打击网络错误信息视频：特征分析、检测与未来方向

Yuyan Bu 中国科学院计算技术研究所 中国科学院大学
buyuyan22s@ict.ac.cn

Qiang Sheng 中国科学院计算技术研究所
shengqiang18z@ict.ac.cn

Juan Cao
计算技术研究所，中国科学院
中国科学院大学科学
caojuan@ict.ac.cn

Peng Qi
中国科学院计算技术研究所
pengqi.qp@gmail.com

Danding Wang
计算技术研究所，中国科学院
wangdanding@ict.ac.cn

Jintao Li
计算技术研究所，中国科学院
jtli@ict.ac.cn

arXiv:2302.03242v3 [cs.CV] 6 Aug 2023

摘要

随着通过在线视频流媒体获取信息日益普及，虚假信息视频对在线信息生态系统的健康构成了新的威胁。尽管此前的研究在检测文本和图像格式中的错误信息方面取得了显著进展，但视频虚假信息给自动检测系统带来了全新且独特的挑战：1）各种模态带来的高度信息异质性；2）误导性视频操纵与无恶意的艺术性视频剪辑之间界限模糊；3）由于推荐系统在在线视频平台上发挥主导作用，虚假信息传播出现了新的模式。为推动这一具有挑战性的任务的研究，我们开展本综述，介绍虚假信息视频检测方面的进展。首先，我们从信号、语义和意图三个层面对虚假信息视频进行分析和特征描述。基于此，我们系统回顾现有检测研究，涵盖各种模态的特征以及线索整合技术。我们还介绍现有资源，包括具有代表性的数据集和实用工具。除总结已有研究外，我们还探讨相关领域并梳理尚待解决的问题与未来方向，以鼓励并指导更多关于虚假信息视频检测的研究。相关代码仓库位于
htt
ps://github.com/ICTMCG/Awesome-Misinfo-Video-Detection。

CCS 概念

• **信息系统** →**多媒体信息系统**。

关键词

虚假信息视频检测；多模态计算；综述

作者也隶属于中国科学院计算技术研究所智能信息处理重点实验室。

允许免费出于个人或课堂用途制作本作品部分或全部内容的数字或纸质副本，前提是这些副本不得用于营利或商业利益，不得为此制作或分发，并且副本须附有本声明及首页上的完整引文。本作品中第三方组件的版权必须得到尊重。其他用途请联系所有者/作者。*MM '23，2023*年*10*月*29*日至*11*月*3*日，加拿大安大略省渥太华© 2023 版权所有，归所有者/作者所有。ACM ISBN 979-8-4007-0108-5/23/10。https://doi.org/10.1145/3581783.3612426

ACM 引用格式：
Yuyan Bu、Qiang Sheng、Juan Cao、Peng Qi、Danding Wang 和 Jintao Li。2023。《打击网络错误信息视频：特征分析、检测与未来方向》。收录于《第*31*届*ACM*国际多媒体会议论文集（*MM '23*）》，*2023*年*10*月*29*日至*11*月*3*日，加拿大安大略省渥太华。ACM，纽约，纽约州，美国，11页。https://doi.org/10.1145/ 3581783.3612426

1 引言

随着在线视频平台日益普及，通过视频流媒体获取信息变得越来越重要。YouTube 和 TikTok 等热门视频分享平台每月吸引了数十亿活跃用户 [66]。关于新闻消费的研究表明，美国 30 岁以下成年人中，约有四分之一经常从这些视频分享平台获取新闻 [37]。

遗憾的是，视频新闻消费的大幅增长也助长了虚假信息视频的迅速传播，对在线信息生态系统构成了日益严峻的挑战。例如，124 个有关新冠疫苗的 TikTok 虚假信息视频获得了超过 2000 万次观看和 30 万次分享 [47]。图 1 展示了一则声称新冠疫苗中含有微芯片的虚假信息视频帖子。与以往主要研究文本和图像格式中的错误信息相比，视频虚假信息更容易误导受众并迅速走红。政治领域的研究表明，当某个事件以视频形式呈现时，人们更容易相信它确实发生过 [79]。另一项实验表明，视频格式会让新闻报道显得更可信，也更有可能被分享 [68]。这种效应使虚假信息视频危害更大，并可能给各种利益相关方带来进一步的负面影响。对个人而言，虚假信息视频更容易误导受众，使其产生虚假记忆并做出基于错误信息的决策。对在线平台而言，虚假信息视频的广泛传播可能引发声誉危机、导致用户活跃度下降，并招致监管审查。对监督机构而言，虚假信息视频可能被用作煽动动荡、甚至削弱民主制度的工具。因此，亟须采取行动，降低虚假信息视频带来的风险。

作为应对措施，在线平台已采取行动，努力遏制虚假信息视频的传播。例如，TikTok

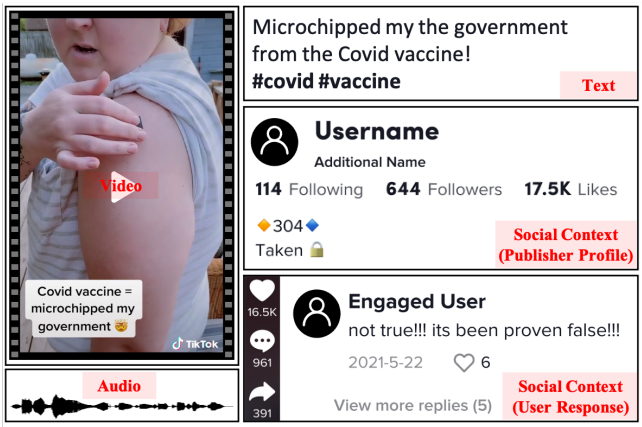


图1：TikTok 上的一条虚假信息视频帖子及其附带的社交背景信息，指出新冠疫苗含有微芯片。出于隐私考虑，我们用占位符替换了用户头像和姓名。

推出了增强版应用内举报功能，以帮助遏制新冠肺炎错误信息的传播 [45]。然而，目前的解决方案在很大程度上依赖专家团队或活跃用户的人力投入，既耗费人力又耗时。它通常无法实现实时检测，因此在新事件出现或此前已被辟谣的错误信息再次传播时，无法迅速做出反应 [51]。此外，这种解决方案可能会引入不确定的个人偏见和错误 [24]。为了解决上述问题，开发虚假信息视频自动检测技术和系统成为一种很有前景的选择。

与基于文本或图文虚假信息检测相比，基于视频的虚假信息检测面临若干独特挑战。首先，多种模态的异构信息激增，给最终预测带来了更多不确定性，甚至噪声。其次，非恶意的视频编辑行为模糊了伪造视频与真实视频之间的界限。第三，在线视频平台推荐主导的内容分发模式，将虚假信息传播从转发等显性行为转变为重新上传等隐性行为。这些挑战要求我们采用新的技术方案来检测视频虚假信息，也凸显了对这一问题进行审慎、具体研究的重要性。

尽管已有许多关于广义错误信息检测的高价值综述，但对视频虚假信息的关注仍然有限。大多数研究将视频视为一种视觉内容，与图像类似，并讨论用于错误信息检测的一般多模态技术 [2, 6, 11]。上述视频虚假信息的独特性并未得到充分考虑。其他相关综述则聚焦于某一特定类型的虚假信息视频，例如伪造视频 [4, 41]，这些综述提供了详细回顾，但缺乏对该问题的综合分析。Venkatagiri等人 [72]讨论了研究视频分享平台上错误信息所面临的挑战及未来研究方向，但未对具体检测方法进行概述。鉴于虚假信息视频可能造成的危害，亟需对这一检测问题开展全面综述。

为改变现状并促进研究社区进一步探索这一具有挑战性的问题，我们在本综述中概述了虚假信息视频检测。我们的主要贡献如下：

- **全面特征分析：**我们从信号、语义和意图三个层面对虚假信息视频的特征进行了全面分析；
- **系统的技术综述：**我们系统综述了现有的多模态错误信息检测技术和方法，并提出了一种原则性方式，对所采用的线索和集成机制进行分类；
- **具体的未来方向：**我们讨论了该领域的若干未解决的问题，并为研究和现实应用场景提出了具体的未来方向。

2 错误信息视频刻画

本节将给出错误信息视频的定义，并从三个层面对其进行分析。遵循 Zhou 和 Zafarani [86]，的做法，我们将错误信息视频定义为：

定义 *2.1*（虚假信息视频）。传播虚假、不准确或误导性信息的视频帖子。

请注意，视频帖子可能不仅包含视频本身。在 Facebook 和 Twitter 等文本主导的社交平台上，帖子可能附带一段文本；而在前文提到的视频主导型平台上，通常会包含标题或简短的文本描述。此外，虚假信息视频这一概念与伪造视频密切相关，但两者并不相同：前者的特点是意图传播虚假或误导性信息，而后者通常指经过 Photoshop 等工具或生成式 AI 技术操纵或制作的视频。伪造视频当然可以用于制作虚假信息视频，但并非所有视频伪造行为都是恶意的（例如用于电影制作）。反过来，未经过操纵、但配有虚假或误导性文本的视频也会被视为虚假信息视频。因此，检测虚假信息视频需要恰当地结合多视角线索。

为了找到有用的检测线索，我们根据虚假信息视频的制作方式对其进行特征描述。从表层到内在，分析分为三个层面：信号、语义和意图。

2.1 信号层面

虚假信息视频通常包含经过操纵或生成的视频与音频内容，而伪造过程往往会在底层数字信号中留下痕迹。能够产生此类痕迹的伪造方法可分为两类：编辑和生成。编辑是指对现有视频与音频模态数据进行视觉改动。典型的编辑操作包括屏幕裁剪、帧拼接、波形剪切、节奏变更等，这些操作可借助编辑软件 [8] 完成。相比之下，生成操作由神经网络执行，这些网络经过训练，能够直接生成完整、生动的视频（大多根据人类动作 [12]或文本 [64]指令生成）。生成的视频可能包含伪造的人脸或声音，以误导观众。

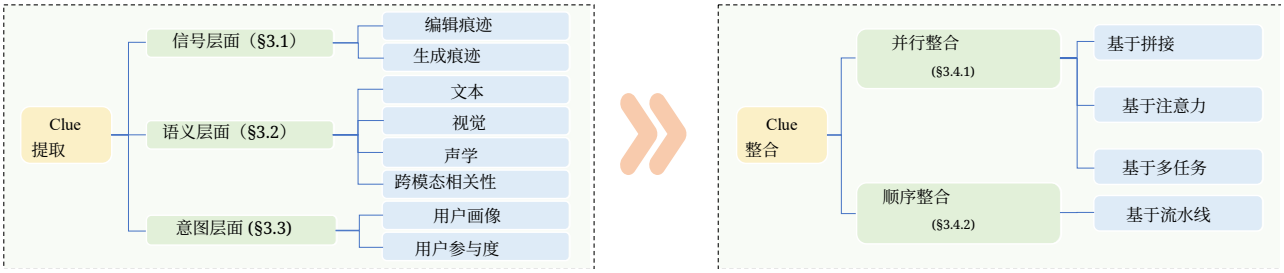


图2：虚假信息视频检测技术概览。

2.2 语义层面

虚假性是通过违背事实的错误语义变化传达的。对于虚假信息视频，这类变化可能出现在某一种特定模态中，也可能出现在多种模态中。前一种情况下，被操纵的元素（例如文本描述中的夸大说法）通过单一模态呈现。后一种情况在多模态情境中更为常见，虚假信息可能通过不同模态中未被伪造的内容之间错误的语义关联来传达。例如，创作者可能上传一段真实的、记录先前发生事件的视频，却配上一段真实描述新近发生事件的文本。

2.3 意图层面

虚假信息的产生通常受到潜在意图的驱动，例如政治影响、经济利益和宣传效果 [77]。为实现这些潜在意图，虚假信息视频通常追求广泛而迅速的传播。这会形成独特的表达、传播和用户反馈模式，与真实视频有所不同。例如，Qi 等人 [51] 发现，与真实新闻视频相比，假新闻视频表现出更明显的情绪偏好，用户参与度更高，也更容易在评论中受到质疑。

3 虚假信息视频检测

虚假信息视频检测问题通常被表述为一个二元分类任务：设 V 和 S 分别表示一个视频帖子及其附带的社交情境。V 由标题/描述 t、视频 v和音频 a等属性构成。社交情境 S 主要由两个部分组成：用户画像 p 和用户参与度 e。用户画像 p 包含一组用于描述上传者账号的特征，例如 IP 地址所示的位置、用户自行撰写且经过验证的简介，以及粉丝数。用户参与度 e 包括评论，以及观看次数、点赞数和收藏数等统计数据。图 1 展示了一个来自 TikTok 的 V 示例。虚假信息视频检测任务旨在根据所有可访问的特征 $E = \{V, S\}$，预测视频帖子 V 是否包含错误信息，即 $F : E \mapsto \{0, 1\}$。现有研究通常采用分类任务中的常见指标，包括准确率、精确率、召回率和 F1 分数，进行性能评估。遵循第 2 节中的特征分析，我们介绍一系列相关研究，探讨如何从不同层级的线索到整合模式检测虚假信息视频。图 2 概述了本节内容。

3.1 信号层面

由于虚假信息视频通常利用伪造技术制作，因此检测视频伪造痕迹可为检测虚假信息视频提供重要线索。相关研究通常归为多媒体取证，过去几十年来一直备受关注。我们分别介绍一些常用于检测编辑痕迹和生成痕迹的技术。鉴于已有关于多媒体取证技术的综述，本文不再详细介绍各项研究。

3.1.1 编辑痕迹。如前所述，编辑是指通过多媒体编辑软件（通常需要手动操作）对视觉或音频内容进行修改。现有的编辑痕迹检测方法主要可分为两类：主动检测和被动检测。

在主动检测方法中，会预先嵌入数字水印或数字签名，并将其提取出来，以检测二次编辑的痕迹。例如，Tarhouni 等人 [69] 提出了一种用于检测的盲式半脆弱视频水印方案。该方案采用组合水印（视频帧水印和音频水印），用于检测两个通道中的篡改。在业界，内容来源与真实性联盟制定了一项基于数字签名技术的技术规范，用于认证媒体内容的来源和真实性。与被动检测相比，主动检测方法通常能够更快地响应，并作出更准确的判断。

然而，实际上大多数视频并未预先嵌入此类附加信息。被动检测方法利用数字视频本身的特性来检测篡改痕迹，因而凸显出其优势。对于视频帧，篡改检测方法会利用帧间和帧内信息。前者检测帧序列中的异常变化，例如视频帧的插入 [85]，删除 [1], 复制 [19], 和乱序 [65] 。后者检测可在单帧中体现的对象篡改，例如区域复制 [5, 9] 和拼接 [31, 56]。此外，压缩过程中产生的伪影、数字摄像机留下的噪声伪影，以及光照、亮度和阴影等不一致性，也会被用作检测线索 [59]。对于音频内容，则利用基于观测属性的统计特征进行伪造检测。其中，电网频率因其围绕标称值随机波动以及电网内一致性的特性，广泛用于取证 [55]。

3.1.2 生成痕迹。 由神经网络（例如生成对抗网络）生成的视频也称为"深度伪造视频"。其中，包含逼真生成的人脸的深度伪造视频已被用于冒充名人，并造成负面影响。过去几年间，视觉深度伪造检测取得了显著进展。参考深度伪造视频的生成方式 [41], ，检测方法利用不同步骤中产生的线索，包括模型指纹 [84], 生物信号 [36], 和时间一致性 [25, 81]。随着文本转语音和语音转换算法的不断进步，生成难以被人类辨别的伪造音频变得更加容易，也因此日益受到关注。为加快深度伪造音频检测的发展，组织了 ASVspoof 挑战赛 [80] 。现有的伪造音频检测研究大多采用手工设计的声学特征，如梅尔频率倒谱系数（MFCC）、线性频率倒谱系数（LFCC）和恒定 Q 倒谱系数（CQCC），并使用高斯混合模型和轻量级卷积神经网络等分类器进行预测。近期研究还尝试利用预训练的 wav2vec 风格模型进行语音表征提取，并构建端到端框架 [35]。

音视频联合深度伪造检测旨在处理视觉和/或听觉模态遭到篡改的情况。检测方法侧重于学习真实视频与伪造视频中视频帧和音频帧之间固有的（非）同步关系 [42, 87]。此外，不同模态也可以相互补充，以作出最终判断 [32]。

鉴于视频压缩广泛应用于在线视频平台的默认上传设置，处理压缩深度伪造视频已成为一个重要问题。压缩操作会抹除某些生成痕迹，从而使视频更难被检测。此外，压缩和解压缩过程中的量化与反量化操作会带来额外的量化噪声和失真。为解决这些问题，Hu 等人 [28] 提出了一种双流方法，通过分析压缩深度伪造视频的帧级和时间级特征进行检测。

3.2 语义层面

尽管信号层面线索能够提供有力证据，但由于便携式编辑工具被广泛使用，这些证据并非决定性的。即使检测到编辑或生成痕迹，也不一定意味着视频传达了错误信息。语义未变的编辑视频的存在，模糊了虚假视频与真实视频之间的界限。例如，为了精简或清晰而编辑的视频，仍可能真实且具有信息价值。相反，未经技术篡改的视频也可能被用于欺骗。与信号层面的线索不同，语义层面线索为识别误导性内容提供了全新视角。因此，近期许多研究着重利用多模态语义线索，不仅从视频内容中提取线索，也从描述性文本信息中获取线索。本节从多模态视角探讨基于语义的方法所利用的特征。

文本特征。 视频内容通常会附带描述性文本信息，例如视频描述和标题。除了这些可直接获取的文本外，从视频中提取的字幕和转录内容也包含有用信息。早期研究大多从这些文本中提取统计特征用于分类。Papadopoulou 等人 [49] 首次利用标题的语言特征，其中包括文本长度等基本属性，以及一些专门设计的指标，例如标题是否包含特定字符（如问号/感叹号以及第一、第二或第三人称代词）、特殊词语（如俚语和情感指示词）的数量，以及可读性分数。其他研究还会考虑特定表达的存在与否，例如标题党短语和暴力词语，以进行检测[33, 48]。Hou 等人 [26] 和 Serrano 等人 [57]利用语料感知特征，如 n 元语法、词频-逆文档频率、词汇丰富度，以及与 LIWC 词典的重合情况。除了人工设计特征，利用深度学习生成的连续表示也日益得到应用。Jagtap 等 [30] 使用 GloVe [50] 和 Word2Vec [40] 生成字幕嵌入。Shang 等人 [58] 和 Choi 和 Ko [13] 训练双向循环神经网络作为文本编码器，对标题、描述和转录等文本信息的语义进行编码。预训练语言模型（如 BERT [17]）的最新进展，也推动最新的多模态检测模型获取上下文化表示 [13, 15, 38, 51, 73]。此外，事件触发词和事件论元等事实元素也被用于提供明确指导，以学习内在的事件语义 [34]。

视觉特征。 视觉内容通常以帧级或片段级方式表示。前者呈现静态视觉特征，后者则呈现额外的时序特征。Shang 等人 [58] 通过均匀采样提取视频帧，并将调整尺寸后的采样帧输入先进的目标检测网络 Fast R-CNN，以获取目标区域的视觉特征。相应的字幕表示用于引导目标区域的整合，从而帮助生成帧视觉表示。Choi 和 Ko [13] 根据视频帧与缩略图的相似度提取视频帧。研究使用预训练的 VGG-19 从视频帧中提取视觉特征。McCrae 等人 [38] 将视频拆分为时长为 32 帧的片段，每个片段均从一个关键帧开始。关键帧通过 FFmpeg 场景检测滤镜检测得到。对于每个片段，分别使用预训练的 FaceNet、ResNet50 和 S3D 网络提取与人脸、物体和活动相关的特征。Wang 等人 [73] 直接将视频拆分为固定时长的片段，并同样使用 S3D 提取视觉特征。Qi 等人 [51] 在帧级和片段级同时表示视觉内容。研究分别使用预训练的 VGG19 模型和预训练的 C3D 模型提取帧特征和片段特征。Liu 等人 [34] 使用预训练的视觉 Transformer ViT [18] 将视频帧编码为特征，该模型应用于 CLIP [54]中。

声学特征。 作为区别于图文虚假信息的独特模态，音频模态（包括语音、环境声音和背景音乐） [51], 在视频信息表达中发挥着至关重要的作用。在检测方面，除了上述提到的转录，当前研究还从声学特征中寻找有用线索。Hou 等人 [26] 首次将情感声学特征引入检测模型，利用了广泛用于原始语音情感识别的预定义特征集。Shang 等人 [58] 设计了一种具备声学感知能力的

语音编码器，并引入梅尔频率倒谱系数特征。Qi 等人 [51] 使用预训练的 VGGish 提取音频特征进行分类。

跨模态关联。模态之间的不匹配，例如视频-文本和视频-音频不匹配，通常是为了误导目的而重新利用视频所致，这会导致跨模态关联发生重大变化。Liu 等人 [34] 分别应用跨模态 Transformer，学习视频与语音、视频与文本以及语音与文本之间的一致关系。随后汇总成对一致性分数，得出最终判断。McCrae 等人 [38]采用一种集成方法，基于字幕的文本分析、自动音频转录、视频语义分析、目标检测、命名实体一致性和人脸验证来识别不匹配。Wang 等人 [73]提出了两种基于对比学习和掩码语言建模的联合表示学习方法，以识别语义不一致。在 [13], 中，利用模态之间的主题分布差异来增强检测模型的鲁棒性。

语义特征日益丰富，推动虚假信息视频检测从挖掘单模态模式发展到建模多模态交互。然而，尽管取得了进展，现有研究仍存在一些局限。例如，大多数研究从识别虚假信息模式的角度提取特征，这可能导致忽视事实信息，也无法利用需要借助外部知识进行推理和判断的线索。尽管 [53] 率先尝试通过可靠的辟谣视频纠正先前检测出的问题新闻视频中的错误预测，以纳入相关事实信息，但多模态验证仍未得到充分探索。此外，与文本-图像建模 [52, 67]不同，模态间一致性和蕴含等逻辑约束尚未得到充分利用。这些局限表明，仍需进一步改进，以提升虚假信息视频检测方法的准确率和鲁棒性。

3.3 意图层面

虚假信息视频往往是出于特定意图而制作和传播的，例如谋取经济和政治利益 [63] 或自我表达 [20]。与信号层面和语义层面的线索相比，意图层面的线索通常更能抵御那些精心制作、旨在逃避检测的虚假信息视频，因为其深层的根本意图及其对社交媒体用户行为的影响，不太可能被大规模改变或操纵。研究人员从虚假信息制作和传播者的动机出发，利用了一些有效特征来检测虚假信息视频。

社会情境指的是信息在具有社交媒体特征的平台上传播时，用户的社交参与和用户画像。如第 2.3 节所述，独特的社会情境可能反映错误信息创作者的传播意图，因此可提供有用的特征 [51]。现有研究主要利用用户评论和用户参与度统计数据。评论通常通过提取人工设计特征 [49] ，或借助深度模型生成通用表示向量 [13, 48, 51]来加以利用。一些研究会进一步挖掘评论。例如，Serrano 等人 [57] 学习评论中的阴谋论特征，而 Choi 和 Ko [14]则关注领域知识。用户参与度统计数据，例如点赞数、评论数和浏览量，通常会直接与其他特征拼接，然后输入分类器 [26]。一些研究也会将统计数值用作重要性权重，以帮助生成嵌入。Choi 和 Ko [14] 通过根据每条评论的点赞数对其嵌入进行加权求和，生成视频评论嵌入。发布者画像为帖子级检测提供有关来源可信度的辅助信息。Papadopoulou 等人 [49] 和 Li 等人 [33] 利用了一系列特征，例如发布者频道的浏览量、已发布视频的数量以及粉丝关注比。Qi 等人 [51] 还指出，地理位置信息、账号是否经过认证等用户画像特征有助于检测，并在其模型中利用了发布者文本描述。

请注意，社交情境并非捕捉意图层面线索的唯一选择。在基于文本以及文本与图像的研究中，研究人员已开始初步尝试直接分析内容本身来挖掘意图 [16, 21]。

3.4 线索整合

在前几小节中，我们从信号、语义和意图层面，对现有虚假信息视频检测研究中使用的各种特征进行了分类。在实践中，现有方法通常会结合来自不同模态的多种特征来做出判断。本节将线索整合方法分为两类（即并行整合和顺序整合），并介绍当前的研究进展。

3.4.1 并行整合。在并行整合中，来自不同模态的所有线索都会参与最终决策过程，尽管它们的作用可能并不相同。整合过程既可以在特征层进行（即将来自不同模态的特征融合后再输入分类器），也可以在决策层进行（即由不同分支独立生成预测，再通过投票等策略合并以得出最终决策）。在现有研究中，特征融合比决策整合更常见。下面将讨论现有研究中使用的三类特征融合技术。

基于拼接：现有大多数多模态错误信息检测研究会将每种模态嵌入为一个表示向量，然后将其拼接为多模态表示。生成的表示可直接用于分类任务，也可以输入深度网络（例如卷积神经网络）中，以进行更深层次的融合和分类 [33]。线性组合是结合不同模态特征向量的另一种简单而有效的方法 [13]。结合不同模态特征的融合过程可以在视频级或帧/片段层进行 [38]。

基于注意力：注意力机制是一种更有效的方法，可用于利用不同模态的嵌入，因为它通过关注特定部分来联合利用多模态特征，并支持对序列数据进行动态融合。Shang 等人 [58] 使用协同注意力模块，同时学习视频帧与口语词各个配对之间的成对关系，以融合视觉和语音信息。Wang 等人 [73] 通过使用一种变体的掩码

表1：虚假信息视频检测方法总结。E/G：编辑/生成痕迹。T：文本。V：视觉。A：声学。CMC：跨模态关联。UE：用户参与度。UP：用户画像。C：基于拼接。ATT：基于注意力。MT：基于多任务。PL：基于流水线。由于篇幅限制，未纳入伪造视频检测方法。

方法		数据集			线索提取						线索整合		
		来源	发布数量?	E/G	T	V	A	CMC	UE	UP	C	ATT	MT
Liu 等人 [34]	2023	Twitter	10,000	N		•	•		•			•		
Qi 等人 [51, 53]	2023	抖音、快手	5,538	Y		•	•	•	•	•	•	•		
Ganti [22]	2022	未指定	-	N	•	•								•
McCrae 等人 [38]	2022	Facebook、YouTube	4,651	N		•	•		•	•		•		
Wang 等人 [74]	2022	Twitter	943,667	N		•	•	•	•			•	•	
Wang 等人 [73]	2022	Twitter	160,000	N		•	•		•			•	•	
Li 等人 [33]	2022	哔哩哔哩	700	N		•				•	•	•		
Choi 和 Ko [14]	2022	YouTube	2,912	N		•	•			•		•		
Choi 和 Ko [13]	2021	YouTube	4,622	N		•	•		•	•		•	•	
Shang 等人 [58]	2021	YouTube	891	N		•	•	•					•	
Jagtap 等 [30]	2021	YouTube	2125	Y		•								
Serrano 等人 [57]	2020	YouTube	180	N		•				•		•		
Hou 等人 [26]	2019	YouTube	250	N		•		•		•		•		
Palod 等人 [48]	2019	YouTube	546	Y		•				•		•		
Papadopoulou 等人 [49]	2017	YouTube、Twitter、Facebook	5,006	Y		•				•	•	•		

语言建模来对视频和文本的联合分布进行建模。训练 Transformer，以根据文本上下文和视频预测每个文本词元。Qi 等人 [51] 和 Wang 等人 [74] 利用跨模态 Transformer 对不同模态之间的相互作用进行建模。

基于多任务：另一种采用的融合架构基于多任务学习。在这种架构下，辅助网络用于更好地学习单独或多模态的表示、空间或参数，并提升分类性能 [2]。例如，Choi 和 Ko [13] 采用主题对抗分类，引导模型学习主题无关特征，以实现良好的泛化。Wang 等人 [73] 采用对比学习，构建视频和文本的联合表示空间。

在实践中，没有哪一种方法表现出压倒性的优势。基于拼接的融合保留了所有可用信息，但会平等对待所有特征，而不明确考虑它们的相对重要性。基于注意力的融合使模型能够关注信息量更丰富的部分，但需要设计更复杂的机制并消耗更多计算资源。基于多任务的融合增强了模型的泛化能力，但需要更多标注数据和计算资源。融合方法的选择应取决于任务的具体要求、可用资源，以及简洁性、计算需求和性能提升之间的权衡。

3.4.2 顺序整合。在顺序整合中，来自不同模态的线索以逐步的方式进行组合，每种模态都逐步为最终决策提供依据。这种整合方式可以提高整体效率和有效性，尤其是在某些模态提供冗余或重叠信息时。我们以 Ganti [22] 提出的双管齐下方法为例，说明如何利用顺序整合检测虚假信息视频。该方法通过查找与给定可疑视频对应的原始视频，并测量两者之间的相似度来进行判断。首先，该方法使用反向图像搜索检索原始视频，并计算两段视频帧之间的相似度，以判断给定视频是否为换脸深度伪造视频。如果是，则将该视频判定为虚假信息视频；否则，该方法会分析视频字幕的语义相似度，以检测两段视频在含义和意图上的变化。语义相似度较高，则判定为真实视频。为避免情感倾向变化对语义相似度的影响，对于语义相似度较低的视频，该方法会比较其视频字幕的情感倾向。如果情感倾向相似，该方法便确认语义差异并非受情感倾向影响，并将给定视频判定为虚假信息视频。

表1总结了虚假信息视频检测方法，并列出了每种方法所使用的数据来源、线索和线索整合技术。大多数方法会利用多种线索，并使用拼接和注意力机制进行线索整合。纳入更多模态并更好地建模模态间关系，有望提升虚假信息视频检测效果。

4 资源

资源往往是开展这项研究的限制因素。这里介绍相关数据集和工具，并重点说明其特征和应用场景。

4.1 数据集

由于视频爬取（大多基于精心挑选的关键词）和人工标注难度较大，许多现有数据集规模较小且针对特定主题。我们在附录B中汇总了这些数据集的信息，包括数据来源、主题

领域、规模、可访问性，以及在这些数据集上最先进（SOTA）方法（及其性能）。遗憾的是，大多数数据集并未发布。下面详细介绍四个规模较大且公开可用的数据集：

• **FVC**[1]. 初始的 FVC 包含来自各种事件类别（例如政治和体育）的视频，其中包括 200 个虚假视频和 180 个真实视频。研究人员以最初收集的视频为种子，在三个平台（YouTube、Facebook 和 Twitter）上进行搜索，将 FVC 扩展为一个多语言数据集，其中包含 3,957 个虚假视频和 2,458 个真实视频，并附有文本新闻内容和用户评论。

• **YouTubeAudit**[2]. 该数据集包含 2,943 个于 2020 年发布在 YouTube 上的视频，涵盖五个热门的错误信息主题 [29]。每个样本均被标注为传播、辟谣或对错误信息持中立态度。该数据集还提供了社交情境信息，例如元数据（如视频 URL、标题和时长）、用户参与度统计数据以及用户画像（如性别和年龄）。

• **FakeSV**[3]。该数据集包含从短视频平台抖音和快手抓取的5,538条中文短视频（1,827条虚假视频、1,827条真实视频和1,884条辟谣视频），涵盖2019年至2022年的738个新闻事件。该数据集还提供社交情境信息，包括用户反馈和发布者画像。

• **COVID-VTS**[4]。该数据集包含1万组与COVID相关的视频-文本对，专门用于研究模态间一致性。其中一半样本未经修改，其余样本则通过部分修改或替换原始样本中的事实信息生成。

4.2 工具

针对特定用途的工具对于核验可疑视频很有价值，因为它们能提供重要的辅助信息，而这些信息很难通过数据驱动模型学习得到。我们列举了三种适用于不同应用场景、具有代表性的公开工具：

• **DeepFake Detector**：一项人工智能服务，由WeVerifyProject[5]开发，用于判断给定视频是否包含深度伪造篡改人脸。该服务以可疑图像或视频的URL为输入，并返回深度伪造概率分数。它有助于检测第3.1.2节讨论的生成痕迹。

• **反向图像搜索**：以图像作为查询条件的网络搜索服务。提交提取出的关键帧后，我们可以检查之前是否有其他地方出现过相似的视频。这有助于通过比较检测篡改 [22]并识别原始来源，正如第3.1节和第3.2节中所讨论的那样。许多通用搜索引擎（如 Google 和百度）都提供此类服务。专用工具包括 TinEye、ImageRaider 和 DupliChecker[6]。

• **视频验证插件**：由 InVID 欧洲项目[7]提供的一款谷歌浏览器扩展程序，用于核查视频。它提供了一套工具，可获取 YouTube 或 Facebook 上下文信息、提取关键帧以进行反向搜索、显示元数据，并进行取证分析，这对于在信号层面和语义层面进行检测很有帮助（第3.1节和第3.2节）。

[1]https://mklab.iti.gr/results/fake-video-corpus/ [2]https://social-comp.github.io/YouTubeAudit-data/ [3]https://github.com/ICTMCG/FakeSV [4]https://github.com/FuxiaoLiu/Twitter-Video-dataset [5]https://weverify.eu/tools/deepfake-detector/ [6]https://tineye.com/, https://infringement.report/api/raider-reverse-image-search/, & https://www.duplichecker.com/ [7]https://www.invid-project.eu/tools-and-services/invid-verification-plugin/

表 2：虚假信息视频检测与相关领域的比较。

错误信息检测与欺骗检测	欺骗性技术可用于传达捏造信息，但非故意制造的错误信息可能并不一定对应欺骗行为。
虚假信息检测与有害内容检测	有害内容更有可能造成对特定人群造成心理伤害或图-ures, 而错误信息旨在误-导读者。
错误信息虚假信息检测与标题党检测	标题党可被视为一种表现形式虚假信息材料。

5 相关领域

本节讨论三个与虚假信息视频检测相关的领域，并阐明虚假信息视频检测与这些领域之间的关系（表 2）。

欺骗检测：其旨在识别欺骗行为是否存在，这对个人与公共安全至关重要。早期研究中，语言线索和非语言线索在欺骗检测中都发挥着重要作用 [7]。语言线索主要指陈述的语言特征，而非语言线索则包括神经学、视觉和声音指标。

有害内容检测：有害内容通常表现为人肉搜索、身份攻击、身份冒用、侮辱、性侵犯和暴力威胁 [10]。检测基于视频的有害内容通常依赖于捕捉多种模态中的指示性特征，例如音频转录文本的语言特征、视频情感以及被标记的有害物体 [6]。

标题党检测：标题党是网络媒体中常用来描述引人注目、吊人胃口的标题（缩略图）的术语 [63]。标题党视频检测旨在判断视频是否如实呈现了其所指事件。基于内容的方法着重分析最初呈现的信息（例如标题和视频缩略图）与整个视频所表达内容之间的语义差距；另一些方法则利用创作者档案和受众反馈 [71]。

6 未解决的问题与未来方向

尽管现有研究在检测虚假信息视频方面已取得显著进展，但仍有许多问题阻碍其应用于现实世界系统。本文将介绍两个未解决的问题和两个未来方向，并提出具体任务，以推动实用检测系统的发展。

6.1 未解决的问题

6.1.1 可迁移性。 可迁移性反映了检测系统应对数据分布偏移的能力，而这种偏移在现实应用中普遍存在且不可避免。尽管这是一个热门研究主题，但这一问题在虚假信息视频检测领域仍未得到充分探索，这也是检测方法难以付诸实践的关键障碍。

为此，我们从不同方面提出三个与可迁移性相关的子问题：

1）多平台检测。平台之间的内容和用户群体差异塑造了不同的社会情境，并提供了额外线索；研究发现，这有助于检测 [39, 51]。鉴于视频在不同平台间跨平台分享（重新上传）十分常见，这对该领域尤为重要。然而，如何应对多平台分发差异的原则仍不明确。

2）多领域检测。不同新闻领域中的错误信息文本在词语使用和传播模式上各不相同，从而导致数据偏移 [44, 60, 89]。因此，研究并缓解视频错误信息的领域差异是一个很有前景的方向。

3）时间泛化。在线视频数据随时间发生分布偏移不可避免。过去数据上有效的特征在当前环境中可能表现不佳 [27, 43, 88]。如何找到稳定特征并快速适应当前的视频数据，仍有待进一步探索。

6.1.2 可解释性。现有方法大多专注于提高准确率，却忽视了提供解释的重要性。如果缺少符合人类预期的解释，用户就很难了解检测系统的优缺点，也无法判断何时应当信任该系统。这个问题应从两个方面着手解决：

1）区分细粒度错误信息类型。除了二元分类外，模型还应进一步预测检测到的错误信息样本的具体类型（例如视频滥用）。这需要建立新的分类体系，并对数据集进行细粒度标注。

2）多模态线索归因。检测模型应将其输出归因于提取出的多模态线索。由于虚假信息视频具有复杂特征和细粒度类型，一方面，单个虚假信息视频可能只包含少数线索；另一方面，真实视频也可能包含前述某些特征，例如编辑痕迹和宣传意图。无论视频是真是假，都应厘清线索整合过程，通过线索归因来解释最终输出。

6.2 未来方向

6.2.1 线索整合与推理。视频帖子涉及多种模态，因此其检测模型需要比基于文本和图像的检测模型具备更强的线索整合与推理能力。在大多数情况下，对错误信息的最终判断既不依赖单一模态，也不依赖所有模态，找出有效的组合并非易事。例如，对于一段先前录制的事故视频，若将其重新用作新事故的现场视频并配上背景音乐，那么判断的关键在于视频与文本之间的不匹配，而非视频与音频之间的不匹配。

然而，该领域的线索整合通常是通过直接对齐并融合从不同模态获得的所有表示向量来实现的，这使模型难以学习模态间的推理。我们认为，让模型能够在模态间进行推理，对于更好地整合线索和实现更灵活的检测至关重要。可能的方向包括：

1）跨模态关系建模。对于视觉问答等需要推理能力的任务 [23], ，可以构建图来引导模态之间的交互。

2）问题分解。将检测转化为若干子问题的组合后，可以利用思维链/思维树 [78, 83] 4 [46] 提示大语言模型（例如 GPT-）进行推理。结合在线百科全书、事实核查报告和相关新闻等外部知识来源，这种方法适用于视频声明核验（与文本和图像格式中的做法类似 [3, 61, 62, 70]）。

6.2.2 推荐与检测协同。基于推荐的视频分发与错误信息视频检测之间的协同，对实际系统至关重要；这类系统的最终目标是在持续向用户推荐其感兴趣的视频的同时，避免向用户传播错误信息。为此，检测系统应包含不同的模型和策略，以利用来自推荐系统的丰富侧信息，并提高推荐的可信度。下面给出三种具体的协同场景：

1）用户兴趣感知检测。视频的观看历史不仅反映了用户的兴趣，也体现了他们容易受特定主题（如选举）影响的程度。因此，我们可以优先处理这些推荐视频，并结合主题信息检测错误信息（文本假新闻中也有类似情况，见 [75]）。

2）用户反馈感知检测。用户向平台提供的群体反馈，可能是可疑视频的重要指标。近期的一个例子是，将用户对错误信息的举报用作文本假新闻检测的弱监督信号 [76]。利用更多来自推荐系统的用户反馈，例如用户因事实性问题而表达的不喜欢，将是一个很有前景的方向。

3）可信度感知推荐。在推荐系统中考虑信息可信度，有助于减少用户接触虚假信息视频的机会，并使推荐更具问责性。一种可行的方案是将虚假信息视频检测纳入辅助任务，或使用训练良好的检测器作为评判器来提供反馈。

7 结论

我们综述了现有的虚假信息视频检测文献，并全面回顾了先进的检测方案，包括信号、语义和意图层面的线索，以及线索整合技术。我们还总结了公开可用的数据集和工具，并讨论了相关领域。此外，我们指出了研究和现实应用中的关键未解决问题与未来方向。我们还开源了一个 GitHub 仓库，并将持续更新，纳入该领域的最新进展。希望本综述能为进一步开展防范虚假信息视频的研究提供启发。

致谢

本研究部分得到国家重点研发计划（2022YFC3302102）、国家自然科学基金（62203425）、中国科学院项目（E141020）、中国博士后科学基金（2022TQ0344）以及中国博士后科学基金会办公室国际博士后交流计划（YJ20220198）的资助。

参考文献

[1] Javad Abbasi Aghamaleki 和 Alireza Behrad。2017年。利用量化效应的时空域分析检测 MPEG 视频中的恶意帧间视频篡改。《多媒体工具与应用》76（2017年），20691–20717。[2] Sara Abdali。2022。多模态错误信息检测：方法、挑战与机遇。arXiv:2203.13883（2022）。[3] Sahar Abdelnabi、Rakibul Hasan 和 Mario Fritz。2022。通过在线资源对脱离语境的图像进行开放域、基于内容的多模态事实核查。载于《电气电子工程师学会/计算机视觉基金会计算机视觉与模式识别会议论文集》。14940–14949。[4] Ronak Agrawal 和 Dilip Kumar Sharma。2021。基于视频的假新闻检测技术综述。载于《2021年第八届可持续全球发展计算国际会议》。663–669。[5] Omar Ismael Al-Sanjary、Ahmed Abdullah Ahmed、Adam Amril Bin Jaharadak、Musab AM Ali 和 Hewa Majeed Zangana。2018。使用光流方法检测克隆对象的移动。载于《2018年电气电子工程师学会计算机应用与工业电子学研讨会》。388–394。[6] Firoj Alam、Stefano Cresci、Tanmoy Chakraborty、Fabrizio Silvestri、Dimiter Dimitrov、Giovanni Da San Martino、Shaden Shaar、Hamed Firooz 和 Preslav Nakov。2022。多模态虚假信息检测综述。载于《第29届国际计算语言学会议论文集》。6625–6643。[7] Haya Alaskar、Zohra Sbaï、Wasiq Khan、Abir Hussain 和 Arwa Alrawais。2022。欺骗检测的智能技术：综述与批判性研究。《软计算》（2022），1–20。[8] Shivangi Aneja、Cise Midoglu、Duc-Tien Dang-Nguyen、Michael Alexander Riegler、Paal Halvorsen、Matthias Nießner、Balu Adsumilli 和 Chris Bregler。2021。MMSys' 21廉价伪造检测大挑战。arXiv:2107.05297（2021）。[9] Neema Antony 和 Binet Rose Devassy。2018。使用蛮力匹配实现图像/视频复制移动伪造检测。载于《2018年第二届电子与信息学趋势国际会议》。1085–1090。[10] Michele Banko、Brendon MacKeen 和 Laurie Ray。2020。有害内容的统一分类体系。载于《第四届在线滥用与危害研讨会论文集》。125–137。[11] Juan Cao、Peng Qi、Qiang Sheng、Tianyun Yang、Junbo Guo 和 Jintao Li。2020。探索视觉内容在假新闻检测中的作用。《社交媒体中的虚假宣传、错误信息与假新闻》（2020），141–161。[12] Caroline Chan、Shiry Ginosar、Tinghui Zhou 和 Alexei A. Efros。2019。Everybody Dance Now。载于《电气电子工程学会/计算机视觉基金会国际计算机视觉会议论文集》。5933–5942。[13] Hyewon Choi 和 Youngjoong Ko。2021。使用主题建模和对抗神经网络进行假新闻视频检测。载于《第30届ACM信息与知识管理国际会议论文集》。2950–2954。[14] Hyewon Choi 和 Youngjoong Ko。2022。在 YouTube 上利用领域知识和多模态数据融合进行有效的假新闻视频检测。《Pattern Recognition Letters》154（2022），44–52。[15] Christos Christodoulou、Nikos Salamanos、Pantelitsa Leonidou、Michail Papadakis 和 Michael Sirivianos。2023。利用 Transformer 模型进行转录文本上下文分析，以识别 YouTube 上的错误信息。arXiv:2307.12155（2023）。[16] Jeff Da、Maxwell Forbes、Rowan Zellers、Anthony Zheng、Jena D. Hwang、Antoine Bosselut 和 Yejin Choi。2021。编辑媒体理解框架：推理视觉错误信息的意图与影响。载于《计算语言学协会第59届年会暨第11届自然语言处理国际联合会议论文集》（第一卷：长篇论文）。2026–2039。[17] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019。BERT：用于语言理解的深度双向Transformer预训练。载于《2019年计算语言学协会北美分会会议：人类语言技术论文集》，第1卷（长篇与短篇论文）。4171–4186。[18] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。2021。一张图像胜过16×16个词：用于大规模图像识别的Transformer。载于《国际学习表征会议》。[19] Sondos M Fadl、Qi Han 和 Qiong Li。2018。监控视频认证：基于残差帧检测帧重复。《法医学杂志》63，4（2018），1099–1109。[20] Ryoya Furukawa、Daiki Ito、Yuta Takata、Hiroshi Kumagai、Masaki Kamizono、Yoshiaki Shiraishi 和 Masakatu Morii。2022。通过社交网站上的偏置用户画像进行假新闻检测。载于《IEEE/WIC/ACM网络智能与智能代理技术国际会议》。136–145。[21] Saadia Gabriel、Skyler Hallinan、Maarten Sap、Pemi Nguyen、Franziska Roesner、Eunsol Choi 和 Yejin Choi。2022。错误信息反应框架：推理读者对新闻标题的反应。载于《第60届年会论文集 计算语言学协会（第*1*卷：长篇论文）. 3108 – 3127.[22] Dhanvi Ganti。2022年。利用反向图像搜索、语义分析和元数据情感比较检测视频中错误信息的一种新方法。https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4128499。[23] Difei Gao、Ke Li、Ruiping Wang、Shiguang Shan 和 Xilin Chen。2020年。用于视觉与场景文本联合推理的多模态图神经网络。载于 IEEE/CVF计算机视觉与模式识别会议论文集。12746–12756。[24] Ankur Gupta、Neeraj Kumar、Purnendu Prabhat、Rajesh Gupta、Sudeep Tanwar、Gulshan Sharma、Pitshou N Bokoro 和 Ravi Sharma。2022年。打击假新闻：利益相关者干预与潜在解决方案。IEEE Access 10 (2022)，78268–78289。[25] Alexandros Haliassos、Konstantinos Vougioukas、Stavros Petridis 和 Maja Pan- tic。2021年。嘴唇不会说谎：一种可泛化且稳健的人脸伪造检测方法。载于 IEEE/CVF计算机视觉与模式识别会议论文集。5039–5049。[26] Rui Hou、Verónica Pérez-Rosas、Stacy Loeb 和 Rada Mihalcea。2019年。迈向在线医疗视频中错误信息的自动检测。载于2019年国际多模态交互会议。235–243。[27] Beizhe Hu、Qiang Sheng、Juan Cao、Yongchun Zhu、Danding Wang、Zhengjia Wang 和 Zhiwei Jin。2023年。立足过去，演进未来：预测假新闻检测的时间趋势。载于计算语言学协会第61届年会论文集（第5卷：产业专题）。116–125。[28] Juan Hu、Xin Liao、Wei Wang 和 Zheng Qin。2021年。利用帧时序双流卷积网络检测社交网络中的压缩深度伪造视频。IEEE视频技术电路与系统汇刊 32, 3 (2021)，1089–1102。[29] Eslam Hussein、Prerna Juneja 和 Tanushree Mitra。2020年。衡量视频搜索平台中的错误信息：一项针对 YouTube 的审计研究。ACM人机交互论文集 4, CSCW1 (2020)，1–27。[30] Raj Jagtap、Abhinav Kumar、Rahul Goel、Shakshi Sharma、Rajesh Sharma 和 Clint P George。2021年。利用视频字幕检测 YouTube 上的错误信息。arXiv:2107.00941 (2021)。[31] Pamela Johnston、Eyad Elyan 和 Chrisina Jayne。2020年。利用从真实内容中学习到的特征定位视频篡改。神经计算与应用 32 (2020)，12243–12257。[32] Hasam Khalid、Minha Kim、Shahroz Tariq 和 Simon S Woo。2021年。利用单模态和多模态检测器评估音视频多模态深度伪造数据集。载于第1届合成多媒体研讨会论文集——视听深度伪造的生成与检测。7–15。[33] Xiaojun Li、Xvhao Xiao、Jia Li、Changhua Hu、Junping Yao 和 Shaochen Li。2022年。基于CNN的误导性视频检测模型。Scientific Reports 12, 1 (2022)，1–9。[34] Fuxiao Liu、Yaser Yacoob 和 Abhinav Shrivastava。2023年。COVID-VTS：短视频平台上的事实提取与核查。载于第17届计算语言学协会欧洲分会会议论文集。178–188。[35] Zhiqiang Lv、Shanshan Zhang、Kai Tang 和 Pengfei Hu。2022年。基于无监督预训练模型的伪造音频检测。载于2022年IEEE声学、语音与信号处理国际会议。9231–9235。[36] Falko Matern、Christian Riess 和 Marc Stamminger。2019年。利用视觉伪影揭露深度伪造和人脸操纵。载于2019年IEEE冬季计算机视觉应用研讨会。83–92。[37] Katerina Eva Matsa。2022年。越来越多的美国人在 TikTok 上获取新闻，这一趋势与其他社交媒体网站相反。https://www.pewresearch.org/fact-tank/2022/10/21/more-americans-are-getting-news-on-tiktok-bucking-the-trend-on-other-social-media-sites/。[38] Scott McCrae、Kehan Wang 和 Avideh Zakhor。2022年。社交媒体新闻帖子中的多模态语义不一致检测。载于多媒体建模：MMM 2022。331–343。[39] Nicholas Micallef、Marcelo Sandoval-Castañeda、Adi Cohen、Mustaque Ahamad、Srijan Kumar 和 Nasir Memon。2022年。跨平台多模态错误信息：文本帖子与视频的分类体系、特征和检测。载于AAAI国际网络与社交媒体会议论文集，第16卷。651–662。[40] Tomas Mikolov、Kai Chen、Greg Corrado 和 Jeffrey Dean。2013年。高效估计向量空间中的词表示。arXiv:1301.3781 (2013)。[41] Yisroel Mirsky 和 Wenke Lee。2021年。深度伪造的制作与检测：综述。ACM计算调查 54, 1 (2021)，1–41。[42] Trisha Mittal、Uttaran Bhattacharya、Rohan Chandra、Aniket Bera 和 Dinesh Manocha。2020年。情感不会说谎：一种利用情感线索的视听深度伪造检测方法。载于第28届ACM多媒体国际会议论文集。2823–2832。[43] Yida Mu、Kalina Bontcheva 和 Nikolaos Aletras。2023年。时间至关重要：使用时间顺序划分重新思考谣言检测基准的评估。载于计算语言学协会论文集：EACL 2023。736–743。

[44] Qiong Nan、Juan Cao、Yongchun Zhu、Yanyan Wang 和 Jintao Li。2021年。MD-FEND：多领域假新闻检测。载于第30届ACM信息与知识管理国际会议论文集。3343–3347。[45] Arjun Narayan。2020年。我们在新冠肺炎疫情期间打击错误信息所做的努力。https://newsroom.tiktok.com/en-in/our-efforts-towards-fighting-misinformation-in-times-of-coronavirus。[46] OpenAI。2023年。GPT-4技术报告。arXiv:2303.08774 (2023)。[47] Ciarán O'Connor。2021年。TikTok声音如何被用来煽动反疫苗恐惧。https://www.isdglobal.org/digital_dispatches/how-tiktok-sounds-are- used-to-fuel-anti-vaccine-fears/。[48] Priyank Palod、Ayush Patwari、Sudhanshu Bahety、Saurabh Bagchi 和 Pawan Goyal。2019年。YouTube 上的误导性元数据检测。载于信息检索进展：ECIR 2019。140–147。[49] Olga Papadopoulou、Markos Zampoglou、Symeon Papadopoulos 和 Yiannis Kompatsiaris。2017年。利用上下文线索进行网络视频验证。载于第2届多媒体取证与安全国际研讨会论文集。6–10。[50] Jeffrey Pennington、Richard Socher 和 Christopher Manning。2014年。GloVe：词表示全局向量。载于2014年自然语言处理实证方法会议论文集。1532–1543。[51] Peng Qi、Yuyan Bu、Juan Cao、Wei Ji、Ruihao Shui、Junbin Xiao、Danding Wang 和 Tat-Seng Chua。2023年。FakeSV：具有丰富社交情境的多模态基准，用于短视频平台上的假新闻检测。载于AAAI人工智能会议论文集，第37卷。14444–14452。[52] Peng Qi、Juan Cao、Xirong Li、Huan Liu、Qiang Sheng、Xiaoyue Mi、Qin He、Yongbiao Lv、Chenyang Guo 和 Yingchao Yu。2021年。通过实体增强框架融合多样化多模态线索，改进假新闻检测。载于第29届ACM多媒体国际会议论文集。1212—1220。[53] Peng Qi、Yuyang Zhao、Yufeng Shen、Wei Ji、Juan Cao 和 Tat-Seng Chua。2023年。三个臭皮匠，胜过一个诸葛亮：通过关联邻居改进假新闻视频检测。载于计算语言学协会论文集：ACL 2023。11947–11959。[54] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。2021年。从自然语言监督中学习可迁移视觉模型。载于第38届国际机器学习会议论文集，第139卷。8748–8763。[55] Paulo Max Gil Innocencio Reis、João Paulo Carvalho Lustosa da Costa、Ri- cardo Kehrle Miranda 和 Giovanni Del Galdo。2016年。基于ESPRIT-Hilbert的音频篡改检测：利用支持向量机分类器和电网频率进行取证分析。IEEE信息取证与安全汇刊 12, 4 (2016)，853–864。[56] Mubbashar Saddique、Khurshid Asghar、Usama Ijaz Bajwa、Muhammad Hussain 和 Zulfiqar Habib。2019年。利用连续帧纹理分析进行空间视频伪造检测与定位。电气与计算机工程进展 19, 3 (2019)，97–108。[57] Juan Carlos Medina Serrano、Orestis Papakyriakopoulos 和 Simon Hegelich。2020年。基于自然语言处理的特征提取，用于检测 YouTube 上的新冠肺炎错误信息视频。载于ACL 2020新冠肺炎自然语言处理研讨会第1届论文集。[58] Lanyu Shang、Ziyi Kou、Yang Zhang 和 Dong Wang。2021年。用于 TikTok 上新冠肺炎短视频的多模态错误信息检测器。载于2021年IEEE大数据国际会议。899–908。[59] Nitin Arvind Shelke 和 Singara Singh Kasana。2021年。数字视频伪造检测被动技术综述。多媒体工具与应用 80 (2021)，6247–6310。[60] Qiang Sheng、Juan Cao、H Russell Bernard、Kai Shu、Jintao Li 和 Huan Liu。2022年。中国微博多领域虚假新闻的特征及潜在用户影响。信息处理与管理 59, 4 (2022)，102959。[61] Qiang Sheng、Juan Cao、Xueyao Zhang、Rundong Li、Danding Wang 和 Yongchun Zhu。2022年。拉远观察：用于假新闻检测的新闻环境感知。载于计算语言学协会第60届年会论文集（第1卷：长篇论文）。4543–4556。[62] Qiang Sheng、Juan Cao、Xueyao Zhang、Xirong Li 和 Lei Zhong。2021 年。利用记忆增强关键句匹配对文章重新排序，以检测已核查事实的声明。载于计算语言学协会第59届年会暨第11届自然语言处理国际联合会议论文集（第1卷：长篇论文）。5468–5481。[63] Kai Shu、Amy Sliva、Suhang Wang、Jiliang Tang 和 Huan Liu。2017年。社交媒体上的假新闻检测：数据挖掘视角。ACM SIGKDD探索通讯 19, 1 (2017)，22–36。[64] Uriel Singer、Adam Polyak、Thomas Hayes、Xi Yin、Jie An、Songyang Zhang、Qiyuan Hu、Harry Yang、Oron Ashual、Oran Gafni、Devi Parikh、Sonal Gupta 和 Yaniv Taigman。2023年。Make-A-Video：无需文本视频数据的文本生成视频。载于第11届国际学习表征会议。[65] K Sitara 和 BM Mehtre。2017年。一种全面揭露帧间视频伪造的方法。载于2017年IEEE第13届信号处理及其应用国际研讨会。73–78。[66] Matthew S Smith。2022年。《TikTok 与 YouTube：哪一个适合你？》https://www.lifewire.com/tiktok-vs-youtube-6665455。[67] Mengzhu Sun、Xi Zhang、Jianqiang Ma 和 Yazheng Liu。2021年。《不一致性至关重要：一种用于多模态谣言检测的知识引导双重不一致网络》。载于《计算语言学协会论文集：EMNLP 2021》研究成果，1412–1423页。[68] S Shyam Sundar、Maria D Molina 和 Eugene Cho。2021年。《眼见为实：视频模态是否更有力地通过在线即时通讯应用传播假新闻？》《计算机媒介传播期刊》26卷6期（2021年），301–319页。[69] Nesrine Tarhouni、Salma Masmoudi、Maha Charfeddine 和 Chokri Ben Amar。2022年。一种基于视频帧和音频水印的新冠肺炎虚假视频检测器。《多媒体系统》（2022年），1–15页。[70] James Thorne、Andreas Vlachos、Christos Christodoulopoulos 和 Arpit Mittal。2018年。《FEVER：用于事实提取与验证的大规模数据集》。载于《2018年计算语言学协会北美分会人类语言技术会议论文集》第1卷（长篇论文），809–819页。[71] Deepika Varshney 和 Dinesh Kumar Vishwakarma。2021年。《利用认知证据检测 YouTube 标题党视频的统一方法》。《应用智能》51卷（2021年），4214–4235页。[72] Sukrit Venkatagiri、Joseph S Schafer 和 Stephen Prochaska。2023年。《危机和大规模聚集事件期间研究视频分享平台错误信息所面临的挑战》。arXiv:2303.14309（2023年）。[73] Kehan Wang、David Chan、Seth Z Zhao、John Canny 和 Avideh Zakhor。2022年。《社交媒体视频帖子中的错误信息检测》。arXiv:2202.07706（2022年）。[74] Kehan Wang、Seth Z Zhao、David Chan、Avideh Zakhor 和 John Canny。2022年。《社交媒体帖子中的多模态语义不匹配检测》。载于《2022年IEEE第24届多媒体信号处理国际研讨会》，1–6页。[75] Shoujin Wang、Xiaofei Xu、Xiuzhen Zhang、Yan Wang 和 Wenzhuo Song。2022年。《真实性感知与事件驱动的个性化新闻推荐，以缓解假新闻》。载于《2022年ACM万维网大会论文集》，3673–3684页。[76] Yaqing Wang、Weifeng Yang、Fenglong Ma、Jin Xu、Bin Zhong、Qiang Deng 和 Jing Gao。2020年。《通过强化学习对假新闻检测进行弱监督》。载于《AAAI人工智能会议论文集》第34卷，516–523页。[77] Claire Wardle。2017年。《假新闻：事情很复杂》。https://firstdraftnews.org/articles/fake-news-complicated/。[78] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Ed Chi、Quoc Le 和 Denny Zhou。2022年。《思维链提示促使大语言模型进行推理》。载于《神经信息处理系统进展》第35卷。[79] Chloe Wittenberg、Ben M Tappin、Adam J Berinsky 和 David G Rand。2021年。《政治视频相较于文本的（微弱）说服优势》。《美国国家科学院院刊》118卷47期（2021年），e2114388118。[80] Junichi Yamagishi。2022年。《从 ASVSpoof 中汲取的经验与尚存挑战》。载于《第一届音频多媒体深度伪造检测国际研讨会论文集》，1–2页。[81] Xin Yang、Yuezun Li 和 Siwei Lyu。2019年。《利用不一致的头部姿势揭露深度伪造》。载于《2019年IEEE声学、语音与信号处理国际会议》，8261–8265页。[82] Yuxing Yang、Junhao Zhao、Siyi Wang、Xiangyu Min、Pengchao Wang 和 Haizhou Wang。2023年。《基于对比学习的多模态短视频谣言检测系统》。arXiv:2304.08401（2023年）。[83] Shunyu Yao、Dian Yu、Jeffrey Zhao、Izhak Shafran、Thomas L Griffiths、Yuan Cao 和 Karthik Narasimhan。[n. d.]。《思维树：利用大语言模型进行审慎的问题求解》（[n. d.]）。[84] Ning Yu、Larry S Davis 和 Mario Fritz。2019年。《将虚假图像溯源至生成对抗网络：学习并分析生成对抗网络指纹》。载于《电气电子工程师学会/计算机视觉基金会国际计算机视觉会议论文集》，7556–7566页。[85] Markos Zampoglou、Foteini Markatopoulou、Gregoire Mercier、Despoina Touska、Evlampios Apostolidis、Symeon Papadopoulos、Roger Cozien、Ioannis Patras、Vasileios Mezaris 和 Ioannis Kompatsiaris。2019年。《利用多媒体取证与深度学习检测篡改视频》。载于《多媒体建模：MMM 2019》，374–386页。[86] Xinyi Zhou 和 Reza Zafarani。2020年。《假新闻综述：基础理论、检测方法与机遇》。《ACM计算调查》53卷5期（2020年），1–40页。[87] Yipin Zhou 和 Ser-Nam Lim。2021年。《音视频联合深度伪造检测》。载于《电气电子工程师学会/计算机视觉基金会国际计算机视觉会议论文集》，14800–14809页。[88] Yongchun Zhu、Qiang Sheng、Juan Cao、Shuokai Li、Danding Wang 和 Fuzhen Zhuang。2022年。《泛化至未来：缓解假新闻检测中的实体偏差》。载于《第45届ACM SIGIR信息检索研究与发展国际会议论文集》，2120–2125页。[89] Yongchun Zhu、Qiang Sheng、Juan Cao、Qiong Nan、Kai Shu、Minghui Wu、Jindong Wang 和 Fuzhen Zhuang。2022年。《记忆引导的多视图多领域假新闻检测》。《IEEE知识与数据工程汇刊》（2022年）。

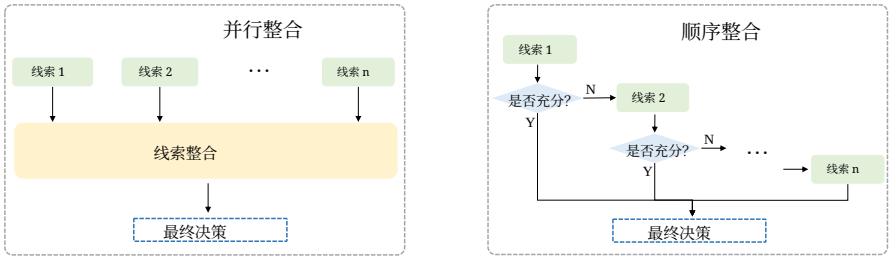
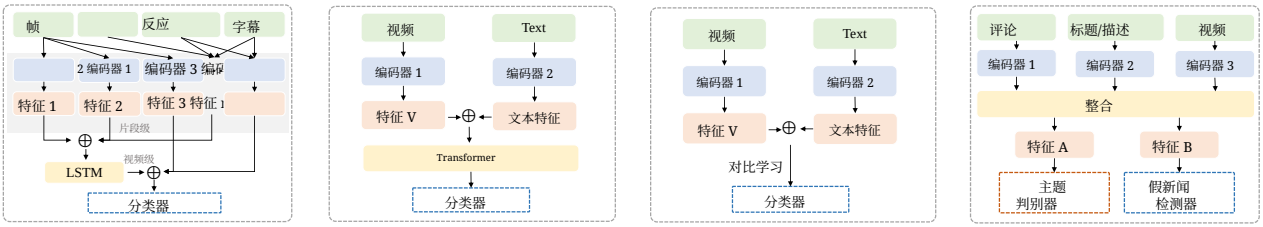


图3： 虚假信息视频检测中并行线索整合与顺序线索整合的比较 .



(a) McCrae 等人 [38] **(b) Wang 等人 [73] (掩码语言模型)** **(c) Wang 等人 [73] (对比学习)** **(d) Choi 和 Ko [13]**

图 4：具有代表性的错误信息视频检测模型示意图。

表 3：用于检测虚假信息视频的数据集概览。请注意，Christodoulou 等人 [15] 在 YouTubeAudit 的一个子集上进行了实验。SOTA：最先进水平。

数据集	来源	领域	Size	是否已发布?	最先进方法	最先进性能（%）	
						准确率	F1
FVC	YouTube、Twitter、Facebook	各种	5,006	Y	FANVM [13]	––	85.84
VAVD	YouTube	各种	546	Y	FANVM [13]	––	86.71
YouTubeAudit	YouTube	阴谋论	2,943	Y	（子集）RoBERTa [15]	94.00	94.00
FakeSV	抖音、快手	各种	5,538	Y	SVFND+需要 [53]	84.62	84.61
COVID-VTS	Twitter	新冠肺炎	10,000	Y	TwtrDetective [34]	68.10	67.90
(Hou 等人，2019年)	YouTube	前列腺癌	250	N	(Hou 等人) [26]	74.41	––
(Serreno 等，2020年)	YouTube	新冠肺炎	180	N	(Serreno 等) [57]	84.40	––
(Jagtap 等，2021年)	YouTube	各种	2,125	N	(Jagtap 等) [30]	92.00	78.00
(Shang 等，2021年)	YouTube	新冠肺炎	891	N	TikTec [58]	72.31	60.51
(Choi 和 Ko，2021年)	YouTube	各种	4,622	N	FANVM [13]	––	86.28
(Choi 和 Ko，2022年)	YouTube	各种	2,912	N	FVDM [14]	––	82.55
(Li 等人，2022年)	哔哩哔哩	健康	700	N	(Li 等人) [33]	90.00	89.00
(Wang 等人，2022a)	Twitter	各种	160,000	N	(Wang 等人) [73]	71.07	––
(Wang 等人 2022b)	Twitter	各种	943,667	N	(Wang 等人) [74]	85.43	––
(McCrae 等人 2022)	Facebook、YouTube	各种	4,651	N	(McCrae 等人) [38]	60.50	––
(Yang等人（2023年）)	抖音、快手	各种	8,213	N	（Yang 等人） [82]	87.71	87.38

A 补充图示

为便于说明，我们在图 3 中对线索整合的两种范式的过程进行了可视化和比较。由于现有研究大多采用并行整合，本文展示了四种代表性模型，帮助读者直观理解检测框架。图 4a 和图 4b 分别展示了基于拼接的融合，以及利用注意力机制对跨模态交互进行建模。图 4c 和图 4d 则分别展示了采用对比学习和主题分类等辅助目标来优化多模态特征学习。

B 数据集概述

在第 4.1 节中，我们详细介绍了四个大型公开可用数据集。不过，值得注意的是，还有一些未开放源代码的小规模和特定主题数据集，但它们仍对研究做出了重要贡献。表 3 总结了所有这些数据集，包括其数据来源、领域、规模、可访问性、最先进的方法，以及基于准确率和 F1 分数的相应性能。