基于分散注意力与路径增强特征金字塔的文本检测

程琦，王国栋*，赵毅
青岛大学计算机科学技术学院，山东 青岛 266071

摘要 为了进一步提升基于卷积神经网络的文本检测器的检测精度，首先，用具有分散注意力机制的特征提取网络替代原始算法的主干网络，如残差网络，以促进通道间的信息交互，最大化地激活文本特征。其次，在原始特征金字塔网络的基础上增加自底向上的路径，以减少文本特征信息的损耗。实验结果表明，本算法在 CTW1500、Total-Text 曲线数据集上的平均精度分别为 78.7%、79.0%，在多方向数据集和多语言数据集的平均精度分别为 82.7%、79.3%，相比其他算法均有一定的提升。

关键词 图像处理；卷积神经网络；主干网络；分散注意力机制；特征金字塔网络
中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP57.241023

Text Detection Based on Split-Attention and Path Enhancement Feature Pyramid

Cheng Qi, Wang Guodong*, Zhao Yi
College of Computer Science and Technology, Qingdao University, Qingdao, Shandong 266071, *China*

Abstract In order to further improve the detection accuracy of the text detector based on convolutional neural networks, first, feature extraction network with split-attention mechanism is used to replace the backbone network of the original algorithm, such as residual network, to promote information exchange between channels and maximize the activation of text features. Second, based on the original feature pyramid network, a bottom-up path is added to reduce the loss of text feature information. Experimental results show that the average accuracy of the algorithm is 78.7% and 79.0% on CTW1500 and Total-Text curve data sets, and 82.7% and 79.3% in multi-directional and multi-language data sets, respectively, which is better than other algorithms.

Key words image processing; convolutional neural network; backbone network; split-attention mechanism; feature pyramid network
OCIS codes 100.2000; 100.4996; 100.3008

1 引 言

近年来，自然场景文本检测作为计算机视觉领域的基础性任务，在自动驾驶、场景理解和文本识别等领域得到了广泛应用。随着卷积神经网络(CNN)[1-7]的飞速发展，人们提出了多种优秀的算法[8-11]。Liao 等[12]通过修改锚框的尺寸和卷积核的形状，解决了不同长宽比的文本检测问题。Liao 等[13]用四边形边界框回归多方向文本实例。He 等[14]引入注意力机制，可粗略识别文字区域。Liao 等[15]提取旋转不变特征并将其用于文本分类，同时选择敏感特征用于文本回归，对多方向和长文本具有鲁棒性。Zhang 等[16]使用语义分割算法提取文本区域。Yao 等[17]将文本块定义为三部分，并用全卷积神经网络(FCNN)[18]分别预测相应部分的热图。Lyu 等[19]使用角点定位表示不规则边界框。Deng 等[20]通过预测像素间的连接，区分相邻文本。Xie 等[21]利用实例分割算法和上下文信息检测曲线

收稿日期：2020-06-09；**修回日期**：2020-06-18；**录用日期**：2020-06-28
基金项目：山东省自然科学基金(ZR2019MF050)

* **E-mail**：doctorwgd@gmail.com

文本。Wang 等[22]提出逐尺度扩张算法，通过设置多尺寸核重建文本实例。一方面，基于分割的算法通常用分类网络提取文本特征，如残差网络(ResNet)[1]、高效网络(EfficientNet)[23]，而分类网络没有注意力机制，直接应用于文本检测任务时，无法有效激活文本特征。因此，在文本检测算法中简单地嵌入分类网络不能获得最优解。另一方面，在文本检测器中，高层对全局特征(大尺度文本行)响应强烈，低层则更容易被局部特征(小尺度文本行)激活，因此，多尺度特征对于文本检测具有重要意义。

传统算法通常用特征金字塔网络(FPN)[24]提取多尺度特征信息，但 FPN 中浅层到高层的路径过长阻碍了小尺度文本特征信号的流动。为了解决该问题，本文提出了一种基于分割方法的检测器。首先，为了增强通道间的特征交互，使检测器能最大化地激活文本特征响应，提出了用具有分散注意力机制的 ResNeSt[25]作为主干网络以提取文本特征。与传统的分类网络相比，ResNeSt 可以明显增强对文本特征的激活响应。其次，为了使低层的小尺度文本特征畅通无阻地流向高层，提出了一种路径增强特征金字塔网络(PEFPN)，通过建立少于 10 层的极短路径，将原始特征图引入特征融合阶段，减少了小尺度文本特征流向高层的损耗；同时，在 PEFPN 中用深度可分离卷积(Depthwise separable convolution)[26]代替传统卷积，以获得更快的推理速度。

2 算法分析

2.1 基于分割与基于边界框的算法

基于深度学习的文本检测算法可分为基于边界框和基于分割的算法。基于边界框的算法将文本视为普通目标，通过 CNN 直接预测文本的边界框；基于分割的算法将文本检测视为语义分割问题，逐像素地预测文本区域，将大于设定阈值的像素作为文本区域，将小于设定阈值的像素当作背景区域。基于边界框的算法通常需要预先定义边界框的尺寸，但文本行长宽比例的变化较大，预定义边界框很难完全覆盖各种极端长宽比例的文本行。而基于分割的算法对每个像素进行预测分类，摆脱了预定义边界框的束缚。自然场景中的文本行具有任意方向、任意形状的特点，如艺术字体，但预先定义的边界框通常是水平和垂直方向的矩形框，无法拟合文本的不同形状。而基于分割的算法通过逐像素地预测文本区域，将大于设定阈值的同一文本像素点进行连通，可表示任意形状的文本行。两种算法的检测效果如图 1 所示，可以看出，基于边界框的算法很难检测形状复杂的文本，而基于分割的算法能克服文本检测中文本行长宽比例多样、形状任意的不利因素，因此实验选用分割算法作为基础算法。

图 1 两种算法的检测结果。(a)基于边界框的算法；(b)基于分割的算法

Fig. 1 Detection results of the two algorithms. (a) Algorithm based on bounding box; (b) algorithm based on segmentation

2.2 注意力机制

注意力机制源于对人类视觉的研究，在认知科学中，人们通常会选择性地关注所有信息中的重要部分，忽略其他不重要的信息。随着深度学习的发展，注意力机制被引入计算机视觉任务以进行视觉信息处理。注意力大致可分为强注意力(hard attention)和软注意力(soft attention)两种，强注意力关注图像中每个像素点的反馈，是一个随机、动态的预测过程，且强注意力不可微，只能通过增强学习来训练；软注意力更关注区域或通道，是确定性、可微的注意力，可通过计算梯度反向求导更新注意力权重，即软注意力经过学习后可直接由 CNN 生成，因此在视觉任务中得到广泛应用。软注意力通常适用于编码(encoder)-解码(decoder)架构的 CNN 中，在编码阶段需要给不同通道上的信息进行加权，从而在解码阶段过滤掉相关度较低的特征信息。

对于文本检测任务，文本行通常处于背景复杂、

遮挡严重的图像中。如果在编码阶段文本特征信息不能获得激活响应，则解码时文本特征很容易被当作无用信息过滤掉。传统的文本检测算法只是简单的用分类网络作为主干网络以编码特征，但分类网络不具备注意力机制，因此，引入具有分散注意力机制的特征提取网络 ResNeSt，以增强文本像素的激活响应；同时，该网络在各通道内进行通道再分散，根据通道间的信息交互，最大化地激活文本像素。

2.3 传统特征金子塔网络

在检测任务中，小尺度目标由于缺乏足够的像素信息，经过多层卷积后，其像素信息容易丢失。FPN 利用主干网络不同阶段输出的不同分辨率特征图，构建了一条自上而下的特征融合路径，将低分辨率特征图上采样到与下一层特征图相同的尺度后进行特征融合，递归该操作直到融合至最大尺度的特征图，从而充分融合低分辨率语义信息较强的特征图和高分辨率空间信息丰富的特征图。但高分辨率特征图中的空间信息到达传统 FPN 的最高层需要经过几十层甚至上百层网络，从而丢失大量的空间信息。

对于文本检测任务而言，自然场景中有很多小尺度文本行，在一张图像中占据的区域较小。使用 FPN 时，小尺度文本的特征信息容易丢失在通向特征金字塔最顶层的路径中，导致文本检测器无法有效检测小尺度文本行，从而降低检测精度。因此，在传统 FPN 的基础上，构建了一条自下而上的路径，使高分率特征图通过自下而上的路径经过几层网络就能到达 FPN 的最顶层，最大程度地保留小尺度文本的特征信息。

3 算法实现

3.1 总体结构

本算法的总体结构如图 2 所示，具体步骤如下。

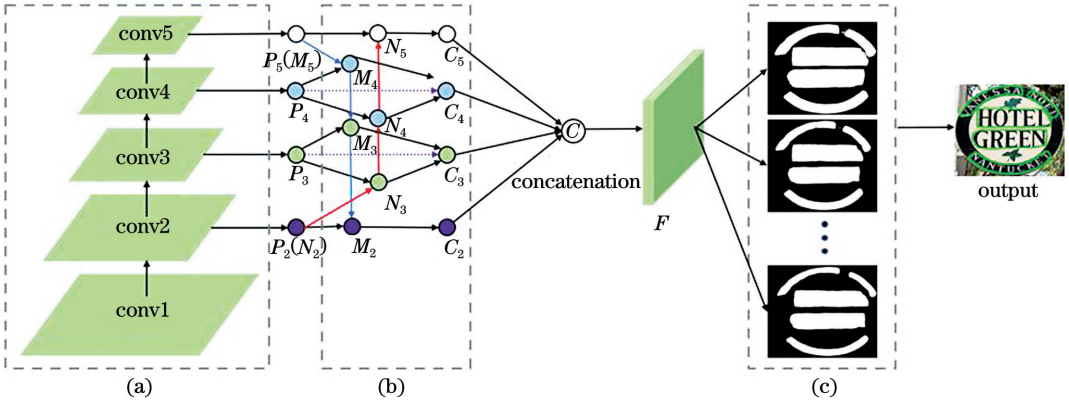

图 2 网络的总体结构。(a)主干网络；(b)PEFPN；(c)后处理算法

Fig. 2 Overall structure of the network. (a) Backbone; (b) PEFPN; (c) post processing algorithm

1) 输入图像被送入图 2(a)中的主干网络提取特征，实验分别用 50 层和 101 层的 ResNeSt 作为主干网络，前者的速度更快，后者的精度更高。两种网络的组件完全相同，都由卷积层(conv)、最大池化层(Max pooling layer)、平均池化层(Average pooling layer)、全连接层(Fully connection layer)和归一化指数函数(Softmax)构成。其中，conv1 的卷积核为 7×7，步长为 2；最大池化层的卷积核为 3×3，步长为 2；conv2～conv5 包含相同的瓶颈结构块(bottleneck block)，即两个 1×1 卷积和一个 3×3 分散注意力卷积，不同的是 conv4 中瓶颈结构块的堆叠次数，50 层网络为 6，101 层网络为 23。

2) 图 2(b)中的 PEFPN 以主干网络中 conv2～conv5 的特征图 $\{P_2,P_3,P_4,P_5\}$ 为输入，首先，将所有卷积层的特征图对齐到同一维度，即 $\{P_2,P_3,P_4,P_5\}$ 分别经过一个卷积核为 1×1、输出维度为 256 的卷积层。然后，特征图在 PEFPN 中通过自上而下和自下而上的路径进行融合，两个路径融合后的特征图可分别表示为 $\{M_2,M_3,M_4,M_5\}$ 和 $\{N_2,N_3,N_4,N_5\}$。在自上而下的路径中，M_i 由上一层特征图 M_{i+1} 执行双线性插值上采样操作并与 P_i 逐像素相加，再经过卷积核为 3×3 的平滑卷积层得到；在自下而上的路径中，N_i 由下一层特征图 N_{i-1} 经过卷积核为 3×3，步长为 2 的卷积层进行下采样操作并与 P_i 逐像素相加，再经过卷积核为 3×3 的平滑卷积层得到。最后，M_i 和 N_i 与对应的 P_i 通过逐像素相加操作进行融合，获得融合特征图 $\{C_2,C_3,C_4,C_5\}$。

3) $\{C_2,C_3,C_4,C_5\}$首先被上采样至同一尺寸后,进行维度连接(concatenation)操作;然后使用Sigmoid函数将特征图归一化到0~1范围内,以获取分割图F;最后用特定阈值对分割图进行二值化处理,其中,背景像素为0,文本像素为1。

4) 使用逐尺度扩张算法(PSEA)[22]处理二值图,以区分二值图中不同的文本行,得到最终的分割结果。

3.2 路径增强特征金字塔

为了解决传统FPN中小尺度文本特征信息的丢失问题,进一步增强网络的定位能力,构建了一条极短的自下而上路径(少于10层),以减少低层文本特征信息在特征金字塔进行融合时的损失,并将该模块命名为PEFPN。PEFPN的主干网络用conv2~conv5的特征图$\{P_2,P_3,P_4,P_5\}$作为输入,并将所有阶段的输入对齐到同一维度。与传统FPN不同,为了提高检测效率,PEFPN将输入维度减少至64维,多尺度特征融合的具体步骤如下。

1) 在自上而下的路径中,执行与FPN相同的操作,可表示为

$$M_i = X_{\text{conv}}\left[X_{\text{up}\times 2}(P_{i+1}) + M_{i+1}\right], \tag{1}$$

式中,M_i为自上而下的路径中第i层的融合特征图,i的值为$\{2,3,4\}$,M_5即P_5,X_{conv}为卷积函数,$X_{\text{up}\times 2}$为2倍上采样。为了进一步提高效率,在conv中用深度可分离卷积代替普通的3×3卷积,自上而下的路径结构如图3(a)所示。

2) 构建一条反向路径,使低层特征流向高层,可表示为

$$M_i = X_{\text{conv}}\left[X_{\text{down}\times 2}(P_{i-1}) + N_{i-1}\right], \tag{2}$$

式中,N_i为自下而上的路径中第i层的融合特征图,i的值为$\{3,4,5\}$,N_2即P_2,$X_{\text{down}\times 2}$和X_{conv}操作均使用深度可分离卷积实现,卷积步长分别为2和1。自下而上的路径结构如图3(b)所示。

3) 经过上述操作,可获得两组融合特征图,分别表示为$\{M_2,M_3,M_4,M_5\}$和$\{N_2,N_3,N_4,N_5\}$。然后将输入PEFPN的原始特征图$\{P_2,P_3,P_4,P_5\}$引入其对应的融合层,并将P_i与其对应的M_i和N_i进行逐像素相加,以增强融合后的特征表示,可表示为

$$C_i = \begin{cases} M_{i+1} + P_i, & i = 2 \\ M_i + N_i + P_i, & 4 \geqslant i \geqslant 3 \\ N_{i-1} + P_i, & i = 5 \end{cases}, \tag{3}$$

式中,C_i为第i层的输出。融合操作的具体步骤如图3(c)所示,可以发现,PEFPN能有效避免低层特征在融合过程中的损失。

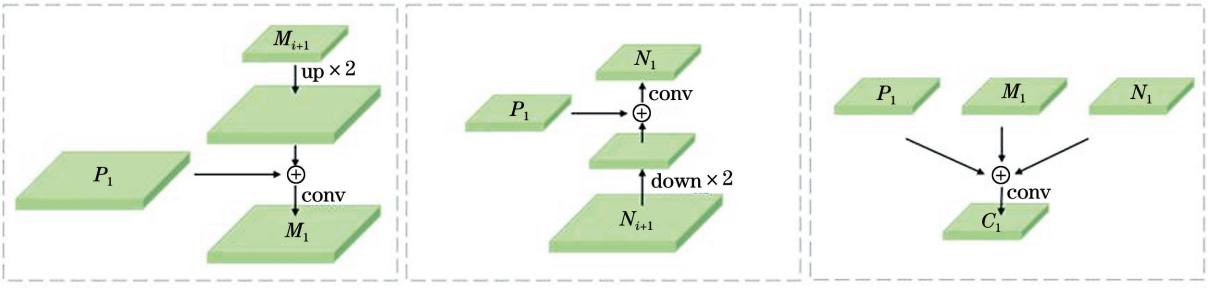

图3 PEFPN的结构。(a)自上而下;(b)自下而上;(c)特征融合

Fig. 3 Structure of PEFPN. (a) Top-down; (b) bottom-up; (c) feature fusion

3.3 优化函数

为了优化本算法的学习效果,引入了两种损失函数,可表示为

$$L = \lambda L_c + (1 - \lambda L_s), \tag{4}$$

式中,L_c为完整文本实例的损失,L_s为压缩文本实例的损失,参数λ可平衡两种损失的重要性。通常情况下,文本实例在一张图像中占据的面积较小,使用二值交叉熵(BCE)[27]损失函数时,对非文本域的预测会产生偏差。因此,采用Dice系数[28]平衡预测偏差,可表示为

$$D(S_i,G_i) = \frac{2\sum_{x,y}(S_{i,x,y}\times G_{i,x,y})}{\sum_{x,y}S_{i,x,y}^2 + \sum_{x,y}G_{i,x,y}^2}, \tag{5}$$

式中,$S_{i,x,y}$和$G_{i,x,y}$分别为像素点(x,y)的分割结果S_i和真实数据G_i。在自然场景中存在大量与文本类似的图案,如栅栏、格子,因此,采用在线难例实例挖掘(OHEM)算法获取L_c[29],以更好地辨别这些非文本图案。L_c更注重文本实例与非文本区域的分割,假设OHEM算法的训练掩码为M,则L_c可表示为

$$L_c = 1 - D(S_n \cdot M, G_n \cdot M), \tag{6}$$

式中，S_n 为分割结果，G_n 为真实数据。由于压缩文本区域被包围在完整文本实例中，为了避免冗余，忽略分割结果 S_n 中的非文本区域，则 L_s 可表示为

$$L_s=1-\frac{\sum_{i=1}^{n-1}D(S_i\cdot M,G_i\cdot M)}{n-1},\tag{7}$$

$$W_{x,y}=\begin{cases}1,\text{if } S_{n,x,y}\geqslant 0.5\\0,\text{otherwise}\end{cases},\tag{8}$$

式中，$W_{x,y}$ 为 S_n 中忽略非文本像素后的掩码，$S_{n,x,y}$ 为 S_n 中像素点(x,y)的像素值。

4 实验结果与分析

4.1 标签生成

在网络最后一步的处理中，使用 PSEA 区分紧密相邻的文本实例。该算法需使用不同尺寸的核，通过深度优先搜索(BFS)算法得到分割结果，因此，在训练时需要不同尺寸的真实标签，然后通过压缩原始文本，获得不同尺寸的真实标签，如图 4 所示。图 4(b)中的边界为原始文本实例的真实标签，对应图 4(c)中尺寸最大的分割标签掩码。为获得压缩掩码，算法将原始多边形 p_n 缩小 d_i 个像素，i 为缩放掩码的序号，如图 4(a)所示。

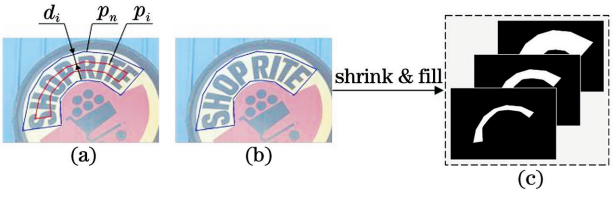

图 4 标签的生成流程。(a)压缩掩码的计算；(b)真实掩码；(c)压缩掩码

Fig. 4 Generation process of label. (a) Calculation of compression mask; (b) real mask; (c) compression mask

将缩小后的多边形 p_i 转换为二值掩码图，并作为分割标签的真实值，分别用 $G_1,G_2,\cdots,G_n$ 表示，设真实值与压缩掩码的比例为 r_i，则 p_n 边框上一点到 p_i 边框上的直线距离可表示为

$$d_i=\frac{X_{\text{Area}}(p_n)\times(1-r_i^2)}{X_{\text{Perimeter}}(p_n)},\tag{9}$$

式中，X_{Area} 为多边形的面积，$X_{\text{Perimeter}}$ 为多边形的周长。此外，真实标签 G_i 的比例 r_i 可表示为

$$r_i=1-\frac{(1-l)\times(n-i)}{n-1},\tag{10}$$

式中，l 为最小比例，其值在[0,1]之间。由(10)式可知，比例 $r_1,r_2,\cdots,r_n$ 由 n 和 l 两个超参数决定。

4.2 数据集

CTW1500 数据集[30]是一个曲线文本数据集，用 14 个顶点标注文本区域，共有 1000 张训练图像和 500 张测试图像。Total-Text 数据集[31]是一个拥有多形状文本的数据集，包括弯曲、水平和多方向文本，共有 1255 张训练图像和 300 张测试图像，用多边形和单词标注。ICDAR2015 数据集[32]由谷歌眼镜捕获，忽略位置、图像质量和视点，包括 1000 张训练图像和 500 张测试图像，仅包含英文，使用四边形框标注文本实例。MSRA-TD500 数据集[33]包含中文和英文两种语言，由 300 张训练图像和 20 张测试图像组成，文本行使用矩形框标注。

4.3 实验参数

实验在训练时用 2 块 NVIDIA GeForce GTX 1080 Ti 显卡，每批次输入 8 张图像，共进行 72000 迭代。初始学习率为 1×10^{-3}，分别在第 24000 和 48000 迭代时，以 10 倍比例进行衰减。用随机梯度下降(SGD)算法优化网络，权重衰减率为 5×10^{-4}，Nesterov 动量为 0.99，用 He 等[34]提出的方法进行权重初始化。

训练期间，用三种数据增广策略：1)将图像尺寸以{0.5,1.0,2.0,3.0}中的一种比例随机缩放；2)将图像以[−10°,10°]的角度进行随机翻转；3)所有图像的尺寸均被调整为 640 pixel×640 pixel。对于四边形文本实例，计算最小矩形面积以提取边界框；对于曲线文本实例，用后处理算法得到的结果作为输出。测试期间，用 1 块 NVIDIA GeForce GTX 1080 Ti 显卡，每批次输入 1 张图像，测试时为所有测试图像设置一个合适的宽度，然后按照宽高比调整图像的高度。

4.4 评估指标

用信息检索中使用的精确度(P)和召回率(R)评价本算法的检测性能，可利用 P 和 R 计算平均得分(F_{mean})，其中，P 和 R 基于 ICDAR2015 交并比(IOU)指标[32]计算，可表示为

$$X_{\text{IoU}}=\frac{X_{\text{Area}}(G_{\text{k}}\cap D_{\text{l}})}{X_{\text{Area}}(G_{\text{k}}\cup D_{\text{l}})},\tag{11}$$

式中，k 为真实文本框，l 为检测框。令正确检测的 IOU 阈值 $X_{\text{IOU}}>0.5$，则平均得分 F_{mean} 可表示为

$$F_{\text{mean}}=2\times\frac{P\times R}{P+R}。\tag{12}$$

4.5 验证实验

4.5.1 主干网络对检测结果的影响

为了更好地分析主干网络对检测结果的影响，

用多种主干网络在相同的超参数下进行训练，并在相同的数据集中进行测试；为了验证 PEFPN 的有效性，将传统 FPN 和 PEFPN 分别与实验中采用的所有主干网络进行结合，包括视觉几何组（VGG）网络、层数为 50 层、101 层的残差网络（Resnet-50 Resnet-101）、压缩激励网络（SENet）以及层数为 50 层、101 层的分散注意力残差网络（ResneSt-50 ResneSt-101），结果如表 2 所示。主干网络的检测效果如图 5 所示，可以发现，用 ResNeSt 作为检测器主干网络的检测精度最优，且 PEFPN 的性能均优于传统 FPN。

表 2 验证实验的结果

Table 2 Result of verification experiment unit：%

Backbone	ICDAR2015					
	FPN			PEFPN		
	P	R	F_{mean}	P	R	F_{mean}
VGG[35]	74.3	69.6	71.9	76.5	72.3	74.3
ResNet-50[1]	80.4	75.9	78.1	81.2	76.4	78.7
ResNet-101[1]	82.1	76.7	79.3	82.5	78.3	80.3
SENet[36]	81.6	77.3	79.4	82.4	78.1	80.2
ResNeSt-50[25]	82.5	79.4	81.0	83.1	80.3	81.7
ResNeSt-101[25]	83.0	80.8	81.9	83.8	81.7	82.7

图 5 验证实验的效果对比。（a）VGG＋FPN；（b）VGG＋PEFPN；（c）Resnet-50＋FPN；（d）Resnet-50＋PEFPN；（e）ResnetSt-50＋FPN；（f）ResnetSt-50＋PEFPN

Fig. 5 Effect comparison of verification experiment. (a) VGG＋FPN; (b) VGG＋PEFPN; (c) Resnet-50＋FPN; (d) Resnet-50＋PEFPN; (e) ResnetSt-50＋FPN; (f) ResnetSt-50＋PEFPN

4.5.2 实验结果分析

网络中的特征图如图 6（a）所示，可以看出，在网络提取特征阶段即编码阶段，使用分散注意力机制增加了文字区域的权重，导致文字特征对卷积核的响应更敏感；同时很好地抑制了不相关的特征，逐渐过滤掉复杂的背景。在解码阶段，分别使用了自上而下和自下而上的特征传递路径，小尺度文本被有效保留。这表明小尺度文本特征在通向特征金字塔顶层的路径中并没有发生信息丢失，能更完整地传递到 FPN 的顶层以进行逐尺度的特征融合操作。

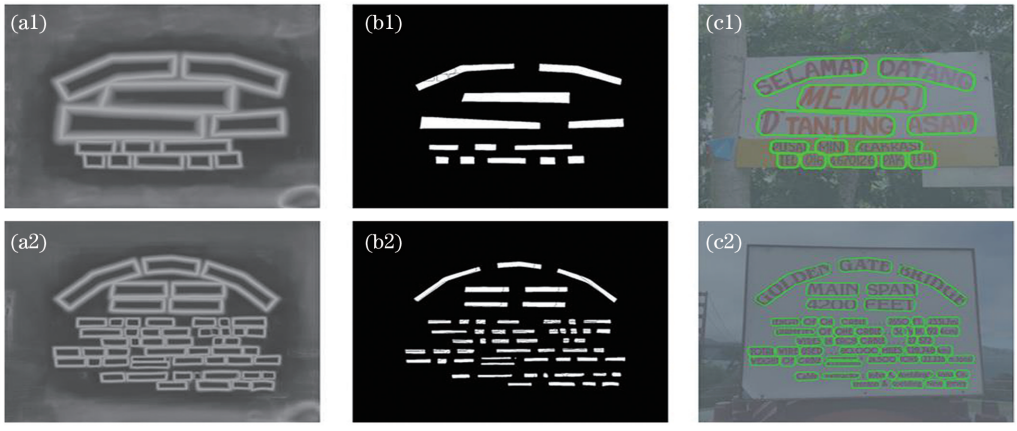

图 6 网络特征图的可视化。（a）特征图；（b）二值图；（c）结果图

Fig. 6 Visualization of network feature maps. (a) Feature map; (b) binary map; (c) result map

4.6 对比实验

为了评估本算法在不同数据集上的泛化能力，分别在曲线数据集 CTW1500、Total-Text，多方向数据集 CDAR2015 和多语言数据集 MSRA-TD500 上进行实验，结果如表 3 所示。可以发现，用 ResNeSt-101 作为主干网络时，在曲线数据集 CTW1500 和 Total-Text 上的平均精度分别为 78.7%和 79.0%，均超过文本连接建议网络(CTPN)、段连接网络(SegLink)、效率与精确文本检测网络(EAST)、文本蛇网络(TextSnake)、像素连接网络(PixeLink)、逐尺度扩张网络(PSENet)；在多方向数据集 ICDAR2015 和多语言数据集 MSRA-TD500 上的平均精度分别为 82.7%、79.3%，这表明本算法在多个数据集中具有较好的泛化能力，部分检测结果如图 7 所示。

表 3 不同算法在多个数据集上的检测结果

Table 3 Test results of different algorithms on multiple data sets unit: %

Method	CTW1500			Total-Text			ICDAR2015			MSRA-TD500		
	P	R	F_{mean}	P	R	F_{mean}	P	R	F_{mean}	P	R	F_{mean}
CTPN[37]	60.4	53.8	56.9	–	–	–	74.2	51.6	60.9	–	–	–
SegLink[38]	42.3	40.0	40.8	30.3	23.8	26.7	73.1	76.8	75.0	86.0	70.0	77.0
EAST[39]	78.7	49.1	60.4	50.0	36.2	42.0	83.6	73.5	78.2	87.3	67.4	76.1
TextSnake[40]	67.9	85.3	75.6	82.7	74.5	78.4	84.9	80.4	82.6	82.7	74.5	78.4
PixeLink[20]	–	–	–	–	–	–	82.9	81.7	82.3	83.0	73.2	77.8
PSENet[22]	80.6	75.6	78.0	81.8	75.1	78.3	81.5	79.7	80.6	–	–	–
Our(50)	80.9	75.9	78.3	82.0	74.7	78.2	83.1	80.3	81.7	83.5	74.8	79.0
Our (101)	81.3	76.2	78.7	82.5	75.7	79.0	83.8	81.7	82.7	84.0	75.1	79.3

图 7 不同数据集上的检测结果

Fig. 7 Detection results on different data sets

5 结 论

针对当前文本检测算法存在的问题，提出了用具有分散注意力机制的 ResNeSt 作为文本检测器的主干网络，以增强跨通道特征的交互，最大化地激活文本特征响应；同时，提出了 PEFPN，在特征金字塔中构建了一条短路径，使特征信息的流动方向由单向变为双向，并将主干网络中输出的原始特征图引入最后的融合阶段，以减少小尺度文本特征信息在特征融合时的损耗。在四个公开文本数据集上的实验结果表明，本算法在四个数据集上的平均精度均高于其他算法，充分验证了本算法的鲁棒性。在下一步研究中，将对算法作进一步优化，使其检测速度能够达到实时效果。

参 考 文 献

[1] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[2] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 2261-2269.

[3] Li X, Wang W H, Hu X L, et al. Selective kernel networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 510-519.

[4] Zhu X Z, Hu H, Lin S, et al. Deformable ConvNets V2: more deformable, better results [C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE 2019: 9300-9308.

[5] Chen L L, Zhang Z D, Peng L. Real-time detection based on improved single shot MultiBox detector[J]. Laser & Optoelectronics Progress, 2019, 56(1):

011001.
陈立里, 张正道, 彭力. 基于改进 SSD 的实时检测方法[J]. 激光与光电子学进展, 2019, 56(1): 011001.

[6] Wang D C, Chen X N, Zhao F, et al. Vehicle detection algorithm based on convolutional neural network and RGB-D images[J]. Laser & Optoelectronics Progress, 2019, 56(18): 181003.
王得成, 陈向宁, 赵峰, 等. 基于卷积神经网络和 RGB-D 图像的车辆检测算法[J]. 激光与光电子学进展, 2019, 56(18): 181003.

[7] Shi F F, Zhang S L, Peng L. Salient object detection based on deep residual networks and edge supervised learning[J]. Laser & Optoelectronics Progress, 2019, 56(15): 151502.
时斐斐, 张松龙, 彭力. 基于深度残差网络与边缘监督学习的显著性检测[J]. 激光与光电子学进展, 2019, 56(15): 151502.

[8] Liu Y L, Jin L W, Zhang S T, et al. Curved scene text detection via transverse and longitudinal sequence connection[J]. Pattern Recognition, 2019, 90: 337-345.

[9] Liu X H, Sun S Y, Gu L P, et al. 3D object detection based on improved Frustum PointNet[J]. Laser & Optoelectronics Progress, 2020, 57(20): 201508.
刘训华, 孙韶媛, 顾立鹏, 等. 基于改进 Frustum PointNet 的 3D 目标检测[J]. 激光与光电子学进展, 2020, 57(20): 201508.

[10] Xue C H, Lu S J, Zhang W. MSR: multi-scale shape regression for scene text detection[EB/OL]. [2020-05-22]. https://arxiv.org/abs/1901.02596.

[11] Tian Z T, Shu M, Lyu P Y, et al. Learning shape-aware embedding for scene text detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 4229-4238.

[12] Liao M H, Shi B G, Bai X, et al. Textboxes: a fast text detector with a single deep neural network[EB/OL]. [2020-05-28]. https://arxiv.org/abs/1611.06779.

[13] 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.

[14] He P, Huang W L, He T, et al. Single shot text detector with regional attention[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 3066-3074.

[15] Liao M H, Zhu Z, Shi B G, et al. Rotation-sensitive regression for oriented scene text detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 5909-5918.

[16] Zhang Z, Zhang C Q, Shen W, et al. Multi-oriented text detection with fully convolutional networks[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 4159-4167.

[17] Yao C, Bai X, Sang N, et al. Scene text detection via holistic, multi-channel prediction[EB/OL]. [2020-05-22]. https://arxiv.org/abs/1606.09002.

[18] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 3431-3440.

[19] Lyu P Y, Yao C, Wu W H, et al. Multi-oriented scene text detection via corner localization and region segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 7553-7563.

[20] Deng D, Liu H F, Li X L, et al. PixelLink: detecting scene text via instance segmentation[EB/OL]. [2020-05-25]. https://arxiv.org/abs/1606.09002.

[21] Xie E Z, Zang Y H, Shao S, et al. Scene text detection with supervised pyramid context network[EB/OL]. [2020-05-25]. https://arxiv.org/abs/1811.08605.

[22] Wang W H, Xie E Z, Li X, et al. Shape robust text detection with progressive scale expansion network[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 9328-9337.

[23] Tan M X, Le Q V. EfficientNet: rethinking model scaling for convolutional neural networks[EB/OL]. [2020-05-25]. https://arxiv.org/abs/1905.11946.

[24] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu,

HI, USA. New York: IEEE, 2017: 936-944.

[25] Zhang H, Wu C R, Zhang Z Y, et al. ResNeSt: split-attention networks [EB/OL]. [2020-05-28]. https://arxiv.org/abs/2004.08955.

[26] Chollet F. Xception: deep learning with depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 1800-1807.

[27] de Boer P T, Kroese D P, Mannor S, et al. A tutorial on the cross-entropy method[J]. Annals of Operations Research, 2005, 134(1): 19-67.

[28] Milletari F, Navab N, Ahmadi S A. V-net: fully convolutional neuralnetworks for volumetric medical image segmentation[C]//2016 Fourth International Conference on 3D Vision (3DV), October 25-28, 2016, Stanford, CA, USA. New York: IEEE, 2016: 565-571.

[29] Shrivastava A, Martens J, Dahl G, et al. On the importance of initialization and momentum in deep learning[C]//Proceedings of the 30th International Conference on Machine Learning, June 16-21, 2013, Atlanta, Georgia, USA. New York: ACM, 2013, 28: 1139-1147.

[30] Liu Y L, Jin L W, Zhang S T, et al. Detecting curve text in the wild: new dataset and new solution[EB/OL]. [2020-05-26]. https://arxiv.org/abs/1712.02170.

[31] Ch′ng C K, Chan C S. Total-text: a comprehensive dataset for scene text detection and recognition[C]// 14th IAPR International Conference on Document Analysis and Recognition, November 9-15, 2017, Kyoto, Japan. New York: IEEE, 2017: 935-942.

[32] Karatzas D, Gomez-Bigorda L, Nicolaou A, et al. ICDAR 2015 competition on robust reading[C]//2015 13th International Conference on Document Analysis and Recognition (ICDAR), August 23-26, 2015, Tunis, Tunisia. New York: IEEE, 2015: 1156-1160.

[33] Yao C, Bai X, Liu W, et al. Detecting texts of arbitrary orientations in natural images[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, June16-21, 2012, Providence, RI, USA. New York: IEEE, 2012: 1083-1090.

[34] He K M, Zhang X Y, Ren S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]2015// IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1026-1034.

[35] Simonyan K, Zisserman A. Very deep convolution networks for large-scale image recognition[EB/OL]. [2020-05-26]. https://arxiv.org/abs/1409.1556.

[36] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 7132-7141.

[37] Tian Z, Huang W L, He T, et al. Detecting text in natural image with connectionist text proposal network[M]//Leibe B, Matas J, Sebe N, et al. Computer Vision-ECCV 2016. Lecture Notes in Computer Science. Cham: Springer, 2016, 9912: 56-72.

[38] Shi B G, Bai X, Belongie S. Detecting oriented text in natural images by linking segments[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 3482-3490.

[39] Zhou X Y, Yao C, Wen H, et al. EAST: an efficient and accurate scene text detector[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 2642-2651.

[40] Long S B, Ruan J Q, Zhang W J, et al. TextSnake: a flexible representation for detecting text of arbitrary shapes[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer Vision-ECCV 2018. ECCV 2018. Lecture Notes in Computer Science. 2018, 11206: 19-35.