数字内容生成、检测与取证技术综述

曹娟[1,2]，朱勇椿[1,2]，亓鹏[1,2]，黄子尧[1,2]，杨天韵[1,2]，王政嘉[1,2]，卜语嫣[1,2]
1. 中国科学院计算技术研究所数字内容合成与伪造检测实验室，北京 100190；
2. 中国科学院大学，北京 100049

摘要
近年来，数字生成内容技术得到了极大的发展，数字内容的检测和取证技术面临新的挑战。首先从自然语言大模型、视觉生成技术、多模态生成技术3个方面介绍数字内容生成技术，从生成文本检测、生成图片检测、生成音视频检测3个方面介绍数字内容检测技术，从利用事实信息和伪造痕迹两方面介绍数字内容取证技术；接着介绍这些技术的应用场景；最后对该研究领域的未来工作进行展望，指出几个需要重点关注的方向。

关键词
数字内容；生成技术；检测应用；取证技术

中图分类号：TP316 文献标志码：A doi: 10.11959/j.issn.2096-0271.2023066

A survey on digital content generation, detection, and forensics techniques

CAO Juan[1,2], ZHU Yongchun[1,2], QI Peng[1,2], HUANG Ziyao[1,2], YANG Tianyun[1,2], WANG Zhengjia[1,2], BU Yuyan[1,2]
1. Media Synthesis and Forensics Lab, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
2. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract
In recent years, the technology of digital content generation has been greatly developed, and the detection and forensic technology of digital content are facing new challenges. This paper firstly introduced digital content generation technology from three aspects: large natural language model, visual generation technology, and multimodal generation technology. Secondly, it introduced digital content detection technology from three aspects: generated text detection, generated image detection, and generated audio and video detection. Thirdly, it introduced digital content forensics technology from two aspects: utilizing fact ual information and forging traces. Then, this paper introduced the application scenarios of these techniques. Finally, it prospected the future work in this research field, and pointed out several directions that need to be focused on.

Key words
digital content, generation technology, detection, application, forensics technology

0 引言

近几年深度学习快速发展，在大模型方面取得显著的技术突破，例如：视觉大模型ViT（vision transformer）[1]、语言大模型BERT（bidirectional encoder representations from transformers）[2]和GPT（generative pre-trained transformer）[3]、多模态大模型CLIP（contrastive language-image pre-training）[4]等。数字内容生成技术指利用人工智能技术自动生成内容，大模型的出现为数字内容生成技术提供了强力的支撑。此外，随着数字经济与实体经济融合程度不断加深，人类对数字内容总量和丰富程度的整体需求不断提高，海量的数字内容供给需求牵引数字内容生成技术应用落地，微软、Meta、百度等多家头部企业投入数字内容生成技术的研发。最新出现的DALL-E 2、ChatGPT、GPT 4等数字内容生成技术掀起了内容创造热潮，重塑甚至颠覆了数字内容的生产方式和消费模式。生成的数字内容具有真实性、多样性、可控性的特点，有助于企业和个人提高内容生产的效率，提供更加丰富多元、动态且可交互的内容，有着广泛的应用前景，例如智能新闻写作可提升新闻资讯的时效；生成商品3D模型用于商品展示和虚拟试用；打造虚拟主播，赋能直播带货；人工智能创作电影；元宇宙数字人等。

然而，科技是发展的利器，也可能成为风险的源头。技术发展与风险挑战相伴而生，数字内容生成技术领域同样如此。虽然，数字内容生成技术的快速发展大大提高了生成文本、语音、图像、视频的逼真度和多样性，催生了AI写作、AI作图、AI语音合成等众多应用。然而，随着生成内容质量的不断提高，机器伪造能力已经超过人类的真假识别能力，生成技术的发展和应用也为虚假和伪造信息的泛滥埋下了新的隐患。数字内容生成相关技术被不法分子快速工具化、普及化甚至武器化，制作和传播虚假伪造内容，影响传播秩序和社会秩序，给网络空间健康发展带来安全威胁。这些伪造信息渗透到社会行为的方方面面，包括政治博弈、军事伪装、经济诈骗、舆论欺骗等，给政治安全、经济安全、公共安全、人身安全、军事安全等带来严重危害。美国、欧盟等已将虚假伪造内容视为重大的国家安全威胁。因此，亟须进一步加强对数字内容生成技术的监管，统筹发展与安全，推进数字内容生成技术依法合理有效利用，促进数字内容生成服务规范发展，维护网络空间良好生态。如图1所示。

数字内容伪造检测是对深度合成技术生成的数字内容进行有效监管和治理的重要一环。只有对生成的数字内容进行精确检测，才能监督并防范有意或无意发布的生成数字内容，进而通过明确标识等方式限制其影响及传播范围。然而数字内容伪造检测面对3个挑战，使得精确检测生成数字内容并非易事：①对抗性，数字内容的生成和检测技术存在类似于矛和盾的对抗关系；②泛化性，由于数字内容生成技

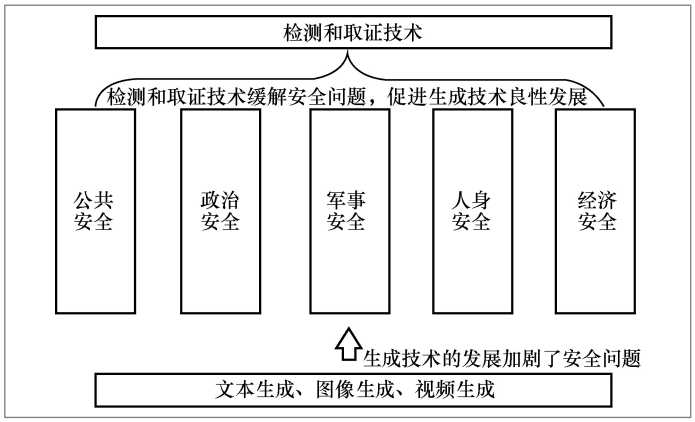


图1 数字内容生成、检测和取证技术

术的不断升级，检测技术需要在不断变化的伪造环境中持续适应和进化；③不确定性，人类无法轻易地分辨数字内容是否由机器自动生成得到，特别是当面对未知的伪造类型时。此外，多模态的伪造数字内容伪造痕迹不同，例如：机器伪造文本内容可能不连贯，伪造图片内容可能局部细节模糊。因此，需要针对不同场景的伪造内容设计特定的检测方法，干预伪造内容的传播。

然而检测模型可解释性差，现有大多数检测模型只能给出最后的检测结果，却不能对结果进行人类可理解的解释。这些伪造检测方法并不能给出鉴别的证据，并且其判别结果可能不确定，导致不能作为一个确切的结论，无法作为后续操作，比如法律程序的可信证据。数字内容取证可辅助人类专家进行深度伪造证据的采集、量化、分析、呈现，最终形成直观可信的证据，保证证伪鉴定结果的科学性与可靠性。如何在检测后对其进行取证，得到检测的依据以及证据，是伪造检测落地以及走入司法程序的关键一环。数字内容取证技术的研究目前还处在早期的阶段，面临着两个挑战。①证据难提取，针对文本内容伪造，需要在海量信息中检索事实证据，并进行比对。对于图像伪造，难以在肉眼可见的图像空间可视化伪造痕迹特征。②取证结果和人类认知难对齐，现有的伪造图像检测模型识别的伪造痕迹特征往往不能解耦并映射到五官扭曲、边缘不一致等人类能理解的概念上。因此，数字内容取证技术亟待进行深入研究。

数字内容生成技术的发展为产业带来革新的同时，引入了更多的安全隐患，威胁网络内容生态健康安全发展，而数字内容伪造检测技术和数字内容取证技术提升了技术监管和内容治理的能力，促进数字内容生成技术的良性发展。本文对数字内容生成、检测与取证技术进行了概述。现有的大多数生成、检测、取证的综述通常仅介绍单个环节的技术，并且从技术角度出发进行概述，而本文从数字内容产业的角度，介绍数字内容生产和监管的不同环节，这不仅可以使读者清楚地了解数字内容生成现状，还可以更加清晰地了解检测并防范恶意伪造的方法及应用。

1 数字内容生成

数字内容生成技术是指利用数字技术生成图像、视频或语言等数字内容的技术，其中AIGC（artificial intelligence generated content），即以人工智能技术来生成内容，在最近几年得到了惊人的发展。在视觉生成领域，继生成对抗网络（generative adversarial network，GAN）之后，Diffusion模型展示了惊人的生成质量，以DALLE2、Stable Diffusion等为代表的文字到图像生成模型具备了根据人类语义精确生成高质量图像的能力，现已在场景设计、角色设计、虚拟偶像等多个方向出现了应用。在语言生成上[5]，2022年OpenAI的ChatGPT横空出世，其凭借上下文理解、知识储备、对话理解能力震惊世人，还能高质量完成翻译、写代码、改论文、写文案等多项任务。可以预见的是，基于Diffusion和ChatGPT等技术的AIGC将会带来一波会深刻改变人类现实生活的应用热潮。然而，数字内容生成技术也需要注意安全反制问题，以防止其被滥用，或者被窃取模型、隐私数据等。虽然数字内容生成技术取得了重大突破，但依然存在一些挑战。

- 多样性：数字内容生成技术需要具备生成多样化内容的能力，以满足不同用户的需求。现有大模型在大量的数据上训

练，但在更多样化的训练集上训练并不总是比精心准备的基础模型上对下游性能更好[6]。因此更好地理解跨领域表示以及它们如何对测试时分布偏移具有弹性，有助于指导训练数据集的设计，从而平衡专业化和泛化性。

● 推理性：数字内容生成技术需要具备推理能力，即从给定的信息中推断出隐含的信息，可以帮助人们做出决策、解决问题。现有的大语言模型虽然在一些推理任务上表现出一定的能力，但有时仍可能在常识推理任务上失败[7-8]。

● 可控性：数字内容生成技术需要具备一定的可控性，以便用户能够控制生成内容的质量和风格。生成模型的可控性一直以来是研究的热点[9-12]，但在现实场景下，可控的内容生成依然不足以满足用户需求。例如：对于角色设计而言，一个角色身上的每一个装饰反映的是设计师对角色的理解和设定，而现有的图像生成模型难以达到这种细节的可控生成。

● 安全性：现有生成模型具有一定的安全问题。一方面，ChatGPT类应用服务生成文本目前存在事实性错误、政治偏见等问题。一旦被别有用心的组织用于舆论引导，大量生成的错误文本流传到互联网上，将对网络信息生态造成重大风险；另一方面，高质量的生成技术也可能被应用于网络暴力、涉黄涉暴网文创作、电信诈骗等危害公共安全的场景。例如：ChatGPT用于模拟人物对话，图像和视频生成技术用于制作逼真的人物视频，声音合成技术用于制作语音等。最后，生成模型本身可能会发生泄露用户隐私、训练数据等情况[13-14]，威胁数据安全。

从数字内容模态的角度，数字内容生成方法可以分为自然语言大模型、视觉生成技术和多模态生成技术方法，如**图2**所示。下面对这3类检测方法进行介绍。

1.1 自然语言大模型

（1）大模型架构

自然语言大模型的成功离不开两个关键要素：模型结构和预训练方法。首先是

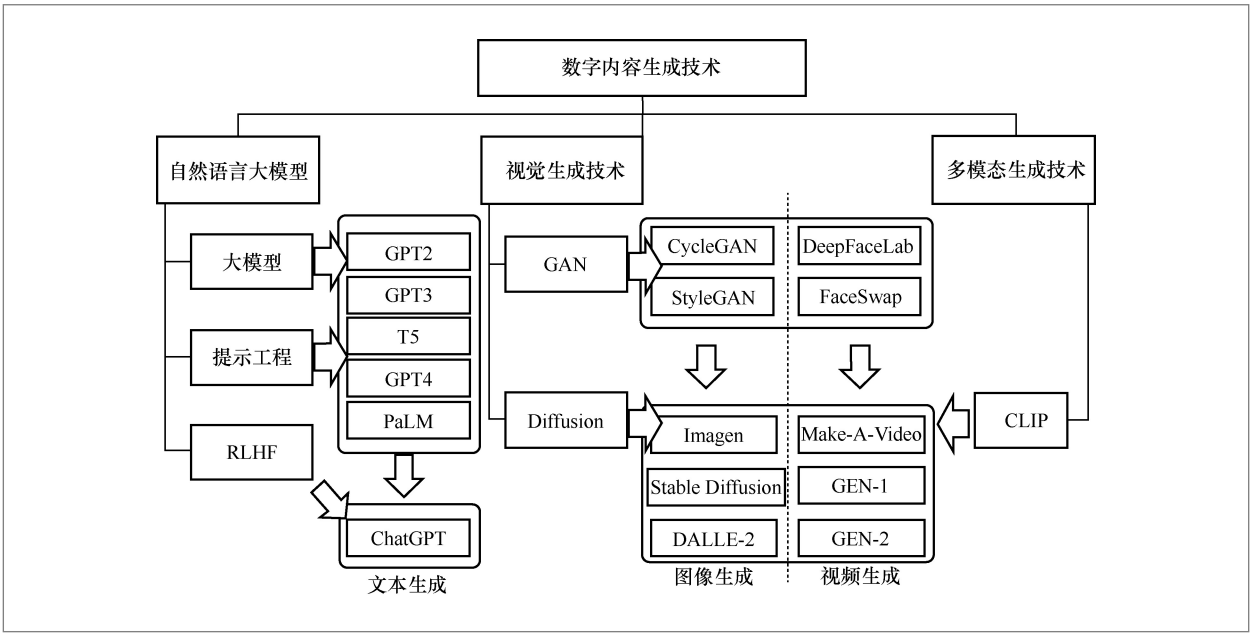


图 2 数字内容生成技术

Transformer模型结构。Transformer[15]模型是自然语言大模型采用的网络结构。Transformer由编码器和解码器组成，其中每一层由多头注意力和前馈神经网络组成。多头注意力是Transformer的核心部件，其根据各词的相关性分配不同权重，能够更好地处理长期依赖关系，具有高度可并行性，并允许数据克服归纳偏置[16]，极为适合大规模的预训练。其次是模型规模的上升。Transformer结构模型参数超过1亿，之后BERT[2]模型达到了3亿的规模。对于GPT系列，GPT-1[3]拥有上亿的参数量，GPT-2[17]达到了15亿，而GPT-3[18]更上一个数量级，参数规模首次突破千亿。大模型背后更是海量的数据集。如GPT-3的训练语料达到总计45 TB，包括CommonCrawl[19]、WebText2[17]和维基百科语料等。而GPT-4则在多个模态的海量数据上训练，取得了跨模态的理解能力。预训练方法是大模型成功的另一个关键。预训练是指在大量无标注的数据上设计训练任务，通过学习大量的数据特征和语言规律来提高模型能力。Bert使用掩码语言模型（masked language model，MLM）[20]的预训练方法，通过随机掩码的方式预测输入中被遮盖的单词，从而学习单词之间的关系和语言规律。自回归模型，例如GPT、OPT[21]等，则通过根据上一次词的输入预测下一个词的方法来进行训练，这种训练方式更适合生成任务。

（2）提示工程与上下文学习

提示工程（prompt engineering）[22]是指在使用自然语言处理大模型时，通过设计高效的输入提示来引导模型进行正确的预测，被广泛应用于提高模型的准确性和效率上。提示工程可以通过多种技术实现，包括生成式提示、填充式提示、控制式提示等。提示工程是构建高质量自然语言处理系统的重要工具之一。在提示工程中，上下文学习（in-context learning）作为一种提高语言模型性能的有效方法受到了广泛关注。它使用预训练的语言模型作为主干，并在提示中添加一些输入标签示范对和说明，进而提高性能。这种技术可以提高生成的文本的连贯性和语境适应能力，从而使生成的文本更加自然和可读。

（3）RLHF

为了使大语言模型产生符合人类意图的输出，人类反馈的强化学习（reinforcement learning from human feedback，RLHF）被应用于InstructGPT[23]、Sparrow[24]和ChatGPT等的微调之中。RLHF是基于人类反馈的强化学习方法。传统的强化学习方法是基于环境反馈的，即智能体通过与环境的交互获得奖励信号来调整其行为。但是，对于语言大模型而言，环境反馈缺乏、不准确或者代价高昂，此时可引入人类反馈来指导智能体的学习过程。RLHF的核心思想是将人类反馈作为额外的奖励信号加入强化学习框架中，同时采用一系列技术来处理反馈的不确定性和多样性。这些技术包括：反馈的采样、筛选和汇聚、反馈的表示和转化，以及反馈与环境奖励的融合等。具体在语言模型中，RLHF的整个流程包括3个步骤：预训练、奖励学习和强化学习微调。对于预训练语言模型回答的不符合人类要求的内容，可以使用训练的奖励模型来编码多样化和复杂的人类偏好，接着利用ELO[25]等算法将成对比较关系转换为逐点奖励标量，最后用强化学习对语言模型θ进行微调以最大化学习到的奖励函数。为稳定强化学习训练，常采用近似策略优化（PPO）[26]算法。

1.2 视觉生成技术

（1）GAN

GAN[27]在图像生成领域广受欢迎。

GAN由生成器和判别器两部分组成。生成器学习真实样本的分布以生成新数据，判别器则确定输入是来自真实数据空间还是来自生成器的数据空间。生成器和判别器的结构对GAN的训练稳定性和性能有很大影响。一些代表性的GAN变体包括DCGAN[28]、ProGAN[29]、BigGAN[30]和StyleGAN[9,31-32]等，除此之外GAN有多种生成器和判别器的结构、目标函数以及用于解决GAN训练稳定性和性能问题的各种技术，例如WGAN[33]和LS-GAN[34]等的目标函数可以稳定GAN的训练过程。其中StyleGAN以风格编码来解耦向量空间，同时获得高质量生成结果，是GAN的代表性工作。得益于条件GAN[35]、GAN inversion[10]等技术的发展，人们可以操纵GAN模型的隐空间，从而能够控制模型的输出。

(2) Diffusion模型

生成扩散模型(generative diffusion model)是一种基于概率的无监督式生成模型，其设计灵感来自非平衡热力学，模仿扩散过程对图像不断加噪以将其转变为近似噪声的隐编码，然后模型学习逆转加噪的过程，从图像相同尺寸的噪声中不断去噪以还原原始图像。生成扩散模型有以下3种主要构建方式。

- DDPM[36]被认为是一种基于马尔可夫链的参数化模型，通过在真实图片上逐步添加高斯噪声的扩散步骤得到噪声图像，模型学习如何反向扩散过程，以便从纯噪声中构建出所需的数据样本。基于分数的生成模型(SGM)直接处理数据对数密度(即分数函数)的梯度。
- NCSN[37]是一种基于SGM的生成模型，它通过对数据进行多尺度的强化噪声扰动，可以更加准确地估计分数函数。NCSN的训练和推理步骤完全解耦，这意味着可以分别进行训练和推理，从而提高生成样本的效率。此外，由于SGM可以直接建模数据对数密度函数的梯度，因此在处理数据时具有很好的优化效果。
- Score SDE[38]是一种将之前的两种形式推广到连续情况的生成模型。该模型中，噪声扰动和去噪过程被描述为随机微分方程的解。通过将概率流ODE应用于逆转过程的建模中，该模型证明了ODE同样可以用于生成模型的构建。

扩散生成模型的训练相对简单且稳定，比起传统的生成对抗网络更容易实现。这是因为扩散生成模型的训练过程不需要对抗式训练，减少了模型训练过程中的稳定性问题。

同时，扩散生成模型的表示能力非常强。其加噪去噪过程的设计适合完成图像到图像的转换，包括图像修复、图像超分辨率、图像风格转换等任务。此外，扩散生成模型的设计也适合完成图像编辑任务，如人脸表情编辑、风格化的头像生成等。另外，扩散生成模型也适用于生成大模型。由于GAN的训练过程中存在梯度消失和梯度爆炸等问题，GAN在生成大模型方面面临较大的困难。而扩散生成模型的训练过程中不会出现这些问题，使其在生成大模型方面表现更加出色。

尽管扩散生成模型在生成高质量图像方面取得了很好的效果，但也存在一些问题。首先，推理速度比较慢，需要较长的时间才能生成一张高质量的图像。这使其在实时性应用上受到了限制。其次，扩散生成模型的隐空间比较难以操纵，很难通过直接操作隐变量来控制图像的某些特征。另外，扩散生成模型生成的结果具有一定的随机性，这使得每次生成的结果都会略有差异，虽然能提升结果的多样性，但会影响某些应用场景(如视频合成)的效果。针对这些问题，目前有很多研究正在进行中，希望能够通过改进模型结构和算法来

解决这些问题，进一步提高扩散生成模型的实用性和性能。

1.3 多模态生成技术

CLIP（contrastive language-image pre-training）[4]是2021年由OpenAI发布的将图像和自然语言处理（NLP）领域相结合的联合训练模型，可以理解文本和图像之间的相似性。CLIP模型的主要思路是将图像和文本对输入模型中，以自监督学习的方式训练，拉近同一个物体或概念在两个编码之间的距离。这个过程可以在大量的图像和文本对数据上进行训练，这样就可以学习到一个具有广泛应用能力的模型。具体来说，CLIP 模型包括两个部分：一个图像编码器和一个文本编码器。对两个编码器获得的两个模态的向量进行对齐，这两个编码器的向量空间是相同的，因此可以通过计算它们之间的相似度来衡量图像和文本之间的相似度。CLIP已经被广泛应用于多模态任务中，特别在生成领域，CLIP应用于连接语言和图像，例如文本到图像生成[11,39]。在其他多模态场景下，CLIP也被用来进行桥接，例如文本与语音的CLIP[40]。

2 数字内容检测

数字内容伪造检测是对深度合成技术生成的数字内容进行有效监管和治理的重要一环。只有对生成的数字内容进行精确检测，才能监督并防范有意或无意发布的生成数字内容，进而通过明确标识等方式限制其影响及传播范围。然而，精确检测生成数字内容并非易事。随着深度合成技术的快速发展，生成数字内容高度逼真，人眼难以辨别其真伪。针对重大事件、重要人物恶意伪造的数字内容往往会在极短的时间内迅速发酵，借助社交媒体获得大量传播，造成严重的消极影响，因此需要对伪造数字内容进行快速响应。而随着深度合成技术的广泛应用，互联网上涌现大量的生成数字内容，使得伪造检测的计算资源需求增加，也进一步对检测算法的实时性和轻量性带来了挑战。因此，需要大力发展针对生成数字内容的快速、准确、高效的自动检测技术，用检测技术来对抗合成技术，从源头上控制生成数字内容的传播，从而更好地抵御深度合成技术带来的风险，促进深度合成服务的良性发展。然而伪造数字内容检测存在以下3个难点。

- 对抗性：数字内容的生成和检测技术存在类似于矛和盾的对抗关系。伪造者会针对检测方法中使用的伪造痕迹对自身进行迭代升级，使检测算法失效，这促使检测技术不断寻找新的伪造痕迹和漏洞，以适应不断变化的伪造算法。伪造者还可能对检测模型进行对抗攻击，通过对生成数字内容进行人眼不可见的微小修改或添加扰动，使得检测模型精度下降，甚至出现直接判错的情况[2]。为保证检测算法的准确性和鲁棒性，检测技术需要采用更加先进的对抗性学习算法，对对抗攻击进行识别和抵御。
- 泛化性：由于数字内容生成技术的不断升级，检测技术需要在不断变化的伪造环境中持续适应和进化。这要求检测技术不仅能够识别已知的、已标注的伪造内容，更需要适应新的、未见过的伪造内容。检测技术需要学习到生成数字内容的本质特征，并借助迁移学习等技术泛化至不同伪造算法、不同数据集、不同领域数据及不同事件主题等，实现在新的伪造环境下仍能保证检测的精度和效果。
- 不确定性：与目标检测等经典的分类任务不同，人类无法轻易地分辨数字内

容是否由机器自动生成得到，特别是当面对未知的伪造类型时。因此，检测技术需要对预测结果的不确定性进行显式建模，同时结合取证技术对预测结果进行解释，以加强生成数字内容检测的可靠性，避免误判、错判带来的消极影响。

从伪造数字内容模态的角度，数字内容的检测方法可以分为生成文本检测方法、生成图片检测方法和生成音视频检测方法，如**图3**所示。下面对这3类检测方法进行介绍。

2.1 生成文本检测方法

机器生成文本是由机器产生、修改或扩展的自然语言文本[41]。机器生成文本检测方法可以分为黑盒检测和白盒检测。黑盒检测是在生成模型未知的情况下，利用统计及语言模式的差异对机器生成文本以及人类书写的文本进行区分。现有研究[42-45]发现，与人类书写的文本相比，机器生成的文本更正式、容易重复和不连贯。因此，研究人员提出一系列统计特征来衡量文本的重复度[46]、连贯性[46]、可读性[47]、分布曲线[48]等，以及依存关系分析、情感分析等词法特征[42]。基于手工特征的方法简单有效、可解释性强，但对不同生成和采样算法迁移性不强[46]，缺乏全面性和灵活性。相比之下，微调后的语言模型（如RoBERTa[49]）往往能够取得更好的检测效果，并且具有更强的泛化能力[50-51]。由于缺乏对内容真实性的约束，语言模型往往会生成包含错误信息的文本[52]，因此可以借助事实核查的方法[53-55]对文本的真实性进行验证，从而辅助机器生成文本的检测。然而，随着大规模语言模型的逐步升级，机器生成文本和人类书写文本在上述层面上的差距逐渐缩小，降低了黑盒检测方法的可用性。白盒检测是指在可以完全访问语言模型的情况下控制模型的生成行为，以达到追踪的目的。具体地，编码方对生成的文本添加水印，解码方根据文本中是否隐藏水印判断该文本是否由给定模型生成。该方法可以从源头上对机器生成文本进行标记，检测准确率更高，但需要创作者进行主动配合。现有研究大多采用在文本生成后添加水印，包

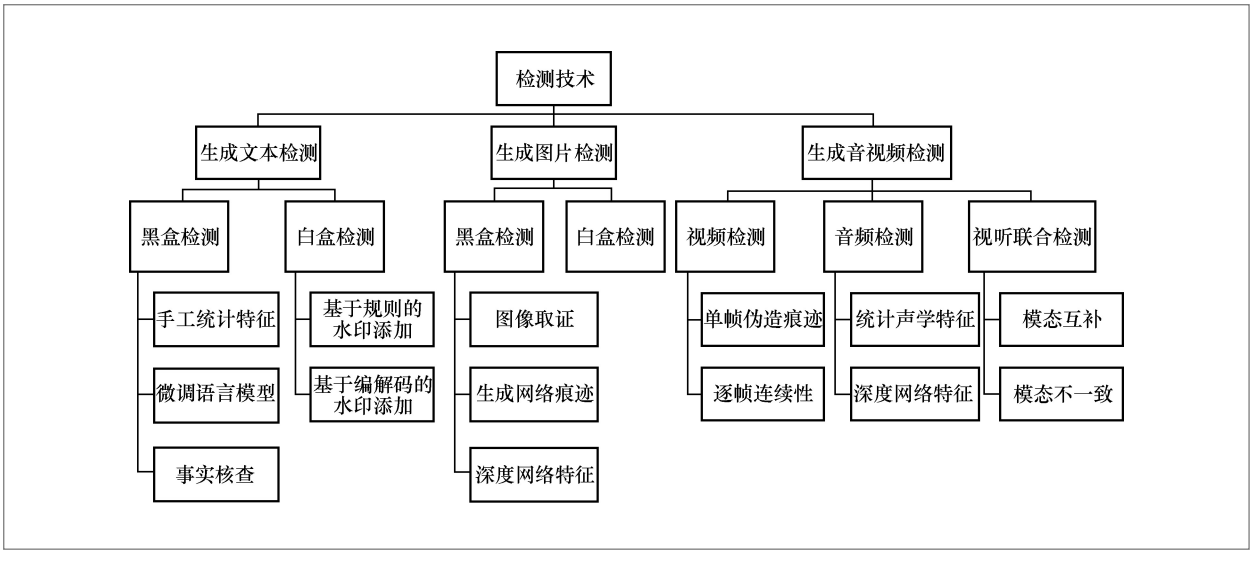


图 3 数字内容检测技术

括在句法树[56]、语法树[57]上使用基于规则的固定替换进行水印添加，以及设计相应的编码解码网络进行水印添加[58-59]。

2.2 生成图片检测方法

生成图片指由深度学习技术自动合成的图片，也被称为深度伪造图片。这类图片主要对人物面部进行篡改，包括人脸替换、人脸生成、表情伪造或面部属性操纵等[60]。与生成文本检测类似，现有的生成图片检测方法也可以分为黑盒检测和白盒检测两大类。其中，白盒检测通过分析预先嵌入的水印或数字签名实现对图片原始性的判断[61]。由于白盒检测要求对生成模型已知，在实际应用中受限，现有方法大多关注无须预先嵌入信息的黑盒检测技术。现有的黑盒检测方法主要分为基于图像取证的检测方法、基于生成网络痕迹的检测方法以及基于数据驱动的检测方法。其中，基于图像取证的检测方法利用特定篡改导致的异常痕迹作为线索，通过手工设计或神经网络拟合提取特征进行检测。例如：针对原图与篡改区域之间像素排列逻辑不一致的现象，Kirchner等人提出利用重采样特征进行检测[62-63]；针对篡改区域图层边缘和原图背景不匹配的现象，Zhou等人提出异常边缘特征进行检测[64-65]；针对自然拍摄和伪造生成区域光学噪声不同的现象，Lukáš等人提出利用光响应非均匀性特征进行检测[66-68]；针对原始图像与篡改区域JPEG压缩次数和压缩系数不同的现象，Lin等人通过分析重压缩特征[69-70]和频域信息[71]进行检测。基于生成网络痕迹的检测方法将生成对抗网络生成的伪造图片中隐藏的痕迹及纹理信息作为生成对抗网络的指纹进行辅助检测[72-73]。这类方法具有很强的模型依赖性和指向性，因此除了用于对伪造图片进行检测外，还可以对生成算法进行溯源[74]，但这类方法对于新出现的生成模型的泛化能力不强。随着伪造数据量规模的不断增加，基于数据驱动的检测方法也得到了广泛应用。这类方法将深度伪造图片检测任务抽象为一个二分类问题，利用一些经典的神经网络架构进行分类，如Xception[75]、VGG[76]、ResNet[77]、GoogLeNet[78]、ViT[1]等，它们在伪造图片检测任务上获得了不错的性能。

2.3 生成音视频检测方法

深度伪造视频通过对多帧伪造图片进行组合得到。与深度伪造图片类似，深度伪造视频主要对人脸面部区域进行篡改，部分深度伪造视频还会对人物动作进行伪造[79]。深度伪造音频包含文本生成语音[80-81]和语音转换[82-83]两种方式，生成音频可以较好地模拟目标人物的音调音色，从而与视觉内容组合成一个完整的伪造视频。现有的深度伪造视频的检测方法主要关注视觉层面的线索。常见的检测思路是将视频逐帧分解，再利用生成图片检测的相关技术分析单帧伪造痕迹。与图片相比，视频放大了伪造生成模型在细节上的瑕疵，主要体现在眨眼[84]、头部运动[85]、唇语[86]、肤色变化[87]等生理特征上。此外，视频特有的时序信息也为伪造检测提供了有效的线索。大多数伪造视频合成时容易忽视帧间的平滑，从而导致多帧伪造图片在时序上的不一致[88]。这类基于视频时序的方法在精度和泛化性上都优于基于单帧的模型[88-90]。近年来，深度伪造音频检测的研究逐渐兴起，相关竞赛ASVspoof[91]推动产出了很多具有实际应用价值的解决方案。大多数方法采用手工提取的梅尔频率倒谱系数（MFCC）、

线性频率倒谱系数（LFCC）、常数Q倒谱系数（CQCC）等声学特征，并应用高斯混合模型和轻卷积神经网络等分类器进行检测[92]。数据驱动的检测方法在深度伪造音频检测任务上同样得到了有效应用，比如Lyu等人利用预训练的XLS-R模型[93]进行语音表示提取和端到端的检测[94]。视听联合的多模态深度伪造检测作为一个新方向，近年来受到越来越多的关注。这类工作可以利用音频与视频之间的不一致进行检测[95-97]，或者利用不同模态间的互补信息对单一模态的伪造检测进行优化[98]。

3 数字内容取证

深度伪造数字内容的逼真程度日益提高，当前主流的基于深度学习的伪造鉴别模型是高度非线性的复杂模型，判别结果难以解释和可视化，导致鉴伪过程的可信性和正确性难以评估。数据内容取证可辅助人类专家进行深度伪造证据的采集、量化、分析、呈现，最终形成直观可信的证据，保证证伪鉴定结果的科学性与可靠性。

数据内容取证技术的难点在于：①证据难提取。对于机器生成假新闻检测领域，大部分工作基于伪造内容的表层和表现特征，缺少对其内容本身的核实和取证。对于伪造图像检测领域，通用视觉任务上的解释方法往往不适用于伪造检测任务，难以在肉眼可见的图像空间可视化伪造痕迹特征。②取证结果和人类认知难以对齐。现有的伪造图像检测模型识别的伪造痕迹特征往往不能解耦并映射到五官扭曲、边缘不一致等人类能理解的概念上。

从取证证据来源的角度，数字内容的取证方法可以分为基于事实信息的取证方法和基于伪造痕迹的取证方法，如**图4**所示。前者主要将外部事实知识作为鉴伪结果的证据，后者从伪造信息本身出发，挖掘伪造内容区分于真实内容的痕迹特征。下面对这两种取证方法进行介绍。

3.1 基于事实信息的取证方法

基于事实信息的取证研究可以追溯到早期人工新闻认证阶段，即事实核查（fact-checking）。根据杜克大学Reporters' Lab

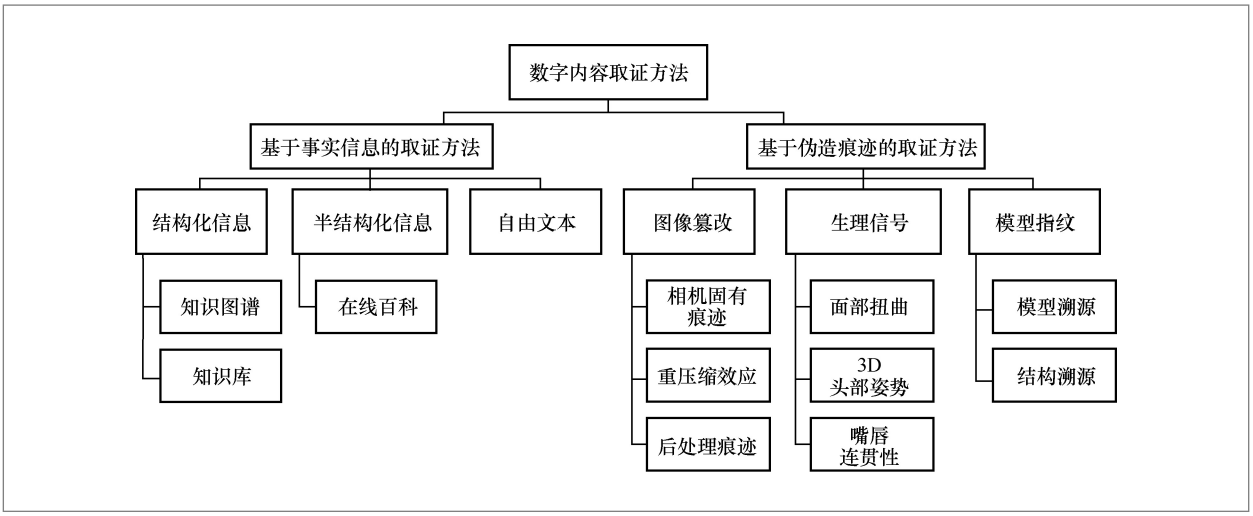


图 4 数字内容取证方法

的统计，全球目前有341个活跃的事实核查项目，分布在102个国家[99]。然而人工事实核查需要高昂的时间和人力成本，为了查证真相，专业事实核查往往需要耗费数小时甚至数天时间[100]，因此基于事实信息的自动虚假内容检测得到了研究者的广泛关注。与基于模式和社交上下文的虚假内容检测方法相比，基于事实信息的检测方法通过对检测内容本身的核实以及对外部事实信息的取证，能够在提升检测性能的同时，提供更好的可解释性、可展示性。基于事实信息的虚假内容取证与人类进行事实核查的思路相近，给定待检测内容，通过检索等手段从事实信息源获取参考信息，根据参考信息对给定内容的支持、反对立场，判断给定内容的真实性。基于事实信息的取证方法按照事实信息源不同，可分为基于知识图谱或知识库、基于在线百科、基于自由文本的方法。

● 基于知识图谱或知识库的方法。这类方法主要是利用构建好的知识图谱或知识库提供的知识，对给定内容进行事实核查。基于知识图谱或知识库的方法适用于检测有明确知识性且涉及内容动态性较弱的内容，例如：健康、科学类内容。Kou等[101]构建了关于COVID-19的知识图谱，并基于该图谱检测与COVID-19相关的不实信息。Hu等[102]通过知识检索，构建了“主题-新闻句子-相关实体”异构图用于检测。除了依靠知识图谱实现实体的概念化，还有一些工作直接考察图文实体的不一致性：Qi等[103]引入百度API识别配图中的人物、标识等，并与文本实体进行比较，其后续工作进一步区分了图像的角色以避免图文实体关系建模对装饰性配图的情况造成负面影响[104]。

● 基于在线百科的方法。这类方法主要是利用维基百科等在线百科提供的知识，对给定内容进行事实核查。2018年FEVER数据集[105]的发布促进了该领域的发展。Nie等[106]首次将神经语义匹配用于事实核查中，Zhou等[107]提出了一种基于图神经网络的证据交互方法用于推理验证环节，Jiang等[108]尝试了使用大规模预训练模型直接进行事实核查。

● 基于自由文本的方法。基于自由文本的方法使用的事实信息源一般为搜索引擎搜索返回的网页结果。Popat等[109]使用新闻内容作为检索词，在互联网搜索引擎中搜索相关网页，并将最匹配前k个片段作为候选事实证据集，之后基于双向长短时记忆网络（Bi-LSTM）和注意力机制对新闻和候选证据进行交互，实现基于证据的新闻真实性预测。此后的诸多类似工作主要围绕如何更好地实现“新闻-事实证据”交互[110-112]、增强模型的推理能力[113-114]、充分利用元信息[115]展开。

3.2 基于伪造痕迹的取证方法

根据取证痕迹的层次，基于伪造痕迹的取证方法可以分为基于传统图像篡改的取证方法、基于生理信号的取证方法和基于模型指纹的取证方法。前两种挖掘伪造痕迹在图像层和生理层的模式特征，后者则从生成模型的固有痕迹出发进行取证。

基于传统图像篡改的取证方法大多基于篡改图像底层特征的分析，包括基于相机固有痕迹[63,116-120]、重压缩效应[69,121-122]和后处理痕迹[123-127]的方法等。基于相机固有痕迹的方法基于篡改区域与未篡改区域来自不同的相机的假设，通过分析区域之间的相机统计特性的差异来定位篡改区域，采用的相机特征包括镜头色差[116-117]、光场响应不均匀性[118-119]和颜色滤波阵列[63,120]等。基于重压缩痕迹的方法通过比较不同区域的压缩痕迹的差异来检测局部的篡改区域，包括基于块状

效应[121]和基于DCT系数[69,122]的方法等。基于后处理痕迹的方法分析篡改区域为了贴合背景图像，进行的重采样[123-124]、滤波[125-126]和色彩变换[127]等操作产生的后处理痕迹。

基于生理信号特征的取证方法以深度伪造视频的生理特性的异常为中心，是构建鉴伪证据的一类重要方法。Matern等[128]发现经过伪造的视频在眼睛、牙齿、面部轮廓等视觉特征上存在的瑕疵。Yang等[129]根据伪造图片将合成的面部区域拼接到原始图像中来创建的事实，认为这样会引入3D头部姿势上的瑕疵。Agarwal等[85]认为每个人在说话时都有其固定的面部以及头部运动习惯，因此可以抽取这些运动习惯作为参考。Haliassos等[86]发现了伪造视频在嘴唇的连贯性上存在的瑕疵，并通过设计唇语识别预训练任务提高伪造检测的鲁棒性。

基于模型指纹的取证方法研究生成模型固有痕迹特征。Yu等[72]和Marra等[130]首次发现并验证了模型指纹的存在性，即生成模型和相机设备一样会在其生成图像上里留下独有的模型指纹，此发现推动了模型溯源工作的研究，为伪造图像的来源判定提供了可能。大多模型溯源工作[72,130-131]在固定有限的多个生成模型上取得了理想的溯源效果。Yang等[74]提出的DNA-Det将模型溯源的场景扩展到结构溯源，希望在改变模型的随机种子、损失函数和训练数据的情况下，还能将深伪模型生成图像溯源到对应的结构上。为了应对真实环境中存在的大量未知模型，Liu等[132]提出了开集模型溯源任务，并提出了基于渐进式开放空间扩展的模型开集溯源方法，通过渐进式增加增强模型的方法来模拟未知模型的潜在开放空间，在溯源已知模型的同时区分已知和未知模型。

4 应用

4.1 元宇宙

元宇宙是人类运用数字技术构建的，由现实世界映射或超越现实世界，可与现实世界交互的虚拟世界，具备新型社会体系的数字生活空间。元宇宙对内容的体量、内容之间的交互以及持续的内容再生有着根本性的需求。互联网的高速发展已经将内容的生产模式从专业生成内容（professional generated content，PGC）阶段带入（user generated content，UGC）阶段。近几年，人工智能技术的发展进一步改变了数字内容生产模式，进入人工智能生成内容（AI generated content，AIGC）阶段，助力内容产能和主流社交形态均实现了跨越式的提升。广大AIGC创作的内容形成了不断膨胀的内容库，能够为元宇宙源源不断地补充内容，拓宽元宇宙内容的边界。

4.2 伪造内容自动检测平台

随着数字内容生成技术的发展，互联网上出现越来越多的生成内容，其中包含恶意伪造的数字内容，因此需要伪造内容自动检测平台实时监控并预警。国内最具代表性的此类平台是“睿鉴识谣”虚假新闻自动检测平台，在国家重大事件的虚假信息治理中发挥重要作用。该平台运行9年，积累了百万级的争议性新闻线索，十万级的精标谣言数据集，针对真实应用场景中面临的信息不完整性、任务不确定性、环境强对抗性问题展开科技攻关，围绕内容可信、传播可信、用户可信多个维度

提出了一系列的信息可信度度量方法。该平台现在不仅监控传统虚假新闻，同时监控伪造数字内容。

4.3 图像视频伪造检测与溯源专用设备

数字内容伪造检测技术在落地应用大规模部署中面临两大“绊脚石”，一是处理的数据大多为高度敏感内容，如何保证环境安全可控；二是面对大并发的现网流量，深伪检测的复杂模型如何高效部署。因此，亟须大力发展针对生成数字内容的快速、准确、高效的专用检测设备。中国科学院计算技术研究所研发基于国产可控设备的软硬跨层优化加速技术，研制出国内首款深伪检测国产专用设备，既满足敏感任务的处理需求，安全可控，又大大提升检测性能，每张图片平均检测耗时10 ms，每秒视频平均检测耗时25 ms以内，处理图像视频数据流量2 GB/s，高效支持针对现网流量的高通量高并发需求，推动了国产加速卡落地应用。目前该设备已经在公安部和工信部的重大任务中实测应用，有效维护了国家网络信息安全与稳定。

5 未来展望

（1）生成技术下游个性化应用

各种生成模型已经达到了高质量的生成效果，但更多的下游任务的应用依然充满了挑战，例如：医疗保健[133]、金融服务[134]、自动驾驶[135]、机器人控制[136]等。这些应用场景要求模型提供的内容具有高可控性、可靠性、准确性、较低的容错性和可解释性特点，同时要求模型具有足够的领域知识，甚至要求模型具有常识的概念。现有高质量生成模型需要针对不同的应用场景收集训练数据和训练策略，甚至是多个对话模型、图像生成模型等协同优化与工作。

（2）视频合成

随着视觉生成模型能力的提升，视频生成能力从之前专精于部分场景，如人脸替换[137]、脸型驱动[138-139]、动作迁移[140]等，开始转向更通用的视频生成和编辑[141-142]，例如，文本到视频生成、视频补帧、视频延伸等。但是现有的通用视频生成方法还处于初期阶段，在生成质量、时序一致性、可控性等方面依然有较大的进步空间。

（3）针对特定领域的检测

深度合成技术的恶意使用在不同领域存在较大差异，很难提出一个通用的方法来检测各类深度合成数据。比如金融行业的人脸身份伪造、证件伪造，媒体行业的新闻配图伪造，国家安全层面的重要人物发言视频伪造等。为了增强伪造检测技术在不同行业真实场景下的实用性，需要结合领域内伪造内容的特点，推测伪造者的意图以及可能带来的风险，将关注点聚焦在最重要且最易被恶意伪造的内容上，发展领域专用的伪造检测技术，提升特定情境下检测算法的准确性和稳定性。

（4）复杂环境下的检测

为聚焦核心问题，现有工作大多在样本数量均衡、高质量的学术数据集上研究检测算法。但在实际应用环境中，伪造检测面临的问题则更加复杂，包括真假样本数目不均衡、互联网数据噪声高质量低、检测时效性要求高等。在学术数据集上训练的模型，在真实应用数据上往往不能取得理想的检测效果。因此，一方面需要研究高泛化性、高鲁棒性、高实时性的检测模型；另一方面要结合多种技术手段和检测维度，构建立体化分层级的检测体系，从而提高模型在复杂的真实场景下对多样化待

测数据的检测能力。

(5)认知对齐的取证方法

多篇国内外学者的综述文章指出，目前虚假检测的可解释方向有很大不足[143-145]，例如：解释不具备实际影响，模型不能有效纠正人对虚假内容的错误认知，需要与认知学心理学结合。认知对齐的可解释方法要求从人的决策模型出发设计模型解释，并通过用户实验评估解释对协同决策的有效性，是未来可解释研究发展的趋势。

(6)基于生成机理的取证方法

通用的可解释方法难以对伪造痕迹特征进行可视化并对齐到人类可理解的概念上，解决这个问题依赖于对伪造痕迹生成机理的研究，如何对生成模型指纹痕迹的形成原因进行还原和解释，是未来基于伪造痕迹的取证方法的研究趋势。

6 结束语

数字内容生成技术已经具备了大规模生成数字内容的能力，促进了大量数字内容相关产业的发展。然而数字内容生成技术高速发展的同时，也引入了更多的风险，不法分子利用前沿的数字内容生成技术进行恶意伪造，对互联网生态造成了极大的威胁。因此，发展数字内容检测和数字内容取证技术对于维护互联网生态极为重要。本文概述了现在数字内容生成、数字内容检测、数字内容取证技术，并介绍了这些技术的应用场景，最后对该研究领域的未来工作进行了展望。可以预见，在不久的将来，数字内容生成技术会有更广泛的应用，同时也需要大力发展检测和取证技术，在安全的前提下使用生成技术，人们生产内容的方式将发生巨大的变化，效率得以显著提升。

参考文献:

[1] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB]. arXiv preprint, 2020, arXiv: 2010.11929.

[2] DEVLIN J, CHANG M, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[EB]. arXiv preprint, 2018, arXiv: 1810.04805.

[3] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]. OpenAI Blog, 2018, 1(8): 9.

[4] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB]. arXiv preprint, 2021, arXiv: 2103.00020.

[5] 万小军. 智能文本生成：进展与挑战[J]. 大数据, 2023, 9(2): 99-109.
WAN X J. Intelligent text generation: recent advances and challenges[J]. Big Data Research, 2023, 9(2): 99-109.

[6] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models[EB]. arXiv preprint, 2021, arXiv:2108.07258.

[7] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[EB]. arXiv preprint, 2022, arXiv:2201.11903.

[8] ZHANG Z, ZHANG A, LI M, et al. Multimodal chain-of-thought reasoning in language models[EB]. arXiv preprint, 2023, arXiv:2302.00923.

[9] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of StyleGAN[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 8107-8116.

[10] SHEN Y J, GU J J, TANG X O, et al. Interpreting the latent space of GANs for semantic face editing[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 9240-9249.

[11] PATASHNIK O, WU Z Z, SHECHTMAN E, et al. StyleCLIP: text-driven manipulation of StyleGAN imagery[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 2065-2074.

[12] MENG C, HE Y, SONG Y, et al. SDEdit: guided image synthesis and editing with stochastic differential equations[EB]. arXiv preprint, 2021, arXiv: 2108.01073.

[13] WU Y, YU N, LI Z, et al. Membership inference attacks against text-to-image generation models[EB]. arXiv preprint, 2022, arXiv:2210.00968.

[14] CARLINI N, TRAMER F, WALLACE E, et al. Extracting training data from large language models[EB]. arXiv preprint, 2020, arXiv: 2012.07805.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all You need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 6000-6010.

[16] ELHAGE N, NANDA N, OLSSON C, et al. A mathematical framework for transformer circuits[J]. Transformer Circuits Thread, 2021.

[17] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 9.

[18] AGHAJANYAN A, OKHONKO D, LEWIS M, et al. Htlm: Hyper-text pre-training and prompting of language models[EB]. arXiv preprint, 2021, arXiv:2107.06955.

[19] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM Press, 2020: 1877-1901.

[20] QIU X, SUN T, XU Y, et al. Pre-trained models for natural language processing: a survey[J]. Science China Technological Sciences, 2020, 63(10): 1871-1897.

[21] ZHANG S, ROLLER S, GOYAL N, et al. OPT: open pre-trained transformer language models[EB]. arXiv preprint, 2022, arXiv: 2205.01068.

[22] LIU P F, YUAN W Z, FU J L, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023, 55(9): 1-35.

[23] OUYANG, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[EB]. arXiv preprint, 2022, arXiv: 2203.02155.

[24] GLAESE A, MCALEESE N, TRĘBACZ M, et al. Improving alignment of dialogue agents via targeted human judgements[EB]. arXiv preprint, 2022, arXiv: 2209.14375.

[25] COULOM R. Whole-history rating: a bayesian rating system for players of time-varying strength[C]//Proceedings of International Conference on Computers and Games. Heidelberg: Springer, 2008: 113-124.

[26] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB]. arXiv preprint, 2017, arXiv: 1707.06347.

[27] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[28] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[EB]. arXiv preprint, 2015, arXiv: 1511.06434.

[29] KARRAS T, AILA, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[EB]. arXiv preprint, 2017, arXiv: 1710.10196.

[30] DONAHUE J, KRÄHENBÜHL P, DARRELL T. Adversarial feature learning[EB]. arXiv preprint, 2016, arXiv: 1605.09782.

[31] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[C]// Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 4396-4405.

[32] KARRAS T, AITTALA M, LAINE S, et al. Alias-free generative adversarial networks[EB]. arXiv preprint, 2021, arXiv: 2106.12423.

[33] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of Wasserstein GANs[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 5769-5779.

[34] QI G J. Loss-sensitive generative adversarial networks on lipschitz densities[J]. International Journal of Computer Vision, 2020, 128(5): 1118-1140.

[35] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 2242-2251.

[36] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM Press, 2020: 6840-6851.

[37] SONG Y, ERMON S. Generative modeling by estimating gradients of the data distribution[EB]. arXiv preprint, 2019, arXiv: 1907.05600.

[38] SONG Y, SOHL-DICKSTEIN J, KINGMA D P, et al. Score-based generative modeling through stochastic differential equations[EB]. arXiv preprint, 2020, arXiv: 2011.13456.

[39] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 10674-10685.

[40] GUZHOV A, RAUE F, HEES J, et al. Audioclip: extending clip to image, text and audio[C]//Proceedings of ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 976-980.

[41] CROTHERS E, JAPKOWICZ N, VIKTOR H. Machine generated text: a comprehensive survey of threat models and detection methods[EB]. arXiv preprint, 2022, arXiv: 2210.07321.

[42] GUO B, ZHANG X, WANG Z, et al. How close is ChatGPT to human experts? comparison corpus, evaluation, and detection[EB]. arXiv preprint, 2023, arXiv: 2301.07597.

[43] GEHRMANN S, STROBELT H, RUSH A. GLTR: statistical detection and visualization of generated text[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Stroudsburg: Association for Computational Linguistics, 2019: 111-116.

[44] HOLTZMAN A, BUYS J, DU L, et al. The curious case of neural text degeneration[EB]. arXiv preprint, 2019, arXiv: 1904.09751.

[45] SEE A, PAPPU A, SAXENA R, et al. Do massively pretrained language models make better storytellers? [C]//Proceedings of the

23rd Conference on Computational Natural Language Learning (CoNLL). Stroudsburg: Association for Computational Linguistics, 2019: 843–861.

[46] FRÖHLING L, ZUBIAGA A. Feature–based detection of automated language models: tackling GPT–2, GPT–3 and Grover[J]. PeerJ Computer Science, 2021, 7: e443.

[47] CROTHERS E, JAPKOWICZ N, VIKTOR H, et al. Adversarial robustness of neural–statistical features in detection of generative transformers[C]//Proceedings of 2022 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2022: 1–8.

[48] ZIPF G K. Human behavior and the principle of least effort; an introduction to human ecology[M]. Cambridge: Addison–Wesley Press, 1949.

[49] LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB]. arXiv preprint, 2019, arXiv: 1907.11692.

[50] RODRIGUEZ J, HAY T, GROS D, et al. Cross–domain detection of GPT–2–generated technical text[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2022: 1213–1233.

[51] BAKHTIN A, GROSS S, OTT M, et al. Real or fake? learning to discriminate machine from human generated text[EB]. arXiv preprint, 2019, arXiv: 1906.03351.

[52] JI Z W, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1–38.

[53] ZHONG W J, TANG D Y, XU Z N, et al. Neural deepfake detection with factual structure of text[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2020: 2461–2470.

[54] MASSARELLI L, PETRONI F, PIKTUS A, et al. How decoding strategies affect the verifiability of generated text[C]//Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 223–235.

[55] SHAKEEL D, JAIN N. Fake news detection and fact verification using knowledge graphs and machine learning[EB]. arXiv preprint, 2021: 10.13140/RG.2.2.18349.41448.

[56] ATALLAH M J, RASKIN V, CROGAN M, et al. Natural language watermarking: design, analysis, and a proof–of–concept implementation[M]//Information hiding. Heidelberg: Springer, 2001: 185–200.

[57] TOPKARA U, TOPKARA M, ATALLAH M J. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions[C]//Proceedings of the 8th workshop on Multimedia and security. New York: ACM Press, 2006: 164–174.

[58] ABDELNABI S, FRITZ M. Adversarial watermarking transformer: towards tracing text provenance with data hiding[C]//Proceedings of 2021 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2021: 121–140.

[59] DAI L, MAO J, FAN X, et al. DeepHider: a covert NLP watermarking framework based on multi–task learning[EB]. arXiv preprint, 2022, arXiv: 2208.04676.

[60] JUEFEI–XU F, WANG R, HUANG Y H, et al. Countering malicious DeepFakes: survey, battleground, and horizon[J]. International Journal of Computer Vision, 2022, 130(7): 1678–1734.

[61] 朱新同, 唐云祁, 耿鹏志. 数字图像篡改检测技术综述[J]. 中国人民公安大学学报(自然科学版), 2022, 28(4): 87–99.
ZHU X T, TANG Y Q, GENG P Z. Survey

on digital image tampering detection technology[J]. Journal of People's Public Security University of China (Science and Technology), 2022, 28(4): 87–99.

[62] KIRCHNER M, BÖHME R. Synthesis of color filter array pattern in digital images[C]//Proceedings of Media Forensics and Security. [S.l.:s.n.], 2009: 191–204.

[63] FERRARA P, BIANCHI T, DE ROSA A, et al. Image forgery localization via fine-grained analysis of CFA artifacts[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(5): 1566–1577.

[64] ZHOU P, HAN X T, MORARIU V I, et al. Learning rich features for image manipulation detection[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1053–1061.

[65] HUH M, LIU A, OWENS A, et al. Fighting Fake News: Image Splice Detection via Learned Self-Consistency[C]//European Conference on Computer Vision. Cham: Springer, 2018: 106–124.

[66] LUKÁŠ J, FRIDRICH J, GOLJAN M. Detecting digital image forgeries using sensor pattern noise[C]//Proceedings of Security, Steganography, and Watermarking of Multimedia Contents. [S.l.:s.n.], 2006: 362–372.

[67] CHIERCHIA G, PARRILLI S, POGGI G, et al. PRNU-based detection of small-size image forgeries[C]//Proceedings of 2011 17th International Conference on Digital Signal Processing (DSP). Piscataway: IEEE Press, 2011: 1–6.

[68] COZZOLINO D, VERDOLIVA L. Camera-based image forgery localization using convolutional neural networks[C]//Proceedings of 2018 26th European Signal Processing Conference (EUSIPCO). Piscataway: IEEE Press, 2018: 1372–1376.

[69] LIN Z C, HE J F, TANG X O, et al. Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis[J]. Pattern Recognition, 2009, 42(11): 2492–2501.

[70] WANG Q, ZHANG R. Double JPEG compression forensics based on a convolutional neural network[J]. EURASIP Journal on Information Security, 2016, 2016(1): 1–12.

[71] QIAN Y, YIN G, SHENG L, et al. Thinking in frequency: face forgery detection by mining frequency-aware clues[C]//Proceedings of Computer Vision-ECCV 2020: 16th European Conference. Cham: Springer, 2020: 86–103.

[72] YU N, DAVIS L, FRITZ M. Attributing fake images to GANs: learning and analyzing GAN fingerprints[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2020: 7555–7565.

[73] GUARNERA L, GIUDICE O, BATTIATO S. Deepfake detection by analyzing convolutional traces[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2020: 666–667.

[74] YANG T Y, HUANG Z Y, CAO J, et al. Deepfake network architecture attribution[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(4): 4662–4670.

[75] CHOLLET F. Xception: deep learning with Depthwise separable convolutions[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 1800–1807.

[76] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB]. arXiv preprint, 2014, arXiv: 1409.1556.

[77] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016

IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770–778.

[78] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 1–9.

[79] CHAN C, GINOSAR S, ZHOU T H, et al. Everybody dance now[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2020: 5932–5941.

[80] SHEN J, PANG R M, WEISS R J, et al. Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions[C]// Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2018: 4779–4783.

[81] KUMAR K, KUMAR R, DE BOISSIERE T, et al. MelGAN: generative adversarial networks for conditional waveform synthesis[EB]. arXiv preprint, 2019, arXiv: 1910.06711.

[82] KAMEOKA H, KANEKO T, TANAKA K, et al. StarGAN-VC: non-parallel many-to-many voice conversion using star generative adversarial networks[C]// Proceedings of 2018 IEEE Spoken Language Technology Workshop (SLT). Piscataway: IEEE Press, 2019: 266–273.

[83] KANEKO T, KAMEOKA H, TANAKA K, et al. CycleGAN-VC3: examining and improving CycleGAN-VCs for mel-spectrogram conversion[EB]. arXiv preprint, 2020, arXiv: 2010.11672.

[84] LI Y Z, CHANG M C, LYU S W. In ictu oculi: exposing AI created fake videos by detecting eye blinking[C]//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway: IEEE Press, 2019: 1–7.

[85] AGARWAL S, FARID H, GU Y, et al. Protecting world leaders against deep fakes[C]//Proceedings of CVPR Workshops. [S.l.:s.n.], 2019: 38.

[86] HALIASSOS A, VOUGIOUKAS K, PETRIDIS S, et al. Lips don't lie: a generalisable and robust approach to face forgery detection[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 5037–5047.

[87] QI H, GUO Q, JUEFEI-XU F, et al. DeepRhythm: exposing DeepFakes with attentional visual heartbeat rhythms[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 4318–4327.

[88] ZHANG D, LI C, LIN F, et al. Detecting deepfake videos with temporal dropout 3DCNN[C]//Proceedings of IJCAI. [S.l.:s.n.], 2021: 1288–1294.

[89] ZHENG Y L, BAO J M, CHEN D, et al. Exploring temporal coherence for more general video face forgery detection[C]// Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 15024–15034.

[90] SUN Z K, HAN Y J, HUA Z Y, et al. Improving the efficiency and robustness of deepfakes detection through precise geometric features[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 3608–3617.

[91] YAMAGISHI J. Lessons learned from ASVSpoof and remaining challenges[C]// Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia. New York: ACM Press, 2022: 1–2.

[92] BU Y, SHENG Q, CAO J, et al. Combating online misinformation videos: characterization, detection, and future directions[EB]. arXiv preprint, 2023, arXiv: 2302.03242.

[93] BABU A, WANG C, TJANDRA A, et al. XLS-R: self-supervised cross-

lingual speech representation learning at scale[EB]. arXiv preprint, 2021, arXiv: 2111.09296.

[94] LYU Z Q, ZHANG S S, TANG K, et al. Fake audio detection based on unsupervised pretraining models[C]// Proceedings of ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 9231–9235.

[95] AGARWAL S, FARID H, FRIED O, et al. Detecting deep–fake videos from phoneme–viseme mismatches[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2020: 2814–2822.

[96] ZHOU Y P, LIM S N. Joint audio–visual deepfake detection[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 14780–14789.

[97] MITTAL T, BHATTACHARYA U, CHANDRA R, et al. Emotions don't lie: an audio–visual deepfake detection method using affective cues[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 2823–2832.

[98] KHALID H, KIM M, TARIQ S, et al. Evaluation of an audio–video multimodal deepfake dataset using unimodal and multimodal detectors[C]//Proceedings of the 1st Workshop on Synthetic Multimedia – Audiovisual Deepfake Generation and Detection. New York: ACM Press, 2021: 7–15.

[99] STENCEL M, LUTHER J. Annual census finds nearly 300 fact–checking projects around the world[Z]. Duke Reporters' Lab, 2020.

[100]MICALLEF N, ARMACOST V, MEMON N, et al. True or false: studying the work practices of professional fact–checkers[J]. Proceedings of the ACM on Human–Computer Interaction, 2022, 6(CSCW1): 1–44.

[101]KOU Z Y, SHANG L Y, ZHANG Y, et al. HC–COVID: a hierarchical crowdsource knowledge graph approach to explainable COVID–19 misinformation detection[J]. Proceedings of the ACM on Human–Computer Interaction, 2022, 6(GROUP): 1–25.

[102]HU L M, YANG T C, ZHANG L H, et al. Compare to the knowledge: graph neural fake news detection with external knowledge[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 754–763.

[103]亓鹏, 曹娟, 盛强. 语义增强的多模态虚假新闻检测[J]. 计算机研究与发展, 2021, 58(7): 1456–1465.
QI P, CAO J, SHENG Q. Semantics–enhanced multi–modal fake news detection[J]. Journal of Computer Research and Development, 2021, 58(7): 1456–1465.

[104]QI P, CAO J, LI X R, et al. Improving fake news detection by using an entity–enhanced framework to fuse diverse multimodal clues[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM Press, 2021: 1212–1220.

[105]THORNE J, VLACHOS A, COCARASCU O, et al. The fact extraction and VERification (FEVER) shared task[EB]. arXiv preprint, 2018, arXiv: 1811.10971.

[106]NIE Y X, CHEN H N, BANSAL M. Combining fact extraction and verification with neural semantic matching networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 6859–6866.

[107]ZHOU J, HAN X, YANG C, et al. GEAR: graph–based evidence aggregating and reasoning for fact verification[C]//

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 892–901.

[108]JIANG K, PRADEEP R, LIN J. Exploring listwise evidence reasoning with T5 for fact verification[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 402–410.

[109]POPAT K, MUKHERJEE S, YATES A, et al. DeClarE: debunking fake news and false claims using evidence-aware deep learning[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 22–32.

[110]WU L W, RAO Y, YANG X, et al. Evidence-aware hierarchical interactive attention networks for explainable claim verification[C]//Proceedings of the 29th International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2020: 1388–1394.

[111]VO N, LEE K. Hierarchical multi-head attentive network for evidence-aware fake news detection[EB]. arXiv preprint, 2021, arXiv: 2102.02680.

[112]XU W Z, WU J F, LIU Q, et al. Evidence-aware fake news detection with graph neural networks[C]//Proceedings of the ACM Web Conference 2022. New York: ACM Press, 2022: 2501–2510.

[113]MA J, GAO W, JOTY S, et al. Sentence-level evidence embedding for claim verification with hierarchical attention networks[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019.

[114]WU L W, RAO Y, SUN L, et al. Evidence inference networks for interpretable claim verification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(16): 14058–14066.

[115]MISHRA R, SETTY V. SADHAN: hierarchical attention networks to learn latent aspect embeddings for fake news detection[C]//Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval. New York: ACM Press, 2019: 197–204.

[116]JOHNSON M K, FARID H. Exposing digital forgeries through chromatic aberration[C]//Proceedings of the 8th Workshop on Multimedia and Security. New York: ACM Press, 2006: 48–55.

[117]MAYER O, STAMM M C. Accurate and efficient image forgery detection using lateral chromatic aberration[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(7): 1762–1777.

[118]CHIERCHIA G, POGGI G, SANSONE C, et al. A Bayesian-MRF approach for PRNU-based image forgery detection[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(4): 554–567.

[119]KORUS P, HUANG J W. Multi-scale analysis strategies in PRNU-based tampering localization[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(4): 809–824.

[120]POPESCU A C, FARID H. Exposing digital forgeries in color filter array interpolated images[J]. IEEE Transactions on Signal Processing, 2005, 53(10): 3948–3959.

[121]LI W H, YUAN Y, YU N H. Passive detection of doctored JPEG image via block artifact grid extraction[J]. Signal Processing, 2009, 89(9): 1821–1829.

[122]BIANCHI T, DE ROSA A, PIVA A.

Improved DCT coefficient analysis for forgery localization in JPEG images[C]// Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2011: 2444–2447.

[123]PRASAD S, RAMAKRISHNAN K R. On resampling detection and its application to detect image tampering[C]//Proceedings of 2006 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2006: 1325–1328.

[124]KIRCHNER M, BOHME R. Hiding traces of resampling in digital images[J]. IEEE Transactions on Information Forensics and Security, 2008, 3(4): 582–592.

[125]YUAN H D. Blind forensics of Median filtering in digital images[J]. IEEE Transactions on Information Forensics and Security, 2011, 6(4): 1335–1345.

[126]CHEN C L, NI J Q, HUANG J W. Blind detection of Median filtering in digital images: a difference domain based approach[J]. IEEE Transactions on Image Processing, 2013, 22(12): 4699–4710.

[127]STAMM M, LIU K J R. Blind forensics of contrast enhancement in digital images[C]//Proceedings of 2008 15th IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2008: 3112–3115.

[128]MATERN F, RIESS C, STAMMINGER M. Exploiting visual artifacts to expose deepfakes and face manipulations[C]// Proceedings of 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). Piscataway: IEEE Press, 2019: 83–92.

[129]YANG X, LI Y Z, LYU S W. Exposing deep fakes using inconsistent head poses[C]//Proceedings of ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2019: 8261–8265.

[130]MARRA F, GRAGNANIELLO D, VERDOLIVA L, et al. Do GANs leave artificial fingerprints? [C]//Proceedings of 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). Piscataway: IEEE Press, 2019: 506–511.

[131]JOSLIN M, HAO S. Attributing and detecting fake images generated by known GANs[C]//Proceedings of 2020 IEEE Security and Privacy Workshops (SPW). Piscataway: IEEE Press, 2020: 8–14.

[132]LIU H, CAO Z J, LONG M S, et al. Separate to adapt: open set domain adaptation via progressive separation[C]// Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020:2922–2931.

[133]REDDY S, ALLAN S, COGHLAN S, et al. A governance model for the application of AI in health care[J]. Journal of the American Medical Informatics Association, 2020, 27(3): 491–497.

[134]QI Y, XIAO J. Fintech: AI powers financial services to improve people's lives[J]. Communications of the ACM, 2018, 61(11): 65–69.

[135]GRIGORESCU S, TRASNEA B, COCIAS T, et al. A survey of deep learning techniques for autonomous driving[J]. Journal of Field Robotics, 2020, 37(3): 362–386.

[136]DRIESS D, XIA F, SAJJADI M S M, et al. PaLM-E: an embodied multimodal language model[EB]. arXiv preprint, 2023, arXiv: 2303.03378.

[137]NIRKIN Y, KELLER Y, HASSNER T. FSGAN: subject agnostic face swapping and reenactment[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2020:7183–7192.

[138]PRAJWAL K R, MUKHOPADHYAY R, NAMBOODIRI V P, et al. A lip sync expert is all You need for speech to lip

generation in the wild[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 484-492.

[139]SIAROHIN A, LATHUILIÈRE S, TULYAKOV S, et al. First order motion model for image animation[EB]. arXiv preprint, 2020, arXiv: 2003.00196.

[140]LIU W, PIAO Z X, MIN J, et al. Liquid warping GAN: a unified framework for human motion imitation, appearance transfer and novel view synthesis[C]// Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2020: 5903-5912.

[141]SINGER U, POLYAK A, HAYES T, et al. Make-A-Video: text-to-video generation without text-video data[EB]. arXiv preprint, 2022, arXiv: 2209.14792.

[142]ESSER P, CHIU J, ATIGHEHCHIAN P, et al. Structure and content-guided video synthesis with diffusion models[EB]. arXiv preprint, 2023, arXiv: 2302.03011.

[143]KOTONYA N, TONI F. Explainable automated fact-checking: a survey[C]//Proceedings of the 28th International Conference on Computational Linguistics. Stroudsburg: International Committee on Computational Linguistics, 2020: 5430-5443.

[144]GUO B, DING Y S, YAO L N, et al. The future of false information detection on social media: new perspectives and trends[J]. ACM Computing Surveys, 2020, 53(4): 1-36.

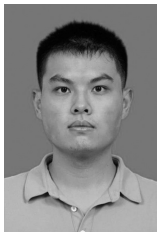
[145]SHU K, SLIVA A, WANG S H, et al. Fake news detection on social media[J]. ACM SIGKDD Explorations Newsletter, 2017, 19(1): 22-36.

作者简介

曹娟（1980- ），女，博士，中国科学院计算技术研究所研究员、前瞻研究实验室主任、数字内容合成与伪造检测实验室主任，中国科学院大学岗位教授，中国科学院计算技术研究所“十四五”规划重点研究方向“数字内容合成与伪造检测”方向牵头人。主要从事多媒体数字内容分析与伪造检测相关的研究工作。作为第一完成人，成果入选2022年世界互联网大会领先科技成果；获得2020年北京市科学技术进步奖一等奖、2020年北京市三八红旗奖章及2021年中国人工智能大赛“创新人物”和“创新之星”称号。作为项目负责人，围绕多媒体内容安全方向承担十余项国家级重要课题。

朱勇椿（1996- ），男，博士，2023年毕业于中国科学院计算技术研究所，主要研究方向为迁移学习、推荐系统、虚假新闻检测。

亓鹏（1996- ），女，博士，2023年毕业于中国科学院计算技术研究所，主要研究方向为虚假信息检测、多媒体内容分析。



黄子尧(1995–),男,中国科学院计算技术研究所博士生,主要研究方向为数字人合成技术。

杨天韵(1997–),女,中国科学院计算技术研究所博士生,主要研究方向为深度生成模型溯源、人工智能安全。

王政嘉(1998–),女,中国科学院计算技术研究所博士生,主要研究方向为可解释虚假信息检测。

卜语嫣(2000–),女,中国科学院计算技术研究所硕士生,主要研究方向为多模态虚假信息检测。

收稿日期: 2023-04-12

通信作者: 曹娟, caojuan@ict.ac.cn

基金项目: 国家自然科学基金资助项目(No.62203425);中国科学院项目(No.E141020);中国博士后科学基金特别资助(No.2022TQ0344);博士后国际交流计划引进项目(No.YJ20220198)

Foundation Items: The National Natural Science Foundation of China(No.62203425), The Project of Chinese Academy of Sciences(No.E141020), The China Postdoctoral Science Foundation(No.2022TQ0344), The International Postdoctoral Exchange Fellowship Program by Office of China Postdoc Council(No.YJ20220198)