arXiv:2305.04609v1 [cs.CV] 8 May 2023

SwinDocSegmenter：一种用于文档实例分割的端到端统一域自适应 Transformer

Ayan Banerjee*[1][0000−0002−0269−2202]、Sanket Biswas*[1][0000−0001−6648−8270]、Josep Llad´os[1][0000−0002−4533−4739] 和 Umapada Pal[2][0000−0002−5426−2618]

[1] 计算机视觉中心与计算机科学系，巴塞罗那自治大学，西班牙{abanerjee、sbiswas、josep}@cvc.uab.es[2] 印度统计研究所 CVPR 单元，印度 umapada@isical.ac.in

摘要。 文档实例级分割旨在为图像中的每个像素分配类别感知且实例感知的标签。这是文档解析与理解的关键步骤。本文提出一种统一的 Transformer 编码器-解码器架构，用于对文档图像中复杂版面进行端到端实例分割。该方法采用对比训练，并在解码器中通过混合查询选择进行锚点初始化。随后，它对获得的查询嵌入与像素嵌入图（由编码器生成）执行点积，以进行语义推理。在 PubLayNet、PRIMA、历史日文（HJ）和 TableBank 等具有竞争力的基准测试上进行的大量实验表明，在参数量低于十亿的情况下，我们采用 SwinL 主干网络的模型分别取得了 **93.72**、**54.39**、**84.65** 和 **98.04** 的平均精度，分割性能优于现有最先进的方法。代码已公开发布，地址为：github.com/ayanban011/SwinDocSegmenter

关键词： 文档版面分析 · 实例级分割 · Swin Transformer · 对比学习。

1 引言

文档智能（DI）系统有助于提供解决方案，自动化大规模文档处理工作流，以提取信息并理解文档内容。商业智能流程（如文档检索、文本识别、内容分类等）在将文档解析为结构化的机器可读格式时，通常需要提取文档中的语义信息。随后，这些提取出的数据可以集成到机器人流程自动化工具的文档处理工作流中。因此，关键工业领域（例如银行业、金融、医疗保健等）已开发出更高效的解决方案 [31,34]。文档版面分析（DLA）已

* 这些作者对本研究贡献相同。

成为文档智能中的一项重要任务，因为任何与文档理解相关的任务，都需要获得结构化表示，以便定位文档中存储的关键信息。最初，基于卷积的算法（卷积神经网络）取得了显著进展，例如用于文档目标检测（DOD）的 Faster RCNN [43] 、用于实例分割的 Mask RCNN [5] ，以及其他专用架构。这些架构实现起来相当简单，在某些特定案例研究中也很有效（例如表格检测 [24], 、科学文章的版面分析 [49] 等），但缺乏处理其他类似任务的泛化能力。近年来，基于 Transformer 的架构 [2,16] 借助全局注意力机制，取得了优于卷积神经网络的性能。然而，这些模型并未统一，因而无法实现检测任务与分割任务之间的协同，影响了模型性能，因为检测模块和分割模块无法相互引导。不仅如此，这些架构还对其预训练数据集存在偏向，在面对类似任务的域偏移时表现不佳。由于这类 Transformer 模型通常使用大量来自相关源领域的数据进行预训练（例如大规模工业文档 [15] 或科学文章 [49]），因此无法处理差异相对较大的任务（例如杂志的版面提取 [10]）。将这种域偏移属性引入文档版面分析模型，有望降低计算开销，并帮助构建一个不依赖特定数据的通用模型。

为解决上述问题，我们提出了 *SwinDocSegmenter* 框架，通过使用基于 Swin Transformer 特征提取骨干网络 [30] 和 Transformer 编码器特征生成的高分辨率像素嵌入图上的内容查询嵌入，对复杂文档版面执行实例级分割。它有助于在更高层次上对特征进行全局语义推理，克服使用 ResNet-FPN [28] 骨干网络的缺点。在此，我们利用编码器密集先验将掩码查询初始化为锚点，并根据排名靠前的词元预测掩码。这有助于在早期阶段执行像素级分割，从而增强边界框。在后续阶段，这些边界框通过构造动态锚框来提升分割性能。这种相互促进的任务协同现象有助于构建用于版面检测与分割的统一模型。我们引入了一种受 [48] 启发的对比去噪训练，通过聚焦低层级实例来加速分割训练。它大幅提升了模型性能，因为 Transformer 的主要缺点之一是在处理无标签数据时会惩罚特征表示数量极少的类别 [4]。最后但同样重要的是，我们采用混合二分图匹配 [21]，以实现更一致的语义匹配，这有助于处理域偏移 ，并利用来自完全不同领域的预训练 Transformer 权重执行类似任务。在本例中，我们利用 MS-COCO目标检测基准 [29] 的预训练权重，对复杂文档版面进行实例级分割。

本工作的总体贡献可概括为以下三个方面：

– 我们提出了一个统一的基于 *Transformer* 的框架，用于执行实例级文档布局分割，该框架采用 Swin Transformer 主干网络、锚框引导的交叉注意力机制以及增强的查询选择策略。–我们引入对比去噪训练，以增强低层实例，从而提升无标签数据集上的性能。– 我们采用混合二分图匹配来利用域偏移特性，以节省预训练时间，并使用来自不同域、用于相似任务的公开预训练权重，从而提升模型的泛化能力。

本文其余部分组织如下：第 2 节回顾文档版面分析领域最先进的方法。第 3 节介绍 $SwinDocSegmenter$ 。第 4 节介绍我们的实验评估以及消融实验。最后，第 5 节总结全文，并展望未来的研究方向。

2 相关工作

为了从数字文档中提取相关信息，布局识别方法通过对不同布局组件（例如表格、文本、图、标题等）之间进行关系推理来获得空间理解。在深度学习时代之前，主流的版面分析算法一直由经典的启发式规则算法主导。随后，卷积框架在解决该任务方面发挥了主导作用，直到基于 Transformer 的架构取得了卓越性能。本节旨在通过分析不同的方法方案，概述该任务的最新研究进展。

基于启发式规则的文档版面分析。 使用启发式方法进行文档版面分割，可以进一步分为三种不同的类别：自顶向下、自底向上和混合策略。自底向上的方法[3,38] 执行像素分组和合并等基本操作，为相似对象创建同质区域，并将其与不相似的对象分离。自顶向下的策略 [17,19] 迭代地将文档图像划分为不同区域，直到在相似对象周围获得明确的区域。尽管自底向上的方法能够处理复杂版面，但其计算成本较高。此外，自顶向下的方法实现速度更快，但会牺牲泛化能力，且仅能在特定类型的文档上有效运行。为了兼顾两者的优势，混合方法 [6,11]结合自底向上和自顶向下的线索，以获得快速且高效的结果。在深度学习时代到来之前，这些方法是表格检测领域的最先进技术。

基于卷积的文档版面分析。 自 2012 年以来，深度学习算法取代了基于规则的算法，卷积神经网络（CNNs）成为解决文档实例分割任务的首选策略。Faster-RCNN [37]提供了强大的文档目标检测能力，可用于解决页面分割问题 [25]。随后，类似的网络 Mask-RCNN [1]为实例级布局分割提供了首个基准测试。

报纸元素的分割。RetinaNet [27] 还提供了另一种卷积基准测试，用于文档图像中的关键词检测。这是一种复杂的方法，但只能帮助检测文本区域。为了提供一个用于表格检测和表格结构识别的全新先进基准测试，DeepDeSRT [40] 采用了一种新颖的图像变换策略，以识别表格结构的视觉特征，并将其输入带有跳跃池化的全卷积网络。类似地，Oliviera 等人 [33] 使用了基于 FCNN 的类似框架，对历史文档页面进行逐像素分割，其性能优于 Chen 等人 [7,8] 先前提出的基于卷积自动编码器的基准测试。Saha 等人 [39] 提供了 ICDAR2017 POD（页面对象检测）基准测试 [12] ，通过采用基于迁移学习的 Faster-RCNN 主干网络，实现了对数学公式、表格和图的先进检测结果。文献 [23] 建立了一项新的跨领域文档目标检测（DOD）基准测试，以应用领域自适应策略解决领域偏移问题。最近，文献 [47] 提出了一项基于视觉的版面检测基准测试，该方法利用 VoVNet-v2 主干网络 [20] 的循环卷积神经网络，通过 ICDAR-2013 和 GROTOAP 数据集生成合成 PDF 文档。该方法建立了一项新的基准测试，用于解决科学文档分割任务。

基于 ***Transformer*** 的文档版面分析。 如今，Transformer 通过利用位置嵌入和自注意力机制，展现出了更加出色的性能 [44]。在此，DiT [22]通过在大规模无标注文档图像上进行自监督预训练，为文档图像分类、版面分析和表格检测建立了新的基准，但这种方法并不适用于 PRIMA 等小规模杂志数据集。类似地，Li 等人 [26] 提出了一种多模态框架，用于理解文档中的结构化文本。然而，该模型在处理文本内容中语义相似的情况时表现非常糟糕。为了提升这些性能，研究人员引入了 TILT [35]机制，该机制通过编码器-解码器 Transformer 同时学习文本语义、视觉特征和布局信息。文献 [46] 采用了类似的 Transformer 编码器-解码器架构，通过 OCR 提取文本信息，为 PubLayNet 数据集提供了新的基准（AP：95.95）。最近，LayoutLmv3 [16] 通过联合学习文本、布局和视觉特征，在视觉文档理解（VDU）任务中取得了最先进的结果。该方法在大规模数据集上表现出色，但在小规模数据集上则表现不佳。DocSegTr [4] 在 Transformer 层上采用了带有自注意力机制的 ResNet-FPN 主干网络，这有助于其在小规模数据集上更快收敛，但仍无法达到最先进的性能。其他近期方法 [2,18,13,14] 也采用了这种联合预训练策略，以解决包括文档视觉问答在内的多种 VDU 任务。这些技术通过统一预训练，对多种下游任务具有显著帮助。然而，这种方法存在预训练偏差，阻碍了模型适应域偏移；此外，由于缺乏权重优先机制，当实例数量较少时，模型也无法有效学习类别信息。

受 Transformer 近期突破的启发，为解决上述问题并提升其性能，我们提出了一种端到端、统一域自适应的文档分割 Transformer，并结合对比训练。该方法不仅在标准实例级分割基准上取得了卓越性能，还为新提出的工业文档版面分析数据集 [34] 提供了首个 Transformer 基线。

3 方法

所提出的*SwinDocSegmenter* 是一种统一的端到端架构，包含一个 Swin Transformer 主干网络 [30], 、一个 Transformer 编码器-解码器对，以及一个通过类别实例映射由多个投影头构成的分割分支。所提出的架构如图 1 所示，模型首先利用 Swin 主干网络提取多尺度特征。随后，特征...

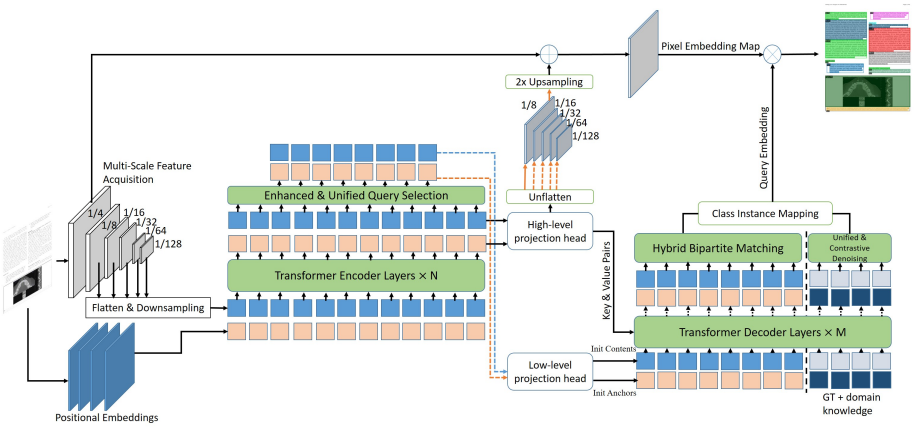


图1：**提出的 SwinDocSegmenter 框架**。 给定来自任意领域的输入文档图像，模型通过统一的检测与分割分支预测分割后的文档布局。

特征在送入 Transformer 编码器之前会被展平并下采样，否则将产生大量可训练参数，在资源有限的情况下无法进行训练。Transformer 编码器将这些特征及其对应的位置嵌入（通过若干个卷积核大小为 3×3 的卷积层获得）作为输入，以执行特征增强。此处采用了一种统一的混合查询选择策略，通过低层投影头初始化位置查询和锚点。这种查询选择策略的主要优势在于，它不会初始化内容查询，而是让它们保持可学习状态，这在发生域偏移时很有帮助。此外，借助低层投影头，它还能聚焦于低维图像特征；由于数据点不足，这些特征在 Transformer 训练中往往会被忽略。这也使解码器能够进行对比去噪训练（CDN） [48]。在解码器中，可变形的

利用注意力机制 [45] 将编码器的输出与逐层查询更新相结合。通过 CDN，还将已分割区域/错误分割区域视为困难负样本，并采用两次前瞻方法 [48] 进行纠正，在早期阶段将梯度传递到相邻层之间。我们采用混合二分图匹配策略，基于有助于生成精确分割区域的动态锚框来优化分割区域。通过类别实例映射将这两部分信息结合起来，得到最终的查询嵌入。我们对最终的查询嵌入与像素嵌入图执行点积运算，从而获得文档图像上的最终实例分割输出。

3.1 分割分支

为了执行掩码分类，我们采用一个关键思想 [9] ，通过结合多尺度特征（由 Swin 主干网络提取）和 Transformer 编码特征来构建像素嵌入图（PEM）。如图 1 所示，PEM 通过融合来自主干网络（S_b）的 $1/4^{th}$ 分辨率特征图与来自 Transformer 编码器（T_e）的上采样 $1/8^{th}$ 分辨率特征图构建而成。输出掩码 M 通过 PEM 与从解码器获得的查询嵌入（ Q_e ）进行点积运算得到（见公式 1 ）。

$$M = Q_e \otimes \delta(\Gamma(S_b) + \psi(T_e)) \tag{1}$$

其中， Γ 是用于将通道维度映射到 Transformer 维度的卷积层， ψ 是用于对T_e进行 $2\times$ 上采样的插值函数，而δ 是分割头。该机制简单且易于实现。

3.2 特征编码技术

特征编码技术由四个重要的子部分组成：查询选择、低层和高层特征投影以及锚点初始化，旨在提升性能并简化解码技术。

查询选择策略 研究发现，编码器的输出包含可在解码器中用作更优先验的密集特征。在此，我们在编码器输出中采用了一个分类头、一个检测头和一个分割头，后接一个低层投影头和一个高层投影头。我们将每个令牌的分类分数作为置信分数，并利用这些分数选出排名靠前的特征，将其作为内容查询输入解码器。选定的特征还通过检测头和分割头回归边界框，并经过高层投影头，通过点积与高分辨率特征图结合，以预测掩码。这些预测掩码和边界框经过低层投影头处理后，被视为解码器的初始锚点。这有助于解码器进行对比训练，因为其中同时存在高层和低层类别实例；通过调整对比损失，我们可以在不影响高层实例性能的情况下，提升低层类别实例的性能。

低层投影头 这是一种浅层多层感知机（MLP），利用投影特征生成低层嵌入，用于低层视图中的对比学习。它有助于学习更细粒度的不变性。具体而言，我们对低层特征应用非线性函数 $F = (f_1, f_2, ... f_s)$，以增强这些特征，然后将其初始化为解码器中的内容查询和锚点。该低层投影头的目标函数定义见公式 2。

$$\mathcal{L}_{low} = \sum_{i=1}^{n} \sum_{j=1}^{n'} -log \frac{exp(f_i \cdot f_j / \tau)}{\sum_{c=1}^{k} exp(f_c \cdot f_j / \tau)} \tag{2}$$

其中， n 和 n' 分别表示从检测头和分割头获得的特征数量， c 表示从分类头获得的排名靠前的特征数量。在这里，我们需要一个温度超参数 τ ，根据所使用的数据集调整各层，以增强特征。对于 PubLayNet、PRIMA、HJ 和 TableBank，该值分别为 $\tau = 0.02$，0。6，0。1，和 0。2。注意：所有这些超参数值均通过实验获得。

高层投影头 它是一个深层多层感知机，用于保留高层特征的高层不变性。基本上，它设置不同数量的原型 $P = (p_1, p_2, ..., p_m)$，以获得不同的键值对 $k_1 v_1$, $k_2 v_2, ..., k_n v_n$，这也丰富了特征表示。这种原型化的目标函数已在公式 3 中定义。

$$\mathcal{L}_{high} = \sum_{i=1}^{n} \sum_{j=1}^{n'} -log \frac{exp(f_i \cdot p_j / \phi_j)}{\sum_{c=1}^{k} exp(f_c \cdot p_j / \phi_j)} \tag{3}$$

其中， p_j 是对应键值对的原型， ϕ_j 是 [32] 表示围绕该原型的表示分布的集中度估计指标。

锚点初始化 文档实例分割是一种像素级分类任务，而目标检测则是一种区域级位置回归任务。因此，尽管分割在初始阶段更容易学习，但由于其粒度更精细，相较于检测而言更具挑战性。例如，利用高分辨率特征图进行点积查询，只需比较每个像素的语义相似度即可预测掩码。然而，对于图像中的目标检测，则必须直接回归边界框坐标。因此，在初始阶段，掩码预测的准确性显著高于边界框预测。为了解码器提供更好的锚点初始化，我们因此在统一查询选择之后，根据预测掩码推导出边界框。得益于这种高效的任务协同，增强后的边界框初始化有望显著提升检测性能。

3.3 用于掩码预测的特征解码

在这一阶段，我们引入统一对比去噪训练，以有效提升低层级实例的性能，并采用混合二分匹配来应对域偏移。下面将详细讨论这两种策略。

统一对比去噪训练 查询去噪 [48]是一种通过加快收敛来提升性能的有效技术。然而，它无法区分两个相近的类别实例。CDN 可以通过拒绝无用的锚点来解决这一问题。在此，我们向真实标签和边界框添加噪声，Transformer解码器接收这些带噪的位置查询和内容查询。在这里，我们有两个超参数 λ_p 和 λ_e，其中 $\lambda_e > \lambda_p$如图2所示。它有助于生成两种类型的查询（正向和

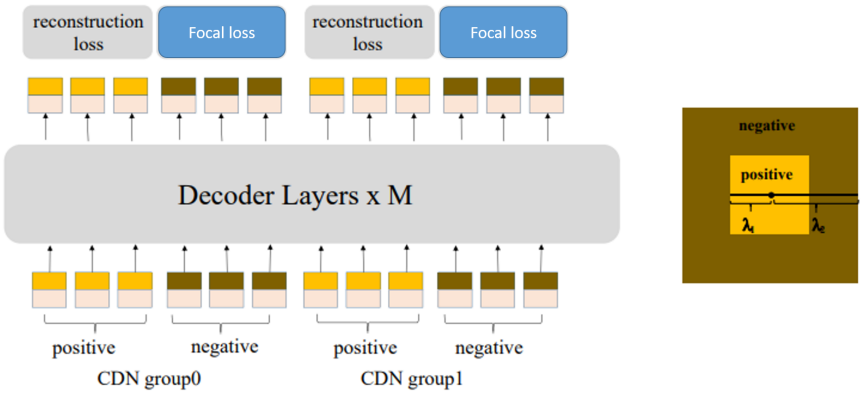


图2：**统一对比去噪训练策略**。与 DINO[48] 的实现类似，但我们引入了焦点损失来替代无目标检测，以优化和增强低层级实例。

负样本）。我们预计，内层方框中的正查询将重构其对应的真实框，因为它们的噪声尺度小于 λ_p。另一方面，负查询的噪声尺度大于 λ_p 且小于 λ_e，并通过焦点损失进行最小化。通常，我们将 λ_e 保持得非常小，因为这有助于使困难负样本靠近真实标注锚点，从而提升性能。每个 CDN组都包含正查询和负查询（见图2）。对于包含 q 个真实框的图像，每个 CDN组将包含 $2 \times q$ 个查询，其中每个真实框都会生成一个正查询和一个负查询。为提高效率，我们还采用多个 CDN组。

为了训练模型，我们利用对象特征经过噪声处理后的版本来重建这些特征。我们也将这种方法应用于涉及分割的任务。边界框与掩码天然相关，因为掩码可以被视为边界框更精细的表示。因此，我们可以将根据边界框预测掩码的过程作为一种去噪任务来训练模型，并将其视为

将边界框作为掩码的噪声版本。为了更有效地训练掩码去噪，用于掩码预测的边界框也会被随机添加噪声。在训练过程中，这些带噪声的对象将被添加到原始解码器查询中，但在推理过程中会被移除。我们进行了大量调优，以获得 λ_p 和 λ_e 的最优值。然而，观察发现，分别采用 $\lambda_p = 0.1$ 和 $\lambda_e = 0.02$ 时，大多数通用性能指标均已达到最佳水平。

混合二分图匹配 该技术通过调整不同预测头所预测的一对掩码的对应权重，有助于消除它们之间的不一致性。基于这一动机，我们将这一概念应用于域偏移。具体而言，我们在二分图匹配中，除了 L1 损失和焦点损失之外，还额外加入了掩码预测损失。该损失能够促使单个查询获得更加准确且一致的匹配结果。因此，当我们使用来自不同域的预训练模型时，会对这一损失施加更大的惩罚，迫使模型对相应权重进行显著调整；随着模型逐渐接近收敛，我们会缓慢降低惩罚强度。同时，该损失也会与 L1 损失和焦点损失共同优化，从而实现统一的域偏移处理。最后，在类别与预测实例之间执行类别实例映射。这是一种简单的一对一映射，用于执行查询嵌入（query embedding），并通过点积运算将其与像素嵌入图有效结合，从而以端到端的方式完成实例分割过程。

4 实验评估

数据集。 文档版面分析（DLA）领域一直以来都关注缺乏标准公共基准测试的问题。在本研究中，我们使用大规模标注数据集，如 PubLayNet [49], 、TableBank [36], 和历史日文（HJ） [41] ，以及小规模数据集 PRIMA [10] ，来评估我们提出的分割方法（详细说明请参阅表 1）。除此之外，我们还将模型与近期发布的标准工业级文档版面分割基准测试 **DocLayNet** [34]进行比较。该数据集包含 91104 个对象实例，涵盖 11 种不同的标签（图注、脚注、公式、列表项、页脚、页眉、图片、章节标题、表格、文本和标题），并覆盖了广泛的文档对象尺寸范围（从大到小）。

评估指标。 交并比（IoU）分数是评估实例级分割任务中预测实例（文档类别）准确性的最通用方式。Microsoft COCO 基准测试的标准实例分割评估采用不同 IoU 阈值（0.5 至 0.95，步长为 0.05）下 APs 的均值，计算整个模型的平均精度均值（mAP）。由于这些方法均在类似环境下计算 mAP，因此将本文提出的方法与现有方法进行比较颇有助益。此外，我们还依据 [4,16,42] 计算了模型对各类别文档实例的评估性能。

表格 1：实验数据集描述（实例级）

PubLayNet			PRIMA			历史日文			TableBank		
对象	训练	Eval	对象	训练	评估	对象	训练	评估	对象	训练	评估
Text	2,343,356	88,625	Text	6401	1531	Body	1443	308	表格	2835	1418
标题	627,125	18,801	图像	761	163	Row	7742	1538	-	-	-
列表	80,759	4239	表格	37	10	标题	33,637	7271	-	-	-
图	109,292	4327	Math	35	7	Bio	38,034	8207	-	-	-
表格	102,514	4769	分隔符	748	155	Name	66,515	7257	-	-	-
-	-	-	其他	86	25	位置 33	,576	7256	-	-	-
-	-	-	-	-	-	其他	103	29	-	-	-
总计 3	**,263,046 120**	**,761**	**总计**	**8068**	**1891**	**总计** 181	,097 31	,866	**总计** 2835	1418	

特征提取骨干网络的选择。针对实例级文档分割，我们开展了广泛的消融实验，以量化模型框架中各组件的重要性，并论证其在分割各种布局元素时的作用。所有消融实验均在 PRIMA数据集上进行，因为该数据集规模最小，且包含复杂布局。本研究采用了不同的卷积神经网络和视觉 Transformer 骨干网络（见表 2）。其中，我们选用 SwinL Transformer 骨干网络进行多尺度特征提取。尽管可训练参数数量有所增加，但与 ResNet、ResNeXt 和视觉 Transformer 相比，性能提升了 8%；与 Swin Tiny 相比，则提升了 5%。卷积骨干网络能够关注局部特征，有利于小目标检测，但缺乏全局注意力，无法对大型对象施加惩罚。另一方面，视觉 Transformer 使用自注意力机制，但需要大量训练数据才能学习多尺度特征。起初，Swin-Tiny 表现良好，但对噪声敏感，因此在后期会对重建施加惩罚，影响总体性能。SwinL 尺寸较大，能够轻松消除噪声，性能也优于其他方法。

表格 2：不同特征提取骨干网络的消融实验

主干网络	参数量	AP	AP@50	AP@75	APs	APm	APl
ResNet-50	52M	36.065	52.362	41.112	20.152	23.327	38.142
ResNet-101	102M	37.112	54.982	41.872	22.242	26.153	41.986
ResNext-101	104M	38.405	58.405	41.916	25.982	29.364	44.129
ViT-S	126M	40.342	59.763	42.158	29.176	33.129	48.526
ViT-B	164M	46.128	62.689	47.358	31.389	33.458	50.508
Swin-T	178M	49.349	65.956	50.317	34.128	36.909	52.049
Swin-L	223M	**54.393**	**69.313**	**52.965**	**39.327**	**42.061**	**60.142**

输入图像分辨率的选择。 除此之外，由于模型规模较大且可训练参数数量庞大，图像分辨率也会影响模型性能。因此，如果使用较低的图像分辨率，那么在训练后期，模型只会学习噪声，从而浪费计算资源。提高图像分辨率能够改善模型性能（见表3），然而，受限于资源，我们无法将分辨率提高到1024以上。此外，从变化趋势可以得出，提高图像分辨率也能提升系统性能，直至达到饱和点。

表3：图像分辨率对实例分割性能的影响

图像分辨率	平均精度 AP@50 AP@0.75 APs APm			APl
256×256	45.022 60.189	46.258	28.372 32.458	53.568
512×512	50.132 66.235	52.317	32.242 36.909	54.148
1024×1024	54.393 69.313	52.965	39.327 42.061	60.142

解码器查询数量的选择。同样，深入分析模型后我们发现，解码器中用于初始化的查询数量会影响总体性能。查询数量较少时，很难生成能够达到相近性能的负样本，这不仅会损害模型性能，还会增加损失函数的优化时间。此外，密集查询可以稳定模型，并有助于修正误分类样本（见表4）。使用 SwinL 主干网络时，查询数量可以扩展到 900–1200，但由于计算资源有限，我们必须将其限制为 300。

表4：Transformer 编码器生成的查询数量消融实验

查询数量	**平均精度 AP@50**	**AP@0.75**	**APs APm**		**APl**
100	50.022	65.189	52.258	32.372 36.458	53.968
150	50.132	66.235	52.317	32.242 36.909	54.148
200	51.393	67.313	52.765	37.312 41.011	60.111
250	52.092	68.212	52.964	37.512 42.060	60.132
300	**54.393**	**69.313**	**52.965**	**39.327 42.061**	**60.142**

学习目标的选择。最后，对损失函数进行了消融实验，以了解重建损失与分类损失的哪种组合最优。从图 3 可以看出，L1 损失与焦点损失的组合在此任务中最有效。L1 损失倾向于将系数压缩至零，这更有利于特征选择；而 L2 损失倾向于均匀地压缩系数。另一方面，焦点损失有助于调整标准交叉熵损失，动态降低易分类样本对应损失的权重，并更加关注难分类样本，从而使系统在难分类样本上也能表现得更好。

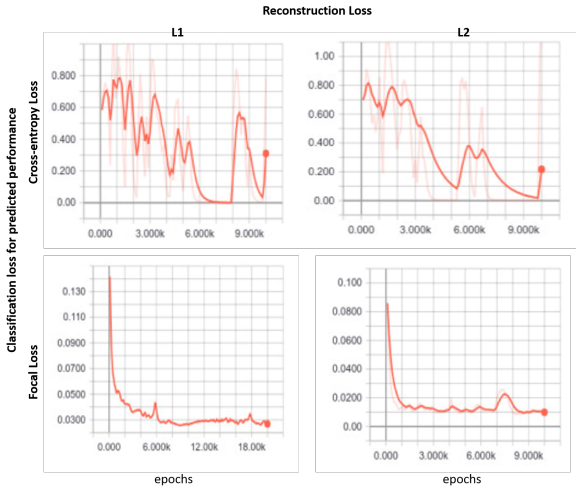


图3：**学习目标的影响**。上述图表研究了不同损失函数组合的有效性

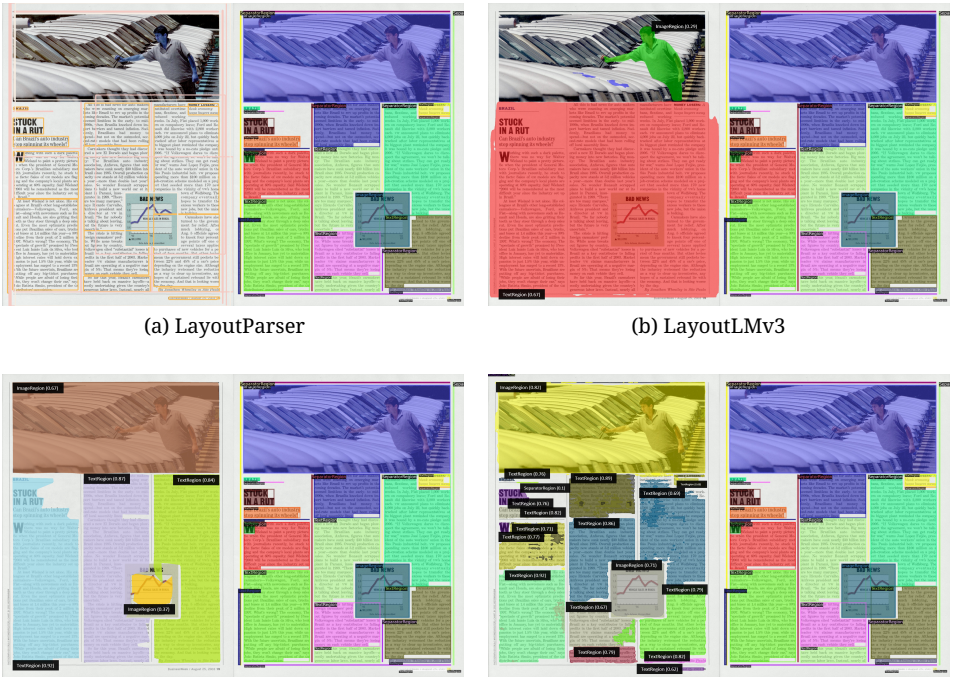
此外，在表5中观察到一个最有趣的事实：在相似数据集上进行预训练会引入偏差，并降低模型的总体性能。在这里，我们使用了一个在PubLayNet数据集上预训练的SwinL 主干网络，以及另一个在MSCOCO数据集上预训练的SwinL 主干网络。我们可以观察到，由于“表格”是两个数据集中的常见类别，模型在该类别上取得了非常高的性能。但是，由于预训练模型不熟悉“分隔符”区域，它会对其施加较大惩罚，从而降低网络的总体性能，因为模型通过关注相似类别并将其他类别视为CDN中的负样本，使损失迅速收敛。相比之下，采用MSCOCO预训练时，借助增强的统一查询选择，使内容查询具有可学习性，并通过混合二分图匹配优化相应的预训练权重，模型能够取得更通用的性能。这有助于消除偏差因素。

表5：预训练如何提供偏置

预训练	总体性能						类别级性能					
	平均精度	AP@50	AP@75	APs	APm	APl	文本	图像	表格	数学	分隔符	其他
PubLayNet	49.36	64.43	51.45	32.94	34.07	54.21	85.55	72.51	**70.68**	56.05	8.55	2.83
Ms-COCO	**54.39**	69.31	52.96	39.32	42.06	60.14	87.72	75.92	49.89	**78.19**	**27.56**	**7.05**

定性分析。 图 4 展示了 *SwinDocSegmenter* 和最先进的方法在 PRIMA数据集上取得的版面分割结果。在这个测试案例中，*SwinDocSegmenter* 能够相当有效地分割不同布局元素的实例。

从图4(a)可以看出，尽管 LayoutParser 在 PRIMA数据集上的表现相当有效，但面对这个复杂案例时却失效了，因为边界框



(a) LayoutParser (b) LayoutLMv3

(c) DocSegTr (d) Ours

图4：SwinDocSegmenter 框架与当前最先进方法的比较分析（**左侧：** 预测布局 **右侧：** 真实标注）

重叠严重，且与真实标注的匹配不准确。不过，LayoutLMv3（图4(b)）在这种情况下的表现很差。它识别出了文本和图形实例，却未能识别其他类别实例。DocSegTr（图4(c)）尝试提升性能，但其分割结果仍与真实标注相去甚远。另一方面，我们的方法能更令人满意地分割复杂文档，并将类别实例映射到真实标注（图4(d)）。

定量分析。 SwinDocSegmenter 的最终平均精度均值表现相当有趣，并有望为文档布局分割提供新的基准测试。在表6和表7中，该方法排名第二，因为 LayoutLMv3 [16] 和 Layout Parser [42] 都在实例分割任务中结合了文本信息与视觉信息。在 PubLayNet 数据集上，我们发现，所提出的方法在文本区域识别方面优于 LayoutLMv3，除“标题”类别外，在其他类别上的表现也与之相当。“标题”中也包含文本，因此若缺少文本信息，就很难处理这些边界案例。此外，仅使用视觉信息时，该方法在 AP@0.5 和 AP@0.75 指标上已超越 LayoutLMv3。不仅如此，它还优于 DiT（平均精度：93.5） [22]，并且与 UDoc（平均精度：93.9） [13] 表现相当；这两种方法为 PubLayNet 数据集提供了标准基准。在 PRIMA 数据集上也观察到了相同的情况。已经发现，LayoutLMv3

表6：在 DocLayNet 基准上的性能

类别	MaskRCNN	FasterRCNN	YOLOv5	我们的方法
图注	71.5	70.1	77.7	**83.56**
脚注	71.8	73.7	**77.2**	64.82
公式	63.4	63.5	**66.2**	62.31
列表项	80.8	81.0	**86.2**	82.33
页脚	59.3	58.9	61.1	**65.11**
页眉	70.0	**72.0**	67.9	66.35
图片	72.7	72.0	77.1	**84.71**
章节标题	69.3	68.4	**74.6**	66.5
表格	82.9	82.2	**86.3**	**87.42**
Text	85.8	85.4	88.1	**88.23**
标题	80.4	79.9	**82.7**	63.27
All	73.5	73.4	**76.8**	**76.85**

检测小对象的能力不足。但 Layout Parser 可以做到这一点，因为它采用卷积骨干网络而非 Transformer 骨干网络，并且还使用 Microsoft OCR 从图像中提取文本信息，再将其与视觉信息结合。所提出的模型在除“表格”和“其他”类别外的所有类别上均优于这一最先进方法。其表现主要受“其他”类别影响，因为该类别没有明确的定义，而且缺少适当的文本信息时，很难将其与其他类别区分开来。

表7：PubLayNet 和 PRIMA 基准测试上的性能分析

PubLayNet					**PRIMA**				
对象	**布局解析器**	**Doc SegTr**	**布局 LMv3**	**Ours**	**对象**	**布局解析器**	**Doc SegTr**	**布局 LMv3**	**Ours**
Text	90.1	91.1	94.5	**94.55**	**Text**	83.1	75.2	70.8	**87.72**
标题	78.7	75.6	90.6	87.15	**图像**	73.6	64.3	50.1	**75.92**
列表	75.7	91.5	95.5	93.03	**表格**	95.4	59.4	42.5	49.89
图	95.9	97.9	97.9	97.91	**Math**	75.6	48.4	46.5	**78.19**
表格	92.8	97.1	97.9	97.25	**分隔符**	20.6	1.8	9.6	**27.56**
					其他	39.7	3.0	17.4	7.054
AP	86.7	90.4	**95.1**	93.72	**AP**	64.7	42.5	40.3	54.39
平均精度@0.5	97.2	97.9		**97.94 平均精度@0.5**		77.6	54.2		69.31
平均精度@0.75	93.8	95.8		**96.28 平均精度@0.75**		71.6	45.8		52.965

在表 8 中可以观察到，所提出的方法优于此前所有最先进的方法。在历史日文数据集上，该方法以微弱优势（1%）超越了 DocSegTr，但在“名称”和“职位”类别中表现出显著提升。另一方面，由于 TableBank 数据集仅包含一个类别，且布局相对简单，因此该方法在 TableBank 数据集的表格检测任务中取得了显著提升（5%）。

表 8：在历史日文和 TableBank 基准测试上的性能分析

历史日文						TableBank			
对象	布局解析器	Doc SegTr	布局 LMv3	我们的方法	对象	布局解析器	Doc SegTr	布局 LMv3	Ours
Body	99.0	99.0	99.0	**99.72**	表格	91.2	93.3	92.9	**98.04**
Row	98.8	99.1	99.0	99.0					
标题	87.6	93.2	92.9	89.5					
Bio	94.5	94.7	94.7	86.26					
Name	65.9	70.3	67.9	**83.8**					
位置	84.1	87.4	87.8	**93.0**					
其他	44.0	43.7	38.7	40.57					
AP	81.6	83.1	82.7	**84.55**	AP	91.2	93.3	92.9	**98.04**
平均精度@0.5		90.1		**90.78**	平均精度@0.5		98.5		**98.95**
平均精度@0.75		88.1		**88.22**	平均精度@0.75		94.9		**98.90**

最后但同样重要的是，我们为新提出的数据集 **DocLayNet** 取得了首个基于 Transformer 的基准测试。该数据集包含工业文档，其布局比 PubLayNet 基准测试更具挑战性。从表6可以看出，我们提出的 SwinDocSegmenter 的表现显著优于基于卷积神经网络的算法（ MaskRCNN、FasterRCNN 等）。

5 结论

本文提出了 *SwinDocSegmenter*，这是一种功能强大的模型，仅利用视觉信息即可进行文档版面分析。其性能优于当前最先进方法，主要得益于增强且统一的查询选择、对比去噪训练以及两次前瞻方法。此外，低层投影头有助于增强低层级实例，从而显著提升整体性能，因为 Transformer 通常不擅长检测小对象。不过，该方法仍有进一步改进的空间。在 PRIMA 上的性能尚未达到仅使用视觉特征时的最先进水平，因为该数据集具有复杂的版面且训练样本极少。未来可以采用少样本设置来提升性能。

致谢

本工作部分得到了西班牙项目 PID2021-126808OB- I00、加泰罗尼亚项目 2021 SGR 01559 以及 AGAUR 的博士奖学金（2021FIB-10010）的支持。计算机视觉中心是 CERCA计划 / 加泰罗尼亚政府 的组成部分。

参考文献

1. Almutairi, A., Almashan, M.：使用 Mask R-CNN 对报纸元素进行实例分割。载于：2019年第18届 IEEE 机器学习与应用国际会议（ICMLA）。第1371–1375页。IEEE（2019年）2. Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.：Docformer：用于文档理解的端到端 Transformer。载于：IEEE/CVF 国际计算机视觉会议论文集。第993–1003页（2021年）3. Asi, A., Cohen, R., Kedem, K., El-Sana, J.：简化历史手稿的阅读。载于：国际文档分析与识别会议论文集（2015年）4. Biswas, S., Banerjee, A., Llad´os, J., Pal, U.：Docsegtr：一种用于文档图像分割的实例级端到端 Transformer。arXiv 预印本 arXiv:2201.11438（2022年）5. Biswas, S., Riba, P., Llad´os, J., Pal, U.：超越文档目标检测：复杂版面的实例级分割。《国际文档分析与识别期刊》（IJDAR）**24**(3)，269–281（2021年）6. Chen, J., Lopresti, D.：噪声离线手写文档中的表格检测。载于：ICDAR（2011年）7. Chen, K., Seuret, M., Hennebert, J., Ingold, R.：用于历史文档图像页面分割的卷积神经网络。载于：2017年第14届 IAPR 国际文档分析与识别会议（ICDAR）。第1卷，第965–970页。IEEE（2017年）8. Chen, K., Seuret, M., Liwicki, M., Hennebert, J., Ingold, R.：利用卷积自编码器对历史文档图像进行页面分割。载于：2015年第13届国际文档分析与识别会议（ICDAR）。第1011–1015页。IEEE（2015年）9. Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., Schwing, A.G.：用于视频实例分割的 Mask2Former。arXiv 预印本 arXiv:2112.10764（2021年）10. Clausner, C., Antonacopoulos, A., Pletschacher, S.：ICDAR2019 复杂版面文档识别竞赛（RDCL2019）。载于：国际文档分析与识别会议论文集。第1521–1526页（2019年）11. Fang, J., Gao, L., Bai, K., Qiu, R., Tao, X., Tang, Z.：一种通过视觉分隔符和表格结构检测多页 PDF 文档中表格的方法。载于：ICDAR（2011年）12. Gao, L., Yi, X., Jiang, Z., Hao, L., Tang, Z.：ICDAR2017 页面对象检测竞赛。载于：国际文档分析与识别会议论文集。第1卷，第1417–1422页（2017年）13. Gu, J., Kuen, J., Morariu, V.I., Zhao, H., Jain, R., Barmpalios, N., Nenkova, A., Sun, T.：UniDoc：用于文档理解的统一预训练框架。《神经信息处理系统进展》**34**，39–50（2021年）14. Gu, Z., Meng, C., Wang, K., Lan, J., Wang, W., Gu, M., Zhang, L.：XYLayoutLM：面向视觉丰富文档理解的布局感知多模态网络。载于：IEEE/CVF 计算机视觉与模式识别会议论文集。第4583–4592页（2022年）15. Harley, A.W., Ufkes, A., Derpanis, K.G.：深度卷积网络在文档图像分类与检索中的评估。载于：2015年第13届国际文档分析与识别会议（ICDAR）。第991–995页。IEEE（2015年）

16. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3：采用统一文本和图像掩码的文档人工智能预训练。arXiv 预印本 arXiv:2204.08387（2022年） 17. Journet, N., Eglin, V., Ramel, J.Y., Mullot, R.：古代印刷文档的文本/图形标注。载于：国际文档分析与识别会议论文集。第 1010–1014 页（2005年） 18. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.：无 OCR 的文档理解 Transformer。载于：计算机视觉——ECCV 2022，第 27 届欧洲会议，2022年，17 23 498 517 2022年：以色列特拉维夫，10 月——论文集，第 XXVIII 部分。第 17–23 页。施普林格（2022年） 19. Kise, K., Sato, A., Iwata, M.：使用区域 Voronoi 图进行页面图像分割。《计算机视觉与图像理解》**70**(3)，370–382（1998年） 20. Lee, Y., Hwang, J.w., Lee, S., Bae, Y., Park, J.：用于实时目标检测的高能效与 GPU 计算高效骨干网络。载于：IEEE/CVF 计算机视觉与模式识别会议研讨会论文集。第 0–0 页（2019年） 21. Li, F., Zhang, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y., 等：Mask DINO：面向目标检测与分割的统一 Transformer 框架。arXiv 预印本 arXiv:2206.02777（2022年） 22. Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., Wei, F.：DiT：文档图像 Transformer 的自监督预训练。载于：第 30 届 ACM 多媒体国际会议论文集。第 3530–3539 页（2022年） 23. Li, K., Wigington, C., Tensmeyer, C., Zhao, H., Barmpalios, N., Morariu, V.I., Manjunatha, V., Sun, T., Fu, Y.：跨域文档目标检测：基准测试套件与方法。载于：IEEE 计算机视觉与模式识别会议论文集（2020年） 24. Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., Li, Z.：TableBank：面向基于图像的表格检测与识别的表格基准测试。载于：第十二届语言资源与评估会议论文集。第 1918–1925 页（2020年） 25. Li, X.H., Yin, F., Liu, C.L.：使用卷积神经网络和图模型进行页面分割。载于：文档分析系统：第 14 届 IAPR 国际研讨会，DAS 2020，中国武汉，2020年7月26–29日，论文集 14。第 231–245 页。施普林格（2020年） 26. Li, Y., Qian, Y., Yu, Y., Qin, X., Zhang, C., Liu, Y., Yao, K., Han, J., Liu, J., Ding, E.：Structext：利用多模态 Transformer 进行结构化文本理解。载于：第 29 届 ACM 多媒体国际会议论文集。第 1912–1920 页（2021年） 27. Lin, G.S., Tu, J.C., Lin, J.Y.：基于 RetinaNet 和迁移学习的关键词检测，用于保护文档图像中的个人信息。《应用科学》**11**(20)，9528（2021年） 28. Lin, T.Y., Doll´ar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.：用于目标检测的特征金字塔网络。载于：IEEE 计算机视觉与模式识别会议论文集。第 2117–2125 页（2017年） 29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P., Zitnick, C.L.：Microsoft COCO：上下文中的常见对象。载于：计算机视觉——ECCV 2014，第 12 届欧洲会议，2014年，13 6 13 740 755 2014年：瑞士苏黎世，9 月——论文集，第 V 部分。第 740–755 页。施普林格（2014年） 30. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.：Swin Transformer：采用移位窗口的分层视觉 Transformer。载于：IEEE/CVF 国际计算机视觉会议论文集。第 10012–10022 页（2021年）

31. Mathur, P., Jain, R., Mehra, A., Gu, J., Dernoncourt, F., Tran, Q., Kaynig-Fittkau, V., Nenkova, A., Manocha, D., Morariu, V.I., 等：Layerdoc：在富视觉文档中逐层提取空间层级结构。载于：IEEE/CVF冬季计算机视觉应用会议论文集。第3610–3620页（2023年） 32. Mo, S., Sun, Z., Li, C.：用于自监督视觉变换器的多层级对比学习。载于：IEEE/CVF冬季计算机视觉应用会议论文集。第2778–2787页（2023年） 33. Oliveira, S.A., Seguin, B., Kaplan, F.：dhsegment：一种用于文档分割的通用深度学习方法。载于：ICFHR（2018年） 34. Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A.S., Staar, P.：Doclaynet：一个用于文档版面分割的大型人工标注数据集。载于：第28届ACM SIGKDD知识发现与数据挖掘会议论文集。第3743–3751页（2022年） 35. Powalski, R., Borchmann, Ł., Jurkiewicz, D., Dwojak, T., Pietruszka, M., Pałka, G.：利用文本-图像-布局变换器，在文档理解中全力以赴。载于：文档分析与识别——ICDAR 2021：第16届国际会议，瑞士洛桑，2021年9月5–10日，论文集，第二部分16。第732–747页。施普林格（2021年） 36. Prasad, D., Gadpal, A., Kapadni, K., Visave, M., Sultanpure, K.：Cascadetabnet：一种从基于图像的文档中端到端检测表格并识别其结构的方法。载于：CVPRW。第572–573页（2020年） 37. Ren, S., He, K., Girshick, R., Sun, J.：Faster R-CNN：利用区域提议网络实现实时目标检测。载于：NIPS（2015年） 38. Saabni, R., El-Sana, J.：使用缝隙裁剪提取语言无关的文本行。载于：国际文档分析与识别会议论文集（2011年） 39. Saha, R., Mondal, A., Jawahar, C.：文档图像中的图形对象检测。载于：ICDAR（2019年） 40. Schreiber, S., Agne, S., Wolf, I., Dengel, A., Ahmed, S.：Deepdesrt：用于检测文档图像中的表格并识别其结构的深度学习。载于：2017年第14届IAPR国际文档分析与识别会议（ICDAR）。第1卷，第1162–1167页。IEEE（2017年） 41. Shen, Z., Zhang, K., Dell, M.：一个具有复杂版面的历史日文文档大型数据集。载于：IEEE计算机视觉与模式识别会议论文集。第548–549页（2020年） 42. Shen, Z., Zhang, R., Dell, M., Lee, B.C.G., Carlson, J., Li, W.：LayoutParser：一个用于基于深度学习的文档图像分析的统一工具包。载于：国际文档分析与识别会议。第131–146页。施普林格（2021年） 43. Sun, N., Zhu, Y., Hu, X.：结合角点定位的基于Faster R-CNN的表格检测。载于：2019年国际文档分析与识别会议（ICDAR）。第1314–1319页。IEEE（2019年） 44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.：Attention is all you need。《神经信息处理系统进展》**30**（2017年） 45. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.：具有可变形注意力的视觉 Transformer。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第4794–4803页（2022年） 46. Yang, H., Hsu, W.：用于文档版面理解的基于变换器的方法。载于：2022年IEEE国际图像处理会议（ICIP）。第4043–4047页。IEEE（2022年）

47. Yang, H., Hsu, W.H.：使用循环卷积神经网络从科学文献中进行基于视觉的版面检测。载于：2020年 第25届国际模式识别会议（ICPR）。第 6455–6462 页。IEEE（2021年）
48. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.：DINO：改进去噪锚框的 DETR，用于端到端目标检测。arXiv 预印本 arXiv:2203.03605（2022年） 49. Zhong, X., Tang, J., Yepes, A.J.：PubLayNet：迄今最大的文档版面分析数据集。载于：国际文档分析与识别会议论文集。第 1015–1022 页（2019年）