任意形状的场景文本检测

2017 年，CTW1500（Liu 等，2017）和 Total-Text（Chng 等，2017）等任意形状场景文本数据集的出现，使得对任意形状场景文本的研究变得火热起来。

为了检测面向或弯曲的场景文本，Lyu 等人（2018）提出了一种 Mask TextSpotter，它巧妙地细化了 Mask RCNN，利用字符级标签同时检测和识别字符和实例掩码。该方法显著提高了面向点定位或曲线场景文本的性能。然而，字符级标签的成本是极其昂贵的，所以该方法难以落实到实际应用中。Liao 等人（2019）对该方法进行改进，显著减轻了对字符级标签的依赖。该方法依赖于区域生成网络，在一定程度上限制了检测速度。

最近，Qin 等人（2019）提出使用 RoI 掩膜来聚焦弯曲的文本区域，但其结果很容易受到离群像素的影响。另外，分割分支增加了计算负担，拟合多边形过程也带来了额外的计算负担。Liu 等人（2019）提出了一种基于金字塔掩模的场景文本检测算法（PMTD）。该检测算法不再预测文本实例的二值掩膜，而是对每个像素进行回归操作，从而使得生成的文本实例掩膜具有更丰富的信息。

2 非局部像素聚合网络

2.1 整体架构

如图 2 所示，Non-local PAN 遵循基于分割的方法流程来检测场景文本实例。为了提高效率，骨干网络必须是轻量级网络。但是，轻量级骨干网络的表示能力通常较弱。因此，本文为骨干网络添加了非局部模块，用于增强其提取特征的能力。如图 3

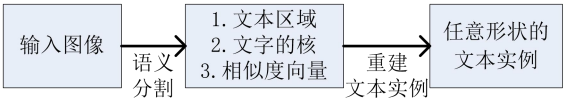

图 2　Non-local PAN 的总体流程

Fig.2　The overall process of non-local PAN

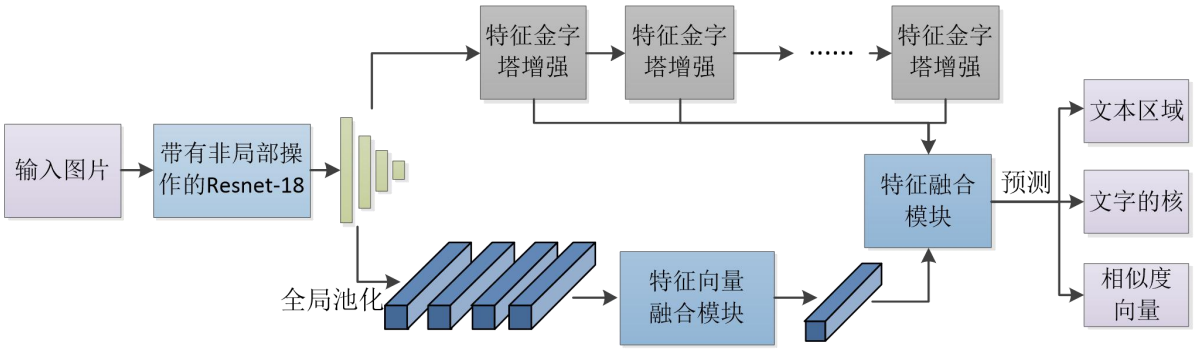

图 3　Non-local PAN 的整体架构

Fig.3　The overall architecture of Non-local PAN

所示，本方法采用 Resnet-18（He 等，2016）作为 Non-local PAN 的骨干网络，并将非局部模块嵌入其第三层的最后一个残差块之前。此外，本文还设计了特征向量融合模块用于融合不同层次的特征向量，以增强不同尺度的场景文字的特征表达。本文还使用了一种轻量的分割头，它在可以有效地以较小的计算成本对特征进行处理。该分割头包含两个关键模块，即功能金字塔增强模块（FPEM）和特征融合模块（FFM）。FPEM 是可级联的，并且计算成本较低，可以将其附着在骨干网络后面，以加深其不同尺度的特征，使其更具表现力。之后，FFM 将不同深度的 FPEM 产生的特征融合到最终的特征中进行分割。Non-local PAN 用预测出的文本区域来描述文本实例的完整形状，并预测文本的核以区分不同的文本实例。网络还预测每个文本像素的相似度矢量，用于指引每个像素聚合到正确的核中。

2.2 非局部模块

捕捉大范围内数据相互之间的依赖关系是一个很重要的问题。一般方法通常使用较大的卷积核来捕捉图片中较远距离的像素之间的关系。然而，传统的卷积神经网络只是在其时间或空间的很小的邻域内进行捕捉，却很难捕获到更远的位置的数据的依赖关系。

非局部网络(Non-local network)（Wang 等，2018）可以很好地捕捉到较远位置的像素点之间的依赖关系。定义在深度神经网络中的非局部操作如下：

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \#(1)$$

其中x表示输入信号，y_i表示输出信号，其尺度大小与x相同。$f(x_i, x_j)$用来计算 i 位置像素和所有可能关联的 j 位置像素之间的内在关系。在本文中，将$f(x_i, x_j)$实现为嵌入高斯核函数，具体算法如下：

$$f(x_i, x_j) = e^{\theta(x_i)^T \varphi(x_j)} \#(2)$$

其中，$\theta(x_i) = W_\theta x_i$和$\varphi(x_j) = W_\varphi x_j$是两个待学习的嵌入空间。该累加后的结果由因子$C(x)$归一化，算法如下：

$$C(x) = \sum_{\forall j} f(x_i, x_j) \#(3)$$

$g(x_j)$用于计算输入信号在 j 位置的特征值，算法如下：

$$g(x_j) = W_g x_j \#(4)$$

其中 W_g 是要学习的权重矩阵。在实现网络时，被实现为空间中的 1×1 卷积。

由以上公式，对非局部操作进行实现，非局部模块的结构如图 4 所示。

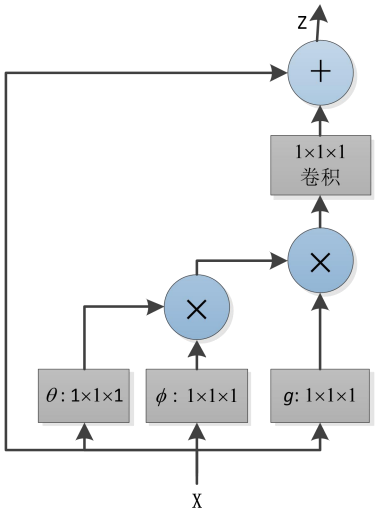

图 4　非局部模块的细节。其中，“×”表示矩阵乘法，“+”表示逐元素求和。

Fig.4　Details of non-local modules. "×" means matrix multiplication, "+" means element-wise summation.

将非局部模块嵌入骨干网络 Resnet-18 时，一般是插入到不同阶段的最后一个残差块之前，并且通过实验发现在第二层、第三次和第四层上嵌入非局部模块的效果好，而在第五层嵌入非局部模块的效果则不明显。在本方法中，非局部模块被嵌入到 Resnet-18 的第三层。嵌入非局部模块的 Resnet-18 第三层网络结构如图 5 所示。

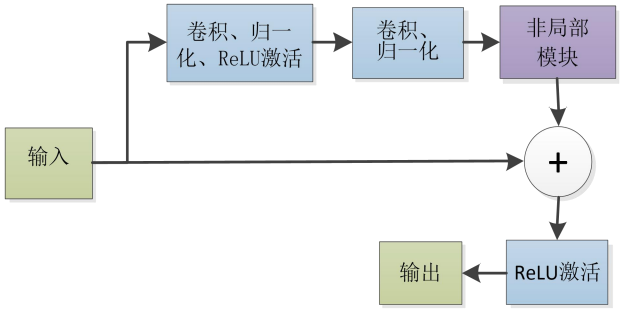

图 5　Resnet-18 第三层的网络结构图

Fig.5　Resnet-18 third layer network structure diagram

2.3 特征向量融合模块

为了融合不同尺度的特征图，增强尺度多变的场景文字的特征表达，本文提出了特征向量融合模块。特征向量融合模块由四个特征向量融合块串联而成。

因果卷积（Oord 等，2016）是特征向量融合块的核心成分，因果卷积首先被 WaveNet（Oord 等，2016）用于生成原始音频。在 WaveNet 中，因果卷积还进行了扩张操作，以实现用较少的网络层数覆盖较大的感受野。而在本方法中，输入的特征向量只有四个，无需使用扩张卷积。因果卷积的输出长度与其输入长度相同，并且当前输出仅取决于当前点及之前的输入信息。将不同层次的特征向量依次输入到特征向量融合模块中，因果卷积会将之前输入的特征向量信息用于本次特征向量的处理，通过训练，可以使最后一个特征向量对应的输出融合所有特征向量的信息。

特征向量融合块的结构如图 6 所示，其输出数据 $x^{(i+1)}$的长度与其输入数据 $x^{(i)}$相同。特征向量融合块公式化如下：

$$temp = Causal\left(x^{(i)}\right) \odot \sigma\left(Causal\left(x^{(i)}\right)\right) \#(5)$$

$$x^{(i+1)} = temp + Conv\left(x^{(i)}\right) \#(6)$$

其中，$Causal(\cdot)$是因果卷积函数，$\sigma(\cdot)$是 Sigmoid 函数，$x^{(i)}$是第 i 个 Feature vector Fusion block 的输入。⊙代表逐元素相乘。

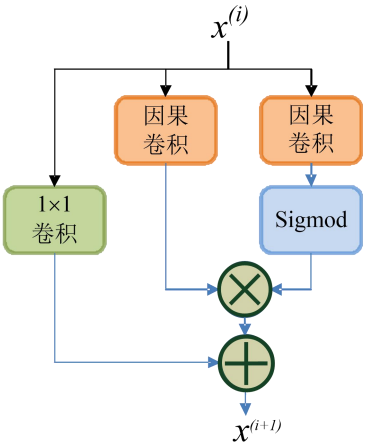

图 6　特征向量融合块的结构细节

Fig.6　Structural details of the feature vector fusion block.

2.4 特征金字塔增强和特征融合

FPEM 能够通过融合低级和高级信息来增强不同尺度的特征。FPEM 是可级联的模块，随着级联层数的增加，不同尺度的特征图会得到更充分地融合，特征图的感受野也随之增大。此外，因为 FPEM 是通过可分解卷积构建的，它的计算开销非常小，仅为 FPN（Lin 等，2017）的 1/5 左右。

图 7 所示的 U 形模块，它由扩大尺度增强和缩小尺度增强两个阶段组成。扩大尺度增强作用在输入的特征图上，它会分别对边长为 32、16、8、4 个像素的特征图进行迭代增强。在缩小尺度增强阶段，输入是由扩大尺寸增强生成的特征金字塔，增强过程与扩大尺寸增强相反。缩小尺寸增强的输出即为 FPEM 的最终输出。

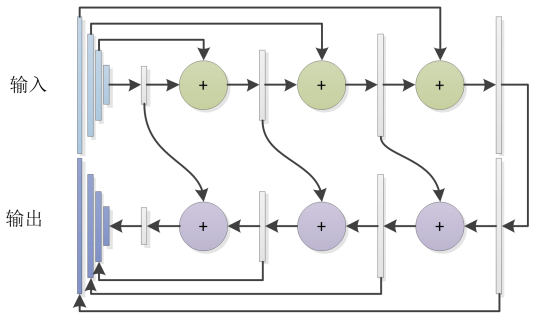

图 7　特征金字塔增强模块的结构细节

Fig.7　Structural details of the feature pyramid enhancement module

特征融合模块用于融合不同深度的特征金字塔以及特征向量融合模块输出的特征向量。如图 8 所示，通过逐元素加法组合相应比例的特征图，并与将特征向量进行上采样得到的特征图一起进行上采

样操作并级联为仅具有 5×128 通道的最终特征图。

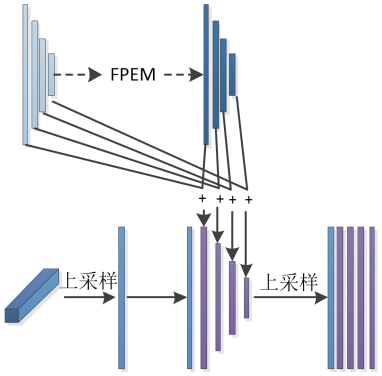

图 8　特征融合模块的结构细节

Fig.8　Structural details of feature fusion module

2.5 像素聚合

文本区域保持了文本实例的完整形状，但是紧密放置的文本实例的文本区域通常是重叠的，所以需要使用核区分文本实例（见图 9 (a)）。文字的核不并是完整的文本实例，如果要重建完整的文本实例（见图 9 (b)），需要将文本区域中的像素合并到核中。本文采用了一种可学习的算法，即像素聚合（Pixel Aggregation），以指导文本像素被分类到正确的核。

在像素聚合中，借鉴了聚类的思想，以从预测的核中重建完整的文本实例。将文本实例的核视为聚类中心，文本像素是要聚类的样本。为了将文本像素聚合到相应的内核，同一文本实例的文本像素与核之间的距离应较小。

在训练阶段，使用聚集损失L_{agg}和判别损失L_{dis}来评判像素聚合的效果，并用以训练。

在测试阶段，使用预测的相似性矢量将文本区域中的像素引导到相应的核。像素聚合的步骤如下：

1) 在核的分段结果中找到连接的组件，每个连接的组件都是一个单独的核。

2) 对于每个内核，有条件地将其相邻文本像素合并到预测文本区域中，使它们的相似度向量的欧几里德距离小于某个阈值。

3) 重复步骤 2)，直到没有符合条件的邻近文本像素。

2.6 损失函数

本方法的损失函数可以表示为：

$$L = L_{tex} + \alpha L_{ker} + \beta\left(L_{agg} + L_{dis}\right)\#(7)$$

其中L_{tex}是文本区域的损失函数，L_{ker}是核的损失函数。L_{agg}是衡量文本实例中的像素和其对应核的损失函数，L_{dis}是分辨不同文本实例的核的一个损失函数。α和β被用来平衡L_{tex}、L_{ker}、L_{agg}和L_{dis}的重要程度，L_{tex}是模型的最终结果，重要程度最高，L_{ker}用于评价核的分割结果，重要程度仅次于L_{tex}，而L_{agg}和L_{dis}重要程度较低。按照其重要程度，α和 β 分别设为 0.5 和 0.25。

(a)文本的核　　(b) 文本实例

图 9　对文本的核进行文本实例重建

Fig.9　Reconstruct the text instance of the core of the text

((a) The kernel of text; (b) The instance of text)

考虑到文本和非文本像素在数量上非常不均衡，可以采用 dice loss（Milletari 等，2016）来监督文本区域的分割结果P_{tex}与核的分割结果P_{ker}，因此L_{tex}和L_{ker}计算方法如下：

$$L_{tex} = 1 - \frac{2\sum_i P_{tex}(i)G_{tex}(i)}{\sum_i P_{tex}(i)^2 + \sum_i G_{tex}(i)^2} \#(8)$$

$$L_{ker} = 1 - \frac{2\sum_i P_{ker}(i)G_{ker}(i)}{\sum_i P_{ker}(i)^2 + \sum_i G_{ker}(i)^2} \#(9)$$

其中$P_{tex}(i)$和$G_{tex}(i)$分别指分割结果的第i个结果以及有标注的文本区域的准确性；类似的$P_{ker}(i)$和$G_{ker}(i)$分别指预测结果的第i个像素值以及核的准确性。

L_{agg}的作用是保证同一文本实例的核和文本实例内其他像素点之间的距离在一定范围内，其公式如下：

$$L_{agg} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|T_i|}\sum_{p\in T_i}\ln\left(D(p,K_i)+1\right) \#(10)$$

其中，N是图像中文本实例的数量，T_i表示第i个文本实例，K_i是该文本实例应的核。$D(p,K_i)$代表文本实例T_i内的像素p到相应的核K_i的距离。

L_{dis}是用于不同文本实例的核的损失，其作用是保证任意两个核之间的距离不至于太小，其计算公式如下：

$$L_{dis} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\ln\left(D\left(K_i,K_j\right)+1\right) \#(11)$$

其中，$D\left(K_i,K_j\right)$代表核K_i与核K_{j}之间的距离。该公式实际上是对于每一个文本实例的核，分别计算与其他核的距离，然后进行累加。

3 实验与分析

本论文提出的方法基于 Pytorch 框架实现，在实验时使用一块 GPU 显卡（Nvidia 1080Ti）进行训练和测试。本方法采用随机梯度下降算法进行优化，训练批大小为 16，初始学习率设为 0.001，迭代训练 500 次，然后将学习率设为 0.00001，迭代训练 100 次。

3.1 数据集

ICDAR 2015 数据集是 ICDAR 发布的场景文本检测数据集。ICDAR 2015 由 1000 张训练集和 500 张测试集组成，该数据集是以四个顶点的边界框的形式标注的。

CTW1500 是一个具有挑战性的曲线文本数据集。它由 1000 张训练图像和 500 张测试图像组成。该数据集使用 14 个点标注的十四边形来表示曲线文本实例。

Total-Text 是一个用于曲线文本检测的数据集。该数据集包括水平文本实例、多方向文本实例和曲线文本实例。它由 1255 幅训练图像和 300 幅测试图像组成。

本方法在各数据集上的检测结果如图 10 所示。

3.2 参数设置实验

为探究非局部模块嵌入位置的不同对结果的影响，本文在 ICDAR 2015 上对非局部模块的作用和嵌入的位置设计了对比实验，实验结果如表 1 所示。

表 1 非局部模块嵌入 Resnet-18 的不同位置时在 ICDAR 2015 上的检测结果对比

Table 1 Comparison of results of ICDAR 2015 when the non-local block is embedded in different places in Resnet-18

嵌入位置	召回率	准确率	F 值	检测速度
无	83.1	80.2	81.6	**26.1**
res2	84.2	81.1	82.6	25.2
res3	**84.5**	**81.4**	**82.9**	24.5
res4	84.1	81.5	82.7	23.1
res5	83.3	80.5	81.8	20.3

注：res2、res3、res4 和 res5 分别代表 Resnet-18 的第二层、第三层、第四层和第五层。F 值用于综合考虑召回率和准确率。检测速度采用 FPS 衡量，即平均每秒处理图片的数量。加粗字体为最优值。

由表 1 可知，嵌入非局部模块之后的效果要优于未嵌入模块的效果，说明非局部模块在检测过程中发挥了积极作用。按照检测精准程度来看，将非局部块嵌入 Resnet-18 的第二层、第三层和第四层的效果显著，而嵌入第五层的效果不明显。其中，将非局部块嵌入第三层的效果最好。按照检测时间来看，检测时间会随着嵌入位置的后移而延长，这是因为 Resnet-18 越往后其规模越大，对其进行非局部操作的复杂度越高。基于此实验结果，在本方法中，非局部模块被嵌入至 Resnet-18 的第三层的最后一个残差块之前。

(a) ICDAR 2015　　(b) CTW 1500　　(c) Total-Text

图 10: 本方法在多个数据集上的检测结果

Fig.10 The detection results of the proposed method on multiple datasets

((a) ICDAR 2015；(b) CTW 1500；(c) Total-Text)

3.3 模型对比实验

本方法与近年出现的其它方法在多个数据集上进行了对比。表 2 展示了多种方法在 ICDAR 2015 上的性能对比。由实验结果可知，PAN 已经在准确度上达到较高水平，在速度上更是远远超过其它方法。而本文提出的 Non-local PAN 的 F 值达到了 83.8，已经超越了 PAN，与其他最新方法相比，本方法可以实现较高的性能并以更快的速度（23.1 FPS）运行。

LOMO 在以上几种模型中准确率最高，但速度也最慢。LOMO 的网络复杂，后处理方法繁琐，远不如 Non-local PAN 的轻量分割头和像素聚合方法简单。由于非局部模块、轻量分割头和像素聚合等方法的共同作用，Non-local PAN 在保证速度远远领先 LOMO 的基础上，还能在准确率上接近 LOMO 的水平。

表 2 不同模型在 ICDAR 2015 数据集上的结果对比

Table 2 Comparison of the results of different models on the ICDAR 2015 dataset

方法	召回率	准确率	F 值	速度
EAST（Zhou 等，2017）	73.5	83.6	78.2	13.2
DeepReg（He 等，2017）	80.0	82.0	81.0	14.3
SegLink（Shi 等，2017）	76.8	73.1	75.0	13.9
SSTD（He 等，2017）	73.9	80.2	76.9	7.7
TextSnake（Long 等，2018）	84.9	80.4	82.6	1.1
ATRR（Wang 等，2019）	83.3	90.4	86.6	15.4
CRAFT（Baek 等，2019）	**84.3**	89.8	86.9	12.5
LOMO（Zhang 等，2019）	83.5	**91.3**	**87.2**	3.4
PAN（Wang 等，2019）	81.9	84.0	82.9	**26.1**
本方法	82.7	85.1	83.8	23.1

注：加粗字体为最优值。

本方法与其他方法在弯曲文本数据集 CTW 1500 和 Total-Text 上的性能对比分别如表 3 和表 4 所示。

表 3 不同模型在 CTW 1500 数据集上的结果对比

Table 3 Comparison of the results of different models on the CTW 1500 dataset

方法	召回率	准确率	F 值	速度
EAST（Zhou 等，2017）	49.1	78.8	60.4	21.2
SegLink（Shi 等，2017）	40.0	42.3	40.8	10.7
SSTD（He 等，2017）	73.9	80.2	76.9	7.7
TextSnake（Long 等，2018）	67.9	85.3	75.6	5.6
ATRR（Wang 等，2019）	80.2	80.1	80.1	22.5
CRAFT（Baek 等，2019）	**81.1**	86.0	**83.5**	19.7
LOMO（Zhang 等，2019）	69.6	**89.2**	78.4	4.4
PAN（Wang 等，2019）	77.4	82.7	79.9	**84.2**
本方法	78.9	83.8	81.3	71.8

注：加粗字体为最优值。

表 4 不同模型在 Total-Text 数据集上的结果对比

Table 4 Comparison of the results of different models on the Total-Text dataset

方法	召回率	准确率	F 值	速度
EAST（Zhou 等，2017）	36.2	50.0	42.0	19.8
SegLink（Shi 等，2017）	23.8	30.3	26.7	9.1
TextSnake（Long 等，2018）	74.5	82.7	78.4	4.7
ATRR（Wang 等，2019）	76.2	80.9	78.5	25.4
CRAFT（Baek 等，2019）	79.9	87.6	83.6	21.6
LOMO（Zhang 等，2019）	75.7	88.6	81.6	4.4
PAN（Wang 等，2019）	81.0	89.3	85.0	**39.6**
本方法	**82.9**	**89.9**	**86.3**	34.3

注：加粗字体为最优值。

如表 3 和表 4 所示，本方法在 CTW 1500 和 Total-text 数据集上都达到的非常高的水平，F 值分别达到了 81.3 和 86.3。在检测速度上，PAN 的检测速度最快，本方法次之，但均远远超过其他检测方法。

3.4 消融研究

为验证本方法中的不同模块在检测过程中发挥了重要作用，本文针对不同模块设计了消融实验，所有的消融实验均在 ICDAR 2015 数据集上进行。

针对非局部模块，本文设计了三组实验：第一组实验不对骨干网络嵌入任何模块，第二组实验嵌入普通卷积模块，第三组实验嵌入非局部模块。第一、二组实验与第三组实验形成对照，分别探究删除、替换非局部模块对结果的影响。若第三组实验效果优于第一、二组的实验效果，则说明非局部模块在文本检测流程中发挥了不可或缺的作用。

为保持控制单一变量原则，普通卷积操作嵌入的位置与非局部模块嵌入位置相同，均嵌入到骨干网络 Resnet-18 第三层的最后一个残差块之前。

本实验的运行环境以及参数设置均与原实验相同。不同嵌入模块的实验结果对比如表 5 所示。由实验结果可知：在准确度上，嵌入普通卷积模块的效果不如嵌入非局部模块的效果，甚至比未嵌入任何模块的效果还要差一些。而在检测速度上，嵌入普通卷积模块要比嵌入非局部模块快一些，但比未嵌入模块要慢。这说明非局部模块不是普通卷积可以代替的，非局部模块表现出的优越性并非是来自于加深了网络，而是来源于它自身的结构特性。另一方面，Resnet-18 中嵌入普通卷积之后，在一定程度上破坏了残差网络的结构，使得它的检测效果甚至不如未嵌入任何模块。在模块的算法复杂度上，普通卷积操作比非局部操作略简单，所以其检测速度比嵌入非局部模块时略快。

表 5 不同嵌入模块效果对比

Table 5 Comparison of effects of different embedded modules

嵌入模块	召回率	准确率	F 值	检测速度
无	80.2	83.1	81.6	**26.1**
普通卷积模块	79.5	82.2	80.8	25.2
非局部模块	**82.7**	**85.1**	**83.8**	23.1

注：加粗字体为最优值。

此外，本文设计了消融实验验证了特征向量融合模块的可行性。表 6 展示了有无特征向量融合模块对最终检测结果的影响。由实验结果可知，特征向量融合模块在场景文字检测过程中发挥了正向的积极作用，它融合了不同尺度的特征图，增强尺度多变的场景文字的特征表达。

表 6 有无特征向量融合模块网络检测结果对比

Table 5 Comparison of results with and without feature vector fusion module

特征向量融合模块	召回率	准确率	F 值	检测速度
无	81.4	84.5	82.9	**24.5**
有	**82.7**	**85.1**	**83.8**	23.1

注：加粗字体为最优值。

4 结 论

本文提出了非局部像素聚合网络，该网络能够实现对任意形状场景文本的实时性检测。针对文本字符的特征，引入了非局部模块，将其嵌入到像素聚合网络的骨干网络中，使其能够捕捉到像素之间的内在关系，大大增强了其提取特征的能力。此外，本文设计了一个特征向量融合模块，用于融合不同尺度的特征图，增强尺度多变的场景文字的特征表达。本方法基于轻量级网络构建，并通过多个模块进行特征增强，在检测速度和检测精度上都达到了较高水平。

本论文提出的基于卷积神经网络的场景文本检测模型，在该方面已取得较好的效果，但仍有一些可改进的地方。比如：本文提出的方法用到了非局部操作，非局部操作是一次全局卷积操作，即卷积核的大小与特征图的大小相等，当特征图较大时，非局部操作的计算量也会非常大。如何简化模型结构还能保持捕捉远距离像素之间的关系，是该模型未来需要优化的方向之一。

参考文献(References)

Xu Y C, Fu M T, Wang Q M, Wang Y K, Chen K, Xia G S and Bai X. 2020. Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, PP(99): 1-1[DOI: 10.1109/TPAMI.2020.2974745]

Lyu P Y, Liao M H, Yao C, Wu W H and Bai X. 2018. Mask TextSpotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes// Proceedings of the European conference on computer vision. Munich: Springer: 71-88[DOI: 10.1007/978-3-030-01264-95]

Liao M H, Lyu P Y, He M H, Yao C, Wu W H and Bai X. 2019. Mask TextSpotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes. IEEE Trans, 43(2): 532-548[DOI: 10.1109/TPAMI.2019.2937086]

Qin S Y, Alessandro B, Michalis R, Yasuhisa F and Xiao Y. 2019. Towards Unconstrained End-to-End Text Spotting// Proceedings of the International Conference on Computer Vision. Seoul: IEEE: 4703-4713[DOI: 10.1109/ICCV.2019.00480]

Baek Y, Lee B, Han D, Yun S and Lee H. 2019. Character Region Awareness for Text Detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE: 9365-9374[DOI: 10.1109/CVPR.2019.00959]

Zhang C Q, Liang B R, Huang Z M, En M Y, Han J Y, Ding E R, and Ding X H. 2019. Look more than once: An accurate detector for text of arbitrary shapes// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE: 10552-10561[DOI: 10.1109/CVPR.2019.01080]

Wang X B, Jiang Y Y, Luo Z B, Liu C L, Choi H, and Kim S. 2019. Arbitrary shape scene text detection with adaptive text region representation// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE: 6449-6458 [DOI: 10.1109/CVPR.2019.00661]

Zhou X Y, Yao C, Wen H, Wang Y Z, Zhou S C, He W R and Liang J J. 2017. East: an efficient and accurate scene text detector// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 2642-2651[DOI: 10.1109/CVPR.2017.283]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE: 3431-3440[DOI: 10.1109/CVPR.2015.7298965]

Liu Z C, Lin G S, Yang S, Feng J S, Lin W S and Goh W L. 2018. Learning markov clustering networks for scene text detection// Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 6936-6944[DOI:10.1109/CVPR.2018.00725]

Liu Y L, Jin L W, Zhang S T, Luo C J and Zhang S. 2017. Detecting curve text in the wild: new dataset and new solution[EB/OL]. [2020-11-07]. http://arxiv.org/abs/1712.02170

Chng C K and Chan C S. 2017. Total text: a comprehensive dataset for scene text detection and recognition// Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. Kyoto: IEEE: 935-942[DOI: 10.1109/ICDAR.2017.157]

Li X, Wang W H, Hou W B, Liu R Z, Lu T and Yang J. 2018. Shape robust text detection with progressive scale expansion network[EB/OL]. [2020-11-07]. http://arxiv.org/abs/1806.02559

Yang Q P, Cheng M L, Zhou W M, Chen Y, Qiu M H and Lin W. 2018. Inceptext: a new inception-text module with deformable psroi pooling for multi-oriented scene text detection[EB/OL]. [2020-11-07]. http://arxiv.org/abs/1805.01167.

Liu J C, Liu X B, Sheng J, Liang D, Li X and Liu Q J. 2019. Pyramid mask text detector[EB/OL]. [2020-11-07]. http://arxiv.org/abs/1903.11800.

He K M, Zhang X Y, Ren S Q, Sun J. 2016. Deep residual learning for image recognition// Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE: 770-778[DOI: 10.1109/CVPR.2016.90]

Wang X L, Girshick R B, Gupta A and He K M. 2018. Non-local neural networks// Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE: 7794-7803[DOI: 10.1109/CVPR.2018.00813]

Oord A V D, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A W and Kavukcuoglu K. 2016. Wavenet: a generative model for raw audio[EB/OL]. [2020-11-07]. http://arxiv.org/abs/1609.03499

Lin T Y, Dollár P, Girshick R B, He K M, Hariharan B and Belongie S J. 2017. Feature pyramid networks for object detection//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE: 936-944[DOI: 10.1109/CVPR.2017.106]

Milletari F, Navab N and Ahmadi S A. 2016. V-net: fully convolutional neural networks for volumetric medical image segmentation// Proceedings of the 2016 Fourth International Conference on 3D Vision. Stanford: IEEE: 565-571[DOI: 10.1109/3DV.2016.79]

Tian Z, Huang W L, He T, He P and Qiao Y. 2016. Detecting text in natural image with connectionist text proposal network// Proceedings of the 14th European conference on computer vision. Amsterdam: Springer: 56-72[DOI: 10.1007/978-3-319-46484-8_4]

He W H, Zhang X Y, Yin F and Liu C L. 2017. Deep direct regression for multi-oriented scene text detection// Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE: 745-753[DOI: 10.1109/ICCV.2017.87]

Shi B G, Bai X and Serge J. Belongie. 2017. Detecting oriented text in natural images by linking segments// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 3482-3490[DOI: 10.1109/CVPR.2017.371]

He P, Huang W L, He T, Zhu Q L, Qiao Y and Li X L. 2017. Single shot text detector with regional attention// Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE: 3066-3074[DOI: 10.1109/ICCV.2017.331]

Long S B, Ruan J Q, Zhang W J, He X, Wu W H and Yao C. 2018. Textsnake: a flexible representation for detecting text of arbitrary shapes// Proceedings of the European conference on computer vision. Munich: Springer: 19-35[DOI: 10.1007/978-3-030-01216-8_2]

Wang W H, Xie E Z, Song X G, Zang Y H, Wang W J, Lu T, Yu G and Shen C H. 2019. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network// Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE: 8439-8448[DOI: 10.1109/ICCV.2019.00853]

作者简介

师广琛，1998 年生，男，硕士研究生，研究方向为计算机视觉。E-mail: shi.guangchen@foxmail.com

巫义锐，1989 年生，男，副教授，主要研究方向为计算机视觉，模式识别与智慧水利。E-mail: wuyirui@hhu.edu.cn