计算机应用研究
Application Research of Computers
ISSN 1001-3695,CN 51-1196/TP

《计算机应用研究》网络首发论文

题目：基于视觉和文本的多模态文档图像目标检测
作者：李玉腾，史操，许灿辉，程远志
DOI：10.19734/j.issn.1001-3695.2022.08.0425
收稿日期：2022-08-04
网络首发日期：2022-11-14
引用格式：李玉腾，史操，许灿辉，程远志．基于视觉和文本的多模态文档图像目标检测[J/OL]．计算机应用研究．
https://doi.org/10.19734/j.issn.1001-3695.2022.08.0425

基于视觉和文本的多模态文档图像目标检测 *

李玉腾，史　操†，许灿辉，程远志

(青岛科技大学 信息科学技术学院，山东 青岛 266061)

摘　要：由于文档图像的布局复杂、目标对象尺寸分布不均匀，现有的检测算法很少考虑多模态信息和全局依赖关系，因此，提出了基于视觉和文本的多模态文档图像目标检测方法。首先探索多模态特征的融合策略，为利用文本特征，将图像中文本序列信息转换为二维表征，在文本特征和视觉特征初次融合之后，将其输入到骨干网络提取多尺度特征，并在提取过程中多次融入文本特征，实现多模态特征的深度融合；其次为保证小物体和大物体的检测精度，设计了一个金字塔网络，该网络的横向连接将上采样的特征图与自下而上生成的特征图在通道上连接，实现高层语义信息和低层特征信息的传播。在大型公开数据集 PubLayNet 上的实验结果表明，该方法的检测精度为 95.86%，与其他检测方法相比有更高的准确率。该方法不仅实现多模态特征的深度融合，还丰富融合的多模态特征信息，具有良好的检测性能。

关键词：多模态；文档图像；目标检测；深度学习

中图分类号：TP391.4　　doi: 10.19734/j.issn.1001-3695.2022.08.0425

Visual and textual based multimodal document object detection

Li Yuteng, Shi Cao†, Xu Canhui, Cheng Yuanzhi

(*School of Information Science & Technology, Qingdao University of Science & Technology, Qingdao Shandong* 266061, *China*)

Abstract: The layout of document images was complex and distribution of object sizes was uneven. Currently, most of detection methods ignored multimodal information and global dependencies. Therefore, this paper proposed a multimodal document object detection method based on vision and text. Firstly, this method explored the fusion strategy of multimodal features. In order to utilize textual features, this paper converted text sequence information of the image into two-dimensional representation. After the initial fusion of text features and visual features, backbone network took the fused features as input to extract multiscale features, and this paper repeatedly integrated textual features during the extraction process, so as to realize deep fusion of multimodal features. Next, to ensure the detection accuracy of small and large objects, this paper designed a pyramid network. The lateral connection could concatenate feature maps of the same spatial size from the bottom-up pathway and the top-down pathway in channel, so as to achieve the propagation between high-level semantic information and low-level feature information. The experimental results on large public dataset PubLayNet show that the detection accuracy of this method reaches 95.86%, and it has a higher accuracy than other methods. This method not only realizes the deep fusion of multimodal features, but also enriches the fused multimodal feature information, and it has good detection performance.

Key words: multimodal; document image; object detection; deep learning

0　引言

随着计算机技术的快速发展，网络上大量的信息以电子文档的形式进行传播，因此，文档成了一种重要的信息传播载体，在人们的生活中发挥着重要的作用。文档图像目标检测在识别文档图像的信息中起着至关重要的作用，目标检测的准确度对于数字化系统的整体成效影响很大，如光学字符识别(OCR)[1]准确性及其提取信息的有用性等。

文档图像目标检测又称页面分割或布局分析，旨在将文档图像自动识别为独立结构和逻辑单元，如文本、表格和图形。对于不同的目标区域有着不同的处理策略。由于文档图像组件的复杂性和多样性，这项工作具有挑战性。自动识别文档的整体结构，具有显著的商业价值和学术价值。国内外有很多研究学者提出了各种用于文档图像检测或分割的方法[2~21]。

文档图像的目标检测方法可以分为传统的方法和深度学习的方法。传统的方法[11~13]对于手工绘制的特征依赖程度高，相关的程序算法复杂，并且难以识别出复杂布局。与传统的方法相比，深度学习的方法具有更强的表征提取与学习能力，更适用于文档图像的目标检测任务。为将文档图像布局分析任务应用于移动端和云服务端，Oliveira 等人[2]提出了一种利用卷积神经网络的快速一维文档检测模型，该模型具有更快的执行时间和更紧凑的数据使用量，并显著提高整体性能。Li 等人[3]提出了一种跨域文档图像目标检测模型，并且设计了三个特征对齐模块用于解决区域偏移的问题。文献[4]提出

收稿日期：2022-08-04；**修回日期：**2022-10-05　　**基金项目：**国家自然科学基金资助项目(61806107，61702135)

作者简介：李玉腾(1997-)，男，山东滕州人，硕士研究生，主要研究方向为计算机视觉和图像处理；史操(1981-)，男(通信作者)，湖南永顺人，讲师，硕导，博士，主要研究方向为深度学习和图像处理(caoshi@yeah.net)；许灿辉(1981-)，女，湖南浏阳人，讲师，硕导，博士，主要研究方向为人工智能、图像处理；程远志(1976-)，男，山东蓬莱人，教授，博导，博士，主要研究方向为医学图像处理.

了一种基于自适应平滑算法的模型，利用 K 均值聚类分析得到合适的阈值，进而实现对文档界面的分割，最后通过识别器区分文本与非文本区域。文献[5]提出了一种采用多特征融合的模型，该模型通过融合来自不同卷积核的特征，并将其输入串并行空间金字塔中实现对特征的进一步优化。为了精确地检测文档图像中的表格，Agarwal 等人[10]提出了利用双主干的深度网络模型，同时在骨干网络中加入可变形卷积，并在较高的 IoU 阈值下获得较高的检测结果。以上的方法虽然表现良好的性能，但是在处理文档图像的特征时局限于视觉特征，却忽略了文档图像中丰富的文本特征，造成信息的浪费。

因此，多模态的方法被应用到文档图像相关的任务[16~19]。Soto 等人[16]将文档图像中的上下文信息融入到 Faster R-CNN[22]网络中，以提升网络检测文档目标区域的性能。Yang 等人[17]通过创建文本嵌入图的方法来利用文本特征，并将其融入端对端的多模态全卷积网络中以提升文档图像的分割精度。Zhang 等人[18]提出了一种基于双流的多模态网络，该网络融合视觉特征、文本特征和组件关系，并在文档的布局分析中表现出良好的性能。同时，2021 年，国际文档分析与识别会议(ICDAR)组织了科学文献解析(SLP)比赛任务 A，其中，入围的方案绝大多数是基于多模态的方法，证明了多模态方法的有效性。相比于基于视觉的方法，多模态的方法实现视觉信息和文本信息的充分利用，有着很大发展空间和应用前景。但是，现有的多模态方法在多模态特征融合方式未实现特征之间的深度融合，以及在后续的处理中没有进一步丰富融合的多模态特征表征信息。

因此，针对上述问题，本文提出了基于视觉和文本的多模态文档图像目标检测方法。为实现不同模态特征的深度融合，本文利用卷积神经网络(如 ResNet[23])将不同的模态空间映射到共享语义子空间，从而融合不同模态的特征，并保留丰富的特征信息；为保证小物体和大物体的检测精度，增强网络的多模态表征能力，设计了一个金字塔网络，该网络将不同尺度的特征在通道上进行连接，使低层特征信息中融入高层的语义信息，进行多模态信息的传递；为了进一步丰富多模态融合特征的信息，特征金字塔网络(FPN[24])被引入到该网络。此外，在处理文本信息的过程中，优化了 PubLayNet[25]数据集的适配，包括插入半结构元素和扩展 Ground Truth 注释，并构建出了层级关系数据集。

1　本文算法

1.1　网络结构

本节详细介绍所提基于视觉和文本的多模态文档图像目标检测网络结构。该网络以 Faster R-CNN[22]为基础，融入了视觉特征和文本特征，旨在利用不同模态之间的补充信息，并将 ResNet-101[23]作为骨干网络用以实现多模态特征的深度融合和保留更多的特征表征信息。深度融合的多模态特征在经过特征增强模块后，不同尺度的特征信息得到传递，使得高层特征信息融入到低层特征中，特征表征得到进一步的增强，从而提升了网络模型的检测精度及鲁棒性。

该网络的结构如图 1 所示，由文本特征提取模块、特征融合模块、特征增强模块、特征金字塔网络(FPN[24])和区域生成网络(RPN)五个模块组成，其中文本特征提取模块主要由 4 个不同的卷积层和正则化层组成，是进行文本特征提取操作的基本组件。特征融合模块以 ResNet[23]网络为主，通过其强大的特征表示能力实现多模态特征的深度融合，并保留丰富的特征信息，从而使二者的信息得到充分利用。特征增强模块主要由卷积层和上采样层组成，外观上与特征金字塔网络相似，主要实现不同尺度特征的表征信息在通道上传递，使得低层特征也包含丰富的语义信息。特征金字塔网络(FPN[24])将相邻层的特征图变换为相同的尺寸，然后对它们执行元素级别加法操作(对应位置元素相加)，目的是为了将高层特征中的强语义信息传递到低层特征中，实现低层次高分辨率信息和高层次强语义信息的结合，从而提升检测性能。区域生成网络(RPN)主要由卷积层、中间层、分类层和回归层组成，其本质是基于滑动窗口和锚框机制在特征图上对目标区域进行分类和回归，并产生一系列的候选区域。

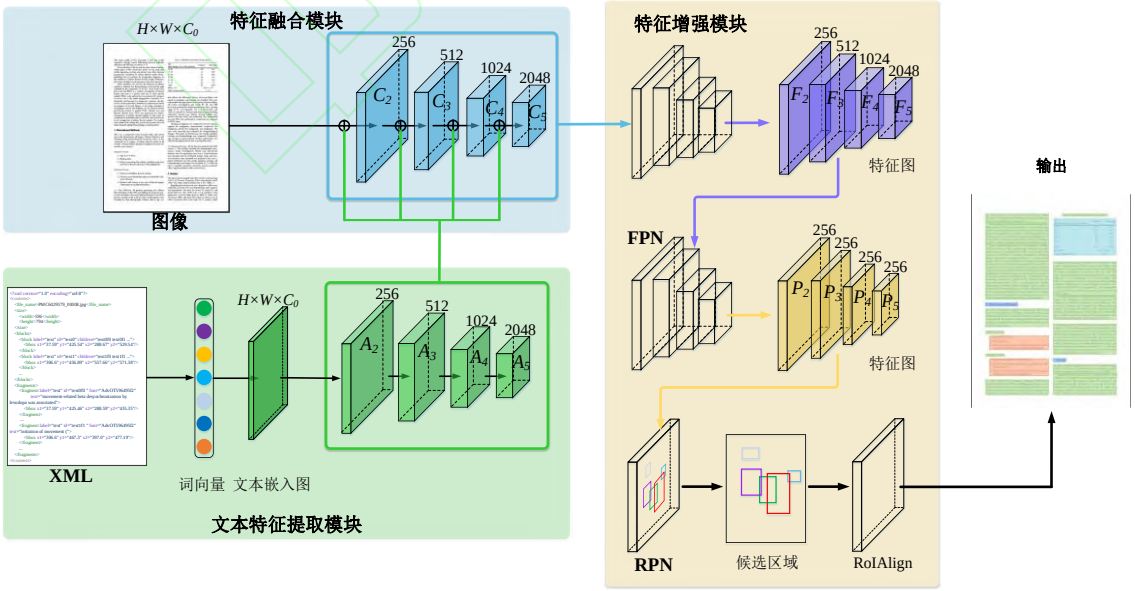

图 1　基于视觉和文本的多模态文档图像目标检测模型架构

Fig. 1　The architecture of visual and textual based multimodal document object detection

1.2　文本特征提取模块

研究表明，文本特征能够区分视觉相似的区域，并对整体检测精度作出贡献[17,18]。所以，为了利用文本信息，需要将文本序列由一维转换为二维，本文构建了行级别的文本嵌入图。本文参考句子转换(sentence-transformers[26])的思想，训练文档图像中的文本信息，得到相应的词向量。用于生成词

向量的文本信息是通过解析文档图像对应的 PDF 文件得到的。考虑到模型的开销和文本信息的复用性，将解析得到的文本信息保存到 XML 文件中，训练得到的词向量也被保存到相应的文件中。解析 PDF 文件得到的信息包含行级别的文本信息和段落级别的文本信息，通过一个比较算法对二者坐标信息进行一个比较，由此构建二者的包含关系。同时，也优化了 PubLayNet[25]数据集的适配，构建了层级关系数据集。

在图像 $x\in\mathbb{R}^{H\times W\times C_0}$ 上有一组行的信息 I_l，其中 H 和 W 分别是图像的高和宽，C_0 表示图像的通道维度。行的信息 $I_l=\{(l_k,b_k)\,|\,k=1,\cdots,n\}$，其中 n 为行的总数目。l_k 是第 k 行的文本信息，$b_k=(x_{ul}^k,y_{ul}^k,x_{br}^k,y_{br}^k)$ 是第 k 行文本的坐标框，(x_{ul}^k,y_{ul}^k) 和 (x_{br}^k,y_{br}^k) 分别代表左上角和右下角的坐标。文本嵌入图 $T\in\mathbb{R}^{H\times W\times C_0}$ 的公式定义如下：

$$T_{i,j}=\begin{cases}E(l_k) & if\,(i,j)\in b_k\\ 0 & otherwise\end{cases}\tag{1}$$

其中，$E(l_k)$是 $l_k\rightarrow\mathbb{R}^{C_0}$ 即一个将文本信息转换为词向量的映射函数；0 表示零向量，其对应于没有文本的区域。每个 b_k 中的所有像素共享相同的行级别词向量。并且文本嵌入图 T 与图像 x 有着相同的空间大小和通道数目。

在得到文本嵌入图 T 后，为实现多模态特征的融合，需要将其输入到文本特征提取模块提取文本特征。由于卷积神经网络善于学习深层特征，同时也可以将文本特征映射到与视觉特征相同的子空间，所以其被用于从文本嵌入图中提取多尺度文本特征图 A_i，其中 A_i 与骨干网络提取的特征有相同的空间大小和通道维度。文本特征提取模块是由 4 个卷积块组成，每个卷积块中包含卷积层和正则化层，其中卷积核的大小为 3×3，步长为 2，边缘填充为 1。

1.3　特征融合模块

由于卷积神经网络具有良好的特征提取能力和学习能力，本文中采用 ResNet[23]作为骨干网络提取特征，并利用其将不同模态空间映射到共享语义子空间中，从而深度融合多模态特征。

来自不同模态的特征表征在确定不同的目标时具有重要作用，视觉信息可以容易地识别较大的目标区域，文本信息对于区分视觉上相似的区域具有重要意义[18]。为充分利用不同模态的补充信息，实现多模态特征的融合就显得十分重要。目前大多数模型[17,18]在通道上叠加多模态特征，以此来实现不同模态信息的融合，但是不同模态特征的占比往往对模型性能起着至关重要的作用。不同于以上的融合策略，本文提出了将文本特征和视觉特征相加，然后将融合后的多模态特征输入骨干网络提取多尺度特征，并在提取的过程中多次融入文本特征，以丰富特征信息并实现多模态特征的深度融合。如图 2 所示，首先从文档图像中提取视觉特征 V_2，然后将文本特征 A_2 与之融合得到多模态特征 C_2。将 C_2 输入到骨干网络中得到特征 V_3，并与文本特征 A_3 融合得到 C_3，通过加入文本特征可以使特征图保留更多的信息。依此类推，C_4 与 C_5 的产生与之相似。特征 C_i 的产生定义如下：

$$\begin{cases}C_i=V_i+A_i & i=2,3,4,5\\ V_2=Conv(x) & \\ V_{i+1}=Conv(C_i) & i=2,3,4\end{cases}\tag{2}$$

其中，$Conv$ 是一个卷积层，文本特征被定义 A_i，x 代表图像。

通过将文本特征融入低层和高层的特征中，使得特征图中的信息更加丰富。卷积神经网络可以将不同的模态空间映射到共享语义子空间中，从而融合不同模态的特征。视觉信息包含较高层次的特征表征，文本信息包含较低层次的特征表征，通过融合二者的补充信息，使得融合后的特征信息比之前单一模态的更加丰富。

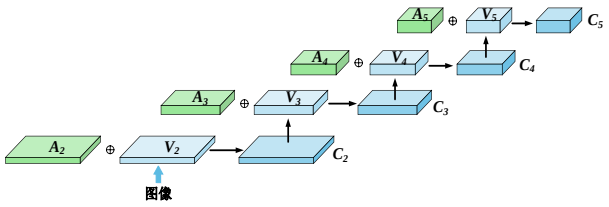

图 2　特征融合模块

Fig. 2　Feature fusion module

1.4　特征增强模块

研究表明，特征金字塔结构可以实现不同尺度特征信息的传递，进而丰富特征信息。因此，本文设计了特征增强模块，外观上与特征金字塔网络(FPN[24])相似。

如图 3 所示，特征增强模块的构造包含自下而上路径、自上而下的路径和横向连接。横向连接将上采样的特征图与自下而上生成的相同大小的特征图在通道上连接，实现高层语义信息和低层特征信息的传播，进而增强特征表征。该过程从 C_5 开始，通过自顶向下和横向连接策略逐步整合层次特征。F_5 由 C_5 直接产生。F_4 的产生首先需要 F_5 经过 1×1 卷积层，使其通道维度降低到原来的一半与 C_4 的通道维度保持一致，之后对其进行上采样使之宽高变为原来的 2 倍。然后，将上采样后的特征与 C_4 在通道维度上连接，再经过 3×3 的卷积层降低其通道维度，使连接后的特征图的通道维度与 C_4 的相同，经过以上操作之后就得到了 F_4。F_3 和 F_2 的产生过程也与之相似。F_i 的生成过程被定义如下：

$$\begin{cases}F_5=C_5\\ F_i=Conv_2(Concat(C_i,U(Conv_1(F_{i+1}))))\quad i=2,3,4\end{cases}\tag{3}$$

其中，$Concat$ 表示通道维度连接操作，U 是上采样函数。$Conv_1$ 和 $Conv_2$ 表示不同的卷积层，卷积核大小分别为 1×1 和 3×3。因此，融合后的多模态特征通过横向连接将上采样的特征和自上而下路径的特征在通道上叠加得到丰富，使得高层的语义信息融入低层特征信息。通过特征增强模块，一组特征表征 F_i 被产生。在得到特征 F_i 后，将其输入特征金字塔网络中进行下一步的操作。

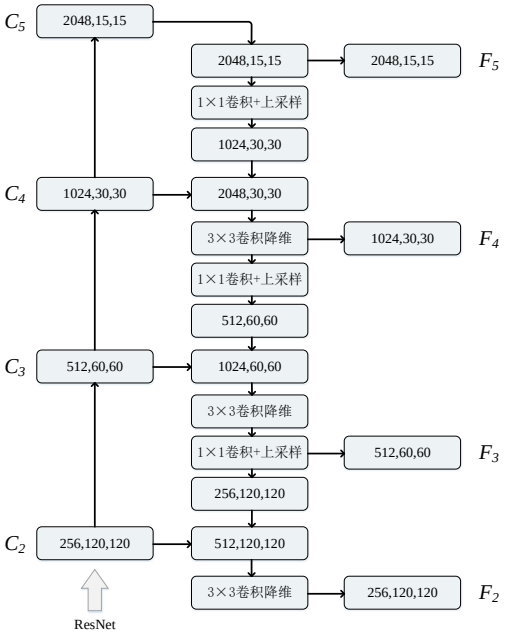

图 3　特征增强模块

Fig. 3　Feature enhancement module

2　实验结果与分析

2.1　数据集

PubLayNet[25]是一个包含超过 36 万张文档图像的大型公开数据集，其标注信息包括边界框标注和多边形分割标注。该数据集被用于 2021 年国际文档分析与识别会议科学文献解析比赛任务 A(ICDAR-SLP-TASK A)。标注文件遵循 MS COCO 对象检测任务的 json 格式。此数据集包含研究论文和文章的图片以及页面上的各种元素的注释。其包含 5 个类别，即 text、title、list、table 和 figure。数据集的类别分布如表 1 所示。

表 1　PubLayNet 数据集中类别分布

Tab. 1　Distribution of categories among publaynet dataset.

类别	数目
text	2523005
title	665269
list	89911
table	112301
figure	118238

2.2　评价指标及实验环境参数

本文实验使用平均精度(AP)、均值平均精度(mAP)和召回率(Recall)作为文档图像目标检测任务的评价标准。AP、mAP 和 Recall 的越高，则算法的性能越好。

本文实验在 NVIDIA TITAN Xp 服务器上搭建 PyTorch 框架下进行的，CUDA 版本为 10.2。PubLayNet[25]数据集的训练周期为 6，初始学习率为 0.001，动量为 0.9，权重衰减为 0.0001，其他对比方法的相关参数配置与之相同。此外，本文方法中的特征金字塔网络(FPN[24])和区域生成网络(RPN)的参数配置与原论文一致。

2.3　对比实验结果与分析

为了证明本文提出网络模型的有效性和合理性，使用上述评价标准，将其与当前主流的检测方法进行比较，包括目前主流的目标检测算法，Faster R-CNN[22]、Mask R-CNN[27]和 ATSS[28]，也包含用于文档图像检测相关的方法，CDeCNet[10]、VSR[18]、DiT[29]和 LayoutLMv3[30]。其中，Faster R-CNN[22]和 Mask R-CNN[27]网络是基于 R-CNN 的两阶段检测网络，ATSS[28]是一阶段的检测网络，CDeCNet[10]是专门用于文档图像表格检测的网络，VSR[18]是融合了视觉特征、文本特征和组件关系的多模态网络，DiT[29]是基于 Transformer 的自监督预训练的文档检测网络，LayoutLMv3[30]是融合视觉和文本的 Transformer 多模态网络。VSR 中使用 ResNeXt-101[31]作为特征提取的骨干网络，为保证公平，除 DiT[29]和 LayoutLMv3[30]以 Transformer 作为骨干网络外，其他方法和本文方法均使用 ResNeXt-101[31]作为骨干网络。此外，本文方法通过加入特征融合模块和特征增强模块分别实现多模态特征的深度融合和传递不同级别特征信息进而丰富特征表征，从而提升网络模型在文档图像数据集上的检测性能。

在相同的参数配置和训练周期的条件下，不同网络模型在 PubLayNet[25]数据集上的检测结果如表 2 所示，其中，VSR[18]的实验结果为其论文所提供，其未提供相应的 Recall 数据。由表 2 可知，本文方法有着优异的性能，在大多数类别表现优于其他检测方法，并且 mAP 和 Recall 均达到最高值分别为 95.86%和 96.91%，这是由于文本信息对于提升文档图像的检测精度起着重要作用，本文的多模态特征融合策略实现不同模态信息之间的深度融合，以及本文设计的金字塔网络使低层特征信息中融入高层的语义信息，进行多模态信息的传递，保证小物体和大物体检测精度。Faster R-CNN[22]、Mask R-CNN[27]和 ATSS[28]在 PubLayNet[25]数据集上的 mAP 都超过 90%，同时 Recall 也超过 92%，本文方法在 mAP 和 Recall 比他们高了大约 3.26%和 3.55%，说明了仅依赖于视觉特征对于提升文档图像的检测性能是有限的。此外，在类别 list、table 和 figure 上本文的方法是高于其他方法的。其中，本文方法在 table 上的 AP 值超过用于检测文档表格的 CDeCNet[10]，比其提高了 1.76%，同时也比 VSR[18]高出 1.06%，这是由于本文方法加入了文本特征使得 table 区域保留更加丰富的特征信息，并且利用 ResNeXt[31]实现多模态特征的深度融合以及金字塔网络实现高层语义信息在通道上传递到低层特征。在 text 和 title 上，VSR[18]高于本文方法，这是由于 VSR[18]中不仅使用了组件关系，而且在创建文本嵌入图时使用了字符级别和行级别的文本信息，不同级别的文本信息对于不同的类别有着不同的功效[18]。本文方法在没有融入组件关系的条件下，多数类别的 AP 值和 mAP 超过 VSR，说明本文多模态深度特征融合策略的有效性及通过金字塔网络对融合多模态特征处理的合理性。DiT[29]和 LayoutLMv3[30]在 PubLayNet[25]数据集上的 mAP 分别达到 94.92%和 95.07%，Recall 分别达到 96.20%和 96.40%，高于 Faster R-CNN[22]、Mask R-CNN[27]、ATSS[28]和 CDeCNet[10]，这表明基于 Transformer 的网络在文档图像目标检测任务中有着良好的性能，本文方法

与他们相比在 mAP 上分别提高 0.94%和 0.79%，并且在绝大多数类别上本文方法是高于他们的，这说明本文的多模态网络与基于 Transformer 的文档检测网络相比性能更好，这是由于本文实现了多模态特征的深度融合，并对融合后的多模态特征进行特征表征增强，使得特征信息更加丰富，进而提升网络的检测性能。

表 2　在 PubLayNet 数据集的性能比较

Tab. 2　Performance comparisons on publaynet dataset.

方法	AP/%					mAP/%	Recall/%
	text	title	list	table	figure		
Faster R-CNN[22](2016)	91.03	82.66	88.27	95.46	93.73	90.23	92.19
Mask R-CNN[27](2018)	91.71	84.11	88.71	96.11	95.01	91.13	92.42
ATSS[28](2020)	93.65	83.73	93.81	96.51	95.29	92.60	93.36
CDeCNet[10](2021)	91.30	84.20	89.70	96.70	90.50	90.48	91.97
VSR[18](2021)	**96.70**	**93.10**	94.70	97.40	96.40	95.66	-
DiT[29](2022)	94.38	89.64	95.67	97.76	97.16	94.92	96.20
LayoutLMv3[30](2022)	94.47	90.57	95.45	97.88	97.00	95.07	96.40
本文	96.22	89.86	**95.71**	**98.46**	**99.05**	**95.86**	**96.91**

图 4 给出了本文方法与其他方法在 PubLayNet[25]数据集上 4 组检测结果的示意图。第一列为标注了 Ground Truth(GT)的样图，第二列为本文方法的检测结果，第三列为 Faster R-CNN[22]的检测结果，第四列为 CDeCNet[10]的检测结果。为了便于观测检测结果，本文在检测框的内部进行了颜色填充，其中淡绿色表示 text 区域，粉红色表示 title 区域，橘黄色表示 list 区域，灰色表示 table 区域和淡蓝色表示 figure 区域。

从图 4(a-3)和(a-4)中可以看出，Faster R-CNN[22]网络虽然识别出了 title 类别，同时也将其识别为 text 类别，造成了误检，而 CDeCNet[10]网络将 title 类别错误识别为 text 类别，并且在 figure 区域检测不精确，使其检测精度下降，相比于二者，从图 4(a-2)中可以得出，本文方法可以准确识别出每个目标区域，这是因为本文方法加入文本特征，其能够区分

相似的目标区域。此外，在图 4(d-3)和(d-4)中，Faster R-CNN[22]不精确地 text 检测框覆盖了多个目标区域，CDeCNet[10]在 text 和 figure 上的检测框也不准确，造成他们检测精度的降低。对比图 4(b-2)和(b-4)，CDeCNet[10]在识别 list 时丢失一部分目标区域，而本文的方法却能够精准地检测出 list 区域，是因为基于视觉的方法在提取特征时，容易丢失其前方的数字或小黑点的特征，在加入文本特征后，这一区域的特征能够被增强。同时，在图 4(c-3)和(c-4)中，Faster R-CNN[22]在 figure 上的检测区域不精确，而 CDeCNet[10]虽然识别出 figure，同时也将其错误的检测为 list 和 text，而在图 4(c-2)中，本文方法能够准确地识别目标区域，这是因为本文不仅实现多模态特征的融合，而且通过金字塔网络实现高层语义信息在通道上传递到低层特征中，进而提升目标区域的检测精度。

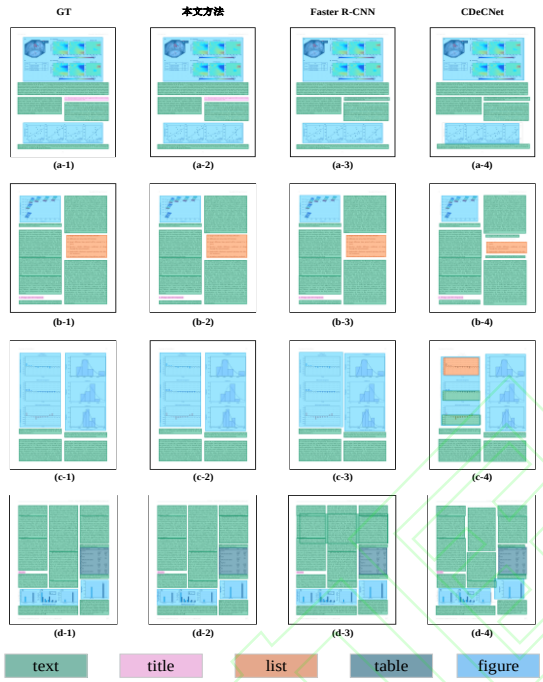

图 4　不同方法在 PubLayNet 数据集的检测样例

Fig. 4　Detection samples of different methods on publaynet dataset

2.4　消融实验

在 PubLayNet[25]数据集上的消融实验结果如表 3 所示，记录的是基线网络 Faster R-CNN[22]在加入不同模块后的实验结果。网络训练步长的设定需要平衡训练时间和检测精度，同时，学习率的设定与优化器以及数据和任务有关，合理地设定学习率可以使模型较快地收敛至最优点。因此，综合考虑，本文将消融实验中模型的训练步长设为 90k，初始学习率设为 0.0025。

从表 3 的实验结果可以得出，在不加入任何优化策略的网络性能是最低的，mAP 和 Recall 分别为 89.20%和 92.22%，与加了特征融合模块和特征增强模块后的网络相比，分别低了大约 3%和 2.29%。在加入特征融合模块之后，网络模型的 mAP 从 89.20% 提高到 92.03%，Recall 从 92.22%提升到 94.51%。从实验结果上看，list 类别的 AP 值比基线网络提升了 5.39%，figure 类别的 AP 值提高了 5.93%。这证明文本信息在提高网络检测精准度方面起到重要作用，同时进一步证明多模态特征融合策略的有效性。通过融合不同模态特征，可以充分利用不同模态之间的补充信息。文本特征能够增强较小区域的特征，比如 list 前面的数字或小黑点在视觉特征提取的过程中容易被丢失，而在加入文本特征之后，该区域的特征信息得到进一步的丰富。当在基线网络中加入特征增强模块后，该网络模型的 mAP 和 Recall 分别提升到了 89.49%和 92.73%，相对于基线网络有较小的提升，这是由于基线网络中只包含视觉特征，使得特征增强模块无法生成更加丰富的特征表征。在基线网络中加入特征融合模块和特征增强模块后，该网络模型的 mAP 提高到 92.22%，相比于基线网络提高了 3%左右。实验结果证明在融合多模态特征后，特征表征得到了进一步的丰富，融合的多模态特征在进入特征增强模块后，通过在通道上连接不同级别的特征，实现了不同级别特征信息的传递，使得低层特征包含了高层的语义信息。

表 3　在 PubLayNet 数据集上的消融实验

Tab. 3　Ablation studies on publaynet dataset.

基线网络	特征融合模块	特征增强模块	AP/% text	title	list	table	figure	mAP /%	Recall /%
Faster R-CNN			91.83	82.53	84.40	95.68	91.58	89.20	92.22
	√		92.81	83.46	89.79	96.60	97.51	92.03	94.35
		√	91.75	82.94	85.01	95.82	91.92	89.49	92.73
	√	√	**92.87**	**83.58**	**90.29**	**96.70**	**97.66**	**92.22**	**94.51**

3　结束语

针对文档图像中组件的复杂多样造成其布局分析难，本文提出了一种基于视觉和文本的多模态检测网络。该网络利用骨干网络强大的特征提取能力和特征表示能力，实现多模态特征的深度融合，从而充分利用不同模态之间的补充信息。融合后的多模态特征进入特征增强模块，使得不同级别的特征信息在通道维度上传递，使低层的特征信息中包含高层的语义信息，从而增强多模态特征表征。实验结果表明，本文所提的方法优于目前主流的方法，能够进一步丰富网络中的特征表征，加入文本特征能增强较小区域的特征，从而提升文档图像目标检测的准确性，减小误差，缩减检测时间。未来研究中，可以将 XML 文件中的层级信息融入到网络，以进一步提升网络的整体性能。

参考文献：

[1] 白翔, 杨明锟, 石葆光, 等. 基于深度学习的场景文字检测与识别 [J]. 中国科学: 信息科学, 2018, 48 (05): 531-544. (Bai Xiang, Yang Mingkun, Shi Baoguang, *et al.* Deep learning for scene text detection and recognition [J]. Scientia Sinica: Informationis, 2018, 48 (05): 531-544.)

[2] Oliveira D A B, Viana M P. Fast CNN-based document layout analysis [C]// Proc of IEEE International Conference on Computer Vision Workshops. Piscataway, NJ: IEEE Press, 2017: 1173-1180.

[3] Li Kai, Wigington C, Tensmeyer C, *et al.* Cross-domain document object detection: benchmark suite and method [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 12912-12921.

[4] 陈园园, 王维兰, 刘华明, 等. 基于自适应游程平滑算法的藏文文档图像版面分割与描述 [J]. 激光与光电子学进展, 2021, 58 (14): 172-179. (Chen Yuanyuan, Wang Weilan, Liu Huaming, *et al.* Layout segmentation and description of Tibetan document images based on adaptive run length smoothing algorithm [J]. Laser & Optoelectronics Progress, 2021, 58 (14): 172-179.)

[5] 应自炉, 赵毅鸿, 宣晨, 等. 多特征融合的文档图像版面分析 [J]. 中国图像图形学报, 2020, 25 (02): 311-320. (Ying Zilu, Zhao Yihong, Xuan Chen, *et al.* Layout analysis of document images based on multifeature fusion [J]. Journal of Image and Graphics, 2020, 25 (02):

311-320.)

[6] 姚佳. 基于深度学习的复杂文档版面分割算法研究 [D]. 北京: 北京交通大学, 2021. (Yao Jia. Complex document layout segmentation based on deep learning [D]. Beijing: Beijing Jiaotong University, 2021.)

[7] Xu Canhui, Shi Cao, Bi Hengyue, *et al.* A page object detection method based on mask R-CNN [J]. IEEE Access, 2021, 9: 143448-143457.

[8] Xu Yiheng, Li Minghao, Cui Lei, *et al.* Layoutlm: pre-training of text and layout for document image understanding [C]// Proc of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2020: 1192-1200.

[9] Xu Canhui, Shi Cao, Chen Yinong. End-to-end dilated convolution network for document image semantic segmentation [J]. Journal of Central South University, 2021, 28 (6): 1765-1774.

[10] Agarwal M, Mondal A, Jawahar C V. Cdec-net: composite deformable cascade network for table detection in document images [C]// Proc of the 25th International Conference on Pattern Recognition. Piscataway, NJ: IEEE Press, 2021: 9491-9498.

[11] Amin A, Shiu R. Page segmentation and classification utilizing bottom-up approach [J]. International Journal of Image and Graphics, 2001, 1 (02): 345-361.

[12] Ha J, Haralick R M, Phillips I T. Recursive XY cut using bounding boxes of connected components [C]// Proc of the 3rd International Conference on Document Analysis and Recognition. Piscataway, NJ: IEEE Press, 1995, 2: 952-955.

[13] Shilman M, Liang P, Viola P. Learning nongenerative grammatical models for document analysis [C]// Proc of the tenth IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2005, 2: 962-969.

[14] Xu Yiheng, Lyu Tengchao, Cui Lei, *et al.* Layoutxlm: multimodal pre-training for multilingual visually-rich document understanding [EB/OL]. (2021-09-09) . https://arxiv. org/abs/2104. 08836v3.

[15] Garncarek Ł, Powalski R, Stanisławek T, *et al.* LAMBERT: layout-aware language modeling for information extraction [C]// Proc of International Conference on Document Analysis and Recognition. Cham: Springer Press, 2021: 532-547.

[16] Soto C, Yoo S. Visual detection with context for document layout analysis [C]// Proc of Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. [S. l.] : Association for Computational Linguistics Press, 2019: 3464-3470.

[17] Yang Xiao, Yumer E, Asente P, *et al.* Learning to extract semantic structure from documents using multimodal fully convolutional neural networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 5315-5324.

[18] Zhang Peng, Li Can, Qiao Liang, *et al.* VSR: a unified framework for document layout analysis combining vision, semantics and relations [C]// Proc of International Conference on Document Analysis and Recognition. Cham: Springer Press, 2021: 115-130.

[19] Barman R, Ehrmann M, Clematide S, *et al.* Combining visual and textual features for semantic segmentation of historical newspapers [EB/OL]. (2020-12-14) . https://arxiv. org/abs/2002. 06144v4.

[20] Shi Cao, Xu Canhui, Bi Hengyue, *et al.* Lateral feature enhancement network for page object detection [J]. IEEE Trans on Instrumentation and Measurement, 2022, 71: 1-10.

[21] Bi Hengyue, Xu Canhui, Shi Cao, *et al.* SRRV: a novel document object detector based on spatial-related relation [EB/OL]. IEEE Trans on Multimedia, 2022. (2022-04-07) [2022-09-23]. doi: 10. 1109/TMM. 2022. 3165717.

[22] Ren Shaoqing, He Kaiming, Girshick R, *et al.* Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2017, 39 (6): 1137-1149.

[23] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Deep residual learning for image recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 770-778.

[24] Lin Tsungyi, Dollár P, Girshick R, *et al.* Feature pyramid networks for object detection [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 2117-2125.

[25] Zhong Xu, Tang Jianbin, Yepes A J. Publaynet: largest dataset ever for document layout analysis [C]// Proc of International Conference on Document Analysis and Recognition. Piscataway, NJ: IEEE Press, 2019: 1015-1022.

[26] Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks [EB/OL]. (2019-8-27) . https://arxiv. org/abs/1908. 10084.

[27] He Kaiming, Gkioxari G, Dollár P, *et al.* Mask R-CNN [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2018, 42 (2): 386-397.

[28] Zhang Shifeng, Chi Cheng, Yao Yongqiang, *et al.* Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 9759-9768.

[29] Li Junlong, Xu Yiheng, Lyu Tengchao, *et al.* Dit: self-supervised pre-training for document image transformer [EB/OL]. (2022-07-19) . https://arxiv. org/abs/2203. 02378.

[30] Huang Yupan, Lyu Tengchao, Cui Lei, *et al.* LayoutLMv3: pre-training for document AI with unified text and image masking [EB/OL]. (2022-7-19) . https://arxiv. org/abs/2204. 08387.

[31] Xie Saining, Girshick R, Dollár P, *et al.* Aggregated residual transformations for deep neural networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 1492-1500.