广播新闻视频命名实体提取的混合确定性框架

Andrea Filiberto Lucas
马耳他大学人工智能系 马耳他姆西达
andrea.f.lucas@um.edu.mt

Dylan Seychell 马耳他大学人工智能系 马耳他姆西达
dylan.seychell@um.edu.mt

arXiv:2602.09154v1 [cs.CV] 9 Feb 2026

摘要——**基于视频的新闻内容日益增多，因此人们愈发需要透明、可靠的方法来提取屏幕信息。然而，图形布局、排版规范和平台特定设计模式各不相同，使人工索引难以切实可行。本文提出一个全面的框架，用于从广播新闻视频和原生社交媒体新闻视频中自动检测并提取人名。该框架构建了一个经过整理且均衡的标注帧语料库，以呈现当代新闻图形的多样性，并提出一套可解释、模块化的提取流程，旨在确定性且可审计的条件下运行。**

本文将该流程与一类截然不同的生成式多模态方法进行评估比较，揭示了确定性可审计性与随机推断之间明显的权衡。底层检测器的 mAP@0.5 达到 95.8%，表明其在图形元素定位方面具有稳健的实际表现。尽管生成式系统的原始准确率略高（F1：84.18% 对比 77.08%），但它们缺乏新闻和分析语境所需的透明数据血缘。所提出的流程实现了均衡的精确率（79.9%）和召回率（74.4%），避免了幻觉，并确保各处理阶段均具备完整的可追溯性。补充的用户调查结果显示，59% 的受访者表示难以读清快节奏广播中的屏幕人名，凸显了这项任务的实际意义。研究结果为现代新闻媒体中的混合式多模态信息提取确立了一个方法严谨且可解释的基线。

索引词—**计算机视觉、人工智能媒体分析、目标检测、光学字符识别、命名实体识别**

本文已被选中在 *2026*年*IEEE*人工智能会议（*CAI*）上发表。

I. 引言

当代媒体生态系统产生了海量视频内容，其数量日益超出受众和分析人员有效处理的能力 [1], [2]。随着短篇幅新闻形式在广播和在线平台上日益占据主导地位 [3], , 对能够自动提取显著上下文信息的系统的需求也随之加剧 [4]。人名仍是新闻叙事中的核心标识，但通常嵌入在动态的图形元素中，其字体排印、空间布局和时间变化性都给机器准确解读带来了困难[5]–[9]。本研究开展的一项调查发现，59% 的受访者觉得难以阅读屏幕上显示的人名，这凸显了无障碍鸿沟，也说明了开发自动化、可靠的提取机制的必要性。

本研究探讨广播视频图形中人名的自动检测与识别，并在此前视觉信息提取工作的基础上进一步拓展[5]–[9]。基于早期系统 [10], , 本研究提出一种可泛化的视觉命名实体识别方法，将计算机视觉（CV）、光学字符识别（OCR）和语言实体建模整合到确定性且可审计的处理流程中 [11]。与此同时，本研究还考察了当代生成式多模态方法的性能特征，揭示了透明、遵循规则的推理与随机生成式推理之间的方法论权衡。

本研究的主要贡献如下：
1）构建了**新闻图形数据集（NGD）**，这是一个经过整理的标注帧语料库，涵盖广播新闻来源和原生于社交媒体的新闻来源中多样的图形表达惯例。2）开发了**可解释的视觉人名提取流程**，其设计围绕确定性处理阶段展开，能够实现透明的数据血缘，并在异构图形环境中稳健地识别人名。3）对**确定性方法与生成式多模态方法**进行了系统的比较分析，量化了可审计性、可靠性与原始提取准确率之间的实证权衡。

II. 相关工作

从新闻视频图形中自动提取文本信息，建立在目标检测（OD）、光学字符识别（OCR）、命名实体识别（NER）以及近期多模态学习等既有工作的基础之上。下三分之一字幕、字幕条和标题等元素充当视觉锚点，传达姓名、职务和地点 [12]–[14]。它们在字体排印、颜色、

本研究是透明媒体监测与理解（*TMMU*）项目的一部分。该项目由 Xjenza Malta 和马耳他数字创新管理局（MDIA）代表科学与技术基金会，通过 FUSION：研究与创新主题计划下的数字技术计划提供资助。

布局和动态效果上因广播机构而异，这给检测和解读带来了重大挑战 [15], [16]。元素重叠、半透明图层以及风格设计上的差异，使分割和识别变得复杂，凸显了对领域专用数据集和稳健处理流程的需求，以适应广播媒体在时间和视觉上的多样性。

现有的视频分析语料库广泛涵盖广播素材，但通常缺少研究图形叠加层所需的细粒度标注，无法满足视觉文本提取任务所要求的详细程度。电视新闻档案库 [17] 以及 Lee 等人 [18]和 Guha 等人 [19] 推出的数据集收录了大量新闻视频，但缺少明确的边界框标注，也未能体现当代屏幕图形所特有的风格多样性。**新闻图形数据集（NGD）**旨在填补这一空白，提供一组经过筛选且精确标注的帧，反映马耳他语、国际新闻来源以及原生社交媒体新闻来源在排版、空间布局和主题方面的差异。这使得研究人员能够系统、可复现地评估从动态视觉元素中提取和解读文本信息的方法。

在视觉元素检测领域，YOLO 等单阶段卷积框架凭借其统一的定位与分类设计，已证明效果良好 [20]–[22]。包括 YOLOv12 在内的后续版本引入了注意力机制和多尺度预测，提升了密集广播场景中的准确率 [23], [24]。Faster R-CNN 等两阶段检测器 [25], [26] 具有较高的精确率，但计算成本也更高，因此不适用于实时部署或资源受限的部署环境。通过基于感知哈希或特征描述符的帧采样和去重技术，还能进一步提高效率 [27]–[29]。

光学字符识别是视频文本提取的核心组成部分。灰度转换、自适应阈值处理和限制对比度自适应直方图均衡化等预处理技术能够提升文本的易读性，但风格化、低分辨率或动态文字仍难以被可靠识别 [30], [31]。透明的开源引擎（包括 Tesseract） [32], [33],具有较强的适应性，可针对特定领域进行调优；而高性能云端系统已被证明能够提高准确率，但代价是增加计算开销和隐私风险 [34]。在后续处理中，命名实体识别模型会从含噪的光学字符识别结果中识别并验证人名。基于 Transformer 的架构（包括 BERT 和 GLiNER）始终优于早期的条件随机场或长短期记忆网络方法，并支持多语言泛化 [35],，尽管光学字符识别噪声仍是一个主要限制因素[36]。

近期的多模态生成系统通过联合分析视觉与文本线索扩展了这些能力 [37],[38]。这些模型在解读复杂图形布局方面展现出灵活性，但其随机且非确定性的行为引发了对事实一致性、可复现性和透明度的担忧 [39]。因此，将确定性目标检测、光学字符识别以及命名实体识别处理流程与多模态推理或验证相结合，已成为一种切实可行的折中方案。这些方法融合了确定性组件的可解释性与生成模型的上下文适应能力，以提高在异构广播环境中的准确率与鲁棒性。

III. 方法

本研究采用多阶段方法框架，涵盖数据集构建、模型训练和比较评估。NGD 旨在捕捉广播新闻和社交媒体新闻叠加层在风格与结构上的多样性。研究人员使用该数据集训练了一组基于 YOLO 系列模型的检测器，并以此作为准确姓名提取流程（ANEP）的基础。研究通过与当代生成式多模态系统进行评估，并采用两种不同的多模态基线，建立了一个公平的鲁棒性和可靠性基准。ANEP 将更广泛的命名实体提取任务细化为专注于新闻图文内容中出现的人名。

NGD 可通过 Roboflow[1]，公开获取，而 ANEP 的实现可在 GitHub[2]上访问。

A. 新闻图形数据集（*NGD*）

NGD 包含 300 个视频，来源涵盖本地、国际和社交媒体新闻媒体，并将每个来源的占比限制在 12.5%，以确保代表性均衡。我们通过时间分段和感知相似度分析选取帧，随后进行人工验证，以确保数据多样性和标注质量。最终数据集包含 1,500 帧，以及涵盖六个类别的 4,749 个区域级标注：突发新闻图形、数字屏幕图形、下三分之一字幕、标题、字幕条和其他新闻图形。所有图像均标准化为 640×640 px，并通过受控调整亮度、曝光度和噪声进行增强。帧提取后，采用分层的 93% / 4% / 3% 训练—验证—测试集划分。尽管划分是在帧级别进行的，时间分段和感知去重显著减少了近似重复帧；此外，广播图形通常变化较大，因此训练—测试泄漏的风险较低。

B. 基于*YOLO*的目标检测

在 NGD 上评估了 YOLO 系列检测器的多个变体，包括 v8、v11、v12 和神经架构搜索（NAS）。YOLOv12 的平均精度均值最高，因此被选用于后续处理。模型从头开始以 640×640 px 的尺寸进行训练，采用余弦退火学习率和早停策略。YOLOv12 的注意力导向设计提升了对小型及重叠图形元素的定位能力，这些元素在广播视频中十分常见。最终的检测结果用于确定文本抽取的感兴趣区域。

[1]https://universe.roboflow.com/ict3909-fyp/news-graphic-dataset [2] https://github.com/AFLucas-UOM/Accurate-Name-Extraction

C. 准确姓名提取流程（*ANEP*）

ANEP 分为五个依次进行的阶段：

1）**帧处理**，以1帧率对帧进行采样，并通过哈希去除冗余；2）**图形检测与预处理**，先应用YOLOv12，再进行对比度增强和自适应阈值处理；3）**文本抽取**，执行光学字符识别并根据置信度过滤输出；4）**命名实体识别与验证**，将经过微调的基于Transformer的NER模型与由启发式过滤器支持的补充语言处理流水线相结合；以及5）**姓名聚类与时间线生成**，使用模糊、杰卡德和基于嵌入的相似度合并变体，以生成规范姓名和时间出现概况。

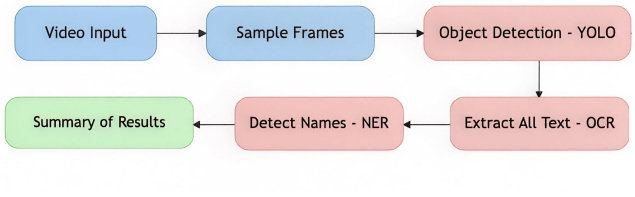

图1. ANEP的简化工作流程，展示了从视频摄入到目标检测、文本抽取、姓名识别以及最终结果综合的顺序推进过程。

该设计可访问透明的中间输出，并在每个处理阶段支持可解释性，从而便于进行诊断性评估和领域特定优化。

D. 生成式人工智能多模态流水线

为进行比较评估，我们实现了两种生成式多模态处理流程。第一种采用 Gemini 1.5 Pro，第二种使用 LLaMA 4 Maverick。两个系统都使用结构化提示进行多模态抽取，仅要求提取真实世界人名。视频帧经感知哈希去重后，以 Base64编码图像的形式提交，用于文本检测和推理。输出以 JSON 格式返回，并采用与 ANEP 相同的指标进行评估。

Gemini 处理流程推理速度更快，精确率也较高，而 LLaMA 4 Maverick 展现出相当的推理能力。不过，这两个系统的延迟变化较大，可复现性有限。相比之下，ANEP 能够对其目标检测、光学字符识别和命名实体识别处理流程的每个阶段进行明确控制，确保输出一致，并使显著性可验证。

IV. 结果评估

评估采用分类、目标检测和姓名抽取的标准指标，以评估基于YOLO的模型的准确性、鲁棒性和效率，ANEP 以及生成式多模态基线。定量分析和对比基准测试对各方法进行了关键评估，指出了其优势与局限性。

A. 目标检测模型的比较性能分析

表 I 汇总了代表性训练运行的性能。基于 YOLOv12(n) 的外部训练配置取得了出色的结果，包括 93.8% 的 mAP@0.5、91.6% 的精确率和 90.8% 的召回率。不过，整体性能最高的是本地训练的 YOLOv12(m)，其 mAP@0.5 达到 95.8%，精确率为 93.9%，召回率为 93.5%。本地训练的 YOLOv8(m) 取得了相近的 mAP@0.5，为 93.7%，但召回率较低，为 86.9%，表明其检测所有目标实例的一致性较差。YOLOv11(n) 的召回率也颇具竞争力，达到 90.4%，但定位准确率略低，mAP@0.5 为 93.1%。YOLOv12(n) Reflect 变体表现不佳，mAP@0.5 为 91.8%；而 YOLO-NAS(n) 的泛化能力最弱，精确率为 85.1%，召回率为 84.3%。

YOLOv12(m) 的优异表现凸显了提升模型容量并针对 NGD 的特性进行训练所带来的优势。其 95.8% 的 mAP@0.5 表明该模型在新闻图形检测方面具有可靠的实际表现，也证明了其适用于实际部署。图2 中相应的损失和置信度曲线显示，模型收敛稳定，并且在整个训练过程中始终保持精确率与召回率之间的均衡。

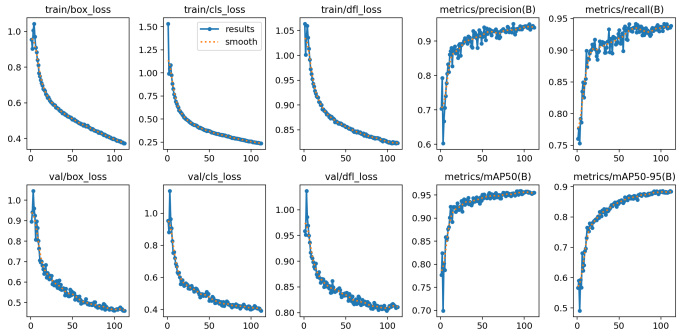

图2. YOLOv12(m) 的训练和验证曲线，展示了边界框损失、分类损失和目标损失的收敛情况，以及精确率、召回率、mAP@0.5 和 mAP@0.5:0.95 指标。结果表明，优化过程稳定，各轮次的性能持续提升，说明模型具有良好的泛化能力，且精确率-召回率表现均衡。

B. 模型注意力与泛化的可视化分析

研究人员采用可视化解释方法，考察检测器的内部行为和泛化能力。Grad-CAM 分析（图3）显示，文本密集区域内的激活较强，尤其是在下三分之一图形和字幕条区域，表明模型捕捉到了与任务相关的空间和语义特征。

表I 在NGD上训练的基于YOLO的目标检测模型的性能比较。表现最佳的配置**YOLOv12*(m)***在所有指标上取得了最高的总体准确率。

模型	精确率（%）	召回率 (%)	mAP@0.5（%）	mAP@0.5:0.95（%）	轮次
本地训练的模型					
YOLOv12*(m)*	**93.9**	**93.5**	**95.8**	**88.7**	**102**
YOLOv8*(m)*	92.6	86.9	93.7	75.2	47
外部训练的配置					
YOLOv12*(n)*	91.6	90.8	93.8	85.4	120
YOLOv11*(n)*	91.2	90.4	93.1	84.9	100
YOLOv12*(n)* (Reflect)	91.4	85.7	91.8	80.4	72
YOLO-NAS*(n)*	85.1	84.3	91.0	61.0	51

研究人员使用此前未见过的广播素材评估泛化能力。如图4所示，检测器成功识别了多种格式的图形叠加层。在更具挑战性的条件下，例如存在眩光的拍摄屏幕（图5），检测器也保持了可靠的性能，不过字幕条和下三分之一字幕等相邻元素之间偶尔会发生合并。

这些结果表明，检测器无需额外训练即可有效迁移到新的广播领域，这对于大规模且持续演进的媒体监测环境至关重要。



图 3. 代表性帧的 Grad-CAM 可视化，展示原始输入（左）、YOLOv12(m) 第 1 层生成的激活图（中）和叠加图（右），突出显示了对富文本区域的高度关注。

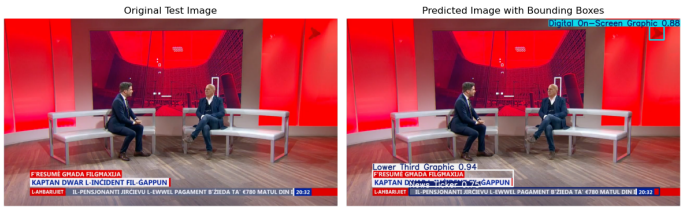

图 4. 来自 TVM 新闻广播的样本检测结果，展示了对所有图形叠加元素的精准定位，以及对所有图形的准确边界框检测。

C. 姓名抽取对比性能

评估以人工构建的真值为基准，其中列出了测试视频中的所有具名人物，并标注了其首次和最后一次出现的帧级时间戳。系统输出通过不区分大小写匹配并采用模糊容差进行对齐，以减轻光学字符识别结果的变化带来的影响；正确性则依据时间重叠

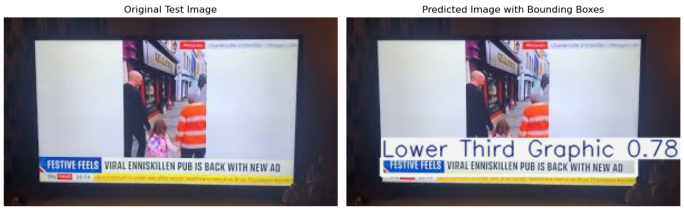

图5. 从电视显示屏拍摄的天空新闻画面的检测示例。尽管存在眩光和压缩伪影，模型仍成功识别了主要图形区域。

表II 姓名抽取模型的对比性能。

模型	精确率	召回率	F1 分数 (%)	运行时间 (s)
ANEP	79.92	74.44	77.08	542.15
Gemini 1.5 Pro	**93.33**	**76.67**	**84.18**	**94.68**
LLaMA 4 Maverick	66.67	50.00	57.14	140.18

进行判定。我们计算了精确率、召回率、F1 分数和运行时间，以评估准确率和效率。所有方法均在 MacBook Air M2 上测量运行时间，并在同一台设备上以统一的 1 FPS采样率进行评估，以确保结果可比。如表II所示，这些结果对三种处理流程进行了全面比较。

1) 定量结果与讨论： 基于 Gemini 1.5 Pro模型构建的生成式多模态基线取得了最优的整体性能，准确率为 93.33%，召回率为 76.67%，F1 分数为 84.18%。其平均运行时间为 94.68 秒，表明其具有良好的计算效率，并显示出适用于近实时监控的潜力。该模型的多模态推理有助于在各种图形布局中实现可靠的提取。尽管准确率较高，但该方法属于黑箱系统，不会公开中间表示，因此限制了其可解释性和可追溯性。

ANEP 的精确率（79.92%）和召回率（74.44%）较为均衡，F1 分数为 77.08%。其性能主要受时间错位和 OCR引起的字符串变化限制。其平均运行时间为 542.15 秒，原因在于

该多阶段流水线按顺序执行，其中包括 1 帧率的帧采样、去重、光学字符识别、基于 Transformer 的命名实体识别和聚类。聚类和别名解析模块因涉及成对相似度计算而成为主要耗时环节。尽管存在这些开销，模块化设计仍能提供透明的中间输出，有助于在新闻和取证语境中开展诊断分析并确保可审计性。

基于 LLaMA 的基线表现最弱，召回率仅为 50.00%，F1 分数为 57.14%。其平均推理时间为 140.18 秒，短于 ANEP，但频繁的误识别和漏检阻碍了信息提取。该模型难以处理重叠或视觉质量受损的文本，这表明，若未经过针对性适配，通用语言模型并不足以胜任结构化视觉信息提取任务。

总体而言，结果凸显了准确率、透明度与计算成本之间的权衡。这些发现进一步表明，对于要求信息提取过程可验证、可问责的应用，模块化且透明的处理流程仍具有重要价值。

2) 错误分析与局限性：对系统输出的分析发现了若干反复出现的错误来源。在 ANEP 中，对于可独立作为英语词汇出现的姓氏（如“Swift”或“Trump”），以及包含变音符号的姓名（如 Ċensu、Żahra 或 Ġużeppi），命名实体识别不一致的情况十分常见。这些问题反映出光学字符识别对正字法变体的敏感性，也体现了常用训练语料库对某些地区姓名的收录有限。别名解析模块成功合并了明显的变体，例如“Donald Trump”和“President Trump”，但在处理间接指称或严重失真的光学字符识别结果时，往往无法将其对应起来，例如“Mickey Rourke”和“Brother Rourke”。

基于 LLaMA 的基线受到底层模型上下文窗口的限制，在较长的视频片段上表现较差。当输入超过约五分钟时，截断效应会削弱时间一致性，并妨碍对具名人物的跟踪，从而限制其在连续广播监测中的适用性。

所有受评估的处理流程在包含图形叠加层重叠、镜头快速切换或风格化视觉布局的场景中，错误率都更高。光学字符识别伪影往往会传播至下游组件，并在实体匹配时同时造成漏检和误报。缺少音频或面部线索等额外模态，也进一步限制了在视觉复杂或信息密集场景中的消歧能力。

D. 用户调查结果

为结合当代媒体消费习惯，明确自动化姓名提取的设计要求，我们对 413 名参与者开展了调查。参与者通过社交媒体平台和在线分发渠道招募，调查以结构化在线问卷的形式进行，问卷包含多项选择题和李克特量表题。参与者的人口统计特征较为广泛，但主要集中在 18 至 34 岁的成年人，男女受访者比例接近均衡，反映的是一般数字新闻消费群体，而非特定领域的专业群体。

调查回复证实，图形叠加层仍是新闻视频内容中重要的上下文信息来源。多数参与者（59%）表示，难以读清屏幕上显示的人名；显示速度、字体变化和视觉清晰度较低被认为是主要原因。许多受访者指出，他们经常暂停或倒回视频以确认某人的身份，这进一步表明，亟需能够提升动态新闻环境中文本信息无障碍性的自动提取系统。

对自动化的态度显示出谨慎的支持。约62.2%的受访者表示，他们对自动化系统的信任将取决于其运行的透明度以及负责其开发的机构的可信度。这些发现凸显了设计优先考虑问责制、可解释性和用户信心的解决方案的重要性，尤其是在事实准确性和公众信任至关重要的新闻语境中。

V. 结论

本研究考察了从新闻视频图形中自动提取人名的问题，该任务需要视觉检测、文本识别和语言学解释的协同运作。所提出的NGD经过整理，以体现当代广播和社交媒体叠加层的风格多样性。这支持了ANEP的开发，这是一个确定性的模块化框架，在透明且可审计的处理流水线中整合了目标定位、字符提取和命名实体消解。针对生成式多模态基线的比较评估，在确定性可审计性与生成式多模态系统所特有的随机推理之间建立了严格的基准。

基于 Gemini 的生成式流水线取得了最高的综合 F1 分数（84.18%），而 ANEP 的精确率和召回率较为均衡（分别为 79.9% 和 74.4%），并提供了明确的数据血缘和可验证的决策路径。底层检测器展现出较强的跨域鲁棒性（mAP@0.5 为 95.8%），表明其适合部署于异质的新闻环境中。补充开展的用户调查也指出，透明度是建立公众信任的决定性因素，这进一步凸显了在新闻业和问责导向型应用中采用可解释系统的必要性。

尽管新型多模态系统不断涌现，其概率式端到端架构在透明度、可复现性和可解释性方面仍存在根本局限。即使这些系统在提取准确率上有所提升，这些限制仍使其不适用于新闻工作和证据用途。本文结果表明，生成式模型在主要指标上可以超越确定性处理流程；然而，在要求数据血缘可验证的场景中，ANEP 框架仍不可或缺。

目前的局限包括：对非拉丁文字体系的处理性能较低，光学字符识别组件容易受到风格化或动画字体排印的影响，以及特定处理阶段依赖外部 API。未来工作应扩大对多语言及跨文字体系语境的覆盖范围，包括从右向左书写的文字和区域字母表，并整合更丰富的语义元数据，例如所属机构和发言者角色。能够联合建模音频、文本和视觉的多模态 Transformer 架构，是实现实时身份消歧的一条有前景的方向。剪枝、量化和边缘部署等以效率为导向的技术，或可降低直播场景中的计算开销。进一步探索自监督和置信度感知学习策略，有望提升大型视频档案中的稳健性和可扩展性。

总体而言，研究结果表明，与最先进的生成式基线进行评估后，透明的目标检测、光学字符识别和命名实体识别流程能够为自动化姓名提取提供准确、可解释且具备实际运营可行性的解决方案。这推动了媒体分析、无障碍和信息问责制等更广泛目标的实现。

参考文献

[1] P. Aussu, “信息过载：应对机制与工具的影响”，载于商业信息处理讲义，2023 年，第 661–669 页。[2] P. G. Roetzel, “信息时代的信息过载：采用文献计量方法与框架构建，回顾工商管理、商业心理学及相关学科的研究文献”，*BuR - 商业研究*，第 12 卷，522，2018 2，479 页。[3] A. A. Naz 和 R. A. Akbar，“媒体在有效教学中的应用及其重要性：一些考量”，《基础教育杂志》，第 18 卷，第 1–2 期，第 35–40 页，2008 年。[4] M. Kenely、D. Seychell、C. J. Debono 和 C. Porter，“用于视觉注意力预测与新闻采访画面分析的深度学习框架”，载于*2025 年 IEEE 人工智能会议（CAI）*，美国加利福尼亚州圣克拉拉，2025 年，第 278–283 页。[5] K. Choroś，“自动视频索引器 AVI 中的视频结构分析与基于内容的索引”，载于多媒体与网络信息系统技术进展。施普林格，2010 年，第 79–90 页。[6] S. Lee 和 K. Jo，“新闻采访视频序列中自动人物索引与检索系统的策略”，载于*2017 年第 10 届国际人机系统交互会议（HSI）*，2017 年，第 212–215 页。[7] H. Zhang 和 Y. Gong，“新闻视频的自动解析”，载于*IEEE 多媒体计算与系统国际会议论文集*，美国马萨诸塞州波士顿，1994 年，第 45–54 页。[8] M. Ghoniem、D. Luo、J. Yang 和 W. Ribarsky，“Newslab：探索式广播新闻视频分析”，载于*IEEE VAST 论文集*，2007 年，第 123–130 页。[9] M. S. Pattichis、V. Jatla 和 A. E. U. Cerna，“应用于视频分析系统的机器学习方法综述”，*arXiv 预印本 arXiv:2312.05352*，2023 年。[10] D. Seychell、G. Hili、J. Attard 和 K. Makantatis，“人工智能作为促进公平新闻业的工具：来自马耳他的案例研究”，载于*2024 年 IEEE 人工智能会议（CAI）*，2024 年，第 127–132 页。[11] J. Attard 和 D. Seychell，“用于自动视频分割的图像、视频与音频分类器比较分析”，*arXiv 预印本 arXiv:2503.21848*，2025 年。[12] O. Järvi，“新闻图形：若干类型学与文本特征”，*LSP 与专业传播*，第 2 卷，第 2 期，第 8–22 页，2002 年。[13] J. S. Foote 和 A. C. Saunders，“网络电视新闻中的图形形式”，《新闻学季刊》，第 67 卷，第 3 期，第 501–507 页，1990 年。[14] L. Buijs、Y. de Haan 和 G. Smit，“在媒体中运用信息可视化”，载于会议论文，2013 年。 [在线]。网址：https://www.internationalhu.com/research/publications/using-information-visualization-in-the-media

[15] R. Borgo 等，“基于视频的图形与视频可视化综述”，载于 Eurographics 2011 - State of the Art Reports，N. John 和 B. Wyvill 编。Eurographics Association，2011。[16] N. H. Li、D. Doermann 和 O. Kia，“数字视频中的自动文本检测与跟踪”，IEEE Transactions on Pattern Analysis and Machine Intelligence，第9卷，第1期，第147–156页，2000。[17] J. Hong、W. Crichton、H. Zhang、D. Fu、J. Ritchie、J. Barenholtz、B. Hannel、X. Yao、M. Murray、G. Moriba、M. Agrawala 和 K. Fatahalian，“对十年美国有线电视新闻中的人脸进行分析”，载于 KDD ’21：第27届ACM SIGKDD知识发现与数据挖掘大会论文集，2021年8月，第3011–3021页。[18] S. H. Lee、J. W. Ahn 和 K. H. Jo，“基于叠加文本的自动姓名条检测与人物索引”，Journal of Multimedia Information System，第2卷，第1期，第163–170页，2015。[19] P. Guha、R. Kannao 和 R. Soni，“电视新闻频道广告检测数据集”，UCI机器学习库，2014，可通过 UCI机器学习库获取。[20] I. Aljarrah 和 D. Mohammad，“使用卷积神经网络进行视频内容分析”，IACS，第1卷，第122–126页，2018。[21] J. Redmon、S. Divvala、R. Girshick 和 A. Farhadi，“You Only Look Once：统一的实时目标检测”，载于 IEEE计算机视觉与模式识别会议（CVPR）论文集，2016，第779–788页。[22] Y. Zhao 和 S. Wang，“基于YOLO算法的实时目标检测研究”，Highlights in Science Engineering and Technology，第7卷，第323–331页，2022。[23] J. Terven、D.-M. Córdova-Esparza 和 J.-A. Romero-González，“计算机视觉中YOLO架构的全面综述：从YOLOv1到YOLOv8和YOLO-NAS”，Machine Learning and Knowledge Extraction，第5卷，第4期，第1680–1716页，2023。[24] Y. Tian、Q. Ye 和 D. Doermann，“YOLOv12：以注意力为中心的实时目标检测器”，2025。 [在线]。网址：https://arxiv.org/abs/2502 .12524[25] R. Sapkota、Z. Meng、M. Churuvija、X. Du、Z. Ma 和 M. Karkee，“YOLOv12、YOLOv11、YOLOv10、YOLOv9和YOLOv8在复杂果园环境中检测和计数幼果的综合性能评估”，arXiv预印本 arXiv:2407.12040，2025。 [在线]。网址：https://arxiv.org/abs/ 2407 12040 G. Preethi、S. Naik、M. Kenchol、S. P. Jakalannanavar 和 M. S. Rachana，“使用 Faster R-CNN、YOLOv7 和 YOLOv8 进行目标检测”，Indiana Journal of Multidisciplinary Research，第4卷，第3期，第136–141页，2024。[27] A. Qin、M. Xiao、B. Huang 和 X. Zhang，“Maze：网络规模下高性价比的视频去重系统”，载于第30届ACM国际多媒体会议论文集，2022，第3163–3172页。[28] N. Lian、Y. Zhang、T. Chen、H. Wang、L. Li 和 J. Wu，“AutoSSVH：探索用于高效自监督视频哈希的自动帧采样”，arXiv预印本，2025。[29] G. U. 等，“基于词袋模型的帧重复检测”，Multimedia Systems，第24卷，第549–567页，2018。[30] K. Hamad 和 M. Kaya，“光学字符识别技术的详细分析”，International Journal of Applied Mathematics Electronics and Computers，第4卷，特刊1，第244页，2016。[31] N. Ashraf、S. Y. Arafat 和 M. J. Iqbal，“OCR方法分析”，International Journal of Computational Linguistics Research，第10卷，第3期，第81–91页，2019。[32] M. Heidarysafa、J. Reed、K. Kowsari、A. R. Leviton、J. Warren 和 D. Brown，“从视频到URL：利用光学字符识别提取用户行为的多浏览器指南”，载于 Advances in Intelligent Systems and Computing，2019。[33] R. Smith，“Tesseract OCR引擎概述”，载于第九届国际文档分析与识别会议（ICDAR），第2卷，2007，第629–633页。[34] K. Smelyakov、A. Chupryna、D. Darahan 和 S. Midina，“面向常见数据源的现代文本识别解决方案与工具的有效性”，International Journal of Computer Applications，2021。[35] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，“BERT：用于语言理解的深度双向变换器预训练”，arXiv预印本 arXiv:1810.04805，2018。 [在线]。网址：https: //arxiv.org/pdf/1810.04805.pdf[36] M. Marrero、J. Urbano、S. Sánchez-Cuadrado、J. Morato 和 J. M. Gómez-Berbís，“命名实体识别：谬误、挑战与机遇”，Computer Standards & Interfaces，第35卷，第5期，第482–489页，2013。

[37] D. Hassabis 和 S. Pichai。（2024）推出 Gemini 1.5：Google 下一代人工智能模型。Google 博客，2024年2月。 [在线]。网址：https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024

[38] A. Vaswani 等，“注意力就是一切”，*arXiv*预印本 *arXiv:1706.03762*，2017。 [在线]。网址：https://arxiv.org/abs/1706 .03762

[39] T. S. Musalamadugu 和 H. Kannan，“用于医学影像分析及应用的生成式人工智能”，*Future Medicine AI*，2023。