ChatGPT

近四年图像与视频伪造检测相关研究综述（2021–2024）

近年来，在图像和视频伪造检测领域出现了大量新方法和技术，尤其聚焦于**人脸Deepfake伪造检测**和**通用图像伪造检测**，并伴随模型可解释性、多模态融合以及攻防对抗等方向的发展。下面我们按类别汇总2021年至2024年间在ECCV、CVPR、ICCV、ICLR、ICML、NeurIPS、AAAI、ACM　MM等顶会发表的相关论文，每篇列出标题、会议年份、核心方法贡献、代码地址及所针对的主要伪造类型。

人脸伪造（Deepfake）检测研究

这一类别涵盖针对人脸替换、表情/动作重演等*Deepfake*视频的检测方法。许多工作侧重于提升跨数据集的泛化能力，有些利用多模态特征（如人脸动作单元、生理信号）或模型注意力机制来提高可解释性。

论文标题	会议（年份）	核心贡献/方法简述	GitHub链接	所涉伪造类型
AltFreezing for More General Video Face Forgery Detection [1] [2] （更通用的视频人脸伪造检测的交替冻结训练策略）	CVPR 2023	提出**交替冻结（AltFreezing）**训练策略，将3D时空卷积模型的空间层和时间层参数交替冻结训练，迫使模型同时学到**空间伪造痕迹**和**时间不连续伪造痕迹**，并配合多种视频级数据增强以增强跨数据集泛化能力 [1] [2]。实验证明该方法在检测未知操纵类型时效果显著提升。	【19†GitHub】（官方实现）	人脸Deepfake（换脸、表情伪造）
AUNet: Learning Relations Between Action Units for Face Forgery Detection [3] （基于面部动作单元关系学习的人脸伪造检测）	CVPR 2023	利用**面部动作单元（AU）**先验提出动作单元关系学习框架，将人脸表情肌肉动作作为辅助特征。包括动作单元关系Transformer（ART）和伪造AU预测模块，实现不同AU区域间相关性的建模，提高对伪造的判别泛化能力 [3]。该方法利用AU语义增强了模型解释性，在跨数据集评估中性能领先。	【11†GitHub】（官方实现）	人脸Deepfake（换脸、表情伪造）

论文标题	会议（年份）	核心贡献/方法简述	GitHub链接	所涉伪造类型
Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization [4] （隐式身份泄漏：阻碍*Deepfake*检测泛化的绊脚石）	CVPR 2023	首次分析并证明Deepfake检测模型泛化不佳的原因在于**意外学习了身份特征**（将伪造人脸的身份信息当作判别依据）。提出一种**ID无关检测模型**，通过定位并仅关注伪造区域的局部特征（引入伪造区域检测模块，弱化全局身份信息）来减少身份泄漏影响 [4]。实验证明该方法在跨数据集测试中显著提升检测准确率。	【20†GitHub】	人脸Deepfake（换脸）
LAA-Net: Localized Artifact Attention Network for Quality-Agnostic and Generalizable Deepfake Detection [5] [6] （局部伪造痕迹注意力网络：与压缩质量无关的通用*Deepfake*检测）	CVPR 2024	针对高质量Deepfake难检测的问题，提出显式**局部伪造痕迹注意力**机制。采用多任务学习框架：一方面结合**热图分支**和**自一致性分支**生成显式注意力，引导模型专注于伪造易发的局部区域；另一方面设计改进的**增强特征金字塔E-FPN**融合多尺度低级特征以捕捉细微伪造痕迹 [5] [6]。在多基准数据集上取得更高AUC和AP，尤其对压缩失真不敏感。	将在GitHub发布（论文承诺公开）	人脸Deepfake（高真实感换脸、表情伪造）
Rethinking the Up-Sampling Operations in CNN-based Generative Network for Generalizable Deepfake Detection [7] [8] （重新思考生成网络上采样操作以提升*Deepfake*检测泛化）	CVPR 2024	从生成模型架构出发，揭示GAN和扩散模型中的**上采样算子会引入局部像素相关性伪造痕迹**，不仅限于频域伪迹 [7]。据此提出伪造结构特征**邻域像素关系（NPR）**作为通用伪造表示，通过捕捉上采样造成的局部像素依赖来区分真伪。基于28种生成模型的大规模实验表明，该NPR表示令检测器泛化性显著提升（跨模型检测准确率提高11.6%），达到新的SOTA [8]。	【43†GitHub】	人脸Deepfake + 通用生成图像

论文标题	会议（年份）	核心贡献/方法简述	GitHub链接	所涉伪造类型
UIA-ViT: Unsupervised Inconsistency-Aware Method based on Vision Transformer for Face Forgery Detection [9] [10] （基于*ViT*的不一致性检测，无监督人脸伪造检测）	ECCV 2022	提出一种**无监督的一致性扰动学习**方法UIA-ViT，只需视频级真伪标签而无需逐帧像素伪造标注 [9]。利用ViT自注意力天然建模全局一致性关系，设计了两个模块：①逐帧对比生成伪造区域的**未监督Patch一致性学习**（UPCL），渐进产生伪造区域的伪标签并训练模型关注这些位置的不一致；②**逐步一致性加权融合**（PCWA），将前序层学到的不一致特征融合到最终判别中 [10]。该方法有效挖掘了伪造区域的时空不一致性，在无真实伪造掩膜的情况下实现了与有监督方法相当的泛化检测性能。	【3†GitHub】（论文附官方实现）	人脸Deepfake（换脸、表情伪造，无监督检测）
ID-Reveal: Identity-Aware DeepFake Video Detection [11] [12] （利用身份特征的人脸*Deepfake*视频检测）	ICCV 2021	面向不同类型人脸伪造方法难以泛化的问题，提出**基于身份一致性的深度伪造检测**。利用每个真实人物的视频建立其说话时**面部动态特征表示**，通过度量测试视频与该人物参考视频的**时间运动特征差异**来判定真伪 [13] [14]。该度量学习方法仅用真人视频训练，无需任何伪造样本也能学习到个人特有的动态模式，从而在**无监督情况下对全新类型的换脸或表情伪造**取得15%以上的准确率提升 [15]。尤其在低清视频上表现出色，对人脸替换和表情重演均具备较强泛化性。	【3†GitHub】（项目代码）	人脸Deepfake（换脸、表情重演）

注：上述方法主要针对人脸伪造（换脸、表情迁移等）的检测。其中AUNet等融入了多模态语义信息（动作单元），LAA-Net等通过显式定位伪造区域提高了结果可解释性（如注意力热图），AltFreezing、ID-Reveal等则注重提升跨不同Deepfake技术的检测泛化能力。

通用图像伪造与合成内容检测研究

这一类别涵盖人脸以外的图像篡改（如拼接、属性编辑）以及纯生成模型合成图像的检测方法。许多工作关注构建对未知伪造类型具鲁棒性的特征表示，包括融合高低层线索、频域与空间信息等。

论文标题	会议（年份）	核心贡献/方法简述	GitHub链接	所涉伪造类型
TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization [16] [17] （*TruFor*：融合全方位线索的图像篡改检测与定位）	CVPR 2023	提出通用图像取证框架TruFor，将**RGB视觉线索**与**噪声级指纹线索**融合 [16]。利用Transformer架构提取图像高层语义和通过自监督学习获得的“相机噪声指纹”特征，将伪造检测视为**偏离真实相机模式的异常检测** [18]。该方法无需限定伪造类型，可**同时输出像素级篡改区域掩码和整图篡改评分**，并提供可靠性评估图减少误报 [19]。实验在多数据集上表明TruFor对各种本地篡改（传统“cheapfake”如拼接篡改及AI深度伪造）均能可靠检测和定位，精度超越以往方法。代码已公开。	【12†GitHub】（官方实现）	图像篡改（拼接、合成脸等）
Towards Universal Fake Image Detectors That Generalize Across Generative Models [20] [21] （通用假图像检测：跨生成模型的泛化）	CVPR 2023	针对传统检测器**难以跨不同生成架构泛化**的问题，提出摒弃训练分类器识别真假的惯用范式，而改用大型预训练模型的**通用特征空间 + 非学习式分类** [20]。具体地，利用CLIP等视觉语言模型的预训练特征，不额外训练判别器，而通过**最近邻检索**或线性探针来判定图像真伪 [22]。如此无需将未见过的生成模型图像归入训练时的“真实”类，从而显著提升对扩散模型、AR模型等新型生成假图的识别（在未见扩散模型上mAP提高15%、准确率提高25%以上） [21]。	【41†GitHub】	生成模型图像（GAN、扩散模型等）

论文标题	会议（年份）	核心贡献/方法简述	GitHub链接	所涉伪造类型
Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection [23] [24] （利用梯度信息学习*GAN*生成图像的通用伪造表征）	CVPR 2023	提出**梯度表示法LGrad**框架，将预训练CNN模型的中间梯度作为刻画GAN图像伪造痕迹的通用特征 [23]。具体做法是利用ImageNet预训模型，将输入图像通过该模型求取梯度图，从中滤除内容，仅保留与模型任务相关的辨别性像素 [25]。这些梯度能够呈现生成图像的泛化伪造模式，再输入简单分类器判别真伪。LGrad将依赖数据的检测转化为依赖预训练模型，从而**首次利用梯度作为GAN伪造特征**，在跨数据集跨GAN模型检测中提升显著（总体精度提高11.4%） [24]。	【0†GitHub】（官方实现）	GAN生成假脸及合成图像
DRCT: Diffusion Reconstruction Contrastive Training towards Universal Detection of Diffusion Generated Images [26] [27] （基于扩散重建对比训练的扩散模型图像检测）	ICML 2024	专注**扩散模型生成图像**检测的泛化难题，提出**Diffusion重建对比训练（DRCT）**框架 [26]。核心思想：通过高保真扩散模型对真实图像进行重建来生成“难判别样本”，使其几乎乱真，如果分类器能识别这些高伪真样本，那么对更明显的伪造也有检测力 [26]。DRCT据此生成**高相似度伪造**作为难例，并设计对比训练目标引导模型学习扩散伪造的独特痕迹。同时构建包含16种扩散模型的大规模数据集DRCT-2M用于评估 [28]。增强主流检测器后，跨未见扩散模型的检测准确率提升超过10% [29]。	【36†GitHub】（即将开放）	扩散模型生成图像

论文标题	会议（年份）	核心贡献/方法简述	GitHub链接	所涉伪造类型
Responsible Disclosure of Generative Models Using Scalable Fingerprinting [30] （利用可扩展指纹实现生成模型负责任披露）	ICLR 2022	提出一种主动防伪策略：在生成模型中嵌入**可检测指纹信息**。方法为训练GAN等生成模型时加入特定**数字水印/指纹**，使得生成图像带有模型独有的频域标记，从而允许日后对图像来源进行鉴别 [30]。这种指纹嵌入具有可扩展性和鲁棒性，支持公开发布生成模型的同时又能负责地区分其输出与真实图像。作者公开了指纹生成代码，证明在不显著影响生成质量下即可检测出合成内容。	【3†GitHub】（官方实现）	生成模型图像（带模型指纹）

注：在通用伪造检测方向，TruFor等方法通过融合图像内容和摄影器材噪声等多线索，实现对各类篡改的**检测+定位** [19]；而针对纯生成图像的新挑战，出现了利用**预训练模型特征空间**（如Universal Fake Image Detector）、**梯度域泛化特征**（LGrad）以及**对抗生成难例训练**（DRCT）等新思路来提升对GAN、扩散模型等**未见生成方法**的识别能力。

对抗攻击与防御方案研究

本类别涵盖为了攻防对抗而提出的*Deepfake*生成攻击和检测防御方案，包括向伪造模型添加扰动以防止生成、对检测模型实施后门攻击，以及提升检测模型鲁棒性和评测标准化等。

论文标题	会议（年份）	核心贡献/方法简述	GitHub链接	所涉伪造类型（攻防场景）
CMUA-Watermark: A Cross-Model Universal Adversarial Watermark for Combating Deepfakes [31] [32] （跨模型通用对抗水印对抗 *Deepfakes*）	AAAI 2022	提出一种**跨模型通用对抗性水印**方案，用于主动防御人脸Deepfake合成。通过迭代攻击多个人脸属性编辑和换脸模型，生成一个可叠加于任意人脸图像的细微水印，使各种Deepfake模型在处理带水印的人脸时产生明显失真 [31] [32]。核心包括：跨模型迭代攻击优化、**两级扰动融合**策略缓解不同图像和模型间冲突，以及自动调整不同模型攻击步长的启发式方法 [33]。实验表明，该水印一次嵌入即可保护大批人脸，对多种Deepfake模型输出均能产生肉眼可辨伪影，成功率和视觉质量均优于现有方法。	【37†GitHub】	主动防御： 人脸Deepfake生成扰动

论文标题	会议（年份）	核心贡献/方法简述	GitHub链接	所涉伪造类型（攻防场景）
Poisoned Forgery Face: Towards Backdoor Attacks on Face Forgery Detection [34] [35] （中毒人脸：针对人脸伪造检测的后门攻击）	ICLR 2024	揭示并实现了**针对Deepfake检测器的后门攻击新威胁** [36]：攻击者可在训练集中植入带特定触发图案的“中毒”伪造样本，使检测模型学习到后门。当测试伪造图像中混入该触发时，模型就输出错误判断（将伪造人脸判为真实） [34]。提出的PFF框架采用干净标签方式进行后门植入，包含：可扩展的**触发生成器**，生成带平移敏感特性的触发图案（利用卷积过程保证触发在图像中的相对位置生效）；以及基于人脸关键点区域的**相对嵌入**策略，提高触发样本的隐蔽性 [37]。结果在多种检测器上成功嵌入后门，攻击成功率较基线提升16%以上且触发无明显可见 [35]。	【38†GitHub】	进攻： 后门攻击人脸伪造检测

论文标题	会议（年份）	核心贡献/方法简述	GitHub链接	所涉伪造类型（攻防场景）
OST: Improving Generalization of DeepFake Detection via One-Shot Test-Time Training [38] （单次测试时训练提升*Deepfake*检测泛化）	NeurIPS 2022	提出一种测试阶段自适应的方法OST：**单样本测试时训练**。针对检测器难以适应新型伪造的问题，在每个测试样本上执行一次小规模梯度更新：从该测试图像**合成伪造样本**并设定辅助任务目标，对模型参数进行一次梯度调整，然后再判别该样本 [39]。同时通过元学习预训练，使模型能**一阶梯度快速收敛**（一次更新即可显著改善性能）[40]。大量实验显示，OST在各种数据集上将现有检测模型对未知伪造的泛化准确率和对抗压缩等鲁棒性都有明显提升 [41]。	【39†GitHub】（官方实现）	防御：增强检测器跨域适应
DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection [42] [43] （*DeepfakeBench*：深度伪造检测综合基准）	NeurIPS 2023 (Datasets/ Benchmarks Track)	发布首个**深度伪造检测综合基准** DeepfakeBench，解决领域内评测不统一的问题 [44]。主要贡献：1）提供统一的数据管理与预处理管线，确保不同检测方法接收**一致输入** [45]；2）集成15种SOTA检测算法的开源实现，方便公平比较；3）制定标准化的评估协议和指标，包含9个公开Deepfake数据集、多种评价指标和分析工具 [42]。Benchmark进行了全面实验分析（如数据增强影响、特征Backbone差异等）[46] 并公布了**模块化代码库**，旨在推动该领域未来研究的透明性和可比性。	【41†GitHub】	评估：Deepfake检测方法全面评测

注：在对抗方向，CMUA-Watermark提供了主动**干预生成**的思路，通过预先在图像加扰动保护隐私 [47] ；Poisoned Forgery Face则暴露了检测器的潜在**数据后门风险**，提醒需要防范训练数据污染；OST方法属于提升检测鲁棒性的**测试时自适应**策略，而DeepfakeBench基准的推出有助于整个领域的**标准化评测**和方法改进。

---

[1] [2] GitHub - ZhendongWang6/AltFreezing: [CVPR 2023 Highlight] Official implementation of the paper: "AltFreezing for More General Video Face Forgery Detection"
https://github.com/ZhendongWang6/AltFreezing

[3] AUNet: Learning Relations Between Action Units for Face Forgery Detection
https://openaccess.thecvf.com/content/CVPR2023/papers/Bai_AUNet_Learning_Relations_Between_Action_Units_for_Face_Forgery_Detection_CVPR_2023_paper.pdf

[4] Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization
https://openaccess.thecvf.com/content/CVPR2023/papers/Dong_Implicit_Identity_Leakage_The_Stumbling_Block_to_Improving_Deepfake_Detection_CVPR_2023_paper.pdf

[5] [6] [2401.13856] LAA-Net: Localized Artifact Attention Network for High-Quality Deepfakes Detection
https://ar5iv.labs.arxiv.org/html/2401.13856

[7] [8] [2312.10461] Rethinking the Up-Sampling Operations in CNN-based Generative Network for Generalizable Deepfake Detection
https://ar5iv.labs.arxiv.org/html/2312.10461

[9] [10] ecva.net
https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136650384.pdf

[11] [12] [13] [14] [15] ID-Reveal: Identity-Aware DeepFake Video Detection
https://openaccess.thecvf.com/content/ICCV2021/papers/Cozzolino_ID-Reveal_Identity-Aware_DeepFake_Video_Detection_ICCV_2021_paper.pdf

[16] [17] [18] [19] [2212.10957] TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization
https://arxiv.org/abs/2212.10957

[20] [21] [22] Towards Universal Fake Image Detectors That Generalize Across Generative Models
https://openaccess.thecvf.com/content/CVPR2023/papers/Ojha_Towards_Universal_Fake_Image_Detectors_That_Generalize_Across_Generative_Models_CVPR_2023_paper.pdf

[23] [24] [25] Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection
https://openaccess.thecvf.com/content/CVPR2023/papers/Tan_Learning_on_Gradients_Generalized_Artifacts_Representation_for_GAN-Generated_Images_Detection_CVPR_2023_paper.pdf

[26] [27] [28] [29] DRCT: Diffusion Reconstruction Contrastive Training towards Universal Detection of Diffusion Generated Images
https://proceedings.mlr.press/v235/chen24ay.html

[30] GitHub - Daisy-Zhang/Awesome-Deepfakes-Detection: A list of tools, papers and code related to Deepfake Detection.
https://github.com/Daisy-Zhang/Awesome-Deepfakes-Detection

[31] [32] [33] [47] [2105.10872] CMUA-Watermark: A Cross-Model Universal Adversarial Watermark for Combating Deepfakes
https://ar5iv.labs.arxiv.org/html/2105.10872

34 35 36 37 openreview.net
https://openreview.net/pdf?id=8iTpB4RNvP

38 39 40 41 OST: Improving Generalization of DeepFake Detection via One-Shot Test-Time Training | OpenReview
https://openreview.net/forum?id=YPoRoad6gzY

42 43 44 45 46 [2307.01426] DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection
https://arxiv.org/abs/2307.01426