中图法分类号： **文献标识码：** **文章编号：** 1006-8961(XXXX)XX-0001-38

论文引用格式： Sang Nong, Huang Kaiqi, Zhao Yao, Gao Changxin, Kao Yueying, Tan Chuangchuang, Wang Xiang, Wu Meiqi, Yin Wenti. Advances in Video and Image Security Research in the Era of Large Models[J/OL]. Journal of Image and Graphics, XXXX:1-38. DOI: 10.11834/jig.250656. (桑农, 黄凯奇, 赵耀, 高常鑫, 考月英, 谭创创, 王翔, 武美奇, 尹文体. 大模型时代的视频与图像安全研究进展[J/OL]. 中国图象图形学报, XXXX:1-38. DOI: 10.11834/jig.250656.)[DOI:10.11834/jig.250656]

大模型时代的视频与图像安全研究进展

桑农[1],黄凯奇[2],赵耀[3],高常鑫[1],考月英[4],谭创创[3],王翔[1],武美奇[2],尹文体[1]

1. 华中科技大学人工智能与自动化学院,武汉 430074; 2. 中国科学院自动化研究所,北京 100190; 3. 北京交通大学计算机科学与技术学院,北京 100044; 4. 北京市科学技术研究院信息与人工智能技术研究所,北京 100089

摘　要： 随着多模态大模型与生成式人工智能技术的快速发展,图像与视频的获取、理解与生成方式正在发生深刻变革。以视觉—语言预训练模型和扩散生成模型为代表的新一代人工智能体系,在语义对齐、跨模态理解与高保真内容生成等方面展现出强大的能力,显著推动了智能安防、内容生产、工业检测和公共治理等应用场景的发展。然而,视觉智能能力的快速扩张也带来了日益突出的安全风险与治理挑战:在理解层面,模型在复杂环境、开放场景和弱监督条件下易产生误判、偏差与鲁棒性不足;在生成层面,高保真合成图像与视频被滥用于深度伪造、虚假信息传播和隐私侵犯,对社会信任与公共安全构成威胁。因此,围绕“大模型时代的视频与图像安全”开展系统性研究具有重要的理论价值与现实意义。本文从图像与视频理解安全和图像与视频生成安全两条主线出发,系统综述了相关技术的研究进展。在理解安全方面,重点总结了全监督、半监督、弱监督和无监督异常检测方法的技术演进,并进一步归纳了基于视觉—语言大模型的零样本、开放词汇和可解释异常检测新范式;在生成安全方面,围绕生成对抗网络与扩散模型的发展脉络,系统分析了图像与视频生成技术的安全风险、深度伪造检测方法及其在政策监管与工程实践中的应用现状。最后,本文讨论了当前研究面临的关键挑战,并展望了大模型时代图像与视频安全研究的未来发展趋势,为相关领域的学术研究与工程应用提供参考。

关键词： 多模态大模型;生成式人工智能;图像视频安全;异常检测;深度伪造检测

Advances in Video and Image Security Research in the Era of Large Models

Sang Nong[1], Huang Kaiqi[2], Zhao Yao[3], Gao Changxin[1], Kao Yueying[4], Tan Chuangchuang[3], Wang Xiang[1], Wu Meiqi[2], Yin Wenti[1]

1. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074; 2. Institute of Automation, Chinese Academy of Sciences, Beijing 100190; 3. School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044; 4. Institute of Information and Artificial Intelligence Technology, Beijing Academy of Science and Technology, Beijing 100089

Abstract: With the rapid advancement of multimodal large models and generative artificial intelligence, the paradigms of image and video acquisition, understanding, and generation are undergoing profound transformations. In recent years, new-generation artificial intelligence systems represented by vision-language pretraining models and diffusion-based generative models have achieved remarkable progress in semantic alignment, cross-modal understanding, and high-fidelity content generation. By leveraging large-scale data and powerful representation learning capabilities, these models have significantly enhanced the performance and flexibility of visual intelligence systems, promoting their widespread adoption in intelligent security, content creation, industrial inspection, and public governance. At the same time, the increasing capability and deployment of visual intelligence systems have also exposed a series of security risks and governance challenges,

收稿日期: 2025-12-29;**修回日期:** 2026-02-12

基金项目: 国家自然科学基金项目(U22B2053,U24B20179,62406030),北京市自然科学基金项目(4264127)

which have become increasingly prominent and cannot be overlooked. From the perspective of image and video understanding, existing visual models are often required to operate in complex and open-world environments characterized by dynamic scenes, background clutter, illumination variation, viewpoint changes, and long-tailed event distributions. In such scenarios, the cost of obtaining large-scale, fine-grained annotations is prohibitively high, leading many practical systems to rely on limited supervision or weak labels. Although large pretrained models exhibit strong generalization ability, they still suffer from misclassification, semantic bias, and insufficient robustness when faced with domain shift, distribution mismatch, and unseen abnormal patterns. These limitations are particularly evident in safety-critical applications, where incorrect predictions or unstable behavior may result in serious consequences. Therefore, improving the reliability, robustness, and interpretability of image and video understanding systems has become a central topic in visual security research. In this context, anomaly detection has emerged as a core task for understanding security, as it aims to identify rare, unexpected, or abnormal events from complex visual data. Existing anomaly detection methods can be broadly categorized into fully supervised, semi-supervised, weakly supervised, and unsupervised paradigms according to the availability and granularity of annotations. Fully supervised approaches rely on precise frame-level or pixel-level labels and typically achieve strong performance under controlled conditions, but their scalability and generalization ability are limited in real-world scenarios. Semi-supervised and unsupervised methods, which assume access only to normal samples during training, attempt to model normal patterns through reconstruction, prediction, or one-class learning, and detect anomalies as deviations from learned normality. Weakly supervised approaches, often formulated under the multiple instance learning framework, strike a balance between annotation cost and detection performance, but still face challenges in accurate temporal localization and semantic interpretation of anomalies. With the emergence of vision-language large models, recent studies have begun to explore new paradigms for anomaly detection and visual understanding security. By leveraging pretrained cross-modal representations and natural language supervision, vision-language models enable zero-shot and few-shot anomaly detection, reducing reliance on task-specific annotations. Open-vocabulary anomaly recognition further allows models to detect and describe abnormal events beyond a fixed set of predefined categories, improving flexibility in open-world environments. In addition, explainable anomaly detection methods based on cross-modal alignment and attention mechanisms provide semantic-level interpretations for detected anomalies, enhancing transparency and trustworthiness. These advances indicate a clear trend toward more general, scalable, and interpretable understanding security frameworks. From the perspective of image and video generation, recent progress in generative adversarial networks (generative adversarial networks, GAN) and diffusion models (diffusion models, DM) has greatly improved the realism and controllability of synthesized visual content. GAN-based methods introduced adversarial learning mechanisms to produce visually plausible samples, while diffusion models further enhanced generation quality and training stability through iterative denoising processes. Building upon these foundations, modern text-to-image and text-to-video generation systems integrate large vision-language models to achieve fine-grained semantic control, enabling the generation of complex scenes that closely resemble real-world data. These developments have brought significant benefits to creative industries and visual content production, but they have also amplified security risks associated with the misuse of generative technologies. High-quality synthetic images and videos can be maliciously exploited for deepfake generation, false information dissemination, identity impersonation, and privacy infringement, posing direct threats to social trust and public security. As a result, generation security has become an essential component of image and video security research. Existing deepfake detection methods have evolved alongside generative models and can be roughly divided into several categories, including approaches based on visual artifacts, frequency-domain characteristics, temporal consistency, and semantic coherence. While early methods focused on detecting low-level inconsistencies introduced by generation algorithms, recent approaches increasingly emphasize higher-level semantic and temporal modeling to cope with the rapid improvement of generative quality. In addition to algorithmic research, the security issues associated with image and video generation have also attracted growing attention in policy regulation and engineering practice. Detection systems are being integrated into real-world platforms to support content moderation, authenticity verification, and risk assessment. Meanwhile, regulatory frameworks and technical guidelines are gradually being established to govern the responsible use of generative models. These efforts highlight the necessity of combining technical solutions with governance mechanisms to address the challenges posed by generative visual technologies. Finally,

despite substantial progress, image and video security in the era of large models still faces several open challenges. For understanding security, improving robustness under complex environmental changes, achieving precise temporal and spatial localization of anomalies, and enhancing semantic interpretability remain key research problems. For generation security, developing generalizable deepfake detection methods that can adapt to rapidly evolving generative models remains an open issue. Moreover, balancing model capability, usability, and security constraints requires further exploration. By systematically reviewing existing research from the perspectives of understanding security and generation security, this paper aims to provide a structured overview of the current landscape and to offer insights into future research directions for image and video security in the era of large models.

Key words: Multimodal Large Models; Generative AI; Image and Video Security; Anomaly Detection; Deepfake Detection

0 引　言

在信息与智能化技术高速发展的当下，图像与视频已成为全球信息传播的主要载体，也是人工智能算法感知世界的核心模态。随着计算机视觉、语义理解和生成式人工智能的突破，图像与视频的获取、分析、生成与分发方式正经历深刻变革。以多模态大模型为代表的新一代人工智能体系，凭借大规模参数化结构和跨模态自监督学习，在语义抽象、视觉表征与知识融合等方面实现统一建模，显著增强了泛化与迁移能力。代表性模型如CLIP（Radford等，2021）、ALIGN（Jia等，2021a）、Florence（Yuan等，2021）、GPT-4V（Yang等，2023）、Gemini（Team等，2024）等在图文对齐和视觉语言理解中展现出强大的语义推理能力，视频生成模型如Sora、Runway Gen-3、Pika、Vidu则在动态内容生成、时序一致性和语义控制方面取得突破。这些技术正推动智能安防、媒体生产、影视娱乐、交通管理、工业质检、教育培训等行业的智能化发展。

与此同时，视觉智能的快速扩张带来了显著的安全挑战。在理解环节，模型在复杂光照、遮挡、伪装和对抗样本等条件下仍易出现误识别与偏差；在开放环境中，大模型可能因训练数据偏差或语义迁移失真而产生误判、歧视或安全漏洞。在生成环节，高保真合成内容被滥用于虚假信息传播、隐私篡改与深度伪造，引发社会信任与伦理风险。因此，围绕“大模型时代的图像视频安全”，从理解安全与生成安全两条主线开展系统研究，已成为学术界与产业界的共同议题。

图1为本文的总体架构与内容组织，主要包括图像与视频理解安全、图像与视频生成安全和总结与展望三个部分。

1）　图像与视频理解安全

图像与视频理解安全旨在保障AI系统视觉感知与理解过程的可靠性、抗干扰性、可信性及合规性，涵盖异常检测、违规内容识别、入侵检测等

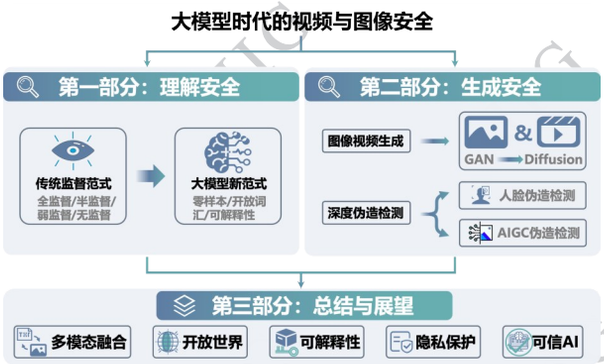

图1　本文的总体架构与内容组织

Fig. 1　Overall architecture and content organization of this paper

核心任务，并延伸至多模态大模型驱动的语义理解、跨域推理与弱监督/零样本学习。

视觉安全研究历经三次关键演进：早期依赖方向梯度直方图（histogram of oriented gradients，HOG）、尺度不变特征变换（scale-invariant feature transform，SIFT）等手工特征与支持向量机（support vector machine，SVM）、Adaboost等传统分类器，在复杂场景与跨域任务中表现受限；在深度学习时代，Faster R-CNN、YOLO系列等图像目标检测网络，以及三维卷积神经网络（3D convolutional neural networks，3D CNN），如卷积3D网络（convolutional 3D，C3D）、膨胀3D网络（inflated 3D，I3D）等视频行为分析网络显著提升了目标检测与行为分析精度，推动智能安防与舆情监测落地；在大模型时代，CLIP、

ALIGN等通过大规模图文预训练实现开放世界语义理解，VideoCLIP、ImageBind等进一步整合多模态数据，构建跨域统一表征，升级异常检测、合规审核的智能化水平。

在应用层面，大模型支持实时检测越界入侵、危险行为等事件，精准识别暴恐、涉黄等敏感内容，赋能工业与交通场景的安全监测与风险预警。当前研究正深度融合模型可信性、对抗鲁棒性与数据治理，学界探索可解释性、置信度校准与对抗训练，产业界构建安全评测体系与可信平台，结合国际国内相关法规要求，推动理解安全从算法研究迈向系统治理与工程化落地，成为多模态时代AI可控可信与数字社会治理协同发展的核心技术基石。

2） 图像与视频生成安全

图像与视频生成是视觉智能的另一条核心主线，其目标是依据文本、草图、姿态或噪声等输入生成逼真且语义一致的视觉内容。研究重点在于保证生成结果的真实性、多样性、可控性与时序一致性。

生成技术的发展经历了从传统方法到深度学习的跨越。早期依赖纹理合成、图像拼接和基于物理的渲染，缺乏多样性和自动化。2014年生成对抗网络（generative adversarial networks，GAN）的提出成为关键转折，变分自编码器（variational autoencoder，VAE）提供了概率生成思路。代表模型包括DCGAN、Pix2Pix、CycleGAN、StyleGAN等，其中StyleGAN实现了高分辨率逼真生成。此后扩散模型崛起（DALL·E 2、Stable Diffusion、Midjourney），通过噪声加减过程生成高质量图像，并结合语言模型实现文本驱动生成，引发人工智能生成内容（artificial intelligence generated content，AIGC）浪潮。视频生成作为图像生成的时序扩展，借助时空注意力与3D卷积建模时序信息，代表系统包括Runway、Pika与Sora。当前研究热点集中在可控生成、3D内容生成、大模型融合与实时生成。

生成技术的普及也带来了深度伪造风险。高保真伪造内容难以通过肉眼辨识，可能引发虚假宣传、信息诈骗与隐私侵犯，威胁社会信任与国家安全。对此，各国相继出台政策加强监管。美国《恶意伪造禁令法案》、欧盟《通用数据保护条例》和《人工智能法案》，以及我国《网络安全法》《人工智能生成内容标识办法》等均明确规范深度伪造内容的生产与传播。产业界亦在检测领域积极布局，谷歌、微软、Meta、英特尔、华为、百度及多家初创企业（SensityAI、AmberVideo、DeepwareAI等）推动伪造检测与溯源技术发展。

深度伪造检测的核心在于识别底层伪造痕迹与高层语义异常以提升模型泛化性与可解释性。早期方法聚焦人脸伪造的纹理与频域特征，现已扩展至多模态内容检测，结合语义一致性分析识别违背物理规律或常识的伪造现象。

综上，图像与视频生成安全已成为人工智能时代的关键研究前沿，关系到生成式模型的可控性与社会信任。如何在释放生成技术潜能的同时建立完善的检测、防护与合规机制，是推动生成式人工智能健康发展的核心问题。本报告后续章节将系统分析该领域的发展态势与技术趋势，重点探讨基于大模型的视频图像异常检测、违规内容识别、跨模态理解、弱监督安全学习及深度伪造检测的最新进展与挑战，为未来研究与工程实践提供参考。

1 国内外研究现状

1.1 图像与视频理解安全

异常检测（anomaly detection，AD）旨在发现偏离标准、正常或预期情况的样本或事件。由于异常情况通常相对稀少，但一旦发生可能带来负面影响，因此AD在金融欺诈检测、网络入侵、工业缺陷检测和视频监控等领域具有广泛的应用潜力。传统的图像与视频异常检测方法在面对复杂、动态、多变的自然场景时，往往面临检测迟滞和误报率高的挑战。然而，随着深度学习时代的到来，尤其是大型预训练模型，如大视觉-语言模型（Sun等，2019，Team等，2024，Yang等，2025b，Yang等，2023，Yu等，2022b，Yuan等，2021）和扩散模型（diffusion model）（Blattmann等，2023a，Blattmann等，2023b，Geng等，2024，Podell等，2023）的兴起，该领域的研究取得了显著进步。以下按照所使用的监督信号和所使用的大模型，对图像、视频异常检测的方法进行分类介绍。

1.1.1 基于监督信号分类的研究进展

根据训练过程中使用的监督信号的多少和性质，现有研究大体可分为全监督、半监督、弱监督和无监督四类。如图2所示。

1）全监督异常检测

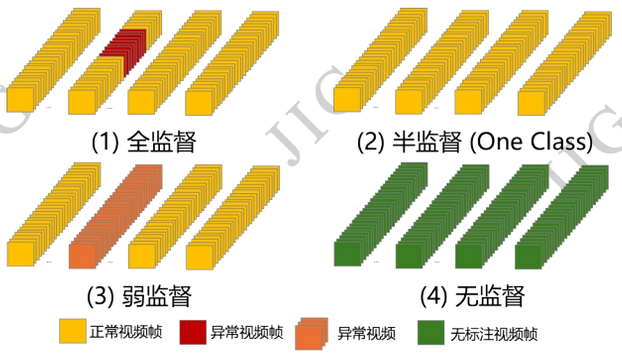

图 2　视频异常检测基于监督信号分类

Fig. 2　Video anomaly detection based on supervised signal classification

全监督异常检测(Li 等，2023b，Sultani 等，2018，Zhu 等，2020)假设训练样本具有完整且精确的监督信号，这意味着每个异常样本(例如视频帧或事件)都附带精确的真值标签。这种方法在拥有充分监督信息的情况下，算法的检测性能会非常出色。然而，由于现实中异常行为的稀缺性和人工标注的密集要求较高，目前针对全监督视频异常行为检测(video anomaly detection，VAD)任务的研究相对较少。典型的方法通常利用对比学习的思路，例如 Yao 等人(2023)提出了基于显式边界引导的半推拉对比学习，以明确地利用所有可用的监督信息来指导特征学习，从而增强异常样本与正常样本之间的分离度。Luo 等人(2025)认为传统的异常检测方法通常依赖于将测试图像与外部训练集中的正常图像进行对比，但由于图像变化和对齐困难，检测效果有限。为了解决这一问题，提出了 INP-Former，通过从测试图像中直接提取正常原型来避免外部依赖，结合正常原型一致性损失和软挖掘损失，有效提升了异常检测的精度，并在多个数据集上实现了最先进的性能，甚至具备零样本泛化能力。另外，异常检测的挑战在于异常类别的分布通常高度不平衡，且可能存在开放集(open-set)问题，即测试时出现训练集中未见过的异常类型。开放集全监督异常检测(Ding 等，2022，Peng 等，2025，Wang 等，2025b，Zhu 等，2024)旨在训练模型来处理包含已知正常样本、已知异常样本，以及在测试时出现未知异常(开放集)的情况。现有方法主要聚焦于对未见异常进行模拟定义和为正常样本学习一个紧凑的空间。DRA(Ding 等，2022)定义了三种异常类型：与有限可见异常类似的异常、与数据增强或外部数据源产生的伪异常类似的异常，以及在某些基于潜在残差的复合特征空间中可检测到的不可见异常。进一步设计了一个多头网络，其中每个头分别负责学习这三种解耦异常。这样，模型学习了多样化的异常表示，而不仅仅是已知的异常，从而可以从正常数据中区分出可见和不可见的异常。AHL(Zhu 等，2024)引入了一种名为异常异质性学习方法，该方法模拟一组多样化的异质性异常分布，然后利用它们在替代开放集环境中学习统一的异质性异常模型。DPDL(Wang 等，2025b)旨在将正常样本封闭在一个紧凑且具有判别性的分布空间中，构建多个可学习的高斯原型，为丰富多样的正常样本创建一个潜在表示空间，并学习一个薛定谔桥，以促进正常样本向这些原型的扩散过渡，同时避开异常样本。近期一些研究(Acsintoae 等，2022，Li 等，2025b，Zhu 等，2022)开始探索视频开放集任务，利用对比学习进行边缘学习，从而获得紧凑的正态分布或者通过异常分布量化不确定性来识别未知异常。

2)半监督异常检测

半监督异常检测(Dong 等，2024，Huang 等，2020，Ruff 等，2019，Tian 等，2023，Yao 等，2023，Zhou 等，2023)假设训练阶段只有正常样本可用，旨在学习正常的模式或范式，并将偏离该学习模式的测试样本视为异常。

方法论方面，半监督 VAD 主要依赖以下几种学习正常范式的方法：

(1) 自监督学习(self-supervised learning)：利用辅助任务(代理任务)从无监督数据中获取监督信号。主流代理任务包括：

① 重建(reconstruction)(Li 等，2024b，Liang 等，2023，Luo 等，2017，Nguyen 和 Meunier，2019，Qi 等，2023，Shan-wu 等，2023，Wang 等，2023a，Zhaobo 等，2025)：将正常数据输入编码-解码网络，激励网络生成与原始输入数据紧密匹配的重建数据。基于重建的方法是半监督 VAD 领域中最常用的。

② 预测(prediction)(Cai 等，2021，Fangyuan 和 Genlin，2024，Hao 等，2022，Jia-xu 等，2022，Liu 等，2022a，Liu 等，2024d，Liu 等，2021b，Shao-nian 等，2023，Yu 等，2022a，Zhou 等，2022a)：利用视频固有的时间相关性，预测下一时刻的数据，基于“正常事件是可预测的，而异常事件是不可预测的”

表1　全监督异常检测方法总结

Table 1　Summary of Fully Supervised Anomaly Detection Methods

核心范式	典型方法	核心机制
对比与原型学习	Yao等人(2023)	利用所有监督信息,通过推拉机制显式分离正常与异常特征边界。
	Luo等人(2025)	从测试图像自身提取正常原型,避免对外部数据集的依赖,提升泛化性。
开放集全监督	Ding等人(2022)	定义并解耦多种异常类型(如可见、伪造、不可见),学习多样化异常表示。
	Zhu等人(2024)	模拟多样化的异质性异常分布,训练统一模型以适应未知异常。
	Wang等人(2025b)	构建紧凑的正常样本原型空间,利用扩散模型或薛定谔桥引导正常样本聚合。
	(Acsintoae等, 2022, Li等, 2025b, Zhu等, 2022)	针对视频数据,通过对比学习界定分布边缘或量化异常分布的不确定性。

这一假设。

③　去噪(denoising)(Barker等, 2023, Gao等, 2025a, Kascenas等, 2023, Li等, 2024a, Wang等, 2025a):通过向输入数据添加噪声并激励网络实现去噪,增强网络对VAD的鲁棒性。

④　对比学习(contrastive learning)(Gao等, 2025a, Jezequel等, 2022, Liu等, 2022b, Wang等, 2022a, Wang等, 2020):通过区分相似和不相似的样本对来学习有用的表征,为正常样本学习可靠的原型。

(2) 单类学习(one-class learning):专注于来自正常类别的样本,不需要设计可行的代理任务。

①　单类分类器(one-class classifiers)(Kim等, 2023, Sharma等, 2022, Wang和Cherian, 2019):假设正常样本位于一个有界集合中,通过优化寻求包含所有正常样本的最小半径超球心。

②　高斯分类器(gaussian classifiers)(Fan等, 2020, Sabokrou等, 2018):假设数据遵循高斯分布,通过训练样本学习高斯分布(均值和方差),与平均值有显著偏差的样本视为异常。如ESAD(Huang等, 2020)探索了另一种半监督异常检测的优化目标,即最大化正态分布和异常类之间的KL散度。考虑到估计异常分布的难度导致直接优化不可行,放宽了基于KL散度的目标函数,并将其进一步分解为两个因子:(i)数据与潜在表示之间的互信息;(ii)潜在表示的熵。

③　对抗学习(Zaheer等, 2020, Zaheer等, 2022a):利用生成器G和鉴别器D之间的对抗训练来学习正态样本的分布。由于正态样本对G来说是可访问的,因此G能够感知正态数据分布。因此,D会明确地决定G的输出是否遵循正态分布。如GANomaly(Akcay等, 2018)是一种通过对抗训练实现的半监督异常检测模型,利用生成对抗网络的框架,通过学习正常数据的分布,并在训练中引入对抗机制,以提升对异常样本的识别能力。

3)弱监督异常检测

弱监督异常检测在图像与视频领域的定义存在显著差异,核心均为通过有限监督信号降低标注成本,其技术路线围绕“特征优化”与“监督信号高效利用”展开,具体进展如下:

(1)弱监督图像异常检测

弱监督图像异常检测的核心假设为:训练数据包含少量标记异常样本与大量未标记样本(Durani等, Zhang等, 2019b, Zhao等, 2024),主流方法通过优化特征分布的“正常-异常分离度”提升检测性能:

偏差损失优化类:DevNet(Pang等, 2019)在潜在空间引入高斯先验与偏差损失,将正常样本向分布中心聚合、异常样本向边缘推离,通过强化相对偏差实现高效区分;FeaWAD(Zhou等, 2022b)进一步融合自编码器与偏差损失,增强特征表示的鲁棒性。

元学习与核机制类:PReNet(Pang等, 2023)从元学习视角出发,结合成对关系与自监督距离度量,强化异常样本的隔离效果;WSAD-DT(Durani等)提

表2　半监督异常检测方法总结

Table 2　Summary of Semi-Supervised Anomaly Detection Methods

核心范式	典型方法	核心机制
自监督学习	(Luo等, 2017, Nguyen和Meunier, 2019, Shan-wu等, 2023, Zhaobo等, 2025)	训练网络重建正常输入，异常样本无法被准确重建故产生高误差。
	(Fangyuan和Genlin, 2024, Liu等, 2024d, Liu等, 2021b, Shao-nian等, 2023, Zhou等, 2022a)	利用时序相关性预测下一帧，正常事件可预测，异常事件不可预测。
	(Barker等, 2023, Gao等, 2025a, Li等, 2024a, Wang等, 2025a)	向输入添加噪声并训练网络去噪，增强对正常模式的鲁棒性。
	(Gao等, 2025a, Jezequel等, 2022, Liu等, 2022b, Wang等, 2022a, Wang等, 2020)	区分相似/不相似样本对，为正常样本学习紧凑的特征表示。
单类学习	(Kim等, 2023, Sharma等, 2022, Wang和Cherian, 2019)	寻找包含所有正常样本的最小半径超球心，异常样本位于球外。
	(Fan等, 2020, Sabokrou等, 2018)	假设正常数据服从高斯分布，学习均值和方差，偏离者为异常。
对抗学习	(Zaheer等, 2020, Zaheer等, 2022a)	生成器学习正常分布，判别器辅助判断是否符合正常分布。

出双尾核机制，通过轻尾核保障类内样本表示紧凑性（相似度随距离快速衰减），重尾核调节类间离散性（为离群点预留更宽边界），实现紧凑性与类外裕度的平衡。

（2）弱监督视频异常检测

弱监督视频异常检测的核心假设为：训练阶段仅提供视频级粗标签（正常/异常），异常事件的时空精确位置无标注（Deyun等，2024，Karim等，2024，Li等，2023a，Si-qian等，2022，Wan等，2020，Wen-hao等，2024，Wu等，2024a，Zhang等，2019b），相比全监督显著降低标注成本，主流技术路线分为两类：

①多示例学习（multiple instance learning，MIL）范式

作为视频级标签处理的经典框架，其核心逻辑为：将视频视为“包”，帧/片段视为“实例”，监督信号仅作用于视频级二值标签（正常视频为负包，所有实例均符合正常模式；异常视频为正包，至少含一个异常实例），通过包间特征分布差异间接学习实例判别性特征，规避帧级/片段级精确标注依赖。

内袋分数差距正则化（Zhang等，2019b）：在MIL基础上提出内包损失函数（inner bag loss，IBL），通过约束包内实例得分差异优化判别能力——正包要求最高异常得分与最低异常得分差距最大化（强化包内异常与正常实例区分），负包要求该差距最小化（保证正常实例分数分布一致性，抑制噪声导致的得分波动）。

动态MIL损失（LDMIL）（Wan等，2020）：改进传统MIL的固定实例选择策略，采用k-max选择方法，从每个视频中选取得分最高的k个片段作为关键实例，其中k值根据视频时长动态调整，提升对不同长度视频异常片段的捕捉灵活性，避免固定k值导致的异常遗漏或误选。

②多模态学习范式

依托多模态预训练模型的跨模态表征能力与知识迁移优势，增强异常事件细粒度理解，主要分为两类技术思路：

视觉-文本多模态融合：基于CLIP（Radford等，2021）等大规模图像-文本预训练模型的强大视觉表征能力（性能优于单一视觉模态方案），结合大语言模型（large language models，LLM）（Jiang等，2023，OpenAi，2023，Touvron等，2023）发展而来的多模态大语言模型（multimodal large language models，MLLM，如LLaVA系列（Liu等，2024a，Liu等，2023）、Qwen-VL（Bai等，2025，Wang等，2024）、InternVL（Chen等，2023b））的跨模态理解能力，构建弱监督检测框架。典型方法包括：CLIP-TSA（Joo等，2023）以CLIP为视频特征提取器，设计时序注意力模块弥补其特征上下文相关性不足的缺陷；VadCLIP（Wu等，2024b）采用双分支结构，通过

CLIP提取视频帧特征与异常类别文本特征(如“Fighting”),一支利用视觉特征完成粗粒度二分类,另一支通过片段级视觉-文本相似度计算实现细粒度分类与监督,达成多模态对齐,提升异常类别区分能力。

视觉-音频跨模态融合:姜迪等人(2025)提出跨模态双曲图注意力网络,在弱监督框架下融合视频RGB特征(I3D提取)与音频特征(VGGish提取),引入双曲图注意力机制,借助双曲空间的层级结构建模局部-全局依赖关系,提供了跨模态表征的新思路。

表3　弱监督异常检测方法总结

Table 3　Summary of Weakly Supervised Anomaly Detection Methods

任务子类	核心范式	典型方法	核心机制
图像异常检测	偏差损失优化	(Pang等, 2019) (Zhou等, 2022b)	引入高斯先验与偏差损失,聚合正常样本,推离异常样本。
	元学习与核机制	(Pang等, 2023) (Durani等)	利用元学习强化隔离效果,或利用双尾核机制平衡紧凑性与类间离散性。
视频异常检测	多示例学习	(Zhang等, 2019b) (Wan等, 2020)	将视频视为“包”,帧为“实例”。通过正负包差异学习实例特征。
	多模态学习	(Joo等, 2023) (Wu等, 2024b)	利用CLIP等图文模型,结合时序注意力或双分支结构对齐视觉与语义。
		姜迪等人(2025)	融合RGB与音频特征,利用图注意力机制建模跨模态依赖。

4)无监督异常检测

无监督异常检测的核心特征为无需任何标注信息,训练集与测试样本集一致,其核心优势在于彻底规避繁重标记负担,支持无人干预下的持续重训练;但受限于缺乏监督信号,存在检测性能相对较弱、误报率与漏报率较高的固有挑战。主流技术路线围绕“伪标签生成”与“变化模式捕捉”两大核心逻辑展开,具体分类及进展如下:

(1)伪标签范式

核心思路为:通过无监督方式从数据中自动生成“正常/异常”伪标签,将无监督问题转化为弱监督学习,以优化模型判别能力,典型方法聚焦伪标签生成的“准确性”与“迭代优化”:

①　两阶段伪标签训练(Wang等, 2018):先通过自适应重建损失阈值训练自编码器,从未标记视频中估计正常事件并作为初始伪标签;再基于该伪标签优化正态性模型,排除异常干扰以提升检测性能。

②　自训练迭代优化(Pang等, 2020):采用“初始检测-伪标签生成-迭代训练”流程,先通过经典单类算法生成异常帧与正常帧的初始伪标签,再利用自训练策略迭代优化端到端异常分数学习器。

③　生成式合作学习(Zaheer等, 2022b):利用异常的低频特性,在生成器与鉴别器间构建交叉监督机制,双方从彼此输出的伪标签中相互学习,强化伪标签可靠性。

④　由粗到细分层生成(Al-Lahham等, 2024):通过分层分裂聚类生成视频级粗伪标签,结合统计假设检验生成片段级细伪标签,基于两级伪标签联合训练异常检测器,提升标签粒度与准确性。

(2)变化检测范式

核心逻辑为:基于“异常事件具有罕见性与非常规性”的先验假设,通过量化数据间的偏差程度捕捉异常变化——异常样本因不符合正常数据的分布规律,会导致模型对其预测误差显著高于正常样本。典型技术以自编码器为核心载体,利用异常样本的重建误差或预测偏差作为异常判定依据,通过度量该偏差的阈值化实现异常检测。

1.1.2　基于大模型的进阶研究进展

在基于监督信号的传统分类之外,学术界和工业界也广泛研究了利用大模型能力实现快速适应和可解释性的新型训练范式,涵盖了零样本与小样本、开放词汇以及可解释性等异常检测方法,如图3所示。

表4　无监督异常检测方法总结

Table 4　Summary of Unsupervised Anomaly Detection Methods

核心范式	典型方法	核心机制
伪标签范式	(Wang等, 2018)	先估计初始伪标签，再基于伪标签优化模型，排除异常干扰。
	(Pang等, 2020)	“检测-生成伪标签-训练”循环，迭代优化异常分数学习器。
	(Zaheer等, 2022b)	利用异常低频特性，生成器与判别器交叉监督互学伪标签。
	AI-Lahham等人(2024)	结合聚类与统计检验，生成视频级（粗）与片段级（细）双层伪标签。
变化检测范式	基础自编码器方法	假设异常具有稀缺性与不规则性，利用自编码器计算重建误差或预测偏差，通过阈值判定异常。

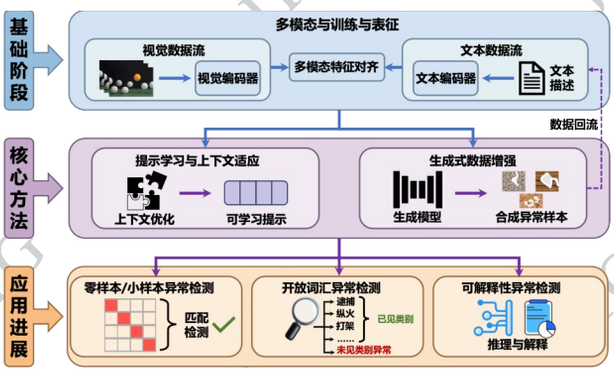

图3　基于大模型的异常检测研究

Fig. 3　Anomaly detection research based on a large model

1）　零样本与小样本异常检测

由于许多实际应用中获取异常数据标签非常困难或成本高昂，零样本(zero-shot)和少样本(few-shot)异常检测应运而生。这些方法通常利用大规模视觉-语言预训练模型(如CLIP)所蕴含的丰富知识。核心思想是通过将图像特征与文本提示(text prompts)进行匹配，利用预训练模型的泛化能力来识别未见过的异常，解决在部署过程中快速适应新场景和新类别的需求。

(1) 零样本异常检测：通过利用预训练的视觉和语言模型(如CLIP)的零样本能力，将图像与文本提示进行匹配，识别未见类别中的异常。挑战在于CLIP的粗粒度图像-文本对齐限制了对细粒度异常的定位和检测性能。研究致力于克服这些限制，例如：

①　AdaCLIP(Cao等，2024)：利用混合可学习提示(hybrid learnable prompts)对CLIP进行适配，以用于零样本异常检测。

②　AnomalyCLIP(Zhou等，2024)：采用与物体无关的提示学习策略进行零样本异常检测。

③　Crane(Salehi等，2025)：提出上下文引导的提示学习和注意力细化，通过将文本编码器的可学习提示以图像上下文为条件，并进行局部到全局的表示融合，以增强对细粒度异常的感知。Crane还可以整合DINOv2(Oquab等，2024)等视觉基础模型来增强空间理解。

④　FiLo(Gu等，2024)：通过细粒度描述和高质量定位实现零样本异常检测。

⑤　GenCLIP(Kim等，2025)：通过多层提示(multi-layer prompting)双分支推理(dual-branch inference)自适应文本提示过滤机制来去除不相关或非典型的类别名称。

⑥　SSMOD-Net(Zhenhua等，2025)：使用一个由状态空间模型驱动的提示编码器，该编码器能够根据输入图像动态生成提示，引导SAM进行分割，而无需对SAM本身进行微调。

(2) 小样本异常检测：小样本异常检测的目标是仅使用少量异常样本来检测先前未见场景中的异常。

①　PromptAD(Li等，2024b)：通过仅使用正常样本学习提示，实现小样本异常检测。

②　CLIP-FSAC++(Zuo等，2024)：针对少正常样本(few-normal shot)异常分类，引入了异常描述符(anomaly descriptor)，通过图像到文本和文本到图像的跨模态注意力，增强视觉和文本嵌入的相关性，并调整CLIP的表示。

③　AnomalyDiffusion(Kong等，2024)/DualAnoDiff(Jin等，2024)：利用扩散模型进行少样本异常图像生成。AnomalyDiffusion利用从大规模数据集中学习到的潜在扩散模型的强先验信息来增强小样本训练数据下的生成真实性，提出了空间异常嵌入，

它由可学习的异常嵌入和从异常掩码编码的空间嵌入组成，将异常信息分解为异常外观和位置信息。此外，基于生成的异常图像与正常样本之间的差异，该模型动态地引导模型将注意力更多地集中在异常生成不太明显的区域，从而生成精确匹配的异常图像-掩模对。

表5　零样本与小样本异常检测方法总结

Table 5　Summary of Zero-Sample and Small-Sample Anomaly Detection Methods

核心范式	典型方法	核心机制
零样本检测	(Cao等, 2024) (Zhou等, 2024)	利用混合可学习提示或物体无关策略适配CLIP，实现未知类别检测。
	(Salehi等, 2025) (Gu等, 2024)	引入上下文引导或细粒度描述，增强对微小异常的定位与感知。
	(Kim等, 2025) (Zhenhua等, 2025)	通过自适应提示过滤去除无关类别，或利用SSM驱动提示引导SAM分割。
小样本检测	Li等人(2024b) Zuo等人(2024)	仅利用正常样本学习提示，或结合异常描述符与跨模态注意力增强特征相关性。
	Kong等人(2024) Jin等人(2024)	利用扩散模型生成逼真的少样本异常图像，通过空间嵌入与注意力引导提升生成质量。

2）开放词汇异常检测

开放词汇异常检测要求模型能够识别训练中可能从未见过的、通常通过自然语言描述的异常类型。这项任务以视觉-语言模型（VLM）为中心，通过将视频/图像与相应的文本标签进行匹配来实现。在图像领域，开放词汇目标检测通过视觉和语言知识蒸馏实现。利用生成式模型生成伪异常样本进行辅助训练也是实现开放词汇异常检测的一种途径。

典型方法包括：Anomize（Li等，2025a）探索了来自基于LLM的多个来源的补充信息，通过利用多层次的视觉数据和匹配的文本信息来减轻检测模糊性。此外，其建议结合标签关系来指导新标签的编码，从而提高新视频与其对应标签之间的一致性，从而有助于减少分类混淆。LaGoVAD（Liu等，2025）利用两种正则化策略动态调整异常定义：通过动态视频合成来多样化异常的相对持续时间，以及通过对比学习和负样本挖掘来增强特征的稳健性。

表6　开放词汇异常检测方法总结

Table 6　Summary of methods for detecting anomalies in open vocabulary

核心范式	典型方法	核心机制
大模型辅助增强	Li等人(Li等, 2025a)	利用LLM获取多层次视觉文本信息，结合标签关系编码减少分类混淆。
	Liu等人(Liu等, 2025)	通过动态视频合成多样化异常持续时间，结合对比学习与负样本挖掘增强稳健性。

3）可解释性异常检测

随着异常检测算法在医疗诊断、金融欺诈和自动驾驶等高风险决策领域的广泛应用，提供对决策的解释已成为道德和监管要求。可解释异常检测（explainable anomaly detection，XAD）（Li等，2023b）旨在从异常检测模型中提取相关知识，以提供对异常事件的洞察和理解。Hinami等人（2017）利用多任务检测器作为通用模型，学习关于视觉概念（例如实体、动作和属性）的通用知识，以人类可理解的形式描述事件，然后设计了一个特定于环境的模型作为异常检测器，用于异常事件的重述和检测。同样，Reiss等人（2022）提取了基于属性的显式表示，如速度和姿态以及隐式语义表示，以做出可解释的异常决策。Yang等人（2024a）提出了第一个基于规则的半监督异常视频检测推理框架，该框架利用大型语言模型（LLM）实现了的异常事件推理能力。Holmes-VAU（Zhang等，2025a）为了实现无偏差且可解释的VAD系统，构建了第一个大规模多模态VAD指令调优基准数据集，即HIVAU-70k。该数据集采用精心设计的半自动标记范式创建。随后，高

效的单帧标注被应用于收集到的未修剪视频，然后使用强大的现成视频字幕生成器和大型语言模型(LLM)将其合成为对异常和正常视频片段的高质量分析。

表7　可解释性异常检测方法总结

Table 7　Summary of Explainable Anomaly Detection Methods

核心范式	典型方法	核心机制
概念与属性学习	Hinami等人(2017) Reiss等人(2022)	学习通用的视觉概念(实体、动作)或属性(速度、姿态)，以人类可理解形式描述异常。
规则推理	Yang等人(2024a)	构建基于规则的半监督推理框架，利用LLM的逻辑能力进行异常事件判定。
多模态大模型	Zhang等人(2025a)	基于大规模指令调优数据集，利用VLM生成高质量的异常分析与视频字幕。

1.2　图像与视频生成技术

1.2.1　生成模型的理论基础

生成的基本任务，在于学习数据分布并从中采样，其内容主要为重建、合成、压缩、推断与下游感知任务的表示迁移。早期的方法如自回归模型(AR)、变分自编码(VAE)将联合分布分解为条件链并以最大似然逐项学习，或通过潜变量与证据下界(ELBO)在“重建—正则”之间权衡，从而获得可解释的密度与稳定训练，但采样往往受序列化或解码器能力所限，表现为生成速度较慢、细节锐度有限。又有如生成对抗网络(GAN)、能量基模型(EBM)，它们分别以判别—生成的极小极大博弈或未归一化能量函数直接匹配数据几何，强调感知质量与支持集对齐，因而能合成高保真样本，但训练动力学敏感、易出现模式崩塌或混合慢、似然评估不便。

而近年来表现突出的扩散方法，则以“跨噪声尺度的去噪—分数学习”为训练准则，在实践中优化变分下界与去噪分数匹配，因而既贴近最大似然范式又保持强表达力与数值稳健性，并可通过采样引导实现高质量的可控生成。扩散模型的核心思想是将“生成”建模为“去噪”的逆过程，其方法并不直接显式构造数据分布，而是定义一条由数据分布逐步“热化”到各向同性高斯的正向随机过程。训练的目标是学习与之相反的“逆过程”——一个将纯噪声逐步还原为数据的去噪动力学。其本质在于将高维生成问题归结为跨噪声尺度的梯度场学习问题；用随机热化保证训练稳定与覆盖性，用去噪分数匹配保证统计一致性；再以逆向积分将“噪声”搬运回“数据”。这一范式兼具优化可解性、统计健壮性与在连续时间、可条件化、可控引导上的提升。

最基础的扩散方法的过程可以做如下描述：学习一条从各向同性高斯先验逐步“还原”到真实数据分布的反向轨迹。为此，模型首先设定一条正向噪声注入链，使观测样本经由马尔可夫过程逐步被高斯噪声取代，并最终在步数 T 处近似成为标准正态样本。假设真实样本 $\sim q(x)$，则其随时间演化的分布为：

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = N\left(\boldsymbol{x}_t;\sqrt{1-\beta_t}\,\boldsymbol{x}_{t-1},\beta_t\mathbf{I}\right),\in(0,1),\#(1)$$

式中，$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ 表示在给定第 $t-1$ 步样本条件下第 t 步样本的条件概率分布；N(·；μ, Σ) 表示均值为 $\boldsymbol{\mu}$、协方差为 $\boldsymbol{\Sigma}$ 的多元高斯分布；表示扩散过程第 t 步的样本；$\sqrt{1-\beta_t}$ 为该条件分布的均值项；$\beta_t\mathbf{I}$ 为该条件分布的协方差项；表示第 t 步加入噪声的方差系数(噪声调度参数)，且$\in(0,1)$；$\mathbf{I}$ 表示与样本维度一致的单位矩阵；t 表示时间步索引。

根据高斯分布的可积性，可以得到从初始样本 x_0 到任意时刻 t 的闭式表达：

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathrm{N}(\boldsymbol{x}_t;\sqrt{\bar{\alpha}_t}\,\boldsymbol{x}_0,\left(1-\bar{\alpha}_t\right)\mathbf{I},\#(2)$$

式中，$q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ 表示由初始样本演化至第 t 步样本的边缘条件分布；表示真实数据分布 $q(\boldsymbol{x})$ 中采样得到的初始样本；表示累计信号保留系数，通常定义为 $=\prod_{s=1}^{t}\alpha_s$，其中 $=1-\beta_s$；$\sqrt{\bar{\alpha}_t}$ 为该分布的均值项；$(1-\bar{\alpha}_t)\mathbf{I}$ 为该分布的协方差项；其余符号同式(1)。

这意味着样本在第 t 步时等价于：

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\,\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}\sim\mathrm{N}(0,\mathbf{I}),\#(3)$$

式中，$\boldsymbol{\varepsilon}$ 表示独立同分布的高斯噪声变量；$\boldsymbol{\varepsilon}\sim\mathrm{N}(0,\mathbf{I})$ 表示 $\boldsymbol{\varepsilon}$ 服从均值为0、协方差为单位矩

阵 $\mathbf{I}$ 的标准多元正态分布；$\sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\varepsilon}$ 表示第 t 步累计噪声项；其余符号同式(2)。

正向扩散就像是一个“逐帧模糊”的过程，随着时间推移，图像信息逐渐被噪声覆盖，直至完全随机；而后续的生成任务就是要一步步“去模糊”并重建原始信息。反向过程同样采用高斯条件分布，其均值与协方差未知、需由神经网络学习。实践中通常把协方差取为解析可计算的固定形式，Ho、Abbeel等人提出让网络直接回归被注入的噪声而非均值本身：随机选取时间步、用闭式公式合成带噪样本，然后让网络把“掺进去的噪声”预测回来。这一“噪声回归损失”结构简单、实现方便，在大规模训练中表现稳健。

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)=\mathrm{N}\left(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t),\sum q(t)\mathbf{I}\right),\#(4)$$

式中，$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ 表示参数为 $\boldsymbol{\theta}$ 的模型所定义的反向扩散条件概率分布；$\boldsymbol{\theta}$ 表示神经网络的可学习参数；$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t)$ 表示由网络预测得到的反向过程均值函数；表示与时间步 t 相关的方差项(通常取为可解析计算的预设形式)；其余符号同式(1)。

在采样阶段，模型从标准高斯 x_t 起步，按时间反向迭代：利用网络对当前带噪样本的噪声估计恢复出一步的去噪均值，再加上少量受控的随机性，逐步得到后续的预测步。直观上，这像是“温度逐步降低、图像逐步显形”的退火过程。方差的形状会影响训练难度和生成质量：在高噪声区保持小步长有利于稳定学习，在低噪声区适当增大步长有助于细节还原与效率平衡。

$$L_s=\mathbb{E}_{t,x_0,\varepsilon}\left[\ /\!/\ \boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t,t)\ /\!/\ _2^2\right],\#(5)$$

式中，表示噪声回归训练损失函数；$\mathbb{E}_{t,x_0,\varepsilon}[\cdot]$ 表示关于时间步 t、真实样本以及噪声 $\boldsymbol{\varepsilon}$ 的期望；$\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t,t)$ 表示由参数为 θ 的网络对噪声的预测值；$/\!/\cdot$ 表示二范数；其余符号同式(3)。

1.2.2 图像生成方法的研究现状

图像生成技术在过去的十年里经历了显著的变革，从最初的生成对抗网络(GAN)到如今的扩散模型(diffusion models)，每一个阶段都推动了人工智能在图像合成领域的进步。本文将从两个主要发展阶段进行详细探讨：GAN时代(2014－2021)与扩散模型崛起(2020年至今)，介绍各阶段的核心思想、代表性模型、贡献与不足，以及相关的应用领域。

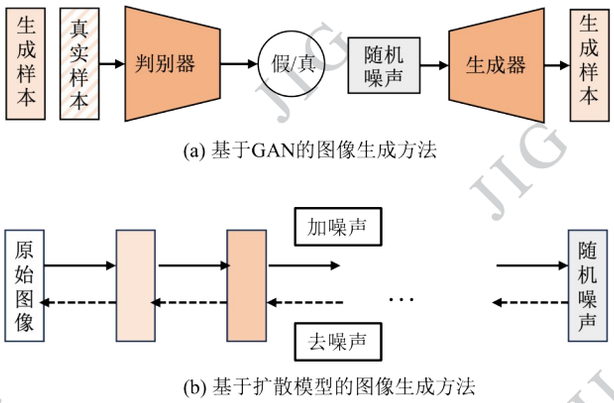

图4 图像生成方法

Fig. 4 Image generation method((a)GAN based image generation method; (b) Diffusion models based image generation method)

1) 基于GAN的图像生成方法

生成对抗网络(GANs)是由加拿大蒙特利尔大学计算机与运筹学系(Department of Computer Science and Operations Research, Université de Montréal) Ian Goodfellow 等人在2014年提出的(Goodfellow等, 2014)。该方法通过两个网络——生成器(generator, G)和判别器(discriminator, D)的对抗博弈来训练模型，从而生成与真实数据分布相似的样本。生成器负责生成假样本，而判别器则试图区分生成的假样本和真实数据。二者的博弈过程可以通过优化算法最终达到纳什均衡，生成器能够输出高质量的假数据。GAN的核心思想来源于博弈论中的纳什均衡。在GAN的框架下，生成器和判别器在训练过程中互相竞争，生成器努力生成越来越真实的图像，而判别器则不断提高区分真实图像与生成图像的能力。生成器和判别器的对抗博弈通过优化目标函数来实现。生成器的目标是使判别器无法分辨生成的图像，判别器的目标是尽可能准确地区分真实图像与生成图像。

其中较为经典的方法包括Vanilla GAN。该模型是由加拿大蒙特利尔大学计算机与运筹学系(Department of Computer Science and Operations Research, Université de Montréal)的Goodfellow等人于2014年提出的。尽管其结构相对简单，但训练过程不稳定，易出现模式崩溃(mode collapse)现象，即生成器仅能生成少量类型的图像，难以全面捕捉真实数据的分布特征。DCGAN(deep convolutional GAN)由Alec Radford等人于2015年提出，采用卷积

神经网络(CNN)架构,显著提升了图像生成的质量,解决了Vanilla GAN中图像模糊的问题(Radford等,2015)。WGAN(Wasserstein GAN)由Martin Arjovsky等人于2017年提出,采用Wasserstein距离(也称地球搬运距离)来衡量生成数据与真实数据的差异,能够缓解GAN的训练不稳定性(Arjovsky等, 2017)。

条件生成对抗网络(conditional GAN, CGAN)由多伦多大学(University of Toronto)的Mehdi Mirza等人提出(2014)。该方法在原始GAN框架中引入额外的输入条件(如类别标签或文本描述),从而实现对生成图像特征的显式控制,显著提升了生成任务的可控性。其后,研究者提出了多种变体:InfoGAN通过最大化互信息量来增强生成样本的多样性(Chen等, 2016);ACGAN(auxiliary classifier GAN)在CGAN基础上进一步利用判别器输出的辅助分类信息,提升了生成图像的质量与判别一致性(Odena等, 2017)。StyleGAN系列由美国NVIDIA研究院(NVIDIA Research)的Tero Karras等人提出(2017, 2019)。该系列模型通过引入新的网络架构与风格映射机制,使生成图像在细节表现和分辨率上均取得突破。其前身ProGAN(progressive growing GAN)采用逐层增长的训练策略,成功实现了高分辨率图像生成;随后发布的StyleGAN v1/v2/v3在特征分离与风格调控方面持续改进,使得生成结果在真实感与多样性上显著优于以往方法,特别在人脸生成任务中表现突出。

BigGAN由英国DeepMind公司(DeepMind Technologies Ltd.)的Andrew Brock等人提出(2018),通过扩大网络规模与类别数量,实现了超高分辨率的图像生成,在ImageNet上取得了当时最优的生成质量。与此同时,SAGAN(self-attention GAN)由Google Brain团队的Han Zhang等人提出(2019a),首次将自注意力机制(self-attention)引入GAN架构,有效提升了生成图像在全局结构与局部细节方面的一致性,推动了基于注意力的生成模型发展。

GAN在图像生成领域的最大贡献是极大提升了图像的清晰度和多样性,使得生成图像的质量几乎可以与真实图像媲美。然而,GAN在训练过程中仍然面临一些问题,如训练不稳定、模式崩溃、生成过程难以控制等。这些问题限制了GAN在实际应用中的普及,尤其是在需要高精度控制生成内容的场景下。

表8　基于GAN的图像生成方法总结

Table 8　Summary of GAN-based image generation methods

核心范式	典型方法	核心机制
典型生成对抗模型	Vanilla GAN(2014)	通过两个网络——生成器和判别器的对抗博弈来训练模型,从而生成与真实数据分布相似的样本。
	DCGAN(Radford等, 2015)	采用卷积神经网络架构。
	WGAN(Arjovsky等, 2017)	采用Wasserstein距离来衡量生成数据与真实数据的差异。
条件生成对抗模型	CGAN(Mirza和Osindero, 2014)	在原始GAN框架中引入额外的输入条件,从而实现对生成图像特征的显式控制,显著提升了生成任务的可控性
	InfoGAN(Chen等, 2016)	通过最大化互信息量来增强生成样本的多样性。
	ACGAN(Odena等, 2017)	在CGAN基础上进一步利用判别器输出的辅助分类信息,提升了生成图像的质量与判别一致性。
	StyleGAN系列(Karras等, 2017, Karras等, 2019)	通过引入新的网络架构与风格映射机制,使生成图像在细节表现和分辨率上均取得突破。
大规模学习	BigGAN (Brock等, 2018)	通过扩大网络规模与类别数量,实现了超高分辨率的图像生成。
自注意力机制	SAGAN(Zhang等, 2019a)	将自注意力机制引入GAN架构,有效提升了生成图像在全局结构与局部细节方面的一致性。

2） 基于扩散模型的图像生成

随着生成对抗网络（GAN）技术的不断发展，其在训练稳定性和可控性方面的局限性逐渐显现。为此，学术界在2020年迎来了新的研究范式——扩散模型（diffusion models）的崛起。该模型由加州大学伯克利分校（University of California，Berkeley）的Jonathan Ho等人首次提出（2020），并迅速成为图像生成领域的核心研究方向。

扩散模型的核心思想源于物理学中的扩散与反扩散过程：模型通过在图像上逐步添加噪声，使其逐渐退化为纯随机噪声（前向过程）；随后再学习如何在反向过程中逐步去除噪声，从而恢复出高质量的原始图像。其训练机制通常建立在马尔可夫链（markov chain）框架上，通过联合建模加噪与去噪两个过程，使模型能够精准地学习数据分布的生成规律。

与传统GAN相比，扩散模型在多个方面展现出显著优势。首先，生成过程高度稳定，有效缓解了GAN常见的模式崩溃（mode collapse）问题。其次，扩散模型具有良好的训练可控性，可通过调整反向去噪过程中的步长与超参数灵活控制生成质量与风格，尤其适用于对内容精度要求较高的应用场景。此外，扩散模型在细节还原与全局一致性方面表现卓越，生成图像在视觉质量上显著超越多数GAN模型。凭借其稳定性、可控性与高保真度，扩散模型迅速成为生成式人工智能的主流技术路线，为后续的潜空间扩散（latent diffusion）、文本到图像生成（text-to-image generation）等方向奠定了重要基础。

在扩散模型的发展脉络中，研究者相继提出了多种具有里程碑意义的模型，构建了完整的技术演化链。最早的DDPM（denoising diffusion probabilistic models）由加州大学伯克利分校（University of California，Berkeley）的Ho等人提出（2020），该模型通过最大化去噪概率实现高质量图像生成，显著改善了生成过程的稳定性与可解释性。随后，Song与Ermon（斯坦福大学）在2020年提出了DDIM（denoising diffusion implicit models）（Song等，2020），引入隐式去噪机制，大幅加快了采样速度，并进一步提升了生成图像的清晰度与一致性。为了降低生成计算开销，德国海德堡大学（Heidelberg University）的CompVis团队提出了LDM（latent diffusion model），即著名的Stable Diffusion（Rombach等，2022）。该模型通过在潜在空间（latent space）中进行扩散与重建，使得生成过程在保持高质量输出的同时显著减少计算成本。与此同时，多模态文本到图像生成模型逐渐成为扩散研究的重要方向。OpenAI提出的GLIDE（Nichol等，2021）与DALL·E 2（Ramesh等，2022）能够根据自然语言描述生成逼真图像，而Google Research的Imagen（Saharia等，2022）则结合了Transformer架构与扩散机制，在细节还原和语义一致性方面取得了更优性能，开创了文本条件图像生成的新纪元。

在生成可控性方面，研究者提出了ControlNet和T2I-Adapter等结构引导型模型（Zhang等，2023a），通过在扩散框架中引入显式结构化信息（如边缘图、深度图或姿态图），实现对生成内容的精细控制。这类模型能够根据输入的草图或轮廓生成符合结构约束的完整图像，广泛应用于建筑设计、角色造型和动画生成等场景。此外，DiT（diffusion transformer）（Peebles和Xie，2023）进一步将扩散建模与Transformer框架统一融合，利用全局注意力机制捕捉图像的长程依赖关系，从而在图像生成、视频生成乃至多模态统一建模任务中展现出强大的潜力。

除了在图像生成领域的突破，扩散模型还展现出广泛的应用拓展性。在图像修复（image restoration）领域，扩散模型能够生成高质量的缺失区域，广泛应用于医学影像补全与老照片修复；在超分辨率重建（super-resolution）任务中，它可以将低分辨率图像恢复为高清版本，为卫星影像分析与视频增强提供技术支撑；在风格迁移（style transfer）场景中，扩散模型可灵活地将一幅图像的艺术风格迁移到另一幅图像上，应用于艺术创作与广告设计。更重要的是，基于扩散模型的文本到图像生成（text-to-image generation）技术已实现从语义描述到视觉内容的高度一致映射，极大拓展了视觉生成的边界。最后，在图像编辑与反演（editing & inversion）方向，扩散模型凭借其可逆性与高维潜空间表达，使图像编辑、替换与重建更加精确，为虚拟现实与交互式内容创作提供了新的可能性。

综上，扩散模型的快速演化不仅推动了生成式模型从对抗学习向概率建模的范式转变，也奠定了多模态生成与可控生成研究的技术基石。其在稳定性、可控性与多样性方面的持续突破，预示着生成式人工智能正迈向更加统一、高效与智能化的发展

阶段。

表9　基于扩散模型的图像生成方法总结

Table 9　Summary of Image Generation Methods Based on Diffusion Model

核心范式	典型方法	核心机制
典型的图像生成扩散模型	DDPM(Ho等, 2020)	通过最大化去噪概率实现高质量图像生成，显著改善了生成过程的稳定性与可解释性。
	DDIM(Song等, 2020)	引入隐式去噪机制，大幅加快了采样速度，并进一步提升了生成图像的清晰度与一致性。
	Stable Diffusion (Rombach等, 2022)	通过在潜在空间中进行扩散与重建，使得生成过程在保持高质量输出的同时显著减少计算成本。
文本条件图像生成扩散模型	GLIDE(Nichol等, 2021) DALL·E 2(Ramesh等, 2022)	能够根据自然语言描述生成逼真图像。
	Imagen(Saharia等, 2022)	结合Transformer架构与扩散机制，在细节还原和语义一致性方面取得了更优性能。

3)小结传统的图像生成方法主要关注于学习像素分布，生成器通过学习输入数据的像素级信息来生成图像。然而，随着技术的进步，图像生成逐渐转向高层次的语义理解，模型不仅仅依赖于像素数据，而是结合文本描述、布局信息、深度条件等高层语义元素，从而生成更加精准和富有结构的图像内容。例如，文本到图像生成技术(Ramesh等, 2022, Rombach等, 2022)通过理解文本中的语义信息，能够生成符合描述的图像，展现了从像素到语义的重大转变。

图像生成的研究也从单一的图像模态扩展到多模态对齐，即通过综合多个模态的信息生成图像。例如，文本到图像(text-to-image)技术结合了文本和图像，能够根据文字描述生成相应的图像。除此之外，模型还开始结合音频、深度信息和姿态等其他模态数据，使生成过程更加多样和精确。通过这些多模态对齐，生成模型不仅能理解图像的外观，还能综合其他信息，生成符合多重条件的图像内容，拓展了图像生成的应用场景。

1.2.3　视频生成方法的研究现状

1)　基于GAN的视频生成

在生成对抗网络(generative adversarial network, GAN)在图像生成领域取得突破性成果之后，研究者开始将其拓展至具有更高时序复杂度的视频生成任务。早期研究主要集中于无条件视频生成方向。VideoGAN(Vondrick等, 2016)首次设计了前景-背景双流生成结构，通过分别采样静态背景与动态前景的潜变量，并采用掩模加权融合(mask-weighted fusion)生成完整视频。随后，Tulyakov等人提出MoCoGAN(Tulyakov等, 2018)，将运动信息与内容信息解耦，引入独立潜变量分别建模时间变化与外观特征，从而显著提升了生成视频的动态一致性与多样性。Saito等人提出TGAN及其改进版本TGANv2(Saito等, 2020)，采用时间卷积(temporal convolution)结构以捕捉长时序依赖，使生成序列在

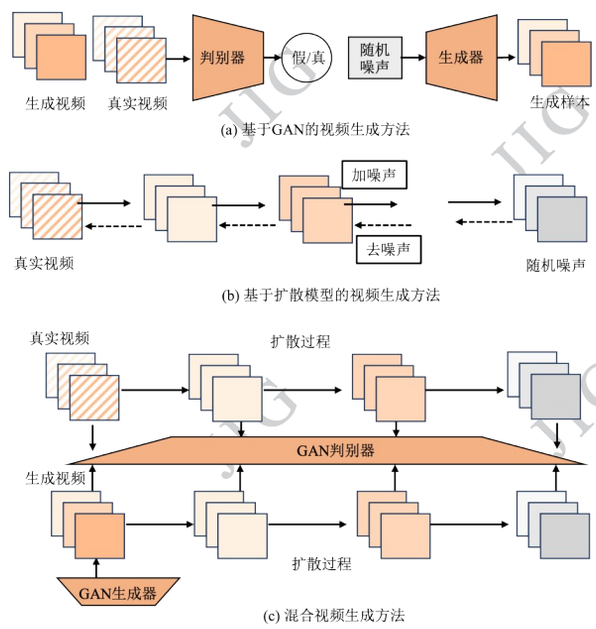

图5　视频生成的方法

Fig. 5　Video generation method ((a) GAN based video generation method; (b) Diffusion models based video generation method; (c) Hybrid video generation method)

时间维度上更加平滑自然。DVD-GAN(Clark等, 2019)与MoCoGAN-HD(Tian等, 2021)进一步引入多尺度生成与判别机制,实现了高分辨率、长时序视频的生成,显著提升了视觉质量与稳定性。

在无条件生成取得阶段性成果后,研究者开始探索将多模态信息作为条件输入,从而实现语义可控的视频生成。其中,Text-to-Video GAN(如TFGAN(Tian等, 2020)、StoryGAN(Li等, 2019))通过将文本描述映射至语义潜空间,引导视频合成过程,实现文本驱动的画面生成。Audio-to-Video GAN(Aldausari等, 2022, Aldausari等, 2024)则通过跨模态学习,使模型能够根据音频信号生成与其时序同步的视觉内容。而Image-to-Video GAN(Zhang等, 2023b)则以单帧图像为输入,通过时序扩展生成连续视频片段,展现出良好的场景动态建模能力。这些研究共同推动了基于GAN的视频生成从无条件到条件、多模态方向的演进,显著增强了生成结果的语义控制与可解释性。

尽管基于GAN的视频生成方法在结构设计和生成质量方面取得了显著进展,但其固有的训练不稳定性与模式坍塌(mode collapse)问题仍制约了模型在高分辨率和长视频生成任务中的表现。这些局限性为后续基于扩散模型的视频生成提供了新的研究契机。

表10　基于GAN的视频生成方法总结

Table 10　Summary of GAN-based video generation methods

核心范式	典型方法	核心机制
无条件视频生成	VideoGAN(Vondrick等, 2016)	设计了前景-背景双流生成结构,通过分别采样静态背景与动态前景的潜变量,并采用掩模加权融合生成完整视频。
	MoCoGAN(Tulyakov等, 2018)	将运动信息与内容信息解耦,引入独立潜变量分别建模时间变化与外观特征,从而显著提升了生成视频的动态一致性与多样性。
	TGANv2(Saito等, 2020)	采用时间卷积结构以捕捉长时序依赖,使生成序列在时间维度上更加平滑自然。
	DVD-GAN (Clark等, 2019) MoCoGAN-HD (Tian等, 2021)	引入多尺度生成与判别机制,实现了高分辨率、长时序视频的生成,显著提升了视觉质量与稳定性。
Text-to-Video GAN	TFGAN(Tian等, 2020) StoryGAN(Li等, 2019)	通过将文本描述映射至语义潜空间,引导视频合成过程,实现文本驱动的画面生成。
Audio-to-Video GAN	(Aldausari等, 2022, Aldausari等, 2024) (Aldausari等, 2022, Aldausari等, 2024)	通过跨模态学习,使模型能够根据音频信号生成与其时序同步的视觉内容。
Image-to-Video GAN	(Zhang等, 2023b)	以单帧图像为输入,通过时序扩展生成连续视频片段,展现出良好的场景动态建模能力。

2)　基于扩散模型的视频生成

随着扩散模型(diffusion models)在图像生成领域取得突破性成果,研究者开始将其推广至视频生成任务,形成了新一代的数据驱动生成框架。早期的代表性工作VDM(video diffusion models)(Ho等, 2022)和PVDM(Yu等, 2023),它们首次将扩散过程扩展至时间维度,通过逐帧或时间块的噪声反演实现视频序列生成。随后提出的AnimateDiff(Guo等, 2023, Lin和Yang, 2024)以及VideoComposer(Wang等, 2023b),在模型结构与控制能力上进一步改进,引入可插拔运动模块与条件控制机制,以实现更灵活的动作编辑与风格保持。

近年来,大规模模型的出现进一步推动了视频扩散技术的发展。OpenAI发布的Sora(2024)(Liu等, 2024e)、中国高性能计算人工智能技术公司(HPC-AI Tech)和北大联合提出的opensora/opensora2.0和opensora-plan(Lin等, 2024, Yang等, 2025a, Zheng等, 2024)、阿里巴巴通义实验室开发的Wan2.1/Wan2.2(Wan等, 2025)、腾讯AI Lab的Hunyuan(Kong等, 2024)以及清华大学与智谱AI联

合提出的Cogvideo、CogVideoX(Hong等，2022，Yang等，2024b)，均将多模态信息(文本、音频、深度与物理信号等)融合入视频扩散框架中，实现了从自然语言描述生成高分辨率、长时序视频片段的能力，显著提升了生成质量与语义一致性。此外，ImagerySearch(Wu等，2025)、NarrLV(Feng等，2025)、VMBench(Ling等，2025)分别从长距离语义一致性、长时序一致性以及运动动态合理性三个维度评估视频生成模型的能力上限，并为后续方法的设计与比较提供了重要参考。

基于扩散的视频生成通常依赖一系列关键技术。首先，时序一致性约束(temporal consistency regularization)用于确保相邻帧之间的运动连续性与内容稳定性，从而缓解随机噪声引入的时序漂移问题。其次，在模型结构方面，多采用空间－时间U-Net或Transformer框架，通过在空间与时间维度上并行建模，实现全局一致的特征生成与动态捕捉。最后，条件控制机制(conditional control)在近年来迅速发展，使模型能够基于多种输入模态(如文本描述、关键帧、深度图、动作轨迹等)实现语义可控的视频内容合成。总体而言，扩散模型的视频生成正经历从逐帧生成向时空并行建模的演进，即从独立帧的局部生成过渡到统一的时空特征建模，以提升生成效率与动态一致性。同时，研究也正从单模态扩散向多模态融合方向拓展，将文本、音频、人体动作及物理约束等信号整合入统一生成框架，使视频生成过程更具语义理解与情境关联能力。这一趋势预示着扩散模型将在未来视频生成领域成为核心范式之一。

表11　基于扩散模型的视频生成方法总结

Table 11　Summary of video generation methods based on diffusion models

核心范式	典型方法	核心机制
维度扩散	Video Diffusion Models (Ho等, 2022) PVDM(Yu等, 2023)	将扩散过程扩展至时间维度，通过逐帧或时间块的噪声反演实现视频序列生成。
模型结构与条件控制	AnimateDiff (Guo等, 2023, Lin和Yang, 2024) VideoComposer(Wang等, 2023b)	在模型结构与控制能力上进一步改进，引入可插拔运动模块与条件控制机制，以实现更灵活的动作编辑与风格保持。
多模态大模型	Sora(2024) (Liu等, 2024e) opensora/opensora2.0、opensora-plan(Lin等, 2024, Yang等, 2025a, Zheng等, 2024) Wan2.1/Wan2.2(Wan等, 2025) Hunyuan(Kong等, 2024) Cogvideo、CogVideoX(Hong等, 2022, Yang等, 2024b)	将多模态信息(文本、音频、深度与物理信号等)融合入视频扩散框架中，实现了从自然语言描述生成高分辨率、长时序视频片段的能力，显著提升了生成质量与语义一致性。

3)　视频生成方法的融合发展

近年来，视频生成领域呈现出多范式融合的发展趋势，研究者正尝试在不同模型架构间取长补短，以突破单一生成范式的性能瓶颈。首先，GAN与Diffusion的混合架构(如Diffusion-GAN(Wang等，2022c)、ScoreGAN(Shehnepoor等，2021))将对抗学习的分布匹配能力与扩散模型的去噪建模机制相结合，在保持训练稳定性的同时显著提升了生成样本的多样性与细节保真度。与此同时，Flow与Diffusion的结合成为另一重要方向，其中Diff2Flow(Schusterbauer等，2025)实现预训练扩散模型(如Stable Diffusion)向流匹配(FM)模型的高效知识迁移，无需额外计算开销，从而显著降低了生成成本并提升推理效率。

在模型架构层面，Transformer驱动的生成统一化(transformer-driven unified generation)正逐渐成为主流趋势。典型代表包括VideoGPT(Yan等，2021)、DiT(diffusion transformer)(Peebles和Xie，2023)。这些模型基于统一的自注意力机制建模时空依赖关系，能够在图像、视频、音频等不同模态下实现跨任务、跨域的统一生成。该类方法通过结构共享和特征耦合，大幅提升了生成模型的泛化能力

与语义一致性。

此外，多模态统一生成（multimodal unified generation）的研究也在快速推进，推动视频生成从静态图像合成迈向具备“世界理解”与“动态推演”能力的生成阶段。代表性工作包括Genie模型（Bruce等，2024），该模型将物理规律与语义约束引入时空生成过程，增强了模型对现实动态的建模能力；另外，VideoCrafter1/2（Chen等，2023a，Chen等，2024a），实现了开放式视频生成与语义可控编辑；以及Sora模型（Liu等，2024e），首次实现了从自然语言文本到长时序、高分辨率视频的端到端生成。

总体而言，视频生成正从单一模态的视觉合成迈向具备认知理解与推理能力的多模态世界建模阶段。这一趋势标志着生成式人工智能正由“感知驱动”向“认知驱动”演进（Huang等，2025a，Huang等，2021），为构建具有世界建模与语义理解能力的下一代通用视频生成系统奠定了基础。

表12　视频生成方法的融合方法总结

Table 12　Summary of video generation method fusion methods

核心范式	典型方法	核心机制
GAN与Diffusion的混合架构	Diffusion-GAN(Wang等, 2022c) ScoreGAN(Shehnepoor等, 2021)	将对抗学习的分布匹配能力与扩散模型的去噪建模机制相结合，在保持训练稳定性的同时显著提升了生成样本的多样性与细节保真度。
Flow与Diffusion的结合	Diff2Flow(Schusterbauer等, 2025)	实现预训练扩散模型（如Stable Diffusion）向流匹配（FM）模型的高效知识迁移，无需额外计算开销，从而显著降低了生成成本并提升推理效率。
Transformer驱动的生成统一化	VideoGPT(Yan等, 2021) DiT(Wang等, 2022b)	基于统一的自注意力机制建模时空依赖关系，能够在图像、视频、音频等不同模态下实现跨任务、跨域的统一生成。
多模态统一生成	Genie模型(Bruce等, 2024)	将物理规律与语义约束引入时空生成过程，增强了模型对现实动态的建模能力。
	VideoCrafter1/2(Chen等, 2023a, Chen等, 2024a)	实现开放式视频生成与语义可控编辑。
	Sora模型(Liu等, 2024e)	实现了从自然语言文本到长时序、高分辨率视频的端到端生成。

4）小结尽管近年来基于GAN和扩散模型的视频生成取得了显著进展，但在实际应用中仍面临多方面的挑战。首先，时序一致性是视频生成任务的核心难题之一，生成模型需要在连续帧之间保持稳定的运动轨迹与外观特征，避免出现帧间闪烁、形变或运动不连贯等问题。其次，空间连贯性同样至关重要，尤其是在复杂场景和高分辨率条件下，模型需同时保证画面的细节一致与结构完整。此外，随着视频生成逐渐向多模态方向发展，多模态同步成为新的研究重点，模型需在图像、动作乃至音频等模态间建立精确的时序对齐关系，实现语义与感知层面的统一。更为根本的挑战在于，视频并非简单的“帧序列叠加”，而是一种动态行为建模与语义连续性生成的过程。模型不仅要捕捉像素层面的运动变化，还需理解场景中角色、物体及事件的逻辑关联，从而生成具有时间逻辑、行为意图和情节连贯的视频内容。这要求视频生成模型同时具备强大的时空建模能力与高层语义理解能力，也为未来的视觉生成研究提出了更高的要求。

1.3　深度伪造检测技术

近年来，随着深度学习生成技术的快速发展，深度伪造技术在多个领域得到广泛应用，如人脸替换和虚拟数字人。然而，技术的滥用也引发了虚假信息传播、身份冒用等社会安全问题。因此，深度伪造检测技术成为计算机视觉和多媒体取证领域的重要研究方向。

深度伪造检测技术通过分析数据中的伪造信息，并结合人工智能技术判断数据的真伪。一些技

术还提供具有解释性的结果，如篡改区域定位和伪造原因分析。当前的伪造信息检测主要从三个方面展开：时间-空间伪造信息、频域伪造信息、预训练模型中的隐式伪造信息。现有研究主要集中在人脸数据的伪造检测以及包括一般内容图像和视频的伪造检测。尽管AIGC图像视频包含人脸数据，但由于早期深伪检测的研究集中在人脸数据上，且人脸数据具有特殊性，因此人脸伪造检测仍然是研究的热点。本文将从人脸伪造检测技术和AIGC图像视频伪造检测技术出发，梳理伪造检测技术的发展与现状，并探讨未来的研究趋势。

1.3.1　人脸伪造检测技术

随着深度伪造技术的不断演进，人脸伪造检测已成为多媒体取证领域的研究热点。为了应对从传统拼接篡改到高逼真神经网络生成的各类攻击，研究者们从物理规律、信号特征以及语义推理等多个维度提出了解决方案。本节将从三个层面系统梳理现有的人脸伪造检测技术，如图6所示。首先是利用混合边界、几何约束及时间连贯性差异的基于时空信息的检测技术。其次是挖掘图像在离散余弦变换或傅里叶变换中残留统计异常的基于频域信息的检测技术。最后是结合多模态大模型以提升决策透明度与逻辑推理能力的可解释检测技术。

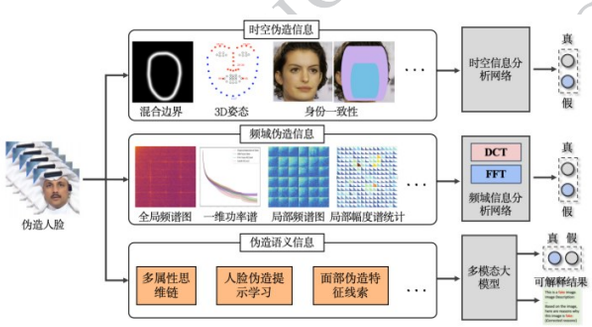

图6　人脸伪造检测方法

Fig. 6　Method for detecting face forgery

1）　基于时空信息的人脸伪造检测技术

针对基于时空信息的人脸伪造检测，相关研究主要致力于挖掘伪造生成过程、后期处理步骤以及人脸属性在图像中残留的时空线索，以此鉴别数据真伪。在各类人脸伪造算法中，将生成面部融合至原始背景是一项核心环节，该操作不可避免地会在图像的融合边界处引入异常痕迹。基于这一观察，FaceX-ray（Li等，2020）提出了一种高效的深伪检测框架。该方法将检测任务解耦为两个阶段：首先利用分割模型定位输入图像中的混合边界，继而对边界区域进行分类以判定真伪。FaceX-ray的显著优势在于其聚焦于必要的后处理操作，而非依赖于特定的人脸生成方法，从而大幅增强了检测模型的泛化能力。此后，Li等人（2018）进一步指出，图像融合过程除产生边界瑕疵外，为实现面部对齐而执行的仿射变换同样会残留特定伪影。据此，他们设计了卷积网络以专门捕捉此类仿射变换痕迹来实现检测。与此同时，Dang等人（2020）通过引入注意力图机制实现了对篡改区域的精确定位，使得检测性能得到显著提升。

尽管基于混合边界或篡改区域定位的方法具备一定的检测效力，但其性能往往严重依赖于训练集中特定的伪造模式，导致在面对未知攻击时泛化能力不足。为突破这一局限，研究者们开始引入各类先验知识以辅助深伪检测，例如几何先验（Yang等，2019，Zhu等，2021），局部相关性（Chen等，2021，Zhao等，2021），和身份一致性（Dong等，2022）。鉴于真实人脸严格遵循物理世界的三维几何约束，而伪造人脸常伴随违反该规律的异常，这些三维特性能够揭示人眼难以察觉的细微伪影。例如，Yang等人（2019）利用人脸图像估计3D头部姿态，通过捕捉伪造图像中关键点的几何偏差来鉴别真伪。FD2Net（Zhu等，2021）则基于3D可变形模型将人脸解耦为形状、纹理及光照等分量，并发现伪造痕迹主要残留于身份纹理和直射光中，但复杂的3D建模过程在一定程度上制约了检测效率。此外，挖掘图像局部区域间的关联性也成为提升泛化性的重要手段。Chen等人（2021）设计了多尺度补丁相似性模块，通过度量局部特征的相似度来学习鲁棒的伪影表征。Zhao等人（2021）侧重于利用伪造图像中的不一致性线索，提出成对自一致性学习策略，通过不一致图像生成器构造训练数据，从而挖掘生成技术特有的源特征残留。针对伪造算法在区域关系学习上的缺失，Xu等人（2021）提出了视觉-语义模型，旨在感知不同区域的语义伪造特征并捕捉异常关联。Fei等人（2022）则提出二阶局部异常学习模块，通过对局部特征邻域的方向与距离分解，计算一阶及二阶异常图，从而提取出更具通用性的伪造痕迹。

除时空与纹理线索外，身份信息的一致性也是提升深度伪造检测鲁棒性的关键维度。在伪造图像

中，内部面部区域与外部轮廓往往存在身份归属冲突。针对这一特性，ICT（Dong等，2022）提出了基于高层语义的身份一致性模型。该方法通过引入额外的身份先验信息，在检测涉及公众人物的面部伪造时表现尤为出色。为解决跨域检测难题，张等人（2025）设计了一种基于多样性负样本生成的检测框架。该模型采用孪生网络架构提取多视图融合特征，并通过引入对比约束机制增强样本特征的可判别性；同时，利用特定的构造规则生成高差异性的负实例，从而有效提升了模型的泛化能力。在骨干网络优化方面，王等人（2024）提出了一种基于注意力机制改进ResNet34的检测方法，旨在应对新型伪造技术。该方法通过集成高效通道注意力模块，引导模型聚焦图像中的伪影区域，显著增强了对局部异常特征的捕捉效能。此外，孟等人（2024）探索了混合架构的潜力，提出了一种结合CNN与ViT的策略，旨在兼顾卷积网络的局部特征提取优势与Transformer的全局关联建模能力，从而提升模型在实际应用场景中的综合性能。

此外，研究表明，人脸视频伪造技术往往会破坏相邻帧之间的时空一致性，导致视频流中出现时序抖动或不连续现象。针对这一特性，基于帧间不连续性的深伪视频检测方法应运而生。早期工作如文献（Güera和Delp，2018，Sabir等，2019）采用长短期记忆网络处理图像帧的特征序列，旨在捕捉视频中的时序异常信号。Amerini等人（2019）则利用光流场来表征帧间运动一致性，以此识别伪造痕迹。随着研究的深入，为进一步增强对视频时空不一致性的建模能力，研究者们相继引入了更复杂的架构，包括双向长短期记忆网络（Masi等，2020）、时空注意力机制（Chen等，2022）、三维卷积（Nguyen等，2021），从而显著提升了视频伪造检测的准确性与鲁棒性。

表13　基于时空信息的人脸伪造检测方法总结

Table 13　Summary of face forgery detection methods based on spatiotemporal information

核心范式	典型方法	核心机制
融合边界与变换痕迹	FaceX-ray(Li等, 2020) Li等人(2018) Dang等人(2020)	聚焦于伪造生成流程中不可避免的后期处理痕迹，通过定位图像混合边界或捕捉仿射变换残留的特定伪影来鉴别真伪，不依赖于特定的人脸生成方法。
基于几何先验与局部一致性	Yang等人(2019) D2Net(Zhu等, 2021)	引入物理世界的3D几何约束（如姿态、光照）以及图像局部区域间的关联性先验，通过挖掘违反物理规律的几何偏差或局部语义/纹理的不一致性来发现异常。
基于身份一致性与架构优化	ICT(Dong等, 2022) 张等人(2025)	利用高层语义检测面部内外部区域的身份归属冲突，并通过改进骨干网络增强对跨域伪造特征的捕捉与泛化能力。
基于帧间时序不一致性	(Güera和Delp, 2018, Sabir等, 2019) (Güera和Delp, 2018, Sabir等, 2019) (Amerini等, 2019)	针对视频伪造破坏相邻帧连续性的特点，利用时间序列模型或光流场分析，捕捉视频流中存在的时序抖动、闪烁或运动不连续痕迹。

2）　基于频域信息的人脸伪造检测技术

基于频域分析的伪造检测方法旨在利用真伪数据在频谱特性上的差异进行取证。现有研究表明，人脸伪造算法在执行区域融合及上采样推理等操作时，不可避免地会引入频域异常，且这种异常在高频分量中尤为显著。针对这一现象，基于频域线索的检测方案应运而生，其核心工具通常涵盖离散傅里叶变换、离散余弦变换及小波变换等。

研究（Frank等，2020）指出频率伪影在不同网络架构、数据集及分辨率下表现出高度的一致性，且伪造图像在高频段呈现出异于真实图像的独特模式。受此启发，Durall等人（2020）提出通过计算DFT

空间中各频段的平均振幅来捕捉伪造模式。F3-Net（Qian等，2020）则设计了一种创新的双流框架，该框架由两部分构成：一是频率感知分解模块，用于从原始输入中提取潜在的频率伪影；二是高层语义分析模块，旨在对局部区域的复杂频率统计特征进行建模，从而精准刻画真伪人脸在频域上的分布差异。此外，HFI-Net（Miao等，2022）通过整合双分支网络与四个全局-局部交互模块，实现了对多层次频率伪影的有效挖掘。Jia等人（Jia等，2021b）提出了一种基于平稳小波分解的不一致性感知网络，通过增强伪造特征来提升检测性能。Luo等人（2021）则另辟蹊径，指出图像噪声能有效抑制颜色纹理等过拟合干扰，进而暴露篡改痕迹。基于此发现，他们构建了包含多尺度高频提取、残差引导空间注意以及跨模态注意的三模块架构，通过高通滤波与RGB空间特征的交互融合，实现了对高频伪造线索的深度利用。

此外，SPSL（Liu等，2021a）的研究进一步指出，作为人脸伪造流程中不可或缺的环节，上采样操作会导致频域特征产生显著的累积异常。针对这一发现，SPSL提出了一种空间-相位浅层学习框架，通过联合分析空间图像与频域相位谱，有效识别了由上采样引入的独特伪影。尽管方法（Liu等，2021a，Luo等，2021）在跨库测试中展现了一定的泛化潜力，但其依赖固定滤波器组与手工设计特征的范式，限制了模型对深层判别性特征的挖掘能力。为突破这一瓶颈，Li等人（2021）设计了一种自适应频率特征生成模块，实现了以完全数据驱动的方式自动发现关键频率线索。

另一方面，为解决过拟合问题并提升检测器的泛化性，FrePGAN（Jeong等，2022b）引入了一种基于频率级扰动的对抗训练策略。该方法通过生成频率扰动图并引导检测器忽略这些非本质的频率伪影，从而迫使模型聚焦于图像级的结构性不规则；其采用的分类器与生成器交替更新机制，被证明能有效增强模型的鲁棒性。此外，董等人（2025）提出了一种空频多特征融合网络，通过将频域动态划分为三个子频带，提取了空域分析难以捕捉的隐蔽伪影，并借助空域信息的补充进一步提升了模型的泛化性能。

表14　基于频域信息的人脸伪造检测方法总结

Table 14　Summary of Face Spoofing Detection Methods Based on Frequency Domain Information

核心范式	典型方法	核心机制
基于频谱分析与特征分解	Durall等人(2020) F3-Net(Qian等, 2020) Luo等人(2021)	利用离散傅里叶/余弦变换或小波分解，挖掘伪造图像在不同频段的异常统计分布；通过双流架构或多尺度提取，捕捉真伪图像在频率幅值及纹理细节上的差异。
基于相位谱与自适应学习	SPSL(Liu等, 2021a) Li等人(2021)	针对上采样操作引入的累积异常，引入相位谱分析以弥补仅依赖幅值的不足；或采用数据驱动的自适应频率特征生成模块，突破固定滤波器组的局限，自动发现关键的频率判别线索。
基于对抗训练与空频融合	FrePGAN(Jeong等, 2022b) 董等人(2025)	为提升模型的泛化性与鲁棒性，引入频率级对抗扰动训练，迫使模型聚焦于本质的结构性不规则；或采用空频多特征融合策略，结合频域子带分析与空域信息，互补捕捉隐蔽伪影。

3）　可解释人脸伪造检测技术

近年来，研究者开始尝试利用多模态模型实现可解释性人脸伪造检测，以提升检测结果的透明度与可理解性，更好地满足用户在安全审查、司法取证及内容监管等场景中的需求。早期工作如DD-VQA（Zhang等，2024）通过众包平台收集人类对深度伪造数据的注释信息，并在此基础上微调BLIP多模态模型，使模型能够生成关于伪造内容的自然语言描述，从而实现检测结果的可解释输出。随着多模态大型语言模型的快速发展，研究者逐渐探索其在伪造检测任务中的潜力。Jia等人（2024）首次系统性评估了GPT在检测被操纵人脸方面的能力，验证了语言模型在识别视觉伪造中的语义推理优势。随后，FFAA（Huang等，2024）利用GPT-4o进行伪造样本注释生成和模型微调，进一步提升了模型在多模态语义对齐与伪造识别方面的性能。X2DFD（Chen

等，2024b)提出了一种自增强机制，通过自监督反馈优化模型的伪造检测能力，显著增强了多模态语言模型在复杂场景下的稳定性与适应性。在可解释性层面，FFTG(Sun等，2025)引入伪造掩码信息，实现了对伪造区域与伪造类型的初步识别，并通过精心设计的提示策略减少语言模型幻觉问题，从而生成更加准确的文本描述结果。M2F2-Det(Guo等，2025)进一步推进了这一方向，通过定制的人脸伪造提示学习和预训练的CLIP模型，提升了对未知伪造样本的泛化能力。此外，M2F2-Det结合了大型语言模型，为伪造检测决策提供了详细的文本解释，通过弥合自然语言与面部伪造细微线索之间的差距，显著增强了可解释性和决策透明度。而Shield(Shi等，2025)则提出多属性链式思维范式，能够围绕图像中的多个属性(包括任务相关与无关特征)进行多层次推理，使模型在推断伪造特征来源与类型时具备更强的逻辑一致性和解释能力。新兴的多模态大模型通过自然语言推理与视觉语义对齐，实现了从检测结果到检测原因的转变。这不仅提升了伪造检测的透明性与可用性，也为实现可信人工智能取证系统奠定了基础。未来的研究可进一步结合视觉解释性与语言推理能力，以实现更高精度、更高可解释度的伪造检测框架。

表15　可解释人脸伪造检测方法总结

Table 15　summarizes methods for explaining face spoofing detection

核心范式	典型方法	核心机制
基于多模态描述与评估	DD-VQA(Zhang等，2024)	利用众包注释微调多模态模型以生成伪造内容的自然语言描述，或系统性评估大语言模型在识别视觉伪造中的语义推理优势。
大模型驱动的对齐与增强	FFAA(Huang等，2024) X2DFD(Chen等，2024b)	利用先进大模型生成注释并微调以提升多模态语义对齐性能，或通过自监督反馈机制实现自我增强，提高模型在复杂场景下的稳定性。
细粒度推理与逻辑解释	M2F2-Det(Guo等，2025) Shield(Shi等，2025)	结合提示学习与CLIP模型弥合视觉线索与语言的差距，或引入多属性链式思维范式，围绕图像属性进行多层次推理，提供具备逻辑一致性的详细解释。

1.3.2　AIGC图像视频伪造检测技术

不同于局部的人脸操纵，基于扩散模型和生成对抗网络等生成算法通常对整幅图像或视频序列的完全合成。这种生成机制在像素分布、纹理细节以及底层频率指纹上留下了不同于传统伪造的独特痕迹。本节将针对通用AIGC生成内容的取证难题，分别从三个技术路线展开论述，如图7所示。包括侧重于捕捉梯度异常与像素级伪影的基于时空信息的检测技术，致力于提取生成模型固有频谱指纹的基于频域信息的检测技术，以及融合视觉与语言模型以实现检测结果溯源与归因的可解释AIGC伪造检测技术。

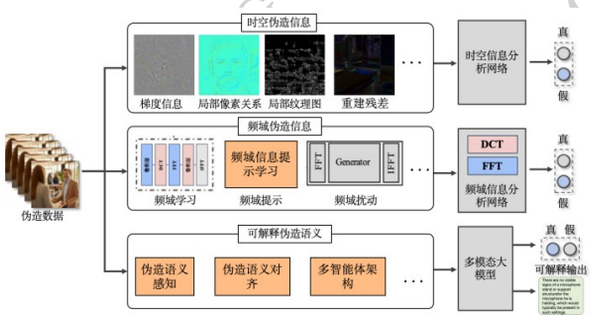

图7　AIGC图像视频伪造检测方法

Figure 7　AIGC Image and Video Forgery Detection Method

1)　基于时空信息的AIGC图像视频伪造检测技术

基于伪影表征的AIGC图像检测旨在通过提取潜在的篡改痕迹来鉴别数据真伪。现有研究主要将此类表征划分为两大范畴：像素级伪影与高维特征表示。在像素级层面，LGrad(Tan等，2023)提出利用梯度信息进行取证，其研究表明生成图像与判别器梯度之间存在内在的信息耦合，梯度图中指示的待优化区域实质上揭示了生成器未能完美拟合的结构性缺陷。局部像素关系(Tan等，2024c)则聚焦于上采样操作引入的局部像素关联性，指出卷积操作会将上采样导致的局部相似性特征扩散至全图，因此提出通过计算局部像素间的相对差异来有效捕获此类痕迹。此外，利用神经网络提取中间层特征作为高维伪影表示也是当前的主流方向。例如，随机

映射方法(Tan等，2024a)利用随机初始化的卷积层将图像投影至高维空间，并结合特定滤波器提取异常模式。UniFD(Ojha等，2023)则直接验证了CLIP模型的预训练特征在实现高效线性检测方面的潜力。针对纹理特征，PatchCraft(Zhong等，2023)观察到生成模型在纹理丰富区域留下的痕迹较平滑区域更为显著，据此设计了一种粉碎与重建预处理策略，通过挖掘不同纹理密度区域间的像素相关性差异来增强检测性能。在基于重建的检测方面，扩散重建误差(Wang等，2023c)利用预训练扩散模型计算输入图像与其重建结果之间的误差来判定真伪。Ma等人(2023)则在此基础上提出了针对扩散生成图像的逐步误差分析方法。

表16　基于空间信息的AIGC图像视频伪造检测方法总结

Table 16　Summary of AIGC Image and Video Forgery Detection Methods Based on Spatial Information

核心范式	典型方法	核心机制
像素级伪影表征	LGrad(Tan等, 2023) 局部像素关系(Tan等, 2024c)	聚焦于生成器未能完美拟合的结构性缺陷或上采样操作留下的痕迹，利用梯度信息揭示待优化区域，或通过计算局部像素间的相对差异捕获相关性异常。
高维特征与纹理分析	UniFD(Ojha等, 2023) PatchCraft(Zhong等, 2023)	利用神经网络提取高维中间层特征，或通过粉碎重建策略挖掘纹理丰富区域与平滑区域间的像素相关性差异，以增强对异常模式的捕捉。
重建残差	扩散重建误差(Wang等, 2023c) Ma等人(2023)	利用预训练扩散模型对输入图像进行重建，通过计算原始图像与重建结果之间的误差来判定数据真伪。

2)　基于频域信息的AIGC图像视频伪造检测技术

与特定结构的人脸图像不同，AIGC生成图像具有极高的内容丰富性与语义多样性，这导致直接依赖频域信息难以实现具有高泛化性的伪造检测。为此，近期研究(Jeong等，2022b，Liu等，2024b，Tan等，2024b)致力于优化频域信息的引入方式以增强检测器的鲁棒性。例如，BiHPF(Jeong等，2022a)采用双高通滤波器策略来显著放大潜在的伪影信号。FrePGAN(Jeong等，2022b)的研究指出，尽管频率级伪影普遍存在，但其表现形式易受GAN模型架构或生成对象类别的影响，这种差异性容易导致检测模型过拟合。针对此问题，该方法通过生成频率级扰动图进行对抗训练，从而消除特定频率伪影的干扰，引导模型关注更本质的特征。另一方面，Tan等人(2024b)提出的FreqNet摒弃了直接输入原始频谱数据的传统思路，转而将图像的高频分量作为输入，并结合高频持续关注模块与频域学习模块，旨在提取与生成源无关的通用特征。FatFormer(Liu等，2024b)则创新性地将频率分析与文本编码器相结合，并将其作为适配器嵌入冻结权重的CLIP视觉模型中，从而有效提升了检测性能。针对扩散模型，Synthbuster(Bammey，2023)专注于挖掘扩散过程残留的固有频率指纹，通过对残差图像进行傅里叶变换，突出了扩散特有的频率伪影以区分真伪。D4(Hooda等，2024)提出了一种分离扩散深伪检测框架，该方法利用集成学习策略对频谱上互不相交的子集进行融合判定；通过显著性划分技术降低对抗子空间的维度，从而显著提升了模型抵御黑箱对抗攻击的能力。此外，Xi等人(2023)开发了一种包含残差流与内容流的双流网络：残差流利用空域富模型精确提取纹理细节，内容流则负责捕捉低频区域的伪造痕迹，最后引入交叉多头注意力机制强化双流间的信息交互。

3)　可解释AIGC图像视频伪造检测技术

近年来，基于多模态大型语言模型的伪造检测方法得到了广泛关注，尤其是在提高可解释性方面表现突出。早期的研究中，ForgeryGPT(Liu等，2024c)提出了利用多模态大语言模型提取图像中的伪造知识，通过引入伪造感知提取器与伪造定位专家，使得不仅能够实现伪造区域的像素级定位，还能生成可解释的输出，显著提升了伪造检测的透明度。紧随其后，SIDA(Li等，2025d)框架则专注于社交媒体图像伪造的检测与解释，结合Mask-Text对齐技术，不仅能精准识别伪造图像，还能清晰解释伪造操作的性质与位置，为社交媒体内容的安全监管提供了有效的技术支持。在多任务学习方面，FakeVLM

表17　基于频域信息的AIGC图像视频伪造检测方法总结

Table 17　Summary of AIGC Image and Video Forgery Detection Methods Based on Frequency Domain Information

核心范式	典型方法	核心机制
频域特征增强与提取	BiHPF(Jeong等, 2022a) FreqNet(Tan等, 2024b)	摒弃直接输入原始频谱的传统思路，采用双高通滤波器放大伪影信号，或聚焦于高频分量并结合注意力机制，旨在提取与生成源无关的通用特征。
对抗训练与适配器学习	FrePGAN(Jeong等, 2022b) FatFormer(Liu等, 2024b)	引入频率级扰动进行对抗训练以消除特定伪影干扰，或将频率分析与文本编码器作为适配器嵌入冻结的CLIP模型中，显著提升对不同生成架构的泛化性。
扩散指纹与多流融合	Synthbuster(Bammey, 2023) D4(Hooda等, 2024)	针对扩散模型挖掘残差图像中的固有频率指纹，利用频谱子集集成学习抵御对抗攻击，或构建残差与内容双流网络，通过交叉注意力融合纹理细节与低频伪造痕迹。

(Wen等, 2025c)通过结合图像-语言对齐和伪造分析，在伪造图像的多模态检测与解释方面表现出色。通过引入FakeClue数据集，FakeVLM不仅能检测伪造内容，还能生成基于文本的伪造线索解释，进一步提升了决策支持的透明度。LEGION(Kang等, 2025)提出了一个多任务学习框架，结合伪造检测与解释生成，并通过引入伪造对齐对比学习模块，显著提升了多模态模型在复杂伪造图像场景中的表现，提出的大规模数据集SynthScars为研究提供了宝贵的资源。

随着技术的进一步发展，FakeReasoning(Gao等, 2025b)结合了伪造检测与推理任务，利用多模态语言模型增强了伪造检测的泛化能力，尤其在复杂伪造类型的识别中表现出了较强的优势。与此同时，AIGI-Holmes(Xu等, 2025)系统通过结合视觉专家模型和MLLMs的语义推理，生成可供人工验证和对齐的解释，进一步增强了模型的泛化能力和可解释性。IVY-FAKE(Zhang等, 2025b)引入了一个统一的框架，支持图像与视频内容的综合检测与解释，弥补了现有方法在处理视频伪造时的不足。该框架通过跨模态的检测与推理，增强了伪造检测的广泛适应性，适用于多种类型的伪造内容。在视频伪造检测方面，BusterX(Wen等, 2025a)提出了结合MLLM和强化学习的框架，专注于视频伪造的检测与解释，并通过跨模态信息的融合提升了检测的准确性和透明度。BusterX++(Wen等, 2025b)在此基础上扩展了多阶段训练和混合推理策略，解决了跨模态伪造检测中的冷启动问题，进一步提升了图像和视频模态联合训练的稳定性和性能。Veritas(Tan等, 2025)提出了模式感知推理的方法，模拟了人类的取证过程。其研究通过增强推理过程，不仅提升了伪造检测的准确性，还能提供更具透明度的推理路径，从而提升了模型的可解释性。ThinkFake(Huang等, 2025c)引入了结构化推理的机制，结合多模态大语言模型，通过推理增强了伪造检测的能力，并生成了具有可解释性的检测报告，有效地支持了复杂伪造内容的识别。最后，RAIDX(Li等, 2025c)提出了结合检索增强生成和群体相对策略优化(group relative policy optimization, GRPO)强化学习框架的深伪检测方法。该方法不仅提升了伪造检测的准确性，还通过生成细粒度的显著性图和文本描述，为模型提供了更加可解释的输出。UniShield(Huang等, 2025b)利用多种伪造检测方法作为感知代理或检测代理来实现伪造图像的检测与定位，其中感知代理智能分析图像特征以动态选择合适的检测模型，检测代理则整合多种专家级探测器生成可解释报告，进一步提高了伪造检测的可靠性与解释能力。

综上所述，当前基于多模态大型语言模型的AIGC图像视频伪造检测技术已取得显著进展。研究不仅关注伪造检测的准确性，还逐步引入可解释性机制，使得伪造检测的过程更加透明。未来的研究可以进一步结合跨模态融合、实时检测和用户友好性等方面的创新，提升模型的泛化能力和应用范围，为构建可信的人工智能取证系统奠定基础。

表18　可解释AIGC图像视频伪造检测方法总结

Table 18　summarizes the AIGC image and video forgery detection methods

核心范式	典型方法	核心机制
多任务学习与精细化定位	ForgeryGPT(Liu等, 2024c) FakeVLM(Wen等, 2025c) LEGION(Kang等, 2025)	将伪造检测与解释生成作为联合任务，利用特制数据集和对齐技术，实现从像素级伪造定位到自然语言线索生成的全链路解释。
类人推理与语义增强	Veritas(Tan等, 2025) AIGI-Holmes(Xu等, 2025) ThinkFake(Huang等, 2025c)	模拟人类取证过程或引入结构化推理机制，结合视觉专家模型与MLLM的语义推理能力，生成具备逻辑透明度、可供人工验证的详细检测报告。
统一框架与代理/强化机制	BusterX(Wen等, 2025a) UniShield(Huang等, 2025b) RAIDX(Li等, 2025c)	拓展至图像与视频的统一检测，引入强化学习、检索增强生成或多智能体架构，动态调用专家模型或优化跨模态融合，提升复杂场景下的解释性。

2　国内外研究进展比较

2.1　图像与视频理解安全

国内外图像与视频异常检测研究在核心方法论上呈现出高度一致的技术轨迹，普遍经历了从传统手工特征提取到深度学习驱动的范式转变，并沿循着自编码器、生成对抗网络到Transformer与大模型的清晰演进路径。然而，在研究关注点与产业生态层面，国内外的发展方向却展现出显著差异。双方虽在数据稀缺、模型泛化能力不足以及评估标准不完善等方面面临共性挑战，但正是这些“瓶颈问题”推动了全球范围内的前沿探索与范式创新。

国际研究生态以北美与欧洲为核心，其学术界普遍强调理论深度与范式批判性。国际学者不仅率先开创了利用生成对抗网络进行异常检测的研究（如AnoGAN），还系统性地反思并揭示了自编码器等主流方法在重构偏差与泛化能力上的结构性缺陷。同时，他们将可解释异常检测（explainable anomaly detection，XAD）正式确立为独立的研究方向，推动了从模型性能导向向模型可解释性与可信性导向的转变。在产业层面，国际企业以NVIDIA Metropolis为代表，普遍采用“平台+生态”的开放赋能模式，致力于构建支持全球开发者的视觉AI基础设施，而非提供单一的终端应用产品。这种战略布局推动了技术的模块化复用与生态共建，加速了从研究成果到产业落地的全球扩散。

与之形成对照，中国的研究生态则展现出产学研一体化与应用驱动的显著特征。国内顶尖高校与科技企业的研究工作更注重模型架构的创新与工程优化，旨在针对具体场景问题实现性能突破与部署落地。例如，国内研究者在视频异常检测、弱监督学习、多模态融合等方向持续提出具备实际适用性的改进方案，在多个国际公开基准上取得领先表现。在产业层面，以华为、阿里巴巴、商汤科技等为代表的企业群体，则通过深度结合庞大的本土市场需求，形成了以“垂直整合、端到端闭环”为特征的解决方案生态。例如，华为聚焦智慧城市与基础设施安全监测，阿里巴巴在电商内容审核中构建行为级异常分析系统，而商汤科技则致力于打造覆盖城市级视频网络的综合监控与分析平台。这一发展模式体现了国内企业在应用导向与工程落地方面的系统性优势。

总体而言，全球视觉异常检测研究在技术发展的主干路径上高度趋同，而在分支方向上各具特色：国际生态以理论创新与平台赋能引领前沿，国内生态以应用牵引与方案集成展现活力。两者并非优劣之分，而是源于不同的市场结构与创新激励机制下的自然分化，形成了互补共生的全球创新格局。随着视觉大模型与多模态预训练框架的普及，国内外研究者在利用大规模知识库与统一表示空间方面已步入同一技术起点。这种同步发展预示着全球视觉异常检测领域将迎来理论与应用深度融合的新阶段，推动智能视觉安全系统迈向更具认知性、解释性与普适性的下一代智能范式。

2.2 图像与视频生成安全

图像与视频生成安全涵盖了“生成技术”本身的发展与“生成内容检测”两个方面。国内外在这两个方面均投入了大量研究资源，但在技术优势和研究焦点上有所分化。

2.2.1 图像与视频生成技术

在生成技术方面，国外研究起步较早，在基础理论框架与模型构建上具有先发优势。早期工作（如VideoGAN、MoCoGAN）聚焦于运动与外观信息的解耦建模，以解决时序一致性问题。后续工作（如DVD-GAN、PVDM）在多尺度判别与时空扩散建模方面取得了进展。近期，以Sora、Genie等模型为代表，国外研究开始引入多模态融合与世界建模理念，探索将视频生成从视觉合成拓展至语义推演与物理感知层面，形成了以Transformer与扩散机制为核心的统一生成框架。

国内在视频生成领域发展迅速，研究重点集中于多模态可控生成和大模型体系的构建。例如，AnimateDiff在可插拔运动控制模块方面进行了探索；VideoComposer关注条件引导与风格保持；Wan2.1/Wan2.2模型在跨模态语义对齐和视觉细节重建方面开展了研究；CogVideoX探索了文本到视频的高分辨率、长时序生成。同时，OpenSora等开源项目的推进，显示了国内在工程实现、训练加速及可扩展性方面的研究积累。

总结而言，国外研究在生成理论框架、物理一致性建模与世界模型构建方面仍处于前沿；国内研究则在工程实现、多模态语义融合与可控生成应用方面形成了快速发展的态势。

2.2.2 深度伪造检测技术

在应对生成内容风险的深度伪造检测技术方面，国内外研究针对不同伪造类型的技术发展阶段有所差异。

在人脸数据深度伪造检测方面，国外研究起步较早，技术积累相对成熟。早期研究（如FaceX-ray、DFDC项目）主要通过分析图像中的底层伪造痕迹（如混合边界、频域特性、高频伪影）来实现检测，这些方法在针对早期GAN模型生成的人脸伪造时取得了较好效果。

在AIGC（人工智能生成内容）通用图像视频深度伪造检测方面，国内外研究几乎同步发展。国外研究者较早开始利用大规模预训练模型提升检测模型的精度和泛化能力。国内研究者则在基于图像梯度、局部像素关系等伪造痕迹检测方面开展了研究。

值得注意的是，在可解释伪造检测领域，国内研究形成了一个集中的研究方向。来自中国科学技术大学、北京交通大学、香港中文大学（深圳）、厦门大学等机构的研究团队，致力于将可解释性与深度伪造检测相结合，提出了基于多模态模型、自然语言处理与图像分析的伪造内容解释方法。此类方法不仅旨在定位伪造区域，还尝试分析伪造生成的具体过程和逻辑，以提升检测结果的透明度与可信性。

3 发展趋势与展望

3.1 图像与视频理解安全

大型视觉语言模型正引领图像与视频异常检测的范式变革。其强大的上下文感知、逻辑推理与零样本能力突破了传统方法依赖统计偏差或重构误差、难以处理语义不一致的局限，使检测从模式识别提升至语义推理层面。未来研究将聚焦三方面：多模态基准构建、开放世界任务推进及大模型深度融合。

1） 大规模、多模态基准的创建

大模型训练依赖多样化数据，传统数据集（如UCSD Pedestrian、CUHK Avenue）场景单一、异常类型有限、环境因素缺乏，难支撑复杂现实任务。新基准（如MSAD）已扩展至14类场景与55种异常类型，并引入环境变化。未来的数据集将融合视觉、语音、深度与热成像信息，提升跨模态关联学习与情境感知，使模型能在信息缺失或模糊场景下保持鲁棒性。

2） 开放世界任务的推进

异常检测正由“封闭世界”向“开放世界”转变。现实中异常类型无限且不断演化，要求模型识别未知异常并适应概念漂移。零样本与增量学习成为关键途径：零样本学习借助大模型的指令理解能力，通过自然语言描述识别未定义异常；增量学习则使模型持续吸收新数据、避免灾难性遗忘，保持长期稳定性能。

3） 大模型的深度融合

融合不同类型大模型是构建高效、可解释检测系统的关键。基础模型既可作为特征编码器提取高质量表征，又可直接执行检测或生成自然语言解释。大型语言模型的引入使检测能识别逻辑层面的“语

义异常”，如“汽车出现在人行道上”。AnomalyGPT（Yang等，2024b）等工作已经展示了利用LLMs解决逻辑异常和结构异常的巨大潜力。

3.2 图像与视频生成安全

1） 图像与视频生成技术

生成技术的快速发展引发数据隐私、内容真实性与责任归属问题。训练数据可能含敏感信息且缺乏授权机制，生成视频的滥用也加剧虚假传播与社会信任风险。未来研究将从“高质量生成”迈向“安全可信生成”，在模型、数据与内容三方面提升安全性。

在模型层面，可通过红队对抗、奖励建模与安全蒸馏强化风险感知，使模型主动规避有害内容。在数据层面，隐私保护、版权溯源与伪造检测是核心议题，可利用联邦学习、隐私增强生成与数字水印实现全流程安全治理。在内容层面，亟需建立统一的真实性验证与追溯机制，结合区块链、数字签名与内容标识标准实现生成内容的身份标识与责任追踪。

2） 深度伪造检测技术

深度伪造内容愈发逼真，使检测技术面临更高挑战。未来研究将向系统化、多模态、对抗防御与可解释性方向发展。

（1）**系统性伪造痕迹分析**：从单一特征检测转向生成过程建模，分析伪造“基因”以提升鲁棒性与适应性。

（2）**增强检测可解释性**：结合多模态大模型利用语义理解发现伪造逻辑不一致，并生成自然语言解释伪造区域与异常原因。

（3）**对抗防御机制**：通过对抗训练与鲁棒特征学习提升抗攻击能力，建立实时检测与防御机制防止绕过攻击。

（4）**自动化与实时检测**：基于自动化模型与分布式计算实现大规模实时筛查，提高效率并减少人工干预。

（5）**法律与伦理框架**：完善全球治理与法律约束，明确隐私与责任界限，保障技术合理应用。

未来的深度伪造检测将依托多模态模型、自动化系统与防御机制，构建高效、可信的多层安全体系，为社会信息安全与个人隐私提供保障。

E-mail：kqhuang@nlpr. ia. ac. cn
E-mail：yzhao@bjtu. edu. cn
E-mail：kaoyueying@bjast. ac. cn
E-mail：tanchuangchuang@bjtu. edu. cn

致谢：本文由中国图象图形学学会视频图像与安全专业委员会组织撰写，该专业委员会链接为https://www.csig.org.cn/16/201704/49324.html。

参考文献

Acsintoae A, Florescu A, Georgescu M-I, Mare T, Sumedrea P, Ionescu R T, Khan F S, and Shah M. 2022. Ubnormal: New benchmark for supervised open-set video anomaly detection// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition (CVPR). New Orleans, USA: IEEE: 20111-20121 [DOI: 10.1109/cvpr52688.2022.01951]

Akcay S, Abarghouei A A, and Breckon T P. 2018. Ganomaly: Semi-supervised anomaly detection via adversarial training// Asian Conference on Computer Vision (ACCV). Perth, Australia: Springer: 622-637 [DOI: 10.1007/978-3-030-20893-6_39]

Al-Lahham A, Tastan N, Zaheer M Z, and Nandakumar K. 2024. A coarse-to-fine pseudo-labeling (c2fpl) framework for unsupervised video anomaly detection// 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa (Hawaii), USA: IEEE: 6779-6788 [DOI: 10.1109/wacv57701.2024.00665]

Aldausari N, Sowmya A, Marcus N, and Mohammadi G. 2022. Cascaded siamese self-supervised audio to video gan// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. New Orleans, USA: IEEE: 4691-4700 [DOI: 10.1109/CVPRW56347.2022.00515]

Aldausari N, Sowmya A, and Mohammadi G. 2024. Multi-head adain audio-to-video motion style transfer// 2024 IEEE International Conference on Robotics And Biomimetics (Robio). Bangkok, Thailand: IEEE: 589-594 [DOI: 10.1109/ROBIO64047.2024.10907605]

Amerini I, Galteri L, Caldelli R, and Del Bimbo A. 2019. Deepfake video detection through optical flow based cnn// Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Seoul, South Korea: IEEE: 0-0 [DOI: 10.1109/ICCVW.2019.00152]

Arjovsky M, Chintala S, and Bottou L. 2017. Wasserstein generative adversarial networks// International Conference on Machine Learning. Sydney, Australia: PMLR: 214-223 [DOI: https://dl.acm.org/doi/10.1145/3746059.3747699]

Bai S, Chen K, Liu X, Wang J, Ge W, Song S, Dang K, Wang P, Wang S, Tang J, Zhong H, Zhu Y, Yang M-H, Li Z, Wan J, Wang P, Ding W, Fu Z, Xu Y, Ye J, Zhang X, Xie T, Cheng Z, Zhang H, Yang Z, Xu H, and Lin J. 2025. Qwen2.5-vl technical report[EB/OL].[2025-02-19]. https://arxiv.org/pdf/2502.13923.pdf

Bammey Q. 2023. Synthbuster: Towards detection of diffusion model

generated images. IEEE Open Journal Of Signal Processing, 5: 1-9 [DOI:10.1109/OJSP.2023.3337714]

Barker J, Bhowmik N, A. Gaus Y F, and Breckon T. 2023. Robust semi-supervised anomaly detection via adversarially learned continuous noise corruption// Proceedings of the 18th International Joint Conference on Computer Vision, Imaging And Computer Graphics Theory And Applications. Lisbon, Portugal: SciTePress: 615-625 [DOI:10.5220/0011684700003417]

Blattmann A, Dockhorn T, Kulal S, Mendelevitch D, Kilian M, Lorenz D, Levi Y, English Z, Voleti V, and Letts A. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets [EB/OL].[2023-11-25]. Https://arxiv.org/pdf/2311.15127.pdf

Blattmann A, Rombach R, Ling H, Dockhorn T, Kim S W, Fidler S, and Kreis K. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Vancouver, Canada: IEEE: 22563-22575 [DOI: 10.1109/CVPR52729.2023.02161]

Brock A, Donahue J, and Simonyan K. 2018. Large scale gan training for high fidelity natural image synthesis[EB/OL].[2018-09-28]. Https://arxiv.org/pdf/1809.11096.pdf

Bruce J, Dennis M D, Edwards A, Parker-Holder J, Shi Y, Hughes E, Lai M, Mavalankar A, Steigerwald R, and Apps C. 2024. Genie: Generative interactive environments// Forty-First International Conference on Machine Learning. Vienna, Austria: PMLR: [DOI:10.5555/3692070.3692255]

Cai R, Zhang H, Liu W, Gao S, and Hao Z. 2021. Appearance-motion memory consistency network for video anomaly detection// AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press:938-946 [DOI:10.1609/AAAI.V35I2.16177]

Cao Y, Zhang J, Frittoli L, Cheng Y, Shen W, and Boracchi G. 2024. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection// European Conference on Computer Vision. Milan, Italy: Springer: 55-72 [DOI: 10.1007/978-3-031-72761-0_4]

Chen B, Li T, and Ding W. 2022. Detecting deepfake videos based on spatiotemporal attention and convolutional lstm. Information Sciences, 601: 58-70 [DOI:10.1016/j.ins.2022.04.014]

Chen H, Xia M, He Y, Zhang Y, Cun X, Yang S, Xing J, Liu Y, Chen Q, and Wang X. 2023a. Videocrafter1: Open diffusion models for high-quality video generation[EB/OL].[2023-10-30]. Https://arxiv.org/pdf/2310.19512.pdf

Chen H, Zhang Y, Cun X, Xia M, Wang X, Weng C, and Shan Y. 2024a. Videocrafter2: Overcoming data limitations for high-quality video diffusion models// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Seattle, USA: IEEE: 7310-7320 [DOI:10.1109/CVPR52733.2024.00698]

Chen S, Yao T, Chen Y, Ding S, Li J, and Ji R. 2021. Local relation learning for face forgery detection// Proceedings of the Aaai Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press: 1081-1088 [DOI:10.1609/aaai.v35i2.16193]

Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, and Abbeel P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets// Advances In Neural Information Processing Systems. Barcelona, Spain: Curran Associates:2172 - 2180 [DOI:10.5555/3157096.3157340]

Chen Y, Yan Z, Cheng G, Zhao K, Lyu S, and Wu B. 2024b. X2-dfd: A framework for explainable and extendable deepfake detection [EB/OL].[2024-10-08]. Https://arxiv.org/pdf/2410.06126.pdf

Chen Z, Wu J, Wang W, Su W, Chen G, Xing S, Zhong M, Zhang Q, Zhu X, Lu L, Li B, Luo P, Lu T, Qiao Y, and Dai J. 2023b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks[EB/OL].[2023-12-21]. https://arxiv.org/pdf/2312.14238.pdf

Clark A, Donahue J, and Simonyan K. 2019. Adversarial video generation on complex datasets[EB/OL].[2019-07-15]. Https://arxiv.org/pdf/1907.06571.pdf

Dang H, Liu F, Stehouwer J, Liu X, and Jain A K. 2020. on the detection of digital face manipulation// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Seattle, USA: IEEE: 5781-5790 [DOI:10.1109/CVPR42600.2020.00582]

Deyun L I U, Ying L I, Zhen Z, and Genlin J I. 2024. Weakly Supervised Video Anomaly Detection Based on Spatiotemporal Dependency and Feature Fusion. Data Collection and Processing, 39(1): 204-214 (柳德云,李莹,周震, and吉根林. 2024. 基于时空依赖关系和特征融合的弱监督视频异常检测. 数据采集与处理, 39(1): 204-214) [DOI:10.16337/j.1004-9037.2024.01.018]

Di J, Huicheng L a I, and Liejun W. 2025. Video Anomaly Detection Based on Cross-Modal Fusion and Hyperbolic Graph Attention Mechanism. Journal of Communications, 46(6): 136-152 (姜迪, 赖惠成, and汪烈军. 2025. 基于跨模态融合与双曲图注意力机制的视频异常检测. 通信学报, 46(6): 136-152) [DOI: 10.11959/j.issn.1000-436x.2025110]

Ding C, Pang G, and Shen C. 2022. Catching both gray and black swans: Open-set supervised anomaly detection// 2022 IEEE/CVF Conference on Computer Vision And Pattern Recognition (CVPR). New Orleans, USA: IEEE: 7378-7388 [DOI: 10.1109/cvpr52688.2022.00724]

Dong H, Frusque G, Zhao Y, Chatzi E, and Fink O. 2024. Nng-mix: Improving semi-supervised anomaly detection with pseudo-anomaly generation. IEEE Transactions On Neural Networks And Learning Systems, 36 (6): 10635 - 10647 [DOI: 10.1109/TNNLS.2024.3497801]

Dong X, Bao J, Chen D, Zhang T, Zhang W, Yu N, Chen D, Wen F, and Guo B. 2022. Protecting celebrities from deepfake with identity consistency transformer// Proceedings of the IEEE/CVF Conference

on Computer Vision And Pattern Recognition. New Orleans, USA: IEEE: 9468-9478 [DOI:10.1109/CVPR52688.2022.00925]

Durall R, Keuper M, and Keuper J. 2020. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Seattle, USA: IEEE: 7890-7899 [DOI:10.1109/CVPR42600.2020.00791]

Durani W, Nitzl T, Plant C, and Böhm C. Weakly supervised anomaly detection via dual-tailed kernel// Forty-Second International Conference on Machine Learning. Vancouver, Canada: PMLR: 1-34 [DOI:None]

Fan Y, Wen G, Li D, Qiu S, Levine M D, and Xiao F. 2020. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. Computer Vision And Image Understanding, 195: 102920 [DOI:10.1016/j.cviu.2020.102920]

Fangyuan G U O, and Genlin J I. 2024. Video Anomaly Detection Method Based on Dual Discriminators and Pseudo Video Generation. Computer Science, 51(8): 217-223 (郭方圆, and 吉根林. 2024.基于双鉴别器和伪视频生成的视频异常检测方法.计算机科学, 51(8): 217-223) [DOI:10.11896/jsjkx.230600148]

Fei J, Dai Y, Yu P, Shen T, Xia Z, and Weng J. 2022. Learning second order local anomaly for general face forgery detection// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. New Orleans, United States: IEEE: 20270-20280 [DOI:10.1109/CVPR52688.2022.01963]

Feng X, Yu H, Wu M, Hu S, Chen J, Zhu C, Wu J, Chu X, and Huang K. 2025. NarrLV: Towards a Comprehensive Narrative-Centric Evaluation for Long Video Generation[EB/OL]. https://arxiv.org/pdf/2507.11245v4.pdf

Frank J, Eisenhofer T, Schönherr L, Fischer A, Kolossa D, and Holz T. 2020. Leveraging frequency analysis for deep fake image recognition// International Conference on Machine Learning. Vienna, Austria: PMLR: 3247-3258 [DOI:10.5555/3524938.3525242]

Gao J, Tao C, Sun Z, Jiang X, and Ma S. 2025a. Semi-supervised anomaly detection through denoising-aware contrastive distance learning// Proceedings of the Acm on Web Conference 2025. Sydney, Australia: ACM: 2111-2119 [DOI: 10.1145/3696410.3714626]

Gao Y, Chang D, Yu B, Qin H, Chen L, Liang K, and Ma Z. 2025b. Fakereasoning: Towards generalizable forgery detection and reasoning[EB/OL].[2025-03-27]. Https://arxiv.org/pdf/2503.21210.pdf

Geng Z, Yang B, Hang T, Li C, Gu S, Zhang T, Bao J, Zhang Z, Li H, and Hu H. 2024. Instructdiffusion: A generalist modeling interface for vision tasks// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Seattle, USA: IEEE: 12709-12720 [DOI:10.1109/CVPR52733.2024.01208]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, and Bengio Y. 2014. Generative adversarial nets// Advances In Neural Information Processing Systems. Montreal, Quebec: Curran Associates: 2672 - 2680 [DOI: 10.5555/2969033.2969125]

Gu Z, Zhu B, Zhu G, Chen Y, Li H, Tang M, and Wang J. 2024. Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization// Proceedings of the 32nd Acm International Conference on Multimedia. Melbourne, Australia: ACM: 2041-2049 [DOI:10.1145/3664647.3680685]

Güera D, and Delp E J. 2018. Deepfake video detection using recurrent neural networks// 2018 15th IEEE International Conference on Advanced Video And Signal Based Surveillance (AVSS). Auckland, New Zealand: IEEE: 1-6 [DOI: 10.1109/AVSS. 2018.8639163]

Guo X, Song X, Zhang Y, Liu X, and Liu X. 2025. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector// Proceedings of the Computer Vision And Pattern Recognition Conference. Nashville, United States: IEEE: 105-116 [DOI:10.1109/CVPR52734.2025.00019]

Guo Y, Yang C, Rao A, Liang Z, Wang Y, Qiao Y, Agrawala M, Lin D, and Dai B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning[EB/OL].[2023-07-10]. Https://arxiv.org/pdf/2307.04725.pdf

Hao L Y U, Peng-Fei Y I, Rui L I U, Dong-Sheng Z, Qiang Z, and Xiao-Peng W E I. 2022. Temporal Multi-Scale Autoencoder for Video Anomaly Detection. Journal of Graphics, 43(2): 223-229 (吕浩,易鹏飞,刘瑞,周东生,张强, and 魏小鹏. 2022.用于视频异常检测的时序多尺度自编码器.图学学报, 43(2): 223-229) [DOI:10.11996/JG.j.2095-302X.2022020223]

Hinami R, Mei T, and Satoh S I. 2017. Joint detection and recounting of abnormal events by learning deep generic knowledge// 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE:3639-3647 [DOI:10.1109/iccv.2017.391]

Ho J, Jain A, and Abbeel P. 2020. Denoising diffusion probabilistic models// Advances In Neural Information Processing Systems. OnLine: 6840-6851 [DOI:10.5555/3495724.3496298]

Ho J, Salimans T, Gritsenko A, Chan W, Norouzi M, and Fleet D J. 2022. Video diffusion models// Advances In Neural Information Processing Systems. New Orleans, USA: Curran Associates: 8633-8646 [DOI:10.5555/3600270.3600898]

Hong W, Ding M, Zheng W, Liu X, and Tang J. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers [EB/OL].[2022-05-29]. Https://arxiv.org/pdf/2205.15868.pdf

Hooda A, Mangaokar N, Feng R, Fawaz K, Jha S, and Prakash A. 2024. D4: Detection of adversarial diffusion deepfakes using disjoint ensembles// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, United States: IEEE: 3812-3822 [DOI:10.1109/WACV57701.2024.00377]

Huang C, Ye F, Zhao P, Zhang Y, Wang Y-F, and Tian Q. 2020. Esad: End-to-end deep semi-supervised anomaly detection[EB/OL].[2020-12-09]. Https://arxiv.org/pdf/2012.04905.pdf

Huang K, Wu M, Chen H, Feng X, and Zhang D. 2025a. The three realms of visual turing: from seeing to imagining in the LLM era. Journal of Graphics, 46(5): 919-930 (黄凯奇,武美奇,陈宏昊, 丰效坤, and 张岱凌. 2025a. 视觉图灵三境界:大模型时代下视觉智能进展与展望. 图学学报, 46(5): 919-930)[DOI: 10.11996/JG.j.2095-302X.2025050919]

Huang K, Zhao X, Li Q, and Hu S. 2021. Visual Turing: the next development of computer vision in the view of human-computer gaming. Journal of Graphics, 42(3): 339-348 (黄凯奇,鑫 赵,李乔哲, and 胡世宇. 2021. 视觉图灵:从人机对抗看计算机视觉下一步发展. 图学学报, 42(3): 339-348)[DOI: 10.11996/JG.j.2095-302X.2021030339]

Huang Q, Xu Z, Zhang X, and Zhang J. 2025b. Unishield: An adaptive multi-agent framework for unified forgery image detection and localization[EB/OL].[2025-10-03]. Https://arxiv.org/pdf/2510.03161.pdf

Huang T-M, Lin W-T, Hua K-L, Cheng W-H, Yamagishi J, and Chen J-C. 2025c. Thinkfake: Reasoning in multimodal large language models for ai-generated image detection[EB/OL].[2025-09-24]. Https://arxiv.org/pdf/2509.19841.pdf

Huang Z, Xia B, Lin Z, Mou Z, Yang W, and Jia J. 2024. Ffaa: Multimodal large language model based explainable open-world face forgery analysis assistant[EB/OL].[2024-08-19]. Https://arxiv.org/pdf/2408.10072.pdf

Jeong Y, Kim D, Min S, Joe S, Gwon Y, and Choi J. 2022a. Bihpf: Bilateral high-pass filters for robust deepfake detection// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, United States: EEE: 48-57 [DOI: 10.1109/WACV51458.2022.00293]

Jeong Y, Kim D, Ro Y, and Choi J. 2022b. Frepgan: Robust deepfake detection using frequency-level perturbations// Proceedings of the Aaai Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press: 1060-1068 [DOI: 10.1609/aaai.v36i1.19990]

Jezequel L, Vu N-S, Beaudet J, and Histace A. 2022. Semi-supervised anomaly detection with contrastive regularization// 2022 26th International Conference on Pattern Recognition (Icpr). Montréal, Canada: IEEE: 2664-2671 [DOI: 10.1109/icpr56361.2022.9956091]

Jia-Xu L, Ming-Pi T a N, Bo H U, and Xin-Bo G a O. 2022. Video anomaly detection based on implicit perspective transformation. Computer Science, 49(2): 142-148 (冷佳旭,谭明圮,胡波, and 高新波. 2022. 基于隐式视角转换的视频异常检测. 计算机科学, 49(2): 142-148)[DOI: 10.11896/jsjkx.210900266]

Jia C, Yang Y, Xia Y, Chen Y-T, Parekh Z, Pham H, Le Q, Sung Y-H, Li Z, and Duerig T. 2021a. Scaling up visual and vision-language representation learning with noisy text supervision// International Conference on Machine Learning. Vienna, Austria: PMLR: 4904-4916 [DOI: 10.48550/arXiv.2102.05918]

Jia G, Zheng M, Hu C, Ma X, Xu Y, Liu L, Deng Y, and He R. 2021b. Inconsistency-aware wavelet dual-branch network for face forgery detection. IEEE Transactions On Biometrics, Behavior, And Identity Science, 3 (3): 308-319 [DOI: 10.1109/TBIOM.2021.3086109]

Jia S, Lyu R, Zhao K, Chen Y, Yan Z, Ju Y, Hu C, Li X, Wu B, and Lyu S. 2024. Can ChatGPT detect deepfakes? A study of using multimodal large language models for media forensics// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Seattle, United States: IEEE: 4324-4333 [DOI: 10.1109/CVPRW63382.2024.00436]

Jiale D, Zhengjie D, Xiyan L, and Shiyun W. 2025. A Deepfake Detection Method Based on Multi-Feature Fusion in Frequency and Spatial Domains. Journal of Graphics, 46(1): 104-113 (董佳乐,邓正杰,李喜艳, and 王诗韵. 2025. 基于频域和空域多特征融合的深度伪造检测方法. 图学学报, 46(1): 104-113)[DOI: 10.11996/JG.j.2095-302X.2025010104]

Jiang A Q, Sablayrolles A, Mensch A, Bamford C, Chaplot D S, Casas D D L, Bressand F, Lengyel G, Lample G, Saulnier L, Lavaud L R, Lachaux M-A, Stock P, Scao T L, Lavril T, Wang T, Lacroix T, and Sayed W E. 2023. Mistral 7b[EB/OL].[2023-10-10]. https://arxiv.org/pdf/2310.06825.pdf

Jin Y, Peng J, He Q, Hu T, Chen H, Wu J, Zhu W, Chi M, Liu J, Wang Y, and Wang C. 2024. Dualanodiff: Dual-interrelated diffusion model for few-shot anomaly image generation[EB/OL].[2024-08-24]. https://arxiv.org/pdf/2408.13509.pdf

Jing Z, Pan X, Wenjun L, Xiaoxuan G, and Fang S. 2025. Cross-domain Face Forgery Detection with Diverse Negative Sample Generation. Journal of Image and Graphics, 30(2): 421-434 (张晶,许盼,刘文君,郭晓萱, and 孙芳. 2025. 多样性负实例生成的跨域人脸伪造检测. 中国图象图形学报, 30(2): 421-434)[DOI: 10.11834/jig.240160]

Joo H K, Vo K, Yamazaki K, and Le N. 2023. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection// 2023 IEEE International Conference on Image Processing (LCIP). Kuala Lumpur, Malaysia: IEEE: 3230-3234 [DOI: 10.1109/icip49359.2023.10222289]

Kang H, Wen S, Wen Z, Ye J, Li W, Feng P, Zhou B, Wang B, Lin D, and Zhang L. 2025. Legion: Learning to ground and explain for synthetic image detection[EB/OL].[2025-03-19]. Https://arxiv.org/pdf/2503.15264.pdf

Karim H, Doshi K, and Yilmaz Y. 2024. Real-time weakly supervised video anomaly detection// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 6848-6856 [DOI: 10.1109/WACV57701.2024.00670]

Karras T, Aila T, Laine S, and Lehtinen J. 2017. Progressive growing of gans for improved quality, stability, and variation[EB/OL].[2017-10-27]. Https://arxiv.org/pdf/1710.10196.pdf

Karras T, Laine S, and Aila T. 2019. A style-based generator architecture for generative adversarial networks// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Long Beach, United States: IEEE: 4401-4410 [DOI: 10.1109/CVPR.2019.00453]

Kascenas A, Sanchez P, Schrempf P, Wang C, Clackett W, Mikhael S S, Voisey J P, Goatman K, Weir A, Pugeault N, Tsaftaris S A, and O'neil A Q. 2023. The role of noise in denoising models for anomaly detection in medical images. Med Image Anal, 90: 102963 [DOI:10.1016/j.media.2023.102963]

Kim D, Park C, Cho S, Lim H, Kang M, Lee J, and Lee S. 2025. Genclip: Generalizing clip prompts for zero-shot anomaly detection[EB/OL].[2025-04-21]. Https://arxiv.org/pdf/2504.14919.pdf

Kim M, Kim J, Yu J, and Choi J K. 2023. Active anomaly detection based on deep one-class classification. Pattern Recognition Letters, 167: 18-24 [DOI:10.1016/j.patrec.2022.12.009]

Kong W, Tian Q, Zhang Z, Min R, Dai Z, Zhou J, Xiong J, Li X, Wu B, and Zhang J. 2024. Hunyuanvideo: A systematic framework for large video generative models[EB/OL].[2024-12-03]. Https://arxiv.org/pdf/2412.03603.pdf

Li F, Liu W, Chen J, Zhang R, Wang Y, Zhong X, and Wang Z. 2025a. Anomize: Better open vocabulary video anomaly detection// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 29203-29212 [DOI: 10.48550/arXiv.2503.18094]

Li J, Xie H, Li J, Wang Z, and Zhang Y. 2021. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Virtual, Online: IEEE: 6458-6467 [DOI:10.1109/CVPR46437.2021.00639]

Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, and Guo B. 2020. Face x-ray for more general face forgery detection// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Seattle, USA: IEEE: 5001-5010 [DOI: 10.1109/CVPR42600.2020.00505]

Li N, Wang D, and Nie X. 2025b. Osad: Open-set supervised anomaly detection in surveillance videos based on margin metric learning. IEEE Transactions On Circuits And Systems For Video Technology: 1-1 [DOI:10.1109/tcsvt.2025.3614193]

Li T, Huang Z, Wen H, He Y, Lyu S, Wu B, and Cheng G. 2025c. Raidx: A retrieval-augmented generation and grpo reinforcement learning framework for explainable deepfake detection [EB/OL].[2025-08-06]. Https://arxiv.org/pdf/2508.04524.pdf

Li W, Wu K, Sun Y, Jiao C, and Xiong S. 2023a. Weakly Supervised Video Anomaly Detection Based on Contrastive Memory Network. Computer Application Research, 40(10): 3162-3167, 3172 (李文中,吴克伟,孙永宣,焦畅, and熊思璇. 2023a.基于对比记忆网络的弱监督视频异常检测.计算机应用研究, 40(10): 3162-3167), 3172) [DOI:10.19734/j.issn.1001-3695.2022.12.0829]

Li X, Tan X, Chen Z, Zhang Z, Zhang R, Guo R, Jiang G, Chen Y, Qu Y, and Ma L. 2025d. One-for-more: Continual diffusion model for anomaly detection// Proceedings of the Computer Vision And Pattern Recognition Conference. Nashville, United States: IEEE: 4766-4775 [DOI:10.48550/arXiv.2502.19848]

Li X, Xiao C, Feng Z, Pang S, Tai W, and Zhou F. 2024a. Controlled graph neural networks with denoising diffusion for anomaly detection. Expert Systems With Applications, 237 (Part B): 121533 [DOI:10.1016/j.eswa.2023.121533]

Li X, Zhang Z, Tan X, Chen C, Qu Y, Xie Y, and Ma L. 2024b. Promptad: Learning prompts with only normal samples for few-shot anomaly detection// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Seattle, USA: IEEE: 16838-16848 [DOI:10.1109/CVPR52733.2024.01594]

Li Y, Gan Z, Shen Y, Liu J, Cheng Y, Wu Y, Carin L, Carlson D, and Gao J. 2019. Storygan: A sequential conditional gan for story visualization// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Long Beach, USA: IEEE: 6329-6338 [DOI:10.1109/CVPR.2019.01014]

Li Y, and Lyu S. 2018. Exposing deepfake videos by detecting face warping artifacts[EB/OL].[2018-11-01]. Https://arxiv.org/pdf/1811.00656.pdf

Li Z, Zhu Y, and Van Leeuwen M. 2023b. A survey on explainable anomaly detection. Acm Transactions On Knowledge Discovery From Data, 18 (1): 1-54 [DOI:10.1145/3609333]

Liang J, Li T, Yang J, Li Y, Fang Z, and Yang F. 2023. Fusion of Self-Attention and Autoencoder for Video Anomaly Detection. Journal of Image and Graphics, 28(4): 1029-1040 (梁家菲,李婷,杨佳琪,李亚楠,方智文, and杨丰. 2023.融合自注意力和自编码器的视频异常检测.中国图象图形学报, 28(4): 1029-1040) [DOI:10.11834/jig.211147]

Lin B, Ge Y, Cheng X, Li Z, Zhu B, Wang S, He X, Ye Y, Yuan S, and Chen L. 2024. Open-sora plan: Open-source large video generation model[EB/OL].[2024-11-28]. Https://arxiv.org/pdf/2412.00131.pdf

Lin S, and Yang X. 2024. Animatediff-lightning: Cross-model diffusion distillation[EB/OL].[2024-03-19]. Https://arxiv.org/pdf/2403.12706.pdf

Ling X, Zhu C, Wu M, Li H, Feng X, Yang C, Hao A, Zhu J, Wu J, and Chu X. 2025. Vmbench: A benchmark for perception-aligned video motion generation// Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 13087-13098 [DOI:10.48550/arXiv.2503.10076]

Liu C, Xue R, Shi L, Li Y, and Gao Y. 2022a. Video Anomaly Detection Using Generative Adversarial Networks with Integrated Gated Self-Attention Mechanism. Journal of Image and Graphics, 27(11): 3210-3221 (刘成明, 薛然, 石磊, 李英豪, and 高宇飞. 2022a. 融合门控自注意力机制的生成对抗网络视频异常检测. 中国图象图形学报, 27(11): 3210-3221) [DOI: 10.11834/jig.210520]

Liu H, Li C, Li Y, and Lee Y J. 2024a. Improved baselines with visual instruction tuning// 2024 IEEE/CVF Conference on Computer Vision And Pattern Recognition (CVPR). Seattle, USA: IEEE: 26286-26296 [DOI: 10.1109/cvpr52733.2024.02484]

Liu H, Li C, Wu Q, and Lee Y J. 2023. Visual instruction tuning// Advances in Neural Information Processing Systems (NeurIPS). New Orleans, USA: Curran Associates, Inc.: 34892 - 34916 [DOI: 10.5555/3666122.3667638]

Liu H, Li X, Zhou W, Chen Y, He Y, Xue H, Zhang W, and Yu N. 2021a. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Nashville, USA: IEEE: 772-781 [DOI: 10.1109/CVPR46437.2021.00732]

Liu H, Tan Z, Tan C, Wei Y, Wang J, and Zhao Y. 2024b. Forgery-aware adaptive transformer for generalizable synthetic image detection// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Seattle, USA: IEEE: 10770-10780 [DOI: 10.1109/CVPR52733.2024.01024]

Liu J, Zhang F, Zhu J, Sun E, Zhang Q, and Zha Z-J. 2024c. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization[EB/OL].[2024-10-14]. Https://arxiv.org/pdf/2410.10238.pdf

Liu Y, Li Z, Pan S, Gong C, Zhou C, and Karypis G. 2022b. Anomaly detection on attributed networks via contrastive self-supervised learning. IEEE Trans Neural Netw Learn Syst, 33(6): 2378-2392 [DOI: 10.1109/TNNLS.2021.3068344]

Liu Y, Liu J, Yang K, Ju B, Liu S, Wang Y, Yang D, Sun P, and Song L. 2024d. Amp-net: Appearance-motion prototype network assisted automatic video anomaly detection system. IEEE Transactions On Industrial Informatics, 20(2): 2843-2855 [DOI: 10.1109/tii.2023.3298476]

Liu Y, Zhang K, Li Y, Yan Z, Gao C, Chen R, Yuan Z, Huang Y, Sun H, and Gao J. 2024e. Sora: A review on background, technology, limitations, and opportunities of large vision models[EB/OL].[2024-02-27]. Https://arxiv.org/pdf/2402.17177.pdf

Liu Z, Nie Y, Long C, Zhang Q, and Li G. 2021b. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 13568-13577 [DOI: 10.1109/iccv48922.2021.01333]

Liu Z, Wu X, Wu J, Wang X, and Yang L. 2025. Language-guided open-world video anomaly detection[EB/OL].[2025-03-17]. https://arxiv.org/pdf/2503.13160.pdf

Luo W, Cao Y, Yao H, Zhang X, Lou J, Cheng Y, Shen W, and Yu W. 2025. Exploring intrinsic normal prototypes within a single image for universal anomaly detection// Proceedings of the Computer Vision And Pattern Recognition Conference. Nashville, USA: IEEE: 9974-9983 [DOI: 10.1109/CVPR52734.2025.00932]

Luo W, Liu W, and Gao S. 2017. Remembering history with convolutional lstm for anomaly detection// 2017 IEEE International Conference on Multimedia And Expo (ICME). Hong Kong, China: IEEE: 439-444 [DOI: 10.1109/icme.2017.8019325]

Luo Y, Zhang Y, Yan J, and Liu W. 2021. Generalizing face forgery detection with high-frequency features// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Nashville, USA: IEEE: 16317-16326 [DOI: 10.1109/CVPR46437.2021.01605]

Ma R, Duan J, Kong F, Shi X, and Xu K. 2023. Exposing the fake: Effective diffusion-generated images detection [EB/OL]. [2023-07-12]. Https://arxiv.org/pdf/2307.06272.pdf

Masi I, Killekar A, Mascarenhas R M, Gurudatt S P, and Abdalmageed W. 2020. Two-branch recurrent network for isolating deepfakes in videos// European Conference on Computer Vision. Glasgow, United Kingdom: Springer: 667-684 [DOI: 10.1007/978-3-030-58571-6_39]

Miao C, Tan Z, Chu Q, Yu N, and Guo G. 2022. Hierarchical frequency-assisted interactive networks for face manipulation detection. IEEE Transactions On Information Forensics And Security, 17: 3008-3021 [DOI: 10.1109/TIFS.2022.3198275]

Mirza M, and Osindero S. 2014. Conditional generative adversarial nets [EB/OL].[2014-11-6]. Https://arxiv.org/pdf/1411.1784.pdf

Nguyen T N, and Meunier J. 2019. Anomaly detection in video sequence with appearance-motion correspondence// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE: 1273-1283 [DOI: 10.1109/iccv.2019.00136]

Nguyen X H, Tran T S, Nguyen K D, and Truong D-T. 2021. Learning spatio-temporal features to detect manipulated facial videos created by the deepfake techniques. Forensic Science International: Digital Investigation, 36: 301108 [DOI: 10.1016/j.fsidi.2021.301108]

Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, Mcgrew B, Sutskever I, and Chen M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models[EB/OL].[2021-12-20]. Https://arxiv.org/pdf/2112.10741.pdf

Odena A, Olah C, and Shlens J. 2017. Conditional image synthesis with auxiliary classifier gans// International Conference on Machine Learning. Sydney, Australia: PMLR: 2642-2651 [DOI: 10.5555/

3305890.3305954]

Ojha U, Li Y, and Lee Y J. 2023. Towards universal fake image detectors that generalize across generative models// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Vancouver, Canada: IEEE: 24480-24489 [DOI: 10.1109/CVPR52729.2023.02345]

Openai. 2023. Gpt-4 technical report[EB/OL].[2023-03-15]. https://arxiv.org/pdf/2303.08774.pdf

Oquab M, Darcet T, Moutakanni T, Vo H V, Szafraniec M, Khalidov V, Fernandez P, Haziza D, Massa F, El-Nouby A, Assran M, Ballas N, Galuba W, Howes R, Huang P-Y, Li S-W, Misra I, Rabbat M, Sharma V, Synnaeve G, Xu H, Jégou H, Mairal J, Labatut P, Joulin A, and Bojanowski P. 2024. Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn. Res., 2024 (1): 1-32 [DOI:10.48550/arXiv.2304.07193]

Pang G, Shen C, Jin H, and Van Den Hengel A. 2023. Deep weakly-supervised anomaly detection// Proceedings of the 29th Acm Sigkdd Conference on Knowledge Discovery And Data Mining. Long Beach (California), USA: ACM: 1795-1807 [DOI: 10.1145/3580305.3599302]

Pang G, Shen C, and Van Den Hengel A. 2019. Deep anomaly detection with deviation networks// Proceedings of the 25th Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. Anchorage, USA: ACM: 353-362 [DOI: 10.1145/3292500.3330871]

Pang G, Yan C, Shen C, Hengel A V D, and Bai X. 2020. Self-trained deep ordinal regression for end-to-end video anomaly detection// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 12170-12179 [DOI: 10.1109/CVPR42600.2020.01219]

Peebles W, and Xie S. 2023. Scalable diffusion models with transformers// Proceedings of the IEEE/CVF international conference on computer vision. Paris, France: IEEE: 4195-4205 [DOI: 10.1109/ICCV51070.2023.00387]

Peng Y, Li X, Liang Z, and Wang Y. 2025. Qsco: A quantum scoring module for open-set supervised anomaly detection// Proceedings of the Aaai Conference on Artificial Intelligence. Philadelphia, United States: AAAI Press: 19884-19894 [DOI: 10.1609/aaai.v39i19.34190]

Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J, Penna J, and Rombach R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis[EB/OL].[2023-07-04]. Https://arxiv.org/pdf/2307.01952.pdf

Qi X, Zeng J, and Ji G. 2023. Dual Cross-Attention Autoencoder Improves Video Anomaly Detection. Journal of Nanjing Normal University (Natural Science Edition), 46(1): 110-119 (戚小莎, 曾静, and 吉根林. 2023. 双交叉注意力自编码器改进视频异常检测. 南京师大学报(自然科学版), 46(1): 110-119) [DOI: 10.3969/j.issn.1001-4616.2023.01.015]

Qian Y, Yin G, Sheng L, Chen Z, and Shao J. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues// European Conference on Computer Vision. Glasgow, UK: Springer: 86-103 [DOI:10.1007/978-3-030-58610-2_6]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, and Clark J. 2021. Learning transferable visual models from natural language supervision// International Conference on Machine Learning. Vienna, Austria: PmLR: 8748-8763 [DOI:10.48550/arXiv.2103.00020]

Radford A, Metz L, and Chintala S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks [EB/OL].[2015-11-19]. Https://arxiv.org/pdf/1511.06434.pdf

Ramesh A, Dhariwal P, Nichol A, Chu C, and Chen M. 2022. Hierarchical text-conditional image generation with clip latents[EB/OL]. [2022-04-13]. Https://arxiv.org/pdf/2204.06125.pdf

Reiss T, and Hoshen Y. 2022. Attribute-based representations for accurate and interpretable video anomaly detection [EB/OL]. [2022-12-01]. https://arxiv.org/pdf/2212.00789.pdf

Rombach R, Blattmann A, Lorenz D, Esser P, and Ommer B. 2022. High-resolution image synthesis with latent diffusion models// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. New Orleans, United States: IEEE: 10684-10695 [DOI:10.1109/CVPR52688.2022.01042]

Ruff L, Vandermeulen R A, Görnitz N, Binder A, Müller E, Müller K-R, and Kloft M. 2019. Deep semi-supervised anomaly detection [EB/OL].[2019-06-06]. Https://arxiv.org/pdf/1906.02694.pdf

Sabir E, Cheng J, Jaiswal A, Abdalmageed W, Masi I, and Natarajan P. 2019. Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (Gui), 3 (1): 80-87 [DOI: 10.48550/arXiv.1905.00582]

Sabokrou M, Fayyaz M, Fathy M, Moayed Z, and Klette R. 2018. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Computer Vision And Image Understanding, 172: 88-97 [DOI:10.1016/j.cviu.2018.02.006]

Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E L, Ghasemipour K, Gontijo Lopes R, Karagol Ayan B, and Salimans T. 2022. Photorealistic text-to-image diffusion models with deep language understanding// Advances In Neural Information Processing Systems. New Orleans, USA: Curran Associates: 36479-36494 [DOI: 10.5555/3600270.3602913]

Saito M, Saito S, Koyama M, and Kobayashi S. 2020. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. International Journal Of Computer Vision, 128 (10): 2586-2606 [DOI:10.1007/s11263-020-01333-y]

Salehi A, Salehi M, Hosseini R, Snoek C G, Yamada M, and Sabokrou

M. 2025. Crane: Context-guided prompt learning and attention refinement for zero-shot anomaly detections [EB/OL]. [2025-04-15]. Https://arxiv.org/pdf/2504.11055.pdf

Schusterbauer J, Gui M, Fundel F, and Ommer B. 2025. Diff2Flow: Training Flow Matching Models via Diffusion Model Alignment// Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, Tennessee, USA: IEEE: 28347-28357 [DOI: 10.1109/CVPR52734.2025.02640]

Shan-Wu Y a N, Hong-Bing X, Yu W, and Mei S U N. 2023. Video anomaly detection by integrating spatiotemporal pedestrian information. Journal of Graphics, 44(1): 95-103 (闫善武,肖洪兵,王瑜, and孙梅. 2023.融合行人时空信息的视频异常检测.图学学报, 44(1): 95-103) [DOI:10.11996/JG.j.2095-302X.2023010095]

Shao-Nian H, Pei-Ran W E N, Qi Q, and Rong-Yuan C. 2023. Lightweight Video Anomaly Detection Based on Multi-Branch Aggregated Frame Prediction. Journal of Graphics, 44(6): 1173-1182 (黄少年,文沛然,全琪, and陈荣元. 2023.基于多支路聚合的帧预测轻量化视频异常检测.图学学报, 44(6): 1173-1182) [DOI:10.11996/JG.j.2095-302X.2023061173]

Sharma R, Mashkaria S, and Awate S P. 2022. A semi-supervised generalized vae framework for abnormality detection using one-class classification// 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Bordeaux, France: IEEE: 1302-1310 [DOI:10.1109/wacv51458.2022.00137]

Shehnepoor S, Togneri R, Liu W, and Bennamoun M. 2021. Scoregan: A fraud review detector based on regulated gan with data augmentation. IEEE Transactions On Information Forensics And Security, 17: 280-291 [DOI:10.1109/TIFS.2021.3139771]

Shi Y, Gao Y, Lai Y, Wang H, Feng J, He L, Wan J, Chen C, Yu Z, and Cao X. 2025. Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. Visual Intelligence, 3 (1): 9 [DOI: 10.1007/s44267-025-00079-w]

Si-Qian W E I, Gen-Lin J I, Zhen X U, and Yu-Jie L I U. 2022. Video Anomaly Detection Using Multi-Instance Learning with Attention Mechanism. Small and Microcomputer Systems, 43(12): 2575-2579 (魏思倩,吉根林,许振, and刘宇杰. 2022.利用注意力机制的多示例学习视频异常检测.小型微型计算机系统, 43(12): 2575-2579) [DOI:10.20009/j.cnki.21-1106/TP.2021-0398]

Song J, Meng C, and Ermon S. 2020. Denoising diffusion implicit models[EB/OL].[2020-10-06]. Https://arxiv.org/pdf/2010.02502.pdf

Sultani W, Chen C, and Shah M. 2018. Real-world anomaly detection in surveillance videos// Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition. Salt Lake City, USA: IEEE: 6479-6488 [DOI:10.1109/CVPR.2018.00678]

Sun C, Myers A, Vondrick C, Murphy K, and Schmid C. 2019. Videobert: A joint model for video and language representation learning// Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE: 7464-7473 [DOI: 10.1109/ICCV.2019.00756]

Sun K, Chen S, Yao T, Zhou Z, Ji J, Sun X, Lin C-W, and Ji R. 2025. Towards general visual-linguistic face forgery detection// Proceedings of the Computer Vision And Pattern Recognition Conference. Nashville, USA: IEEE: 19576-19586 [DOI: 10.1109/CVPR52734.2025.01823]

Tan C, Liu P, Tao R, Liu H, Zhao Y, Wu B, and Wei Y. 2024a. Data-independent operator: A training-free artifact representation extractor for generalizable deepfake detection[EB/OL].[2024-03-11]. Https://arxiv.org/pdf/2403.06803.pdf

Tan C, Zhao Y, Wei S, Gu G, Liu P, and Wei Y. 2024b. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning// Proceedings of the Aaai Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press: 5052-5060 [DOI:10.1609/aaai.v38i5.28310]

Tan C, Zhao Y, Wei S, Gu G, Liu P, and Wei Y. 2024c. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Seattle, United States: IEEE: 28130-28139 [DOI: 10.1109/CVPR52733.2024.02657]

Tan C, Zhao Y, Wei S, Gu G, and Wei Y. 2023. Learning on gradients: Generalized artifacts representation for gan-generated images detection// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Vancouver, Canada: IEEE: 12105-12114 [DOI:10.1109/CVPR52729.2023.01165]

Tan H, Lan J, Tan Z, Liu A, Song C, Shi S, Zhu H, Wang W, Wan J, and Lei Z. 2025. Veritas: Generalizable deepfake detection via pattern-aware reasoning[EB/OL].[2025-08-28]. Https://arxiv.org/pdf/2508.21048.pdf

Team G, Georgiev P, Lei V I, Burnell R, Bai L, Gulati A, Tanzer G, Vincent D, Pan Z, and Wang S. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context[EB/OL].[2024-03-08]. Https://arxiv.org/pdf/2403.05530.pdf

Tian Q, Chen Y, Zhang Z, Lu H, Chen L, Xie L, and Liu S. 2020. Tfgan: Time and frequency domain based generative adversarial network for high-fidelity speech synthesis[EB/OL].[2020-11-24]. Https://arxiv.org/pdf/2011.12206.pdf

Tian S, Dong J, Li J, Zhao W, Xu X, Song B, Meng C, Zhang T, and Chen L. 2023. Sad: Semi-supervised anomaly detection on dynamic graphs[EB/OL].[2023-05-23]. Https://arxiv.org/pdf/2305.13573.pdf

Tian Y, Ren J, Chai M, Olszewski K, Peng X, Metaxas D N, and Tulyakov S. 2021. A good image generator is what you need for high-resolution video synthesis[EB/OL].[2021-04-30]. Https://arxiv.org/pdf/2104.15069.pdf

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, and Lample G. 2023. Llama: Open and efficient foundation language models[EB/OL].[2023-02-27]. https://arxiv.org/pdf/2302.13971.pdf

Tulyakov S, Liu M-Y, Yang X, and Kautz J. 2018. Mocogan: Decomposing motion and content for video generation// Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition. Salt Lake City, USA: IEEE: 1526-1535 [DOI: 10.1109/CVPR.2018.00165]

Vondrick C, Pirsiavash H, and Torralba A. 2016. Generating videos with scene dynamics// Advances in neural information processing systems: Curran Associates, Inc.: 613 - 621 [DOI: 10.4018/IJSI.309732]

Wan B, Fang Y, Xia X, and Mei J. 2020. Weakly supervised video anomaly detection via center-guided discriminative learning// 2020 IEEE International Conference on Multimedia And Expo (ICME). Shenzhen, China: IEEE: 1-6 [DOI: 10.1109/icme46284.2020.9102722]

Wan T, Wang A, Ai B, Wen B, Mao C, Xie C-W, Chen D, Yu F, Zhao H, and Yang J. 2025. Wan: Open and advanced large-scale video generative models[EB/OL].[2025-03-26]. Https://arxiv.org/pdf/2503.20314.pdf

Wang C, Zhu H, Peng J, Wang Y, Yi R, Wu Y, Ma L, and Zhang J. 2025a. M3dm-nr: Rgb-3d noisy-resistant industrial anomaly detection via multimodal denoising. IEEE Trans Pattern Anal Mach Intell, 47 (11): 9981-9993 [DOI: 10.1109/TPAMI.2025.3592089]

Wang F, Zhang T, Wang Y, Qiu Y, Liu X, Guo X, and Cui Z. 2025b. Distribution prototype diffusion learning for open-set supervised anomaly detection// Proceedings of the Computer Vision And Pattern Recognition Conference. Nashville, United States: IEEE: 20416-20426 [DOI: 10.1109/CVPR52734.2025.01901]

Wang G, Zhan Y, Wang X, Song M, and Nahrstedt K. 2022a. Hierarchical semi-supervised contrastive learning for contamination-resistant anomaly detection// European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer: 110-128 [DOI: 10.1007/978-3-031-19806-9_7]

Wang J, Chen D, Wu Z, Luo C, Zhou L, Zhao Y, Xie Y, Liu C, Jiang Y-G, and Yuan L. 2022b. Omnivl: One foundation model for image-language and video-language tasks// Advances In Neural Information Processing Systems. New Orleans, USA: Curran Associates: 5696-5710 [DOI: 10.5555/3600270.3600682]

Wang J, and Cherian A. 2019. Gods: Generalized one-class discriminative subspaces for anomaly detection// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE: 8200-8210 [DOI: 10.1109/iccv.2019.00829]

Wang P, Bai S, Tan S, Wang S, Fan Z, Bai J, Chen K, Liu X, Wang J, Ge W, Fan Y, Dang K, Du M, Ren X, Men R, Liu D, Zhou C, Zhou J, and Lin J. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution[EB/OL].[2024-09-18]. https://arxiv.org/pdf/2409.12191.pdf

Wang S, Zeng Y, Liu Q, Zhu C, Zhu E, and Yin J. 2018. Detecting abnormality without knowing normality// Proceedings of the 26th Acm International Conference on Multimedia. Seoul, South Korea: ACM: 636-644 [DOI: 10.1145/3240508.3240615]

Wang T, Xuan S, and Zhou J. 2023a. Anomaly Detection Combining Wavelet Transform and Codec Attention. Computer Application Research, 40(7): 2229-2234, 2240 (王婷, 宣士斌, and 周建亭. 2023a. 融合小波变换和编解码注意力的异常检测. 计算机应用研究, 40(7): 2229-2234), 2240) [DOI: 10.19734/j.issn.1001-3695.2022.10.0527]

Wang X, Yuan H, Zhang S, Chen D, Wang J, Zhang Y, Shen Y, Zhao D, and Zhou J. 2023b. Videocomposer: Compositional video synthesis with motion controllability// Advances In Neural Information Processing Systems. New Orleans, USA: Curran Associates: 7594-7611 [DOI: 10.48550/arXiv.2306.02018]

Wang Z, Bao J, Zhou W, Wang W, Hu H, Chen H, and Li H. 2023c. Dire for diffusion-generated image detection// Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22445-22455 [DOI: 10.1109/ICCV51070.2023.02051]

Wang Z, Zheng H, He P, Chen W, and Zhou M. 2022c. Diffusion-gan: Training gans with diffusion[EB/OL].[2022-06-05]. Https://arxiv.org/pdf/2206.02262.pdf

Wang Z, Zou Y, and Zhang Z. 2020. Cluster attention contrast for video anomaly detection// Proceedings of the 28th Acm International Conference on Multimedia. Seattle, USA: ACM: 2463-2471 [DOI: 10.1145/3394171.3413529]

Wen H, He Y, Huang Z, Li T, Yu Z, Huang X, Qi L, Wu B, Li X, and Cheng G. 2025a. Busterx: Mllm-powered ai-generated video forgery detection and explanation[EB/OL].[2025-05-19]. Https://arxiv.org/pdf/2505.12620.pdf

Wen H, Li T, Huang Z, He Y, and Cheng G. 2025b. Busterx++: Towards unified cross-modal ai-generated content detection and explanation with mllm[EB/OL].[2025-07-19]. Https://arxiv.org/pdf/2507.14632.pdf

Wen S, Ye J, Feng P, Kang H, Wen Z, Chen Y, Wu J, Wu W, He C, and Li W. 2025c. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation[EB/OL].[2025-03-19]. Https://arxiv.org/pdf/2503.14905.pdf

Wenhao Z, Hongtao H U, Xu C, and Chunhui Z. 2024. Weakly Supervised Video Anomaly Detection Based on Dual Dynamic Memory Network. Computer Science, 51(1): 243-251 (周文浩, 胡宏涛, 陈旭, and 赵春晖. 2024. 基于双重动态记忆网络的弱监督视频异常检测. 计算机科学, 51(1): 243-251) [DOI: 10.11896/jsjkx.

230300134]

Wu M, Zhu J, Feng X, Chen C, Zhu C, Song B, Mao F, Wu J, Chu X, and Huang K. 2025. ImagerySearch: Adaptive Test-Time Search for Video Generation Beyond Semantic Dependency Constraints[EB/OL].[2025]. https://arxiv.org/pdf/2510.14847.pdf

Wu P, Zhou X, Pang G, Yang Z, Yan Q, Wang P, and Zhang Y. 2024a. Weakly supervised video anomaly detection and localization with spatio-temporal prompts// Proceedings of the 32nd Acm International Conference on Multimedia. Melbourne, Australia: ACM: 9301-9310 [DOI:10.1145/3664647.3681442]

Wu P, Zhou X, Pang G, Zhou L, Yan Q, Wang P, and Zhang Y. 2024b. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection// AAAI Conference on Artificial Intelligence (AAAI). Vancouver, Canada: AAAI Press: 6074-6082 [DOI:10.1609/AAAI.V38I6.28423]

Xi Z, Huang W, Wei K, Luo W, and Zheng P. 2023. Ai-generated image detection using a cross-attention enhanced dual-stream network// 2023 Asia Pacific Signal And Information Processing Association Annual Summit And Conference (Apsipa Asc). Taipei, Taiwan: IEEE: 1463-1470 [DOI: 10.1109/APSIPAASC58517.2023.10317126]

Xu L, Han D, Li G, Zhou M, Wan J, and Li M. 2025. Multimodal feature cooperative refinement for few-shot anomaly detection. Advanced Engineering Informatics, 68: 103792 [DOI: 10.1016/j.aei.2025.103792]

Xu Y, Jia G, Huang H, Duan J, and He R. 2021. Visual-semantic transformer for face forgery detection// 2021 IEEE International Joint Conference on Biometrics (Ijcb). Shenzhen, China: IEEE: 1-7 [DOI:10.1109/IJCB52358.2021.9484407]

Yakun W, Baolin Z, Zhaocheng W, Wanyun M, and Shijia G. 2024. Research on Deep Face Forgery Detection Method Based on Improved ResNet34. Journal of Hebei University of Technology, 53(6): 44-51 (王雅坤,张宝林,王兆成,马琬雲, and 郭仕佳. 2024. 基于改进resnet34的深度人脸伪造检测方法研究. 河北工业大学学报, 53(6): 44-51) [DOI:10.14081/j.cnki.hgdxb.2024.06.005]

Yan W, Zhang Y, Abbeel P, and Srinivas A. 2021. Videogpt: Video generation using vq-vae and transformers[EB/OL].[2021-04-20]. Https://arxiv.org/pdf/2104.10157.pdf

Yang C-A, Peng K-C, and Yeh R A. 2025a. Toward long-tailed online anomaly detection through class-agnostic concepts[EB/OL].[2025-07-22]. Https://arxiv.org/pdf/2507.16946.pdf

Yang T, Chang L, Yan J, Li J, Wang Z, and Zhang K. 2025b. A survey on foundation-model-based industrial defect detection [EB/OL].[2025-02-26]. Https://arxiv.org/pdf/2502.19106.pdf

Yang X, Li Y, and Lyu S. 2019. Exposing deep fakes using inconsistent head poses// ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech And Signal Processing (ICASSP). Brighton, UK: IEEE: 8261-8265 [DOI:10.1109/ICASSP.2019.8683164]

Yang Y, Lee K, Dariush B, Cao Y, and Lo S-Y. 2024a. Follow the rules: Reasoning for video anomaly detection with large language models// European Conference on Computer Vision (ECCV). Milan, Italy: Springer: 304-322 [DOI: 10.1007/978-3-031-73004-7_18]

Yang Z, Li L, Lin K, Wang J, Lin C-C, Liu Z, and Wang L. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision)[EB/OL].[2023-09-29]. Https://arxiv.org/pdf/2309.17421.pdf

Yang Z, Teng J, Zheng W, Ding M, Huang S, Xu J, Yang Y, Hong W, Zhang X, and Feng G. 2024b. Cogvideox: Text-to-video diffusion models with an expert transformer[EB/OL].[2024-08-12]. Https://arxiv.org/pdf/2408.06072.pdf

Yao X, Li R, Zhang J, Sun J, and Zhang C. 2023. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Vancouver, Canada: IEEE: 24490-24499 [DOI:10.1109/CVPR52729.2023.02346]

Yu J, Lee Y, Yow K C, Jeon M, and Pedrycz W. 2022a. Abnormal event detection and localization via adversarial event prediction. IEEE Trans Neural Netw Learn Syst, 33(8): 3572-3586 [DOI:10.1109/TNNLS.2021.3053563]

Yu J, Wang Z, Vasudevan V, Yeung L, Seyedhosseini M, and Wu Y. 2022b. Coca: Contrastive captioners are image-text foundation models[EB/OL].[2022-05-04]. Https://arxiv.org/pdf/2205.01917.pdf

Yu S, Sohn K, Kim S, and Shin J. 2023. Video probabilistic diffusion models in projected latent space// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Vancouver, Canada: IEEE: 18456-18466 [DOI: 10.1109/CVPR52729.2023.01770]

Yuan L, Chen D, Chen Y-L, Codella N, Dai X, Gao J, Hu H, Huang X, Li B, and Li C. 2021. Florence: A new foundation model for computer vision[EB/OL].[2021-11-22]. Https://arxiv.org/pdf/2111.11432.pdf

Yuan M, and Xiyuan W. 2024. A Dual-Stream Deep Detection Method for Forged Faces Based on Feature Fusion. Journal of Ningxia University (Natural Science Edition), 45(3): 299-306 (孟媛, and 汪西原. 2024. 一种特征融合的双流深度检测伪造人脸方法. 宁夏大学学报(自然科学版), 45(3): 299-306) [DOI:10.20176/j.cnki.nxdz.000051]

Zaheer M Z, Lee J-H, Astrid M, and Lee S-I. 2020. Old is gold: Redefining the adversarially learned one-class classifier training paradigm// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 14171-14181 [DOI: 10.1109/CVPR42600.2020.01419]

Zaheer M Z, Lee J H, Mahmood A, Astrid M, and Lee S I. 2022a. Stabilizing adversarially learned one-class novelty detection using pseudo anomalies. IEEE Trans Image Process, 31: 5963-5975 [DOI:10.1109/TIP.2022.3204217]

Zaheer M Z, Mahmood A, Khan M H, Segu M, Yu F, and Lee S-I. 2022b. Generative cooperative learning for unsupervised video anomaly detection// 2022 IEEE/CVF Conference on Computer Vision And Pattern Recognition (CVPR). New Orleans, USA: IEEE:14724-14734 [DOI:10.1109/cvpr52688.2022.01433]

Zhang H, Goodfellow I, Metaxas D, and Odena A. 2019a. Self-attention generative adversarial networks// International Conference on Machine Learning. Long Beach, United States: PMLR: 7354-7363 [DOI:10.48550/arXiv.1805.08318]

Zhang H, Xu X, Wang X, Zuo J, Han C, Huang X, Gao C, Zhang S, Yu L, and Sang N. 2025a. Holmes-VAU: Towards Long-term Video Anomaly Understanding at Any Granularity// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Nashville, United States: IEEE: 13843-13853 [DOI: 10.1109/CVPR52734.2025.01292]

Zhang J, Qing L, and Miao J. 2019b. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection// 2019 IEEE International Conference on Image Processing (LCIP). Taipei, Taiwan: IEEE: 4030-4034 [DOI: 10.1109/icip.2019.8803657]

Zhang L, Rao A, and Agrawala M. 2023a. Adding conditional control to text-to-image diffusion models// Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3836-3847 [DOI:10.1109/ICCV51070.2023.00355]

Zhang S, Wang J, Zhang Y, Zhao K, Yuan H, Qin Z, Wang X, Zhao D, and Zhou J. 2023b. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models[EB/OL].[2023-11-07]. Https://arxiv.org/pdf/2311.04145.pdf

Zhang W, Jiang C, Zhang Z, Si C, Yu F, and Peng W. 2025b. Ivy-fake: A unified explainable framework and benchmark for image and video aigc detection[EB/OL].[2025-06-01]. Https://arxiv.org/pdf/2506.00979.pdf

Zhang Y, Colman B, Guo X, Shahriyari A, and Bharaj G. 2024. Common sense reasoning for deepfake detection// European Conference on Computer Vision. Milan, Italy: Springer: 399-415 [DOI: 10.1007/978-3-031-73223-2_22]

Zhao H, Zi C, Liu Y, Zhang C, Zhou Y, and Li J. 2024. Weakly supervised anomaly detection via knowledge-data alignment// Proceedings of the Acm Web Conference 2024. Singapore, Singapore: ACM: 4083-4094 [DOI:10.1145/3589334.3645429]

Zhao T, Xu X, Xu M, Ding H, Xiong Y, and Xia W. 2021. Learning self-consistency for deepfake detection// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 15023-15033 [DOI: 10.1109/ICCV48922.2021.01475]

Zhaobo C, Lin Z, and Xiaoxuan M A. 2025. Research on Improved Attention-Mixed Autoencoder for Video Anomaly Detection. Computer Engineering and Science, 47(1): 130-139 (陈兆波,张琳, and马晓轩. 2025.改进注意力混合自动编码器视频异常检测研究.计算机工程与科学, 47(1): 130-139) [DOI:10.3969/j.issn.1007-130X.2025.01.014]

Zheng Z, Peng X, Yang T, Shen C, Li S, Liu H, Zhou Y, Li T, and You Y. 2024. Open-sora: Democratizing efficient video production for all[EB/OL].[2024-12-29]. Https://arxiv.org/pdf/2412.20404.pdf

Zhenhua X U E, Qiang L I, and Chao H. 2025. Pixel-level image anomaly detection method driven by visual foundation models. Computer Applications, 45(3): 823-831 (薛振华,李强, and 黄超. 2025.视觉基础模型驱动的像素级图像异常检测方法.计算机应用, 45(3): 823-831) [DOI: 10.11772/j. issn. 1001-9081.2024091398]

Zhong N, Xu Y, Li S, Qian Z, and Zhang X. 2023. Patchcraft: Exploring texture patch for efficient ai-generated image detection[EB/OL].[2023-11-21]. Https://arxiv.org/pdf/2311.12397.pdf

Zhou F, Wang G, Zhang K, Liu S, and Zhong T. 2023. Semi-supervised anomaly detection via neural process. IEEE Transactions On Knowledge And Data Engineering, 35 (10): 10423-10435 [DOI:10.1109/TKDE.2023.3266755]

Zhou Q, Pang G, Tian Y, He S, and Chen J. 2024. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection// International Conference on Learning Representations (ICLR). Vienna, Austria: ICLR:1-33 [DOI:10.48550/arXiv.2310.18961]

Zhou W, Li Y, and Zhao C. 2022a. Object-guided and motion-refined attention network for video anomaly detection// 2022 IEEE International Conference on Multimedia And Expo (ICME). Taipei, Taiwan: IEEE:1-6 [DOI:10.1109/icme52920.2022.9859927]

Zhou Y, Song X, Zhang Y, Liu F, Zhu C, and Liu L. 2022b. Feature encoding with autoencoders for weakly supervised anomaly detection. IEEE Trans Neural Netw Learn Syst, 33 (6): 2454-2465 [DOI:10.1109/TNNLS.2021.3086137]

Zhu J, Ding C, Tian Y, and Pang G. 2024. Anomaly heterogeneity learning for open-set supervised anomaly detection// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Seattle, United States: IEEE: 17616-17626 [DOI:10.1109/CVPR52733.2024.01668]

Zhu S, Chen C, and Sultani W. 2020. Video anomaly detection for smart surveillance[EB/OL].[2020-04-01]. https://arxiv.org/pdf/2004.00222.pdf

Zhu X, Wang H, Fei H, Lei Z, and Li S Z. 2021. Face forgery detection by 3d decomposition// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. Nashville, United States: IEEE: 2929-2939 [DOI: 10.1109/CVPR46437.2021.00295]

Zhu Y, Bao W, and Yu Q. 2022. Towards open set video anomaly detection// Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition. New Orleans, USA: IEEE: 395-412 [DOI:10.1007/978-3-031-19830-4_23]

Zuo Z, Dong J, Wu Y, Qu Y, and Wu Z. 2024. Clip-fsac++: Few-shot anomaly classification with anomaly descriptor based on clip [EB/OL].[2024-12-05]. Https://arxiv.org/pdf/2412.03829.pdf

作者简介

桑农,男,教授,主要研究方向为模式识别和计算机视觉。

黄凯奇,通信作者,男,研究员,主要研究方向为计算机视觉、模式识别、博弈决策。

赵耀,通信作者,男,教授,主要研究方向为计算机视觉、AIGC鉴伪、AI视频编码、多媒体智能理解。

高常鑫,男,教授,主要研究方向为图像/视频理解与生成、多智能体协同。

考月英,女,高级工程师,主要研究方向为计算机视觉和模式识别。

谭创创,男,讲师,主要研究方向为计算机视觉、深度伪造检测。

王翔,男,博士研究生,主要研究方向为视频行为分析、视频生成。

武美奇,女,博士研究生,主要研究方向为计算机视觉、视频生成和手势交互。

尹文体,男,硕士研究生,主要研究方向为视频异常检测。