基于增强FPN的Vision Transformer在文档布局分析任务中的应用研究

张法,李艳红*,吴龙雨,龙焓
(中南民族大学 计算机学院,湖北 武汉 430074)

摘 要 相较于传统基于卷积神经网络(CNN)的方法,基于视觉变换器(Vision Transformer)的文档布局分析模型通过多模态预训练机制,能够为下游任务提供鲁棒的语义与视觉表征.当前,多尺度特征生成模块以及跨分辨率特征融合过程,容易引发类别属性与边界细节的丢失,进而引发类别混淆与边界模糊问题.针对这一瓶颈,提出了局部特征增强生成(LFEG)和全局到局部特征增强融合(GLEF)技术,用于构建增强的特征金字塔网络(FPN)结构,以实现新型的特征优化.其中,局部特征增强生成模块优化4个分辨率修改模块,生成多尺度特征,全局特征增强融合则对传统自顶向下的融合方式进行了优化.实验结果表明:所提出的增强FPN结构能够有效提升多尺度特征图的类别一致性与边界清晰度,为基于Vision Transformer的文档布局分析的精度优化提供关键技术支撑.

关键词 视觉变换器;特征金字塔网络;特征融合;文档布局分析
中图分类号 TP399 **文献标志码** A **文章编号** 1672-4321(2026)04-0548-11
doi:10.20056/j.cnki.ZNMDZK.20260708

Application research of Vision Transformer with enhanced FPN in document layout analysis tasks

Zhang Fa, Li Yanhong*, Wu Longyu, Long Han
(College of Computer Science, South-Central Minzu University, Wuhan 430074, Hubei China)

Abstract Compared with traditional methods based on Convolutional Neural Networks (CNN), the document layout analysis model based on Vision Transformer can provide robust semantic and visual representations for downstream tasks through multi-modal pre-training mechanisms. However, the current multi-scale feature generation module and cross-resolution feature fusion process are prone to causing the loss of category attributes and boundary details, which in turn leads to issues such as category confusion and blurred boundaries.To address this bottleneck, Local Feature Enhancement Generation (LFEG) and Global-to-Local Feature Enhancement Fusion (GLEF) techniques are proposed to construct an enhanced Feature Pyramid Network (FPN) structure for achieving novel feature optimization. Specifically, the LFEG module optimizes four resolution modification modules to generate the multi-scale feature, while the GLEF module optimizes the traditional top-down fusion approach. Experimental results demonstrate that the proposed enhanced FPN structure can effectively improve the category consistency and boundary clarity of multi-scale feature maps, providing key technical support for optimizing the accuracy of document layout analysis based on Vision Transformer.

Keywords Vision Transformer; FPN; feature fusion; document layout analysis

文档是用于记录信息、传递思想的重要工具. 同时,伴随着大型语言模型和检索增强生成系统的

收稿日期 2025-06-03
* **通信作者** 李艳红(1973-),女,教授,博士,研究方向:时空数据处理,网络大数据,DB与AI融合,E-mail: liyanhong@mail.scuec.edu.cn
基金项目 湖北省自然科学基金资助项目(2017CFB135);中央高校基本科研业务费专项资金资助项目(CZY23019);网络创新及应用型人才课程实践教学研究项目(2019年第一批)

进步，自动读取、解析和理解高质量文档的技术变得越来越重要.然而，由于本身的特点，文档通常具有复杂的结构.

文档布局分析（Document Layout Analysis，DLA）是一项旨在识别和理解文档中不同元素及其空间与逻辑关系的技术任务，其核心目标是将文档内容分割为具有语义意义的区域（如文本块、表格、图像等，如图 1 所示），并分析这些区域的布局结构和层次关系.其最终目的是将非结构化的文档（如扫描件、PDF）转化为结构化数据，便于后续大型语言模型或检索增强生成等系统理解.

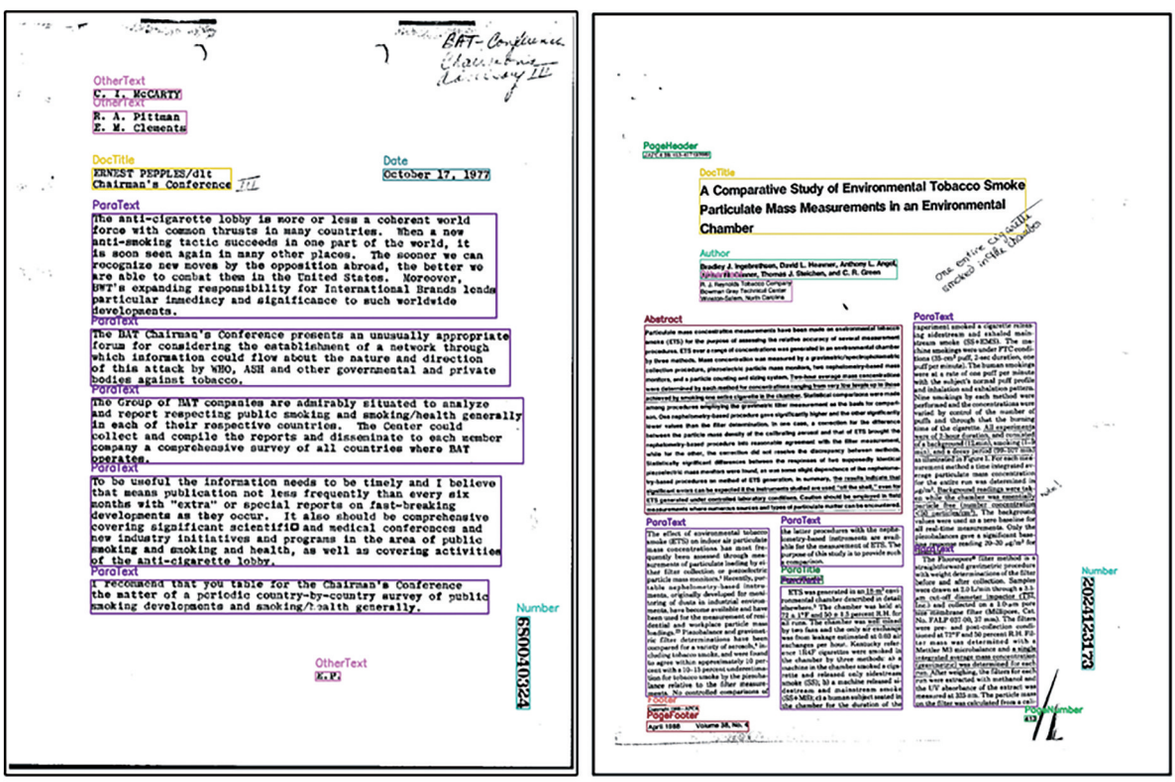


图 1　文档布局分析任务可视化

Fig. 1　Visual representation of document layout analysis task

然而，在实际场景中开展文档布局分析任务面临着许多挑战：不同类型的文档具有不同的格式和排版规则，这会增加元素识别的难度；文档可能存在复杂的布局，如多层嵌套、图文混排等，可能导致区域分割不准确；低质量的图像，其文本和图像元素难以清晰辨识，从而影响布局分析的结果.

在面对众多挑战时，需要模型提供更加精细的特征来表示复杂的文档页面.目前，研究人员为了充分利用文档中的丰富资源来表示复杂的文档页面，提出了多种方法，这些方法可以分为基于单模态特征和基于多模态特征的方法.通常的单模态方法对于复杂的页面表现不佳，为了能够实现更高的准确率，在过去的几年中，DLA 算法逐渐倾向于基于 CNN 融合[1]和基于 Vision Tranformer 融合[2-5]的多模态方法.

为了比较两种方式在布局分析任务中的性能差异，本文总结了几种重要模型的实验结果，如表 1 所示.相较于基于 CNN 的方式，基于 Vision Transformer[6]的文档布局分析方法展现出了显著优势.得益于其架构中的注意力机制，Vision Transformer 能够在大量未标注的数据集上预训练，这种预训练模型能够高效且鲁棒地学习文档中复杂的多模态信息.因此当基于 Vision Transformer 的方法作为文档布局分析任务的主干网络时，其蕴含的丰富语义和视觉上下文信息能够显著提升模型的性能.

为了能够进一步提升模型的检测精度，本文发现，基于 Vision Transformer 的现代文档布局分析模型通常使用 4 个分辨率修改模块来生成多尺度特征，然后通过特征金字塔网络（FPN）[7]融合不同尺度的特征图.但是，4 个分辨率修改模块在生成多尺度特征和后续的融合阶段中，通常会导致部分特征信息的丢失.在文档图像中，一些重要的边界信息

表1 DLA任务中基于CNN与基于Transformer方式结果对比

Tab. 1 Comparison of CNN-based andTransformer-based approaches in DLA Tasks

model	PubLayNet	DocBank	D^4LA
ResNeXt-101	93.5	77.4	65.1
DiT-Base	94.5	79.6	67.7
LayoutLMv3-Base	95.1	78.3	60.5
VGT	96.2	84.1	68.8
DoPTA-HR	94.9	–	70.7

和类别特征能够较大程度地影响检测的准确性.具体来说,在原始特征金字塔的自顶向下方式中,往往通过上采样方式扩大高层特征图,再与低层特征相融.但是,简单的上采样粗特征融合会严重导致类别不一致、特征边界模糊问题,使得特征图丢失大量至关重要的信息.

在Chen提供的方法中[8],使用自适应低通滤波器(ALPF)生成器、偏移生成器、自适应高通滤波器(AHPF)生成器3个组件能够有效地解决类别内不一致和边界位移问题.所以,为了进一步提高文档的特征表示,结合CHEN提供的方法,本文在分辨率修改模块扩大特征图时,恰当地加入融合策略,使得各个尺度的特征图更清晰地表示出各个类别和边界;同时,文档中每个元素的大小有着不小的差异,比如单个标题与大段的文段或者整页的图像、表格.为了使各个尺度的特征图能够适应这些差异,本文在特征金字塔网络的融合阶段使用不同的参数,控制模型在不同大小的特征图中有不同的感受野,这样模型能够在不同的尺度下捕捉图像的特征,使模型能够有效地检测不同尺度的目标.本文的贡献如下:

(1)本文提出使用局部特征增强生成来缓解生成FPN输入时的特征图信息丢失;

(2)在FPN融合阶段设计融合器有不同参数,使整个结构能感受到不同尺度的元素,增强了模型检测不同尺度元素的能力的同时,使得特征图更加清晰地表示类别特征和边界;

(3)融入本文的基于Vision Transformer文档布局方法后,现有模型的检测精度均有不同程度的提高.

1 相关工作

1.1 文档布局分析方法

文档布局分析方法已被广泛研究.早期基于规则的方法[9-13]一般只能用来处理版面相对固定并清晰的文档图像.随着文档越来越复杂,传统方法面临难以设计和容易出错的问题.随后,深度学习被引入文档布局分析任务中.基于深度学习的方法大致可以分为基于单模态特征的方法和基于多模态特征的方法.对于基于单模态特征,早期工作直接使用卷积神经网络(CNN)[14]编码的视觉特征进行布局单元检测,并已被证明是有效的.Faster-RCNN架构[15]的提出成为文档目标检测领域的重要里程碑,其创新的区域建议网络(RPN)机制显著提升了页面分割任务的处理效率与定位精度.随后发展的Mask-RCNN[16]架构通过集成掩膜预测模块实现了创新性改进,在应对复杂版面布局(如具有密集多栏排版结构的报纸版面)时展现出卓越的像素级分割精度,为文档图像解析技术树立了新的性能标杆.Saha团队[17]在ICDAR2017 POD基准测试中构建了基于迁移学习的Faster-RCNN改进框架,专门用于检测数学公式、表格和图表等细粒度页面元素.YOLO[18]是一种在物体检测方法中领先的算法,在自然图像上的准确性和速度方面都表现出色.为了进一步提高YOLO在文档图像处理上的性能,Zhao等[19]将网络调整为适合文档数据的特定特征,同时引入了全局到局部可控感受野模块(GL-CRM),使模型能够有效地检测不同尺度的目标.

然而,仅依靠视觉特征可能会忽略文本语义和全局结构信息,导致对复杂排版(如多栏混杂、嵌套表格)的解析能力有限.为了克服仅依靠视觉特征的不足,研究人员开始探索同时考虑视觉和语义特征的方法,这就是基于多模态特征的方法.Zhang等[1]结合计算机视觉和自然语言处理的方法,为文档布局分析提出了一个统一的框架VSR.VSR通过文档图像引入视觉,通过文本嵌入映射引入语义;然后,使用双流网络提取特定于模态的视觉和语义特征,这些特征被自适应融合以充分利用互补信息.为了融合两个模态,李玉腾[20]提出一种单粒级文本特征与视觉特征的融合方法,即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,并在提取过程中多次融入文本特征,以丰富特征信息,实现多模态特征的深度融合,该方式把文本特征同时融入到低层和高层特征中,可以使特征图中的信息更加丰富.两个模态的特征信息互补,有效利用了不同模态的特征.该方式被证明性能大大优于以前的单模态模型.

近年来,文档预训练已取得显著成功.其中

Document Image Transformer（DiT）[2]使用图像进行预训练，不依赖于任何人工标记的文档图像，仅从大规模未标记数据中学习每个文档图像内的全局补丁关系；同时，它在 DLA 上也获得了良好的性能．但是，该方式仅依赖图像的像素信息，仍然只能利用视觉信息进行文档布局分析的微调．为了更大程度地利用文档中有价值的信息，LayoutLMv3[3]首次在没有 CNN 的情况下统一文本和图像两个不同的模态，降低了时间和空间复杂度，构建了一个简单而高效的统一多模态框架；同时，针对文本和图像模态的预训练目标存在嵌入不对齐以及模型参数量大等问题，温绍杰等[21]提出了一种基于 Transformer 的多模态级联文档布局分析网络（MCOD-Net），使用了 3 种预训练目标即掩码语言建模、掩码图像建模和词块对齐，以促进模型在文本和图像模态上的表征学习能力．Da 等[4]提出了 VGT，一种双流多模态 Vision Grid Transformer 文档布局分析模型，其中 Vision Transformer 用于视觉流，Grid Transformer 被用来直接建模 2D 语言信息．为了增强 token 级和段级别的语义理解，提出了用于 GiT 预训练的掩码网格语言建模（MGLM）和段语言建模（SLM）目标；同时，GiT 将语言信息渲染为 2D 网格特征，这些特征在空间位置上与图像信息自然对齐．为了更好地融合图像特征与文本特征，DoPTA[5]提出了一种新颖的图像-文本对齐技术，专门用于利用文档图像中的文本信息来提高视觉任务的性能．最终，两个模态信息的互补提升了 DLA 任务性能．

1.2　特征金字塔网络（FPN）

文档元素的大小差异显著——从简短的标题到长篇的段落，甚至整页的图像和表格．准确检测出这些不同大小的组件要求网络学习丰富且语义上有意义的多尺度特征表示．一种广泛采用的解决方案是使用特征金字塔网络（FPN）[7]，该网络是专门为解决目标检测中的尺度变化挑战而设计的．FPN 于 2017 年提出，它以最小的额外计算成本增强了多尺度特征表示．这是通过引入一个带有横向连接的、自顶向下的路径来实现的，该路径融合了来自网络不同深度的特征．具体来说，较高层次（语义丰富但分辨率较低）的特征图通过最近邻插值等方法进行上采样，然后通过横向连接与较低层次（高分辨率但语义较少）的特征进行合并．这些横向连接使用卷积层来匹配较低层次特征的通道维度与上采样后的较高层次特征的通道维度，从而实现逐元素的融合．通过结合浅层网络的空间细节信息和深层网络的丰富语义信息，FPN 能够有效地捕捉多尺度的目标特征，从而显著提升检测性能——特别是在处理文档图像中的小型元素和可变大小元素时．

2　方法论

2.1　局部特征增强生成

目前大多数 Vision Transformer 文档布局分析模型基于 VIT-Base[6]、采用了 12 层的 Transformer 编码器构建，使用第[3,5,7,11] Transformer 块输出构建特征金字塔网络．将第 3 层、第 5 层分别使用反卷积进行两次上采样，一次上采样，使得对应特征图分辨率扩大为原始编码块输出的四倍、两倍，用来当作特征金字塔网络的下两层．使第 7 层分辨率保持不变，第 11 层使用最大池化将对应特征图分辨率缩小两倍，用来当作特征金字塔网络的上两层．通过观察变化的特征图，发现该过程通常会导致部分特征信息的丢失．如图 2 所示，首先观察子图 2(a)中的作者信息栏，进行特征提取后能够获取到具有代表性的特征图（图 2(b)），发现该部分特征图能够清晰地表示作者信息栏板块．但后续使用原始上采样进行操作后，观察子图 2(c)可以看出信息丢失十分严重，已经无法清晰反应出作者信息栏板块．在文档图像中，一些重要的边界信息和类别特征，往往能够较大程度地影响检测的准确性．

因此，本文在 Vision Transformer 中重构了特征金字塔网络的生成过程，如图 3 所示．

具体来说，首先重新处理第[3,5,7,11] Transformer 编码器输出．为了能够最大程度地减少特征图在变化过程中的信息丢失，本文提出局部特征增强生成模块（LFEG）．局部特征增强生成模块的核心是使用融合策略，恰当的融合方式，能够同时捕捉到低级高分辨率特征和高级初略特征特有的细节和语义信息．对于融合模块，本文采用了频率感知特征融合模块，该模块通过自适应地平滑高层特征、在附近重采样类别一致的特征以及增强低层特征的高频分量，有效解决信息丢失的问题．对于需要扩大分辨率的特征图，首先使用反卷积进行初始扩张．将原始分辨率特征图送入自适应低通滤波器生成器（ALPF Generator），用于平滑高级特征，减轻特征不一致；然后，对该特征图进行上采样，并使其与扩展后的特征图对齐．将扩大分辨率的特征图送入自

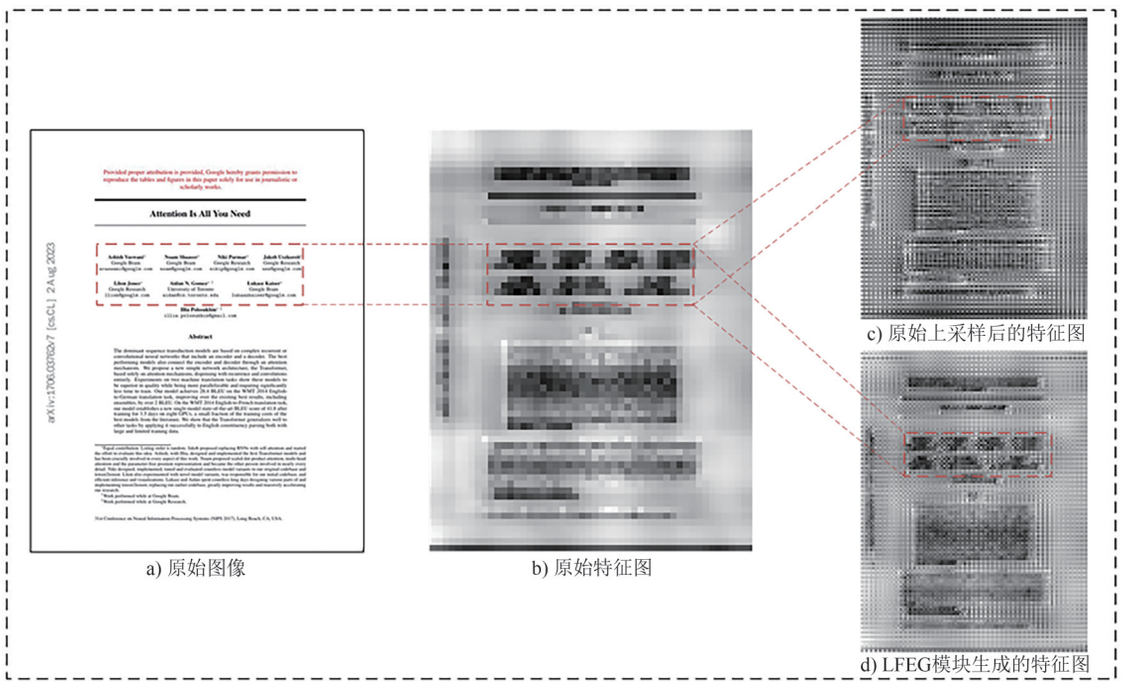


图 2　原始特征图、上采样后的特征图、局部增强后的特征图可视化

Fig. 2　Visualization of original feature maps, upsampled feature maps, and locally enhanced feature maps

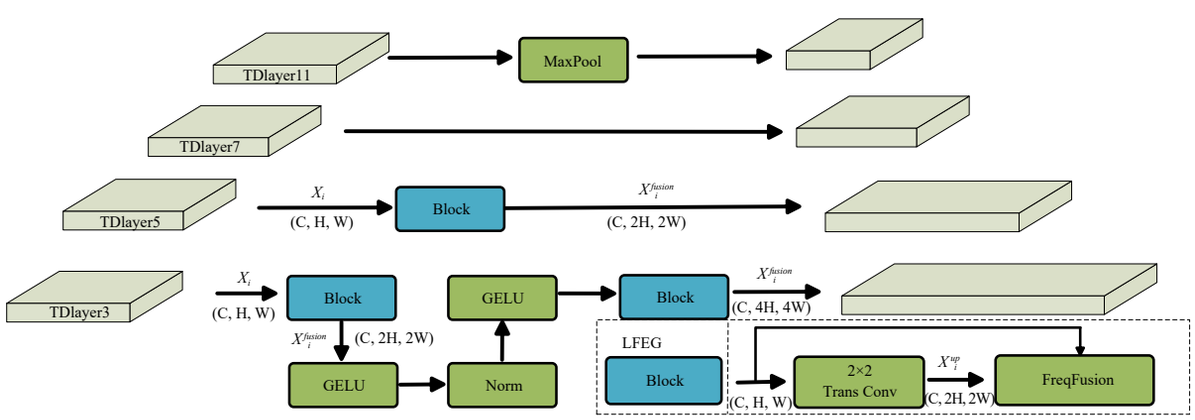


图 3　使用局部特征增强生成模块生成特征金字塔网络

Fig. 3　Generating a feature Pyramid network using a local feature enhancement generation module

适应高通滤波器生成器(AHPF Generator),以增强详细边界信息;之后,将其与发送到自适应高通滤波器前的特征图进行融和;最后,将通过自适应低通滤波器生成器和自适应高通滤波器处理后的特征进行相加,得到最终的融合特征.通过观察图 2 中使用局部特征增强模块后的特征图 d),可以看出,使用该模块后,特征图分辨率的扩大能够最大程度地保留原始特征信息.这使得本文可以构建出综合的金字塔网络输入特征集合 X,它包含了一组保留了大量原始特征信息的输入特征,具体见下式:

$$X=\{X_{\text{TDlayer3-up4}\times},X_{\text{TDlayer5-up2}\times},X_{\text{TDlayer7}},X_{\text{TDlayer11-dw2}\times}\}. \tag{1}$$

以下讨论第[3, 5, 7, 11] Transformer 编码器输出.

第 3 层 Transformer 块输出处理:使用该层构建特征金字塔网络的最底层,需要把特征图分辨率扩大四倍,即对原始特征图进行两次反卷积上采样操作,所以需要进行两次融合,具体实现如下:

$$X_{\text{up2}\times}=\text{ConvT2D}(X_{\text{TDlayer3}}), \tag{2}$$

$$\tilde{X}=F^{\text{UP}}(F^{\text{LP}}(X_{\text{TDlayer3}})),\tilde{X}_{\text{up}}=F^{\text{HP}}(X_{\text{up2}\times})+X_{\text{up2}\times}, \tag{3}$$

$$X_{\text{fusion}}=\text{GELU}(\text{BN}(\tilde{X}+\tilde{X}_{\text{up}})), \tag{4}$$

$$X_{\text{up4}\times}=\text{ConvT2D}(X_{\text{fusion}}), \tag{5}$$

$$X_{\text{TDlayer3-up4}\times}=\{F^{\text{UP}}(F^{\text{LP}}(X_{\text{fusion}}))\}+\{F^{\text{HP}}(X_{\text{up4}\times})+X_{\text{up4}\times}\}. \tag{6}$$

第 5 层 Transformer 块输出处理:该层只需要扩大两倍,只需进行一次融合,具体实现如下:

$$X_{\mathrm{up2\times}} = \mathrm{ConvT2D}(X_{\mathrm{TDlayer5}}), \tag{7}$$

$$X_{\mathrm{TDlayer5-up2\times}} = \{F^{\mathrm{UP}}(F^{\mathrm{LP}}(X_{\mathrm{TDlayer5}}))\} + \{F^{\mathrm{HP}}(X_{\mathrm{up2\times}}) + X_{\mathrm{up2\times}}\}. \tag{8}$$

第 7 层 Transformer 块输出处理：该层需要保持分辨率，无需进一步处理，与 Transformer 块输出一致：

$$X_{\mathrm{TDlayer7}} = X_{\mathrm{TDlayer7}}. \tag{9}$$

第 11 层 Transformer 块输出处理：该层构建特征金字塔最顶层，为最小分辨率，因此需要使用最大池化将分辨率进行缩小，具体实现如下：

$$X_{\mathrm{TDlayer11-dw2\times}} = \mathrm{MaxPool}(X_{\mathrm{TDlayer11}}), \tag{10}$$

其中，F^{UP} 表示上采样，F^{LP} 表示由 ALPF 生成器通过 3 × 3 卷积层和 Softmax 层对初步融合特征 Z^l 进行预测，得到空间变化的低通滤波器，其过程表示为：

$$\bar{V}^l = \mathrm{Conv}_{3\times3}(Z^l), \tag{11}$$

$$\bar{W}_{i,j}^{l,p,q} = \mathrm{Softmax}(\bar{V}_{i,j}^{l}), \tag{12}$$

式中 $\bar{V}^l \in R^{\bar{k}^2\times2H\times2W}$ 表示空间变化的滤波器权重；$\bar{k}$ 表示低通滤波器的内核大小，l 表示层级索引，q 表示位置索引，(i,j) 表示坐标信息．其中，F^{HP} 表示由 AHPF 生成器通过 3 × 3 卷积层、Softmax 层和滤波器反转操作对初步融合特征 Z^l 进行预测，得到空间变化的高通滤波器，其过程表示为：

$$\hat{V}^l = \mathrm{Conv}_{3\times3}(Z^l), \tag{13}$$

$$\hat{W}_{i,j}^{l,p,q} = E - \mathrm{Softmax}(\hat{V}_{i,j}^{l}), \tag{14}$$

式中 $\hat{V}^l \in R^{\hat{k}^2\times H\times W}$ 表示高通滤波器，E 表示基础权重．

2.2　全局到局部特征增强融合

在原始特征金字塔的自顶向下方式中，往往通过上采样方式扩大高层特征图再与低层特征相融合，可以表述为：

$$X_{\mathrm{fusion}} = F^{\mathrm{UP}}(X_j) + X_i, \tag{15}$$

其中 $X_i \in R^{C\times2H\times2W}$，$X_j \in R^{C\times H\times W}$，分别为主干网第 i 层和第 j 层的输出特征，在 VIT 中，能够确保它们有相同的通道数．

尽管这类方法在学术界和工业界得到广泛采用，但其采用的基础特征融合范式仍存在两个制约密集预测性能的关键缺陷：类别内不一致和边界位移．标准融合在纠正这些不一致的特征方面存在不足，并且其中的简单插值甚至可能通过将单个不一致的特征上采样到多个不一致的像素而加剧问题．已有研究进一步揭示[22-24]，直接插值方法容易引发预测边界的过度平滑效应，应用到文档图像时这种伪影会显著降低几何敏感区域（如文字行边缘、表格交线）的定位精度．此外，当前方法对浅层特征中蕴含的高分辨率边界线索缺乏有效的跨层级整合机制，导致细粒度结构信息的利用率不足．

图 4 展示了一个文档图像中不同元素在尺寸上相差较大的案例，所以不同于自然图像，处理文档图像时金字塔融合阶段需要处理这种尺寸差异问题．



图 4　文档图像中的元素大小差异对比

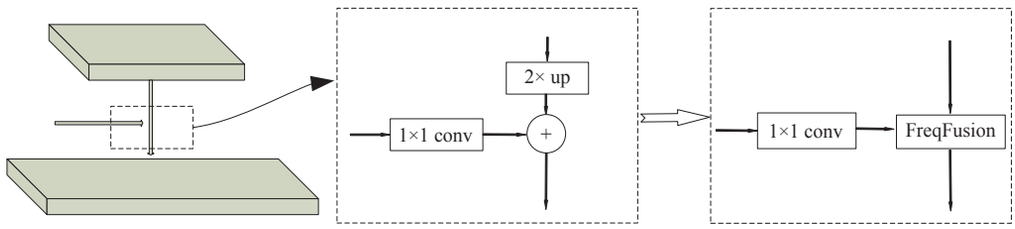
Fig. 4　Comparison of element size differences in document images

针对上述两个问题，本文首先把原始特征金字塔的自顶向下方式中的标准融合范式替换为频率感知特征融合模块．具体来说，当上层特征处理结束后不使用传统的上采样操作，直接与低层经过侧向卷积后的特征一起送入频率感知特征融合模块进行处理，如图 5 所示，这在图 1 的中间层特征图可视化中可以证明是有效的；其次为了更好地处理文档中的尺度变化挑战，在融合模块的编码器中，本文定义了不同的核大小来控制编码器的感受野大小，$k = [k_1, k_2, ..., k_n]$．这种方法可以获得一组不同编码器表示：

$$\mathrm{Encoder} = \{E_{(k_1)}, E_{(k_2)}, ..., E_{(k_n)}\}, \tag{16}$$

其中 E 表示编码器，k_i 表示不同大小核．

为了使用不同参数的编码器处理元素变化，本文使用了全局到局部的设计思想[19]，如图 6 所示，该方法把整个文档页面定义为 3 个级别：全局级别-块



a) 原始FPN融合 b) FreqFusionFPN融合

图5 原始FPN融合结构与FreqFusionFPN融合结构

Fig. 5 Original FPN fusion structure and FreqFusionFPN fusion structure

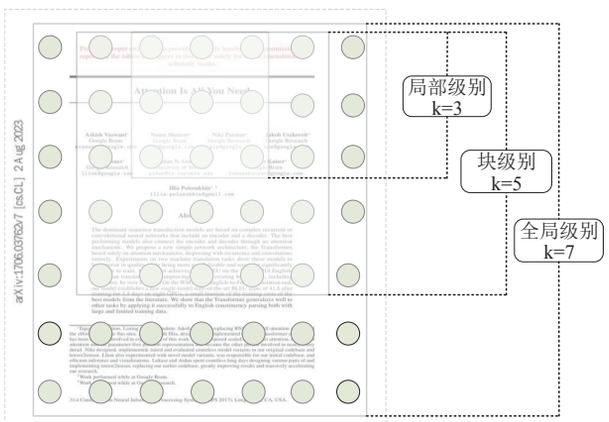


图6 全局到局部设计思想示意图

Fig. 6 Illustration of the global-to-local design philosophy

级别-局部级别.该架构能够从全局上下文到子块区域再到局部语义信息感知不同级别的特征表示.

对于全局级别,本文使用特征金字塔网络上层表示,该层捕获全局语义.所以在上层的融合模块中使用$k=7$参数控制编码器,对于低分辨率特征图,大内核通常能够覆盖整个特征图,因此能够做到全局信息聚合,保留整个页面元素细节.

对于块级别,本文使用特征金字塔网络中间层表示,同时使用$k=5$参数控制编码器,适当的核大小能够感知类似文档子块的中尺度元素.

对于局部级别,本文使用特征金字塔网络底层表示,对于高分辨率特征图,包含丰富的细节信息.较小的卷积核能更精细地处理局部区域,避免过度平滑,因此使用$k=3$的参数控制编码器捕获更多纹理细节.

通过这种方法能使整个金字塔特征网络结构具有从全局上下文到局部语义信息的层次感知能力.本文把上述整个过程定义为全局到局部特征增强融合(GLEF),具体表现为:

$$X=\mathrm{FreqFusion}_{(k_i)}(X_i,\mathrm{Conv}(X_j)), \tag{17}$$

以自顶向下为例,其中X_i,X_j分别表示特征金字塔网络中前一层输出的特征和后一层经过横向卷积输出的特征.k表示对应级别的参数值.FreqFusion表示频率感知特征融合模块.

3 实验

3.1 数据集和评估标准

为了评估本文方法的有效性,在目前最复杂的公开数据集D^{4}LA(https://www.modelscope.cn/datasets/iic/D4LA)上进行实验,该数据集由VGT作者公布,包含11092张带噪声图像,涵盖了12种不同的文档类型,包含了27个类别,训练集包含8868张图像,测试集包含2224张图像.为了评估指标,本文使用了COCO[25]风格的标准平均精度(mAP)指标来评估,采用交并比(IoU)阈值区间来评估检测精度,该区间覆盖从0.50到0.95的范围,以0.05为间隔逐步递增.然后计算这些IoU层级上平均精度的算术平均值,确保与其他模型的对比具有规范性和可比性.变量及公式定义如下:

(1) True Positives(TP):文档元素中的正例被正确地划分为正例的样本数量;

(2) False Positives(FP):文档元素中的负例被错误地划分为正例的样本数量;

(3) False Negatives(FN):文档元素中的正例被错误地划分为负例的样本数量;

(4) True Negatives(TN):文档元素中的负例被正确地划分为负例的样本数量.

精确率P表示被预测为正样本中实际为正样本的概率,计算公式为:

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \tag{18}$$

召回率R表示预测结果中正样本被正确预测的概率,计算公式为:

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{19}$$

AP指处于不同置信度时,不同的P和R组成的点连起来所构成曲线围成的面积;AP50指交并比为0.5时,由不同的P和R组成的点所构成曲线围成的面积.

mAP 就是文档所有元素类别的 AP 之和除以类别数目.mAP 作为衡量文档布局分析任务的精度的重要指标，计算公式为：

$$\mathrm{mAP}=\frac{1}{|\mathrm{classes}|}\sum_{c\in\mathrm{classes}}\frac{|\mathrm{TP}_c|}{|\mathrm{FP}_c|+|\mathrm{TP}_c|}. \tag{20}$$

3.2　实施细节

本文整个实验基于 Detectron2 框架[26]，使用 Cascade R-CNN[27]作为检测器，采用 AdamW 算法进行优化.为了测验本文方法的有效性，分别使用目前现有开源的基于 Vison Tranformer 的文档布局方法作为主干网络，比如 VGT[4]、DiT-Base[2]、LayoutLMv3-Base[3]，沿用其基本配置：采用 12 层 Transformer 编码器，12 头注意力，输入图像尺寸为 224×224，编码维度 D=768，patch 大小为 16，然后替换其中的 FPN 为本文的方式.在整个 D^{4}LA 数据集下进行的实验，采用批量大小为 6，学习率设置为 0.0001，进行 35000 次迭代.实验环境如表 2 所示.

表 2　实验环境

Tab. 2　Experimental environment

名称	配置信息
操作系统	Ubuntu 22.04
框架	Pytroch 2.0.1+Cuda 11.7
CPU	Intel(R) Xeon(R) Silver 4108
GPU	RTX A6000

4　主要结果

表 3 和图 7 总结了本文在 VGT、DiT-Base、

表 3　D^{4}LA 验证集上 mAP @ IOU [0.50:0.95]

Tab. 3　mAP @ IOU [0.50:0.95] on D^{4}LA validation set

Model	DocTitle	ListText	LetterHead	Question	RegionList	TableName
LayoutLMv3-Base	67.2	56.0	80.4	38.3	72.1	63.1
DiT-Base	73.1	70.6	82.2	55.0	80.1	68.4
VGT	72.6	71.3	82.3	63.9	80.2	68.4
LayoutLMv3-Base-EnFPN	68.8	62.2	78.5	48.3	72.9	67.7
DiT-Base-EnFPN	73.9	70.7	81.0	66.7	80.1	67.8
VGT-EnFPN	72.7	71.4	84.5	72.1	80.7	67.6
Model	FigureName	Footer	Number	ParaTitle	RegionTitle	LetterDear
LayoutLMv3-Base	34.8	73.9	82.1	54.2	59.6	69.6
DiT-Base	51.8	81.2	83.2	63.2	67.5	74.5
VGT	46.6	79.7	83.2	63.0	67.2	76.7
LayoutLMv3-Base-EnFPN	41.4	75.3	82.8	58.0	60.5	70.6
DiT-Base-EnFPN	52.1	79.6	83.6	62.9	66.6	73.7
VGT-EnFPN	54.6	80.6	83.0	64.1	67.0	75.7
Model	OtherText	Abstract	Table	Equation	PageHeader	Catalog
LayoutLMv3-Base	50.9	64.9	82.1	12.1	52.8	48.6
DiT-Base	59.2	73.8	86.2	9.2	56.5	44.8
VGT	60.0	80.4	86.0	19.9	56.9	40.9
LayoutLMv3-Base-EnFPN	54.2	67.1	82.5	5.6	56.2	38.9
DiT-Base-EnFPN	60.2	73.5	87.3	23.5	57.2	41.5
VGT-EnFPN	60.9	81.4	87.7	19.1	57.1	56.7
Model	ParaText	Date	LetterSign	RegionKV	Author	Figure
LayoutLMv3-Base	80.4	62.6	60.4	59.2	56.6	72.5
DiT-Base	86.4	69.7	71.6	68.8	66.0	77.2
VGT	86.2	71.3	75.5	70.1	67.6	76.7
LayoutLMv3-Base-EnFPN	82.8	64.3	62.1	61.9	60.5	74.5
DiT-Base-EnFPN	86.4	68.9	72.4	68.5	66.5	77.3
VGT-EnFPN	85.7	70.2	74.7	69.9	66.7	76.8
Model	Reference	PageFooter	PageNumber	mAP		
LayoutLMv3-Base	77.4	61.6	51.3	60.9		
DiT-Base	83.4	65.5	58.3	67.7		
VGT	85.6	66.5	58.7	68.8		
LayoutLMv3-Base-EnFPN	78.6	64.3	54.8	62.8		
DiT-Base-EnFPN	83.5	66.8	58.2	68.5		
VGT-EnFPN	84.6	66.7	58.9	70.0		

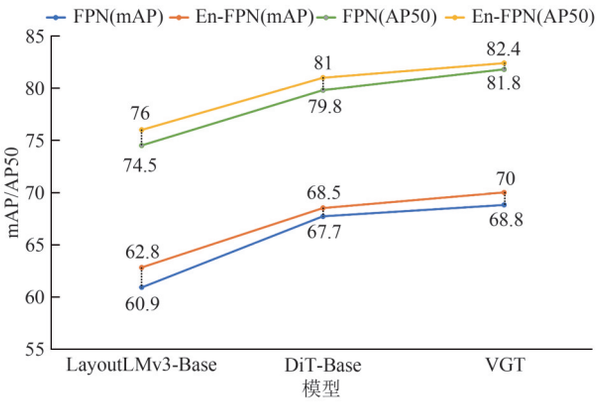


图7 各模型FPN和增强FPN中mAP和AP50变化趋势

Fig. 7 The change trend of mAP and AP50 in FPN and enhanced FPN of various models

LayoutLMv3-Base中结合增强FPN之后与原始FPN性能的对比,结果以不同文档元素的平均精度(mAP)来衡量.

根据表3和图7可以得出,通过本文方法进行特征金字塔网络的增强后,能够在现有基于Vision Transformer的文档布局分析方法中实现mAP以及AP50指标的显著改进.对于LayoutLMv3-Base实现了1.9的mAP和1.5的AP50提升,DiT-Base实现了0.8的mAP和1.2的AP50提升,VGT实现了1.2的mAP和0.6的AP50提升.

为了更直观地展现出本文方法的提升效果,实验中对经过整个FPN结构的特征图部分通道进行了可视化,如图8所示,可以观察到,文档在经过传统FPN中的上采样扩大分辨率融合后,一定程度上造成了类别特征不一致和边界位移导致的信息丢失,而本文方法在经过局部特征增强生成和全局到局部增强融合后,能够有效地解决文档中信息丢失的问题,维持了相同类别特征的一致性,获得更为清晰的边界,使得整个特征图具有更为精细的特征表示.

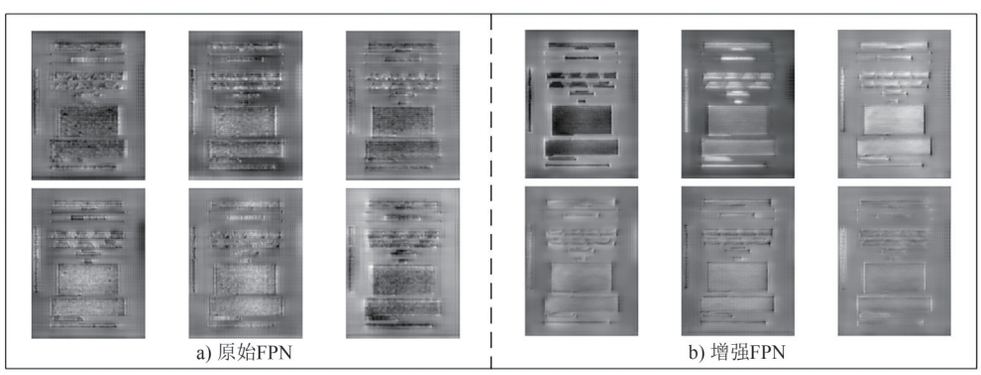


图8 特征图可视化

Fig. 8 Feature map visualization

5 消融实验

最后对提出的整个增强结构在LayoutLMv3-Base中进行消融实验,结果如表4所示.

表4 消融实验结果

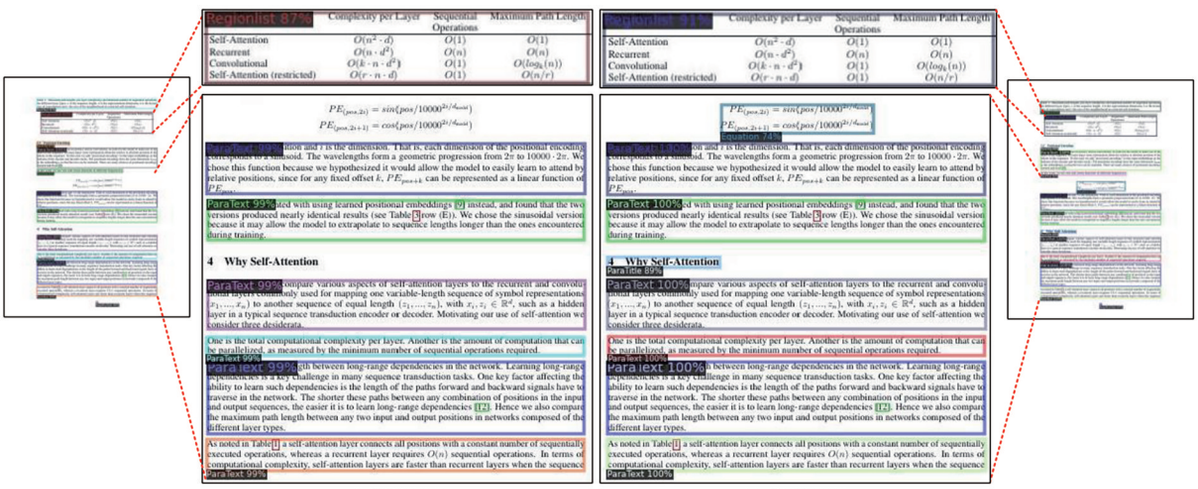
Tab. 4 Ablation experiment results

Ablation		D^4LA	
LFEG	GLEF	mAP	AP50
		60.9	74.5
√		62.4	75.9
		↑1.5	↑1.4
	√	62.7	75.8
		↑1.8	↑1.3
√	√	62.8	76.0
		↑1.9	↑1.5

基于消融实验分析,本研究验证了局部特征增强生成(LFEG)和全局到局部特征增强融合(GLEF)两个模块对基于Vision Transformer的文档布局分析模型具有性能优化的作用.实验表明,独立引入LFEG或GLEF均能显著提升基准模型的平均精度(mAP)与交并比为0.5的精度(AP50).进一步联合两个模块后,性能增益效果显著超越单模块,表明LFEG与GLEF在特征学习层面具有互补性,二者的协同作用实现了多尺度特征生成与跨层次特征融合的联合优化,从而能够在复杂文档页面中显著提升综合检测精度.

6 可视化案例

图9展示了使用VGT在原始FPN(a)和增强FPN(b)两种情形下的检测结果,相较于原始的FPN结构,本文的方法能够对表格、公式这些具有复杂结构的文档元素进行高置信度的检测,不会产生漏



a) 原始FPN　　b) 增强FPN

图9　检测可视化

Fig. 9　Visualization of detection

检情况；同时观察文档中类似文段标题的元素，因为夹杂在大段文段文字中间，原始方式通常会因为无法适应这种程度的尺度变化而出现无法识别．而本文方法得益于特征图的增强，同样能够在高置信度的情况下预测出该部分精确的检测框．这些定性结果证明了本文方法对这些基于 Vision Transformer 的文档布局方法具有很强的特征增强效果．

7　结论

以往文档布局分析任务中使用 Vision Transformer 方式时易出现特征信息丢失的问题，对此本文使用了局部特征增强生成和全局到局部特征增强融合方式来加强特征金字塔网络在处理文档时生成多尺度特征的效果．实验结果和特征图可视化表明：本文方法能够为 Vision Transformer 用于文档布局分析任务时实现更为精细的特征表示，为提升检测精度提供了支撑．

参　考　文　献

[1] Zhang P, Li C, Qiao L, et al. VSR: A unified framework for document layout analysis combining vision, semantics and relations [C]//Document Analysis and Recognition-ICDAR 2021. Cham: Springer, 2021: 115-130.

[2] Li J, Xu Y, Lyu T, et al. DiT: Self-supervised pre-training for document image transformer [C]//Proceedings of the 30th ACM International Conference on Multimedia., Lisboa: ACM, 2022: 3530-3539.

[3] Huang Y, Lyu T, Cui L, et al. LayoutLMv3: Pre-training for document AI with unified text and image masking [C]// Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022: 4083-4091.

[4] DA C, Luo C, Zheng Q, et al. Vision grid transformer for document layout analysis [C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE, 2023: 19405-19415.

[5] Nikitha S R, Menta T R, Sarkar M. DoPTA: Improving document layout analysis using patch-text alignment [J]. arXiv preprint arXiv: 2024, 2412.12902.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv: 2020, 2010.11929.

[7] Lin T Y, Dollar P, Girshick R, et al. Feature pyramid networks for object detection [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 936-944.

[8] Chen L, Fu Y, Gu L, et al. Frequency-aware feature fusion for dense image prediction [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 10763-10780.

[9] Bukhari S S, Breuel T M, Asi A, et al. Layout analysis for Arabic historical document images using machine learning [C]//2012 International Conference on Frontiers in Handwriting Recognition. Bari: IEEE, 2012: 639-644.

[10] Garz A, Sablatnig R, Diem M. Layout analysis for historical manuscripts using sift features [C]//2011 International Conference on Document Analysis and Recognition. Beijing: IEEE, 2011: 508-512.

[11] Ha J, Haralick R M, Phillips I T. Document page decomposition by the bounding-box project [C]// Proceedings of 3rd International Conference on Document Analysis and Recognition. Montreal: IEEE,

1995: 1119-1122.

[12] Ha J, Haralick R M, Phillips I T. Recursive X-Y cut using bounding boxes of connected components [C]// Proceedings of 3rd International Conference on Document Analysis and Recognition. Montrea: IEEE, 1995: 952-955.

[13] Mehri M, Nayef N, Heroux P, et al. Learning texture features for enhancement and segmentation of historical document images [C]//Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing. Gammarth: ACM, 2015: 47-54.

[14] Sermanet P, Chintala S, Lecun Y. Convolutional neural networks applied to house numbers digit classification [C]//Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). Tsukuba: IEEE, 2012: 3288-3291.

[15] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[16] Cai Z, Vasconcelos N. Cascade R-CNN: High quality object detection and instance segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1483-1498.

[17] Saha R, Mondal A, Jawahar C V. Graphical object detection in document images [C]//2019 International Conference on Document Analysis and Recognition (ICDAR). Sydney:IEEE, 2019: 51-58.

[18] Chen H, Chen K, Ding G, et al. YOLOv10: Real-time end-to-end object detection [C]//Advances in Neural Information Processing Systems. Vancouver: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 107984-108011.

[19] Zhao Z, Kang H, Wang B, et al. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception[J]. arXiv preprint arXiv:2024,2410.12628.

[20] 李玉腾. 基于视觉和文本的文档图像布局分析方法[D]. 青岛: 青岛科技大学, 2023.

[21] 温绍杰, 吴瑞刚, 冯超文, 等. 基于Transformer的多模态级联文档布局分析网络[J]. 浙江大学学报(工学版), 2024(2): 317-324, 369.

[22] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015: 3431-3440.

[23] Lu H, Liu W, Ye Z, et al. SAPA: Similarity-aware point affiliation for feature upsampling[J]. Advances in Neural Information Processing Systems, 2022, 35: 20889-20901.

[24] Lu H, Dai Y, Shen C, et al. Index networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(1): 242-255.

[25] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[C]//Computer Vision-ECCV 2014. Cham: Springer, 2014: 740-755.

[26] Wu Y, Kirillova, Massa F, et al. Detectron2[EB/OL]. 2019 [2025-08-07]. https://github.com/facebookresearch/detectron2.

[27] Cai Z, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE, 2018: 6154-6162.

(责编 曹东,校对 雷建云)