自动化学报
ACTA AUTOMATICA SINICA

基于边缘特征增强的任意形状文本检测网络

白鹤翔 王浩然

A New Arbitrary-shaped Text Detection Network by Reinforcing Edge Features

BAI He-Xiang, WANG Hao-Ran

在线阅读 View online: https://doi.org/10.16383/j.aas.c220429

您可能感兴趣的其他文章

自然场景图像中的文本检测综述

Text Detection in Natural Scene Image: A Survey

自动化学报. 2018, 44(12): 2113-2141 https://doi.org/10.16383/j.aas.2018.c170572

面向边缘计算应用的宽度孪生网络

Broad Siamese Network for Edge Computing Applications

自动化学报. 2020, 46(10): 2060-2071 https://doi.org/10.16383/j.aas.c200555

基于短文本情感增强的在线学习者成绩预测方法

Short-text Sentiment Enhanced Achievement Prediction Method for Online Learners

自动化学报. 2020, 46(9): 1927-1940 https://doi.org/10.16383/j.aas.c190008

基于多尺度先验深度特征的多目标显著性检测方法

Multiple Salient Objects Detection Using Multi-scale Prior and Deep Features

自动化学报. 2019, 45(11): 2058-2070 https://doi.org/10.16383/j.aas.c170154

基于全卷积神经网络与低秩稀疏分解的显著性检测

Saliency Detection via Full Convolution Neural Network and Low Rank Sparse Decomposition

自动化学报. 2019, 45(11): 2148-2158 https://doi.org/10.16383/j.aas.2018.c170535

基于多尺度特征融合反投影网络的图像超分辨率重建

Multiscale Feature Fusion Back-projection Network for Image Super-resolution

自动化学报. 2021, 47(7): 1689-1700 https://doi.org/10.16383/j.aas.c200714

基于边缘特征增强的任意形状文本检测网络

白鹤翔 [1] 王浩然 [1]

摘 要 在场景文本检测方法中, 文本实例的边缘特征与其他特征在大多数模型中都是以同样的方式进行处理, 而准确检测相邻文本边缘区域是正确识别任意形状文本区域的关键之一. 如果对边缘特征进行增强并使用独立分支进行建模, 必能有效提高模型的标识准确率. 为此, 提出了三个用以增强边缘特征的网络模块. 其中, 浅层特征增强模块可有效增强包含更多边缘特征的浅层特征; 边缘区域检测分支将普通特征和边缘特征进行区分以对目标的边缘特征进行显式建模; 而分支特征融合模块可将两种特征在识别过程进行更好的融合. 在将这三个模块引入渐进尺度扩张网络 (Progressive scale expansion network, PSENet) 之后, 相关消融实验表明这三个模块的单独使用及其组合均可进一步增加网络的预测准确率. 此外, 在三个常用公开数据集上与其他十个最新模型的比较结果表明, 改进后得到边缘特征增强网络 (Edge-oriented feature reinforcing network, EFRNet) 的识别结果具有较高的 F1 值.

关键词 场景文本检测, 任意形状, 边缘区域, 浅层特征, 渐进尺度扩张网络

引用格式 白鹤翔, 王浩然. 基于边缘特征增强的任意形状文本检测网络. 自动化学报, 2023, **49**(5): 1019−1030

DOI 10.16383/j.aas.c220429

A New Arbitrary-shaped Text Detection Network by Reinforcing Edge Features

BAI He-Xiang[1] WANG Hao-Ran[1]

Abstract In the detection of scene texts areas, the text instances′ edge features are processed in the same way as other features. Nevertheless, the accurate detection of adjacent text edges is crucial in the correct identification of arbitrary-shaped text regions in natural scenes. Obviously, the identification accuracy increases if edge features can be enhanced and modeled through independent branches in the network. To this end, three network modules are proposed to enhance the edge features in this paper. These modules are the shallow feature enhancement module which effectively enhances the shallow features with more edge features, the edge region detection module which decouples the original features into edge features and text features to explicitly model the edge features of the object, and the branch feature fusion module which effectively fuses these two types of features in the recognition process. After the proposed modules are added to the progressive scale expansion network (PSENet), the ablation experiments show that both the independent application and the synthetic application of these modules increase the prediction accuracy. In addition, the comparison experiments on three commonly used public datasets with ten state-of-the-art methods show that the improved edge-oriented feature reinforcing network (EFRNet) has higher F1-measure accuracy.

Key words Scene text detection, arbitrary-shaped, edge region, shallow feature, progressive scale expansion network (PSENet)

Citation Bai He-Xiang, Wang Hao-Ran. A new arbitrary-shaped text detection network by reinforcing edge features. *Acta Automatica Sinica*, 2023, **49**(5): 1019−1030

从复杂多变的自然场景中快速有效地检测文本信息区域, 即场景文本检测, 在实际应用中具有重要的应用价值. 其主要任务是将自然场景图像中的文本区域标注出来, 以便为其他应用提供基础. 例如, 正确地识别场景文本区域能够有效提高即时翻译[1] 和场景理解[2] 的准确率, 并为其后续应用如自动驾驶[3] 等提供更为有效的支撑.

场景文本检测方法的准确率与文本区域的特征有效建模密切相关. 经典的检测方法主要使用了文本区域的形状、对比度、以及边缘强度和边缘密度等特征在不同的框架下进行文本区域检测[4−6]. 深度学习场景文本检测方法能够挖掘更深层的文本区域特征, 特别是多尺度的复杂特征, 因此其相较于经

收稿日期 2022-05-23 录用日期 2022-09-26
Manuscript received May 23, 2022; accepted September 26, 2022
国家自然科学基金 (41871286, 62072291), 国家重点研发计划课题 (2017YFB0503501) 资助
Supported by National Natural Science Foundation of China (41871286, 62072291), and National Key Research and Development Plan (2017YFB0503501)
本文责任编委 白翔
Recommended by Associate Editor BAI Xiang
1. 山西大学计算机与信息技术学院 太原 030006
1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006

典方法有着更高的准确率[7−8].

目前, 基于深度学习的文本检测方法使用了两种不同的策略进行文本区域的识别. 第一种策略是通过逐像元多尺度特征预测候选文本框, 再对候选文本框进行分类从而挑选出正确的文本区域位置, 此类方法被称为基于回归的检测方法[9−12]. 另一种策略也被称为基于语义分割的检测方法[13−15], 这一策略直接使用多尺度进行逐像元的文本/非文本二分类, 再进行区域融合以得到准确的文本区域.

由于实际场景中的文本区域复杂多变, 因此对任意形状文本区域的正确检测是提高识别准确率的关键. 如果缺乏对这一因素的考虑, 会导致识别结果的各种异常. 例如, 仅使用矩形候选框的基于回归的检测方法难以避免地会出现不同文本实例之间的混杂交错. 基于语义分割的方法虽然不再受限于矩形文本框, 但是如何利用文本区域复杂形状特征区分相邻文本的边缘区域仍然是一个有待解决的难题. 因此, 对于任意形状文本区域的检测成为了近年来的一个研究热点.

为此, 很多研究工作在上述两种策略的基础上进行相应改进, 以应对任意形状文本区域检测的挑战. 在基于回归的文本检测方法方面, 通常是利用文本区域的特征改进候选框的形状, 而不再局限于矩形的候选框. 例如, TextRay 利用任意形状文本区域中心点与其轮廓在不同方向上距离的不同以识别非矩形的候选框[16]; ABCNet (Adaptive bezier-curve network) 使用文本区域边缘特征点通过贝塞尔曲线以重建文本实例的非矩形轮廓[17]; FCENet (Fourier contour embedding network) 则在频率域预测傅里叶特征向量, 然后通过这些特征向量在图像空间域中重建文本实例轮廓点的位置坐标[18]. 此外, 还有一些方法使用分割子网络的方式对文本区域进行二次调整以得到精确的文本实例边界[19−20] 或者使用自底向上聚合识别出的文本组件以重建文本实例[21].

在基于语义分割的文本检测方法方面, 主要是通过各种方式增加不同文本实例之间以及文本实例同背景之间的区分度以提高文本区域检测的准确度. 其中, 一些方法通过逐步扩张的方式从文本实例的核心逐步添加区分度高的像元或区域以最终确定文本实例的轮廓, 如渐进尺度扩张网络 (Progressive scale expansion network, PSENet)[22]、Centripetal Text[23] 以及 CSENet (Conditional spatial expansion network)[24]. 一些方法将像元特征嵌入到一个新的空间以增加实例像元的内聚性[25]. 此外还有一些方法通过注意力机制强化近邻信息以更好地区分不同文本实例[26].

实际上, 不论采用何种改进方法以识别任意形状的文本区域, 文本区域的边缘识别以及骨干网络获取的像元特征都有着不可替代的作用. 在基于深度学习的模型中, 边缘信息和像元特征的提取是相辅相成的. 更丰富的像元特征有助于更好地确定文本区域的边缘, 同时利用训练集中定义良好的边缘信息也能够帮助神经网络提取出更易于识别边缘区域的像元特征. 因此, 在深度学习模型中通过引入边缘信息增强像元特征的方式可以更好地从原始图像中提取出同文本边界相关的像元特征, 从而增加不同文本实例之间以及文本实例同背景之间的区分度, 并进一步提高任意文本区域的检测准确度.

基于上述讨论, 本文以 PSENet 为例, 通过引入边缘预测模块和特征增强模块以进一步提高任意形状场景文本检测的准确度. 其中, 边缘预测模块增加了文本边缘区域作为预测目标, 使得网络在训练过程中可以引入文本边缘结构信息; 而特征增强模块将像元的局部特征进行强化, 从而使其更易于提取文本的边缘区域特征. 本方法的优点在于所增加的文本边缘预测分支增强了同边缘密切相关的特征, 使其更易于识别自然场景下的任意形状文本框. 通过在三个公开数据集上与 TextSnake、PSENet、FCENet 等 10 个最新任意形状场景文本检测方法的对比实验表明, 在引入这两个模块后的 PSENet, 即边缘特征增强网络 (Edge-oriented feature reinforcing network, EFRNet) 的 F1 值比原始 PSENet 提高了至少 7%, 比基于 PSENet 的 PAN++ 也提高了 1.4% 到 4.7%, 同时相对于其他方法也具有较高的 F1 值. 此外, 相关消融实验表明, 引入本文所提出的不同模块后 F1 值可提高 1% 到 3%.

1 边缘特征增强网络

边缘信息在任意形状场景文本检测中扮演着重要的角色. 在特征提取阶段对文本对象的边缘特征进行增强, 必然能够为文本轮廓的准确识别提供更有代表性的特征, 从而提高检测准确度. 目前, 大多数场景文本检测方法都使用 ResNet[27] 作为骨干网络进行特征提取, 并且通过特征金字塔网络 (Feature pyramid network, FPN) 进行特征融合. 融合后的特征再经过一系列的处理后最终用作文本区域检测的特征. 然而, 在这一特征提取过程并未显式地引入文本框的边缘特征. 本节以 PSENet 为基础, 在特征提取阶段, 有针对性地对文本框边缘特征的提取进行了增强, 提出了 EFRNet 模型, 以期提高文本对象检测的准确度.

PSENet 是一个有代表性的基于语义分割的场

景文本检测方法. 这一方法可以分为三个阶段. 在特征提取阶段, 这一方法也是以 ResNet 作为骨干网络结合 FPN 进行了特征融合, 并对融合后的特征进行了上采样、拼接和卷积. 在识别阶段, 这些特征被用以预测不同尺度的文本核心区和完整文本区. 在后处理阶段, PSENet 在多尺度识别结果的基础上将核心区域逐步扩张得到文本对象的轮廓.

EFRNet 在 PSENet 特征提取阶段增加了两个模块, 分别为特征增强模块和分支特征融合模块. EFRNet 的网络结构如图 1 所示, 在 ResNet 骨干网络和 FPN 模块之后该网络增加了特征增强模块. 这一模块可按照顺序分为两个子模块. FPN 模块的输出首先经过第一个子模块, 即浅层特征增强模块, 以对 FPN 中最小尺度细节特征信息进行增强. 紧接浅层特征增强子模块的为分支特征融合子模块. 该子模块包括文本特征提取和边缘特征提取两个部分, 其提取出的特征分别记为 TF (Text feature) 和 EF (Edge feature). TF 和 EF 都是通过对浅层特征增强模块输出的特征卷积而得到, 并分别用以预测完整文本实例和文本实例的边缘区域.

经过训练后, TF 更倾向于包含场景图像中完整实例的特征, 而 EF 则更倾向于包含场景图像中文本对象边缘区域的特征. 在得到 TF 和 EF 后, 再对两者的特征进行融合以得到二者的混合特征 (Mixed feature, MF), 并同时用以强化完整文本实例和文本实例的边缘区域的检测. EFRNet 的识别阶段包含了三个分支, 分别为完整文本实例检测分支、文本实例边缘区域检测分支、以及内核图 (Kernel map) 预测分支. 在对完整文本实例进行检测时, EFRNet 同时使用了 TF 和 MF 特征对文本实例区域进行预测. EFRNet 所增加的文本实例边缘区域检测分支分别从 EF 和 MF 中提取出边缘信息并进行拼接以对文本边缘区域进行预测. 这两个分支的具体实现细节请参见第 1.3 节. 最后在内核图预测时, EFRNet 也使用与 PSENet 相同的方法计算得出了 n 个不同尺度的内核图 $K_1, \cdots, K_{n-1}, K_n$. 即对文本框标注区域使用不同参数进行腐蚀操作. 与 PSENet 不同的是, EFRNet 使用的是文本实例的边缘区域特征 (EF), 而不是通过 FPN 得到的融合特征, 内核图的具体计算方式请参见文献 [22]. 三个不同的预测分支所采用的真实值均来源于文本实例的边框标注. 其中训练数据的文本实例区域和内核图标注采用了同 PSENet 相同的方式. 文本边缘区域的标注方法请参见实验部分.

在对实际数据进行预测时, 当目标场景图像通过特征增强模块后, EFRNet 使用完整文本实例检测和内核图预测两个分支计算出图像中的所有文本实例区域的多尺度识别结果. 在此基础上, 与 PSENet 相同, EFRNet 也使用其相应的渐进扩展后处理算法得到文本区域的最终预测结果. 在这一过程中,

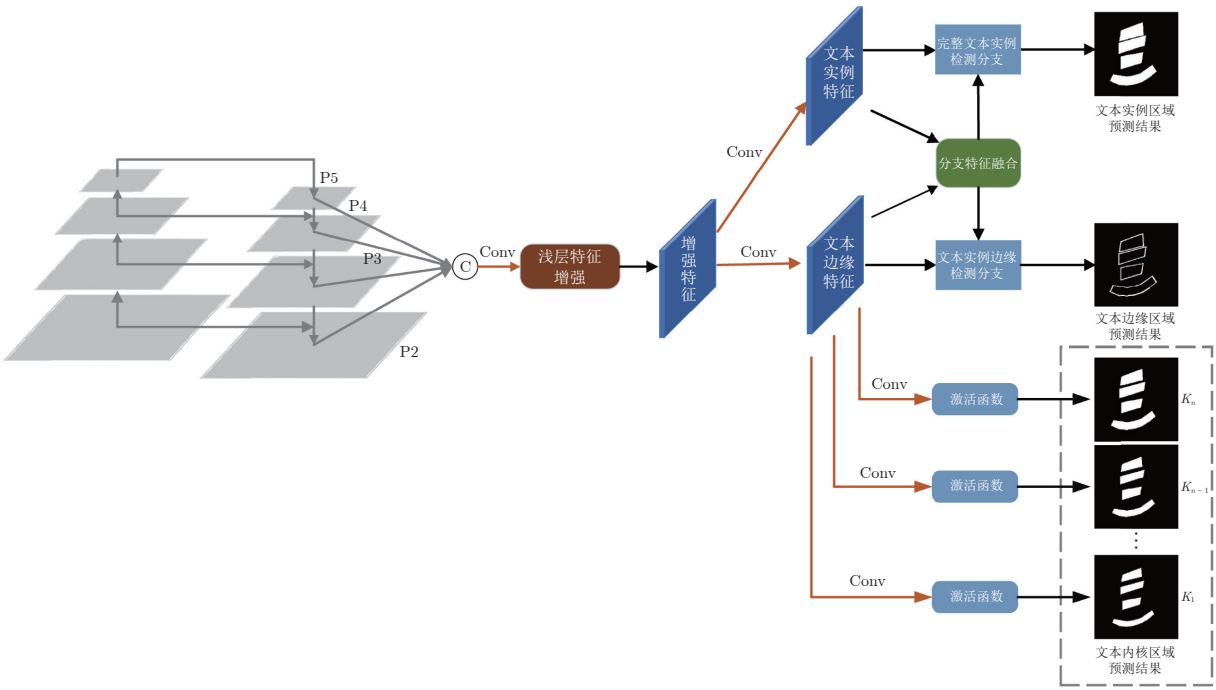


图 1　EFRNet 网络整体架构

Fig. 1　Overall pipeline of EFRNet

边缘检测结果并未参与文本区域预测. 边缘预测模块的主要作用是在特征提取过程中能够帮助 EFR-Net 提取文本对象边缘区域的特征. 这些特征在预测阶段, 可用以提高完整文本实例检测和内核图预测的准确度, 从而为渐进扩展后处理算法提供更可靠的依据.

1.1 浅层特征增强模块

通过将每个分辨率的特征与上采样的低分辨率特征逐像元相加的方式, FPN 可以将大尺度的语义信息逐层传递到高分辨率特征, 并从而提取出更有区分性的特征. 因此, FPN 可有效地对语义分割网络中深层特征, 即语义信息, 进行建模. 然而, 在场景文本检测中, 浅层特征, 特别是同文本框边缘密切相关的浅层特征, 对于正确检测文本区域有着重要的价值. 因此, 除了要融合大尺度的语义信息之外, 应当对最浅层特征进行单独处理和增强, 以避免融合后的文本区域特征中细节信息的缺乏.

浅层特征增强模块如图 2 所示, 与 PSENet 相同的是, 该模块首先也需要使用卷积融合 FPN 所有分辨率的特征以得到融合后的特征 (图 2 中的融合特征). 融合特征中包含了目标图像在多个尺度上的特征. 这些特征既包含图像的浅层特征, 也包含图像的深层特征. 紧接着, 该模块对融合特征再次进行了二次卷积, 并得到图 2 中的深层特征.

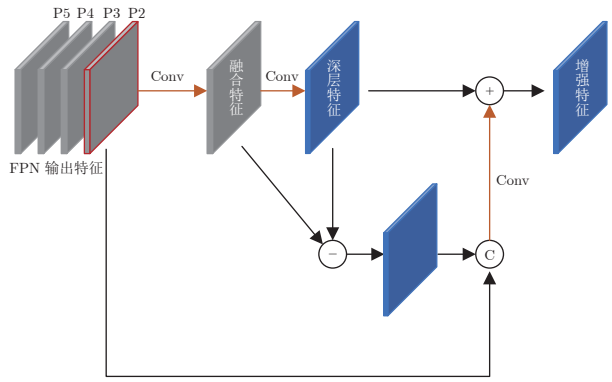


图 2 浅层特征增强模块

Fig. 2 Shallow feature enhancement module

由于经过了多层卷积, 这一深层特征往往包含了更多的语义信息[28]. 当得到融合特征和深层特征后, 再使用融合特征减去对应的深层特征. 这一操作在一定程度上过滤掉图像深层特征的同时保留了融合特征中的浅层特征. 然而, 经过层层处理后得到的这一图像特征必然会存在不同程度浅层特征信息的损失. 为了能够弥补这一损失, 该模块将包含了丰富细节信息的 P2 与这一图像特征拼接, 以期能够在识别过程中对图像的浅层特征进行增强. 最后, 将强化后的浅层特征卷积后与深层特征再逐像元相加, 作为后续文本检测的图像特征. 这一图像特征与 FPN 输出特征相比具有更强的浅层特征, 有助于更准确地识别出文本实例的边缘区域.

1.2 分支特征融合模块

为了能够更好地利用文本实例的边缘信息, 在得到浅层信息增强特征后, 将该特征通过不同的卷积以构建面向文本实例与其边缘的特征, 即图 3 中的 TF 和 EF. 此外, 为了突出最具有代表性的特征并抑制无关特征, TF 和 EF 都分别使用了注意力机制对不同特征赋予不同的权重. 该权重的获取主要是通过压缩提取 (Squeeze and extract, SE)[29] 通道注意模块完成 (如图 3 中的 SE 所示).

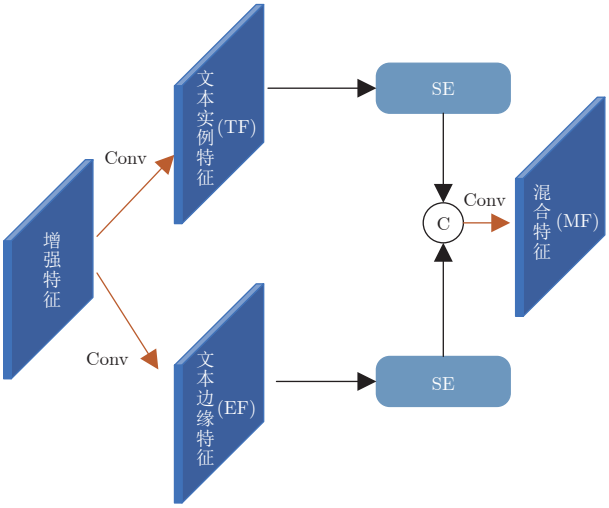


图 3 分支特征融合模块

Fig. 3 Branch feature fusion module

紧接着, 经过注意力机制处理的 TF 和 EF 再通过拼接和二次卷积后进行融合以得到同时包含两者信息的混合特征 MF. 随后, TF 和 EF 将分别用于检测文本实例和文本实例边缘, 而这一混合特征将被同时用于文本实例及其边缘区域的检测.

1.3 文本实例及边缘区域检测分支

文本实例区域检测分支的主要任务是检测出完整的文本实例区域. 图 4 的上半部分即为文本实例区域检测分支. 同 PSENet 中相应部分最大的区别在于, 为了能够得到更准确的分割结果, 该分支采用了文献 [30] 中的可微二值化方法, 利用 MF 对 TF 得到的文本实例区域和背景区域预测结果进行了微调. 该分支首先根据对 TF 进行卷积激活后得到目标图像的文本实例区域和背景区域的初始结果, 并

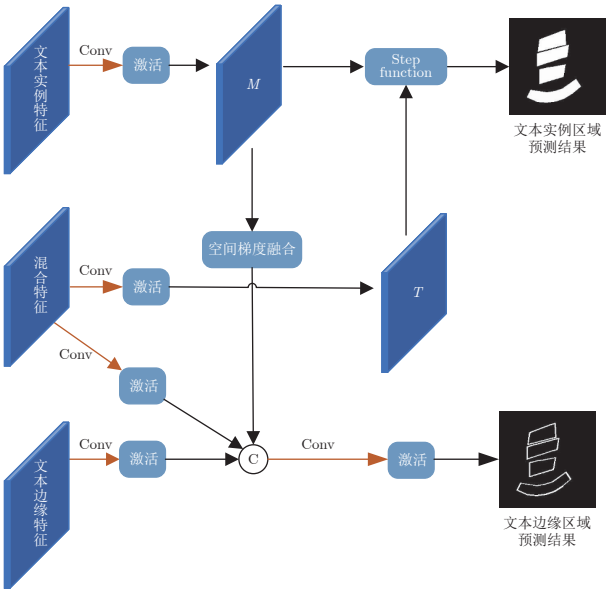


图 4　文本实例及边缘区域检测分支

Fig. 4　Detection branch of text instance and text edge region

记为 M. 随后再对 MF 进行卷积和激活得到中间结果 T 并将其用作 M 的逐像元调整阈值. 最后使用式 (1), 即图 4 中的 Step function, 根据阈值 T 对 M 进行微调以得到最终的文本实例区域预测结果.

$$M = \frac{1}{1+\exp(-k(M-T))} \tag{1}$$

其中 k 表示放大因子. 由于 T 是通过卷积激活的方式从 MF 中构建的, 因此其在网络训练过程中会不断地根据损失情况进行更新[30]. 参数 k 在这一更新中具有重要的作用. 如果 k 取值过小会导致梯度变小, 收敛缓慢; 而 k 过大则会将部分背景区域误识别为文本实例区域, 从而导致较大的预测误差[30]. 式 (1) 已被广泛地应用于 DB (Differentiable binarization) 网络模型及其各种改进中. 在这些模型中, 放大因子 k 在实验中均被设置为 50, 并且获得了较为理想的实验结果. 因此, 后续实验中也采用了这一经验值, 并取得了较好的检测精度.

同 DB 方法一样, EFRNet 利用式 (2) 的自动二值化功能, 通过对前景背景的区分进行优化以区分距离较近的文本实例. 除此之外, 该公式还可以过滤边缘区域部分被错误分割的像元. EFRNet 与 DB 方法中对于式 (2) 应用的最大不同点在于阈值图 T 的计算和训练方法. DB 方法主要通过计算文本边缘标注预测值的误差来对阈值图 T 进行更新. 而本文所提出的方法并没有计算边缘标注的预测误差, 而是对前面得到的混合特征进行卷积的方式直接得到阈值 T, 并且在训练过程中不断更新卷积核的参数以选取最优卷积核.

图 4 的下半部分为文本实例边缘检测分支. 这一分支的目的是区分文本实例的边缘区域和非边缘区域, 并对提取边缘特征相关的参数进行优化, 以期从原始图像中提取出更具有代表性的边缘特征. 这一分支首先采用文献 [31] 中的空间梯度融合方法, 通过式 (2) 对文本实例区域检测分支的初始结果 M 进行池化和激活等操作以计算出文本实例区域检测分支的边缘信息.

$$\nabla M = \sigma\left(\left|M - pool_u\left(M\right)\right|\right) \tag{2}$$

其中, σ 是激活函数; $pool_u$ 对应于内核大小为 u 的自适应平均池化操作. 在实验过程中参考文献 [31] 将 u 设置为 3.

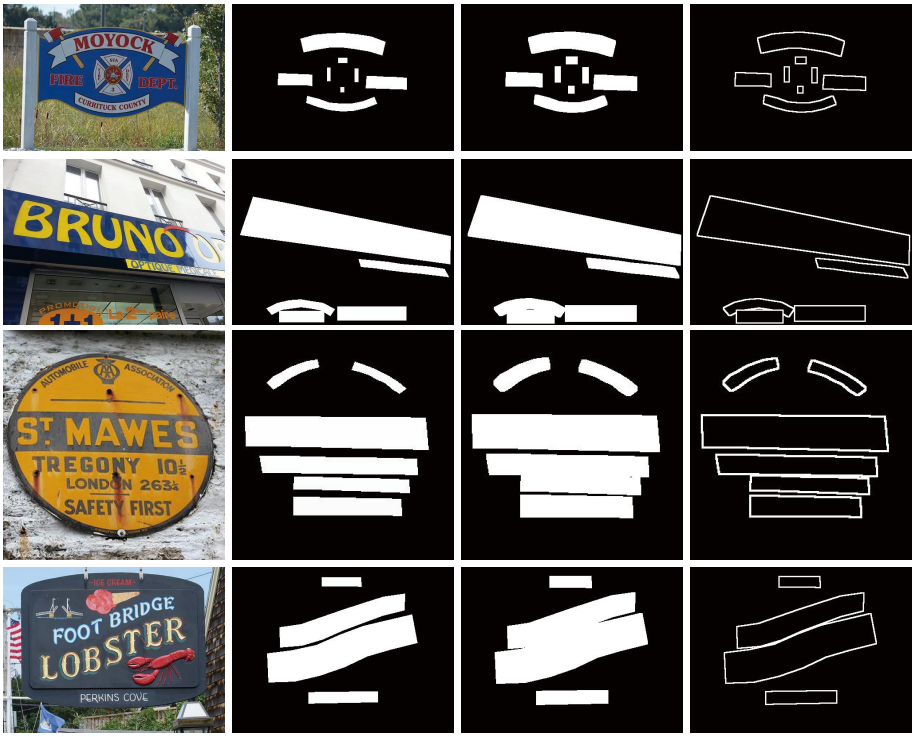
式 (2) 的具体计算过程如图 5 所示. 图 5 中第 1 列为原始图像; 第 2 列为相应原始图像通过文本实例检测得到的初始检测结果, 即式 (2) 中的 M; 第 3 列为使用最大池化函数对 M 进行处理后得到的文本实例区域, 也就是式 (2) 中的 $pool_u$ 函数. 该操作相当于对 M 中的文本实例区域进行了一次膨胀操作. 随后, 对 M 和 $pool_u(M)$ 的差的绝对值进行激活后得到式 (2) 的最终结果, 即从文本特征计算出的文本边缘区域, 如图 5 第 4 列所示.

此外, 边缘检测分支分别对 EF 和 MF 进行卷积和激活以得到从其所对应的初始文本边缘区域检测结果. 然而, 从单一特征中得到的初始检测结果仍然存在较大的误差. 为了得到更准确的文本边缘检测结果, 该分支将来自 M 、EF 以及 MF 使用卷积得到的文本边缘区域进行融合. 在融合过程中, 该分支将三者得到的文本边缘区域先进行拼接, 再通过卷积激活后, 得到最终的文本边缘区域的预测结果.

1.4　损失函数

通常, 在自然场景文本检测中, 文本对象所占区域相对较小, 因此在训练过程中存在着正负样本不均衡问题. 而这一问题会导致网络的预测偏向于背景. 在深度学习中 dice 损失函数[32] 可以有效解决这一问题. 目前该损失函数已经被广泛地用于解决医学图像分割中前景背景的样本不平衡问题. 同时, OHEM (Online hard example mining)[33] 方法能够进一步平衡前景和背景样本. EFRNet 同时采用了这两种方法设计相应的损失函数. 其一是根据 OHEM 的要求, 将正负样本的比例根据数据情况设置为 1 : 3; 其二是使用了 dice 损失函数.

在训练过程中所涉及到的损失函数共有三个, 分别对应于完整文本实例的损失函数 L_{T} 、文本实



(a) (b) (c) (d)

图 5 文本分割图空间梯度计算过程 ((a) 原始图像; (b) 初始检测结果; (c) 最大池化; (d) 探测到的边缘区域)

Fig. 5 Spatial gradient calculation for text segmentation graph ((a) Original images; (b) Initial detection results; (c) Max pool results; (d) Detected edges)

例的内核图损失函数 L_{K}, 以及文本实例边缘区域的损失函数 L_{E}. 对于每个损失函数 $L_{\mathrm{O}} \in \{L_{\mathrm{T}}, L_{\mathrm{K}}, L_{\mathrm{E}}\}$, 首先使用 OHEM 方法根据正负样本比例将分割结果和实际标签进行掩膜处理, 并得到掩膜后的结果为 S 和 G, 再使用

$$L_{\mathrm{O}}(S, G) = 1 - \frac{2\sum\limits_{x,y}(S_{x,y} \times G_{x,y})}{\sum\limits_{x,y}(S_{x,y}^2 + G_{x,y}^2)} \tag{3}$$

计算其损失, 其中 $S_{x,y}$ 和 $G_{x,y}$ 分别为分割结果和实际标签掩膜后的像素值. 当计算内核图损失时, 需要计算不同尺度下的损失并进行综合以得到对应的损失值. 在计算 K_i 尺度的内核图损失时, $S_{x,y}$ 表示坐标为 (x, y) 的像元通过网络所预测出的结果, 即该像元是否处于 K_i 尺度下文本区域内部, 并用 1 和 0 分别表示处于文本区域内部和外部. $G_{x,y}$ 表示在该尺度下, 通过 OHEM 掩膜处理后, 坐标为 (x, y) 的像元是否处于文本区域内部的真实值. 根据这两个信息, 使用式 (3) 即可计算出该尺度内核图的损失. 当计算出所有尺度下内核图的损失后, 使用其均值作为内核图的综合损失 L_{K}.

最后, 如式 (4) 所示, 整个模型的损失函数可以表示为文本实例内核损失、文本边缘损失以及文本内核损失的加权求和.

$$L = \lambda_1 L_{\mathrm{T}} + \lambda_2 L_{\mathrm{K}} + \lambda_3 L_{\mathrm{E}} \tag{4}$$

其中, λ_1、λ_2 和 λ_2 分别为文本实例、文本内核以及文本边缘的损失权重. 在具体实现过程中, λ_1 和 λ_2 两个参数的比例被设置为 2 : 1, 这一比例同 PSENet 中文本实例损失权重和内核损失权重的比例相当. 此外, 经过多次实验表明, 只有当文本边缘与文本对象权重相当时, 才能够充分发挥文本边缘特征在文本检测中的作用, 得到较高的检测精度. 因此, 按照 2 : 1 : 2 的比例, λ_1、λ_2 和 λ_3 分别设置为 0.4, 0.2 和 0.4.

2 实验

本节首先介绍了用以评价 EFRNet 有效性的三个场景文本检测数据集 (CTW1500、Total-Text 以及 ICDAR 2015) 并详细阐明了本方法在实验中的相关实现细节. 同时, 通过相关模块的消融实验进一步说明了边缘检测分支、浅层特征融合模块以及分支特征融合模块在 EFRNet 中所起到的作用. 最后, 在三个文本检测数据集上同近期 10 个具有代表性的最新方法的对比实验表明, EFRNet 在 CTW1500 和 Total-Text 两个数据集上具有最高

的 F1 值, 在 ICDAR 2015 数据集上与具有最高 F1 值的方法相比其差异在 0.3% 以内. 最后, 本节还对 EFRNet 的局限性进行了深入讨论.

2.1　数据集

ICDAR 2015[34] 数据集中主要包含了大量多方向文本实例以及部分小区域低分辨率的文本实例. 数据集中的文本实例均使用四边形进行标注. 该数据集中包含了 1 500 幅图像, 其中 1 000 幅图像用于训练, 500 幅图像用于测试.

Total-Text[35] 是最为常用的任意形状场景文本检测数据集之一. 同 ICDAR 2015 相比, 其数据更为复杂多样, 不仅包含了多方向文本实例, 而且还有大量任意形状的文本实例. 每个文本实例均使用 10 个顶点围成的多边形进行标注. 该数据集中的 1255 幅图像用于训练, 300 幅图像用于测试.

CTW1500[36] 是另一个常用的中英文任意形状场景文本数据集. 该数据集中还存在一些长文本实例. 每个文本实例均使用 14 个顶点围成的多边形进行标注. 该数据集中的 1 000 幅图像用于训练, 500 幅图像用于测试.

2.2　实现细节

在后续实验中, 所有的模型训练均在一台配置了一张 GeForce 3090 GPU 显卡的台式机上完成. 如无特殊说明, 各模型中均使用的骨干网络 ResNet50 是在 ImageNet 数据集[37] 上经过预训练的网络. 在 EFRNet 的训练过程中, 采用了文献 [38] 中的方法对其进行权重初始化, 并使用随机梯度下降方法进行模型优化. 在优化过程中, 权重衰减参数和非阻尼动量分别设置为 5×10^{-4} 和 0.99. 此外, 初始学习率设置为 1×10^{-3}, 在每 24k 次迭代后就将学习率修正为之前的 1/10, 并且将输出内核图数目设置为 3. 在训练阶段, 数据集 ICDAR 2015 上进行了 96k 次迭代, 在数据集 CTW1500 和 Total-Text 上进行了 72k 次迭代. 此外, 用以进行模型结果对比的指标均为常用的准确率 (Precision)、召回率 (Recall)、以及 F1 值[39].

为了避免过拟合现象, 每张图像在训练前均通过数据预处理进行了数据增强. 增强方式主要有: 1) 随机缩放为原来的 0.5 到 3 倍; 2) 随机水平翻转; 3) 在正负 10° 范围内以图像中心为原点进行随机旋转; 4) 经过前述方法处理过的图像中再随机裁剪出 640×640 像元大小的子图. 此外, 在测试阶段, 需要对图像进行等比例缩放, 以使其短边为 736 像元.

由于所有数据集均缺少文本实例边缘区域标注, 因此实验中的另一个重要步骤为通过标注多边形构建文本实例边缘区域. 在边缘区域的构建过程中, 首先对文本实例的标注多边形进行膨胀和腐蚀, 然后再使用 Vatti 算法[40] 对膨胀和腐蚀后的区域进行裁剪以得到文本边缘区域. 同时, 因为标注多边形的大小不一, 若使用同样的偏移量进行膨胀和腐蚀, 可能会导致一些较小的标注多边形边缘区域过大, 或较大的标注多边形边缘区域过小从而影响模型的有效性. 因此, 需要根据多边形的面积和周长以及缩放比例来设置偏移量 ($offset$), 具体偏移量计算公式为

$$offset = \frac{Area\,(P) \times \left(1 - r^2\right)}{Perimeter\,(P)} \tag{5}$$

其中 P 为标注多边形, $Area(P)$ 为标注多边形的面积, $Perimeter(P)$ 是标注多边形的周长, r 为控制边缘大小的参数. 在计算文本边缘区域标注时, 经过实验发现当 r 小于 0.9 时会导致过多的文本实例内部区域被包含到边缘区域中, 导致精度下降. 而当 r 接近于 1 时, 比较小的文本实例则难以形成闭合的边缘区域, 也会导致预测精度偏低. 因此, 在所有测试数据集上的 r 均统一设置为 0.9. 在计算文本内核区域标注时, 其参数 r 是根据数据集本身特点得到的, 具体计算方法可以参见文献 [22]. 根据文献 [22] 的计算结果, 在数据集 CTW1500 和 Total-Text 上, r 分别设置为 0.9, 0.8 和 0.7 以得到不同尺度上的内核图标注. 而在 ICDAR 2015 数据集上, r 则被设置为 0.8, 0.6 和 0.4 以计算不同尺度内核图的标注.

2.3　消融实验

EFRNet 在 PSENet 上所增加模块的有效性可通过在三个数据集上的消融实验进行验证. 在消融实验过程中 PSENet 与前述 EFRNet 相同模块部分均采用相同的参数设置. 以 PSENet 在三个数据集上的结果作为比较基准. 表 1 为在 PSENet 上增加不同模块组合的网络在不同数据上的预测准确率、召回率以及 F1 值.

如表 1 的第 3 行所示, 当在 PSENet 上增加浅层特征增强模块后, 除了在 ICDAR 2015 数据集上的召回率没有变化外, 所有数据集的相应指标有所增加. 这表明浅层特征增强模块中 P2 信息的二次引入所增加的浅层特征信息对于文本区域的识别有一定的帮助, 从而提高了相关评价指标. 这一实验结果也表明, 仅通过增强浅层特征所能起到的作用非常有限. 然而, 如表 1 的最后两行所示, 当采用了分支特征融合模块和边缘检测分支之后, 再增加浅层特征增强模块可以将三个数据集上的 F1 值再分

表 1 CTW1500、Total-Text 和 ICDAR 2015 数据集上的消融实验结果, 其中 P 表示准确率, R 表示召回率
Table 1 Ablation experimental on CTW1500, Total-Text and ICDAR 2015 datasets, P represents accuracy, R represents recall

浅层特征增强模块	分支特征融合模块	边缘检测分支	CTW1500			Total-Text			ICDAR 2015		
			R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)
—	—	—	77.8	83.2	80.4	78.8	88.5	83.4	75.8	84.2	79.8
√	—	—	78.1	83.9	80.8	79.3	88.6	83.7	75.8	85.3	80.3
—	—	√	83.4	85.3	84.3	81.9	87.7	84.7	83.1	87.6	85.2
√	—	√	83.8	86.3	85.0	82.0	87.9	84.9	83.3	87.8	85.5
—	√	√	84.1	86.6	85.2	83.1	88.3	85.6	84.0	88.1	86.0
√	√	√	85.9	86.8	86.3	84.0	88.9	86.4	85.7	89.5	87.6

别提升 1.1%, 0.8% 和 1.6%, 这也验证了特征增强模块的有效性.

此外, 边缘检测分支的引入可以带来较大的性能提升. 如表 1 的第 4 行所示, 当引入这一分支后, 网络在三个数据集上的召回率分别提高了 5.6%, 3.1%, 7.3%, F1 值分别提高了 3.9%、1.3% 和 2.6%, 在 CTW1500 和 ICDAR 2015 数据集的准确度也分别提高了 2.1% 和 3.4%. 仅在 Total-Text 数据集上的准确度有轻微下降 ($< 1\%$). 由此可见, 边缘检测分支对 EFRNet 的性能提升起到了非常关键的作用.

2.4 对比实验

为了评价 EFRNet 对任意形状文本的检测性能, 分别在 CTW1500、Total-Text 和 ICDAR 2015 三个数据集上将其与最近提出的一些代表性方法进行了对比. 这些方法既包括基于回归的方法, 如 TextSnake、DRRG、ContourNet、PCR、MOST 和 FCENet, 也包括基于分割的方法, 如 PSENet、PAN、DB 和 DBNet++. 除了在 TextSnake 和 DRRG 方法中使用了 Vgg16 外, 其他方法均使用 ResNet 作为骨干网络, 并且所有参数均采用其在相应论文中所述的最优参数.

表 2 为所有数据集上各个模型的召回率、准确率和 F1 值. 根据表 2 同采用 ImageNet 进行预训练的 EFRNet 比较结果可以得出, 在 CTW1500 和 Total-Text 两个任意形状文本数据集上, EFRNet 结果的召回率和准确率均较高, 特别是其 F1 值优于其他所有的方法. 以 CTW1500 数据集为例, 其召回率和 F1 值均为最高, 其准确率仅小于 DBNet++ 1.1%, 然而 DBNet++ 的召回率要比 EFRNet 低 3.1%. 因此从整体来看, 其性能仍优于其他模型. 在 Total-Text 数据集上, EFRNet 的召回率仅小于 DRRG 不到 1%, 但其准确率比 DRRG 方法要高 2.4%; EFRNet 的准确率虽然比 FCENet 低 0.4%, 但其召回率比 FCENet 要高 1.5%. 因此, 综合了准确率和召回率后, EFRNet 的 F1 值高于

表 2 CTW1500、Total-Text 和 ICDAR 2015 数据集模型性能对比
Table 2 Performance comparison on CTW1500, Total-Text and ICDAR 2015 dataset with state-of-the-art models

方法	CTW1500			Total-Text			ICDAR 2015		
	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)
TextSnake[41]	85.3	67.9	75.6	74.5	82.7	78.4	80.4	84.9	82.6
PAN++[42]	81.2	86.4	83.7	81.0	89.3	85.0	81.9	84.0	82.9
PSENet[22]	75.6	80.6	78.0	75.1	81.8	78.3	79.7	81.5	80.6
DB[30]	80.2	86.9	83.4	82.5	87.1	84.7	83.2	91.8	87.3
DRRG[21]	83.0	85.9	84.5	84.9	86.5	85.7	84.7	88.5	86.6
ContourNet[43]	84.1	83.7	83.9	83.9	86.9	85.4	86.1	87.6	86.9
FCENet[18]	83.4	87.6	85.5	82.5	89.3	85.8	82.6	90.1	86.2
MOST[11]	79.4	83.6	81.4	80.0	86.7	83.2	87.3	89.1	88.2
PCR[12]	83.3	87.2	84.7	82.0	88.5	85.2	84.1	89.6	86.7
DBNet++[15]	82.8	87.9	85.3	83.2	88.9	86.0	83.9	90.0	87.3
EFRNet (ImageNet)	85.9	86.8	86.3	84.0	88.9	86.4	85.7	89.5	87.6
EFRNet (SynthText)	85.9	86.8	86.3	84.3	89.2	86.7	86.3	89.6	87.9

所有其他方法.

在 ICDAR 2015 数据集上, 综合考虑准确率和召回率后, EFRNet 的性能要优于除 MOST 以外的其他模型. 在该数据集上 MOST 方法比 EFRNet 的 F1 值高 0.6%. 但是在 CTW1500 和 Total-Text 数据集上的 EFRNet 的 F1 值与 MOST 方法相比分别获得了 3% 到 5% 的提升. 这与 MOST 方法本身的特点密切相关. MOST 方法主要是针对具有较大长宽比的文本进行优化处理. 并且由于其基于 EAST 方法, 因此主要针对的是不同方向的四边形文本实例数据. 因此, 其局限于处理四边形标注的数据, 如 ICDAR 2015. 但是这一方法并不适合具有弯曲特征的文本实例识别. 因此, 其在另外两个数据集上的精度均较低. 此外, 当使用与 MOST 方法同样的 SynthText 数据集[44] 而不是 ImageNet 数据集对 ResNet-50 骨干网络进行预训练后, EFRNet 在 ICDAR 2015 数据集上的 F1 值达到了 87.9%, 与 MOST 相比仅有 0.3% 的误差, 这也表明两者使用同样的预训练数据集后, 预测精度相当.

在进行对比的文本检测方法中, 基于 PSENet 的 PAN++ 模型在特征提取过程中通过对 FPN 进行堆叠的方式以得到更具区分性的特征. 这一改进获得了显著的性能提升. EFRNet 则通过引入和加强边缘特征的方式对 PSENet 进行了改进. 同 PAN++ 相比, EFRNet 在三个数据集上的召回率分别提高了 4.7%, 3.0%, 3.6%, F1 值分别提高了 2.6%, 1.4% 和 4.7%; 在数据集 CTW1500 和 ICDAR 2015 上 EFRNet 的准确率分别提高了 0.4% 和 5.5%. 这一实验数据也说明, 对于文本实例有针对性地引入边缘特征更容易选择出区分性较强的特征, 从而更好地对文本实例进行识别.

图 6 展示了 EFRNet 和 PSENet 对于一些有代表性的图片的识别结果. 其中第 1 列为原始图像, 第 2 列为训练数据, 第 3 列为 PSENet 的识别结果, 第 4 列为 EFRNet 的识别结果, 最后一列为边缘预测模块的预测结果. 从图中可以观察到, 一些被 PSENet 所未检测到的、检测错误的以及边框不够准确的情况都在 EFRNet 中得到了修正. 例如, 第 1 行至第 5 行中, 很多被漏检的实例被 EFRNet 所检测到, 既包含较大的文字区域 (如第 1 行), 也包含很多较小的文本区域 (如第 2、第 4 和第 5 行), 甚至一些训练数据中漏标的文本区域也被正确识别和标注. 这些变化都在图中使用了不同颜色进行标注.

此外, PSENet 中还存在一些边框错误的例子. 以图 6 的第 7 行为例, 整个图像包含了两个相邻的文本区域, 然而 PSENet 将其识别为了一个区域, 同样的错误可在第 8 行看到. 另一个典型的边框错误的例子是图 6 的第 6 行, PSENet 将一大一小两个区域错误地合并到了一起. 同时, 还有一些文本实例虽然被正确识别, 但是其边框并不准确. 如图 6 第 1 行中的 BEEF 和 SANDWICH, 一个边框过大, 一个边框过小. 当 EFRNet 将边缘信息引入后, 这些都补正确的识别和标注.

实际上, 这些修正都同边缘预测模块的引入密切相关. 如图 6 的最后一列所示, 所有被 PSENet 漏检误检以及边框错误的区域的边缘信息都在边缘预测模块的预测结果被正确识别. 这也表示模型所提取的边缘信息是正确有效的, 并且其提取出的边缘特征在预测过程中起到了积极的作用.

2.5 本模型在利用文本内容标注上的局限性

目前 EFRNet 的目的是从图片中直接识别出文本实例的位置和覆盖区域. 如果要进一步识别文本内容, 还需要对文本框中的内容进行单独处理. 目前, 一些方法 (ABCNet V2, Mask textspotter v3[45]) 直接实现了从图像到文本内容的识别, 即端对端的识别. 此类模型大多同时使用了文本区域标注和文本内容标注以提高预测识别精度. 例如, ABCNet V2 模型不仅使用文本区域标注生成控制点进行 Bezier 曲线框生成, 还要使用目标文本内容进一步筛选出错误识别的文本区域. ABCNet V2 的消融实验表明, 设计专用网络模块以额外使用文本标注信息可提高 2% 至 6% 的预测精度. 由于使用了额外的文本内容作为训练数据, 使用这一模型可以得到较高的预测精度. 例如, 其在 Total-Text 数据集上召回率、准确率和 F1 值分别为 84.1%, 90.2% 和 87.0%, 在 ICDAR 2015 数据集上召回率、准确率和 F1 值分别为 86.0%, 90.4% 和 88.1%. 由于 ABCNet V2 使用了类似于 SynthText 的数据集, 而不是 ImageNet, 进行了骨干网络的预训练. 因此在与 ABCNet V2 进行比较时, EFRNet 使用了 SynthText 数据集进行骨干网络的预训练以获得同 ABCNet V2 尽可能相似的骨干网络. 显然, 同使用 SynthText 数据集进行预训练的 EFRNet 相比, ABCNet V2 在两个数据集上 F1 值分别高 0.3% 和 0.2%.

然而, 使用文本数据标注也会额外带来其他的问题, 例如 ABCNet V2 受限于文本的重度弯曲、阅读排列方向和不同字体的影响. 从而影响到其预测精度, 并导致其在某些数据集上的预测精度偏低. 例如, 在 Total-Text 和 ICDAR 2015 数据集上 ABCNet V2 的 F1 值仅比 EFRNet 分别高 0.3% 和 0.2%. 从实际应用角度而言, 其与 EFRNet 模

图 6 在数据集 CTW1500 上的预测结果比较 ((a) 原始图像; (b) 真实值; (c) PSENet; (d) EFRNet; (d) 边缘预测结果)

Fig. 6 Predicted results comparison on dataset CTW1500 ((a) Original image; (b) Ground truth; (c) PSENet; (d) EFRNet; (e) Predicted edge)

型的预测精度是相当的. 此外, ABCNet V2 在 CTW1500 数据集上的召回率、准确率和 F1 值仅分别为 83.8%、85.6% 和 84.7%, 使用了额外的文本内容标注的同时反而比 EFRNet 分别低 2.1%、1.2% 和 1.6%. 综上比较, 与端对端的方法相比较, EFRNet 的优势在于无需文本内容标注并且不受限于文本内容标注的数据特点, 对于各种形状文本实例均有较好的适应性. 但是, 如果能够在 EFRNet 的基础上合理引入文本内容标注, 必然还能够在特定的数据集上获得更好的预测精度.

3 结论

为了在任意形状文本检测中更充分地利用文本实例的边缘信息, 本文通过对边缘特征进行增强的方式以 PSENet 为基础提出了一种新的深度学习模型 EFRNet. 在特征提取模块, EFRNet 对 FPN 所提取的浅层特征进行增强以避免忽略一些与边缘密切相关的图像特征. 在预测模块, 该模型通过增加边缘预测分支的方式对文本实例特征和文本边缘特征进行分流, 以更好地提取出图像中文本区域的边缘特征. 此外, 还在最终预测文本框位置和形状时将文本实例特征和文本边缘特征进行融合以更好地利用所得的边缘特征. 消融实验表明, 这些模块的引入进一步提高了任意形状场景文本检测的准确度. 此外, 在 ICDAR 2015、Total-Text 以及 CTW1500 数据集上与其他 10 种代表性方法的比较结果表明, EFRNet 具有更优的识别准确率.

尽管同其他方法相比 EFRNet 可以提供更准

确的识别结果, 该模型的参数量为 PSENet 的 1.14 倍, 因此其需要更多的时间才能完成训练. 此外, 该模型还需要对训练数据中文本实例进行预处理以得到其边缘部分, 为边缘预测分支提供支撑. 在未来的工作中, 我们将进一步优化这一模型以降低其参数量, 同时设计更好的数据预处理方法. 此外, 现有模型只是隐含地使用了文本实例的边缘信息, 并未对其进行显式建模. 可以设想, 当将边缘信息引入这些网络模型后, 其模型性能必将会有进一步的提高. 然而, 不同的网络具有不同的结构和特点, 因此需要在以后的工作中, 对如何进行边缘信息的有效建模进行深入研究, 以使其适合于不同结构的深度学习模型. 最后, 一些最新的模型引入了文本内容标注以进一步提高模型的预测精度, 如 ABCNet v2, Mask textspotter v3[45] 等. 因此, 未来的工作中也可以适当考虑在引入文本标注内容的基础上避免文本内容误识别其所带来的副作用, 以进一步提高 EFRNet 的分类精度.

References

1 Lyu P Y, Liao M H, Yao C, Wu W H, Bai X. Mask TextSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 71−88

2 He T, Huang W L, Qiao Y, Yao J. Text-attentional convolutional neural network for scene text detection. *IEEE Transactions on Image Processing*, 2016, **25**(6): 2529−2541

3 Qin S Y, Manduchi R. Cascaded segmentation-detection networks for word-level text spotting. In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto, Japan: IEEE, 2017. 1275−1282

4 Cho H, Sung M, Jun B. Canny text detector: Fast and robust scene text localization algorithm. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 3566−3573

5 Tian S X, Pan Y F, Huang C, Lu S J, Yu K, Tan C L. Text flow: A unified text detection system in natural scene images. In: Proceedings of the International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 4651−4659

6 Wang Run-Min, Sang Nong, Ding Ding, Chen Jie, Ye Qi-Xiang, Gao Chang-Xin, et al. Text detection in natural scene image: A survey. *Acta Automatica Sinica*, 2018, **44**(12): 2113−2141 (王润民, 桑农, 丁丁, 陈杰, 叶齐祥, 高常鑫, 等. 自然场景图像中的文本检测综述. 自动化学报, 2018, **44**(12): 2113−2141)

7 Liu Y L, Jin L W. Deep matching prior network: Toward tighter multi-oriented text detection. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 3454−3461

8 Zhang Z, Zhang C Q, Shen W, Yao C, Liu W Y, Bai X. Multi-oriented text detection with fully convolutional networks. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 4159−4167

9 Zhong Z Y, Jin L W, Huang S P. DeepText: A new approach for text proposal generation and text detection in natural images. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, USA: IEEE, 2017. 1208−1212

10 Tian Z, Huang W L, He T, He P, Qiao Y. Detecting text in natural image with connectionist text proposal network. In: Proceedings of the 14th European Conference on Computer vision. Amsterdam, The Netherlands: Springer, 2016. 56−72

11 He M H, Liao M H, Yang Z B, Zhong H M, Tang J, Cheng W Q, et al. MOST: A multi-oriented scene text detector with localization refinement. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 8809−8818

12 Dai P W, Zhang S Y, Zhang H, Cao X C. Progressive contour regression for arbitrary-shape scene text detection. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021. 7389−7398

13 He P, Huang W L, He T, Zhu Q L, Qiao Y, Li X L. Single shot text detector with regional attention. In: Proceedings of the International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 3066−3074

14 Deng D, Liu H F, Li X L, Cai D. PixelLink: Detecting scene text via instance segmentation. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI, 2018. 6773−6780

15 Liao M H, Zou Z S, Wan Z Y, Yao C, Bai X. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(1): 919−931

16 Wang F F, Chen Y F, Wu F, Li X. TextRay: Contour-based geometric modeling for arbitrary-shaped scene text detection. In: Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM, 2020. 111−119

17 Liu Y L, Chen H, Shen C H, He T, Jin L W, Wang L W. ABCNet: Real-time scene text spotting with adaptive Bezier-curve network. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 9806−9815

18 Zhu Y Q, Chen J Y, Liang L Y, Kuang Z H, Jin L W, Zhang W. Fourier contour embedding for arbitrary-shaped text detection. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 3122−3130

19 Zhang C Q, Liang B R, Huang Z M, En M Y, Han J Y, Ding E R, et al. Look more than once: An accurate detector for text of arbitrary shapes. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 10544−10553

20 Qin X G, Zhou Y, Guo Y H, Wu D Y, Tian Z H, Jiang N, et al. Mask is all you need: Rethinking mask R-CNN for dense and arbitrary-shaped scene text detection. In: Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: ACM, 2021. 414−423

21 Zhang S X, Zhu X B, Hou J B, Liu C, Yang C, Wang H F, et al. Deep relational reasoning graph network for arbitrary shape text detection. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 9696−9705

22 Wang W H, Xie E Z, Li X, Hou W B, Lu, T, Yu G, et al. Shape robust text detection with progressive scale expansion network. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 9328−9337

23 Sheng T, Chen J, Lian Z H. CentripetalText: An efficient text instance representation for scene text detection. In: Proceedings of the 34th Advances in Neural Information Processing Systems. Cambridge, MA, USA: NIPS, 2021. 335−346

24 Liu Z C, Lin G S, Yang S, Liu F Y, Lin W S, Goh W L. Towards robust curve text detection with conditional spatial expansion. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Long

Beach, USA: IEEE, 2019. 7261−7270

25 Tian Z T, Shu M, Lyu P, Li R Y, Zhou C, Shen X Y, et al. Learning shape-aware embedding for scene text detection. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 4229−4238

26 Li J C, Lin Y, Liu R R, Ho C M, Shi H. RSCA: Real-time segmentation-based context-aware scene text detection. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPRW). Nashville, USA: IEEE, 2021. 2349−2358

27 He K M, Zhang X Y, Ren S Q, Sun J. Identity mappings in deep residual networks. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 630−645

28 Liu W, Liao S C, Ren W Q, Hu W D, Yu Y N. High-level semantic feature detection: A new perspective for pedestrian detection. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 5182−5191

29 Hu J, Shen L, Sun G, Wu E. Squeeze-and-excitation networks. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 7132−7141

30 Liao M H, Wan Z Y, Yao C, Chen K, Bai X. Real-time scene text detection with differentiable binarization. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 11474−11481

31 Zhen M M, Wang J L, Zhou L, Li S W, Shen T W, Shang J X, et al. Joint semantic segmentation and boundary detection using iterative pyramid contexts. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 13663−13672

32 Milletari F, Navab N, Ahmadi S A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the 4th International Conference on 3D Vision (3DV). Stanford, USA: IEEE, 2016. 565−571

33 Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 761−769

34 Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, et al. ICDAR 2015 competition on robust reading. In: Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR). Tunis, Tunisia: IEEE, 2015. 1156−1160

35 Ch′ng C K, Chan C S. Total-Text: A comprehensive dataset for scene text detection and recognition. In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto, Japan: IEEE, 2017. 935−942

36 Liu Y L, Jin L W, Zhang S T, Luo C J, Zhang S. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition*, 2019, **90**: 337−345

37 Deng J, Wei D, Socher R, Li J, Kai L, Li F F. ImageNet: A large-scale hierarchical image database. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE, 2009. 248−255

38 He K M, Zhang X Y, Ren S Q, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1026−1034

39 Yin X C, Yin X W, Huang K Z, Hao H W. Robust text detection in natural scene images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, **36**(5): 970−983

40 Vatti B R. A generic solution to polygon clipping. *Communications of the ACM*, 1992, **35**(7): 56−63

41 Long S B, Ruan J Q, Zhang W J, He X, Wu W H, Yao C. TextSnake: A flexible representation for detecting text of arbitrary shapes. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 19−35

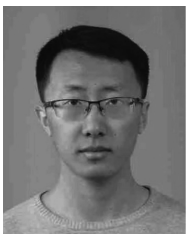
42 Wang W H, Xie E Z, Li X, Liu X B, Liang D, Yang Z B, et al. PAN++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(9): 5349−5367

43 Wang Y X, Xie H T, Zha Z J, Xing M T, Fu Z L, Zhang Y D. ContourNet: Taking a further step toward accurate arbitrary-shaped scene text detection. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 11750−11759

44 Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 2315−2324

45 Liao M H, Pang G, Huang J, Hassner T, Bai X. Mask TextSpotter v3: Segmentation proposal network for robust scene text spotting. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 706−722

白鹤翔 山西大学计算机与信息技术学院副教授. 主要研究方向为空间数据挖掘和图像处理. 本文通信作者.
E-mail: baihx@sxu.edu.cn
(**BAI He-Xiang** Associate professor at the School of Computer and Information Technology, Shanxi University. His research interest covers spatial data mining and image processing. Corresponding author of this paper.)



王浩然 山西大学计算机与信息技术学院硕士研究生. 主要研究方向为深度学习和场景文本检测.
E-mail: wanghr_sxu@163.com
(**WANG Hao-Ran** Master student at the School of Computer and Information Technology, Shanxi University. His research interest covers deep learning and scene text detection.)