用于临床数据提取的多模态管道：将视觉语言模型应用于输血反应报告扫描

Henning Schäfer[1,2], Cynthia S. Schmidt[1,4], Johannes Wutzkowsky[2], Kamil Lorek[2], Lea Reinartz[2], Johannes Rückert[2], Christian Temme[1], Britta Böckmann[2], Peter A. Horn[1], and Christoph M. Friedrich, *Member, IEEE*[2,3,*]

***Abstract*— 尽管电子健康记录的采用越来越广泛，许多流程仍依赖于纸质文档，这反映了医疗服务提供的异质性现实条件。将基于纸质的数据转移到数字格式时，手动转录过程耗时且容易出错。为了简化这一工作流，本研究提出了一个开源管道，用于从扫描文档中提取和分类复选框数据。在输血反应报告中进行了演示，其设计支持适应其他复选框丰富的文档类型。提出的方法整合了复选框检测、多语言光学字符识别（OCR）和多语言视觉-语言模型（VLMs）。与 2017 年至 2024 年年度编制的黄金标准相比，该管道实现了高精度和召回率。结果是减少行政工作量和准确的监管报告。该管道的开源可用性鼓励对复选框表单进行自托管解析。**

I. 介绍

输血反应报告用于评估输血医学中的患者安全与质量保证 [1] 。尽管许多医疗机构已经转向完全数字化的文件记录系统 [2] ，但是数字化程度差异很大。长期以来的临床工作流程和偏好经常影响纸质文档的持续存在 [3] 。在参与这项研究的德国大学医院，由于将患者和产品特定的标签实际贴在标准化的输血反应表上的实际好处，使工作人员倾向于保持纸质输血反应报告。这在数据必须系统收集和报告时带来了挑战 [4] 。

其中一个行政任务是将汇总的输血反应数据年度报告给像保罗 · 埃尔利希研究所 (PEI) [5] 这样的国家机构。历史上，负责编制这些年度摘要的毕业计划官员执行手动数据输入到结构化数字格式中，并验证每一项输入。这个过程虽然准确，但是费力、耗时，并且可能引入错误。随着数据量的持续增长，这种方法的扩展也变得越来越困难。

从临床角度来看，拥有一个可搜索的、结构良好的输血反应数据集，有助于识别趋势和模式，进而采取有针对性的干预措施以改善患者的结果。在行政和监管层面，收集的数据可以增强基于证据的研究计划。

为了支持这一过程，提出了一种管道方案，该方案将扫描表单上的多语言复选框区域读取为结构化数据，前提是表单上的类别可以被预定义。该管道利用了计算机视觉 (CV) 和自然语言处理 (NLP) 的最新进展：视觉语言模型 (VLMs) [6] 来理解和解释复杂的视觉和文本布局，并通过预定义的类别映射作为一层额外的机制来减少噪音，确保临床发现和疑似诊断被准确分类。虽然评估集中在输血反应报告上，但这些技术具有普适性，可以被其他语言版本的复选框丰富的纸质文档使用，这些语言由 Pixtral 支持 [7] 。

尽管 VLMs 具有潜力，自动化解决方案必须应对扫描伪影、不同形式布局以及非标准化的复选框位置等实际问题。传统的 OCR 技术常常在面对这些复杂性时遇到困难 [8] 。自动化方法必须证明其可靠性、透明性和可解释性，以赢得习惯于完全手动核查的临床工作人员的信任 [9] 。通过借鉴 VLM 文档解析的最新进展，这项工作解决了从扫描文件重建输血反应数据的独特挑战。

通过对跨越 2017 年至 2024 年的输血反应报告语料库进行测试，结果表明该流程能够实现高精度和召回率，与人工工作流非常接近。这不仅减少了人工劳动和潜在错误，还提高了效率。此外，该流程作为一个开源库提供，鼓励更广泛的使用、适配和改进 [1] 。

本文提出了一种通过结合的方法来解决这些挑战的流程：

- 复选框检测算法，用于识别文档中的复选框区域，从而实现后续的数据提取。
- 视觉语言模型（VLMs），用于在复选框区域内整体匹配已选类别，利用视觉和文本理解将提取的文本映射到标准化类别。

本文的其余部分结构如下：II 节介绍相关工作。III 节描述了本研究中使用的材料和数据，IV 节介绍了所使用的方法、基于 VLM 的复选框检测和类别映射，?? 节报告了结果和性能指标以及定性分析，V 节讨论了影响、限制和潜在改进，最后 VI 节总结了贡献和可能的未来方向。

[1]Institute for Transfusion Medicine, University Hospital Essen, Hufelandstraße 55, Essen, Germany.
[2]Department of Computer Science, University of Applied Sciences and Arts Dortmund (FHDO), Emil-Figge Str. 42, Dortmund, Germany.
[3]Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Hufelandstraße 55, Essen, Germany.
[4]Institute for AI in Medicine (IKIM), University Hospital Essen, Girardetstraße 2, Essen, Germany.
[5]Institute of Interventional and Diagnostic Radiology and Neuroradiology, University Hospital Essen, Hufelandstraße 55, Essen, Germany.
[6]Department of Dermatology, University Hospital Essen, Hufelandstraße 55, Essen, Germany.
Contact: christoph.friedrich@fh-dortmund.de

[1]https://github.com/ReMeDi-Blut/Checkbox-Detection-in-Clinical-Documents (最后访问日期：2025 年 7 月 2 日)

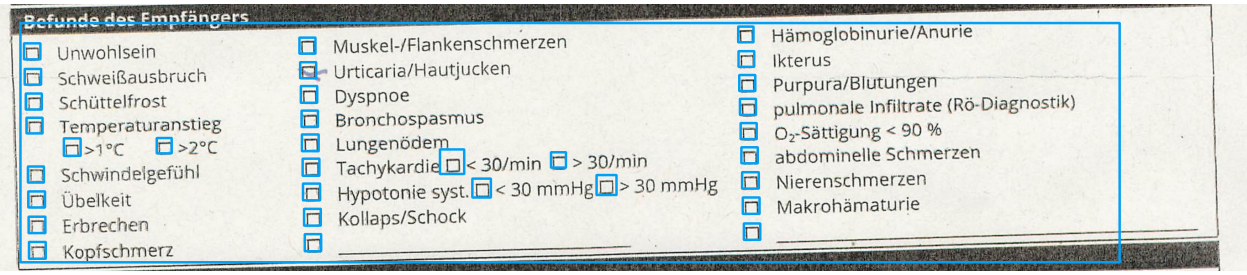

Fig. 1. 展示了基于 YOLO 的复选框检测方法应用于完整的输血反应报告。每个蓝色边界框表示检测到的复选框区域。该模型用于提取相关的连续复选框区域的边界框，以输入到视觉语言模型中。

II. 相关工作

纸质临床文档的数字化是一个长期存在的挑战，其中光学字符识别（OCR）是一项历史悠久的技术 [10] 。OCR 系统通常依赖于手工设计的规则或有限的统计模型 [11] ，难以应对不规则的布局、低质量的图像和特定领域的术语。最近的临床 OCR 工作在专门应用中显示出希望 [12], [13], [14] 。基于深度学习的 OCR 框架，例如使用 LSTM [15] 后端的 Tesseract [16] 或利用卷积和 transformer 架构的商业解决方案 [17] ，提高了准确性和鲁棒性。现有的解决方案仍然采用基于模板的方法或需要大量的手动调整，严重限制了它们的普适性。

A. 用于文档理解的视觉-语言模型

视觉语言模型（VLMs）整合文本和视觉信号，使得图像和文档能够得到整体的解释 [18] 。如 LayoutLM [19] 和 LayoutLMv2 [20] 等模型使用 transformer 架构来对文本内容及其空间布局进行编码，在表单理解数据集上表现优于传统的仅 OCR 流程 [21] 。更近期的方法，例如 Donut [22] 和 PaLM-E [23] ，则采用图像到序列 transformer，直接从像素输入中解析整个文档页面而无需显式的 OCR 预处理，在统一表示中捕捉到文本和图形元素。

传统方法依赖于图像处理启发式或模板匹配来识别选中的框。现代方法则采用经过标注数据集训练的目标检测框架（例如，Faster R-CNN、YOLO、DETR）来稳健地定位复选框并分类其状态。除了复选框之外，表单解析技术还结合了布局分析和表格提取来重建逻辑文档结构。最近的研究表明，结合 VLMs 可以通过利用语义上下文和视觉问答（VQA）提示来总结这些步骤。这相比为单独的表单创建模板而言加快了工作速度。

III. 数据

A. 数据来源和特征

数据集由德国某大学医院在八年期间（从 2017 年至 2024 年）收集的年度输血反应报告组成。每个条目对应一个扫描的文件，其中包含多个复选框用于指示发现结果和疑似诊断，如图 1 所示。除了这些类别之外，这些文件还包含在报告时放置在表格上的物理标签。

总共有 387 个经过验证的输血反应被包括在内。这些反应与 488 个输血血液产品相关，表明有些情况下多个单位共同导致了报告的反应。平均每份报告包含大约 1.26 个血液产品和 3.9 项报告的发现。数据集中反映了性别分布平衡，并涵盖了广泛的时间范围，确保了文档实践的代表性变异性。因为只有当工作人员启动怀疑输血反应的协议时才会生成记录，而这在日常实践中是一个发生率较低的事件，所以绝对数量仍然有限。多年来，表格有一些轻微的变动，例如增加了额外类别，需要配置更大的参考词典。

这些报告由毕业计划官员每年编制和汇总，他手动提取相关信息以完成强制性的报告义务。为了开发和验证自动提取方法，收集了由毕业计划官员制作的年度汇总表。该数据包含所有输血反应的汇总计数及其分类结果以及疑似诊断，作为可以与提取数据进行比较的黄金标准。

大约有 10 % 的反应报告中出现了手写注释，通常为已勾选的类别提供了额外的细节（例如，为笼统的“不适”指定症状）。只有手写反应且没有相应复选框标记的情况很少且排除在数据集中。

所有患者识别信息均根据相关数据保护法规进行处理，并在伦理批准下进行评估。

IV. 方法

该流程处理从 2017 年到 2024 年选择的扫描文档，首先进行血液产品和患者标识符的验证。然后使用基于 YOLOv8 的检测模型来识别扫描文档中的连续复选框区域（见图 1 ）。此模型是公开可用的，并在使用复制-粘贴增强技术生成的自定义数据集上训练，其中结合文档布局分析以确保复选框的合理位置 [2] 。训练在超过 10,000 张合成图像上进行，验证是在 150 份人工标注的文档上进行的，使用的是 YOLO 格式的标注 [24] 。随后，应用了两种互补的方法来提取和映射识别出的复选框到预定义的类别中，既适用于受者中观察到的发现（例如，不适、出汗），也适用于疑似诊断（例如，溶血反应）。这两种方法均用于评估在处理临床文档中常见挑战的性能，例如模糊或不明确的勾选标记、扫描质量差以及多样的标记风格。

[2]https://github.com/LynnHaDo/Checkbox-Detection（最后访问日期：07.02.25）

A. 条形码检测

每张输血表格都有一个条形码贴纸编码血产品编号。为了准确将报告的输血反应与血产品连接起来，这些条形码使用 pyzbar [3] 进行解码。提取的条形码字符串包括一个 3 位数字的国家代码（例如，德国的 “276”），一个 3 位数字的机构代码，以及一个 9 位数字的内部序列号。为了验证这个号码，还有一个 MOD 11.10 校验位。

在第一个基线方法中，采用了一种基于 OCR 的方法。通过一个可配置的阈值，将 Paddle OCR [25] 应用于由 YOLO 模型检测到的已填选框旁边的每个文本区域。因此，OCR 用于提取与每个选中的复选框相关联的文本标签。为了增加 OCR 输出与预定义类别的匹配概率，计算识别出的文本与每个预定义类别之间的 Levenshtein 距离 [26] 。选择编辑距离最小的类别作为该已填复选框的匹配项。

a) 方法二：基于视觉-语言模型（VLM）的提示： Pixtral-Large-Instruct-2411 [7] 被应用于整个裁剪后的复选框区域。该方法并不是处理单个复选框，而是专注于与发现或疑似诊断相关的复选框所在的整个区域。通过让 VLM 更有效地解释并关注相关的文档片段，这种有针对性的策略减少了噪声和复杂性。

为了在提示中嵌入更多的上下文，在用于发现和怀疑诊断的指令中添加了单独的预定义类别：

VLM Prompt for Findings of the Recipient

You are a medical document analysis assistant to extract the text of marked checkboxes. The following image snippet contains checkboxes from a transfusion reaction report. Each checkbox corresponds to a recipient finding. Possible FINDINGS categories include: [...]. Identify which of these categories are checked in the provided image snippet. Do not include any category that is not checked. Provide the checked categories as a simple list.

User: [Image snippet of the findings checkboxes attached here]

[3]https://github.com/NaturalHistoryMuseum/pyzbar（最后访问日期：07.02.25）

VLM Prompt for Suspected Diagnoses

You are a medical document analysis assistant to extract the text of marked checkboxes. Each checkbox corresponds to a suspected diagnosis. Possible SUSPECTED DIAGNOSIS categories include: [...]. Identify which of these categories are checked in the provided image snippet. Do not include any category that is not checked. Provide the checked categories as a simple list.

User: [Image snippet of the suspected diagnoses checkboxes attached here]

从 2017 年到 2024 年，总共汇编了 387 份输血反应报告，这些报告可用于相关扫描的评估目的。每份报告包括多达 24 种可能的受体发现（例如，发热、不适、荨麻疹）和多达 13 种可能的疑似诊断（例如，过敏反应、溶血反应、TRALI）。

B. 条形码检测

表 I 总结了条形码提取的性能。

TABLE I
条码检测和验证结果

Metric	Value
Blood Product Barcode Accuracy	93.21 %
Patient Sticker Barcode Accuracy	89.75 %

虽然这些关于条形码检测的结果不会影响复选框检测，但它们仍在此报告，以提供对扫描中血液产品和患者标签条形码检测准确性的实际评估。

C. 发现和疑似诊断类别映射

比较了两种将这些已审核的类别提取并映射到预定义标签的方法：

- OCR + Levenshtein 匹配：使用 PaddleOCR 读取由 YOLO 模型检测到的每个已填选框区域中的水平文本，然后使用 Levenshtein 距离将提取的文本匹配到最接近的预定义类别。
- 基于 VLM 的提示：一个视觉-语言模型（Pixtral-Large-Instruct-2411）被提供了复选框区域的图像片段及一个已知类别的列表；它根据提示推断哪些类别被选中了。

表 II 总结了基于 OCR 的方法。虽然编辑距离匹配有助于纠正轻微的文本失真，但如果 OCR 输出严重损坏或截断，一些扫描样本可能会增加找到正确类别的混淆风险。

表 III 展示了基于 VLM 方法的性能。尽管类别众多，该方法在应对淡化或部分标记的复选框以及轻微扫描失真方面仍保持较强的鲁棒性。

为了提供关于发现和怀疑诊断的汇总视图，表 IV 报告了每种方法的平均准确率。

TABLE II
光学字符识别 + 莱文斯坦类别提取

Category Set	Precision	Recall	F1-Score
Findings (24)	88.27 %	84.39 %	86.27 %
Suspected Diagnoses (13)	89.04 %	85.46 %	87.21 %

TABLE III
基于 VLM 的类别提取

Category Set	Precision	Recall	F1-Score
Findings (24)	93.21 %	89.24 %	91.18 %
Suspected Diagnoses (13)	94.08 %	91.64 %	92.84 %

V. 讨论

结果表明，所提出的基于 VLM 的方法在扫描输血反应表格上始终优于 OCR + Levenshtein 基准。类目映射中视觉语言管道的高准确性表明了更广泛应用的良好方向。在评估过程中出现的三个主要讨论点：

- 部分勾选标记：基于 VLM 的方法展示了其解释模糊或不完整勾选标记的能力。通过利用整体图像分析而不是单单依赖于检测到的复选框区域的填充阈值，该模型可以在不理想的条件下识别出已勾选框的视觉线索。这种能力尤为重要，因为快速填写表单和不同的操作人员可能会导致勾选行为的不一致。对 24 张标记模糊、被改正或模棱两可的表单进行分析显示，VLM 正确识别了 21 张（87.5 %），而 OCR 仅为 16 张（66.7 %）。
- 扩展的类别集：由于接收者发现最多可达 24 个，疑似诊断可达 13 个，潜在类别的数量增加了这两种方法的难度。分析表明，即使存在因需要区分大量选项而可能引发类别混淆的长上下文提示，VLM 方法依然保持稳健。
- 扫描质量差和伪影：在质量下降的扫描中，例如，污迹和低分辨率，另一个挑战出现了。这些条件导致了 OCR 识别的显著失败。在许多这些困难的情况下，基于 VLM 的方法能够通过利用扫描中保持完整的任何视觉和文本信息来推断正确的选择。这表明 VLMs 可以为文档解析工作流程增加一层韧性，特别是在扫描质量无法保证的真实世界临床环境中。

除了用于数字化以外，该应用程序还可以直接集成到需要强制双重控制原则以确保至少由两名人工操作员进行验证的输血医学工作流程中。该系统可以在此充当第三级验证，通过检测差异并在存在不确定情况时启动审核，以进一步增强安全性。在基础设施不可用或有限的情况下，例如大量伤亡事件或系统故障时，能够快速从纸质表单中提取数据，以支持业务连续性。

TABLE IV
类别映射准确性：VLM vs. OCR + Levenshtein

Approach	Accuracy (Avg)
VLM-Based	92.04 %
OCR + Levenshtein	85.17 %

A. 局限性

尽管性能指标表现强劲，但一些限制需要予以承认。准确性仍然依赖于扫描质量和复选框标记的清晰度。文本严重扭曲或标记方式不寻常的文件可能总是需要人工审核。虽然这样的情况很少见，但它们仍然是潜在错误的来源。没有观察到将严重反应（例如，过敏性休克）误分类为轻微反应的实例，尽管这样的错误会产生重大的临床影响，并需要人类参与的验证进行干预。另一个挑战源于存在非必填手写条目，这不在本次评估研究的范围内：虽然两种方法都可以在某种程度上检测手写文本，但在时间压力下工作的专业人士所写的多语言、术语繁重的笔记极难检测。

尽管 VLM 方法可能适用于各种文档，但当前配置仅通过一种表单进行评估。将此解决方案扩展到其他机构可能需要调整检测和分类组件，以适应不同的布局和类别命名法。虽然这些方法具有通用性，但在其他地方应用它们可能涉及微调模型、添加自定义词典以及修改部分源代码。

VI. 结论

本研究提出了一种技术上可靠的、全自动的流程，用于从纸质扫描件中提取和分类输血反应数据。这个开源流程将多选框丰富的文件转换为结构化的、机器可读的输出。在进行的评估中，结果数据与毕业计划官员每年编制的金标准数据集一致。

在这里，基于 VLM 的流程展示出在减少人工数据输入负担和提高输血反应报告准确性方面的强大潜力。然而，扩展到更复杂的表格并确保在图像畸变下可靠性能仍然需要进一步研究。未来的工作将重点整合更先进的错误校正技术，目标是实现一个全自动的、端到端的解决方案，将结构化数据导入医院信息系统，从而将输血反应与患者其他数据联系起来，使之能够执行诸如预测输血结果与可能出现的不良反应和耐受的风险有关的任务。

开源方法鼓励在专有集成之外进行定制和扩展，例如适应其他临床表格和医学报告任务。所提出的自托管工作流程作为开放和可调整的文档处理解决方案的蓝图，在真实世界的情境中经过验证，希望能够改进质量管理和数据驱动的临床数据收集。

References

[1] R. R. P. de Vries, J.-C. Faber, P. F. W. Strengers, and M. of the Board of the International Haemovigilance Network, "Haemovigilance: an effective tool for improving transfusion practice," *Vox Sanguinis*, vol. 100, no. 1, pp. 60–67, 2011.

[2] L. A. Baumann, J. Baker, and A. G. Elshaug, "The impact of electronic health record systems on clinical documentation times: A systematic review," *Health Policy*, vol. 122, no. 8, pp. 827–836, 2018.

[3] J. J. Saleem, A. L. Russ, C. F. Justice, H. Hagg, P. R. Ebright, P. A. Woodbridge, and B. N. Doebbeling, "Exploring the persistence of paper with the electronic health record," *International Journal of Medical Informatics*, vol. 78, no. 9, pp. 618–628, 2009.

[4] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," *Journal of Business Research*, vol. 70, pp. 263–286, 2017.

[5] B. Keller-Stanislawski, A. Lohmann, S. Günay, M. Heiden, and M. B. Funk, "The German Haemovigilance System–reports of serious adverse transfusion reactions between 1997 and 2007," *Transfusion Medicine*, vol. 19, no. 6, pp. 340–349, 2009.

[6] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-Language Models for Vision Tasks: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5625–5644, 2024.

[7] P. Agrawal, S. Antoniak, E. B. Hanna, B. Bout, D. S. Chaplot, J. Chudnovsky, D. Costa, B. D. Monicault, S. Garg, T. Gervet, S. Ghosh, A. Héliou, P. Jacob, A. Q. Jiang, K. Khandelwal, T. Lacroix, G. Lample, D. de Las Casas, T. Lavril, T. L. Scao, A. Lo, W. Marshall, L. Martin, A. Mensch, P. Muddireddy, V. Nemychnikova, M. Pellat, P. von Platen, N. Raghuraman, B. Rozière, A. Sablayrolles, L. Saulnier, R. Sauvestre, W. Shang, R. Soletskyi, L. Stewart, P. Stock, J. Studnia, S. Subramanian, S. Vaze, T. Wang, and S. Yang, "Pixtral 12B," *CoRR*, vol. abs/2410.07073 v2, 2024.

[8] T. Sato, T. Kanade, E. K. Hughes, M. A. Smith, and S. Satoh, ""Video OCR: indexing digital news libraries by recognition of superimposed captions"," *Multimedia Systems*, vol. 7, no. 5, pp. 385–395, Sep 1999.

[9] A. Chavaillaz, D. Wastell, and J. Sauer, "System reliability, performance and trust in adaptable automation," *Applied Ergonomics*, vol. 52, pp. 333–342, 2016.

[10] C. Thorat, A. Bhat, P. Sawant, I. Bartakke, and S. Shirsath, ""A Detailed Review on Text Extraction Using Optical Character Recognition"," in *ICT Analysis and Applications*, S. Fong, N. Dey, and A. Joshi, Eds. Singapore: Springer Nature Singapore, 2022, pp. 719–728.

[11] S. Mori, C. Y. Suen, and K. Yamamoto, "Historical review of OCR research and development," *Proc. IEEE*, vol. 80, no. 7, pp. 1029–1058, 1992.

[12] Q. Tian, M. Liu, L. Min, J. An, X. Lu, and H. Duan, "An automated data verification approach for improving data quality in a clinical registry," *Computer Methods and Programs in Biomedicine*, vol. 181, p. 104840, 2019, sI: Data Quality Assessment.

[13] E. Murphy, S. Samuel, J. Cho, W. Adorno, M. Durieux, D. Brown, and C. Ndaribitse, "Checkbox Detection on Rwandan Perioperative Flowsheets using Convolutional Neural Network," in *2021 Systems and Information Engineering Design Symposium (SIEDS)*, 2021, pp. 1–6.

[14] M. A. Zaryab and C. R. Ng, "Optical Character Recognition for Medical Records Digitization with Deep Learning," in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 3260–3263.

[15] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] R. Smith, "An Overview of the Tesseract OCR Engine," in *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 629–633.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.

[18] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-Language Models for Vision Tasks: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5625–5644, 2024.

[19] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of Text and Layout for Document Image Understanding," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, R. Gupta, Y. Liu, J. Tang, and B. A. Prakash, Eds. ACM, 2020, pp. 1192–1200.

[20] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, M. Zhang, and L. Zhou, ""LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding"," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 2579–2591.

[21] G. Jaume, H. K. Ekenel, and J. Thiran, "FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents," in *2nd International Workshop on Open Services and Tools for Document Analysis, OST@ICDAR 2019, Sydney, Australia, September 22-25, 2019*. IEEE, 2019, pp. 1–6.

[22] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, "OCR-Free Document Understanding Transformer," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., vol. 13688. Springer, 2022, pp. 498–517.

[23] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, "PaLM-E: An Embodied Multimodal Language Model," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 8469–8488.

[24] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2015.

[25] Y. Du, C. Li, R. Guo, C. Cui, W. Liu, J. Zhou, B. Lu, Y. Yang, Q. Liu, X. Hu, D. Yu, and Y. Ma, "PP-OCRv2: Bag of Tricks for Ultra Lightweight OCR System," *CoRR*, vol. abs/2109.03144 v2, 2021.

[26] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet physics. Doklady*, vol. 10, pp. 707–710, 1965.