E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)06-1872-81

论文引用格式:Ma Y Z, Zhang Y F, Jia W, Liu J Y, Gan T, Yang W H, Zhuo J B, Liu W and Ma H M. 2025. Recent advances in data generation and its applications in computer vision. Journal of Image and Graphics, 30(6):1872-1952(马愈卓, 张永飞, 贾伟, 刘家瑛, 甘甜, 杨文瀚, 卓君宝, 刘武, 马惠敏. 2025. 面向计算机视觉的数据生成与应用研究进展. 中国图象图形学报, 30(6):1872-1952)[DOI:10.11834/jig.250085]

面向计算机视觉的数据生成与应用研究进展

马愈卓[1],张永飞[1*],贾伟[2],刘家瑛[3],甘甜[4],杨文瀚[5],卓君宝[6],刘武[7],马惠敏[6]

1. 北京航空航天大学计算机学院,北京　100191; 2. 合肥工业大学计算机与信息学院,合肥　230009;
3. 北京大学王选计算机研究所,北京　100080; 4. 山东大学计算机科学与技术学院,济南　250100; 5. 鹏城实验室,深圳　518055;
6. 北京科技大学计算机与通信工程学院,北京　100083; 7. 中国科学技术大学信息科学技术学院,合肥　230027

摘　要: 大规模图像和视频数据集是驱动计算机视觉算法发展的核心要素。面向计算机视觉任务,构建大规模图像和视频数据集是一项重要但复杂的任务。基于生成对抗网络和扩散模型等数据生成方法可以可控地生成大规模、多样性的图像和视频数据,有效替代或弥补真实图像和视频数据集,为计算机视觉技术领域的发展提供了新的动力。本文在对面向计算机视觉的图像和视频数据生成与应用背景简介的基础上,首先,从以几何变换等为代表的传统数据增广和生成、以虚拟引擎和神经辐射场等为代表的基于三维渲染的数据生成方法、以生成对抗网络和扩散模型等为代表的基于深度生成模型的生成方法等3方面系统调研典型的图像和视频数据生成技术与模型;其次,梳理了典型的图像和视频数据生成技术与模型在图像增强,目标检测跟踪与姿态动作识别等个体分析,基于图像和视频的生物特征识别、人员计数与人群行为分析等群体行为分析、自动驾驶、视频生成、具身智能等典型计算机视觉相关任务中的应用;最后,分析面向计算机视觉的数据生成与应用中存在的问题,并展望未来发展趋势,以期促进图像和视频数据生成及计算机视觉技术的发展。

关键词: 计算机视觉;数据生成与应用;传统数据生成;三维渲染;深度生成模型;图像增强;个体分析;生物特征识别;群体分析;自动驾驶;视频生成;具身智能

Recent advances in data generation and its applications in computer vision

Ma Yuzhuo[1], Zhang Yongfei[1*], Jia Wei[2], Liu Jiaying[3], Gan Tian[4], Yang Wenhan[5], Zhuo Junbao[6], Liu Wu[7], Ma Huimin[6]

1. School of Computer Science and Engineering, Beihang University, Beijing　100191, China; 2. School of Computer Science and Information Engineering, Hefei University of Technology, Hefei　230009, China; 3. Wangxuan Institute of Computer Technology, Peking University, Beijing　100080, China; 4. School of Computer Science and Technology, Shandong University, Ji'nan　250100, China; 5. Pengcheng Laboratory, Shenzhen　518055, China; 6. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing　100083, China; 7. School of Information Science and Technology, University of Science and Technology of China, Hefei　230027, China

Abstract: Large-scale image and video datasets are indispensable for the development of computer vision algorithms, mainly because they provide the necessary resources to train and evaluate various models. Constructing such datasets for different computer vision tasks is a crucial but complex task, because it involves considerable challenges in data collec-

收稿日期:2025-02-28;**修回日期**:2025-03-12;**预印本日期**:2023-03-19

*通信作者:张永飞　yfzhang@buaa.edu.cn

基金项目:国家自然科学基金项目(62472016,62076086, 62476077)

Supported by:National Natural Science Foundation of China(62472016,62076086,62476077)

tion, annotation, and preservation of data diversity. Traditionally, acquiring large, high-quality image and video datasets has been a resource-intensive task, requiring manual labeling, data collection in real-world settings, and the use of specialized hardware for capturing high-quality images and videos. As deep learning methods increasingly rely on large-scale labeled data, the need for innovative data generation techniques has become more prominent. In recent years, generative models, such as generative adversarial network (GAN) and diffusion models have emerged as powerful tools for generating synthetic datasets. These models can create diverse, controllable, and highly realistic image and video data, offering an effective alternative or supplement to traditional data collection methods. By using these techniques, vast amounts of data can be generated to represent various scenarios and conditions, which are essential for training robust computer vision models. Generative models provide a flexible solution that can generate data without the need for real-world data acquisition, unlike traditional data collection, which is often constrained by geographic, financial, and logistical limitations. This review begins by introducing the significance and background of image and video data generation in computer vision. Image and video data play a critical role in the development and training of computer vision algorithms, as large-scale, diverse datasets are essential for building robust models. Moving on, the review categorizes the key data generation techniques into three broad approaches: traditional data augmentation methods, 3D rendering-based generation methods, and deep generative models. First, traditional data augmentation techniques, including geometric transformations, color adjustments, and cropping, are commonly used to improve model generalization by expanding existing datasets. Although these methods are relatively simple and computationally inexpensive, their ability to generate diverse and realistic datasets is limited. In comparison, 3D rendering technologies, such as virtual engines and neural radiance fields (NeRF), enable the creation of highly realistic synthetic data by simulating real-world environments. These technologies have the advantage of generating diverse datasets by adjusting environmental factors, such as lighting, object interactions, and camera angles. Furthermore, deep generative models, such as GAN and diffusion models, have shown remarkable effectiveness in generating high-quality synthetic data. On the one hand, GAN work by training two neural networks in a competitive manner in which a generator creates synthetic data, while a discriminator evaluates its realism. Over time, as the generator improves its output, it creates increasingly realistic data. Diffusion models, on the other hand, iteratively refine noisy data into clear and realistic images or videos, enabling the generation of diverse, high-quality datasets. Next, the review discusses the diverse applications of these generative models across a wide range of computer vision tasks. These tasks include image enhancement, object detection, tracking, pose or action recognition, biometric identification, crowd behavior analysis, and more recently, emerging fields like autonomous driving and embodied artificial intelligence. In particular, synthetic data have been instrumental in training models for tasks that are challenging to address using only real-world data. For example, in biometric identification, synthetic data can generate a wide variety of samples for fingerprints, faces, irises, and palmprints, thus providing more diverse training examples and reducing reliance on real biometric data, which are often difficult to acquire. Similarly, in autonomous driving, synthetic data can generate various driving scenarios, including different road conditions, weather patterns, and traffic behaviors, thereby training autonomous vehicle models in safe and controlled environments. In recent years, synthetic data have also been proven invaluable in fields like pose and action recognition, where diverse datasets are essential for accurately detecting human actions across different settings and contexts. However, despite the considerable progress made in image and video data generation, several challenges remain. One of the primary issues is ensuring the realism and diversity of generated data, which is crucial for training models that can generalize well to real-world scenarios. Furthermore, despite significant advances in generative models, there remains a lack of research on how to effectively evaluate the quality of synthetic data and use feedback mechanisms to guide the generation process. In addition, ethical considerations surrounding the use of synthetic data, especially in sensitive applications, such as biometric recognition, must be carefully addressed. Among others, the use of synthetic data raises concerns regarding privacy, consent, and potential misuse, which must be handled responsibly. Looking ahead, as generative models continue to evolve, they are expected to produce even more realistic and diverse datasets, thus offering new possibilities for training computer vision models. The future of image and video data generation holds great promise, with advancements in generative technologies poised to drive further innovation in computer vision, AI, and many other fields.

Key words: computer vision; data generation and application; conventional data generation; 3D rendering; deep genera-

tive model; image enhancement; individual analysis; biometric recognition; crowd analysis; autonomous driving; video generation; embodied artificial intelligence

0 引 言

计算机视觉是人工智能(artificial intelligence, AI)的核心技术之一,它通过模拟人类视觉系统,实现图像与视频信息的智能识别、理解与分析,具有极其广泛的应用前景。高质量、标注精确的大规模图像和视频数据是驱动计算机视觉技术发展的核心要素。在深度学习时代,模型的效果常常与训练数据的质量和数量成正比。一个强大的计算机视觉系统,不仅依赖于先进的模型架构和高效的学习算法,更离不开全面且大规模的图像或视频训练数据集。大规模且涵盖不同场景、环境和条件的、标注准确的图像和视频训练数据集,能够为模型训练提供丰富的信息,驱动模型从中学习到图像和视频的深层次特征,直接决定了模型的泛化能力和应用效果。规模定律(scaling law)也表明,模型的泛化误差与训练集的大小呈现幂律关系,即随着训练数据集的增大,模型的泛化误差会以一定的幂次下降。

然而,构建大规模图像或视频数据集是一项复杂而艰巨的任务,涉及到采集、标注、处理和存储等多个环节。首先,数据的采集需要考虑多样性和代表性,确保数据集能够覆盖各种实际场景和条件,如不同的光照、天气、角度和场景变化等。这一过程不仅需要大量的硬件设备支持,如高质量的摄像机等,还可能涉及到复杂的场景设置和拍摄任务,以保证数据的多样性和真实感。其次,数据标注是数据集构建中的一个关键步骤。图像或视频标注不仅要求标注人员准确识别图像中的目标,还需处理运动物体、目标交互和动态场景中的复杂情况,往往非常烦琐且时间密集,尤其是在大规模数据集的构建过程中,标注质量的控制至关重要。为确保高效标注,许多项目采用了半自动化标注工具或众包平台,但即便如此,仍然需要人工校验和优化,以保证数据的准确性。总体来看,从真实世界采集数据构建大规模数据集成本高昂且费时、费力。另一方面,随着公众和政府部门对隐私保护及国家安全的日益重视,真实数据集的采集与公开问题变得愈发敏感。

随着2014年生成对抗网络(generative adversarial network, GAN)的提出,生成式人工智能(artificial intelligence generated content, AIGC)取得突破性发展。GAN通过生成器和判别器的对抗训练,实现了高质量数据的生成。随后,变分自编码器(variational auto encoder, VAE)和自回归模型等技术的提出进一步推动该领域的发展。2020年后,AIGC迎来爆发式增长,以OpenAI的GPT-3(generative pre-trained transformer-3)和DALL·E为代表,前者展现了强大的自然语言生成能力,后者实现了文本到图像的生成。近年来,扩散模型(diffusion model)和多模态大模型进一步提升了生成质量和多样性。AIGC已广泛应用于艺术创作、内容生成和科学研究等领域,成为人工智能发展的重要方向。借助三维渲染和深度生成模型,AIGC能够可控生成大规模、多样化的图像和视频数据集,几乎无须标注、没有安全或隐私泄露问题,且能够弥补现实世界的不足,逐步成为推动人工智能技术创新与产业升级的关键力量。

生成的图像和视频数据在训练和评估各种计算机视觉算法模型方面极具价值。在模型训练方面,生成数据既可用于传统数据的增广,扩充数据集规模并提升数据多样性,帮助模型更好地学习数据分布,也能替代真实数据,解决数据稀缺、隐私保护及成本问题。此外,生成数据还可用于解决类别不平衡问题,支持模型初始化、预训练与正则化,促进弱监督学习以及跨领域迁移学习,从而为模型训练提供更加灵活和高效的解决方案。在模型测试评估方面,通过生成数据,可以模拟各种极端或罕见场景,有效提升测试的全面性和可靠性。例如自动驾驶中的恶劣天气条件或医疗影像中的罕见病例,从而评估模型在复杂环境下的鲁棒性和泛化能力。此外,生成数据还可以创建多样化的测试集,覆盖不同的数据分布和特征,帮助发现模型在特定场景下的潜在缺陷。在隐私保护领域,生成数据可以替代真实数据用于测试,避免用户的隐私泄露。例如特斯拉利用AIGC技术生成仿真模拟环境,用于自动驾驶算法的训练和测试。这种仿真模拟环境能够模拟各种复杂的道路场景、交通情况和天气条件,为自动驾驶算法提供大量的训练数据,既可以大幅降低实际道路测试的成本和风险,同时提高自动驾驶算法的安

全性和可靠性。据市场研究机构 Gartner 预测：到2030年，生成数据预计将在 AI 模型中完全超越真实数据。

综上，大规模高质量图像和视频数据集的构建，及其在计算机视觉任务中的应用已成为人工智能领域的热点。鉴于此，本报告围绕典型数据生成技术与模型，及其在典型计算机视觉任务中的应用，进行了系统的调研、分析和综述，希望能为数据生成和计算机视觉技术领域的专家学者和研究人员提供参考，促进数据生成和计算机视觉技术的发展与应用。本文第 1 节从传统数据生成方法、三维渲染技术以及深度生成模型 3 个方面介绍典型数据生成技术与模型。第 2 节讨论典型计算机视觉任务中的应用。最后总结面向计算机视觉的数据生成与应用中目前尚存在的问题，并展望相关技术与应用的发展趋势。

1　典型图像和视频数据生成技术

随着计算机视觉技术的快速发展，面向计算机视觉的数据生成技术在过去数年间经历了显著的演变。从近 5 年的相关文献中，对关键词进行了聚类分析，并绘制了散点图和词云图，如图 1 所示。其中，左侧的散点图展示关键词在降维空间中的聚类分布，颜色代表不同的聚类类别；右侧的词云图通过词频大小直观呈现各关键词的重要程度。结果表明，这些关键词可以划分为 3 大类，大致代表 3 个主要研究方向。第 1 类关键词（如贝叶斯定理、图像增强、蒙特卡洛方法）主要与基于规则和统计模型的传统数据生成技术相关；第 2 类关键词（如三维成像、计算机图形学、虚拟现实）聚焦于三维渲染技术；第 3 类关键词（如深度学习、计算机神经网络、“机器学习）对应深度生成模型技术。

基于此，本文将数据生成技术的发展归纳为 3 个主要阶段：传统数据生成阶段、三维渲染阶段，以及深度生成阶段。在早期的传统方法中，数据生成主要依赖手工设计的规则、几何变换、图像处理算法和统计模型，这些方法在满足基本需求的同时，在生成数据的复杂性、多样性和真实感方面存在显著局限。随后，随着计算能力的提升和虚拟现实技术的发展，三维渲染技术逐步兴起。通过构建虚拟环境或三维模型，该方法能够生成多样性更高且更加真实的图像数据，但因其依赖复杂的模型设计和高计算成本，扩展性受到限制。近年来，深度生成模型（如变分自编码器、生成对抗网络、扩散模型等）的崛起为数据生成技术带来了革命性变化。这些方法基于强大的学习能力和无监督特性，能够在多种场景下高效生成逼真的数据样本，显著提升了数据生成的多样性和真实感。同时，AIGC 的持续进步进一步推动了数据生成技术的智能化和自动化，为计算机视觉领域的创新提供了强有力的支持（严昊 等，2023）。

为了更清晰地呈现该领域的发展趋势，在谷歌学术上统计近 5 年来与计算机视觉数据生成技术相关的论文数量变化，如图 2 所示。可以看出，相关研究逐年增长，体现了该领域的持续关注与快速发展。其中，2023 年达到最高峰，出版量达 5 790 篇，随后 2024 年略有下降，但整体趋势仍维持在较高水平。这表明，数据生成技术在计算机视觉领域的重要性

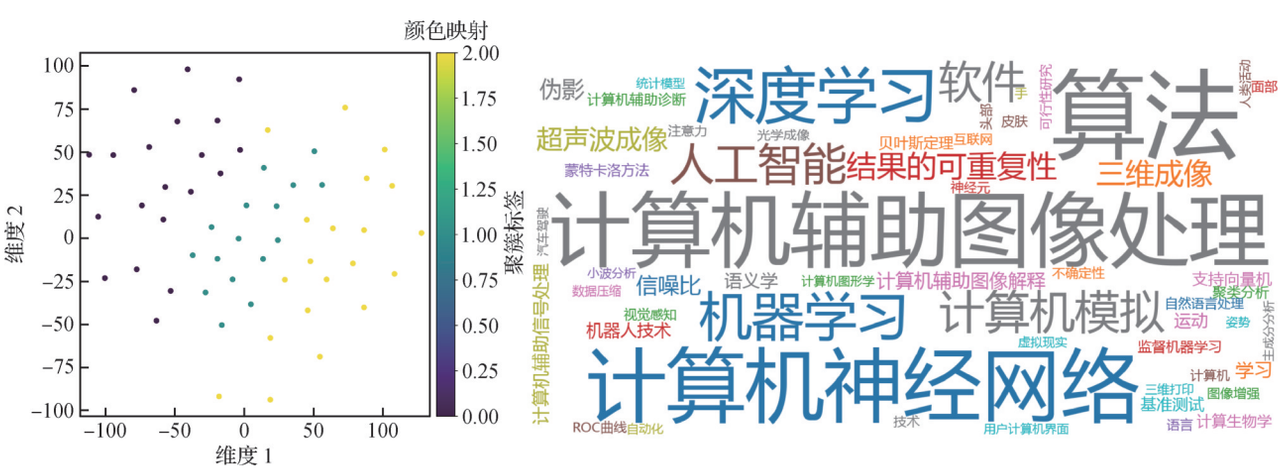

图 1　近 5 年相关文献关键词聚类可视化分析及词云统计图

Fig. 1　Cluster visualization and word cloud statistics of keywords in related literature from the last five years

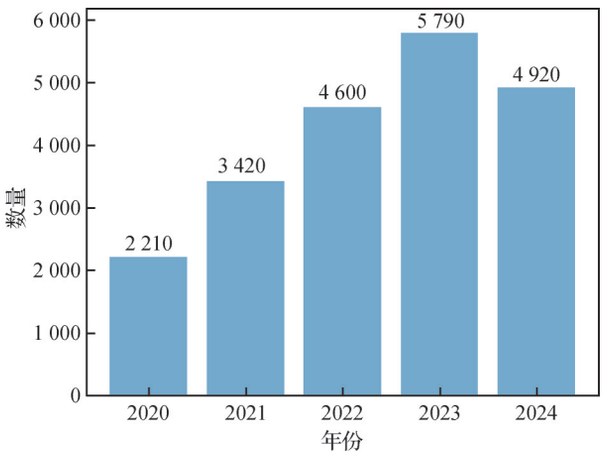

图2　视觉数据生成相关论文数量变化

Fig. 2　Variation in the number of papers on visual data generation

日益凸显，研究热度不断攀升。

同时，根据文献来源和技术类型对近5年与计算机视觉数据生成技术相关的研究进行分类统计。图3(a)展示这些文献在不同学术平台分布情况。可以看出，AAAI(Association for the Advancement of Artificial Intelligence)会议中的相关文献占比最高，达40.8%，体现出该会议对数据生成技术研究的广泛关注和高度认可。紧随其后的是计算机视觉领域的3大顶级会议——CVPR(IEEE Conference on Computer Vision and Pattern Recognition)、ICCV(IEEE International Conference on Computer Vision)和ECCV(European Conference on Computer Vision)，分别占据18.2%、10.8%和7.4%，进一步表明数据生成技术已成为计算机视觉领域的重要研究方向之一，吸引了大量高质量的学术成果投稿。此外，结合图1中的聚类结果，进一步统计了传统数据生成、三维渲染和深度生成模型这3类技术的文献占比，结果如图3(b)所示。从分布上来看，深度生成模型技术占据主导地位，占比37.9%，反映出近年来深度学习驱动的数据生成方法在计算机视觉领域的广泛应用与重要性。其次是三维渲染技术，占比32.1%，显示出虚拟现实、计算机图形学等技术在视觉数据生成中的关键作用。最后，传统数据生成技术占比30.0%，尽管比例略低，但作为基础技术，仍在许多场景中发挥着不可或缺的支撑作用。

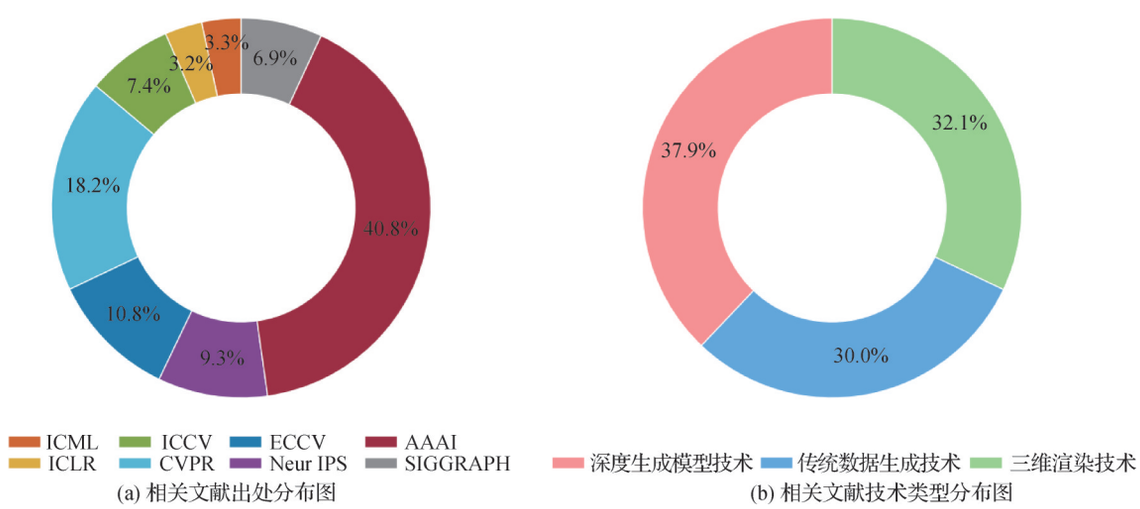

图3　相关文献出处分布图以及相关文献技术类型分布图

Fig. 3　Distribution map of sources of relevant literature and distribution map of technology types of relevant literature ((a)distribution map of relevant literature sources; (b)distribution chart of related literature technology types)

综上，该领域已有很多相关工作，本节着重为这些技术提供一个较为全面的概述，主要从传统数据生成方法、三维渲染技术以及深度生成模型3个方面进行系统调研与分析，如图4所示。通过对这3个方向的深入分析，旨在全面展示计算机视觉数据生成技术的演进与现状，并探讨未来发展趋势。

1.1　传统数据生成技术

传统数据生成技术通过一系列确定性或规则化操作，对现有图像数据进行编辑和扩展，从而生成符合特定需求、多样化的新图像样本，已广泛应用于数据增强、训练集构建，为复杂计算机视觉任务的成功奠定了坚实基础。传统数据生成技术的实现路径多

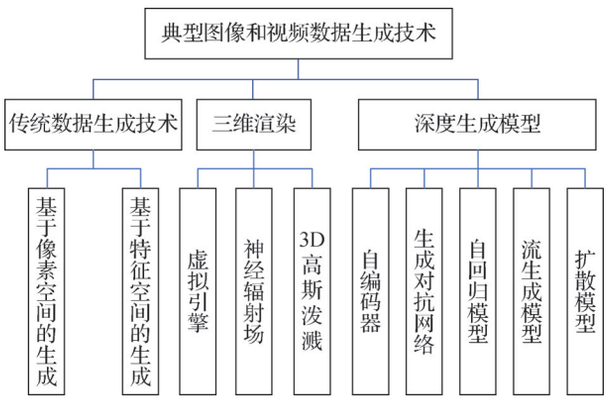

图4　典型图像和视频数据生成技术

Fig. 4　Typical image and video data generation technologies

样，包括但不限于几何变换、图像扰动、图像粘贴、形态学操作、物理渲染、数学建模，以及特征融合。

根据编辑和扩展对象所在的空间不同，本节将传统数据生成方法分为基于像素空间的数据生成和基于特征空间的数据生成两大类，介绍其思路和相关工作。

1.1.1　基于像素空间的生成

基于像素空间的图像数据生成方法专注于直接在像素级对图像进行编辑操作，通常通过在像素级调整图像的几何形态、色彩属性等基本视觉元素，对现有数据进行加工和扩展，从而实现多样化的数据生成。本小节从几何变换、图像扰动、图像粘贴和形态学操作等方面介绍基于像素空间的数据生成方法。

1）几何变换。图像几何变换是图像处理和计算机视觉中的一个基础操作，它通常涉及对图像中的像素位置进行变换，以实现图像的剪切、翻转、旋转、缩放等效果。以下是一些常见的图像几何变换算法。

（1）随机裁剪。通过从图像中随机选择区域裁剪生成新图像，增强数据多样性并丰富场景。例如，RICAP（random image cropping and patching）方法（Takahashi等，2020）将4幅图像的裁剪区域拼接生成新的样本，但其有效性依赖于裁剪策略和任务需求，可能丢失关键信息或引入噪声（Zhang等，2018）。

（2）随机翻转或旋转。通过水平或垂直翻转和任意角度旋转模拟多样化视角，提高模型对方向和旋转变化的鲁棒性。例如，Alomar等人（2023）提出通过随机选择图像中的圆形区域并施加随机旋转，从而增强了模型对图像局部特征的识别能力。然而，翻转或旋转需限制范围，避免失真。

（3）随机擦除。通过遮挡图像某一区域（如随机像素或固定值）减少模型对特定特征的依赖（Zhong等，2020）。优化策略如利用神经网络引导擦除位置（Sun等，2020），使其更具任务相关性，进一步增强对遮挡的适应性。

（4）变形缩放。通过调整图像大小和形状以增加数据多样性，例如Chen等人（2019）提出的基于线段拉伸或收缩的方法以提高模型对形变的鲁棒性。当然，由于变形可能会引入非自然形态，在使用时需权衡形变与图像自然性的关系。

（5）仿射变换。通过线性映射保持共线性和直线平行性，常用于图像的几何处理。但普通仿射变换可能引起图像的局部特征和比例关系发生失真，引入莫比乌斯变换（Zhou等，2021）可以更好地保留局部特征和比例关系，提升对几何变化的识别能力。

2）图像扰动。图像扰动广泛用于生成合成数据。图像扰动可以通过基于经典或手工方法引入，也可以通过基于学习的方法引入。

（1）基于经典或手工方法。通过直接对图像像素进行操作来模拟真实世界中的噪声和环境变化，引入的扰动通常包括高斯模糊、颜色抖动等。高斯模糊通过应用高斯核来平滑图像，减少图像的清晰度，从而模拟图像在不同成像条件下的模糊效果；颜色抖动则通过改变图像中每个通道的像素值来改变亮度、对比度和饱和度等，从而模拟不同光照和颜色变化的情况。例如，He等人（2022b）使用颜色抖动对乳腺癌病理图像进行数据增强，有效减少模型训练过程中的过拟合。

（2）基于学习的方法。通过训练模型来生成特定于输入图像的扰动，这些扰动通常通过对抗性训练机制引入，旨在制造可能导致分类或检测任务错误预测的干扰，从而有助于模型正则化并提高其鲁棒性。例如，Wang等人（2021a）使用对抗增强的方法生成混合不同类型噪声的人体位姿图像，从而提升了人体姿态估计任务在复杂场景中的表现。

3）图像粘贴。图像粘贴是一种简单有效的图像编辑技术，通过将一幅图像或其特定区域复制并整合到另一幅图像中，实现视觉信息的合成和增强。具体而言，图像粘贴方法通过粘贴具有不同背景和前景的图像样本，丰富了训练数据的多样性。在计算机视觉任务中，这种技术通常用于数据增强，从而

提高模型的泛化能力。例如，Dwibedi等人(2017)提出一种构建大型实例检测数据集的方法，该方法通过从对象级实例中剪切出物体并将其粘贴到随机选择的背景上，从而得到新的合成图像。将这些合成图像和真实图像一起用于训练，显著提高了模型的性能。此外，Ghiasi等人(2021)的研究进一步验证了简单的图像复制粘贴在实例分割任务中的有效性。他们的实验证明，这一技术不仅有助于改善模型在分割任务中的表现，还表明其在半监督学习环境中的潜力。总的来说，图像粘贴作为一种传统数据生成技术，通过简单的复制粘贴操作，显著增强了实例检测和分割模型的鲁棒性和泛化能力，尤其是在数据稀缺或标签有限的场景中展现出巨大的应用潜力。

4)形态学操作。形态学操作基于集合论的概念，通过像素级的元素与图像交互，实现图像形状和结构的修改。基本的形态学操作包括腐蚀和膨胀，分别通过减少或增加目标物体的尺寸，来强调或抑制图像中的特定形状特征。腐蚀操作通过“侵蚀”目标区域的边界，使物体缩小，而膨胀操作则通过“扩展”物体的边界，使其变大，从而改变图像中物体的几何形态。例如，Maltoni等人(2009)利用形态学操作，如膨胀和腐蚀，来改变指纹脊线的厚度，从而生成不同压力条件下的指纹图像。

5)物理渲染。物理渲染是一种基于物理规律的图像生成方法，通过模拟光线与物体交互的过程生成高保真图像，常用方法包括光线追踪(ray tracing)和辐射度算法(radiosity)。

光线追踪是一种从观察者视角出发模拟光线传播的渲染方法。光线从摄像机发射，沿路径与物体表面发生反射、折射和吸收等交互，随后继续传播，直到被吸收或达到最大反弹次数。该方法能够生成真实感光影效果，适用于电影特效、建筑渲染等高保真场景。常用的工具有Blender Cycles、NVIDIA OptiX、V-Ray等。

辐射度算法通过计算场景中各个表面的辐射交换来模拟光照，尤其适用于均匀散射的光照条件。与光线追踪不同，辐射度算法更多地关注表面之间的光能传递。它通常通过将场景划分为多个小面片，并计算这些面片之间的光照交互来生成图像。该方法擅长表现柔和光照，常见于建筑和室内设计中的光照模拟。常用的工具有Radiance和RenderMan。

1.1.2　基于特征空间的生成

与基于像素空间的数据生成直接在像素层面进行操作不同，基于特征空间的传统数据生成方法通过图像抽象特征层面的处理，实现多样化新数据的生成。通过在特征空间进行操作，这些方法能够捕捉和利用图像的高级属性，从而为计算机视觉任务提供更为丰富和有意义的数据。本小节从数学建模和图像特征融合两方面介绍基于特征空间的数据生成。

1)数学建模。数学建模作为数据生成的早期方法，核心在于对图像特征空间的分布进行建模，从特征层面生成新数据样本，而非简单依赖原始像素。其过程包括对图像高层特征(如边缘、纹理、形状等)及其统计分布的深入分析，结合统计学与概率论构建生成模型，再通过采样实现新数据的生成。这一方法依赖对现实现象的精确描述，为数据生成提供了理论支持。以下是主要方法：

(1)蒙特卡罗方法。基于随机采样，通过从概率分布中生成样本，模拟复杂系统的行为。它常用于生成具有随机光照、纹理等特性的图像。例如，在纹理合成中，该方法通过随机采样生成与原始图像统计特性相似的新纹理，增加数据的多样性。

(2)马尔可夫蒙特卡罗(Markov chain Monte Carlo，MCMC)(Ho等，2020)。利用马尔可夫链从复杂概率分布中采样，适合高维空间的数据生成。在图像生成任务中，MCMC可以生成具有特定纹理或形状特征的图像。尽管它能够有效处理复杂分布，但其较高的计算复杂度限制了在大规模数据生成中的应用。

(3)朗之万动力学。通过模拟粒子在势场中的运动进行采样，生成具有复杂分布的数据样本(Vahdat等，2021；Song和Ermon，2020)。这种方法可用于自然纹理图像生成或图像风格化，如生成油画效果的图像。虽然其生成样本的多样性较高，但计算复杂度大，生成过程耗时。

(4)马尔可夫随机场(Markov random field，MRF)。通过图模型描述特征间的依赖关系，生成局部一致性较高的图像。其常用于纹理生成，能够生成与样本图像具有相似统计特性的纹理。然而，MRF对模型参数的依赖性较强，生成的样本多样性有限。

(5)Gibbs采样。通过迭代更新每个变量的条件分布,从联合分布中生成样本。该方法适用于生成高维数据,如图像和视频。但同样存在计算复杂度大的问题。

这些数学建模方法在特征层面的数据生成中提供了重要工具,但通常面临计算复杂度高和生成速度较慢的问题。如何平衡生成质量与效率仍是亟待解决的挑战。

2)图像特征融合。图像特征融合通过整合多个数据源或视角的高级语义信息(如边缘、纹理、形状等)生成包含丰富信息的新图像。其核心步骤包括特征提取、选择、匹配、融合和图像重构。相比图像粘贴,特征融合在特征空间中处理图像信息,可有效保留原图重要特征,并通过优化融合策略提升生成图像的质量和多样性,适应特定任务需求。

在学习方式上,图像特征级融合方法大体可以分为无监督学习和自监督学习两类。

在无监督学习类方法中,网络通常通过编码器和解码器结构(Prabhakar 等,2017;Li 和 Wu,2019)提取图像的特征(如图 5 所示),并使用特定的融合规则融合源图像之间的特征,进而生成新的图像。

在自监督学习类方法中,通过将源图像分解为共享特征和独特特征,再融合这些特征生成新图像。如图 6 所示,DeFusion 框架(Liang 等,2022)通过将源图像分解成所有图像的共享特征和独特特征,然后简单地组合这些成分来生成目标融合图像,不需要任何配对数据和复杂的损失函数设计(Zhang 等,2020b)。

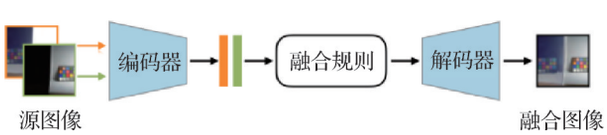

图 5　通过无监督方式进行图像融合(Liang 等,2022)

Fig. 5　Image fusion via unsupervised approach (Liang et al,2022)

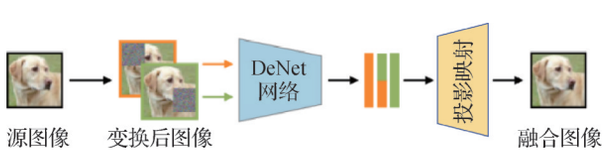

图 6　通过自监督方式进行图像融合(Liang 等,2022)

Fig. 6　Image fusion via self-supervised approach (Liang et al,2022)

此外,TransMEF 架构(Qu 等,2022)通过多任务学习训练编码器,提取多曝光图像的通用特征并在特征空间融合生成新图像,增强了图像的表达能力和模型的鲁棒性。

在融合策略上,图像特征级别融合方法主要包含特征金字塔、注意力机制和多任务学习 3 种方法。特征金字塔是一种多尺度特征提取方法,通过将图像分解为不同分辨率的特征图,以捕捉全局和局部信息;注意力机制方法通过动态权重分配,突出图像中关键区域或显著特征,同时抑制不重要的部分;而多任务学习则通过设计多个相关任务(如图像重建、分割、分类)共同训练模型,从而学习更加通用的特征表示。

1.1.3　小结

总体来说,作为最基础的图像数据生成手段,传统生成技术主要具备 3 个显著优点。首先,它们允许研究者精确控制合成数据的统计分布,从而在实验设计中实现高度的定制化和预测性。这一特点特别适合需要严格变量控制的实验研究场景。其次,与基于深度学习的方法相比,这些方法通常依赖明确的物理规则或数学模型,因此具有更高的可解释性,有助于结果的分析和验证。最后,传统数据生成方法对计算资源的需求较低,使其在资源受限的环境中依然高效运行,从而广泛适用于开发阶段或低成本任务中。

然而,这些方法也存在一些局限性。为了实现控制和生成的便利性,传统方法可能过度简化现实世界的复杂性,从而导致生成数据与真实数据分布之间存在偏差,可能对模型训练结果产生负面影响。此外,对于不同的下游任务,使用一些基于几何变换的方法进行数据生成在增加多样性的同时可能会破坏对象间的空间关系,从而引入人为偏差,进一步降低数据的真实性。例如,在目标检测任务中,随机裁剪或变形缩放可能导致目标在图像中的比例或位置不自然,从而无法准确反映实际应用场景中的目标分布特性(Du 等,2021)。更重要的是,在处理复杂或极端情况时,传统方法通常需要深厚的专业知识和复杂算法的支持,这提高了方法的实现门槛和开发成本。因此,尽管传统数据生成方法在可解释性和资源节约性方面具有一定的优势,但其在模拟复杂场景和捕捉数据分布细节上的能力不足,决定了其适用范围的局限性(Wang 等,2024c)。

1.2 三维渲染

1.2.1 虚拟引擎

为了应对真实世界中采集困难导致的数据稀缺问题，研究者们开始尝试向“虚拟世界”寻求解决方案。在游戏、影视制作行业中，技术人员会借助Blender、Unity、Unreal Engine等虚拟引擎构建三维虚拟数字世界。作为“创世者”，人们拥有在该虚拟世界中执行任意操纵和获取所有精确信息的权限，可以根据需求在其中设置场景、资产，并通过模拟传感器（例如相机、激光雷达）的采集机理，将三维虚拟世界的信息映射到指定的数据媒介（例如图像、点云）中。通过虚拟引擎生成的数据可以涵盖从简单的图像识别（Sun和Zheng，2019；Wang等，2020，2022d；Zhang等，2021b）到复杂的场景理解（Gaidon等，2016；Richter等，2016；Ros等，2016；Hu等，2019；Straub等，2019；Li等，2021b；Roberts等，2021）等多个层面，为研究人员提供了无限的可能性。

以行人再识别为例。行人再识别是一个跨相机行人图像检索任务，因涉及个人隐私问题，其数据的收集十分敏感。如图7上半部分所示，行人再识别真实数据的采集需要收集监控相机网络录制的视频，并对视频帧进行行人边界框检测，然后对裁剪出的行人图像的身份进行手动标注。其中，跨相机的身份标注十分考验标注人员的记忆力和专注力。尤其是身着相似款式服装的行人的区分，现有真实数据集或多或少存在错误标注。鉴于此，研究人员（Sun和Zheng，2019；Wang等，2020，2022d；Zhang等，2021b）借助虚拟引擎模拟监控场景实现数据采集。如图7下半部分所示，UnrealPerson（Zhang等，2021b）首先借助MakeHuman工具批量生成数字人，通过控制输入参数，MakeHuman可以制造出不同肤色、形态、着装的数字人，这保证了行人身份的多样性。之后，UnrealPerson（Zhang等，2021b）利用Unreal Engine中提供的示例场景，并根据经验部署相机进行监控记录。至于数据的标注，UnrealPerson（Zhang等，2021b）调用Unreal CV（Qiu等，2017）工具获取实例分割图像，并以行人分割掩码为依据进行行人图像裁取。因为在该场景中，行人身份是全局一致且固定的，因而可以直接获取行人图像的身份标注。实验表明，在该生成数据上训练的模型在现实场景中具备良好的泛化能力。

虚拟引擎的优势在于可控性和绝对的信息获取权限，这意味着该技术方案可以采集到现实场景中受制于传感器而难以获取的数据。人体的形状姿态估计需要精确的三维人体形态标签（通常是SMPL格式），而这一信息在真实世界中需要借助专业动作捕捉设备采集。为规避这一限制，新加坡南洋理工大学的Yang等人（2023c）基于Unreal Engine开发了一套人体合成数据采集管线：根据需求创建人体模型、将数据库中的动作重定向到生成的行人模型、合理地摆放人体、场景和相机，最终收集渲染的各种信息配置和输出，汇总得到最终的数据集。得益于虚拟引擎的优良特性，基于该管线可以采集准确的人体形状、姿态和标注。德国马克斯·普朗克研究所的Black等人（2023）和微软的Hewitt等人（2024）也采用类似的管线。相关实验证明，合成数据极大地增强了现有模型的性能。除了以人为中心的任务，在场景理解领域也有工作采用类似的管线进行数据采集。Hypersim（Roberts等，2021）是室内场景合成数据集的代表性工作，它基于V-ray开发设计了一套管线自动规划相机轨迹对专业艺术家设计的461个室内场景进行渲染，得到大量像素级标注的数据。至于室外场景，MatrixCity（Li等，2023d）基于Unreal Engine提供的黑客帝国城市场景，设计空拍和街景视角进行渲染，为大规模城市重建与新视角生成提供了宝贵基准。这两者都提供了精准的深度、反照率和材质标注，而这些标注在真实世界中是无法获取的。这为基于深度学习的场景理解提供了新的可能。

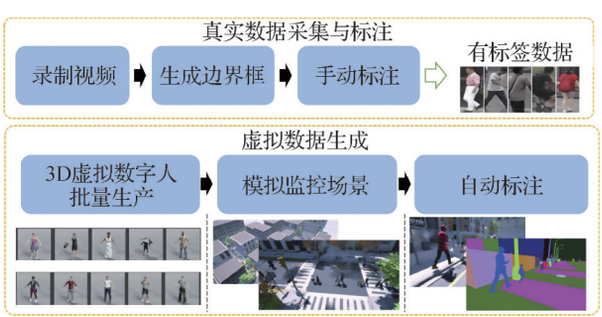

图7 UnrealPerson的数据采集流程（Zhang等，2021b）

Fig. 7 Data collection process of UnrealPerson (Zhang et al., 2021b)

除了渲染静态数据集的获取，虚拟引擎创作的三维场景还可以用于仿真器的基础环境。Unreal Engine和Unity甚至还自带物理仿真能力。常用的自动驾驶仿真引擎Airsim（Shah等，2018）和Carla（Dosovitskiy等，2017）就是基于Unreal Engine进行

开发的。这些仿真环境中的三维场景都是由艺术家精心设计得到的，这导致相关研究的高门槛，阻碍了相关研究的推广和普及。为此，美国普林斯顿大学 Raistrick 等人（2023，2024）的一系列工作尝试借助程序化内容生成技术实现三维场景的快速设计与开发。程序化内容生成技术（procedural content generation，PCG）是指利用算法自动创建内容的过程，它可以按照特定规则组合预设的高质量资产，高效地生成大量多样化且合理的数据或媒体内容。从某种程度上讲，程序化生成技术将美术设计任务转化为代码编程任务，这仍然需要开发者熟练掌握相关程序语言。为了进一步自动化程序设计的生产力，研究人员尝试借助大语言模型强大的分析规划能力来驱动 PCG 技术生成所需场景。3D-GPT（Sun 等，2024a）基于 Infinigen（Raistrick 等，2023）进行开发。受限于 Infinigen（Raistrick 等，2023）本身的建模质量，其渲染的视觉效果并不尽如人意。中国科学院自动化研究所开发的 SceneX（Zhou 等，2024b）和 CityX（Zhang 等，2024b）则以 Blender 社区丰富且专业的插件生态为基础，设计了全自动的程序化生成管线。以 CityX 为例，如图 8 所示，首先设计了一种管理协议来兼容具有不同接口和功能的插件，使得后续的统一调用成为可能。之后，Zhang 等人（2024b）基于大语言模型设计了一套多智能体框架，能够处理多模态输入指令（例如 OSM（Open Street Map）格式文件、文字描述、语义分割图和卫星图）。其中的每个智能体分别负责任务规划、规划验证和任务执行。初步搭建场景的渲染结果将输入到一个视觉语言大模型中分析生成缺陷，相关信息将反馈回多智能体框架，引导场景的精进和微调。该工作可以实现高效、多样、逼真和可控的三维场景生成。PCG 技术创作的场景遵循工业标准，这些三维的虚拟场景可以很方便地导入 Issac Sim 等仿真器中，为具身智能研究提供了近乎无限的训练和验证环境。

借助虚拟引擎进行数据生成的范式生成并非没有缺点。其中最大的难题之一是如何确保虚拟数据与真实世界的匹配度。尽管现代虚拟引擎已经能够在视觉效果上达到非常逼真的水平，但在某些细节处理上仍存在差距，如物理行为的真实感、纹理的细腻程度等。这些差异可能会导致模型在虚拟环境中表现良好，但在实际应用中表现不佳的问题。因此，如何缩小虚拟与现实之间的差距（sim-to-real）是未来研究的一个重要方向。

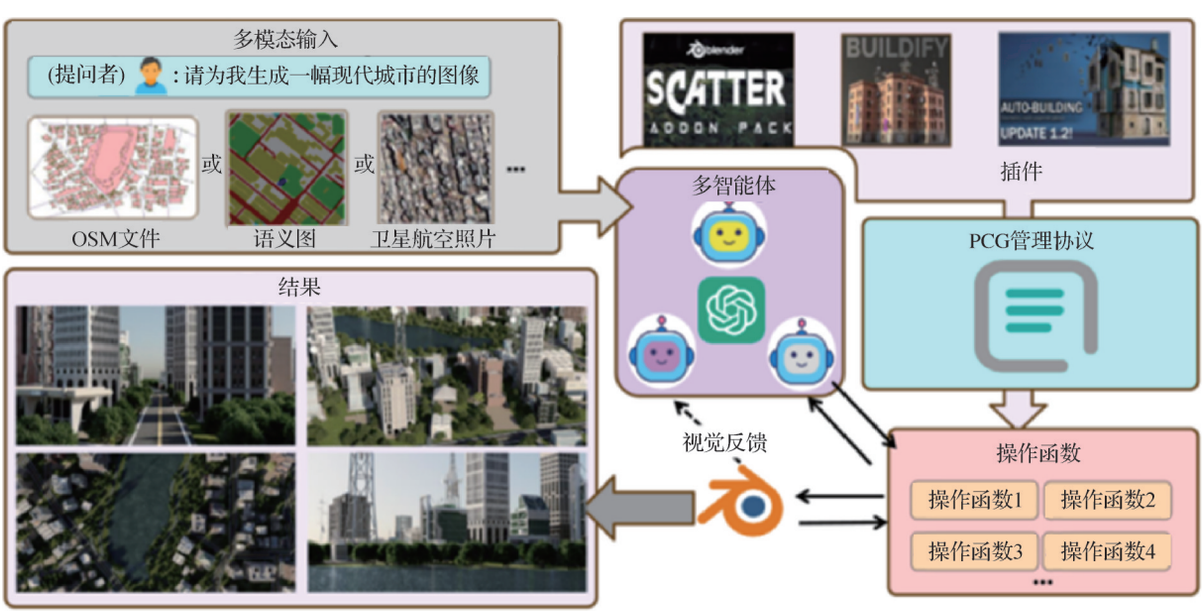

图 8　CityX 管线

Fig. 8　Pipeline of CityX

1.2.2　神经辐射场

神经辐射场（neural radiance fields，NeRF）由 Mildenhall 等人于 2020 年首次提出，用于新视图合成（Mildenhall 等，2021）。NeRF 采用体积渲染和隐式神经网络表示来学习 3D 场景的几何和光照，在复杂场景下也能合成具有高度照片真实感的视图。围

绕神经辐射场的数据生成方法可以大致分为原生三维方法和基于蒸馏先验的方法。前者通过将现有常用生成网络应用到神经辐射场表示，结合多目图像、三维模型等数据提供监督，训练适用于三维数据的生成网络；后者依赖于预训练的二维图像扩散模型，借助神经辐射场的可微渲染，构建三维表示与二维图像之间的桥梁，通过图像扩散模型的监督来反向优化神经辐射场的生成。下面从这两方面分别介绍一些相关工作。

1）原生三维方法。在二维图像生成中，VAE（Kingma 和 Welling，2013）、GAN（Goodfellow 等，2014）和扩散模型（Rombach 等，2022）已经展现了不错的生成能力。因此对于三维数据生成，将这些二维生成网络与具有高真实感图像合成能力的三维神经辐射场表示进行结合，训练前馈式三维生成网络，是一个可行的思路。其中，GAN 作为一个无监督训练生成网络，在三维数据缺乏的背景下，被大量研究应用于三维神经辐射场的生成。GRAF（generative radiance fields）（Schwarz 等，2020）将神经辐射场的渲染器看做一个以随机变量和相机参数作为条件的生成网络，引入图像判别器，利用生成式对抗网络架构实现给定视角下的图像生成，构建了一个三维感知的多目图像生成器。pi-GAN（Chan 等，2021a）进一步将 SIREN（sinusoidal representation networks）（Sitzmann 等，2020）作为核心表示，并引入了基于 StyleGAN（Karras 等，20219）的映射网络。遵循与 GRAF 类似的思路，CVPR2021 的最佳论文 GIRAFFE（Niemeyer 和 Geiger，2021）对场景中的物体采用不同神经辐射场进行解耦表示，不仅实现了对场景的无监督解耦生成，也实现了可控生成，可以增删物体、平移或旋转物体等。这些工作严格意义上并没有构建生成网络直接生成神经辐射场表示，而是利用可微渲染和 GAN 的训练机制来反向优化神经辐射场表示。为了能够直接生成作为隐式表示的神经辐射场，需要构建显式特征表示。EG3D（efficient geometry-aware 3D generative adversarial network）（Chan 等，2022）直接利用 StyleGAN（Karras 等，2019）生成 3 个正交的特征平面，由此形成了三维特征空间，并构建了双判别器架构来保证生成结果的三维一致性。Gao 等人（2022）进一步将该三平面表示扩展到一般三维数据的生成。三平面表示使二维卷积网络能够应用于三维数据的生成。因此，Rodin（Wang 等，2023c）和 SSDNeRF（single-stage diffusion NeRF）（Chen 等，2023b）将二维图像扩散模型（Rombach 等，2022）与三平面表示结合，利用三维合成数据作为监督实现了三维生成。除了三平面表示，GRAM（Deng 等，2022）和 GRAM-HD（Xiang 等，2023）的多个二维曲面构成的分层特征表示，PanoHead（An 等，2023b）的三网格表示都能与 StyleGAN 结合，实现高质量三维数据生成。HoloFusion（Karnewar 等，2023a）和 HoloDiffusion（Karnewar 等，2023b）将多尺度特征体素（Müller 等，2022）与扩散模型结合，实现三维数据生成。而 DiffRF（Müller 等，2023）直接在定义了体密度和颜色的三维体素上训练扩散模型。基于 GAN 的神经辐射场生成思路也被用于三维数字人生成（Noguchi 等，2022；Zhang 等，2022）和室内场景生成（DeVries 等，2021）。NeRF-VAE（Kosiorek 等，2021）则通过与变分自编码器的结合，实现了简单三维场景的生成。

另外，考虑到多数现实场景中很难为渲染对象采集大量的多视角数据，稀疏视角合成技术也得到了广泛研究。PixelNeRF（Yu 等，2021）通过卷积神经网络（convolutional neural network，CNN）提取图像特征作为条件输入，大大降低了对多视角数据的依赖，提高了 NeRF 的泛化能力。VisionNeRF（Lin 等，2023c）在 PixelNeRF 的基础上进一步考虑视图的全局特征，并用视觉 Transformer 模型（vision Transformer，ViT）（Dosovitskiy 等，2021）对特征进行层次化建模，可以进行效果良好的单视角图像渲染，并显著提高了模型在不同类别物体上的泛化能力。G-NeRF（Huang 等，2024c）则尝试引入几何先验，从单视角数据合成多视角数据，并利用深度图像监督来提升模型的深度感知，实现以单幅图像渲染出更真实的 3D 效果。同时，图像和视频扩散模型的发展以及多目数据（Yu 等，2023a；Wu 等，2024b）的增多也催生了利用多目图像生成实现三维数据生成的工作（Liu 等，2024c）。如 Wonder3D（Long 等，2024）通过在图像扩散模型中添加相机条件控制，同时生成多目图像和法向；SV3D（Voleti 等，2025）通过微调视频扩散模型实现环绕视角的多目图像生成。这些多目生成结果被进一步用于神经辐射场重建，完成最终的三维数据生成。

2）蒸馏二维先验方法。训练一个大规模的原生三维模型需要大量三维数据，目前可用的三维数据

仍然无法保证生成的泛化性，且公开数据集中的数据需要进一步被清洗和筛选才能保证生成质量。与三维数据不同，二维图像数据，尤其是文本—图像配对数据的规模相当大（Schuhmann 等，2022）。借助这些数据，文本—图像语义对齐模型 CLIP（contrastive language-image pre-training）（Radford 等，2021）被提出，并且可以作为监督，反向优化神经辐射场生成，如 DreamFields（Jain 等，2022）和 CLIP-NeRF（Wang 等，2022a）。而二维图像扩散模型已经能实现从文本到高真实图像的跨模态生成，因此通过蒸馏图像的生成先验实现三维数据的生成成为探索的方向。先驱性工作 DreamFusion（Poole 等，2022）引入了分数蒸馏采样（score distillation sampling，SDS）损失函数，通过对神经辐射场的渲染图像添加随机噪声，并在文本引导下预测噪声，构建梯度反向优化神经辐射场，实现跨模态三维生成。在 SDS 的基础上，MVDream（Shi 等，2024）通过微调扩散模型引入三维先验来增强生成结果的三维一致性。SJC（score Jacobian chaining）（Wang 等，2023a）和 Prolifc-Dreamder（Wang 等，2023h）等一些工作（Liang 等，2024b；Yu 等，2023b；Zhu 等，2024a）尝试对 SDS 的公式进行改进，增强生成质量和一致性。Magic3D（Lin 等，2023a）、Fantasia3D（Chen 等，2023c）、RichDreamer（Qiu 等，2024）等工作则将生成过程分步骤（先神经辐射场后网格）或分属性（几何、材质等）解耦，从而提升生成质量和可控性。基于这些文本到三维数据的生成方法，将图像作为额外的输入可以实现图像到三维数据的生成（Deng 等，2023；Gu 等，2023）。这类方法主要包括两种方式：一是将图像输入到经过微调后的扩散模型中，例如 Zero123（Liu 等，2023）；二是将图像监督作为额外的约束信号，例如 Magic123（Qian 等，2023）和 Make-it-3D（Tang 等，2023）。

1.2.3 3D 高斯泼溅

3D 高斯泼溅（Kerbl 等，2023）技术能够从多视角图像中快速恢复高质量三维场景，并能够对进行场景高质量的实时渲染，在三维重建领域内获得了广泛的关注。近年来在三维重建与视图合成任务中首先取得成功的方法是隐式神经辐射场（Mildenhall 等，2021；Barron 等，2022）。隐式神经辐射场使用神经网络编码场景的几何与颜色信息，通过在空间中进行离散的采样，使用体渲染绘制图像，仅通过图像损失函数即可完成重建任务。与隐式神经辐射场不同，3D 高斯点云通过显式的、离散的点云表达场景信息，每个高斯点绑定有参数 $\{x_i, r_i, s_i, o_i, C_i\}$，依次表示高斯点的三维位置、旋转、尺度、透明度和用球面谐波函数编码的颜色。

如图 9 所示，3D 高斯点云以 SfM（structure from motion）稀疏点云与相机参数为输入，通过可微分的光栅化渲染器对高斯点云进行体渲染，根据图像损失优化每个高斯点的参数。其中自适应密度控制（图 10）能够有效地对重建不足或重建过度的区域进行高斯点稠密化，并去除漂浮的低透明度高斯点。得益于显式点云表达，3D 高斯泼溅相较隐式神经辐射场在训练速度与渲染速度上都有十足的提升，能够将重建训练的时间从数小时缩短到几分钟，并将渲染速度提高了近百倍，支持对较大场景的实时自由视角渲染。

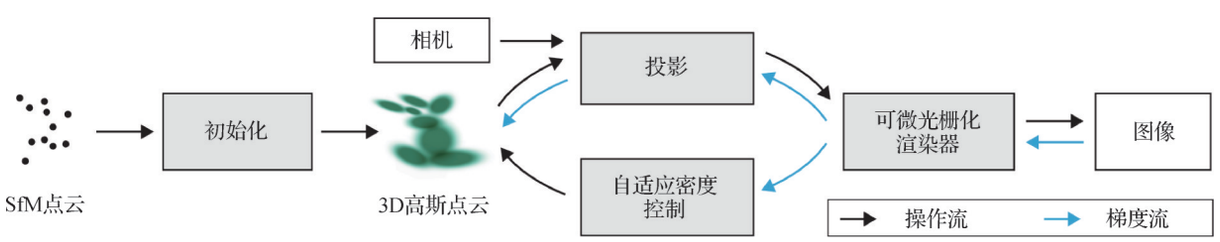

图 9　3D 高斯泼溅总体流程（Kerbl 等，2023）

Fig. 9　Overall process of 3D Gaussian splash (Kerbl et al., 2023)

针对 3D 高斯泼溅方法中仍然存在的不足，一系列后续研究尝试在存储效率与训练速度方面对 3D 高斯泼溅进行优化。3D 高斯泼溅用独立的点属性去表示场景的几何与外观，导致所需存储的数据量较大。通常一个中等规模的户外场景（Barron 等，2022）需要数百 MB 到几 GB 的存储空间。近期研究通过向量量化，将不同高斯点中相似的属性值在密码本中存储一次来降低模型占用的存储空间。Fan

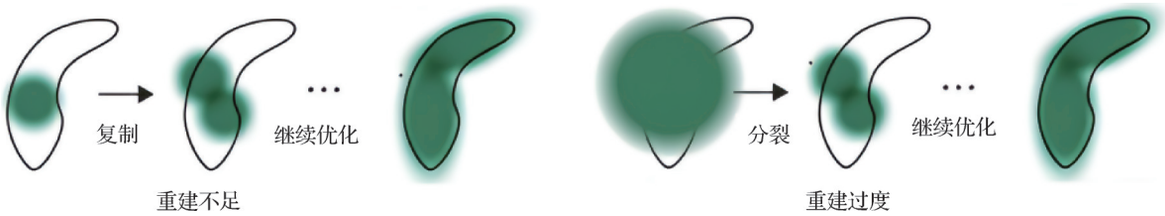

图10　3D高斯泼溅中的自适应稠密化过程(Kerbl等,2023)

Fig. 10　Adaptive densification process in 3D Gaussian splashing(Kerbl et al., 2023)

等人(2024)、Girish等人(2024)、Lu等人(2024)发现3D高斯泼溅中的点云稠密化过程带来了大量的冗余点。Girish等人(2024)、Navaneet等人(2025)、Niedermayr等人(2024)设计了更有效的自适应分裂与删除规则,实现用更少的点对场景进行相同质量的建模,同样能有效降低模型的存储和渲染的开销。虽然3D高斯泼溅已经达到较快的训练速度,在导数逆向传播过程中的像素级别原子操作仍然占据了较多的时间开销,因此有研究(Durvasula等,2023;Mallick等,2024)设计了更快的并行策略,减少原子操作的次数进而加快了3D高斯泼溅的重建训练速度。

高质量的3D高斯泼溅建模通常需要上百幅训练图像,在训练视角较为稀疏的情况下3D高斯模型的渲染质量退化明显,因此提升稀疏视角下的重建质量也成为一个有价值的研究方向。一些方法(Charatan等,2024;Chen等,2025a;Li等,2024b;Xu等,2025;Zhang等,2025b;Liu等,2025)将3D高斯点云表达与深度预测模块结合,通过在大量不同类型场景数据中训练学习跨场景的先验信息,从而实现在新场景中基于少量图像推理高斯点云,大幅提高稀疏视角下的重建质量。如图11所示,此类方法能够在原版3D高斯泼溅无法工作的极端情况下(仅有2幅输入图像)进行场景重建和新视图合成。由于少量的监督视角对三维场景缺乏足够的约束,3D高斯模型难以学习到三维一致的结果,导致在新视角下的渲染质量降低。因此,Zhu等人(2025)使用2D扩散模型从少量的输入图像生成具有三维一致性的更多新视角图像作为3D高斯点云重建时的额外监督,提高了稀疏视角下的重建质量。类似地,Chen等人(2025b)使用Point Transformer(Zhao等,2021)为3D高斯点云引入空间先验约束,增强高斯点云的三维一致性,大幅提升在外插视角下的渲染质量。

一系列后续工作专注于改善3D高斯泼溅方法在反走样、几何精度、反射光与大场景重建、排序误

图11　仅2幅输入图像下的稀疏重建(Xu等,2025)

Fig. 11　Sparse reconstruction with only 2 input images (Xu et al., 2025)

差等不同方面的重建效果。Song等人(2024)、Yan等人(2024b)、Yu等人(2024d)通过多尺度的学习和改进3D高斯泼溅渲染中的滤波过程,改善了3D高斯点在光栅化时的走样问题,提升了在远近不同尺度下的重建与渲染质量。Dai等人(2024)、Huang等人(2024a)将3D高斯椭球扁平化为2D椭圆,并引入基于法向量的几何约束,提高了重建的表面几何准确性。Jiang等人(2023b)、Ye等人(2024)分别应用法向预测模块与法向局部传播策略提升3D高斯模型几何法向的准确度,进而提升了对高光与镜面反射区域的建模质量。Kerbl等人(2024)对高斯点云建立层级结构,能够对不同细节层次的高斯点云进行高效训练和渲染,提高了3D高斯泼溅方法在大规模场景重建任务中的效率和质量。Chen等人(2024a)通过对空间进行划分,同样使用不同细节层次的表达方式对室外大场景进行建模,并结合扁平化的高斯点提升了大场景下几何表面的重建质量。3D高斯泼溅根据高斯中心点的深度排序进行体渲染,而中心点排序与光线实际接触到的高斯点顺序不一定一致,这导致在视角变化时渲染结果存在因为排序误差带来的颜色突变,Mai等人(2024)、Radl等人(2024)通过改进对高斯点排序与累加的方式改善了训练和渲染中的此类问题。

得益于3D高斯点云表达优秀的重建效率与质量,许多后续工作将对其进行扩展后应用到了动态场景重建、三维模型生成、自动驾驶和物理模拟等不同的领域中。一些研究通过在3D高斯表达中加入

时序信息实现了对动态场景的重建，其中，Xu 等人（2024b）、Wu 等人（2024a）、Yang 等人（2024a）使用空间变形网络隐式地表达高斯点属性在时序上的变化，Duan 等人（2024）、Yang 等人（2024b）使用四维的高斯点显式地对时序信息进行建模。一些研究将 3D 高斯泼溅与扩散模型结合，能够从文字描述或图像参考中生成用 3D 高斯点云表达的高质量物体（Tang 等，2024；Yi 等，2024）与场景（Liang 等，2024a；Zhou 等，2024d）模型。Yan 等人（2024a）、Zhou 等人（2024c）结合深度与语义视觉模型，使用 3D 高斯泼溅对大规模道路与车辆进行重建，实现了能够用于自动驾驶任务中的高质量场景表达。Xie 等人（2024）将 3D 高斯点云应用于物理仿真，实现了符合牛顿力学规律且所见即所得的运动效果合成。

1.3 深度生成模型

近年来，以生成对抗网络和扩散模型等代表的深度生成模型通过生成高质量图像数据，为视觉任务提供了更丰富的数据资源，推动了计算机视觉的发展。下面将展开介绍这些典型模型。

1.3.1 自编码器/变分自编码器

自编码器（auto encoder，AE）是一种旨在通过无监督学习提取数据低维表征的神经网络结构。其发展脉络如图 12 所示。1985 年，Rumelhart 等人（1986）首次提出自编码器的概念，作为一种学习和重构输入数据的神经网络结构。自编码器的基本结构由一个编码器和一个解码器构成。编码器接受高维数据输入，通过线性层或更为复杂的神经网络，将复杂高维分布中的数据映射到一个低维的潜在空间；随后解码器通过逐步扩大低维数据的维度，尝试恢复出接近于输入数据的高维表示。通过调节编码器与解码器的参数，使得从潜在空间的表示重建出的数据尽可能接近输入数据，从而完成对数据的压缩及有效重构。2010 年，Vincent 等人（2010）提出去噪自编码器（denoising autoencoder，DAE），通过在训练过程中引入噪声，增强模型的鲁棒性，能够从噪声数据中恢复出清晰的数据。2011 年提出稀疏自编码器（sparse autoencoder，SAE），通过限制神经元的激活数量实现对输入数据的稀疏表示。

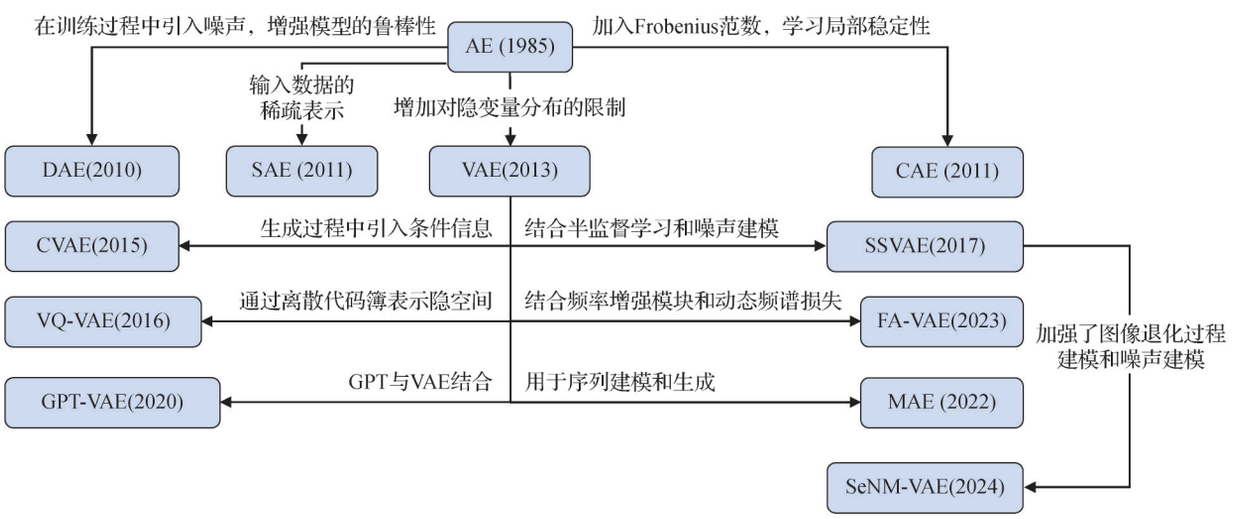

图 12 AE/VAE 模型发展脉络图

Fig. 12 AE/VAE model development map

VAE（variational auto encoder）由 Kingma 和 Welling（2013）于 2013 年提出，是一种将自编码器与变分推理相结合以用于新数据生成的技术。VAE 对输入数据进行编码映射到潜在空间，通过解码器复原。与自编码器不同的是，VAE 学习数据在潜在空间中服从的某种形式已知的概率分布（通常为高斯分布），以便在潜在空间中进行编辑、采样等操作，从而实现对原数据的编辑或对新数据的生成。

随后，VAE 技术迅速在数据生成领域得到广泛应用，产生了多种不同数据表示和不同网络结构的 VAE。在数据表示上，Sohn 等人（2015）提出的条件变分自编码器（conditional variational autoencoder，CVAE）在生成过程中引入条件信息（如类别标签）。通过利用这些条件，CVAE 能够生成符合特定条件的数据，从而提高生成结果的可控性和多样性。Van den Oord 等人（2017）提出的 VQ-VAE（vector

quantized-variational autoencoder)则通过引入一个离散的代码簿表示数据的隐空间(latent space),使得每个数据点的隐向量可以被量化为代码簿中的一个离散代码,增强了生成模型捕捉数据中的离散结构的能力,如生成语言、图像和音频中的序列数据等。2016年,β-VAE被提出(Higgins等,2017),用于高效描述二维黏性流体的周期和混沌状态。2017年,半监督变分自编码器(semi-supervised variational autoencoder,SSVAE)被提出,结合了半监督学习和噪声建模,用于高质量的图像修复任务(Xu等,2017b)。此外,VAE在网络结构上也衍生出了众多不同模型。GPT-VAE将GPT(generative pre-trained Transformer)与VAE相结合,使用预训练的Transformer(Vaswani等,2017)作为VAE的解码器,生成更高质量的文本和序列数据。

近年来,基于各式变分自编码器的模型被广泛地应用于处理生成图像数据或其他模态数据的下游任务中。进入2019年以后,VAE技术不断拓展并应用于各个领域。2022年,He等人(2022a)提出MAE(masked autoencoder),这是一种用于序列建模的自动编码器变体,具有生成连贯且上下文适当的文本或视频的能力。2023年,Lin等人(2023b)提出频率增强VAE(frequency augmented variational autoencoder,FA-VAE),通过频率增强模块和动态频谱损失进一步提升了图像重建的质量。2024年,Zheng等人(2024a)推出了改进版SeNM-VAE,其在图像退化建模和噪声建模任务中表现出色。通过将VAE应用于图像的隐空间表示,SeNM-VAE能有效生成逼真的图像数据,特别是在图像修复任务(如去噪和超分辨率)中具有优异的表现。此外,2024年,Solera-Rico等人(2024)提出一种结合β-变分自编码器和Transformer的方法,应用于流体动力学领域,成功地实现了对二维黏性流体周期和混沌状态的高效建模。

1.3.2 生成对抗网络

生成对抗网络自从2014年由Goodfellow等人(2014)提出以来,已经成为生成模型领域的核心研究方向之一。GAN的基础思想非常直观,借鉴了博弈论中的对抗思维,将两个神经网络对抗训练,生成高质量的伪造数据。GAN的独特性在于其生成模型的能力,即从随机噪声中生成逼真的样本,同时也开辟了图像生成、图像修复、风格迁移和文本生成等多领域的应用。

GAN由两个核心部分组成:生成器(generator)和判别器(discriminator)。两者之间进行对抗训练,使得生成器能够逐渐生成与真实数据非常相似的伪造数据,而判别器则尽力将真实数据和伪造数据区分开。这个对抗过程持续进行,生成器与判别器交替训练,直到生成器能够生成足够逼真的样本,使得判别器无法轻松区分。

GAN作为一种强大的生成模型,自2014年提出以来,在多个领域展现了出色的应用能力。图13展

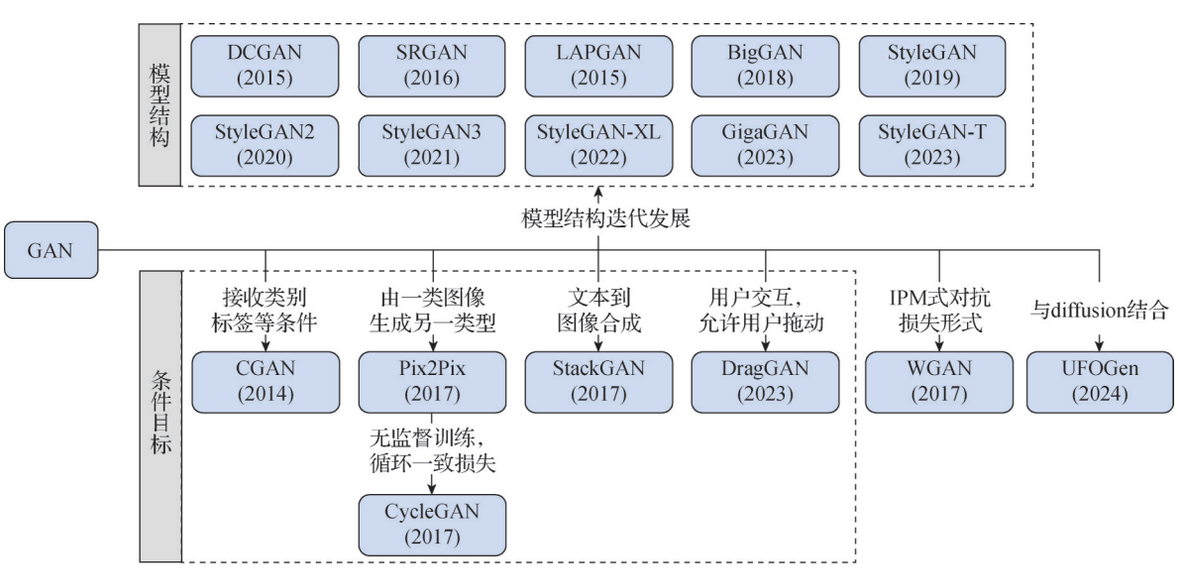

图13 GAN生成模型的代表性工作及发展

Fig. 13 Representative work and development of GAN generation models

示了以GAN为基础的生成模型代表性工作的发展脉络。2014年提出的条件生成对抗网络(conditional GAN,CGAN)(Mirza,2014)为GAN模型引入了条件生成的概念,使得生成过程能够基于额外的条件信息(如标签)进行控制。这为生成模型在实际应用中提供了更高的灵活性。2015年,DCGAN(deep convolution generative adversarial network)提出利用CNN替代传统的全连接层,显著提高了图像生成质量和生成稳定性。紧接着,LAPGAN(Laplacian pyramid of generative adversarial network)(Denton等,2015)引入了多尺度生成结构,使得GAN能够生成更高分辨率的图像,而SRGAN(super-resolution generative adversarial network)(Ledig等,2016)则专注于低分辨率图像的超分辨率(super-resolution,SR)重建,为图像质量的提升做出了重要贡献。

进入2017年,GAN模型的训练稳定性成为研究的关键问题,WGAN(Wasserstein GAN)(Gulrajani等,2017)通过引入Wasserstein距离,解决了原始GAN训练中常见的梯度消失问题,从而显著提高了训练的稳定性和收敛性。同年,StackGAN(Zhang等,2017)提出通过分阶段的生成过程合成文本到图像的能力,开启了文本驱动图像生成的新方向。Pix2Pix(Isola等,2017)通过监督学习实现了图像到图像的转换,推动了图像翻译技术的发展,而CycleGAN(Zhu等,2017)则通过无监督学习实现了从一种图像风格到另一种图像风格的转换,极大地扩展了GAN在跨域图像生成中的应用。BigGAN(Brock等,2018)作为一种大规模的GAN模型,通过引入更深层次的网络结构和更大的训练数据集,能够在ImageNet数据集上生成更高质量、多样化的图像,显示了GAN在高分辨率图像生成方面的潜力。GAN逆映射(Zhu等,2016)是指将生成对抗网络中的生成器反转,使其能够从给定的图像中找到对应的潜在向量,即将图像映射回生成器的潜在空间。研究者针对CGAN、StyleGAN等GAN结构分别提出相应的逆映射设计(Perarnau等,2016;Abdal等,2019)。这一技术为许多基于GAN的应用提供了基础支持,如图像编辑、风格迁移和高质量图像重建等。

随着GAN技术的成熟,更高质量和精细控制的生成模型应运而生。StyleGAN模型(Karras等,2019)是GAN的一大进步,它在图像生成任务中引入了样式变换,使得生成器可以控制生成图像的不同属性,如发型、表情、光照等,从而生成高质量的图像。随着技术的发展,StyleGAN2(Karras等,2020)通过改进生成网络结构,进一步减少了假象(artifact)和模式崩溃问题,使得生成图像的质量大幅提升。StyleGAN3(Karras等,2021)在效率和视觉效果方面进一步优化,能够生成更加稳定和高质量的图像。StyleGAN-XL(Sauer等,2022)通过扩展模型的规模和复杂度,进一步提升了图像分辨率,能够生成极高分辨率的图像,适应更高质量的创作需求。

随着生成模型应用的多样化,越来越多的创新出现在GAN的应用领域。StyleGAN-T(Sauer等,2023)通过结合CLIP模型,使得低分辨率图像的生成速度得到了显著提升,并能够实现更高效的文本引导图像生成。GigaGAN(Kang等,2023)在文本到图像的合成上实现了突破,不仅提升了生成效率,还能够生成更高分辨率的图像,特别适合应用于高质量创作与内容生成中。DragGAN(Pan等,2023)进一步扩展了生成模型的交互性,允许用户通过简单地拖动图像中的点,精确控制图像的姿态、形状和表情等细节,为用户提供了更加灵活的图像编辑与操控方式。2024年,UFOGen(Xu等,2024c)将扩散模型与GAN目标相结合,通过创新的生成策略,在文本到图像的生成上实现了高质量的图像输出,并具备更快的生成速度。

1.3.3 自回归模型

自回归模型(autoregressive model,AR)是一类机器学习模型,通过对序列中先前的输入进行测量来自动预测序列中的下一个分量。自回归通常被用做一种时间序列分析的统计技术,它假设时间序列的当前值是其过去值的函数,即将当前时刻的输出建模为过去时刻输出的概率函数来生成数据。

视觉自回归模型的代表性工作及发展历史如图14所示。典型自回归模型如循环神经网络(recurrent neural network,RNN),其直接通过神经网络来循环建模这一条件概率分布。然而在过长时间序列的建模中,根据梯度反向传播的链式法则,传统RNN会出现梯度消失或爆炸的问题。为此,后续研究者陆续提出长短时记忆(long short-term memory,LSTM)、门控循环单元(gated recurrent unit,GRU)等工作缓解这一问题。在计算机视觉的应用中,代表性工作包括PixelRNN(van den Oord等,2016b)和

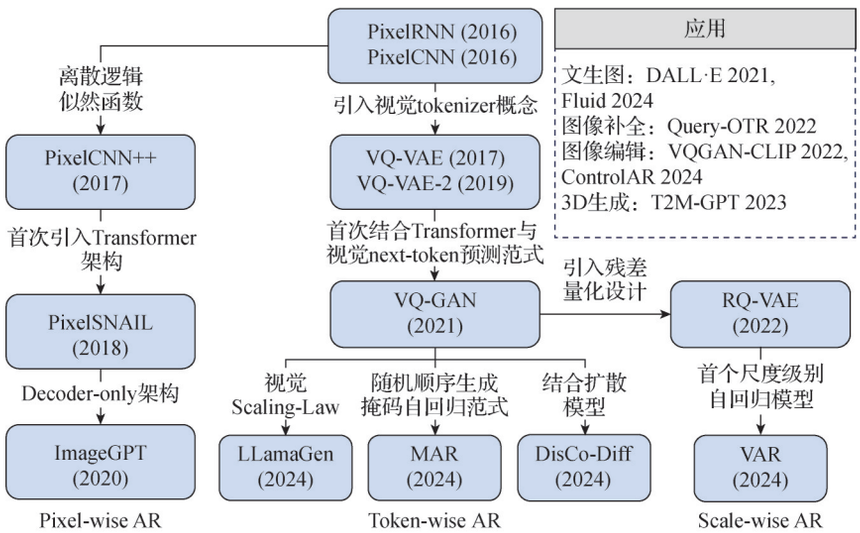

图14　视觉自回归模型的代表性工作及发展

Fig. 14　Representative work and development of visual autoregressive models

PixelCNN（van den Oord等，2016a），后续2018年PixelSNAIL（Chen和Hays，2018）首次引入Transformer架构并应用至自回归视觉生成。

上述工作主要基于像素层面进行自回归生成。然而像素对象由于缺少有意义的语义信息，往往无法达到令人满意的效果。因此，VQ-VAE系列工作（van den Oord等，2017）提出视觉tokenizer的概念，并催生了后续VQ-GAN（vector quantized generative adversarial network）（Esser等，2021）的这一里程碑式的工作。VQ-GAN首次结合Transformer与视觉next-token预测范式，在高分辨率图像生成上取得了令人印象深刻的结果。同时，近年来仍然有许多研究者在探索token层面的自回归生成范式。例如，香港大学Sun等人（2024b）提出LLamaGen，进一步优化了图像分词器的细节设计，通过高质量的训练数据和模型参数的提升，实现了图像生成质量的突破，通过Scaling Law证明了自回归生成模型的潜力，并超越了同期主流的基于Diffusion模型的工作；来自美国麻省理工学院等单位的研究者（Li等，2024d）重新考虑自回归生成模型与图像向量量化之间的关系，提出一种无向量量化的自回归图像生成方法，指出了其相比离散化令牌编码的优势，同时进一步将标准自回归模型和掩码生成模型统一到一个广义自回归框架下，称为掩码自回归（masked autoregressive，MAR）模型，其以随机顺序自回归并可以同时预测多个输出令牌，保持了“基于已知预测下一个令牌”的自回归本质，且能无缝集成扩散损失，实现高效的图像生成；DisCo-Diff（Xu等，2024c）则通过增加可学习的离散潜势来增强扩散模型，采用自回归Transformer模拟这些离散潜向量的分布，在ImageNet基准测试中获得最先进的FID（Fréchet inception distance）分数。

此外，作为近年来的关键进展，最新工作VAR（visual autoregressive modeling）（Tian等，2024）证明了除token级别上的自回归，图像尺度上的自回归可以达到比肩甚至超越Diffusion模型的生成图像质量。其背后的基本技术来自于韩国浦项科技大学研究者（Lee等，2022）提出的一种残差量化变分自编码器（RQ-VAE）方法，它使用多层残差量化从粗到细地精确逼近特征图，同时通过残差量化Transformer网络预测下一个残差编码，能够进一步降低特征的空间分辨率，有效提升图像生成的质量和效率。VAR（Tian等，2024）在RQ-VAE分层和残差式量化概念的基础上，引入了一种更简洁的尺度量化方法。VAR采用一种普通的VQ-GAN（Esser等，2021）对潜在空间中的特征映射进行编码，并进行插值计算残差，将特征量化为多个分辨率逐渐递增，即多尺度的token映射，最终达到与原始特征分辨率匹配的结果，并解码获得最终图像。与RQ-VAE类似，VAR对每个尺度相对于前一个尺度的残差进行量化，残差在不同尺度之间共享相同的码本。通过这种分层方法，VAR将图像表示为一个从粗到细的比例顺序。该序列的量化过程完全符合因果关系，同时每个自回归单元保持空间局部性，在大规模训练

中取得了比肩甚至超越Diffusion的生成质量。

同时，有许多工作从应用层面拓展了自回归的潜力。DALL-E（Ramesh等，2021）采用变分自编码器，将文本通过预训练的文本编码器映射到隐空间，再通过解码器将文本向量映射到图像空间，从而完成了文本控制下的图像生成；Query-OTR（Yao等，2022）则探索了自回归在图像补全任务下的应用；ControlAR（Li等，2024e）关注图像编辑这一任务，通过在解码过程中引入空间控制信号（如边缘信息和深度图），实现对图像标记的细致且精确的操控；此外，一些工作（Zhang等，2023a）证明了自回归范式在3D数据，例如3D人体运动生成领域也可以取得令人瞩目的结果。

1.3.4 流生成模型

概率密度估计在许多机器学习问题中有着重要应用，但有很大的实现难度。例如，在深度学习模型中，需要进行反向传播，因此嵌入的概率分布（如后验分布）应足够简单，以便于计算导数。所以在隐变量生成模型中常使用高斯分布。

正则化流（normalizing flow）的数学概念最早于2010年由Tabak和Vanden-Eijnden（2010）提出。2015年，Rezende和Mohamed（2015）在变分推断框架下首次将正则化流推广开来。正则化流通过一系列可逆的变换，将简单分布转化为复杂分布，经过一连串变换后，最终得到目标变量的概率分布，实现对复杂分布的更好近似。

规范化流模型通过可逆变换将简单分布转化为复杂分布，用于密度估计和生成式模型的训练。如图15所示，早期的流模型主要采用简单的线性形式的流函数，例如对角矩阵、上三角矩阵等，优点在于其逆函数和雅可比行列式易于计算，缺点则在于流的表达能力受到变换形式的限制，难以构造更为复杂的分布到高斯分布的映射。Householder流（Tomczak和Welling，2016）利用Householder变换实现了对正交矩阵的参数化，使得训练正交形式的线性流成为可能，提高了流的表达能力。但是，线性形式的变换仍然难以满足实际应用映射复杂分布的需求。

为了突破线性变换对流的表达能力的掣肘，残差流、耦合流等更多形式的流变换逐渐得以研发。Planar and Radial Flow（Rezende和Mohamed，2015）是残差流的最早尝试，通过构造残差形式的流实现分布间更为复杂的变换。但Rezende和Mohamed（2015）构造的残差流的逆变换无法得到闭式解，导致采样生成涉及到的计算十分烦琐。Sylvester流（van den Berg等，2018）通过在线性层上先后作用非线性函数和新的线性变换层实现了含有类似“隐藏层”的流函数，在保证残差带来的更强表达能力的同时使流的逆变换易于计算。在此基础上，RevNets模型（Gomez等，2017）首次在残差流引入网络形式的流函数，进一步拓展了表达能力。

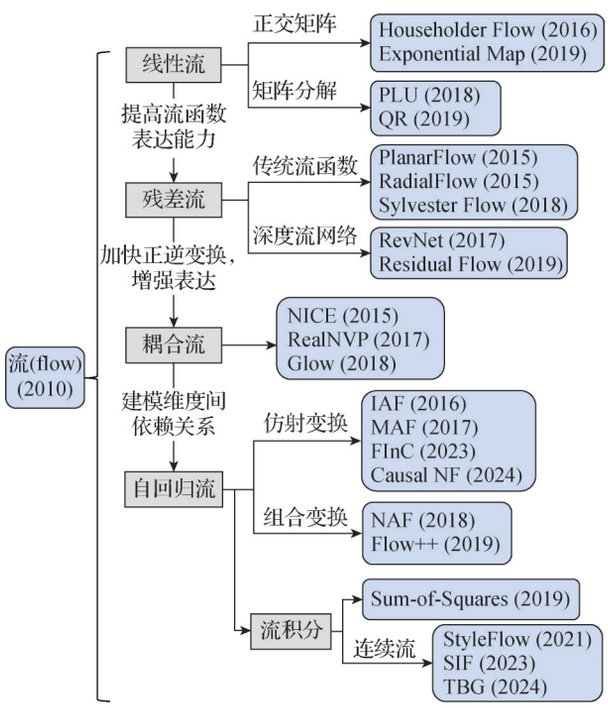

图15 基于流的生成模型的代表性工作及发展

Fig. 15 Representative work and development of flow-based generative models

耦合流（coupling flow）指用于实现分布变换的流函数由前序分布决定的流，具有强大的分布变换和表达能力。代表性的耦合流如RealNVP模型（Dinh等，2017）通过堆叠仿射耦合层实现分布间的转换，确保易于计算雅可比行列式，首次实现了较高质量的自然图像的生成。NICE（non-linear independent component estimation）模型（Dinh等，2015）是RealNVP的前身，使用的是加性耦合层，不包含缩放项。Glow模型（Kingma和Dhariwal，2018）进一步优化，通过可逆1×1卷积和激活归一化，使得模型在多层结构中更高效地捕获复杂分布。

近年来，正则化流的研究主要集中于在耦合流思想的基础上诞生的自回归流（autoregressive flow）。MAF（masked autoregressive flow）（Papamakarios等，

2017)在流生成的过程中引入了掩码(masking)技术,用层叠自回归流实现了快速密度估计。NAF(neural autoregressive flow)(Huang等,2018a)将自回归流函数统一为单调神经网络。Grathwohl等人(2018)利用ODE(ordinary differential equation)生成时序连续流,实现了高效快速采样生成。Flow++(Ho等,2019)进一步提升了生成视觉效果,达到接近自回归模型(autoregressive model)的生成性能。2021年,StyleFlow(Abdal等,2021)被提出,在GAN的潜在空间中实例化连续流,提升了对人脸的生成能力。

正则化流已在多个领域中展现了强大的应用能力。它通过增强卡尔曼滤波模型(de Bézenac等,2020),改善多元时间序列的预测精度,并应用于句子嵌入任务(Li等,2020),将BERT(bidirectional encoder representations from transformers)的嵌入分布转化为高斯分布,提升文本相似性任务的表现。此外,正则化流还在解决逆问题和量化不确定性方面表现突出。如,SRFlow模型(Lugmayr等,2020)利用正则化流在图像超分辨率任务中取得了显著效果。同时,基于流的更多模态数据生成也得到了探索,例如ManiFlow模型(Postels等,2022)将正则化流应用于3D点云生成,Klein和Noé(2025)将流与Boltzmann生成器结合,应用于分子系统近似采样生成等。

1.3.5 扩散模型

扩散模型是多类模型的理论统一,包括去噪扩散概率模型、分数匹配模型、一致性模型和流匹配模型。扩散模型的理论演进与工程实践历程如图16所示。其中去噪扩散概率模型的建模形式的流行程度相对最高。去噪扩散概率模型首先通过逐步向原始图像加入高斯噪声的方式构建一系列条件概率分布,形成了从自然图像分布到高斯噪声分布的映射,再通过一个神经网络拟合去噪过程的条件概率函数,形成由高斯噪声分布到自然图像分布的映射。从自然图像分布到高斯噪声分布的映射称为前向过程,从高斯噪声分布到自然图像分布的映射称为反向过程,如图17所示。

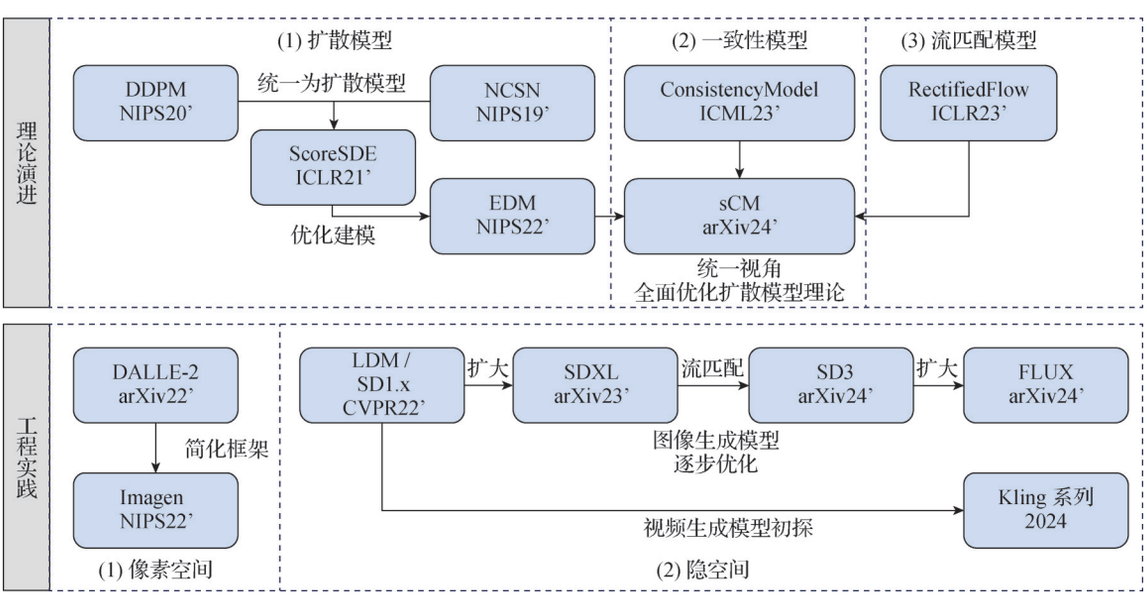

图16 扩散模型的代表性工作和发展

Fig. 16 Schematic diagram of the development process of diffusion models

扩散模型构建的从高斯分布到图像分布的映射,可以精准拟合复杂而多样的未知分布,从而实现生成高质量图像的目标。

除了去噪扩散概率模型之外,分数匹配模型(Song和Ermon, 2020)的建模方式也是扩散模型的一类重要分支。得益于其更加精准的建模形式,许多扩散模型的理论推进是基于分数匹配模型而展开的。后续,Song等人(2021)将分数匹配模型与去噪扩散概率模型完成了理论统一,并进一步提出扩散模型采样过程的常微分方程建模方式(Diffusion-ODE),提出扩散模型的稳定采样方法。针对Diffusion的常微分建模,DPM-Solver(Lu等,2022)给出了其一阶、二阶、三阶近似解,大幅减少了扩散模型所需的采样步数,并证明了DDIM(denoising diffusion

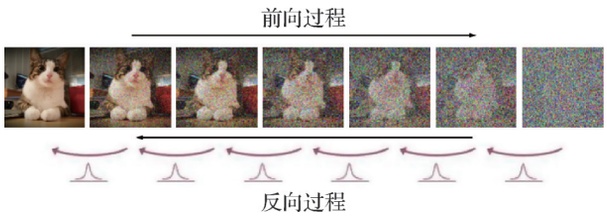

图 17　去噪扩散概率模型示意图(前向:加噪;反向:去噪)

Fig. 17　Schematic diagram of denoising diffusion probability model. The figure shows the forward process of noise addition and the reverse process of denoising

implicit model)(Song 等,2022)是 DiffusionODE 的一阶近似,完成了 DDIM 和扩散模型的理论统一。进一步地,Consistency Model(Song 等,2023)简化了扩散模型所建模的从源域到目标域的转换过程,从理论角度减少了扩散模型的采样所需步数。另一方面,RectifiedFlow(Liu 等,2022)设计了一种流匹配模型,探索了线性的从源域到目标域的转换流程,训练模型拟合这一线性过程,并从理论角度统一了扩散模型与线性的流匹配模型。最终,Lu 和 Song(2025)完整统一了扩散模型、一致性模型和流匹配模型,并从理论角度全面优化了一致性模型的稳定性,标志着扩散模型的理论发展臻于成熟。

扩散模型近年来已在计算机视觉的多个领域中取得卓越成就。在像素空间的图像生成领域,代表性工作 DALLE-2(Ramesh 等,2022)和 Imagen(Saharia 等,2022)有能力生成语义复杂的高质量开放世界图像。考虑到图像的冗余特性,潜在扩散模型(latent diffusion models,LDM)(Rombach 等,2022)提出先将图像转化为精致的隐空间特征,再在隐空间构建扩散模型的方式,提高扩散模型的训练与推断效率。SDXL(stable diffusion XL)(Podell 等,2023)将 LDM 的参数量与训练数据进行扩大,取得了更高的生成质量。进一步地,SD3(stable diffusion 3)(Esser 等,2024)将流匹配建模方式应用在隐空间,FLUX 则将 SD3 的质量进一步提升。在视频方面,Kling 系列模型(可灵)将基于隐空间的扩散模型应用在多样灵活引导的视频生成任务上,有能力生成具有较高的帧间一致性与较正确的物理交互性的视频,成为基于扩散模型的视频生成技术的初探。

2　典型计算机视觉任务中的数据生成与应用

基于第 1 节中的各种典型图像或视频数据生成技术,能够可控生成大规模、多样化的图像或视频数据。这些生成数据不仅可用于计算机视觉模型训练,提升其性能和泛化能力,也可以模拟各种极端或罕见场景,有效提升测试的全面性和可靠性。本节将根据典型计算机视觉任务,深入探讨图像增强、生物特征识别、个体分析、群体分析、自动驾驶、视频生成以及具身智能中的数据生成与应用,涉及的计算机视觉相关任务如图 18 所示。

2.1　图像增强

图像增强是计算机视觉任务的基础。本节将从图像超分辨率、暗光图像增强和雨雾图像增强 3 方面介绍数据生成在图像增强任务中的典型应用。

2.1.1　图像超分辨率

图像超分辨率重建旨在提升图像的分辨率,使图像更加清晰。通过将低分辨率图像重建为高分辨率图像,该技术可广泛应用于安全监控、医学成像和卫星摄影等领域。此外,这项技术可用于改善数字媒体的图像缩放质量、提高面部识别系统的准确性,

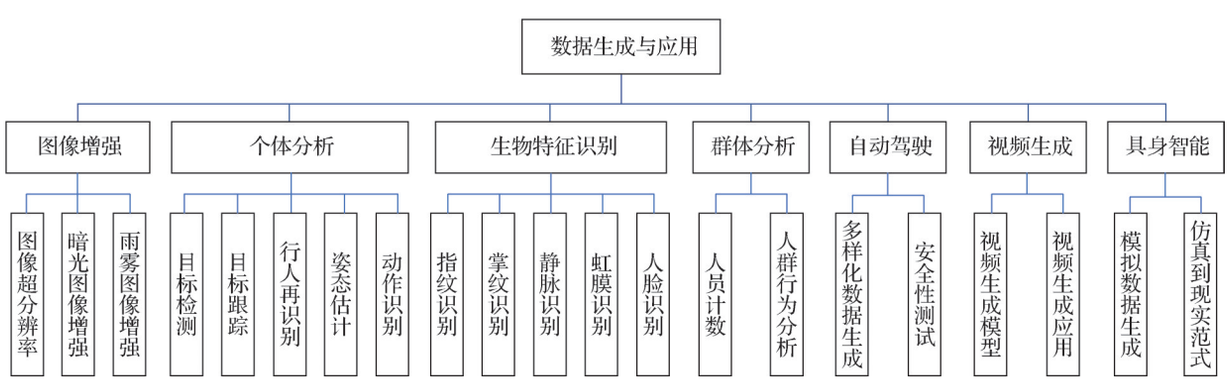

图 18　典型计算机视觉任务中的数据生成与应用

Fig. 18　Data generation and applications in typical computer vision tasks

以及增强AR(augmented reality)和VR(virtual reality)应用中的图像清晰度。

基于深度学习的超分辨率重建技术起源于2014年,Dong等人(2014)提出第1个基于卷积神经网络的超分辨率模型(super-resolution convolution neural network,SRCNN)。这一模型通过学习低分辨率(low resolution,LR)和高分辨率(high resolution,HR)图像之间的映射关系,展示了深度卷积神经网络在超分辨率重建中的有效性,为后续的研究奠定了基础。

随后,深度学习技术在超分辨率领域迅速发展,出现了多种创新的网络架构。Kim等人(2016)提出的VDSR(very deep super-resolution network)通过使用非常深的网络结构,进一步提升了重建图像的质量。此外,Tong等人(2017)提出的SRDenseNet利用密集连接机制,增强了网络内部的信息流动,进一步提升了重建精度。这些技术有效提高网络的学习能力和效率,大幅提升了超分辨率重建的质量。此外,研究者们还探索了如何利用多尺度学习、递归网络和GAN(Goodfellow等,2014)等技术增强模型的性能。

Transformer架构的引入为超分辨率重建领域带来新的研究机遇。Liang等人(2022)提出的SwinIR基于Swin Transformer,展示了在图像重建任务中的有效性。此外,Zamir等人(2022)提出的Restormer通过在Transformer中嵌入CNN,实现了多尺度的局部—全局学习,进一步提升了图像重建的质量。这些研究表明,Transformer架构在捕捉图像全局信息和长距离依赖关系方面具有显著优势,为超分辨率重建技术的发展提供了新的动力。

基于GAN的超分辨率工作在提升视觉感知质量、处理复杂降质场景以及应用泛化性方面均取得了显著成果。SRGAN(Ledig等,2016)通过生成器和判别器的对抗训练,成功生成感知上更加真实的高分辨率图像。ESRGAN(Wang等,2019b)对生成器进行了增强,采用了残差块结构并引入了对抗性感知损失函数,大幅提升了图像的细节保真度。Real-ESRGAN(Wang等,2021b)进一步在真实场景下优化,采用了更加鲁棒的降质模型,使得超分辨率技术能够处理复杂的现实降质情况。在此基础上,GLEAN(Chan等,2021b)利用预训练生成器与判别器,结合对图像内容的语义理解,实现了输入保真与细节重建的平衡,特别是在高模糊图像重建中表现出色。KernelGAN(Bell-Kligler等,2019)则专注于自监督的卷积核估计,通过无监督学习恢复图像退化过程中的卷积核,进一步提升了超分辨率的恢复效果。Maeda(2020)提出一种无需成对训练数据的生成对抗网络超分辨率方法,通过噪声和核校正网络以及伪成对超分辨率网络实现对真实低分辨率图像的超分辨率重建。

基于流模型的超分辨率方法在生成高质量图像、处理复杂图像变换以及保持生成过程的可逆性方面展现了独特优势。SRFlow(Lugmayr等,2020)是一种基于正则化流的超分辨率方法,通过学习条件分布解决了问题的多解性,能够生成多样的高质量图像。FKP(flow-based kernel prior)(Liang等,2021)是一种基于正则化流的核先验方法,通过学习各向异性高斯核分布与易处理的潜在分布之间的可逆映射,优化核的潜在空间,从而提高盲超分辨率结果。FlowSR-LP(Tsao等,2024)通过引入条件学习先验,解决了流模型超分辨率中的伪影、逆变爆炸等问题,提升了生成效果。

近年来,基于扩散模型的超分辨率技术在图像质量提升领域取得了显著进展,SRDiff模型(Gao等,2023)利用去噪扩散概率模型和马尔可夫链将高斯噪声转化为超分辨率图像。SR3模型(Xia等,2023)通过迭代精炼高斯噪声生成更真实的超分辨率图像,优于传统的基于GAN的方法。IDM(implicit diffusion models)模型结合隐式神经表示和去噪扩散模型,以实现连续的图像分辨率提升。DiffBIR方法(Lin等,2024b)通过引入预训练模型的先验知识,在较少迭代次数下实现准确估计。StableSR(Wang等,2024b)通过微调预训练的文本到图像扩散模型,利用其先验知识实现真实世界的超分辨率。DiffIR方法(Xia等,2023)同样利用预训练模型的先验知识,以较少的迭代次数达到准确的图像恢复。尽管这些基于扩散的模型在生成高保真度图像和减少伪影方面展现出潜力,但它们通常需要大量训练新样本,并且可能存在收敛速度慢的问题,限制了在某些场景下的应用。

2.1.2 暗光图像增强

随着拍摄设备的普及和个人自媒体的发展,在恶劣场景下拍摄图像和视频成为常见的拍摄需求。黑夜的低光照环境是最常见的恶劣场景之一,与之

伴随的便是低光照环境下图像偏暗、噪声严重以及存在色偏等成像挑战。暗光图像增强旨在使用后处理算法对暗光图像进行光照提亮、细节恢复以及噪声去除，可广泛应用于自动驾驶、视频监控和夜间摄影等领域。此外，这项技术可用于改善数字媒体的呈现质量，提高视觉系统的鲁棒性。

Lore 等人（2017）首先将深度学习方法应用到图像低光照增强问题上，提出深度自编码器 LLNet（low-light net）进行增强和去噪。该方法在正常光照图像上使用随机伽马变换合成带噪声的低光照图像，将正常光照图像和低光照图像组成图像对，从而构建数据集，并利用该数据集监督 LLNet 的训练。

随着深度学习技术在各项计算机视觉领域迅速发展，研究者尝试了多种不同的网络架构。Cai 等人（2018）提出使用深度卷积网络根据 Retinex 理论将图像拆分为高频和低频，分别使用网络进行训练的方法。最终网络可以达到多曝光图像结合增强算法的性能。该工作建立了一个大规模的具有多种曝光程度图像的数据集，并从中训练增强网络。Zhang 等人（2019c）在 Retinex 的基础上，提出不同场景下的反射部分应该结构相同，但是实际上低光照情况下的反射图比明亮场景下的降质严重，因此所提出的 KinD 网络对反射图进行了优化，此外本工作指出由于正常光照图像的重建目标实际上是不明确的，因此在 KinD 网络中加入了任意调整光照图的子模块，给予使用者选择的自由度。Guo 等人（2020）则提出首个在训练过程中不需要任何配对或不配对的正常光照—低光照数据的无监督模型。该模型通过预测一个曲线提亮低光照模型，采用一个轻量的神经网络结构来估计给定图像的动态调整范围。Liu 等人（2021）将基于 Retinex 模型的迭代过程进行展开，并通过神经网络结构搜索寻找合适的网络结构，其轻量级框架以及快速精调的属性使其算法可以迅速部署。

对比于感受野有限的可学习卷积核，得益于注意力机制，Transformer 结构在计算机视觉的各类任务上展现出更为优越的性能。在暗光图像增强领域，Dudhane 等人（2023）提出使用 Transformer 结构对于多幅暗光的输入帧进行融合及增强。该增强框架结合了邻域提取的特征以及全局特征，并针对性设计多尺度多层次层对齐模块完成多图像预处理。Cai 等人（2023）提出一个基于 Transformer 架构的单阶段 Retinex 框架，该框架首先估计照明信息，以照亮低光照图像，然后恢复损坏，以生成增强后的图像，以及一个照明引导变换器，利用照明表示来引导不同光照条件区域的非局部交互建模。

GAN 能有效完成分布间的映射，应用于视觉领域能有效提升图像感知质量。EnlightenGAN（Jiang 等，2021）采用了对抗学习的机制，不使用配对的真值正常光照图像监督学习，而是在不配对的正常光照—低光照训练数据集上施加约束。这种约束的信息从输入图像自身提取。模型采用了全局—局部的判别器设计、自归一化的感知损失函数融合以及注意力机制。Yang 等人（2020）采用了半监督的训练范式，在第 1 阶段使用端到端框架进行训练完成光照恢复，第 2 阶段使用对抗学习范式进行图像质量增强，从而完成主客观质量提升。

作为生成模型的另一范式，Flow 模型通过构建可逆神经网络完成分布间的一一映射。Wang 等人（2022c）提出的 LLFlow 则成功地将 Flow 模型应用于暗光图像增强领域，该工作使用归一化流模型建模成对数据。可逆网络将低光照图像和特征作为条件，学习将正常曝光图像的分布映射到高斯分布。这种方式可以有效建模正常曝光图像的条件分布。而可逆网络的另一推理方向代表增强过程，通过一个更好地描述自然图像流形结构的损失函数进行约束。

基于扩散模型的暗光图像增强技术同样取得了显著进展。研究者们通常将降质过程与扩散过程相关联，通过深度网络完成正常光照高质量图像的生成。Zhou 等人（2023b）提出首个性能超越其他算法的基于扩散模型的低光照算法。其主要针对扩散模型速度较慢以及应用到低光照领域容易导致色偏的问题进行了优化，对于前者，该方法提出在采样的前期阶段使用较小分辨率的图像进行扩散过程以降低算法复杂度；对于后者，该工作设计了一个全局的校正模块对扩散过程的结果进行处理，校正后的结果明显改善了色偏问题。Wang 等人（2023f）则针对原始数据 Raw 域的增强引入了扩散模块，并使用基于物理模型的噪声引导扩散过程，且对扩散模块加入了自适应残差层以提升性能。与扩散模型假设的高斯噪声不同，该模型使用了基于物理信号的噪声模型。Wang 等人（2024h）则提出一组光照无关的四元先验组，使用正常光照的大规模数据集进行训练，借

助预训练扩散模型完成光照恢复和细节重建。

2.1.3 雨雾图像增强

雨雾条件下的图像增强技术旨在去除雨水与雾层带来的图像遮挡、模糊与颜色失真，提升图像在恶劣天气下的质量。该技术可广泛应用于自动驾驶、交通监控和无人机导航等领域。此外，这项技术还可用于提高雨雾天气下安全监控的可靠性、改善环境感知系统的图像质量等。

2017年以前，雨雾条件下的图像增强典型方法是基于模型的策略。此阶段的方法主要由以图像分解（Kang等，2012）、稀疏编码（Luo等，2015）以及基于先验的高斯混合模型（Li等，2016）等代表思路推动。基于深度学习的雨雾图像增强技术起源于2017年，Yang等人（2017）在JORDER（joint rain detection and removal）中构建了一个联合的雨水检测与去除网络，能够处理强降雨、重叠的雨条纹以及雨水积累。该网络通过CNN预测二进制雨水掩码检测雨水位置，采用递归框架逐步去除雨条纹并清除雨水积累。随后，更多采用更先进的网络架构、引入了新的与雨相关的先验知识的方法被提出。然而，这些方法都受到监督学习范式的局限，而在去雨任务中难以获得真实雨图像及其对应的无雨图像作为成对训练数据，需要使用合成的雨图像与原图像作为成对数据训练，因而使得这些方法在处理训练中从未见过的真实雨水条件时，往往表现不佳。

为了捕捉一些无法建模和合成的雨水视觉特性，对抗学习生成的方法被引入去雨任务中以减少生成结果与真实干净图像之间的域差异（domain gap）。Zhang等人（2020a）将条件生成对抗网络（CGAN）应用于单幅图像的雨水去除任务，将雨图作为条件变量输入生成对应的无雨图像。该方法相较传统CNN方法能够捕捉超越信号保真度的视觉特性，呈现出在照明、颜色和对比度分布上更好的结果。Li等人（2019a）针对不同强度的雨图像增强提出两阶段的网络架构：初始基于物理模型的子网络，后接由深度引导的GAN细化子网络，增强了对强降雨图像的恢复能力。受到图像解耦策略的启发，Ye等人（2021）提出一个基于CycleGAN的联合雨水生成和去除框架，通过在更简单的雨水空间中进行图像转换来完成任务。同时，诸如附着在镜头上的雨滴、降雪等更多形态的雨条件下的图像增强也得到了关注。Qian等人（2018a）开发了一种注意力生成对抗网络（attentive GAN），将视觉注意力引入至生成网络和判别网络中，不仅引导判别网络更加关注恢复的雨滴区域的局部一致性，还使生成网络更加关注雨滴区域周围的上下文信息。除此之外，Jin等人（2019）提出一种无监督去雨生成对抗网络，通过引入自监督约束和从未配对的雨水图像与干净图像中提取的内在先验实现了无监督学习。基于GAN的模型通过对抗训练生成更高质量的图像，并降低了模型对合成雨图和真实雨图间域差异的敏感性，提升了去雨效果，但由于其生成能力的限制，在细节保留和图像自然度方面有待提升。

与基于GAN的方法相比，基于流的模型减少了训练期间的不稳定性，因为它们不依赖于鉴别器网络。目前基于流的生成模型还没有完全应用于去雨的工作。Kulikov等人（2023）提出利用规范化流模拟目标内容的分布，作为最大后验（maximum a posteriori，MAP）优化过程的先验。该模型有效地解决了各种类型的图像退化问题，这种方法非常适合去噪等任务，并且可能扩展到去雨问题，因为它具有在未知退化情况下恢复丢失细节的强大能力。

近年来，受到扩散模型的启发，诞生了众多基于扩散模型原理的去雨模型。Özdenizci和Legenstein（2023）提出一种基于去噪扩散隐式模型的新型单图像去雨方法。该方法采用基于SR3的U-Net全卷积网络架构变体，通过在通道维度上连接任意尺寸的雨图像和噪声图像，迭代生成高分辨率的无雨图像。WeatherDiffusion（Liu等，2024a）则引入了一种基于块的扩散恢复方法，通过对重叠块的平滑噪声估计引导去噪过程，从而优化扩散模型的采样过程，实现了与图像尺寸无关的处理能力。残差去噪扩散模型（RDDM）（Liu等，2024b）将传统的单一去噪扩散过程解耦为残差扩散和噪声扩散，令残差扩散更明确低指导图像恢复的逆生成过程。Helminger等人（2021）提出任务嵌入（task-plugin）模块，为模型提供包括雨雾条件下图像增强任务在内的任务特定的先验知识，引导扩散过程保留图像内容，从而减小了扩散过程随机性对图像细节的丢失。

扩散模型的引入使增强模型具有更稳定的训练过程和更强的样本多样性，能够生成更高质量的雨雾条件增强图像，并在捕捉图像细节和保留原始图像结构信息方面具有更为出色的表现，尤其是在处理复杂背景和多种降雨条件时。但扩散模型具有相

对较高的计算成本，同时其带来的随机性的影响也难以完全消除。

2.2 个体分析

本节将从目标检测、目标跟踪、行人再识别、姿态估计和动作识别等方面介绍数据生成在个体分析任务中的典型应用。

2.2.1 目标检测

数据生成方法能够为目标检测任务提供增强数据，以在有限真实数据的基础上提升所训练的检测器的性能。Copy-paste（Dvornik 等，2018；Ghiasi 等，2021）是早期的一种简单且有效的为检测器提供合成数据的方法，通过将已有图像的部分信息复制到其他图像的方法创建新的合成数据，然而这一类方法一方面无法合成新的内容，只能重复利用已有的信息，另一方面通过简单的复制粘贴无法获得真实图像。后续的工作引入 DALL-E（Ge 等，2022）和 Stable Diffusion（Zhao 等，2023）以生成前景物体并粘贴到已有背景中构建合成数据，虽然能够生成新的内容，但是这类方法提供的合成数据在前景的边界处容易出现伪影，仍然难以获取真实图像。DatasetGAN（Zhang 等，2021c）依靠 GAN（Goodfellow 等，2014）生成带有精细标签的图像数据集，是完全依靠生成方法获取合成数据的早期尝试，在人脸和汽车结构等较为简单的数据集上取得了不错的生成效果和精准度。扩散模型（Ho 等，2020）目前已成为图像生成的主流方法，也被用于合成训练数据以辅助目标检测（Chen，2024b；Fang 等，2024；Feng 等，2024；Zhang 等，2023c；Zhu 等，2024b）、图像分类（Azizi 等，2023；He 等，2023a；Li 等，2023e；Sarıyıldız 等，2023；Trabucco 等，2023；Dunlap 等，2023）和语义分割任务（Jahn 等，2021；Gong 等，2023；Li 等，2023d；Peng 等，2023；Wu 等，2023b，c；Xu 等，2023；Yang 等，2023b；Kondapaneni 等，2024）。这些可用于生成带标注图像的方法大致可以分为 3 类。第 1 类工作（Gong 等，2023；Li 等，2023f；Wu 等，2023b；Xu 等，2023；Zhang 等，2023c；Feng 等，2024；Kondapaneni 等，2024）首先生成图像，然后引入预训练的感知模型为合成数据生成伪标签；第 2 类工作（Chen 等，2024b；Cheng 等，2023；Jahn 等，2021；Li 等，2023c；Peng 等，2023；Wang 等，2024e；Yang 等，2023c；Zhang 等，2023b；Zheng 等，2023；Zhou 等，2024a；Zhu 等，2024b）使用伪标签作为生成条件，利用可控图像生成模型获取合成数据，利用以边界框为代表的布局信息作为条件生成图像的方法为代表；第 3 类工作（Ciampi 等，2020；Fabbri 等，2021；Kerim 等，2021）利用游戏引擎进行合成数据生成，通过模拟环境和人物行为，为目标检测任务提供高质量的合成数据。这类方法基于游戏引擎（如 Unity、虚幻引擎等）进行环境建模，生成具有真实感的图像和视频序列，以供训练目标检测模型。

第 1 类工作中，InstaGen（Feng 等，2024）引入开放世界检测模型 Grounding DINO（Liu 等，2024e）为合成图像生成伪标签。Diffusionengine（Zhang 等，2023c）使用预训练的扩散模型和检测适配器以合成数据用于目标检测任务。

第 2 类工作中，LayoutDiffusion（Zheng 等，2023）将布局和类别的嵌入信息进行融合，并建立前景物体感知的交叉注意力机制实现布局可控的图像生成。LayoutDiffuse（Cheng 等，2023）在梯度扩散模型中引入布局注意力模块实现布局可控的图像生成。Li 等人（2023c）在预训练的梯度扩散模型中插入门控自注意力层以适应包括布局可控图像生成在内的多样化可控生成任务。ReCo（Yang 等，2023d）和 GeoDiffusion（Chen 等，2024b）将前景物体位置信息进行编码后加入到文本输入中，然后重新训练文本编码器和扩散模型以实现布局可控的图像生成。ControlNet（Zhang 等，2023b）利用梯度扩散模型中 U-Net 的编码部分网络作为 backbone 编码条件信息实现可控图像生成。Zhou 等人（2024a）将多实例的生成分解为单实例生成的子任务并加入额外的注意力模块实现布局可控的图像生成。InstanceDiffusion（Wang 等，2024e）综合利用边界框、掩码和简单涂鸦等条件实现精确的实例级别的控制。ODGEN（Zhu 等，2024b）首先提出利用合成数据作为引导实现布局可控的图像生成方法，利用预训练的扩散模型生成前景物体的图像并依据布局信息构建条件图像列表实现布局控制，并引入可训练的模块拟合复杂的文本输入条件，显著提升了布局可控的复杂场景生成效果，并为目标检测任务带来有效增益。

第 3 类工作中，ViPeD（virtual pedestrian-detection dataset）（Ciampi 等，2020）是一个适用于行人检测任务的合成数据集，使用《GTA 5》游戏引擎生成，共包含来自 512 个不同城市的约 500 000 幅图像，每个行人边界框的坐标都是自动生成的，显著

减小了数据标注的成本。实验表明，使用合成数据训练的检测器在未见过的场景中具有更好的泛化性能，相比仅使用真实世界数据集训练的模型，平均精度均值（mean average precision，mAP）提升了0.085和0.163。MOTSynth（Fabbri等，2021）是另一个大规模合成数据集，包含1 382 K帧不同天气条件下的城市场景，用于行人检测、跟踪和分割。此数据集包含边界框、分割和深度蒙版。每个视频为25 FPS（frames per second），平均每帧包含29.5人，最多125人，总计超过40 M个边界框，超过1.3 M个密集注释帧。该数据集使用《侠盗猎车手Ⅴ》游戏引擎生成，通过定义一组行人群体必须遵循的轨迹来手动预先规划行人流量，并依靠防撞算法获得每个自然行人的行为。实验表明，基于MOTSynth训练的模型性能优于基于COCO（common objects in context）训练的模型。其中YOLOv3（you only look once）、CenterNet、Faster R-CNN和Mask R-CNN的平均精度分别提高了2.49%、3.48%、2.30%和1.87%。

2.2.2　目标跟踪

数据收集通常需要大量的手动工作。随着不断需要更多数据来训练不断增长的模型，标记此类数据集的成本增大，变得令人望而却步。这种负担可能会限制可用数据的质量或数量并阻碍进展。解决上述问题的一个可能办法是使用生成数据。数据生成方法能够为目标跟踪任务提供增强数据，以在有限真实数据的基础上提升目标跟踪模型的性能。在早期工作中，SIT（Liu等，2025）使用Pix2Pix（Isola等，2017）方法根据RGB图像生成带有跟踪标签的红外图像和视频，提升了在红外视频进行目标跟踪的性能。LRPD（Xu等，2017a）使用生成红外可见光图像的网络预训练权重，然后将这些权重应用到RGB数据的网络中，目的是改进对行人的目标跟踪。Virtual KITTI（Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago）（Gaidon等，2016）基于游戏引擎生成虚拟世界场景，使用现代计算机图形技术和新颖的虚拟克隆方法生成完全标记的、动态的且逼真的虚拟世界，并且其能自动进行准确的标注，可用于目标跟踪等其他高级语义任务。NOVA（Kerim等，2021）基于Unity引擎，生成带标注的模拟世界与人类，支持多种特征和动作混合，并提供多种Ground Truth（如光流、深度图等）。研究者通过NOVA构建了两个用于人员跟踪的合成数据集。第1个数据集包括108个序列，每个序列的难度级别不同，部分难度很高，例如在拥挤的场景或夜间跟踪。第2个数据集包含97个正常天气条件下的序列，难度正常。实验表明，在合成数据和真实数据的混合上微调基线可以提高性能，相比只用真实数据微调两个跟踪器的平均重叠分数分别提高了4.89%和1.76%。同时，与仅用真实数据训练相比，仅使用生成数据可获得更高性能，两个跟踪器分别提升34.28%和2.15%。验证了NOVA生成的合成数据在计算机视觉任务中的有效性。

随着GAN网络以及扩散模型的发展，促使文本转视频任务不断进步，其旨在生成视频以应用于各种视觉任务（如目标跟踪）。早期的T2V（text-to-video）模型（例如Sync-DRAW（Mittal等，2017）以及图像GAN到视频的扩展）主要解决更简单的问题，例如移动数字或特定动作。Sync-DRAW将VAE与递归注意机制结合起来，在时间上形成一系列渐进式的视频帧；DVD-GAN（Clark等，2019）将大型图像生成模型BigGAN（Brock等，2018）扩展到视频领域，从而能够生成高分辨率和具备时间一致性的视频。最近的进展（例如GODIVA（Wu等，2021）和NÜWA（Wu等，2022））主要采用2D VQ-VAE和稀疏注意力以及多任务学习方案生成更真实的场景。之后，越来越多的工作利用了图像先验，如MoCoGAN-HD（Tulyakov等，2018）和CogVideo（Hong等，2022b），通过利用预先训练的图像模型简化视频生成。最近的工作（Blattmann等，2023b；Luo等，2023）利用视频扩散模型和大规模文本视频对获得了显著的性能，但仍然没有细粒度的轨迹控制，这些方法通常缺乏生成复杂转换和多样化事件序列的能力。然而，LVD（LLM-grounded video diffusion）（Lian等，2024）等工作采用大型语言模型创建视频传播的动态场景布局，专注于文本驱动的布局生成。VideoDirectorGPT（Lin等，2024c）和DriveDreamer（Wang等，2023e）进一步推进了多场景视频生成，展示了该领域的进步。虽然取得了这些进步，但在连续多对象跟踪（MOT）视频数据的生成方面仍然存在巨大差距。TrackDiffusion（Li等，2024a）首次提出一种专为多对象跟踪（multiple object tracking，MOT）复杂需求设计的视频生成方法，填补了领域空白，这一方法不仅能够捕获不同的动作和事件，还确保运动的一致性和实例的连贯性，能够从轨迹生成视频，并对复杂对象动态

进行精确控制，这对于高级MOT算法的训练至关重要。实验结果强调了TrackDiffusion在增强MOT系统训练方面的潜力，从而标志着合成视频数据生成为目标跟踪服务的领域向前迈出了重要一步。

2.2.3　行人再识别

数据生成方法能有效提升行人再识别数据集规模，增强行人再识别模型的检索能力和泛化性能。本节分别从传统数据生成方法、基于虚拟引擎的数据生成方法、基于深度生成模型的数据生成方法3方面介绍典型的数据生成技术在行人再识别领域的应用。

传统数据生成方法应用数字图像处理技术对行人图像进行几何变换、图像粘贴、随机擦除和颜色抖动等操作生成新行人图像。McLaughlin等人(2015)最先应用平移、旋转和裁剪等几何变换方法生成行人图像，提升行人再识别模型的泛化能力，并应用图像粘贴方法实现行人图像背景变换，提升行人再识别模型对图像背景的鲁棒性。Huang等人(2018b)和Zhong等人(2020)随机擦除行人图像中的局部区域，避免行人再识别模型依赖局部信息进行检索。Gong等人(2021)将行人图像或图像中的随机区域转换为灰度图像，缓解因颜色信息引起的过拟合。Lin等人(2021b)通过对比度调整方法和流形学习方法，对行人图像进行光照风格迁移，增强行人再识别模型对光照的鲁棒性。传统数据生成方法能够以较低的开销改变行人外观并生成行人图像。然而，该生成方法仅在现有行人图像的基础上添加扰动，对行人再识别任务的增益有限。

基于虚拟引擎的数据生成方法应用虚拟引擎创建人物模型并渲染行人图像。早期工作(Barbosa等，2018；Bąk等，2018；Sun和Zheng，2019)将行人模型独立导入虚拟引擎中渲染行人图像生成虚拟行人数据集。此后，为保证行人模型与背景结合更自然，TAGPerson方法(Chen等，2022)结合背景风格渲染行人图像。以上工作主要采用公开获取或手工制作的三维人物模型，导致虚拟人物数量很难扩展，且这些工作生成的行人图像缺乏人与人之间的遮挡关系和摄像机之间的转移规律，与真实场景采集的行人图像有较大差异。为解决此问题，基于复杂场景模拟的数据生成方法(Wang等，2020；Zhang等，2021b；Li等，2021a；Wang等，2022d)通过脚本大规模生成人物模型，为人物模型绑定骨骼和动画使其在虚拟场景中随机移动，并架设多个虚拟摄像头采集行人图像。基于虚拟引擎的数据生成方法几乎无需标注、没有隐私问题且能弥补现实世界的不足，可以创造出全新的、独立的行人图像。然而，基于该方法生成的行人图像与真实行人图像存在较大域差异。

基于深度生成模型的数据生成方法应用GAN模型、扩散模型等深度生成模型生成新行人图像。基于GAN模型的生成方法主要分为风格迁移方法和随机生成方法。风格迁移方法生成与原图像行人身份一致、图像风格不同的行人图像以提升行人再识别模型的鲁棒性。Zhong等人(2018)、Liang等人(2018)针对图像相机参数和光照条件进行风格迁移，提升行人再识别模型对相机风格的鲁棒性。Huang等人(2019)、Pang等人(2022)针对图像背景进行风格迁移，增强行人再识别模型对图像背景的鲁棒性。上述方法直接在几个图像集合之间进行风格迁移，Qian等人(2018b)、Borgia等人(2019)、Zhang等人(2019a)、Zheng等人(2019)则提取行人姿态作为先验信息，针对行人姿态进行风格迁移，增强行人再识别模型对行人姿态的鲁棒性。上述方法生成行人身份确定的行人图像，无需重新对图像进行标注。随机生成方法(Zheng等，2017；Ainam等，2019；Salem Hussin和Yildirim，2021；Eom和Ham，2019)则随机生成身份未知的行人并应用离群值标签平滑正则算法(label smoothing regularization for outliers，LSRO)等分类方法为生成图像进行标注。此后，扩散模型被证明生成质量显著高于GAN模型并成为图像生成领域的主流方法。Niu等人(2024)设计提示词控制扩散模型生成行人图像，并证明使用生成图像预训练能有效提升行人再识别模型的表现。仅使用提示词难以精确控制生成行人图像，Kim等人(2024)引入行人姿态信息控制扩散模型生成姿态丰富的行人图像，增强行人再识别模型对行人姿态的鲁棒性。得益于深度模型强大的特征提取能力和生成能力，基于深度生成模型的数据生成方法能很好地保持行人身份信息并生成高质量行人图像。

在当前行人图像生成领域，基于虚拟引擎的数据生成方法生成图像真实性不足，而基于深度生成模型的生成方法可控性仍存在一定提升空间。目前对三维重建技术和扩散模型的研究方兴未艾，有望进一步提升生成行人图像的可控性和真实性，助力

行人再识别模型改进与性能提升。

2.2.4 姿态估计

生成模型在基准数据生成和模型评估方面的实用性之前已在对象分类和检测任务中得到认可，由真实数据（Human3.6M（Ionescu等，2014）和3DPW（von Marcard等，2018））组成的基准无法提供这样的控制——在不同属性的情况下捕捉足够多的人以相同的精确姿态是不可行的。此外，此类真实数据很难捕获，并且感兴趣的属性会根据任务的应用而变化。计算机图形生成的综合数据（Varol等，2017；Cai等，2024；Patel等，2021；Bazavan等，2022；Black等，2023）提供对姿势、相机位置和服装等的高度控制。然而，为了实现真实感和多样性，它们需要大量高质量的3D资源，必须设计服装并对其进行物理模拟（Black等，2023），并且必须对位置进行建模或捕获。此外，有效利用计算机图形学的方法需要专业知识，而这可能是人体姿态估计从业者所欠缺的。总体而言，当前的计算机图形学方法生成的数据无法对人体姿态估计模型进行全面的评估。

相比之下，文本到图像模型的文本输入是一个更灵活且易于使用的界面，用于自定义基准生成和控制服装、位置等属性。可控人体图像生成方法可以分为基于二维姿态控制的方法（Brooks和Efros，2022；Ju等，2023；Ostrek等，2024；Liu等，2024d；Mou等，2024）和基于三维姿态控制的方法（Grigorev等，2021；Bergman等，2022；Hong等，2022a；Dong等，2023；Cao等，2024）。对于二维姿态为条件的方法，如ControlNet（Zhang等，2023b）、T2I-Adapter（Mou等，2024）、HumanSD（Ju等，2023）和HyperHuman（Liu，等，2024d），通过二维人体关键点的控制扩展了SD（stable diffusion）（Rombach等，2022）方法。它们都表现出良好的文本控制能力，但无法精确控制3D人体姿态。虽然ControlNet能够在免训练过程中组合多个控制信号（例如深度和2D姿态），但生成的模型仅实现有限的3D可控性；T2I作为独立插件与预训练模型结合，不改变原有模型；Adapter可以利用额外的控制信息，如颜色、结构，进行更精细化的生成控制；HumanSD（Ju等，2023）引入了一种由骨骼引导的扩散模型，以提高姿态控制的准确性，然而其生成的内容质量较低；HyperHuman（Liu等，2024d）通过学习人体结构的内在规律，将人体表示为一个隐式函数，从而避免了传统参数化模型在细节表达上的局限性。对于三维姿态控制方法，可以大致将其分为基于GAN的方法（Bergman等，2022；Noguchi等，2022；Dong等，2023；Hong等，2022c；Yang等，2023a）和基于Stable Diffusion的方法。基于GAN的方法学习从二维单视图图像集合中生成三维人像。通过诸如反蒙皮等技术，实现了出色的姿态控制。如AG3D（Dong等，2023）通过采用整体3D生成器并集成高效灵活的关节模块捕捉身体和宽松衣服的形状和变形，为了提高真实性，使用多个鉴别器训练，同时还以预测的2D法线图的形式整合几何线索。然而，它们生成的人物图像没有背景，并且不提供文本控制功能。基于Stable Diffusion的方法（Poole等，2022；Kolotouros等，2023；Cao等，2024；Liao等，2024）或CLIP（Radford等，2021；Hong等，2022a；Youwang等，2022）的方法提供文本控制，如DreamAvatar（Cao，等，2024）利用可训练的NeRF预测3D点的密度和颜色特征，并使用预训练的文本到图像扩散模型来提供2D自我监督。其提升了可控的生成效果，并可为人体姿态估计任务带来有效增益。

2.2.5 动作识别

近年来，人们创建了许多合成视频数据集用于训练人体动作识别深度学习模型。当需要快速生成大量带注释的数据时，合成数据已被证明是一种有用的解决方案。模拟人体运动可以追溯到20世纪80年代。Badler等人（1993）提供了早期方法的广泛概述，在之前关注合成人类数据的工作中，很少涉及动作识别。最近，合成2D人体姿态序列（Lyu和Nevatia，2007）和合成点轨迹（Rahmani和Mian，2015；Rahmani等，2018；Zhang等，2018）已用于视图不变的动作识别任务中。然而，基于RGB的动作识别综合训练相对较新，De Souza等人（2017）的工作是最早的尝试之一。De Souza等人（2017）手动定义35个动作类别，并在多任务设置中联合估计真实类别和合成类别。然而，它们的类别不容易扩展，并且不一定与目标类集相关。与De Souza等人（2017）的工作不同，Varol等人（2021）自动从真实数据中提取运动序列，使该方法能够灵活地适应新类别。Puig等人（2018）生成了VirtualHome数据集，这是一个使用众包以编程方式定义合成活动的模拟环境。与Varol等人（2021）的工作不同，Puig等人（2018）的重点不是对真实数据的概括。Liu等人（2019a）生成合

成训练图像，以在看不见的视点上获得更好的动作识别性能。Liu 等人（2019a）的工作是对 Rahmani 和 Mian（2016）工作的扩展，其采用 RGB-D 数据作为输入，而不仅仅依赖于深度信息。这两个工作都在其合成数据上制定了基于帧的姿态分类问题，然后将其用作动作识别的特征。但这些特征不一定能够区分目标动作类别。与这个方向不同的是，Varol 等人（2021）明确地为合成视频分配一个动作标签，并直接在动作分类上定义监督，取得了不错的效果。像 Everybody Dance Now（Chan 等，2019）和 LWG（liquid warping GAN）（Liu 等，2019b）这样的模型通过将身体姿势从一个人执行活动的视频转移到另一个视频或图像中的新人来生成视频，可以根据输入的视觉源图像生成具有不同外观的动作视频。然而，这些模型依赖于单眼视频或图像的身体姿势估计，本质上不如运动捕捉准确。Panev 等人（2024）将 MoSh（Loper 等，2014）（一种用于从运动捕捉标记集生成 3D 人类 SMPL 网格的模型）集成到 LWG 架构中。因此，扩展了其从动作捕捉数据生成视频的功能。

近年来，自然语言处理（natural language processing，NLP）模型已经展示了根据文本提示生成不同数据类型的能力，TEMOS（Petrovich 等，2022）、MotionCLIP（Tevet 等，2022a）、MDM（motion diffusion model）（Tevet 等，2022b）、T2M-GPT（Zhang 等，2023a）和 MotionGPT（Jiang 等，2023a）等模型可以通过活动描述（文本到运动）生成逼真的人体运动。然而，很难将此类模型合并到动作识别数据生成中，因为它们无法对生成的输出提供细粒度的控制，也无法跨不同的合成集建立精确的运动匹配。

然而，一些工作通过将合成数据用于数据增强取得了不错的效果，动作识别中的标准数据增强技术包括水平翻转和裁剪，其中通过在每一帧选择一个框来创建新视频，然后调整生成的视频大小以使其与原始视频具有相同的大小。虽然此策略有所帮助，但生成的视频并没有为训练集增加太多的多样性。ActorCut（Zou 等，2023）和 VideoMix（Yun 等，2020）通过将一个视频的前景裁剪并粘贴到另一个视频上来增加新视频样本的多样性，这种组合两个数据样本的通用技术已被证明非常有效。然而，生成的视频不太真实，并且无论其质量如何都用于训练，会引入噪声数据。Zhang 等人（2020c）更进一步使用 GAN 合成新样本，并使用“自定进度选择”进行训练，从简单样本开始，逐步选择更难的样本。相反，Gowda 等人（2022）建议通过分割、修复并将一个视频的前景混合到另一个视频的背景上来创建真实的数据样本，比较重要的是，其学会丢弃预计对分类无用的新视频样本，从而总体上生成更准确的数据，为动作识别任务带来了有效增益。

2.3 生物特征识别

生物特征识别通过测量和分析指纹、掌纹、静脉、虹膜和人脸等独特的物理或行为特征，可以准确识别或验证个人身份。表 1 列出并对比了指纹识别、掌纹识别、静脉识别、虹膜识别和人脸识别等 5 种常用的视觉生物特征识别技术，以及对应的数据集规模。由于大量收集生物特征识别相关数据不仅成本高昂，而且引发人们对用户隐私和潜在法律侵犯的担忧，因此很多公开数据集的使用遵守更严格的限制，其中一些数据集变得不可访问。如表 1 所示，目前掌纹、掌静脉和虹膜的公开数据集仅有数百到几千的规模，指纹数据集约几千到一万左右，这使得相关识别模型的训练变得较为艰难。在人脸识别领域，尽管存在拥有数百万幅图像的公开数据集，但大多都缺少较为丰富的类内变化。这些因素都极大限制了生物特征识别领域的发展。

生物特征数据生成技术随着深度学习的不断发展逐渐趋向成熟，从传统的数学建模生成方法到深度学习模型，如 VAE、GAN、扩散模型等，从粗糙的低质量图像生成到精细可控的高质量图像生成，生物特征数据生成方法取得了飞速进步。在深度学习还未全面占领生成领域之前，人们依靠各种数学工具实现图像的生成和编辑操作。不同的生物特征都有其手工设计的纹理模型，例如，Cappelli 等人（2000，2002，2004）使用由多个数学模型组成的 SFinGe 对指纹纹理进行建模，Zhao 等人（2022）使用多条 Bezier 曲线对掌纹进行建模，Hillerström 等人（2014）、Yang 等人（2020）使用生长算法从多个节点逐步生成静脉纹理，而人脸因为存在多个素描数据库因此无需对其纹理进行建模，可直接从素描图生成真实人脸。得到纹理图案之后，生成任务就执行从纹理图案到真实图像的风格转移任务，人们可以通过各种渲染方法来实现这种转换。

随着深度学习的发展，VAE、GAN 和扩散模型等一系列生成模型相继提出，人们不再需要建立复

杂且缺少泛化能力的数学模型，生物特征的纹理图案和其他细节特征完全交给模型进行学习，可以从先验分布直接生成相关图像。但这受限于相关领域的公开可用的数据集规模，因为大部分深度生成模型需要基于庞大的数据集才足以学习到真实数据的分布情况，而这与生成生物特征图像的初衷相矛盾，即缺乏大量的真实数据集。同时，以GAN为代表的生成模型往往容易遇到模式崩塌、难以训练的问题，因此为了提升生成效果，人们通常会在模型架构、损失函数和训练策略等多个方面对其进行改进。之后，为了进一步改进模型在小数据集上很难学习到真实纹理分布的问题，人们将上述两种方法相结合，先通过对纹理建模生成逼真的纹理图案，再使用生成模型将其转移到真实图像领域，从而将生成任务转化成了一个相对简单的图到图的风格转移问题。

表1　不同生物特征识别技术及数据集规模

Table 1　Different biometric identification technologies and data set sizes

模态	采集特性	稳定性	准确性	应用范围	数据集规模
指纹识别	接触式采集、操作简单方便	易受皮肤情况、外界环境等因素影响	高	广泛	几千到一万
掌纹识别	非接触式采集，操作简单方便	易受皮肤情况、外界环境等因素影响	极高	较少	几百到几千
掌静脉识别	非接触式采集、对设备要求高	内部特征，不易受外界影响，但手背静脉尚未证实终生不变性	极高	较少	几百到几千
虹膜识别	非接触式采集、对设备要求高	终生不变	极高	较广泛	几百到几千
人脸识别	非接触式采集，操作简单方便	易受年龄、表情、化妆等因素影响	高	广泛	几千到上百万

2.3.1　图像质量评估

图像质量评估是通过一系列指标和方法衡量图像的清晰度、准确性和视觉感知的过程。对于生成的生物特征图像，首先需要保证其与真实图像的相似性，即具有较高的生成质量；其次，生成的图像需要包含真实图像的特征信息，这些特征是识别算法进行身份识别的重要依据；最后，对于不同身份生成的生物特征图像，需要保证其唯一性，即具有较高的多样性。基于此，可将现有评估方法分为3类，分别是生成图像质量评估、生成细节评估和生成多样性评估，如表2所示。

1）生成图像质量评估。判断生成图像真实性的最简单方法便是人体视觉实验，即将真实图像与生成图像混合到一起，邀请领域内的专家来判断这些图像的真假，通过判断的准确率评估合成图像的真实性。除了人眼判断外，还可以训练分类器对真假图像进行分类，分类的准确率也可作为评估指标。显然，这些方法过于简单粗暴，无法考虑到合成图像与真实图像之间的细微差距，因此一些研究者提出更精细的评估方法，采用相似性指标度量真实图像与合成图像的差异，如FID、余弦相似度、MS-SSIM（multi scale structural similarity index measure）、KS（Kolmogorv-Smirnov）检验等。其中，FID是从计算机视觉特征的统计方面衡量两组图像的相似度，是计算真实图像和生成图像的特征向量之间距离的一种度量。余弦相似度是将图像表示成一个向量，通过计算向量之间的余弦距离来表示两幅图像的相似度。MS-SSIM是一种多尺度的结构相似性指标，该方法可以考虑图像在多个尺度上的像素多样性，并从亮度、对比度、结构这3个方面对整幅图像进行比较，可以更好地模拟人类对图像的感知。KS检验通过计算从真实图像中提取的每个度量的分布和合成图像之间的差异衡量其相似性。这4个指标分别从视觉特征、向量空间、结构和分布等方面描述两个图像间的相似性。此外，为了更方便高效地衡量指纹图像质量，美国国土安全部、科技局等多个部门联合研发了NFIQ 2.0（NIST fingerprint image quality 2.0）进行指纹质量评分，其分数值在0～100之间，分数越高代表图像质量越高。

2）生成细节评估。生物特征中的细节特征被认为是识别中最具鉴别力的特征，因此需要从合成图像中提取细节特征进行评估。Cao和Jain（2018）通过计算成对的细节距离来构建2DMH（2D morphological histogram），使用Ave-2DMH间的欧氏距离评估真实图像与合成图像在细节配置方面的相似性。此外，还可以提取出图像中的细节数量，判断是否接

表 2　常用的生成图像评价指标

Table 2　Commonly used synthetic image evaluation indicators

评估方向	评估指标	评估方法
生成质量评估	视觉评估	邀请领域内专家来判断图像的真假
	分类器评估	训练分类器来判断图像的真假
	KS 检验	度量两幅图像的分布差异
	FID	计算两幅图像的特征向量间的距离
	NFIQ	计算图像的质量得分
	余弦相似度	计算两幅图像向量间的余弦距离
	GCF	基于全局图像对比度的通用图像质量评估度量
	GLCM	使用来自一对像素的灰度级二维直方图的二阶统计量来测量图像纹理
	HSNR/Wang17	静脉图像质量评估度量
生成细节评估	细节 2DMH	计算细节间的距离来构建细节 2DMH
	细节度量	对比两幅图像中的细节数量
生成多样性评估	MS-SSIM	度量两幅图像的结构相似性
	Verifinger SDK 6.3	计算成对的比较分数
	大规模搜索实验	在数据库中通过计算相似度搜索出相应图像
	假匹配分布	假匹配分数越低，代表图像的唯一性越高

注：GCF：global contrast factor；GLCM：gray-level co-occurrence matrix。

近真实图像。

3）生成多样性评估。评估合成图像的多样性，即评估图像间的相似度，相似度越低，则多样性越好。生成质量评估中通过相似度指标评估真实图像与合成图像的差异，在这里还可以使用这些相似度指标评估合成图像间的差异，越高的差异代表越好的多样性，其中 MS-SSIM 指标多用于这里。此外，除了直接计算相似度外，还有一些间接方法，如搜索精度、假匹配分布等。搜索实验利用搜索精度评估多样性，给定一幅图像，搜索算法通过计算相似度来找出数据库中与其最相似的图像，若相似度分数大于某一阈值，则搜索成功，否则失败。其本质还是使用相似度指标，搜索精度越低，代表相似性越低，即多样性越好。假匹配分布可通过专业的匹配器，如指纹匹配器 Verifinger，计算匹配分数得到，通过对比真实数据库与合成数据库的假匹配分布差异评估图像的唯一性，即多样性。

2.3.2　指纹生成方法

目前，大部分指纹识别模型和分类模型都是通过特征提取算法提取指纹的显著特征进行识别和分类任务的，因此在进行指纹合成时，不仅要考虑真实性，还需要生成符合真实指纹分布的特征。此外，指纹数据集根据采集方式的不同，可分为接触式指纹和非接触式指纹。相比非接触式采集，接触式采集方法由于其采集方便、成本较低等优势而得到广泛使用，因此大部分的合成指纹也都为接触式指纹，在本文调研中，只有一篇文献研究非接触式指纹生成。接触式指纹由于手指与传感器直接接触，易受各种噪音因素影响而发生扭曲变形，因此在生成真实的指纹特征后，还需要考虑到指纹的扭曲变形和环境噪音等因素。表 3 中汇总了目前关于指纹生成的方法，并在下文分别展开介绍。

1）指纹纹理建模方法。传统的指纹生成方法一般分为两步：主指纹生成和指纹压痕生成。第 1 步是生成指纹的全局特征（指纹区域、指纹类型等）和局部特征（核心点、三角点等），得到无噪声的指纹图案，即主指纹；第 2 步是模拟手指在不同压力、接触面积等各种噪声条件下的指纹压痕图像。

（1）主指纹生成。Cappelli 等人（2000）最早进行指纹生成研究，并开发了指纹生成软件 Synthetic Fingerprint Generation（SFinGe），基于其成熟的算法体系和简单易行的生成过程，现已得到广泛使用，目

表3　指纹生成方法汇总

Table 3　Summary of fingerprint generation methods

论文	图像大小/像素	方向图模型	脊线图案模型	细节模型	渲染方法	类内变化	分析
Cappelli 等人(2000)	NaN	零极点模型	Gabor滤波器	Gabor滤波器	随机添加均匀噪声	是	开创性工作
Cappelli 等人(2002)	NaN	零极点模型的变体	Gabor滤波器	Gabor滤波器	随机添加均匀噪声	是	改进方向图模型,引入更多自由度应对方向场变化
Cappelli 等人(2004)	NaN	零极点模型的变体	Gabor滤波器	Gabor滤波器	随机添加非均匀噪声	是	使用Perlin噪声函数来代替之前的均匀噪声
Zhao 等人(2012)	NaN	零极点模型和余弦外围模型	AM-FM	空间分布模型	随机添加均匀噪声	是	可以控制合成指纹中特征细节的生成
Johnson 等人(2013)	NaN	零极点模型的变体	Gabor滤波器	空间分布模型	添加真实指纹特征	是	提出的渲染方法可以保留真实指纹数据库的统计特性
Priesnitz 等人(2022)	NaN	零极点模型的变体	Gabor滤波器	Gabor滤波器	添加真实指纹特征	是	非接触式指纹生成的首次尝试
Minaee 和 Abdolrashidi (2018b)	64×64	DCGAN	–	–	–	否	损失函数加正则项以保证生成线条的连贯性
Striuk 和 Kondratenko (2021)	NaN	DCGAN	–	–	–	否	改进了模型架构
Zhong 等人(2021)	128×128	DCGAN	–	–	–	否	用典型CNN改进模型架构
Bontrager 等人(2018)	128×128	WGAN	–	–	–	否	使用WGAN模型稳定训练
Cao 和 Jain(2018)	512×512	IWGAN	–	–	–	否	用训练好的CAE初始化生成器提高生成质量
Mistry 等人(2020)	512×512	IWGAN	–	–	–	否	在Cao和Jain(2018)方法的基础上引入身份损失,提高类间差异
Fahim 和 Jung(2020)	256×256	GAN	–	–	–	否	使用平均残差连接和loss-doping稳定训练
Bahmani 等人(2021)	512×512	多分辨率GAN	–	–	–	否	通过渐进增长训练来稳定训练过程,提高生成质量
Riazi 等人(2020)	256×256	WGAN	–	SR模型	–	否	通过WGAN生成低质量指纹图像,再使用超分辨率模型提升图像质量
Sams 等人(2022)	256×256	StyleGAN	–	–	CycleGAN	否	可以生成包含很多真实细节特征的指纹图像
Kim 等人(2019)	256×256	多阶段GAN	–	–	–	是	在使用深度学习方法中第1次对类内多样性进行研究
Engelsma 等人(2023)	512×512	多阶段GAN	–	–	–	是	通过多个GAN分别学习指纹的不同特征
Wyzykowski 等人 (2021)	256×256	Impoved SFinGe	–	–	CycleGAN	是	传统方法和深度学习方法混合使用
Attia 等人(2019)	256×256	VAE	–	–	–	否	使用VAE学习指纹的复杂分布
杨光锴 (2023)	128×128	Diffusion model	–	–	–	否	首次使用扩散模型进行指纹生成

注:使用GAN等生成模型是端到端生成,没有使用额外算法进行背线建模、细节建模和渲染操作。“–”表示无相应数据。

前有很多指纹生成算法都是基于SFinGe生成主指纹后再做改进。Cappelli等人(2000)介绍了生成主指纹的一般步骤:指纹区域生成、方向场生成、密度图生成和脊线图案生成。指纹区域即指纹图像的尺寸和轮廓,他们提出一种基于4个椭圆弧和一个矩形并由5个参数控制的简单模型,可以模拟指纹形状的变化。接着基于指纹类型和奇异点(核心点、三角点等),通过零极点模型(Sherlock和Monro,1993)生成不同指纹类型的方向图,以控制指纹脊线的总体流向。由于弓型指纹不包含奇异点,无法使用该模型生成方向图,因此Cappelli等人(2020)使用正弦函数,通过调整其频率和幅度以控制弓型的曲率和长宽比来生成弓型指纹的方向图;之后,他们通过观察大量真实指纹图像,总结出其脊线密度分布来生成密度图,以控制指纹脊线的稀疏程度;基于方向图和密度图,通过迭代使用Gabor 滤波器(Fogel和Sagi,1989)生成主指纹,同时不同类型的细节特征会随机出现。

但该方法也存在一些问题。首先真实指纹的方向场并不完全由奇异点类型和位置确定,零极点模型的建模效果并不符合真实的方向场。针对这一点,Cappelli 等人(2002)采用一种零极点模型的变体(Vizcaya和Gerhardt,1996),该模型引入了更多的自由度应对方向场的变化,实验结果显示,在拟合真实指纹方向场方面,该模型优于零极点模型。Zhao等人(2012)对此也做出了一些改进,他们将方向场分解为奇异分量和残差分量,这些分量分别由零极点模型和余弦外围模型(Wang和Hu,2011)近似,以更好地模拟真实指纹的方向场。

其次,由于合成指纹图像中的细节(脊分叉和脊末端)是在Gabor滤波器的迭代过程中随机形成的,因此该方法无法控制细节的数量和位置,且其细节分布不一定遵循真实指纹的分布。因此,Zhao等人(2012)提出一种从采样特征重建主指纹的方法,旨在解决合成指纹时特征细节无法控制的问题。首先为不同指纹类型建立统计特征模型,基于统计模型采集指纹特征,如奇异点、方向场和细节等,其中使用空间分布模型(Chen和Jain,2009)对细节进行采样,最后使用AM-FM模型(Feng和Jain,2011)来从采样特征中重建主指纹,从而对合成指纹中的特征提供了更多的控制,且生成的细节更遵循真实分布。Zhao 等人(2012)比较了 COTS(commercial-off-the-shelf)匹配器提取的细节与预先指定的细节,从而评估对细节的保留程度,结果显示许多真实的细节被保留在合成指纹中,但依然有不少缺失和伪造的细节出现,因此该方法不能做到绝对的控制。

通过上述方法后,可以通过设置不同的参数来生成相应的指纹纹理图案,接下来需要一些渲染方法为其增加真实性。

(2)指纹压痕生成。在采集真实指纹时,由于各种噪声因素会导致指纹的类内变化,因此需要模拟这些噪声因素。在采集时常见的噪声因素有平移和旋转、与传感器的接触面积、皮肤干湿程度、皮肤的非线性变形以及其他噪声等。根据这些因素需要对上一步生成的主指纹进行改变脊线粗细、扭曲变形和扰动等渲染操作。

Cappelli等人(2002)通过对主指纹使用形态算子(Gonzales和Wintz,1987)来模拟手指在不同情况下脊线的粗细变化,如在手指潮湿或高压情况下,脊线会更厚一点。但该方法生成的脊线厚度是统一变化的,并不能动态改变局部厚度,这一点在下一节基于深度学习的方法中得到改善。此外,Cappelli等人(2001)提出一种皮肤扭曲模型用于映射指纹细节点来改进指纹匹配,在这里可通过调整映射来模拟指纹压痕的扭曲变形。最后,通过使用随机选择的噪声参数渲染主指纹,以合成逼真的指纹压痕图像。但是其添加的是均匀分布的噪声,而真实指纹的噪声是非均匀的,因此Cappelli等人(2004)引入Perlin噪声函数(Perlin,1985)产生相干噪声,使得对主指纹的渲染更接近真实图像。

事实上,随机添加噪声的方法并不能保证产生统计上代表真实指纹纹理的指纹纹理,因此Johnson等人(2013)提出一种添加真实指纹特征来渲染主指纹的方法。通过从真实指纹图像中提取特征并建模,然后映射到合成指纹图像上,可以将主指纹渲染成具有相似特征的指纹图像,该方法可以保留真实指纹数据库的统计特性。作者通过KS检验计算不同特征分布在真实指纹和合成指纹之间的差异,结果表明合成指纹的特征分布与真实指纹非常相似。

Priesnit等人(2022) 第1次提出非接触式指纹的生成方法,与前面的生成方法不同,非接触式指纹合成旨在生成真实的无变形扭曲的立体指尖图像。该方法基于SFinGe生成的接触式主指纹,首先通过变形函数将原本平面的主指纹图案变形为立体的非

接触式指纹图案，由于非接触式设备的景深通常只有几毫米，这会导致手指边缘出现失焦模糊，因此通过对手指边界区域进行模糊和扭曲来实现该效果。之后依次添加身份特征（如肤色）和环境特征（如亮度、阴影等）来生成真实的非接触式指纹。该方法可以生成较为真实的脊线纹理，但从视觉上来看，与真实指纹仍有很大差距。

2）指纹深度生成方法。由于指纹图像复杂多变的纹理结构以及采集时可能遇到的各种噪声条件，传统的指纹生成方法需要对生成过程中遇到的各种情况进行建模，生成过程较为复杂，且真实性较差。随着深度学习的快速发展，端到端的生成模型逐渐应用于指纹生成领域。其中，GAN由于其简单易行的训练策略和高质量的生成效果而深得研究者的广泛使用。GAN、VAE和扩散模型等生成模型逐渐被用于生物特征合成中，该类模型从潜在向量直接或间接生成图像，不需要对中间过程进行复杂的建模即可生成与真实图像相近的合成图像，接下来本文将分别介绍基于这3种模型进行指纹生成的方法。

GAN模型通过训练生成器和判别器来使二者达到纳什均衡，但是往往很难判断何时达到纳什均衡，以及定量给出停止的标准。由于GAN的特殊训练机制，在很多情况下都会面临着难以收敛、训练不稳定以及模型崩塌等问题，因此一些研究者对GAN模型做出了许多改进，旨在提高生成质量和训练稳定性。总的来说，人们对GAN模型的改进主要分为几个方面，分别是改进模型架构（Minaee和Abdolrashidi，2018b；Zhong等，2021；Striuk和Kondratenko，2021）、优化训练策略（Bontrager等，2018；Cao和Jain，2018；Fahim和Jung，2020；Mistry等，2020；Bahmani等，2021）、多阶段GAN生成（Riazi等，2020；Sams等，2022）和与传统方法结合（Wyzykowski等，2021）等。

Minaee和Abdolrashidi（2018b）是第一个使用基于深度学习的方法进行指纹生成的，为了提高生成质量，他们采用了DCGAN（Radford等，2016）架构，DCGAN将卷积神经网络和GAN结合到一起，生成网络和鉴别网络都运用到了深度卷积神经网络，提高了训练的稳定性和生成质量。此外，Minaee和Abdolrashide（2018b）在损失函数中加入了正则项——总方差，总方差会惩罚相邻像素之间变化较大的情况，从而使生成的脊线线条更连贯。但最后生成的样本有噪音且不具备真实的指纹特征。Striuk和Kondratenko（2021）、Zhong等人（2021）也是基于DCGAN模型，对生成器和判别器进行改进，以生成逼真的指纹图像。但最后都遭到了模式崩溃的影响，生成图像的多样性随迭代次数增加而减少，且图像中有很多黑点，质量较差。

为了避免上述模式崩塌带来的影响，Arjovsky等人（2017）提出WGAN（Wasserstein GAN）模型，该模型引入Wasserstein距离来代替原来的JS（Jensen-Shannon）散度距离，用于测量真实分布和生成分布之间的差异，可以更好地训练判别器以防止模式崩溃，后续的很多研究（Bontrager等，2018；Cao和Jain，2018；Mistry等，2020）都采用了该方法。Bontrager等人（2018）首次使用WGAN进行指纹生成，但由于其使用的数据集是从真实数据集中随机提取的128×128像素的部分指纹图像，因此GAN的生成器很难学习到完整指纹的全局形状，导致生成的图像中含有很多模糊区域，且相比典型尺寸512×512像素，其生成的尺寸仅为小尺寸128×128像素的指纹块。为此，Cao和Jain（2018）使用改进的WGAN模型（Gulrajani等，2017），旨在生成大尺寸512×512像素的高质量完整指纹图。为了解决直接使用GAN进行大尺寸图像生成而导致的质量较差的问题，Cao和Jain（2018）首先训练了一个卷积自动编码器CAE来对真实图像进行编码和解码重构，接着用训练好的解码器来初始化IWGAN的生成器，从而提高其生成图像的质量。为了生成新样本，在输入的潜在向量中加入随机噪声，从而可保证生成新的高质量的指纹图像。但该方法并没有很好地避免模型崩塌，最终生成的图像中仍然含有很多黑点，且多样性较差。因此Mistry等人（2020）在Cao和Jain（2018）的基础上引入了身份损失，它指导生成器在训练过程中合成具有更明显特征的指纹图像。具体做法为用DeepPrint的深度网络（Engelsma等，2021）提取生成器生成的图像的固定长度表示，然后最小化小批量中每对表示之间的相似性，以引导生成器生成具有不同身份的指纹图像。Mistry等人（2020）通过与Cao和Jain（2018）中的合成指纹对比相似度分数证明了其生成的多样性。

Fahim和Jung（2020）针对GAN的训练不稳定问题采用了很多措施；通过使用平均残差连接，有效缓解了梯度消失问题；引入了loss-doping机制，以条件

方式强化生成器的损失函数，使训练过程避免陷入停滞，从而加快模型收敛速度；此外，还对超参数的不同选择进行了很多消融实验，选择出最有利于模型性能的超参数。通过对比不同方法的MS-SSIM值发现所提出的方法可以获得较低的值，即生成的图像相似度低，这说明了该模型得到了稳健的训练，不易崩塌。尽管该方法提高了训练的稳定性，但生成图像的真实性有待提高。

可以看出，GAN似乎很难达到训练稳定性和生成质量之间的平衡，针对这一问题，Bahmani等人(2021)提出一种基于渐进式增长的GAN模型，该方法基于StyleGAN的架构(Karras等，2019)，采用多分辨率训练的方法，在训练过程中通过逐渐增加空间分辨率来训练生成器和判别器，可以生成大尺寸512×512像素的指纹图像，且没有受到模式崩塌的影响，生成的多样性和质量均高于之前的结果。

除了改进GAN的模型架构和训练策略外，一些研究者发现比起直接从潜在空间生成图像，通过多阶段GAN生成图像似乎能达到更好的效果(Riazi等，2020；Sams等，2022)。Riazi等人(2020)提出一种两阶段过程生成指纹，第1阶段基于WGAN方法生成低质量的指纹图像，该阶段主要学习真实指纹的分布情况；第2阶段使用超分辨率(SR)模型ESRGAN将低质量图像转换为真实的高质量图像，该阶段可以生成逼真的细节特征。最后通过训练多个分类器来分类真实指纹和合成指纹，证明了合成指纹的不可区分性，但没有对生成的多样性进行分析。Sams等人(2022)也是基于两阶段GAN生成的方法，首先从真实指纹数据库与合成指纹数据库中提取出6种指纹架构，用于训练StyleGAN2(Karras等，2020)生成不同类型的指纹架构，生成的图像包含核心点、三角点、山脊、分叉和交叉等特征。之后用真实数据集训练CycleGAN(Zhu等，2017)来对指纹架构加入空隙、噪声和模糊等，转换为逼真的指纹。该方法可以生成包含很多真实细节特征的指纹图像，看起来更接近真实图像，并通过多项指标，如NFIQ分数分布和MS-SSIM分数证明了该方法生成图像的质量和多样性。

尽管目前基于深度学习的方法似乎已经取得了很好的生成效果，但比起传统方法，这些方法都没有在类内多样性上进一步讨论，因此一些研究者开始研究用深度学习模型为同一身份生成多个压痕的方法(Kim等，2019；Wyzykowski等，2021；Engelsma等，2023)。Kim等人(2019)首先通过两阶段GAN生成指纹图像并提取其细节信息组成主细节集，再使用条件GAN从细节图生成指纹图像。通过对同一细节图加入旋转、偏移等变换生成多个压痕，增加类内多样。Engelsma等人(2021)通过3个GAN模型分别学习指纹的身份信息、扭曲效果、纹理细节等特征。之后可通过固定身份信息，改变扭曲和纹理信息为给定身份的指纹生成多个压痕，该方法可以生成高质量的大尺寸512×512像素指纹图像。Wyzykowski等人(2021)提出一种传统方法和GAN结合的方法，首先使用改进的Anguli(SFinGe的开源实现)生成主指纹，前面提到过SFinGe生成的指纹脊线是统一变化的，因此该方法基于正弦函数动态地改变脊的厚度来模拟真实脊线的粗细变化。接着参考真实指纹的空隙、划痕分布为主指纹添加这些特征。为了模拟类内变化，Wyzykowski等人(2021)结合真实数据中的类内差异，对主指纹进行了切割、旋转和变形等操作。最后使用CycleGAN将主指纹转换为真实指纹，其中引入了身份映射损失以避免转换时丢失或改变主指纹的细节特征。该方法可以生成更精细的3级指纹，为指纹识别提供了更精细的特征。

变分自动编码器(VAE)是一种采用变分推断思想的生成方法，可以使隐变量服从某一特定分布。Attia等人(2019)提出使用VAE学习指纹的复杂分布，通过编码器将真实指纹映射到潜在空间中，得到两个变量mu、sigma，接着通过这两个变量采样得到潜在向量z；最后编码器通过z重建输入图像。通常，VAE是通过最小化编码器与解码器两个分布的KL散度和真实输入与重建输入的均方误差进行优化的，为了保持边缘和细节的视觉合理性，本文还使用了结构相似性度量(SSIM)作为优化函数来提高生成质量，但最后生成图像中存在失真模糊图像和非完整指纹图像(partial fingerprint)，这可能是由于训练数据集中包含低质量图像所造成的。

扩散模型是最近新兴的一种生成模型，通过对图像正向加噪再逆向去噪生成新图像，在图像生成方面表现出很好的前景。杨光锴(2023)首次使用去噪扩散模型进行指纹生成，得到了具有较高质量的指纹图像，但由于扩散模型的去噪过程需要迭代上千步，因此该方法的生成速度很慢。

2.3.3 掌纹生成方法

掌纹中包含丰富的特征信息，具体可分为纹线特征和细节特征，其中纹线特征是指掌纹中的多条掌线，也是最为直观的特征，细节特征是指掌纹中的细节点，如分叉点、脊末端等。纹线特征中最清晰的几条纹线可以保持终生不变，且在低质量的图像中不被遮盖，而细节特征易受污渍、采集设备等影响，需要在高质量的图像中才能清晰辨认。因此大部分的掌纹识别算法都是基于掌纹的纹线特征进行识别，同理，其合成图像的生成也主要以纹线特征生成为主。根据纹线的分布，可以将其分为两级：一级主线是掌纹中最粗、最清晰的掌线，通常为2～4条；二级主线为掌纹中的细小纹线，通常为5～10条。掌纹的纹理模型即对这些主线进行建模，之后再通过生成模型将其渲染为包含皮肤纹理的真实掌纹。表4汇总了目前关于掌纹生成的方法，并在下文分别展开介绍。

表4 掌纹生成方法汇总

Table 4 Summary of palmprint generation methods

论文	掌线模型	渲染方法	类内变化	分析
Wei 等人(2008a)	提取算法	采样模型	是	开创性工作
Minaee 等人(2020)	DCGAN	无	否	直接使用GAN生成
Zhao 等人(2022)	几何模型	无	是	首次提出使用贝塞尔曲线生成掌线
Shen 等人(2023)	几何模型	GAN	是	使用改进的生成控制身份的掌线，通过GAN生成真实掌纹图像
Jin 等人(2024)	几何模型	GAN	是	引入PCE域降低生成难度
Jin 等人(2025a)	几何模型	扩散模型	是	扩散模型在掌纹生成领域的首次应用

1)掌纹图像的传统纹理建模方法。关于掌纹生成的研究首次出现在Wei等人(2008a)的方法中，Wei等人(2008a)使用Canny算子从真实掌纹图像中提取主线作为合成图像的纹线特征，再对真实图像进行模糊主线操作，以便使用基于图像块的采样方法生成细节特征，如皮肤纹理、山脊等，最后将纹理特征与细节特征进行融合即可得到掌纹图像。对于类内变化，Wei等人(2008a)参考了Cappelli等人(2002)合成指纹的操作，使用变形网络模拟皮肤扭曲。但是该方法使用的是从真实掌纹中提取的纹线，无法生成新的身份，因此存在身份泄露的风险。

研究表明，掌纹的纹线特征对基于CNN的掌纹识别方法起到关键作用(Zhao等，2022)。因此纹线特征的生成无疑是掌纹生成的关键问题，如何生成新的纹线特征成为一个研究难题，而贝塞尔曲线的出现似乎很好地解决了这个问题。Zhao等人(2022)提出一个简单有效的几何模型BezierPalm，通过控制贝塞尔曲线参数生成多条贝塞尔曲线组成掌纹纹线，并通过随机在参数中添加高斯噪声和选择自然图像作为合成样本的背景来丰富类内多样性。最后得到了由多条主线和褶皱线组成的几何掌纹图。通过使用几何掌纹图对识别模型进行预训练，再在真实数据上微调，其效果优于完全由真实图像训练出来的模型。但是该方法生成的图像为几何掌纹线条，与真实图像有着很大的域差距。

2)掌纹图像的深度学习生成方法。随着深度学习的不断发展，一些研究者提出使用GAN模型直接从潜在空间进行生成掌纹，可以避免掌纹纹线生成的难题。Minaee等人(2020)使用了与Minaee和Abdolrashidi(2018b)中指纹生成类似的方法，采用DCGAN架构，并加入了正则项来使生成的掌线更连贯。但是直接从潜在空间进行生成，显然忽视了掌线分布的先验条件，使得生成的掌线纹理杂乱无章，不符合真实分布。

受BezierPalm和深度学习方法的启发，一些研究者提出新的融合方法。Shen等人(2023)提出RPG-Palm，使用改进的BezierPalm生成纹线特征作为身份条件，再通过条件GAN将其转化为掌纹图像。为了引入类内多样性，其设计了一个条件调制生成器，使用从随机噪声编码的潜在控制向量生成多样化的类内纹理和照明条件。该方法合成的图像数据用于预训练识别模型，可使其精度超越原来的BezierPalm，有效地将对真实数据的依赖降低90%。

由于BezierPalm使用自定义的规则随机生成掌线纹理，其分布不一定符合真实图像的分布，且贝塞尔图像与真实图像之间存在较大的域差距，导致使用RPG-Palm合成图像预训练后的识别模型的小样本学习能力较差。为了进一步提升生成质量和减少对真实图像的依赖，Jin等人（2024）提出PCE-Palm。该方法引入PCE image作为贝塞尔图像和真实掌纹图像的桥梁以减少域差距，其中PCE image是从真实掌纹中提取的掌线纹理图案，其纹理分布与真实图像一致。该方法使用两阶段生成，先将贝塞尔图像映射到PCE领域，使其学习到真实掌纹的纹理分布，再从PCE image生成掌纹图像。与Shen等人（2023）提出的合成方法相比，PCE-Palm预训练的识别模型达到了更高的识别精度，且其小样本学习能力大大提升。

为了进一步提高生成掌纹的真实性，减少在掌纹识别过程中对真实数据的依赖，Jin等人（2025a）提出Diff-Palm方法。在该方法中，一方面为了获得更加真实准确的掌纹线条，一种基于多项式的掌纹表征被提出，取代了Bezier线条；另一方面，为了获得类内更加丰富多样的掌纹，一种可控的扩散模型被提出。该方法不仅可以生成更加真实的、类内丰富的掌纹图像，同时能确保身份特征的一致性。

2.3.4　静脉生成方法

静脉是指全身的各个器官输送血液回到心脏的血管，遍布人体全身，一般用于生物特征识别的主要为手背静脉、指静脉和掌静脉。其中，指静脉和掌静脉可以与指纹和掌纹一起采集，用于多模态识别。由于静脉属于人体内部特征，因此其采集图像相比掌纹、指纹等有很大差异，其图像的合成过程也有所不同。在采集时，由于光的散射和吸收，近红外辐射在人体组织中的穿透很小。在红外辐射下，只有手背上的一些静脉是完全可见的，而其他静脉则无法检测到。由于静脉生物特征识别系统只能检测浅静脉，因此模拟图像必须具有相同的特征。

由于静脉识别应用较少，因此关于其生成图像的研究也较少，根据调研，目前的生成方法按静脉类型可分为手背静脉生成（Crisan等，2008）、指静脉生成（Hillerström等，2014；Yang等，2020）、掌静脉生成（Salazar等，2021a，b），以及一种通用的静脉生成方法（Ou等，2022）。表5汇总了目前关于静脉生成的方法，并在下文展开介绍。

1）静脉纹理建模方法。针对指纹生成提出的SFinGe已经可以达到很好的效果，其合成的指纹数据库也广泛应用于指纹验证竞赛（FVC2000，FVC2002，FVC2004），在手背静脉生成领域，Crisan（2008）提出VEINSIM，旨在达到SFinGe的生成效果。该方法从真实图像中提取关键点，并使用改进的crossing number method（Maltoni等，2009）将其分为端点、节点和分段点，基于重建算法来重新计算这些关键点的位置，得到新的静脉图案。之后使用膨胀算法重建静脉的厚度。为了模拟皮肤纹理与采集时的噪声，Crisan（2008）使用了一系列数学模型和图像处理技术来实现。但是该方法相比指纹生成的SFinGe略显稚嫩，对于其生成效果，Crisan（2008）没有给出任何评估实验进行证明。

Hillerström等人（2014）受叶脉模式生成的启发，使用Runions等人（2005）的生长算法，该算法基于两条主血管，利用源节点引导新的静脉节点的生长。之后对这些静脉节点的位置加入扰动，使用膨胀算法和变薄算法将静脉节点连接在一起得到由多条血管组成的静脉图案。为了模拟采集时皮肤和血液的吸收与散射，使用滤波器和数学模型来建模一些噪声因素，最后得到一个包含50 000个数据的指静脉数据库（包含5 000个个体，每个个体10个样本）。但受限于图像处理规则，该方法生成的图像不符合真实数据分布。

Yang等人（2020）通过观察真实指静脉图像发现，每个静脉图案都由两根较粗的主静脉和延伸出的分支静脉组成，因此提出一种由生长节点引导静脉节点逐步生成静脉图案的算法。

Salazar等人（2021a）受Physarum生成模式的启发，使用Physarum的基于代理的生长算法（Liu等，2017）生成血管网络的轨迹，并将高斯模糊核加入该生成器中模拟采集过程中近红外照明的效果。

可以发现，目前静脉图案的生成大多是受自然界生物的启发，通过模拟其生长模式或结合静脉特性而提出生长算法。Ou等人（2022）提出一种新的静脉图案生成方法，使用了随机块合成技术（random block composition，RBC），即将真实的静脉图案分成多个图像块，通过随机组合不同身份相同位置的图像块来得到新的静脉图案。通过RBC可以以简单有效的方式生成任意部位的静脉图案，包括指静脉、掌静脉等，是一种通用的静脉生成方法，评估

表5　静脉生成方法汇总

Table 5　Summary of vein generation methods

方法	静脉类型	静脉图案	静脉厚度	渲染方法	类内变化	分析
Crisan 等人 (2008)	手背静脉 vein	重建算法	膨胀算法	图像处理技术	是	开创性工作
Hillerström 等人 (2014)	手指静脉 vein	叶脉模式生长算法	Murray' law	数学模型	是	首次将自然界的生物特性引入静脉图案生成
Yang 等人 (2020)	手指静脉 vein	生长算法	Murray' law	GAN	是	GAN 在指静脉生成领域的首次尝试
Salazar 等人 (2021b)	手掌静脉 vein	StyleGAN	–	–	是	完全基于深度学习方法进行静脉图像生成
Salazar 等人 (2021a)	手掌静脉 vein	Physarum 生长算法	–	GAN	是	受 Physarum 生长的启发生成静脉图案
Ou 等人 (2022)	所有类型的静脉 veins	RBC+CycleGAN	–	GAN	否	提出一种全新的通用静脉图案生成方法
Shang 等人 (2025)	手掌静脉 vein	PVTree	–	GAN	是	提出一种 3D 手掌血管树建模方法

注：使用 GAN 等生成模型是端到端生成，没有使用额外算法计算静脉厚度和进行渲染操作。"–"表示无相应数据。

实验结果证明了该方法生成任意图案的静脉图像的可行性。

2）静脉深度生成方法。在将上述纹理图案渲染为真实的静脉图像时，为了实现更真实的渲染效果，Yang 等人（2020）和 Ou 等人（2022）都使用 CycleGAN 学习真实数据的分布，从而实现两种风格的转移。其中，由于静脉采集的特殊性，一些微小血管因缺少血红蛋白而在采集图像中很难成像，为了实现这种效果，Yang 等人（2021）没有使用传统的 U-Net 架构（Ronneberger 等，2015），该架构会从浅层网络中提取一些遗漏的特征，将微小分支转化为清晰图像，而是使用了 Jonnson 等人（2016）提出的生成器架构，旨在凸显静脉的主要分支而不是所有分支。Yang 等人（2021）使用多个度量指标证明了该方法生成的图像分布与真实图像的相似性，且用于预训练识别模型后，取得了很低的等错误率（equal error rate，EER）。在 Ou 等人（2022）的方法中，为了解决直接拼接导致的图像块状明显、线条不连贯等问题，使用 CycleGAN 对其进行细化，同时保持静脉图案不变。得到连通平滑的静脉图案后，再使用 GAN 将其渲染为真实的静脉图像。

Salazar 等人（2021a）在渲染时为了模拟采集静脉时掌纹对其的影响，提出一种掌纹生成方法，使用具有 StyleGAN2 架构的自训练 GAN 模型创建，将生成的掌纹细节与静脉图案混合即可得到真实的静脉图像。为了确保生成图像的身份唯一性，建立了唯一性阈值，若新生成的图像与合成数据库中所有图像的相似度大于该阈值则被丢弃。最后通过随机组合对比度、模糊效果、平移和旋转等视觉变化为单个身份生成多个样本。

此外，Salazar 等人（2021b）使用了一种完全基于 GAN 的方法生成静脉，对真实数据进行数据增强和扩充后，直接使用 StyleGAN2 进行掌静脉生成，其唯一性判别方法和类内变化与 Salazar 等人（2021a）的方法类似，最终得到包含 60 000 幅图像的掌静脉数据集 Synthetic-sPVDB（包含 10 000 个个体，每个个体 6 个样本）。

为了提升掌静脉图像的身份独特性和类内变化可控性，Shang 等人（2025）提出 PVTree 模型。该方法可以建模出 3D 手掌血管树，得到一个独特的身份。接着通过多视角投影将其映射到二维空间中，从而得到自然的类内变化。之后与 Bezier 曲线模拟的掌纹线条混合得到最终的掌静脉纹理。PVTree 不仅满足了身份一致性和类内多样性的需求，而且其训练的识别模型性能首次超越了真实数据的效果。

2.3.5　虹膜生成方法

相比指纹、掌纹等生物特征，虹膜有一个很大的

缺陷是易于伪造，常见的呈现攻击的伪虹膜图像包括打印的高质量真实虹膜图像、手工制作的假眼和带有特定纹理的隐形眼镜。因此其生成类型也可分为真实虹膜图像生成（Cui 等，2004；Makthal 和 Ross，2005；Zuo 等，2007；Wei 等，2008b；Wecker 等，2010；Kohli 等，2017；Minaee 和 Abdolrashidi，2018a；Wang 等，2022b）和伪虹膜图像生成（Zou 等，2018；Yadav 等，2019；Yadav 和 Ross，2021），这些合成图像可用于提高虹膜识别算法和 PAD 算法的性能，表 6 中对这些方法进行了总结。

目前用于真实虹膜图像生成的方法可分为传统方法（Cui 等，2004；Makthal 和 Ross，2005；Zuo 等，2007；Wei 等，2008b；Wecker 等，2010）和深度学习方法（Kohli 等，2017；Minaee 和 Abdolrashidi，2018a；Wang 等，2022b）。在传统方法中，Cui 等人（2004）、Wecker 等人（2010）通过重组原始虹膜的特征信息生成新的图像，Makthal 和 Ross（2005）、Wei 等（2008b）利用原始虹膜的纹理图像块引导生成新的纹理图案，Zuo 等人（2007）从解剖学的角度从零开始生成。

1）虹膜纹理建模方法。主成分分析法（principal components analysis，PCA）可以提取全局特征，且适合于图像构建，因此 Cui 等人（2004）受 PCA 方法的启发，首次对虹膜图像进行生成。PCA 是一种特征压缩方法，通过比较方差大小，选择大方差即包含信息量多的特征实现压缩。Cui 等人（2004）基于这一点，提出一种通过控制特征系数构建虹膜图像的方法。首先给定阈值将空间划分为不同的类，对于每个类内的图像生成，通过对真实虹膜图像的特征系数进行小范围修改得到新的虹膜纹理。由于在实践中，每个虹膜图像的向量会被下采样，使得生成的图像质量较差，因此使用 Huang 等人（2003）提出的超分辨率方法进行图像增强。最后通过类内和类间图像的距离分布证明了该方法可以实现很好的聚类效果。但是由于 PCA 提取的是全局特征，因此最后生成的图像缺乏局部特性，真实性较差。

与之类似的方法为 Wecker 等人（2010）提出的一种多分辨率方法，该方法将虹膜图像分解为多个分量，再重新组合它们得到新的虹膜图像。首先对真实图像进行预处理得到只包含虹膜信息的矩形图像，之后使用 Chaikin 反向细分滤波器，对图像进行 4 级分解再重建。文章证明了对于该分辨率图像，4 级分解足以捕捉到虹膜图像的所有特征。每个虹膜图像可以被分解为 5 个新的分量，在合成时为了创造新身份，会使用来自多个真实虹膜图像的细节得到独特逼真的图像。由于从各种虹膜中提取的特征可能不兼容，因此并非所有的组合都能产生具有高逼真度的合成虹膜。本文将虹膜分为两种类型：单区域和双区域，只组合来自兼容虹膜（来自同一组）的成分，从而提高结果的真实性。生成的图像与

表 6　虹膜生成方法汇总

Table 6　Summary of iris generation methods

目的	方法	类内变化	分析
真实虹膜图像生成	PCA+SR（Cui 等，2004）	是	通过控制 PCA 提取的特征的系数合成虹膜图像
	多分辨率方法（Wecker 等，2010）	否	使用 Chaikin 反向细分滤波器对虹膜图像进行分解再重组
	马尔可夫随机场（Makthal 和 Ross，2005）	是	使用 MRF 引导初始化噪声图像逐步生成特定图案
	基于图像块的采样方法（Wei 等，2008b）	否	根据合成图像与真实图像边界区域距离逐步更新每一个图像块
	基于解剖学的建模（Zuo 等，2007）	是	基于解剖学将虹膜分为多层，分别对每一层进行建模生成
	DCGAN（Minaee 和 Abdolrashidi，2018a）	否	首次使用 DCGAN 进行虹膜生成
	DCGAN（Kohli 等，2017）	否	引入质量评估函数作为先验条件提高生成质量
	StyleGAN2（Wang 等，2022b）	是	使用解纠缠的方式控制类内变化
伪虹膜图像生成	CycleGAN（Zou 等，2018）	否	额外加入两个判别器增加类间多样性
	CycleGAN（Yadav 和 Ross，2021）	否	引入 Styling Network 学习不同领域的风格
	RASGAN（Yadav 等，2019）	否	使用相对性特性旨在提高合成图像比真实图像更真实的概率

真实图像具有良好的统计相似性，但多分辨率方法会提取所有的特征包括反射，使得生成的图像中存在反射，这在视觉上增加了真实性，但可能会影响识别算法的性能。

由于虹膜内容主要包含在虹膜块中，是虹膜纹理的有效表示，因此出现了一些利用原始虹膜图案进行生成的方法。Makthal 和 Ross（2005）提出一种基于马尔可夫随机场（MRF）的虹膜生成方法，该方法可以描述控制特定邻域的像素值的概率分布。首先从真实图像中提取多种不同的原始纹理图案，对初始化噪音图像中的每个像素，比较其所有邻域与原始纹理图案的距离，选择距离最近的像素来更新该像素，引导噪声图像逐步得到局部与原始图像相似的纹理图案。为了实现合成的随机性，根据图案类型出现的位置为它们添加不同的权重参数，根据权重大小决定是否生成该图案。生成新的身份即图案后，考虑了旋转、弹性变形和相机噪音等因素，对同一身份的虹膜图像进行一定的变换操作，增加类内变化。最后通过聚类实验证明了该方法合成的虹膜图像不同于其他随机纹理图像，且与真实图像的纹理图案很相似。

Wei 等人（2008b）使用虹膜模式作为基本元素，引入基于图像块的采样方法来合成各样的纹理。首先对真实图像做预处理得到归一化的去除眼睑、睫毛等噪音的矩形图像，之后使用基于图像块的采样方法合成新的虹膜模式。为了增加类内变化，使用 Wei 等人（2007）的非线性扭曲变形模拟瞳孔扩张和收缩时纹理的扭曲效果，高斯滤波器用于平滑图像实现散焦效果，此外还通过加入噪音、随机移动图像像素、极坐标水平平移来实现扰动效果和旋转效果。最后使用双线性插值将矩形图像变为环形图像得到最终的合成虹膜图像。

上面这些方法都利用了原始虹膜图像的信息，这有可能会造成一定程度的身份泄露。Zuo 等人（2007）使用解剖学原理提出一种从零开始的生成方法，首先将虹膜分成多个层，之后对每一层都进行建模以得到最后的虹膜图像。首先使用一组随机参数生成纤维的三维立体结构，再投影到二维平面内得到正视图像。之后使用余弦函数和平滑高斯噪声对不规则边缘的顶层和凹凸不平的半透明层进行建模，最后通过提亮睫状区实现最后一层。实验结果表明，识别性能与纤维结构有很大关系，对于其他参数不太敏感。

2）虹膜深度生成方法。传统方法大多步骤繁多、计算量大，且生成的图像真实性较差，因此与其他生物特征类似，基于深度学习的虹膜生成方法很快涌现。Minaee 和 Abdolrashidi（2018a）直接使用 DCGAN 进行虹膜生成，实现了 FID = 42.1 的效果。之后，Kohli 等人（2017）在 DCGAN 的基础上引入质量评估函数作为先验知识来引导生成器生成质量比较高的图像。具体做法是，生成器从潜在空间生成图像后还会计算其质量分数，并去除质量分数低于第 1 四分位数的合成图像。在判别器进行判别时，也会去除质量分数较低的真实图像，这使得生成的图像质量大大提升。

这两种方法尽管实现了较高质量的图像生成，但无法为同一身份生成多个样本，因此 Wang 等人（2022b）引入对比学习来实现它。使用对比学习方法将潜在空间根据虹膜特性分为 3 个子空间，分别控制虹膜的身份特性、类内变化和纹理特性。使用双通道的方式输入生成器进行生成，先输入控制身份特性和类内变化的潜在向量用于生成整体架构，接着输入控制纹理特性的潜在向量生成纹理细节。这种输入方式可以使生成器在保持纹理不变的情况下为同一身份的虹膜图像生成多个类内变化。在模型训练时使用虹膜验证网络判断一对图像是否属于同一类，通过相对损失进行优化，使同一类的图像，即身份特性、纹理特性相同，类内变化不同的图像距离更近，不同类的图像，即身份特性、纹理特性和类内变化都不同的图像距离更远。其中 GAN 使用改进的 StyleGAN2 架构。该方法生成的合成图像在视觉上比之前方法生成的图像更清晰真实，且实现了 FID = 5.27 的效果，相比 Minaee 和 Abdolrashidi（2018a）的方法有了很大的提升。

对于伪虹膜生成，研究者主要利用 CycleGAN 在两种领域的风格转换能力来实现从真实虹膜到伪虹膜的风格转换。Zou 等人（2018）旨在在真实虹膜上添加纹理隐形眼镜，由于 CycleGAN 进行生成时，其训练机制使得模型为了降低损失值而生成比较相似的图像，导致多样性较差，因此受 D2GAN（dual discriminator generative adversarial nets）（Nguyen 等，2017）启发，引入两个判别器，分别用于评估生成图像与真实图像的相似性，从而更有效地引导生成器提升图像质量。额外的对抗关系同时也导致模型的

收敛较慢，但对比实验证明了该方法相比CycleGAN在多样性生成上有较大提升。受CycleGAN的局限性影响，该方法只能实现两个领域间的风格转换，即生成一种类型的伪虹膜图像，若需要其他类型，则需要重新训练，较为麻烦。因此，Yadav和Ross（2021）提出CIT-GAN（cyclic image translation GAN），实现所有类型的呈现攻击的图像生成。CIT-GAN可以视做CycleGAN和Styling Network的结合，可以实现真实图像在多个目标领域的风格转移。Styling Network用于学习不同领域的风格编码，之后风格编码与真实图像输入生成器得到具有该风格的领域图像。判别器有多个分支，用于判别输入图像是真实图像还是其他领域的图像。使用合成的数据训练呈现攻击检测方法，识别性能得到提升。

但是传统GAN的鉴别器旨在提高区分真实图像与伪图像的能力，这种方法在生成低分辨率图像时表现良好，但在高分辨率上效果不佳。为了提高生成质量，Yadav等人（2019）提出RASGAN（relative average standard GAN）进行虹膜生成。RASGAN使用了相对性的特性，其生成器在生成图像时不再单独生成，而是需要与真实图像进行对比，旨在提高合成图像比真实图像更真实的概率。实验结果表明，该方法生成的图像在用于训练虹膜识别算法时，显著提升了模型对伪造攻击的泛化能力，使其能够检测出训练阶段未出现的伪虹膜图像。

2.3.6　人脸生成方法

如果在网络上搜索指纹、掌纹等图像，通常会得到很少的信息，因为这些生物特征很难被摄像头捕捉到，而人脸恰恰相反，随着社交媒体的广泛发展，人脸图像随处可见，但是可用于各类人脸应用的标准数据集不仅需要足够多的身份数量，还需要广泛的类内变化。因此针对人脸的生成方法大部分集中于细粒度人脸属性生成，即在保持身份的情况下改变人脸的姿态、表情等。对于人脸的纹理建模方法，由于存在大量的素描数据库，因此无需对其进行建模。目前的人脸生成任务从应用角度主要分为3个领域：素描转人脸生成、细粒度人脸生成和高质量人脸生成。表7汇总了关于人脸生成的方法。

1）素描转人脸生成方法。与之前介绍的生成方法类似，这是一种图生图的风格转移任务，需要在两个领域之间实现很好的风格转移。由于大部分素描图比较稀疏和粗糙，因此Wang等人（2018）使用了距离变换的稠密表示，其提取的特征图对于不完整和噪音更具有鲁棒性。除了对素描图进行增强外，大部分基于GAN的方法的改进主要在模型架构和损失函数这几个方面。

Chen等人（2018）的SketchyGAN提出一种网络构建块MRU（masked residual unit），在将输入的线条图与特征图进行混合时，使用掩码来动态决定需要保留的信息，有助于提高生成质量，GAN模型的生成器与判别器都由该模块堆叠而成。SketchyGAN的判别器是传统的图像块判别器，在局部图像块中逐步区分真假，这无法捕获全局信息，因此后续出现了多尺度判别器。

Wang等人（2018）基于CycleGAN框架，使用多判别器网络，在不同分辨率上进行监督，优化其潜在表示，从而产生高质量的合成，减少伪影。Li等人（2019b）受SAGAN（self-attention GAN）（Zhang等，2019b）的启发，提出一种条件自注意力模块（conditional self-attention mechanism，CSAM），将线条图与特征图拼接在一起后，使用自注意力机制建立长程依赖性，充分利用条件信息。此外，该方法也引入多尺度判别器，与Chen和Hays（2018）对不同尺度的图像进行判别的方式不同，Li等人（2019b）固定真假图像的尺寸，改变判别器的深度以实现不同大小的感受野，实验证明该方法可以获得更高的图像质量。Li等人（2022）使用两阶段生成，第1阶段采用自注意力模块获得全局信息，初步生成粗糙的人脸图像；第2阶段使用多个具有不同下采样因子的多尺度判别器和图像块局部判别器改进细节信息。

Chao等人（2019）使用深度残差U-Net作为生成器、带有残差块的PatchGAN作为判别器进行人脸生成。Bai等人（2022）与上述方法不同，并非从素描图到真实人脸图像的风格转移任务，其采用双编码器架构，分别对输入的任意风格图，如素描、语义图等和指定身份的图像提取其风格信息和身份特征，基于StyleGAN架构将两种特征进行融合，从而生成相应的图像。

除了改进模型架构之外，构建不同的损失函数约束对于生成高质量的人脸图像也至关重要。对于以GAN为架构的生成模型，对抗损失可以使生成器和判别器在生成图像的真实性方面互相竞争，但对于sketch2face任务而言，生成的真实性不足以约束两个领域之间的风格转移，因此需要更多的更细致

表7　人脸生成方法汇总

Table 7　Summary of face generation methods

目的	年份	方法	分析
素描转人脸	2018	GAN with MRU(Chen 和 Hays, 2018)	提出一种网络构建块MRU,可以提高生成质量
	2018	GAN with multi-D(Wang 等, 2018)	使用多判别器网络,在不同分辨率上进行监督
	2019	GAN with multi-D(Li 等, 2019b)	引入条件注意力模块以充分利用条件信息
	2021	GAN with multi-D(Li 等, 2022)	使用两阶段生成方法,从粗糙到细致
	2019	Improverd GAN(Chao 等, 2019)	对生成器和判别器的模型架构进行改进
	2022	StyleGAN(Bai 等, 2022)	可以解决素描图过于模糊无法提供身份信息的问题
可控生成	2017	CGAN+VAE(Bao等, 2017)	提出一种均值特征匹配目标来提升训练稳定性
	2018	CGAN(Choi等, 2018)	使用域条件来实现多个属性的转移
	2023	Diffusion Model(Huang等, 2023)	提出一种动态影响因子来融合不同模态的条件
	2023	Diffusion Model(Kim等, 2023)	融合不同的人脸特征来控制扩散模型生成人脸
	2017	CGAN(Yin等, 2017)	利用3DMM得到先验条件来加速模型收敛
	2018	CGAN(Shen 等, 2018b)	使用多个GAN来学习3DMM提取到的特征分布
	2018	CGAN(Shen 等, 2018a)	将分类器加入到GAN的博弈游戏中
	2020	CGAN(Deng等, 2020)	使用多个VAE来学习3DMM提取到的特征分布
	2018	CycleGAN(Gecer 等, 2018)	从3DMM生成的人脸到2D人脸的风格转移
	2016	GAN(Chen等, 2016)	提出一种互信息度量以充分利用条件信息
	2017	VAE+GAN(Tran等, 2017)	判别器需要同时判别真实性和人脸特征以加强解纠缠
	2018	SDGAN(Donahue 等, 2018)	判别器需要同时判别真实性和id特征以加强解纠缠
	2018	GAN(Bao等, 2018)	可以生成训练集外的人脸身份
提升生成质量	2017	DCGAN(Curtó 等, 2020)	使用BatchNorm和SELU来代替ReLU层
	2018	Progress GAN(Karras 等, 2021b)	在训练时逐渐向模型添加层以提高分辨率
	2020	GAN with multi-D(Karnewar 和 Wang, 2020)	允许梯度在多个尺度上从判别器传递到生成器

的约束。常见的损失函数有类别损失、重建损失、感知损失和纹理损失等。其中类别损失主要用于条件GAN,对于生成图像的类别,生成器和判别器需要进行额外的竞争。重建损失多用于监督训练,对于生成器生成的伪图像,使用真实图像对其做像素级别的约束,以提升生成质量。像素特征损失可以保持低水平的内容相似,而感知损失定义为特征之间的距离,可以使高水平信息相似。此外,为了保持素描图中的纹理细节在风格转移时不被改变,使用滤波器或其他方法提取生成图像的边缘图与素描图做约束,丰富人脸的纹理细节。除了这些常见的额外约束之外,Chen和Hays (2018)引入一种多样性损失,对于同一素描图输入,生成器从不同的噪声生成的图像距离应尽可能大,以丰富类内多样性。Li等人(2022)使用语义损失在第1阶段约束生成图像与语义图的距离,使模型更好地学习人脸全局信息。此外还提出一种颜色损失,约束生成图像与原始图像的颜色差距。

2)人脸可控生成方法。人脸的可控生成大多服务于姿势不变的人脸识别(pose-invariant face recognition,PIFR),PIFR的解决方法一般分为两种:提取姿势不变特征和人脸正面化。可控生成属于第2种方法,旨在在保护身份不变的情况下,生成任意属性的人脸图像,显然该方法大多不能生成新的身份,只能生成类内变化。人脸特征包含身份特征和属性特征,如表情、姿势、照明和背景等,大部分研究者通过条件GAN和解纠缠来实现对不同特征的控制,如图19所示。

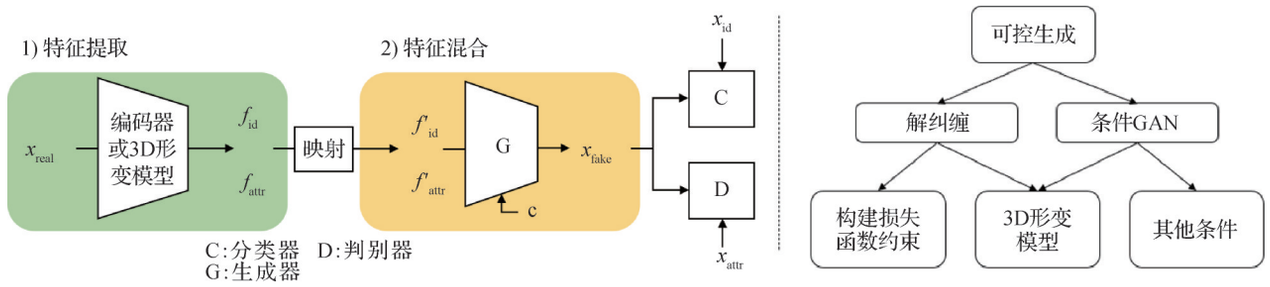

图 19　可控生成的一般流程

Fig. 19　General process of controllable generation

要生成特定属性的人脸图像，最常用的方法是使用额外的类别标签进行引导。Bao 等人（2017）将类别标签引入编码器—解码器结构的生成器中生成特定类别的图像，并引入分类器度量类别信息，同时提出一种均值目标损失函数来提高 GAN 的训练稳定性。StarGAN（Choi 等，2018）使用了一种信息更丰富的域条件，它包含所有的属性信息，但对数据集的要求很高，大部分数据集只包含部分标签信息，因此作者提出一种掩码向量使模型只关注已有的标签信息。

对潜在空间的解纠缠一般通过构造损失函数或三维可形变人脸模型（3D morphable model，3DMM）来实现。InfoGAN（Chen 等，2016）是该领域的一个创新性工作，提出一种无监督解纠缠的方法。该方法将潜在空间分成两个部分 $Z=(z, c)$，其中，z 是随机噪声，c 是结构化语义特征（可以理解为各类条件），构造了一种互信息损失 $I(c, G(z, c))$，可以度量从 $G(z, c)$ 中学习到的关于 c 的信息量。通过最大化互信息使模型充分利用 c 的信息，实现可控生成。该方法旨在提供一种解纠缠思想，在生成质量上较差。Donahue 等人（2018）提出的 SDGAN 采用成对训练的方法，使用相同身份但不同姿势的两幅人脸图像进行训练，其判别器需要判别其真实性和身份独特性。DRGAN（Tran 等，2017）的生成器采用编码器—解码器架构，编码器从真实图像中提取身份特征，之后与姿势特征和噪声一起输入解码器生成对应姿势的人脸图像，生成器与判别器在生成图像与真实图像的身份和姿势分类上进行竞争，以促使编码器提取到更多的身份特征，从而加强解纠缠。可以发现这些方法都是提取训练集中的人脸身份信息，生成新的属性特征，因此 Bao 等人（2018）提出一种开放集人脸生成的方法，可以生成训练集外的人脸身份。该方法使用人脸分类网络从任意人脸图像中提取身份特征，之后将身份特征与属性特征一起送入生成器得到相应人脸图像。其中，为了得到属性特征，他们使用权重较小的像素重建损失来保持伪图像与给定图像的属性相同，但是该方法依然不能生成新身份。

3DMM 是一种强大的 3D 人脸重建工具，可以从给定图像中得到身份、表情和纹理等特征，因此一些研究者开始使用它来辅助人脸的可控生成。Yin 等人（2017）提出一种任意姿势正面化人脸的方法，该方法需要真实的正面人脸进行监督，可用数据集较少，因此使用 3DMM 提取先验特征作为条件输入生成器来加速收敛。特别地，Yin 等人（2017）基于人脸对称性的观察，提出一种掩码损失来提高生成质量。直接使用 3DMM 提取的特征可能会在多样性方面有些损失，因此一些研究者开始学习提取出的特征的分布，通过从分布中采样得到更具多样性的属性特征。Shen 等人（2018b）和 Deng 等人（2020）分别通过训练多个 GAN 和 VAE 学习 3DMM 提取到的特征分布，Shen 等人（2018a）通过一个变换来改变真实特征。其中 Deng 等人（2020）还学习了身份特征分布，因此可以采样出新的身份特征，从而得到新的人脸身份。除了学习其分布外，还可以通过 3DMM 随机采样系数得到新的身份和属性特征。Gecer 等人（2018）通过随机采样系数得到一张新的 3D 人脸，之后使用类似 CycleGAN 的架构实现从 3D 人脸到 2D 人脸的风格转移，其中生成器使用了成对的真实数据进行训练，以增强不同域的转换。

对于服务于 PIFR 的人脸生成任务，在改变属性的同时保护身份不变是一个很重要的任务，上述方法大都使用额外的分类器对真实图像和伪图像的身份特征进行约束，分类器在这里只是起到了一个提

取特征的作用，而Shen等人（2018a）将分类器也纳入GAN的博弈游戏，其与判别器分别在身份识别和图像质量上区分两个域来与生成器竞争，从而进一步缩小真实图像与合成图像的域差距。

扩散模型作为近年来新兴的生成模型，已经展现出卓越的生成能力，因此也逐渐用于人脸的可控生成。Kim等人（2023）使用两阶段生成方法得到人脸数据集，第1阶段训练扩散模型生成新的身份图像；第2阶段使用提取到的身份信息和风格特征作为条件使用交叉注意力机制输入扩撒模型来生成特定风格的身份不变图像，其中风格图从训练集中任意选择。在生成类内图像时，为了使身份特征在生成时不被改变，Kim等人（2023）提出一种动态的身份损失来保护身份不变。Huang等人（2023）利用现有的不同模态的条件扩散模型，提出一种动态影响因子，用于预测每一步不同的扩散模型对预测噪音的贡献程度，将不同模态的条件进行融合，一起协作预测最后的噪声。

3）高质量人脸生成方法。对于小尺寸的图像生成，GAN已经达到了很好的效果，但当用于高分辨率图像生成时，总是会伴随着训练不稳定、模式崩塌的问题，因此很多研究者对模型架构进行改进以稳定训练。Curtó等人（2020）基于DCGAN，对其激活函数进行改进，使用BatchNorm和SELU（scaled exponential linear unit）代替ReLU（rectified linear unit）层，可以稳定高分辨率图像的训练，在CelebA数据集上达到了很好的生成效果，但在其构建的数据集Curto上效果较差。这是由于CelebA数据集几乎为白人，分布比较简单，而Curto数据集包含了不同种族的人，分布较为复杂，因此模型很难学习到它的分布。针对这一点的理论解释是：当真实分布与虚假分布没有足够的重叠时，从判别器传递到生成器的梯度变得没有信息。因此Karnewar和Wang（2020）提出多尺度GAN，允许梯度在多个尺度上从判别器传递到生成器，该方法可以在多个数据集上都达到较好的效果。类似地，Karras等人（2019）提出渐进增长GAN，在训练时，逐渐向模型添加层，从而提高生成图像的分辨率，这种增量性质允许训练首先发现图像分布的大规模结构，然后将注意力转移到越来越精细的尺度细节上，而不必同时学习所有尺度，该方法最高可生成1 024×1 024像素分辨率的图像。

2.4 群体分析

随着人工智能技术的发展，群体分析成为一个重要的研究方向。它涉及到人员计数和人群行为分析等方面，这些技术在视频监控、公共安全等领域有着广泛的应用。然而，由于真实数据的获取存在隐私、成本和场景复杂性等问题，合成数据集的创建和使用成为解决这些问题的关键。本小节将综述合成数据在群体分析中的应用，探讨其如何帮助研究人员和开发者克服真实数据采集的挑战，并提升模型的性能。

2.4.1 人员计数

由于人群场景的广泛应用（如视频监控、公共安全），人群计数成为热门研究话题。但人员计数是一项艰巨的任务：环境多变且人数范围大，导致现有方法无法很好地发挥作用。此外，由于数据稀缺，许多方法都不同程度地存在过拟合。为了解决上述问题，许多方法使用数据生成技术合成人群场景的图像和视频。

Ekbatani等人（2017）使用传统数据生成方法，构建了100万幅158×158像素的合成图像，每幅图像包含多达29个行人。他们首先通过从每个视频帧中减去背景，并提取视频的中值背景。随后从中值背景中减去每幅图像，从中提取行人。通过随机更改图像的全局照明和向背景添加一些随机高斯噪声生成背景。然后使用感兴趣区域（region of interest，ROI）过滤器对图像进行遮罩，最后通过将行人添加到遮罩的背景中创建合成图像。此外，为了使图像尽可能逼真，Ekbatani等人（2017）手动删除了非行人的异常值，并通过图像编辑去除行人身边的光晕和调整行人与街道的透视关系。研究中设计的深度卷积神经网络模型在合成数据集上训练后，在真实人群计数数据集上测试，其性能与依赖于详细标记和专门特征工程的传统方法相当，证明了合成数据的有效性。

尽管传统数据生成方法能够低成本地增加样本多样性，但是存在一些局限性，例如产生的变化有限、难以充分模拟真实世界场景的复杂性和缺乏可扩展性。然而虚拟引擎能够提供精确且高度逼真的虚拟环境，例如模拟复杂的光照、阴影、反射和纹理效果。这些特性使得虚拟引擎在数据生成方面具有显著的优势。因此大多数工作使用虚拟引擎合成用于人员计数的数据。

GCC数据集(Wang等,2019a)是通过《侠盗猎车手Ⅴ》游戏引擎合成的人群场景数据。研究人员通过脚本控制虚拟世界中的行人和车辆,以及设置天气和时间等属性,构建了拥挤的人群场景。考虑到游戏引擎的限制,人数必须少于256人。对于拥挤的人群场景,研究者首先分割几个不重叠的区域,然后在每个区域放置人员,最后将多个场景整合为一个场景。GCC数据集包含15 212幅图像,分辨率为1 080×1 920像素,共计包含7 625 843人,平均每幅图像包含501人。数据集共有100个场景,每个场景有4个不同的视图。GCC数据在人群规模方面略有不平衡,显示密集人群的图像数量低于显示中等或稀疏人群的图像数量。为了验证合成数据对数据增强的效果,实验人员在GCC数据集上对人员检测模型进行预训练,然后使用真实数据集对预训练的模型进行微调。与从头训练的方法相比,使用合成数据能够可以减少约30%的估计误差。

Zhang等人(2021a)提出一种跨视图跨场景的多视图人群计数范式,然而现有的数据规模较小,难以训练CVCS模型。因此,研究者使用《侠盗猎车手Ⅴ》游戏引擎合成相应的数据,首先选择并设置与场景相关的特征,即位置、RoI、天气条件、人体模型和姿势等。然后,将摄像机放置在不同角度拍摄人群场景。共合成280 000幅分辨率为1 920×1 080像素的图像,这些图像来自31个位置,每个位置有60～120个不同的视图。每幅图像包含90～180名行人。相比之下CVCS的行人数量范围相对较小,缺乏拥挤的场景。为了验证合成数据的有效性,研究者在合成数据集上训练跨视角跨场景的模型,在具有不同摄像机视角(跨视图)的真实场景(跨场景设置)上进行测试模型。实验结果表明使用合成数据能够使模型的平均绝对误差减小2.08%,证明合成数据能够丰富训练场景,减小过拟合现象。

CrowdX(Hou等,2022)是使用Unity3D引擎构建的数据集。研究者使用3个包含建筑物、城市道路和交通信号灯等的城市场景和5个纯色背景构建数据。对于每个场景,相机俯仰角变化为30°、50°、70°和90°,行人数量从1到1 000随机抽样,步长为100。CrowdX包含24 000幅人群图像,分辨率为1 024×768像素。利用所合成的数据集CrowdX作为数据增强的实验结果表明,模型性能可提升8.4%。在CrowdX上预训练的另外两个人群检测的经典异构架构也表现出明显的性能提升。为了评估视频人群计数算法,CrowdX扩展为合成视频数据集,成为CrowdXV(Hou等,2023)。它包含10 000个视频片段,每个视频有5帧,帧速率为30帧/s,分辨率为1 024×768像素。每帧平均人数为250人。行人在场景中随机实例化,并以分别遵循高斯分布方向和均匀分布的速度移动。

相比传统数据生成的方法,使用虚拟引擎(如游戏引擎、Unity3D引擎)的方法能够更灵活、更低成本地创建各种复杂场景,自动生成的数据能够具有精确的标注,人员的密集程度上也更加拥挤。但这类方法也存在局限性,例如虚拟引擎无法完全复制现实世界的复杂性,合成数据和真实数据之间可能存在一定的域差异,并且行人的行为模式也可能过于简化。

2.4.2　人群行为分析

人群行为分析主要在人群检测的基础上研究不同行为模式,例如异常行为检测、人流方向预测等,在公共安全领域具有广泛的应用。近年来,基于深度学习的方法在这一领域取得了显著进展。然而,构建深度学习模型所需的数据集往往依赖于大量人工工作,而人群行为分析的特性导致数据采集面临诸多挑战,如隐私问题和复杂场景下的数据稀缺性,这在一定程度上限制了模型的发展与推广。

为了解决这一问题,部分研究者开始尝试利用合成数据来替代真实数据。一类常见的方法是使用程序化的合成数据方式,结合商业软件或游戏引擎对场景和人物进行渲染,生成具有高自由度和大规模的数据集。相比真实数据集,合成数据集减少了人工成本,还能更灵活地控制各种场景和行为。

Aranjuelo等人(2021)通过利用合成数据训练的监控系统实现人群检测。他们使用3D建模软件3ds Max和渲染引擎V-Ray生成合成图像,并使用插件3ds Max Populate创建人物角色。流程包括场景生成、光照和材质模拟、3D资产和人物角色的添加、摄像头位置模拟,以及图像畸变矫正算法。合成训练数据集包含28 000幅图像,细分为6组,其特点是难度级别不断增加。图像是从不同的照片般逼真的场景中获得的,这些场景具有不同的位置、背景、失真、照明条件、行人外观和位置、物体和相机位置。结合真实数据和合成数据的训练方法能够使检测模型的平均精度从70%提升到82%。

Di Benedetto等人(2021)通过合成数据进行人群的安全设备检测。他们使用《侠盗猎车手V》游戏引擎中在游戏地图的不同位置部署一系列带和不带安全设备的行人。在游戏地图的各个位置添加带有所选装备的行人,将相机放置要拍照的地方,检查物体是否在视野范围内且没有被遮挡,从渲染引擎恢复3D网格边界框,并保存游戏截图(即数据集图像)及其相关注释(边界框和类别)。数据集包含180幅图像,个人安全设备包括高能见度背心、头盔和焊接面罩等。除了佩戴这些类型设备的人之外,还包括没有防护的行人,在其中注释人、裸露的头部、裸露的胸部。实验结果表明在所生成的图像上进行训练,并使用有限数量的真实图像执行域自适应步骤是有效的。此外,所生成的数据带来两个主要积极影响:一是由于数据引擎提供的准确注释,对边界框预测更紧密;二是对高变异性类别(例如头部、胸部)、遮挡物体和极端情况(例如蹲伏的人)的改进,网络可以利用引擎提供的所有变异性为其提供更好的高级表示。

Courty等人(2014)对合成数据在群体行为分析中的应用进行了初步探索。他们基于Helbing等人(2020)提出的社会力模型,将行人视为受力的粒子模拟运动轨迹,并考虑行人的意愿、疲劳、社会互动和障碍物的影响。为提高视觉逼真度,Nicolas Courty对社会力模型的参数进行了优化,并借助商业软件Maya和Mental Ray进行渲染,构建了AGORASSET数据集。该数据集由一系列分辨率为640×480像素、帧率为30 Hz的视频组成,涵盖了多种人群拓扑场景。数据集提供了丰富的标注信息,包括行人的运动参数、密度、流量以及分割掩码。研究还表明,由于合成数据不受摄像机角度的限制,其生成的多视角数据比真实数据更有助于提升模型的预测性能。

另一项研究由Cheung等人(2019)提出,他们开发了一个名为LCrowdV的程序化框架,用于生成大量带有标签的人群视频。该框架包括两个模块:一个用于模拟人群运动和行为,另一个用于渲染多样化的视频和图像。LCrowdV能够根据环境、行人数量、密度、行为、流动、照明条件、视角以及噪声等因素,自动生成任意数量的带标注数据。通过使用该框架生成的10 000幅图像进行实验,研究者发现,结合LCrowdV和真实数据集训练的检测器平均精度提升了3%。此外,他们通过逐步去除某一参数的变化,分析了不同因素对检测器性能的影响,结果表明相机角度的变化对平均精度影响最大,达到了36%。这进一步证明了合成数据在多视角应用场景中的优越性,而传统的真实数据集往往难以包含丰富的相机角度变化。

PHAV(procedural human action videos)(de Souza等,2017)是一个大型合成视频数据集,旨在用于动作识别,基于程序化生成模型构建。该数据集包含39 982个视频,涵盖35个动作类别,每个类别都有超过1 000个示例。其中,21个类别基于现有的运动捕捉数据,而另外14个类别则完全通过程序化定义的合成动作生成。de Souza等人(2017)利用程序化生成技术创建了逼真的视频场景,通过定义参数和规则生成不同的动作和背景,并使用Unity® Pro游戏引擎合成了该数据集。de Souza等人(2017)还将真实数据集与PHAV数据集按不同比例混合,实验结果表明,即便仅使用50%的真实数据,模型性能的损失也非常小,这进一步验证了合成数据集的有效性。

除了程序化设计与渲染引擎的结合,还有许多研究选择依托视频游戏《侠盗猎车手V》进行数据收集。这款游戏能够轻松实现现实世界中难以获取的异常行为样本,例如恐慌、斗殴和逃离等,这很好地弥补了该领域数据的不足。

Montulet和Briassouli(2021)制作了用于无监督异常人群行为检测的数据集,采用《侠盗猎车手V》中的Rockstar高级游戏引擎(rockstar advanced game engine,RAGE)创建了一个密集注释且极为逼真的数据集。该引擎提供了极大的灵活性,可以生成个人和人群在不同室内和室外环境、照明及天气条件下的广泛且真实的活动视频。该数据集包含来自54个不同位置的视频,分辨率为2 560×1 440像素,每个视频有450帧,由静态摄像机在不同高度录制。同时,为每一帧提供详细的真实数据,包括天气条件、一天中的时间、人物分割、骨骼坐标、深度图和人群行为类型。与其他工作不同的是,该数据集中的视频以不同的FPS渲染,以模拟常见安全摄像机上的不同帧速率。研究者采用累积和(cumulative sum control chart,CUSUM)方法分析人群光流的统计分布随时间的变化,通过统计分布的异常变化无监督地检测异常事件,并以具体示例证明了该方法与

数据集的有效性。

人群异常行为分析对公共安全和监控等领域至关重要。然而，现有数据集在密集人群场景下的异常行为样本稀缺，且收集此类数据面临隐私和法律挑战。为此，Lazaridis 等人（2018）利用《侠盗猎车手V》游戏引擎创建一个详细的合成数据集，以模拟和分析人群的异常行为。该数据集包括 14 个分辨率为 1 920 × 1 080 像素、帧率为 60 帧/s 的视频，每个视频时长 3～5 min，并附有真实标签。其中 7 个视频展示正常的人群活动，其余则包含战斗和恐慌等异常行为的片段。Lazaridis 等人（2018）将该数据集与真实数据集的训练结果进行了比较，观察到明显的性能提升。正常行为与战斗行为的检测准确率提高约 5 %，而恐慌行为的准确率提高了 12. 2%，这充分说明了该数据集的实用性。

Lin 等人（2021a）也利用《侠盗猎车手Ⅴ》的游戏引擎模拟并记录现实世界中难以捕获的异常事件，制作了用于人群异常行为分析的另一个数据集 SHADE。通过游戏内的脚本原生函数设计和录制各种异常事件场景，自动生成了包含多种异常行为的视频数据集。具体而言，Lin 等人（2021a）首先从背景数据库中选择一个事件场景，然后添加随机的天气、时间、人物与事件，最后使用录制工具获取不同事件对应的视频。SHADE 数据集包含 2 149 个视频剪辑，总共超过 870 000 帧，每个视频时长范围为 2～34 s，帧率为 60 帧/s。不同类型的动作视频长度各异，例如，击倒场景的视频通常持续 2～3 s，而追逐场景的视频平均长度约为 13 s。考虑到现实世界中黑暗环境下异常事件发生的概率大幅提升，Lin 等人（2021a）据此调整了数据集的相关概率分布。

除了通过游戏引擎模拟场景外，还有方法选择从真实数据中自动生成标注数据。Noghre 等人（2022）提出一种自动生成数据集的方法 ADG-Pose，专门用于真实世界中的人体姿态估计。该方法使用自上而下的人体姿态估计模型从未标记的数据中提取人体关键点。用户还可以自定义数据集，以适应不同的人物距离、拥挤程度和遮挡分布。Neghre 等人（2022）利用该方法创建了名为 Panda-Pose 的定制数据集，专门针对停车场监控场景。使用 HRNet-w32 模型进行自上而下的标注，并选择具有极高分辨率和广阔视野的 PANDA 数据集作为基础。实验表明，使用 Panda-Pose 数据集训练的模型在真实世界的骨架基础动作识别上的端到端准确率更高，尤其是中等距离和遮挡场景中。同时，在远距离场景中，使用 Panda-Pose 训练的模型的准确率是在 COCO 数据集上训练的模型的 4 倍。

许多人群行为分析方法基于光流实现，光流算法的性能取决于内容的特定性和应用场景。尽管人群行为分析在实际应用中大量使用光流，但现有光流数据集并未涵盖与人群行为分析密切相关的内容。为此，Schröder 等人（2018）推出了专注于人群行为分析中的光流信息的 CrowdFlow 数据集，旨在缓解这一矛盾。具体而言，利用 Unreal Engine 游戏引擎创建虚拟城市环境，模拟数千个同时移动的个体，从而生成具有真实感的合成视频序列，并获取精确的光流场和个体轨迹。CrowdFlow 数据集包含 10 个高清分辨率（1 280 × 720 像素）的序列，帧率为 25 Hz，每个序列包含 300～450 帧，并提供精确的光流场和个体轨迹作为真实值。这些序列模拟了 371～1 451 个独立移动个体，覆盖了静态和动态摄像机视角，以模拟传统监控和无人机监控场景。在 CrowdFlow 数据集上的实验结果表明，一些在现有基准测试中排名较高的算法在处理复杂人群场景时的性能并未如预期般优秀，而一些在现有基准测试中表现一般的算法却在 CrowdFlow 数据集上展现出较好的性能。这种排名的变化揭示了现有数据集可能无法充分评估算法在复杂人群场景下的实际表现。

总的来看，随着深度学习在公共安全领域的应用不断扩展，生成数据的使用为人群行为分析提供了新的可能性。这些方法通过利用程序化设计、渲染引擎和游戏引擎等技术，不仅克服了真实数据集在采集成本、隐私保护和场景复杂度等方面的限制，还能提供更丰富的多视角数据和异常行为样本，显著提升了模型的性能和泛化能力。未来，生成数据将继续作为人群行为分析中的重要补充，推动该领域的进一步发展与创新。

2. 5　自动驾驶

数据生成在自动驾驶技术中扮演着越来越重要的角色，不仅能够生成丰富的训练样本和多样化场景，从而增强模型的性能，还能生成极端场景和对抗场景数据对系统的安全性进行边界测试。本节从多样化训练数据生成和安全性测试两方面深入探讨数据生成在自动驾驶中的作用。

2.5.1　多样化数据生成

自动驾驶系统往往需要大量路测，以应对复杂的道路、天气和交通情况。然而，实地测试不仅耗时、成本高，还受法规限制。基于此，研究人员开始探索不同的数据生成技术，以模拟真实世界的复杂性、减少对实测的依赖。

1）规则化数据生成。数据生成提供一种高效的替代方案。传统的数据增强技术，如翻转与裁剪，常用于扩充数据集、增强训练多样性以及提高泛化能力。然而，自动驾驶数据集（如nuScenes（Caesar等，2020））因固定的摄像头视角和位置，使得这些技术的应用受限，例如产生不符合实际情况的数据。

相比之下，基于物理引擎的仿真技术如CARLA（Dosovitskiy等，2017）和AirSim（Shah等，2018）等则能灵活模拟复杂环境，降低测试成本，提供可重复的测试环境。尽管如此，仿真数据与真实世界在细节上仍有差距。

2）基于生成模型的数据扩展。生成模型如VAE（Kingma和Welling，2013）和GAN（Goodfellow等，2014）通过预设条件生成新数据场景，弥补了传统数据增强的局限。例如，Yurtsever等人（2022）的方法通过CGAN生成高保真的RGB图像。Swerdlow等人（2024）提出的BEVGen则基于鸟瞰图布局和给定的初始多视图像，生成全新街景图像。此外，TrafficGen（Feng等，2023）作为一种专注于交通场景生成的方法，并不生成感知模块所需的感知图像，而是在给定高清地图及初始条件后，生成新的车辆位置和运动轨迹。

3）复杂驾驶场景构建。随着生成技术的发展，结合大语言模型（large language model，LLM）和扩散模型（diffusion models）（Ho等，2020）的世界模型为复杂驾驶场景生成提供了更高的灵活性。世界模型能基于自然语言等高层输入自主生成完整的场景，提升训练效率和适应性。例如，Drive-WM（Wang等，2024f）能生成多视角画面并保证视角一致性。DriveDreamer-2（Zhao等，2024）则通过大语言模型解析文本提示生成车辆轨迹，再以扩散模型生成多视角驾驶视频。然而，世界模型在生成多模态数据时仍面临一致性挑战。如果各模态间生成数据逻辑不一致，可能影响自动驾驶系统对场景的理解与决策。

2.5.2　基于生成数据的安全性测试

自动驾驶系统不仅需要应对常规驾驶环境，还需要具备抵御对抗攻击和处理极端场景的能力。生成模型在安全性测试中发挥重要作用，有助于识别模型弱点并提升系统鲁棒性。

1）对抗样本生成。对抗样本是那些含有微小扰动的能让深度学习模型做出错误的判断的输入样本，因此能够揭示模型的安全边界，对于测试模型的鲁棒性至关重要。在自动驾驶领域，对抗样本被广泛应用，以帮助识别系统的潜在弱点。例如，LanEvil基准测试（Zhang等，2024a）通过设计多种环境幻觉类型，覆盖了真实世界中影响车道检测的各种因素，为车道检测模型提供一种新的安全性测试方法。Xiao等人（2018）提出的AdvGAN通过使用生成器产生扰动，再将这些扰动叠加到原始样本上，从而生成对抗样本。整个过程通过有效利用GAN的生成能力，大幅提升对抗样本生成的效率和攻击的成功率。

2）极端场景生成。自动驾驶极端场景的数据往往难以获取，且采集过程中可能对采集者和设备造成不可逆的损伤，这些场景涵盖如碰撞、“鬼探头”等复杂且危险的情况，而这些恰恰又是模型训练中最为欠缺的部分，模型往往在此类场景中表现不佳。Rempe等人（2022）提出的STRIVE方法通过一个基于图的CVAE学习复杂的交通运动，并在隐空间中优化场景，从而生成颇具挑战性的碰撞场景。这一方法不依赖于特定的对抗车辆，而是优化所有车辆的轨迹，以生成多样化且贴近现实的碰撞场景。

2.5.3　小结

数据生成技术在自动驾驶领域具有重要作用，能够显著提升模型性能、系统鲁棒性和安全性，并通过严格的边界测试确保系统可靠性。然而，当前技术仍面临真实性和精确性挑战，难以精确复制车辆动力学、传感器噪声等微观细节，导致生成数据与真实场景存在偏差。通过融合大语言模型和扩散模型，世界模型能够基于高层次输入自主构建复杂场景，从而提高数据生成的灵活性、训练效率和适应性。随着技术进步，数据生成技术将为自动驾驶研究和应用提供更全面的支持，推动系统在多样化、复杂化场景中实现安全、可靠的运行，并助力智能化和自主性的大幅提升。

2.6　视频生成

2.6.1　基于扩散模型的视频生成

随着扩散模型和基于Transformer架构的视觉生

成模型的发展，其在图像生成领域已经取得显著成果并得到广泛应用，很多工作也将扩散模型应用于视频生成。视频生成模型旨在根据文本描述或可控条件自动生成相应的视频，具有广泛的应用，包括影视制作、广告营销、自媒体创作、游戏开发、教育培训、虚拟现实、具身智能、自动驾驶和世界模型等。早期的视频生成模型在基于U-Net架构的扩散模型基础上进行改进，尽管一些早期的研究已成功生成低分辨率且时长较短的视频，但是面对巨大的算力与数据成本，高质量且长序列视频生成的挑战仍然很少有人解决。

OpenAI在2024年初发布的视频生成模型Sora提出全新的高质量视频生成方法，将模型架构从U-Net转变为具有更强可扩展性和更大参数量的DiT（diffusion transformer）模型，扩展数据规模并细化训练策略，从而生成具有更高分辨率、更合理的运动、更精准的视觉语言对齐度以及更高可控性的长序列视频。后期很多工作基于DiT架构进行探索，提高数据质量和数量，扩大模型参数量，实现高质量的视频生成。下面整理了基于扩散模型的视频生成领域的研究现状，包括早期工作和2024年最新的DiT架构的视频生成模型，希望能够为视频生成领域的研究者提供参考与启发。

2.6.2　基于U-Net的视频生成模型

作为T2V视频扩散模型的早期工作，VDM（stable video diffusion）（Blattmann等，2023a）首次提出将图像扩散模型中的U-Net架构扩展到3D U-Net结构，引入空间3D卷积和时空分离注意力，使UNet可以应用在可变序列长度上，并对图像和视频进行联合训练。MagicVideo（Zhou等，2023a）最早使用LDM（Rombach等，2022）生成视频，通过在低维潜在空间中利用扩散模型降低计算复杂度，引入的逐帧轻量级adaptor对齐了图像和视频的分布，使所提出的定向注意力能够更好地建模时间关系以确保视频的时序一致性。LVDM（He等，2023b）也使用LDM作为骨干，利用分层框架建模视频潜在空间，使用掩码采样、条件潜在扰动和无条件引导等技术以生成更长的视频。与逐帧压缩视频的MagicVideo不同，LVDM沿着时间轴压缩冗余信息以获得更紧凑的潜在空间，之后的很多工作也将扩散模型应用于潜在空间以提升效率。ModelScope（Wang等，2023b）将时空卷积和注意力整合到LDM中，采用图像数据集LAION和视频数据集WebVid混合训练文生视频模型。此外，VideoFactory（Wang等，2024d）引入交换式时空交叉注意力机制促进时间和空间模块之间的交互，提出大规模高质量视频数据集HD-VG-130M并在其上进行训练，能够生成高分辨率视频。Latent-Shift（An等，2023a）在卷积模块中移动相邻帧之间的通道进行时间建模，在生成视频的同时保持了原始的T2I能力。

为了提高生成视频的分辨率和帧数，很多工作设计了多阶段生成方法，生成关键帧，然后使用时空超分辨率模块进行填补。Imagen Video（Ho等，2022a）使用级联视频扩散模型扩展了T2I模型用于视频生成，包括基础视频扩散模型、时间超分辨率扩散模型和空间超分辨率扩散模型，采用多阶段训练技术以生成高质量视频。Make-A-Video（Singer等，2022）在T2I扩散模型基础上采用2D+1D的时空卷积层和注意力层，从成对的图像—文本数据中学习视觉—文本相关性，并从无监督视频数据中捕获视频运动先验，通过时间和空间超分辨率模型以及插值网络，实现了更高的分辨率和帧率的视频生成。Video LDM（Blattmann等，2023b）和Lavie（Wang等，2023g）分3个阶段训练级联视频扩散模型，包括关键帧T2V生成、视频帧插值和空间超分辨率模块。Show-1（Zhang等，2025a）设计了具有4个阶段的框架，包括在低分辨率像素级上运行的关键帧生成、帧插值和超分辨率模块，以及潜在超分辨率模块，在增强视频分辨率的同时降低开销成本。Seine（Chen等，2023d）从短视频序列逐步扩展生成长视频，且能够实现两个不同场景之间的平滑过渡。然而，多阶段生成策略在保持全局时间一致性上仍然具有挑战性，特别是在快速运动的情况下可能导致时间混叠。此外，超分辨率模块可能会受到域间差异的影响，因为它们是在真实帧上训练的，但在推理过程中应用于生成帧。

AnimateDiff（Guo等，2024）将运动模块引入预训练的T2I模型中，以学习视频动态先验。SVD（stable video diffusion）（Blattmann等，2023a）、Emu-Video（Girdhar等，2024）、I2VGen-XL（Zhang等，2023d）、Dynamicrafter（Xing等，2025）、VideoGen（Li等，2023b）、VideoCrafter（Chen等，2023a）和Pixel-Dance（Zeng等，2024b）等工作都提出图像作为控制条件的T2V方案，通过额外的特征提取网络或条件

潜在变量连接将图像控制条件注入到T2V生成过程中，利用参考图像提高视觉保真度，使模型能够专注学习视频动态。其中SVD等工作验证了增大数据规模能够提高视频扩散模型的能力，它收集并标记了数亿个视频文本数据用于模型训练，开源模型已在许多后续工作中使用。此外，Lumiere（Bar-Tal等，2024）使用时空U-Net架构，通过模型的单次传递生成视频的整个时间持续期。还有很多无需训练的零样本视频生成方法，例如Text2Video-Zero（Khachatryan等，2023）、Tune-A-Video（Wu等，2023a）和DirecT2V（Hong等，2024）等，不依赖于大规模数据集，且能降低训练成本。

2.6.3　基于DiT的视频生成模型

基于Transformer的视频生成模型Sora展现了DiT架构的可扩展性和强大性能，Snap-Video（Menapace等，2024）、Latte（Ma等，2024）等很多工作也使用DiT架构而非U-Net作为模型主干生成视频。W.A.L.T（window attention latent Transformer）（Gupta等，2025）使用因果编码器将图像和视频压缩到统一的隐空间，用于基于DiT架构的扩散模型的联合训练。此外，模型还使用专为联合空间和时空生成建模而定制的窗口注意力架构，以提高记忆和训练效率。CMD（content-motion latent diffusion model）（Yu等，2024b）将内容和运动解耦，首先训练一个自动编码器将视频编码为内容帧和运动隐变量，然后使用DiT架构生成运动隐变量，从而以较低的计算成本生成高分辨率视频。Vchitect-2.0是一个2 B参数量的支持多分辨率24 帧/s的一体化视频超分插帧增强模型，能够生成5～20 s的视频，同时还提出首个兼容长视频的视频生成评测框架VBench（Huang等，2024b）。

NUS最新的Open-Sora v1.2（Zheng等，2024b）在30 M个视频数据上训练了一个1.1 B参数的视频生成模型，支持0～16 s、144 p到720 p、各种宽高比的视频生成。其关键改进包括：1）采用3D视频压缩网络：首先在空间维度上将视频分辨率压缩至原始尺寸的1/64，然后在时间维度上将视频时序信息压缩至原始长度的1/4；2）采用基于SD3（Esser等，2024）的修正流模型代替去噪扩散概率模型（denoising diffusion probabilistic models，DDPM），提出多个模型适应策略在小数据集上微调高分辨率T2I模型以适应视频生成设置；3）采用更多的数据、更好的注释和3阶段多尺度训练。

北京大学提出的Open-Sora-Plan（Lin等，2024a）实现了一个类Sora的大型视频生成模型，能够根据用户的多种输入生成高分辨率且时长较长的视频，是第1个将因果视频VAE和全3D注意力机制应用到DiT架构的开源视频模型。其最新的v1.3版本利用全3D注意力架构，增强了对联合时空特征的捕获。关键改进包括：1）采用更高效、更稳健的基于小波变换的流变分自编码器WFVAE：使用小波变换分解视频，捕获不同频域的信息，学习更好的压缩视觉表示；2）更好的视频生成架构：使用具有全3D注意力架构的DiT扩散模型，联合图像—视频进行训练，同时采用具有跳跃稀疏注意力机制，在保证足够性能的同时加速训练；3）提出多维数据处理流程：包括优化视频文本注释，采用高质量数据清洗策略等，以确保能够获得所需的高质量数据；4）设计多种高效的训练与推理辅助策略，提出多种条件控制器，在定性和定量评估中都取得了显著的视频生成结果。

Allegro（Zhou等，2024e）基于Open-Sora-Plan v1.2构建并训练了2.8 B参数的DiT架构的文生视频模型，能够生成长达6 s、帧率15 帧/s、720 p分辨率的高质量视频。关键改进包括：1）设计处理大规模视频数据的系统用于进行数据处理和过滤；2）设计一个VideoVAE用于将原始视频编码到时空潜在空间中，VideoVAE建立在预训练的图像VAE之上，并通过时空建模层进行扩展，以有效利用空间压缩功能。3）在DiT架构基础上，引入3D RoPE位置嵌入和3D全注意力机制，以有效捕捉视频数据中的空间和时间关系。

EasyAnimate（Xu等，2024a）最新的v5版本将模型参数规模扩展至12 B，并引入MMDiT架构以提升生成性能。该版本支持多种控制条件输入，可生成分辨率为1 024×1 024像素、共49帧、时长为6 s、帧率达8 帧/s的视频内容。EasyAnimate v5的输入可以是文本、图像和视频，使用奖励反向传播来训练Lora并优化视频，使其更好地符合人类偏好，同时引入了混合运动模块，以保证具有帧间一致性的运动生成和过渡。

CogVideoX（Yang等，2025）支持文生视频、视频续写和图生视频，以720×480像素的分辨率生成48帧8 帧/s的视频，v1.0模型开源了2 B和5 B两个版本，之后更新的v1.5版本5 B系列模型支持10 s长度的

视频和更高的分辨率。CogVideoX设计了大规模DiT架构模型从文本提示生成视频，包括一个3D因果VAE和一个配备自适应LayerNorm的专家Transformer，采用增强文本—视频对齐、显式均匀采样和渐进式训练等策略，以产生具有动态运动的连贯长视频。此外，还引入新的文本视频数据处理流程以提高视频注释、视频质量和语义对齐度。

Mochi-1开源了10 B参数的DiT架构模型，其VAE将视频数据压缩至原始体积的1/96，并提出非对称MMDiT（AsymmDiT），使用多模态自注意力机制共同关注文本和视觉标记，并为每种模态学习单独的多层感知机（multilayer perception，MLP）层。但由于隐藏维度较大，视觉流的参数数量几乎是文本流的4倍。为了统一自注意力机制中的模态，AsymmDiT使用非方形QKV和输出投影层的非对称设计，通过简化文本处理并将神经网络集中在视觉推理上，以高效处理用户提示和压缩视频标记，减少推理内存需求。

最近，与传统的扩散模型相比，基于流匹配的流模型表现出卓越的性能和更快的生成速度，对噪声调度选择的鲁棒性也有所增加。Pyramid Flow（Jin等，2025b）采用自回归生成，基于先前帧预测生成后续帧，确保视频内容的连贯性和流畅性。通过引入统一的金字塔流匹配算法，为每个金字塔分辨率设计了一个分段流，它们通过单个DiT中的统一流匹配目标进行联合优化，允许同时生成和解压缩视觉内容，从而实现更高效的视频生成。

Meta发布的具有30 B参数量的视频生成模型Movie Gen（Polyak等，2025）使用流匹配在最大上下文长度为73 K的视频标记上进行训练，Movie Gen遵循多阶段训练，包括联合图像—视频训练，然后用一组精选的高质量文本—视频对进行微调。主要技术包括：时间自编码器（temporal autoencoder，TAE）的设计与优化、基于流匹配的训练目标、联合生成的骨干网络架构、文本嵌入和视觉—文本生成方法、空间上采样技术、模型扩展和训练效率优化等。此外，Movie Gen还在技术报告中详细介绍了其预训练数据的准备过程，包括视觉筛选、运动筛选、内容筛选和字幕生成等步骤，但是目前没有开源代码。

腾讯的HunyuanVideo（Kong等，2025）是目前开源的具有最大参数量的视频生成模型，其DiT参数量达到13 B，架构与之前的主流类Sora模型相似。主要设计包括：1）采用基于因果卷积的3D VAE对视频的时间和空间特征进行压缩，以视频和图像按照4∶1的比例混合渐进训练3D VAE，训练采用L1重建损失、KL损失、感知损失以及对抗损失。2）DiT采用“双流到单流”的混合模型设计，以捕捉视觉和语义信息之间的复杂交互，增强整体模型性能。在双流阶段，视频和文本token通过并行的Transformer块独立处理，使得每个模态可以学习适合自己的调制机制而不会相互干扰；在单流阶段，将视频和文本token连接起来，并将它们输入到后续的Transformer块中进行有效的多模态信息融合。3）使用了一个预训练的多模态大语言模型（multimodal large language model，MLLM）作为文本编码器，在特征空间中具有更好的图像—文本对齐能力，在图像的细节描述和复杂推理方面表现出更强的能力。4）分别以正常模式和导演模式对提示进行改写，将用户输入的提示词改写为更适合模型偏好的写法。

2.6.4　视频生成模型的应用

很多工作在文生视频或图生视频的基模型基础上进行改进和微调，解决了多种视频生成的下游应用任务，包括可控视频生成、定制化视频生成、长时长视频生成以及给定首尾帧的视频生成等。对于可控的视频生成，输入的控制条件包括人体姿态或表情、物体运动轨迹、相机轨迹、深度、音频、视频和图像等；对于定制化视频生成，包括单主体一致的生成、多主体一致的生成、风格一致的生成和视频生视频等任务。此外，很多视频生成模型用于辅助3D物体或场景的生成，包括SV3D（Voleti等，2025）、ViewCrafter（Yu等，2024c）和DimensionX（Sun等，2024c）等工作微调视频生成模型，以利用其3D先验生成新视角视频。对比基于不同基模型改进的工作可以发现，视频基础模型的能力对下游应用的效果具有显著影响，目前尚不清楚基础模型的改进是否会解决许多现有的研究挑战，以及随着这些模型继续扩大规模，是否会出现新的挑战。

由于视频数据的数量和质量、模型的容量以及预训练和优化策略的改进，Sora等基于的DiT架构视频生成模型具有很多优点：1）具有更强的长序列建模能力，能够更好地捕捉视频中复杂的时空关系，生成更连贯和更自然的运动，保持较长视频序列的时间一致性；2）具有更高的视觉质量和运动质量，基于DiT的模型随着扩大数据量和模型大小的增加，

能更准确地捕捉复杂的细节和整体结构；3)能够处理来自不同模态的信息，并从文本或图像中捕获局部到全局的语义信息，有效地处理复杂的上下文关系和长序列描述。

视频生成模型对算力和数据的要求都比较高，很多企业训练的视频生成模型都没有开源，且在近一年的时间内发展迅速。2023年，Runway和Pika发布其文生视频模型早期版本产品，并在2024年推出更新版本Gen-3和Pika-1.5。2024年初，OpenAI发布功能强大的Sora，随后快手的可灵、Luma的Dream Machine、生数的Vidu（Bao等，2024）、智谱的清影、字节的即梦（PixelDance和Seaweed）、Minimax的海螺AI、阿里的通义万相、Adobe的Firefly、Meta的Movie Gen、爱诗科技的PixVerse等众多商业版视频生成模型相继发布。

随着商业版视频生成模型的发展，生成视频的画质及稳定性、运动幅度及合理性、多样性和泛化性、一致性和可控性都不断提高，同时也扩充了更多的功能，例如：可灵的运动笔刷、Runway的镜头控制和视频生视频、Vidu的多主体一致性和画质增强、MovieGen的视频编辑和个性化生成、通义万相和MovieGen的同步视频的音频生成、Pika的特效I2V等。

在VideoGen-Eval（Zeng等，2024a）的最新评测中，闭源模型始终表现出比开源模型更高的视觉和运动质量，并超越了以前的基于U-Net的模型，能够处理自然的动态运动，丰富的多镜头场景和情感表达，以及电影级质感的场景模拟。其中，Gen-3、Kling v1.5、Movie Gen和Minimax等模型在T2V任务上表现出卓越的性能：Minimax在文本控制方面表现出色，尤其是在描述人类表达、相机运动、多镜头生成和主题动态方面；Gen-3在控制照明、纹理和电影技术方面脱颖而出；Kling v1.5在视觉、可控性和运动能力之间显示出良好的权衡。由于数据分布、模型大小和训练策略的不同，每个模型都具有各自的运动表示特征：Luma具有更广泛的相机运动，但主体运动受限；Vidu具有更大幅度和更快速度的主体运动；清影在文本对齐生成方面更加突出。基于强大的T2V基础模型，闭源I2V模型可以根据给定图像生成更合理、更强时间一致性的运动。Kling和Gen-3能够生成高质量的角色动画，还可以进行图像到视频的修复、外涂、插值和超分辨率等通用增强任务；Vidu和Luma分别表现出高度动态的主体和相机运动。

虽然闭源商业模型显著提高了整体质量，但它们在许多方面仍存在局限性。对于T2V生成任务，在沿空间和时间维度的文本对齐生成不佳、局部细节差且分辨率低（例如较小的人脸）、动态运动、推理能力、长序列的ID一致性（10 s及以上的持续时间）、多镜头场景、组合时空关系、复杂的物理交互以及遵循物理规则、多语言文本生成和稳定性等方面仍存在缺陷。对于I2V任务，存在难以理解输入图像的细节和语义信息、不能准确地刻画现有对象的运动而具有引入新对象的趋势、高动态运动时难以保持物体和时间的一致性等问题。这些I2V模型在场景、单物体运动、重照明、创造性场景和微动画中表现良好，在涉及人类角色动画、复杂物理运动、多物体运动、上下文计数或逻辑变化、多镜头变化中保持一致性和自然转换等应用中仍然面临巨大挑战。

2.6.5　小结

在工业界，很多商业版视频生成模型已经应用于影视制作和广告设计，通过AI降低视频制作成本，生成更多创意视频，将视频创作带入新的时代。在学术界，更多的视频生成模型和高质量数据陆续开源，用于学术研究探索和下游任务创新。开源模型和商业版闭源模型之间仍然存在显著性能差距，尤其是在大规模计算资源的模型训练，以及广泛数据集的收集和注释方面，工业界相对学术界具有更大的优势。尽管如此，目前的视频生成模型仍面临很多挑战和不足，诸如人物或运动的生成不够稳定、难以保证很好的一致性以及生成视频画质较低等问题。下面是对视频生成相关技术的一些未来展望：

1)降低视频生成的训练和推理成本，提高生成速度，为交互式和实时视频生成设计新架构，创建更高质量更广泛的数据集，探索更合理的评价指标和基准；

2)统一多模态的生成和理解，不仅限于生成视频内容，还包括理解复杂场景语义和物理规律，打造世界模型和通用人工智能（artificial general intelligence，AGI）；

3)探索更多视频生成的下游应用，包括更加多样化的可控视频生成（镜头控制、动作控制、运动笔刷、光照控制、材质控制等）、更加准确的定制化视频生成（风格定制化、主体一致性）、更长时长的视频生成、给定首尾帧的视频生成、视频延长和续写、视频

画幅扩展和自由长宽比控制、具有ID一致性的长序列多镜头视频生成、具有复杂语义的视频生成、结合三维表示的4D生成等任务；

4）提高生成视频的稳定性和画质，生成大幅度且合理运动的视频，包括提升基模型的能力、多阶段视频画质增强以及局部视频编辑等；

5）结合实际影视制作的工作流程，探索更丰富的功能，包括AI生成场景音效和配乐、根据人物配音驱动角色、分镜转场和智能剪辑、多图层的视频生成和交互式编辑等。

2.7 具身智能

在具身智能的研究中，数据生成技术起着至关重要的作用。传统的数据收集方法依赖于真实世界的交互数据，这种方式不仅成本高昂，而且效率低下。为了解决这一问题，模拟环境成为一种理想的替代方案，能够通过自动化的方式生成大量的视觉交互数据，极大地节省了时间和资源。通过高保真模拟平台，研究人员能够创建丰富的虚拟环境，在其中训练智能体完成任务，并进行复杂的物理交互。此外，近年来深度生成模型的发展使得视觉语言生成模型被引入到机器人策略学习中，从而生成更多样化的演示数据，以提升机器人在多样化场景下的表现。然而，尽管模拟环境提供了大量高质量的数据，这些数据的质量仍然受到环境设计、物理引擎和物体类型等因素的影响。为了使得从模拟到现实的知识转移更加顺利，研究者们也探索了多种仿真到现实的转移方法，这些方法通过多种技术手段缩小模拟与现实环境之间的差距，从而增强了具身智能体的泛化能力和适应性。

本小节将从两方面深入探讨数据生成在具身智能中的作用：一方面，分析模拟数据生成的技术和平台，及其如何为智能体的训练提供丰富的数据支持；另一方面，探讨仿真到现实的转移方法，重点讨论如何利用数据生成技术实现智能体在现实环境中的有效应用。

2.7.1 模拟数据生成

在具身智能领域，智能体需要大量的视觉交互数据来提高其智能化水平。之前大部分的数据收集方法直接在真实世界中收集演示数据并训练模型，但是这种收集方法通常需要大量的人力、物力资源和时间，会导致效率低下。因此，在大多数情况下，研究人员可以选择在模拟环境中收集数据集进行模型训练，而且在模拟环境中收集数据不需要大量资源，通常可以由程序自动化，节省了大量的时间。

DeepMind Lab（Beattie等，2016）是一个第一人称三维模拟平台，可用于研究自主人工智能体如何在大型、部分观察和视觉多样化的世界中学习复杂任务。AI2-THOR（Kolve等，2022）由近乎照片般逼真的3D室内场景组成，智能体可以在场景中学习如何导航并与对象交互以执行任务。SAPIEN（Xiang等，2020）模拟平台可以生成逼真的3D场景，并允许模拟复杂的物理交互，旨在为机器人和计算机视觉领域提供一个高度逼真的物理交互环境。VirtualHome（Puig等，2018）是一个为智能家居场景中的智能体学习和任务执行而开发的模拟平台，其可以创建丰富的视频数据集，从而能够训练和测试视频理解模型。VRKitchen（Gao等，2019）是一个虚拟现实系统，该系统不仅集成了具身智能模拟现实环境中执行各种复杂任务的功能，同时也允许人类作为教师进行演示以训练智能体。ThreeDWorld（Gan等，2021）是一个用于交互式多模态物理仿真平台，该平台支持在丰富的3D环境中模拟高保真传感数据以及移动代理和对象之间的物理交互，同时具备近乎现实照片级别的实时图像渲染功能。CHALET（Yan等，2019）是一款支持导航和操作的3D房屋模拟器，该模拟器支持一系列常见的家庭活动，包括移动物品、控制电器以及将物品放入可封闭的容器中，为具身智能生成了有效的模拟视觉交互数据。IGibson（Xia等，2020）用于训练具身智能在大规模现实场景中的交互式任务，同时提供了一个虚拟显示界面，将人类演示数据收集以强化具身智能的能力。Habitat-Sim（Savva等，2019）是一个专注于大规模环境的高性能三维仿真器，能够渲染逼真的室内环境，同时支持RGB相机和深度相机。

CLIPORT（Shridhar等，2022）和Transporter Networks（Shinn等，2023）是在Pybullet模拟器中收集的，用于端到端网络模型训练，并成功将这些模型从模拟环境转移到真实世界。GAPartNet（Geng等，2023）构建了一个大规模的基于部件的交互式数据集，提供了丰富的部件级别标注，用于感知和交互任务。他们提出一个领域泛化的3D部件分割和姿态估计的流程，能够很好地泛化到模拟器和真实世界中未见过的物体类别。SemGrasp（Li等，2024c）构建了一个大规模抓取文本对齐数据集CapGrasp，这是

一个从虚拟环境中提取的语义丰富的灵巧手抓取数据集。

视觉模拟数据的生成质量对具身智能的能力提升起到了决定性作用，不同的环境类别、物理学特性和物体类型都影响着生成数据的作用效果。构建具身智能模拟器环境的方法主要有两种：基于游戏的场景构建和基于世界的场景构建。基于游戏的场景构建通常具有内置的物理特性和物体类别，具有良好的分割性，如PartNet（Mo等，2019）中提供的三维模型。相比之下，基于真实世界的环境提供了更高的保真度和更准确的真实世界表现，有利于更好地将智能体性能从模拟转移到真实世界，如IGibson和Habitat-Sim模拟器由真实环境构建。视觉模拟数据不仅需要构建逼真的环境，还需要模拟真实世界物理特性的智能体与物体或物体与物体之间的逼真交互。大多数基于游戏场景的模拟器都内置了物理引擎，因此它们都具备基本的物理功能，如碰撞、刚体动力学和重力建模。不仅如此，对于像ThreeDWorld这样的模拟器，其目标是了解复杂的物理环境如何影响具身智能体在环境中的决策，因此配备了包括软体动力学在内的更高级的物理特征。这些物理学特征均对视觉数据产生了影响。具身智能观测到的物体类型主要分为场景物体驱动和数据集驱动两种。场景物体驱动的数据主要来源于模拟环境中的物体，收集成本低但质量难以保证；数据集驱动的数据主要来源于现有数据集，如SUNCG数据集（Song等，2017）、Matterport3D数据集（Chang等，2017）和Gibson数据集（Xia等，2018）。

深度生成模型的发展逐渐成熟，研究人员开始使用视觉语言生成模型为机器人的策略学习生成更多演示数据，从而提升机器人的表现。Gen2Sim（Katara等，2024）是一种通过使用图像扩散模型将二维物体图像扩展为三维资产，并自动生成任务描述、任务分解和奖励函数，利用大规模预训练生成模型来扩展机器人操控技能学习的模拟方法。在三维资产生成方面，该方法通过将3种类别的二维物体图像包括在机器人的环境中拍摄的真实的图像、由Google搜索在相关类别名称下提供的真实的图像和由预训练的文本条件扩散模型生成的图像映射到三维资产。具体而言，Gen2Sim利用扩散模型通过噪声预测网络生成图像，能够基于条件提示（如文本描述或相机姿态）生成特定视角下的三维对象新视图。在此基础上，Gen2Sim使用评分蒸馏采样（Poole等，2022）方法从图像生成可微分的三维模型，通过优化视角匹配度逐步更新三维表示。在生成三维模型后，Gen2Sim通过TEXTure（Richardson等，2023）方法增强纹理，并利用GPT-4查询物理属性来生成合理的尺寸和质量，丰富模型信息以用于模拟。RoboGen（Wang等，2024g）是一个利用大型语言模型和多模态生成技术，自动生成多样化任务、场景和训练监督，用于大规模机器人技能学习的系统。该系统的核心是自我引导的“提出—生成—学习”循环，通过生成多样化的任务、场景和训练监督，来大规模自动化具身智能的技能学习。具体而言，RoboGen使用Sentence-BERT（Reimers和Gurevych，2019）从Objaverse（Deitke等，2023）数据库中检索语言描述相似的对象，并通过视觉—语言模型验证检索到的资产。视觉—语言模型生成对象的图像描述，结合任务和目标资产描述，通过GPT-4进一步验证资产的适用性。如果检索到的资产不合适，系统则使用Midjourney网站进行文本生成图像，再使用Zero-1-to-3（Liu等，2023）进行图像到3D网格的生成。AGENTGEN（Hu等，2025）专注于生成环境和任务以增强基于大语言模型的智能体的规划能力。在环境生成方面，该框架在利用大语言模型生成随机环境的基础之上建立了一个灵感语料库，通过多样化的数据片段以生成复杂环境，从而提升具身智能在不同场景下的能力表现。

2.7.2　仿真到现实范式

具身智能在之前提到的模拟数据和环境中进行广泛学习之后，需要直接将学到的知识迁移到现实世界设置中，本节将介绍5种不同的仿真到现实转移范式。

1）Real2Sim2real（Torne等，2024）。通过在“数字孪生”模拟环境中利用强化学习（reinforcement learning，RL）增强模型在真实世界场景中的模仿学习。该方法首先在模拟环境中通过RL强调具体策略的学习，然后将这些学习到的策略转移到真实世界以解决数据稀缺问题，并实现有效的机器人操作模仿学习。该方法涉及在模拟中通过RL加强策略，然后将这些策略转移到真实世界以解决数据稀缺问题，并实现有效的机器人操作模仿学习。最初，这些方法通过使用NeRF和VR技术进行场景扫描和重建，并将构建的场景资产导入模拟器中，进而获得了

具有真实场景高保真度的模拟环境。随后，使用RL在模拟环境中微调模型从真实世界稀疏专家示范中获得的初始策略。最后，将经过优化的策略被转移到真实世界设置中。

2）TRANSIC（Jiang等，2024）。通过允许人类实时干预修正真实世界场景中机器人的行为，缩小了仿真到现实的差距。该方法通过几个步骤提升仿真到现实的性能转移：首先，机器人在模拟环境中使用RL进行训练，以建立基础策略。随后，这些策略在真实机器人上执行，当出现错误时，人类通过远程控制进行实时干预和纠正。干预过程中收集的数据用于训练残差策略。通过整合基础策略和残差策略，确保仿真到现实转移后的轨迹更加平滑。这种方法显著减少了对真实世界数据收集的需求，从而减轻了负担，并实现了有效的仿真到现实转移。

3）Domain Randomization（Tobin 等，2017；Matas等，2018；Andrychowicz等，2020）。通过引入参数随机化，增强了在模拟环境中训练的模型对真实世界场景的泛化能力，涵盖了在真实世界设置中可能发生的各种条件。尽管模拟和真实环境都依赖于通过摄像头捕获的视觉图像进行感知，但诸如物体摩擦和光泽等因素的差异使得精确模拟具有挑战性。通过在模拟训练期间随机化参数，可以覆盖广泛的条件变化，潜在地涵盖在真实世界环境中可能出现的不同情况。这种方法增强了训练模型的鲁棒性，使其能够顺利从模拟环境部署到真实世界中。

4）System Identification（Yu 等，2017；Kaspar 等，2020）。构建了真实世界物理场景的精确数学模型，涵盖了动态参数和视觉渲染等因素。其目标是使模拟环境与真实世界设置高度相似，从而促进在模拟中训练的模型顺利过渡到真实世界环境中。

5）Lang4sim2real（Yu 等，2024a）。使用自然语言作为桥梁，通过图像的文本描述作为跨域统一信号，解决仿真到现实的差距。该方法有助于学习领域不变的图像表示，从而提高在模拟和真实环境中的泛化性能。最初，在带有跨域语言描述的图像数据上预训练一个编码器。随后，基于领域不变的表示，训练了一个多领域、多任务的语言条件行为克隆策略。这种方法通过从丰富的模拟数据中获取额外信息，弥补了真实世界数据的稀缺性，从而增强了仿真到现实的转移效果。

总的来看，数据生成技术在具身智能领域中发挥了至关重要的作用。通过模拟环境的构建与虚拟数据的生成，研究者们能够有效克服现实世界数据收集中的时间和资源限制，提供丰富的视觉交互数据，进而提升智能体的学习效率和表现。同时，随着深度生成模型的进步，基于视觉语言的生成方法为机器人策略学习带来了新的可能性，进一步扩展了数据生成的应用范围。然而，仅仅依赖模拟数据并不足以完全解决具身智能的现实应用问题，仿真到现实的转移仍然是一个关键挑战。通过多种创新的转移方法，研究者们逐步缩小了模拟环境与现实世界之间的差距，增强了智能体的泛化能力和适应性。未来，随着数据生成技术和转移方法的不断完善，具身智能将在更多实际场景中展现出更强的能力，为机器人技术的广泛应用提供坚实的基础。

3 存在问题与发展趋势展望

数据生成为计算机视觉带来了新的机遇与挑战。前面两节中已经系统详细地介绍了典型的数据生成技术与模型，及其在典型计算机视觉任务中的应用。本节基于前文对国内外数据生成技术与应用的梳理，展望面向计算机视觉任务的数据生成技术应用的未来发展趋势。

1）生成数据的真实性有待进一步提升。尽管数据生成技术为计算机视觉任务提供了丰富的生成数据，但生成数据与真实数据之间的差距仍然存在。生成数据往往缺乏一定的复杂性和细节，特别是在图像的纹理细腻程度、光照变化的真实性、背景干扰和物理行为的真实感等方面，这些都可能影响模型在实际应用中的泛化能力。例如，在自动驾驶领域，生成数据虽然可以生成各种交通场景，但很难精确复现复杂的光影变化、天气状况和行人行为等细节，因此生成数据在某些高精度应用中的实用性受到限制。面对如何提升生成数据真实性的问题，未来的发展方向可以是通过改进生成模型，使得生成数据能够更好地捕捉到真实世界的特征，也可以通过引入更多的真实场景、增强模型的生成能力，以及结合多模态数据（如视频、音频、传感器数据等），提高生成数据的真实性。此外，探索结合真实数据和生成数据的混合训练方式，也可能帮助克服这一问题，使得生成数据能够更好地补充真实数据的不足，提升计算机视觉任务的效果。

2)生成数据的质量评估欠缺。在数据生成过程中,面临的另一个挑战是如何有效评估合成数据的质量,并反馈指导数据的生成。除真实性外,数据的多样性和代表性对于训练有效且实用的计算机视觉模型至关重要。生成数据的多样性不仅包括视觉内容的丰富性(如不同的颜色、形状、姿态等),还包括在环境、背景和光照等方面的多变性。如果生成数据缺乏这些多样性,模型训练可能出现过拟合或在真实环境下的泛化能力差。然而。当前的生成技术大多基于主观经验生成尽可能多的数据,往往没有一个统一的标准评价生成数据集的多样性和代表性。尤其在面临复杂的场景或动态变化时,生成数据的质量可能无法完美模拟现实世界的多样性。此外,还需评估生成数据是否能够覆盖目标计算机视觉任务中所有潜在的应用场景和挑战,尤其是在数据稀缺或长尾分布情况,例如自动驾驶场景下的特殊甚至极端情况。为了解决这些问题,需要探索新的数据质量评价指标和方法。传统的评价方法往往依赖于人为的视觉检查和经验性判断,但这种方法无法保证客观性和全面性。未来,基于自动化指标(如多样性度量、分布差异和对抗性评估等)的方法有望为数据生成的质量提供量化评估。通过引入评估标准和自动化工具,可以进一步提升合成数据在计算机视觉任务中的应用效果和可信度。

3)身份泄露与隐私保护问题。尽管生成数据在一定程度上减少了直接使用真实数据可能导致的身份泄露和隐私风险,但在某些情况下,生成数据仍然可能隐含真实数据中的个人身份信息。例如,没有参考图像和视频的数据生成技术难以保证生成数据的真实性,而为了保证生成数据的真实性,很多图像或视频生成技术需要原始图像或视频作为基础,通过生成不同场景、背景和光照等条件下的合成数据来扩展数据集或满足特定需求。这就不可避免地会存在真实数据所面临的身份泄露与隐私保护问题。这个问题在一些敏感领域(如医疗影像、人脸识别等)中尤其需要关注。为了解决这一问题,未来的数据生成技术应加强对隐私保护的关注,尤其是开发隐私保护技术(如差分隐私机制),以确保生成的数据无法反向推导出个人的敏感信息。此外,开展更加全面的实验来评估合成数据的隐私风险,提出一套标准化的隐私保护方案,能够为数据生成的安全性提供保障,推动其在更广泛领域的应用。

4)生成数据集公开与生态建设。目前,生成数据集的共享和公开仍然是一个亟待解决的问题。与真实数据集大都公开共享不同,目前大多生成数据集都未公开和共享。研究者在构建生成数据集时难以进行公平的对比实验,导致不同数据生成技术在效果上的优劣难以比较。由于合成数据集往往由不同的实验室或团队单独构建,缺乏统一的标准和规范,使得在实际应用中很难对比其优缺点。这种缺乏统一标准的局面,不仅阻碍了技术的快速发展,也增加了实际应用中的不确定性。为了解决这一问题,未来的研究应更加注重合成数据集的开放和共享,推动数据生成技术的标准化建设。通过建立开放的生成数据集平台,进行透明、公开和公正的对比实验,能够帮助社区成员进行有效的交流与合作。此外,数据生成技术的社区建设也应加强,推动学术界与工业界的合作,使得合成数据集和技术可以快速应用于实际项目中,促进人工智能技术的发展和普及。

5)进一步扩大和深化应用领域。生成数据在计算机视觉中虽然已经取得了显著进展,但仍面临许多亟待解决的问题和挑战。首先,当前生成数据主要集中在数据增强和数据补充上,这虽然解决了部分数据稀缺问题,但生成数据的应用场景仍然较为单一,特别是在复杂视觉任务中的应用尚未得到充分挖掘。尽管生成模型,如GAN和扩散模型,能够创造高质量的图像,但如何将这些技术有效地扩展到动态场景、视频生成以及实时应用等更复杂的任务中,仍然是一个关键问题。未来,生成数据的应用需要更广泛地涵盖多模态数据生成,如视频、深度图和三维模型等,以应对更为复杂的计算机视觉挑战。其次,跨域适应是生成数据应用中的另一个瓶颈。虽然生成数据能有效解决某些特定场景下的数据不足问题,但它仍然面临着难以适应不同领域或任务的挑战。例如,自动驾驶、医疗影像和遥感图像等领域的任务具有高度的专业性,现有的生成数据技术往往难以准确模拟这些领域的复杂性和特异性。因此,未来的生成数据技术需更好地满足多领域、复杂应用的需求,提升其跨域适应性和准确性。最后,生成数据在高效训练和模型优化方面的潜力尚未完全挖掘。在传统的深度学习任务中,合成数据用做补充数据,但随着计算能力的提高,生成数据可以更广泛地应用于主动学习、知识蒸馏等任务中,进一步优

化模型性能。在这些新兴应用中，生成数据不仅是用于增加样本量，还能根据任务需求生成具有挑战性的样本，推动模型在少样本或无监督环境中的有效训练。总的来说，生成数据在计算机视觉中的应用方式和领域正在不断拓展，未来需要更深入地探索如何将生成技术应用于更复杂、多样的视觉任务中，以推动计算机视觉技术的创新和进步。

4 结 语

本文首先从传统数据生成方法、三维渲染技术以及深度生成模型3个方面系统介绍了典型的数据生成技术与模型；其次深入总结分析了数据生成在图像增强、个体分析、生物特征识别、群体分析、自动驾驶、视频生成以及具身智能等典型计算机视觉任务中的应用；最后总结了面向计算机视觉的数据生成与应用中目前尚存在的问题，并展望数据生成相关技术与应用的未来发展趋势。

致谢：本文由中国图象图形学学会多媒体专业委员会组织撰写，该专业委员会链接为 https://www.csig.org.cn/16/201612/49319.html。撰写过程中得到中国科学院自动化研究所张兆翔研究员、中国科学院计算技术研究所高林研究员、北京大学信息工程学院袁粒助理教授、浙江大学CAD&CG国家重点实验室许威威研究员、北京航空航天大学计算机学院刘艾杉副教授等多位老师的大力支持并提供了部分章节素材，对本文顺利完成起到重要作用，在此一并致谢。

参考文献（References）

Abdal R, Qin Y P and Wonka P. 2019. Image2StyleGAN: how to embed images into the StyleGAN latent space?//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4431-4440 [DOI: 10.1109/ICCV.2019.00453]

Abdal R, Zhu P H, Mitra N J and Wonka P. 2021. StyleFlow: attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. ACM Transactions on Graphics, 40(3): #21 [DOI: 10.1145/3447648]

Ainam J P, Qin K, Liu G S and Luo G C. 2019. Sparse label smoothing regularization for person re-identification. IEEE Access, 7: 27899-27910 [DOI: 10.1109/ACCESS.2019.2901599]

Alomar K, Aysel H I and Cai X H. 2023. Data augmentation in classification and segmentation: a survey and new strategies. Journal of Imaging, 9(2): #46 [DOI: 10.3390/jimaging9020046]

An J, Zhang S Y, Yang H, Gupta S, Huang J B, Luo J B and Yin X. 2023a. Latent-shift: latent diffusion with temporal shift for efficient text-to-video generation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2304.08477.pdf

An S Z, Xu H Y, Shi Y C, Song G X, Ogras U Y and Luo L J. 2023b. PanoHead: geometry-aware 3D full-head synthesis in 360°//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 20950-20959 [DOI: 10.1109/CVPR52729.2023.02007]

Andrychowicz O M, Baker B, Chociej M, Józefowicz R, McGrew B, Pachocki J, Petron A, Plappert M, Powell G, Ray A, Schneider J, Sidor S, Tobin J, Welinder P, Weng L L and Zaremba W. 2020. Learning dexterous in-hand manipulation. The International Journal of Robotics Research, 39(1): 3-20 [DOI: 10.1177/0278364919887447]

Aranjuelo N, García S, Loyo E, Unzueta L and Otaegui O. 2021. Key strategies for synthetic data generation for training intelligent systems based on people detection from omnidirectional cameras. Computers and Electrical Engineering, 92: #107105 [DOI: 10.1016/j.compeleceng.2021.107105]

Arjovsky M, Chintala S and Bottou L. 2017. Wasserstein GAN [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1701.07875.pdf

Attia M, Attia M H, Iskander J, Saleh K, Nahavandi D, Abobakr A, Hossny M and Nahavandi S. 2019. Fingerprint synthesis via latent space representation//Proceedings of 2019 IEEE International Conference on Systems, Man and Cybernetics. Bari, Italy: IEEE: 1855-1861 [DOI: 10.1109/SMC.2019.8914499]

Azizi S, Kornblith S, Saharia C, Norouzi M and Fleet D J. 2023. Synthetic data from diffusion models improves imagenet classification [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2304.08466.pdf

Badler N I, Phillips C B and Webber B L. 1993. Simulating Humans: Computer Graphics, Animation, and Control. New York, USA: Oxford University Press

Bahmani K, Plesh R, Johnson P, Schuckers S and Swyka T. 2021. High fidelity fingerprint generation: quality, uniqueness, and privacy//Proceedings of 2021 IEEE International Conference on Image Processing. Anchorage, USA: IEEE: 3018-3022 [DOI: 10.1109/ICIP42928.2021.9506386]

Bai Q Y, Xia W H, Yin F and Yang Y J. 2022. Identity-guided face generation with multi-modal contour conditions//Proceedings of 2022 IEEE International Conference on Image Processing. Bordeaux, France: IEEE: 1881-1885 [DOI: 10.1109/ICIP46576.2022.9897459]

Bąk S, Carr P and Lalonde J F. 2018. Domain adaptation through synthesis for unsupervised person re-identification//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer International Publishing: 193-209 [DOI: 10.1007/978-3-030-01261-8_12]

Bao F, Xiang C D, Yue G, He G D, Zhu H Z, Zheng K W, Zhao M,

Liu S L, Wang Y L and Zhu J. 2024. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2405.04233.pdf

Bao J M, Chen D, Wen F, Li H Q and Hua G. 2017. CVAE-GAN: fine-grained image generation through asymmetric training//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2764-2773 [DOI: 10.1109/ICCV.2017.299]

Bao J M, Chen D, Wen F, Li H Q and Hua G. 2018. Towards open-set identity preserving face synthesis//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6713-6722 [DOI: 10.1109/CVPR.2018.00702]

Barbosa I B, Cristani M, Caputo B, Rognhaugen A and Theoharis T. 2018. Looking beyond appearances: synthetic training data for deep CNNs in re-identification. Computer Vision and Image Understanding, 167: 50-62 [DOI: 10.1016/j.cviu.2017.12.002]

Barron J T, Mildenhall B, Verbin D, Srinivasan P P and Hedman P. 2022. Mip-NeRF 360: unbounded anti-aliased neural radiance fields//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5460-5469 [DOI: 10.1109/CVPR52688.2022.00539]

Bar-Tal O, Chefer H, Tov O, Herrmann C, Paiss R, Zada S, Ephrat A, Hur J, Liu G H, Raj A, Li Y Z, Rubinstein M, Michaeli T, Wang O, Sun D Q, Dekel T and Mosseri I. 2024. Lumiere: a space-time diffusion model for video generation//Proceedings of 2024 SIGGRAPH Asia Conference Papers. Tokyo, Japan: Association for Computing Machinery: #94 [DOI: 10.1145/3680528.3687614]

Bazavan E G, Zanfir A, Zanfir M, Freeman W T, Sukthankar R and Sminchisescu C. 2022. HSPACE: synthetic parametric humans animated in complex environments [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2112.12867.pdf

Beattie C, Leibo J Z, Teplyashin D, Ward T, Wainwright M, Küttler H, Lefrancq A, Green S, Valdés V, Sadik A, Schrittwieser J, Anderson K, York S, Cant M, Cain A, Bolton A, Gaffney S, King H, Hassabis D, Legg S and Petersen S. 2016. DeepMind lab[EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1612.03801.pdf

Bell-Kligler S, Shocher A and Irani M. 2019. Blind super-resolution kernel estimation using an internal-GAN//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 284-293

Bergman A W, Kellnhofer P, Wang Y F, Chan E R, Lindell D B and Wetzstein G. 2022. Generative neural articulated radiance fields//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 19900-19916

Black M J, Patel P, Tesch J and Yang J L. 2023. BEDLAM: a synthetic dataset of bodies exhibiting detailed lifelike animated motion//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 8726-8737 [DOI: 10.1109/CVPR52729.2023.00843]

Blattmann A, Dockhorn T, Kulal S, Mendelevitch D, Kilian M, Lorenz D, Levi Y, English Z, Voleti V, Letts A, Jampani V and Rombach R. 2023a. Stable video diffusion: scaling latent video diffusion models to large datasets [EB/OL]. [2025-02-28]. 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 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 22563-22575 [DOI: 10.1109/CVPR52729.2023.02161]

Bontrager P, Roy A, Togelius J, Memon N and Ross A. 2018. DeepMasterPrints: generating MasterPrints for dictionary attacks via latent variable evolution//Proceedings of the 9th IEEE International Conference on Biometrics Theory, Applications and Systems. Redondo Beach, USA: IEEE: 1-9 [DOI: 10.1109/BTAS.2018.8698539]

Borgia A, Hua Y, Kodirov E and Robertson N. 2019. GAN-based pose-aware regulation for video-based person re-identification//Proceedings of 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 1175-1184 [DOI: 10.1109/WACV.2019.00130]

Brock A, Donahue J and Simonyan K. 2018. Large scale GAN training for high fidelity natural image synthesis [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1809.11096.pdf

Brooks T and Efros A A. 2022. Hallucinating pose-compatible scenes//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 510-528 [DOI: 10.1007/978-3-031-19787-1_29]

Caesar H, Bankiti V, Lang A H, Vora S, Liong V E, Xu Q, Krishnan A, Pan Y, Baldan G and Beijbom O. 2020. nuScenes: a multimodal dataset for autonomous driving//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11618-11628 [DOI: 10.1109/CVPR42600.2020.01164]

Cai J R, Gu S H and Zhang L. 2018. Learning a deep single image contrast enhancer from multi-exposure images. IEEE Transactions on Image Processing, 27(4): 2049-2062 [DOI: 10.1109/TIP.2018.2794218]

Cai Y H, Bian H, Lin J, Wang H Q, Timofte R and Zhang Y L. 2023. Retinexformer: one-stage Retinex based transformer for low-light image enhancement//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 12470-12479 [DOI: 10.1109/ICCV51070.2023.01149]

Cai Z A, Zhang M Y, Ren J W, Wei C, Ren D X, Lin Z Y, Zhao H Y, Yang L, Loy C C and Liu Z W. 2024. Playing for 3D human recovery [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2110.07588.pdf

Cao K and Jain A. 2018. Fingerprint synthesis: evaluating fingerprint search at scale//Proceedings of 2018 International Conference on

Biometrics. Gold Coast, Australia: IEEE: 31-38 [DOI: 10.1109/ICB2018.2018.00016]

Cao Y K, Cao Y P, Han K, Shan Y and Wong K Y K. 2024. DreamAvatar: text-and-shape guided 3D human avatar generation via diffusion models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 958-968 [DOI: 10.1109/CVPR52733.2024.00097]

Cappelli R, Erol A, Maio D and Maltoni D. 2000. Synthetic fingerprint-image generation//Proceedings of the 15th International Conference on Pattern Recognition. Barcelona, Spain: IEEE: 471-474 [DOI: 10.1109/ICPR.2000.903586]

Cappelli R, Maio D and Maltoni D. 2001. Modelling plastic distortion in fingerprint images//Proceedings of the 2nd International Conference on Advances in Pattern Recognition. Rio De Janeiro, Brazil: Springer: 371-378 [DOI: 10.1007/3-540-44732-6_38]

Cappelli R, Maio D and Maltoni D. 2002. Synthetic fingerprint-database generation//Proceedings of 2002 International Conference on Pattern Recognition. Quebec City, Canada: IEEE: 744-747 [DOI: 10.1109/ICPR.2002.1048096]

Cappelli R, Maio D and Maltoni D. 2004. An improved noise model for the generation of synthetic fingerprints//Proceedings of the 8th ICARCV Control, Automation, Robotics and Vision Conference. Kunming, China: IEEE: 1250-1255 [DOI: 10.1109/ICARCV.2004.1469025]

Chan C, Ginosar S, Zhou T H and Efros A. 2019. Everybody dance now//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5932-5941 [DOI: 10.1109/ICCV.2019.00603]

Chan E R, Monteiro M, Kellnhofer P, Wu J J and Wetzstein G. 2021a. pi-GAN: periodic implicit generative adversarial networks for 3D-aware image synthesis//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5795-5805 [DOI: 10.1109/CVPR46437.2021.00574]

Chan E R, Lin C Z, Chan M A, Nagano K, Pan B X, de Mello S, Gallo O, Guibas L, Tremblay J, Khamis S, Karras T and Wetzstein G. 2022. Efficient geometry-aware 3D generative adversarial networks//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16102-16112 [DOI:10.1109/CVPR52688.2022.01565]

Chan K C K, Wang X T, Xu X Y, Gu J W and Loy C C. 2021b. GLEAN: generative latent bank for large-factor image super-resolution//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14240-14249 [DOI: 10.1109/CVPR46437.2021.01402]

Chang A, Dai A, Funkhouser T, Halber M, Nießner M, Savva M, Song S R, Zeng A and Zhang Y D. 2017. Matterport3D: learning from RGB-D data in indoor environments//Proceedings of 2017 International Conference on 3D Vision. Qingdao, China: IEEE: 667-676 [DOI: 10.1109/3DV.2017.00081]

Chao W T, Chang L, Wang X G, Cheng J, Deng X M and Duan F Q. 2019. High-fidelity face sketch-to-photo synthesis using generative adversarial network//Proceedings of 2019 IEEE International Conference on Image Processing. Taipei, China: IEEE: 4699-4703 [DOI: 10.1109/ICIP.2019.8803549]

Charatan D, Li S L, Tagliasacchi A and Sitzmann V. 2024. PixelSplat: 3D Gaussian splats from image pairs for scalable generalizable 3D reconstruction//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 19457-19467 [DOI: 10.1109/CVPR52733.2024.01840]

Chen G L, Zhao H Y, Pang C K, Li T L and Pang C Y. 2019. Image scaling: how hard can it be?. IEEE Access, 7: 129452-129465 [DOI: 10.1109/ACCESS.2019.2940353]

Chen H X, Xia M H, He Y Q, Zhang Y, Cun X D, Yang S S, Xing J B, Liu Y F, Chen Q F, Wang X T, Weng C and Shan Y. 2023a. VideoCrafter1: open diffusion models for high-quality video generation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2310.19512.pdf

Chen H S, Gu J T, Chen A P, Tian W, Tu Z W, Liu L J and Su H. 2023b. Single-stage diffusion NeRF: a unified approach to 3D generation and reconstruction//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 2416-2425 [DOI: 10.1109/ICCV51070.2023.00229]

Chen J Y, Ye W C, Wang Y F, Chen D P, Huang D, Ouyang W L, Zhang G F, Qiao Y and He T. 2024a. GigaGS: scaling up planar-based 3D Gaussians for large scene surface reconstruction [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2409.06685.pdf

Chen K, Chen W H, He T, Du R, Wang F, Sun X Y, Guo Y C and Ding G G. 2022. TAGPerson: a target-aware generation pipeline for person re-identification//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 560-571 [DOI: 10.1145/3503161.3548013]

Chen K, Xie E Z, Chen Z, Wang Y B, Hong L Q, Li Z G and Yeung D Y. 2024b. Geodiffusion: text-prompted geometric control for object detection data generation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2306.04607.pdf

Chen R, Chen Y W, Jiao N X and Jia K. 2023c. Fantasia3D: disentangling geometry and appearance for high-quality text-to-3D content creation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22189-22199 [DOI: 10.1109/ICCV51070.2023.02033]

Chen W L and Hays J. 2018. SketchyGAN: towards diverse and realistic sketch to image synthesis//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9416-9425 [DOI: 10.1109/CVPR.2018.00981]

Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I and Abbeel P. 2016. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 2180-2188

Chen X Y, Wang Y H, Zhang L J, Zhuang S B, Ma X, Yu J S, Wang Y L, Lin D H, Qiao Y and Liu Z W. 2023d. SEINE: short-to-long video diffusion model for generative transition and prediction [EB/OL]. [2025-04-22]. https://arxiv.org/pdf/2310.20700.pdf

Chen Y and Jain A K. 2009. Beyond minutiae: a fingerprint individuality model with pattern, ridge and pore features//Proceedings of the 3rd International Conference on Biometrics. Alghero, Italy: Springer: 523-533 [DOI: 10.1007/978-3-642-01793-3_54]

Chen Y D, Xu H F, Zheng C X, Zhuang B H, Pollefeys M, Geiger A, Cham T J and Cai J F. 2025a. MVSplat: efficient 3D Gaussian splatting from sparse multi-view images//Proceedings of the 18th European Conference on Computer Vision-ECCV 2024. Milan, Italy: Springer-Verlag: 370-386 [DOI: 10.1007/978-3-031-72664-4_21]

Chen Y T, Mihajlovic M, Chen X Y, Wang Y M, Prokudin S and Tang S Y. 2025b. SplatFormer: point transformer for robust 3D Gaussian splatting [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2411.06390.pdf

Cheng J X, Liang X, Shi X J, He T, Xiao T J and Li M. 2023. LayoutDiffuse: adapting foundational diffusion models for layout-to-image generation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2302.08908.pdf

Cheung E, Wong A, Bera A, Wang X G and Manocha D. 2019. LCrowdV: generating labeled videos for pedestrian detectors training and crowd behavior learning. Neurocomputing, 337: 1-14 [DOI: 10.1016/j.neucom.2018.08.085]

Choi Y, Choi M, Kim M, Ha J W, Kim S and Choo J. 2018. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8789-8797 [DOI: 10.1109/CVPR.2018.00916]

Ciampi L, Messina N, Falchi F, Gennaro C and Amato G. 2020. Virtual to real adaptation of pedestrian detectors. Sensors, 20(18): #5250 [DOI: 10.3390/s20185250]

Clark A, Donahue J and Simonyan K. 2019. Adversarial video generation on complex datasets [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1907.06571.pdf

Courty N, Allain P, Creusot C and Corpetti T. 2014. Using the agoraset dataset: assessing for the quality of crowd video analysis methods. Pattern Recognition Letters, 44: 161-170 [DOI: 10.1016/j.patrec.2014.01.004]

Crisan S, Târnovan I G and Crişan T E. 2008. A hand vein structure simulation platform for algorithm testing and biometric identification//Proceedings of the 16th IMEKO TC4 Symposium. Florence, Italy: [s.n.]

Cui J L, Wang Y H, Huang J Z, Tan T N and Sun Z. 2004. An iris image synthesis method based on PCA and super-resolution//Proceedings of the 17th International Conference on Pattern Recognition, 2004. Cambridge, UK: IEEE: 471-474 [DOI: 10.1109/ICPR.2004.1333804]

Curtó J D, Zarza I C, de la Torre F, King I and Lyu M R. 2020. High-resolution deep convolutional generative adversarial networks [EB/OL]. [2025-02-28]. http://arxiv.org/pdf/1711.06491.pdf

Dai P X, Xu J M, Xie W X, Liu X G, Wang H M and Xu W W. 2024. High-quality surface reconstruction using Gaussian surfels//Proceedings of 2024 ACM SIGGRAPH Conference Papers. Denver, USA: Association for Computing Machinery: #22 [DOI: 10.1145/3641519.3657441]

de Bézenac E, Rangapuram S S, Benidis K, Bohlke-Schneider M, Kurle R, Stella L, Hasson H, Gallinari P and Januschowski T. 2020. Normalizing Kalman filters for multivariate time series analysis//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 2995-3007

Deitke M, Schwenk D, Salvador J, Weihs L, Michel O, VanderBilt E, Schmidt L, Ehsanit K, Kembhavi A and Farhadi A. 2023. Objaverse: a universe of annotated 3D objects//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 13142-13153 [DOI: 10.1109/CVPR52729.2023.01263]

Deng C Y, Jiang C M, Qi C R, Yan X C, Zhou Y, Guibas L and Anguelov D. 2023. NeRDi: single-view NeRF synthesis with language-guided diffusion as general image priors//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 20637-20647 [DOI: 10.1109/CVPR52729.2023.01977]

Deng Y, Yang J L, Chen D, Wen F and Tong X. 2020. Disentangled and controllable face image generation via 3D imitative-contrastive learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5153-5162 [DOI: 10.1109/CVPR42600.2020.00520]

Deng Y, Yang J L, Xiang J F and Tong X. 2022. GRAM: generative radiance manifolds for 3D-aware image generation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10663-10673 [DOI: 10.1109/CVPR52688.2022.01041]

Denton E, Chintala S, Szlam A, and Fergus R. 2015. Deep generative image models using a laplacian pyramid of adversarial networks//Proceedings of the 29th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 1486-1494

de Souza C R, Gaidon A, Cabon Y and López A M. 2017. Procedural generation of videos to train deep action recognition networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2594-2604 [DOI: 10.1109/CVPR.2017.278]

DeVries T, Bautista M A, Srivastava N, Taylor G W and Susskind J M. 2021. Unconstrained scene generation with locally conditioned radi-

ance fields//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 14284-14293 [DOI: 10.1109/ICCV48922.2021.01404]

Di Benedetto M, Carrara F, Meloni E, Amato G, Falchi F and Gennaro C. 2021. Learning accurate personal protective equipment detection from virtual worlds. Multimedia Tools and Applications, 80(15): 23241-23253 [DOI: 10.1007/s11042-020-09597-9]

Dinh L, Krueger D and Bengio Y. 2015. NICE: non-linear independent components estimation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1410.8516.pdf

Dinh L, Sohl-Dickstein J and Bengio S. 2017. Density estimation using Real NVP. arXiv:1605.08803 [DOI: 10.48550/arXiv.1605.08803]

Donahue C, Lipton Z C, Balsubramani A and McAuley J. 2018. Semantically decomposing the latent spaces of generative adversarial networks [EB/OL]. [2025-02-28]. http://arxiv.org/pdf/1705.07904.pdf

Dong C, Loy C C, He K M and Tang X O. 2014. Learning a deep convolutional network for image super-resolution//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 184-199 [DOI: 10.1007/978-3-319-10593-2_13]

Dong Z J, Chen X, Yang J L, Black M J, Hilliges O and Geiger A. 2023. AG3D: learning to generate 3D avatars from 2D image collections//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 14870-14881 [DOI: 10.1109/ICCV51070.2023.01370]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale[EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2010.11929.pdf

Dosovitskiy A, Ros G, Codevilla F, López A and Koltun V. 2017. CARLA: an open urban driving simulator[EB/OL]. [2025-02-28]. http://arxiv.org/pdf/1711.03938.pdf

Du X Z, Zoph B, Hung W C and Lin T Y. 2021. Simple training strategies and model scaling for object detection[EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2107.00057.pdf

Duan Y X, Wei F Y, Dai Q Y, He Y H, Chen W Z and Chen B Q. 2024. 4D-rotor Gaussian splatting: towards efficient novel view synthesis for dynamic scenes//Proceedings of 2024 ACM SIGGRAPH Conference Papers. Denver, USA: Association for Computing Machinery: 87 [DOI: 10.1145/3641519.3657463]

Dudhane A, Zamir S W, Khan S, Khan F S and Yang M H. 2023. Burstormer: burst image restoration and enhancement transformer//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 5703-5712 [DOI: 10.1109/CVPR52729.2023.00552]

Dunlap L, Umino A, Zhang H, Yang J Z, Gonzalez J E and Darrell T. 2023. Diversify your vision datasets with automatic diffusion-based augmentation//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 79024-79034

Durvasula S, Zhao A, Chen F, Liang R F, Sanjaya P K and Vijaykumar N. 2023. DISTWAR: fast differentiable rendering on raster-based rendering pipelines [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2401.05345.pdf

Dvornik N, Mairal J and Schmid C. 2018. Modeling visual context is key to augmenting object detection datasets//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 375-391 [DOI: 10.1007/978-3-030-01258-8_23]

Dwibedi D, Misra I and Hebert M. 2017. Cut, paste and learn: surprisingly easy synthesis for instance detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1310-1319 [DOI: 10.1109/ICCV.2017.146]

Ekbatani H K, Pujol O and Seguí S. 2017. Synthetic data generation for deep learning in counting pedestrians//Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods. Porto, Portugal: SciTePress: 318-323 [DOI: 10.5220/0006119203180323]

Engelsma J J, Cao K and Jain A K. 2021. Learning a fixed-length fingerprint representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(6): 1981-1997 [DOI: 10.1109/TPAMI.2019.2961349]

Engelsma J J, Grosz S A and Jain A K. 2023. PrintsGAN: synthetic fingerprint generator. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5): 6111-6124 [DOI: 10.1109/TPAMI.2022.3204591]

Eom C and Ham B. 2019. Learning disentangled representation for robust person re-identification//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 5297-5308

Esser P, Rombach R and Ommer B. 2021. Taming Transformers for High-Resolution Image Synthesis//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12868-12878 [DOI: 10.1109/cvpr46437.2021.01268]

Esser P, Kulal S, Blattmann A, Entezari R, Müller J, Saini H, Levi Y, Lorenz D, Sauer A, Boesel F, Podell D, Dockhorn T, English Z, Lacey K, Goodwin A, Marek Y and Rombach R. 2024. Scaling rectified flow transformers for high-resolution image synthesis [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2403.03206.pdf

Fabbri M, Brasó G, Maugeri G, Cetintas O, Gasparini R, Osep A, Calderara S, Leal-Taixé L and Cucchiara R. 2021. MOTSynth: how can synthetic data help pedestrian detection and tracking?//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10829-10839 [DOI: 10.1109/ICCV48922.2021.01067]

Fahim M A N I and Jung H Y. 2020. A lightweight GAN network for large scale fingerprint generation. IEEE Access, 8: 92918-92928 [DOI: 10.1109/ACCESS.2020.2994371]

Fan Z W, Wang K, Wen K R, Zhu Z H, Xu D J and Wang Z Y. 2024. LightGaussian: unbounded 3D Gaussian compression with 15 × reduction and 200+ FPS [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2311.17245.pdf

Fang H Y, Han B R, Zhang S, Zhou S, Hu C X and Ye W M. 2024. Data augmentation for object detection via controllable diffusion models//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 1246-1255 [DOI: 10.1109/WACV57701.2024.00129]

Feng C J, Zhong Y J, Jie Z Q, Xie W D and Ma L. 2024. InstaGen: enhancing object detection by training on synthetic dataset//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14121-14130 [DOI: 10.1109/CVPR52733.2024.01339]

Feng J J and Jain A K. 2011. Fingerprint reconstruction: from minutiae to phase. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(2): 209-223 [DOI: 10.1109/TPAMI.2010.77]

Feng L, Li Q Y, Peng Z H, Tan S H and Zhou B L. 2023. TrafficGen: learning to generate diverse and realistic traffic scenarios//Proceedings of 2023 IEEE International Conference on Robotics and Automation. London, United Kingdom: IEEE: 3567-3575 [DOI: 10.1109/ICRA48891.2023.10160296]

Fogel I and Sagi D. 1989. Gabor filters as texture discriminator. Biological Cybernetics, 61(2): 103-113 [DOI: 10.1007/BF00204594]

Gaidon A, Wang Q, Cabon Y and Vig E. 2016. VirtualWorlds as proxy for multi-object tracking analysis//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4340-4349 [DOI: 10.1109/CVPR.2016.470]

Gan C, Schwartz J, Alter S, Mrowca D, Schrimpf M, Traer J, De Freitas J, Kubilius J, Bhandwaldar A, Haber N, Sano M, Kim K, Wang E, Lingelbach M, Curtis A, Feigelis K, Bear D M, Gutfreund D, Cox C, Torralba A, DiCarlo J J, Tenenbaum J B, McDermott J H and Yamins D L K. 2021. ThreeDWorld: a platform for interactive multi-modal physical simulation [EB/OL]. [2025-02-22]. https://arxiv.org/pdf/2007.04954.pdf

Gao J, Shen T C, Wang Z A, Chen W Z, Yin K X, Li D Q, Litany O, Gojcic Z and Fidler S. 2022. GET3D: a generative model of high quality 3D textured shapes learned from images//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 31841-31854

Gao S C, Liu X H, Zeng B H, Xu S, Li Y J, Luo X Y, Liu J Z, Zhen X T and Zhang B C. 2023. Implicit diffusion models for continuous super-resolution//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 10021-10030 [DOI: 10.1109/CVPR52729.2023.00966]

Gao X F, Gong R, Shu T M, Xie X, Wang S and Zhu S C. 2019. VRKitchen: an interactive 3D virtual environment for task-oriented learning [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1903.05757

Ge Y H, Xu J S, Zhao B N, Itti L and Vineet V. 2022. DALL-E for detection: language-driven compositional image synthesis for object detection [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2206.09592.pdf

Gecer B, Bhattarai B, Kittler J and Kim T K. 2018. Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3D morphable model//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 230-248 [DOI: 10.1007/978-3-030-01252-6_14]

Geng H R, Xu H L, Zhao C Y, Xu C, Yi L, Huang S Y and Wang H. 2023. GAPartNet: cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 7081-7091 [DOI: 10.1109/CVPR52729.2023.00684]

Ghiasi G, Cui Y, Srinivas A, Qian R, Lin T Y, Cubuk E D, Le Q V and Zoph B. 2021. Simple copy-paste is a strong data augmentation method for instance segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2917-2927 [DOI: 10.1109/CVPR46437.2021.00294]

Girdhar R, Singh M, Brown A, Duval Q, Azadi S, Rambhatla S S, Shah A, Yin X, Parikh D and Misra I. 2024. Emu video: factorizing text-to-video generation by explicit image conditioning [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2311.10709.pdf

Girish S, Gupta K and Shrivastava A. 2024. EAGLES: efficient accelerated 3D Gaussians with lightweight encodings [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2312.04564.pdf

Gomez A N, Ren M, Urtasun R, and Grosse R B. 2017. The reversible residual network: backpropagation without storing activations//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 2211-2221

Gong Y P, Zeng Z Y, Chen L W, Luo Y F, Weng B and Ye F. 2021. A person re-identification data augmentation method with adversarial defense effect [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2101.08783.pdf

Gong R, Danelljan M, Sun H, Mangas J D and Van Gool L. 2023. Prompting diffusion representations for cross-domain semantic segmentation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2307.02138.pdf

Gonzales R C and Wintz P. 1987. Digital image processing. Boston, USA: Addison-Wesley Longman Publishing Co., Inc.

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680 [DOI:10.1145/3422622]

Gowda S N, Rohrbach M, Keller F and Sevilla-Lara L. 2022.

Learn2Augment: learning to composite videos for data augmentation in action recognition//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 242-259 [DOI: 10.1007/978-3-031-19821-2_14]

Grathwohl W, Chen R T Q, Bettencourt J, Sutskever I and Duvenaud D. 2018. FFJORD: free-form continuous dynamics for scalable reversible generative models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1810.01367.pdf

Grigorev A, Iskakov K, Ianina A, Bashirov R, Zakharkin I, Vakhitov A and Lempitsky V. 2021. StylePeople: a generative model of full-body human avatars//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5147-5156 [DOI: 10.1109/CVPR46437.2021.00511]

Gu J T, Trevithick A, Lin K E, Susskind J, Theobalt C, Liu L J and Ramamoorthi R. 2023. NeRFDiff: single-image view synthesis with NeRF-guided distillation from 3D-aware diffusion//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 11808-11826

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V and Courville A C. 2017. Improved training of Wasserstein GANs//Proceedings of the 34th International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 5769-5779

Guo C L, Li C Y, Guo J C, Loy C C, Hou J H, Kwong S and Cong R M. 2020. Zero-reference deep curve estimation for low-light image enhancement//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1777-1786 [DOI: 10.1109/CVPR42600.2020.00185]

Guo Y W, Yang C Y, Rao A Y, Liang Z Y, Wang Y H, Qiao Y, Agrawala M, Lin D H and Dai B. 2024. Animatediff: animate your personalized text-to-image diffusion models without specific tuning [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2307.04725.pdf

Gupta A, Yu L J, Sohn K, Gu X Y, Hahn M, Li F F, Essa I, Jiang L and Lezama J. 2025. Photorealistic video generation with diffusion models//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 393-411 [DOI: 10.1007/978-3-031-72986-7_23]

He K M, Chen X L, Xie S N, Li Y H, Dollár P and Girshick R. 2022a. Masked autoencoders are scalable vision learners//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 15979-15988 [DOI: 10.1109/cvpr52688.2022.01553]

He R F, Sun S Y, Yu X, Xue C H, Zhang W Q, Torr P, Bai S and Qi X J. 2023a. Is synthetic data from generative models ready for image recognition? [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2210.07574.pdf

He Y Q, Yang T Y, Zhang Y, Shan Y and Chen Q F. 2023b. Latent video diffusion models for high-fidelity long video generation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2211.13221.pdf

He Z, Lin M W, Xu Z S, Yao Z Q, Chen H, Alhudhaif A and Alenezi F. 2022b. Deconv-transformer (DecT): a histopathological image classification model for breast cancer based on color deconvolution and transformer architecture. Information Sciences, 608: 1093-1112 [DOI: 10.1016/j.ins.2022.06.091]

Helbing D, Farkas I and Vicsek T. 2000. Simulating dynamical features of escape panic. Nature, 407 (6803): 487-490 [DOI: 10.1038/35035023]

Helminger L, Bernasconi M, Djelouah A, Gross M and Schroers C. 2021. Generic image restoration with flow based priors//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville, USA: IEEE: 334-343 [DOI: 10.1109/CVPRW53098.2021.00043]

Hewitt C, Saleh F, Aliakbarian S, Petikam L, Rezaeifar S, Florentin L, Hosenie Z, Cashman T J, Valentin J, Cosker D and Baltrusaitis T. 2024. Look ma, no markers: holistic performance capture without the hassle. ACM Transactions on Graphics, 43 (6): #235 [DOI: 10.1145/3687772]

Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S and Lerchner A. 2017. β-VAE: learning basic visual concepts with a constrained variational framework//Proceedings of 2017 International Conference on Learning Representations. Toulon, France: OpenReview.net: 1-22

Hillerström F, Kumar A and Veldhuis R. 2014. Generating and analyzing synthetic finger vein images//Proceedings of 2014 International Conference of the Biometrics Special Interest Group. Darmstadt, Germany: IEEE: 1-9

Ho J, Chan W, Saharia C, Whang J, Gao R Q, Gritsenko A, Kingma D P, Poole B, Norouzi M, Fleet D J and Salimans T. 2022a. Imagen video: high definition video generation with diffusion models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2210.02303.pdf

Ho J, Chen X, Srinivas A, Duan Y and Abbeel P. 2019. Flow++: improving flow-based generative models with variational dequantization and architecture design//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 2722-2730

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 6840-6851

Hong F Z, Chen Z X, Lan Y S, Pan L and Liu Z W. 2022c. EVA3D: compositional 3D human generation from 2D image collections [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2210.04888.pdf

Hong F Z, Zhang M Y, Pan L, Cai Z, Yang L and Liu Z W. 2022a. AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars. ACM Transactions on Graphics, 41(4): #161 [DOI: 10.1145/3528223.3530094]

Hong S, Seo J, Shin H, Hong S and Kim S. 2024. DirecT2V: large language models are frame-level directors for zero-shot text-to-video generation [EB/OL]. [2025-02-28].

https://arxiv.org/pdf/2305.14330.pdf

Hong W Y, Ding M, Zheng W D, Liu X H and Tang J. 2022b. CogVideo: large-scale pretraining for text-to-video generation via transformers [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2205.15868.pdf

Hou Y, Li C Y, Lu Y H, Zhu L P, Li Y, Jia H Z and Xie X D. 2022. Enhancing and dissecting crowd counting by synthetic data//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, Singapore: IEEE: 2539-2543 [DOI: 10.1109/ICASSP43922.2022.9747070]

Hou Y, Zhang S H, Ma R, Jia H Z and Xie X D. 2023. Frame-recurrent video crowd counting. IEEE Transactions on Circuits and Systems for Video Technology, 33(9): 5186-5199 [DOI: 10.1109/TCSVT.2023.3245678]

Hu M K, Zhao P, Xu C, Sun Q F, Lou J G, Lin Q W, Luo P and Rajmohan S. 2025. AgentGen: enhancing planning abilities for large language model based agent via environment and task generation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2408.00764.pdf

Hu Y T, Chen H S, Hui K X, Huang J B and Schwing A G. 2019. SAIL-VOS: semantic amodal instance level video object segmentation — a synthetic dataset and baselines//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3100-3110 [DOI: 10.1109/CVPR.2019.00322]

Huang B B, Yu Z H, Chen A P, Geiger A and Gao S H. 2024a. 2D Gaussian splatting for geometrically accurate radiance fields//Proceedings of 2024 ACM SIGGRAPH Conference Papers. Denver, USA: Association for Computing Machinery: #32 [DOI: 10.1145/3641519.3657428]

Huang C W, Krueger D, Lacoste A and Courville A. 2018a. Neural autoregressive flows//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 2078-2087

Huang H J, Li D W, Zhang Z, Chen X T and Huang K Q. 2018b. Adversarially occluded samples for person re-identification//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5098-5107 [DOI: 10.1109/CVPR.2018.00535]

Huang J Z, Ma L, Tan T N and Wang Y H. 2003. Learning based resolution enhancement of iris images//Proceedings of 2023 British Machine Vision Conference. Norwich, UK: [s.n.] [DOI: 10.5244/C.17.16]

Huang Y, Wu Q, Xu J S and Zhong Y. 2019. SBSGAN: suppression of inter-domain background shift for person re-identification//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9526-9535 [DOI: 10.1109/ICCV.2019.00962]

Huang Z Q, Chan K C K, Jiang Y M and Liu Z W. 2023. Collaborative diffusion for multi-modal face generation and editing//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 6080-6090 [DOI: 10.1109/CVPR52729.2023.00589]

Huang Z Q, He Y N, Yu J S, Zhang F, Si C Y, Jiang Y M, Zhang Y H, Wu T X, Jin Q Y, Chanpaisit N, Wang Y H, Chen X Y, Wang L M, Lin D H, Qiao Y and Liu Z W. 2024b. VBench: comprehensive benchmark suite for video generative models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 21807-21818 [DOI: 10.1109/CVPR52733.2024.02060]

Huang Z X, Chen Q, Sun L B, Yang Y F, Wang N Z, Wu Q and Tan M K. 2024c. G-NeRF: geometry-enhanced novel view synthesis from single-view images//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10117-10126 [DOI: 10.1109/CVPR52733.2024.00964]

Ionescu C, Papava D, Olaru V and Sminchisescu C. 2014. Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(7): 1325-1339 [DOI: 10.1109/TPAMI.2013.248]

Isola P, Zhu J Y, Zhou T H and Efros A A. 2017. Image-to-image translation with conditional adversarial networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5967-5976 [DOI: 10.1109/CVPR.2017.632]

Jahn M, Rombach R and Ommer B. 2021. High-resolution complex scene synthesis with transformers [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2105.06458.pdf

Jain A, Mildenhall B, Barron J T, Abbeel P and Poole B. 2022. Zero-shot text-guided object generation with dream fields//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 857-866 [DOI: 10.1109/CVPR52688.2022.00094]

Jiang B, Chen X, Liu W, Yu J Y, Yu G and Chen T. 2023a. MotionGPT: human motion as a foreign language//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 20067-20079

Jiang Y F, Gong X Y, Liu D, Cheng Y, Fang C, Shen X H, Yang J C, Zhou P and Wang Z Y. 2021. EnlightenGAN: deep light enhancement without paired supervision. IEEE Transactions on Image Processing, 30: 2340-2349 [DOI: 10.1109/TIP.2021.3051462]

Jiang Y F, Wang C, Zhang R H, Wu J J and Li F F. 2024. TRANSIC: sim-to-real policy transfer by learning from online correction [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2405.10315.pdf

Jiang Y W Q, Tu J D, Liu Y, Gao X F, Long X X, Wang W P and Ma Y X. 2023b. GaussianShader: 3D Gaussian splatting with shading functions for reflective surfaces [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2311.17977.pdf

Jin J, Zhao C L, Zhang R X and Jia W. 2025a. Diff-Palm: realistic

palmprint generation with polynomial creases and intra-class variation controllable diffusion models//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE

Jin J L, Shen L, Zhang R X, Zhao C L, Jin G, Zhang J Y, Ding S H, Zhao Y and Jia W. 2024. PCE-palm: palm crease energy based two-stage realistic pseudo-palmprint generation//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 2616-2624 [DOI: 10.1609/aaai.v38i3.28039]

Jin X, Chen Z B, Lin J X, Chen Z K and Zhou W. 2019. Unsupervised single image deraining with self-supervised constraints//Proceedings of 2019 IEEE International Conference on Image Processing. Taipei, China: IEEE: 2761-2765 [DOI: 10.1109/ICIP.2019.8803238]

Jin Y, Sun Z C, Li N Y, Xu K, Jiang H, Zhuang N, Huang Q Z, Song Y, Mu Y D and Lin Z C. 2025b. Pyramidal flow matching for efficient video generative modeling [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2410.05954.pdf

Johnson J, Alahi A and Li F F. 2016. Perceptual losses for real-time style transfer and super-resolution//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 694-711 [DOI: 10.1007/978-3-319-46475-6_43]

Johnson P, Hua F and Schuckers S. 2013. Texture modeling for synthetic fingerprint generation//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Portland, USA: IEEE: 154-159 [DOI: 10.1109/CVPRW.2013.30]

Ju X, Zeng A L, Zhao C C, Wang J N, Zhang L and Xu Q. 2023. HumanSD: a native skeleton-guided diffusion model for human image generation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 15942-15952 [DOI: 10.1109/ICCV51070.2023.01465]

Kang L W, Lin C W and Fu Y H. 2012. Automatic single-image-based rain streaks removal via image decomposition. IEEE Transactions on Image Processing, 21(4): 1742-1755 [DOI: 10.1109/TIP.2011.2179057]

Kang M, Zhu J Y, Zhang R, Park J, Shechtman E, Paris S and Park T. 2023. Scaling up GANs for text-to-image synthesis//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 10124-10134 [DOI: 10.1109/cvpr52729.2023.00976]

Karnewar A, Mitra N J, Vedaldi A and Novotny D. 2023a. HoloFusion: towards photo-realistic 3D generative modeling//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22919-22928 [DOI: 10.1109/ICCV51070.2023.02100]

Karnewar A and Wang O. 2020. MSG-GAN: multi-scale gradients for generative adversarial networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7796-7805 [DOI: 10.1109/CVPR42600.2020.00782]

Karnewar A, Vedaldi A, Novotny D and Mitra N J. 2023b. HOLODIFFUSION: training a 3D diffusion model using 2D images//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 18423-18433 [DOI: 10.1109/CVPR52729.2023.01767]

Karras T, Aittala M, Laine S, Härkönen E, Hellsten J, Lehtinen J and Aila T. 2021. Alias-free generative adversarial networks//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual: Curran Associates Inc.: 852-863

Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(12): 4217-4228 [DOI: 10.1109/TPAMI.2020.2970919]

Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J and Aila T. 2020. Analyzing and improving the image quality of StyleGAN//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8107-8116 [DOI: 10.1109/CVPR42600.2020.00813]

Kaspar M, Muñoz Osorio J D and Bock J. 2020. Sim2Real transfer for reinforcement learning without dynamics randomization//Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas, USA: IEEE: 4383-4388 [DOI: 10.1109/IROS45743.2020.9341260]

Katara P, Xian Z and Fragkiadaki K. 2024. Gen2Sim: scaling up robot learning in simulation with generative models//Proceedings of 2024 IEEE International Conference on Robotics and Automation. Yokohama, Japan: IEEE: 6672-6679 [DOI: 10.1109/ICRA57147.2024.10610566]

Kerbl B, Kopanas G, Leimkühler T and Drettakis G. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4): #139 [DOI: 10.1145/3592433]

Kerbl B, Meuleman A, Kopanas G, Wimmer M, Lanvin A and Drettakis G. 2024. A hierarchical 3D Gaussian representation for real-time rendering of very large datasets. ACM Transactions on Graphics, 43(4): #62 [DOI: 10.1145/3658160]

Kerim A, Aslan C, Celikcan U, Erdem E and Erdem A. 2021. NOVA: rendering virtual worlds with humans for computer vision tasks. Computer Graphics Forum, 40(6): 258-272 [DOI: 10.1111/cgf.14271]

Khachatryan L, Movsisyan A, Tadevosyan V, Henschel R, Wang Z Y, Navasardyan S and Shi H. 2023. Text2Video-zero: text-to-image diffusion models are zero-shot video generators//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 15908-15918 [DOI: 10.1109/ICCV51070.2023.01462]

Kim H, Cui X N, Kim M G and Nguyen T H B. 2019. Fingerprint generation and presentation attack detection using deep neural networks//Proceedings of 2019 IEEE Conference on Multimedia Information Processing and Retrieval. San Jose, USA: IEEE: 375-378

[DOI: 10.1109/MIPR.2019.00074]

Kim I H, Lee J, Jin W, Son S, Cho K, Seo J, Kwak M S, Cho S, Baek J, Lee B and Kim S. 2024. Pose-dIVE: pose-diversified augmentation with diffusion model for person re-identification [EB/OL]. [2025-02-28]. http://arxiv.org/pdf/2406.16042.pdf

Kim J, Lee J K and Lee K M. 2016. Accurate image super-resolution using very deep convolutional networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1646-1654 [DOI: 10.1109/CVPR.2016.182]

Kim M, Liu F, Jain A and Liu X M. 2023. DCFace: synthetic face generation with dual condition diffusion model//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 12715-12725 [DOI: 10.1109/CVPR52729.2023.01223]

Kingma D P and Dhariwal P. 2018. Glow: generative flow with invertible 1 × 1 convolutions//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 10236-10245

Kingma D P and Welling M. 2013. Auto-encoding variational bayes [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1312.6114.pdf

Klein L and Noé F. 2025. Transferable Boltzmann generators [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2406.14426.pdf

Kohli N, Yadav D, Vatsa M, Singh R and Noore A. 2017. Synthetic iris presentation attack using iDCGAN//Proceedings of 2017 IEEE International Joint Conference on Biometrics. Denver, USA: IEEE: 674-680 [DOI: 10.1109/BTAS.2017.8272756]

Kolotouros N, Alldieck T, Zanfir A, Bazavan E G, Fieraru M and Sminchisescu C. 2023. DreamHuman: animatable 3D avatars from text//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 10516-10529

Kolve E, Mottaghi R, Han W, VanderBilt E, Weihs L, Herrasti A, Deitke M, Ehsani K, Gordon D, Zhu Y K, Kembhavi A, Gupta A and Farhadi A. 2022. AI2-THOR: an interactive 3D environment for visual AI [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1712.05474.pdf

Kondapaneni N, Marks M, Knott M, Guimaraes R and Perona P. 2024. Text-image alignment for diffusion-based perception//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13883-13893 [DOI: 10.1109/CVPR52733.2024.01317]

Kong W J, Tian Q, Zhang Z J, Min R, Dai Z Z, Zhou J, Xiong J F, Li X, Wu B, Zhang J W, Wu K, Lin Q, Yuan J K, Long Y X, Wang A, Wang A D, Li C L, Huang D J, Yang F, Tan H, Wang H M, Song J, Bai J W, Wu J B, Xue J B, Wang J, Wang K, Liu M Y, Li P Y, Li S, Wang W Y, Yu W Q, Deng X C, Li Y, Chen Y, Cui Y T, Peng Y B, Yu Z T, He Z Y, Xu Z Y, Zhou Z X, Xu Z N, Tao Y Y, Lu Q L, Liu S T, Zhou D, Wang H F, Yang Y, Wang D, Liu Y H, Jiang J and Zhong C. 2025. HunyuanVideo: a systematic framework for large video generative models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2412.03603.pdf

Kosiorek A R, Strathmann H, Zoran D, Moreno P, Schneider R, Mokrá S and Rezende D J. 2021. NeRF-VAE: a geometry aware 3D scene generative model//Proceedings of the 38th International Conference on Machine Learning. Virtual: OpenReview.net : 5742-5752

Kulikov V, Yadin S, Kleiner M and Michaeli T. 2023. SinDDM: a single image denoising diffusion model//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 17920-17930

Lazaridis L, Dimou A and Daras P. 2018. Abnormal behavior detection in crowded scenes using density heatmaps and optical flow//Proceedings of the 26th European Signal Processing Conference. Rome, Italy: IEEE: 2060-2064 [DOI: 10.23919/EUSIPCO.2018.8553620]

Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z H and Shi W Z. 2016. Photo-realistic single image super-resolution using a generative adversarial network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 105-114 [DOI: 10.1109/CVPR.2017.19]

Lee D, Kim C, Kim S, Cho M and Han W S. 2022. Autoregressive image generation using residual quantization//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11513-11522 [DOI: 10.1109/CVPR52688.2022.01123]

Li B H, Zhou H, He J X, Wang M X, Yang Y M and Li L. 2020. On the sentence embeddings from pre-trained language models//Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. Online: Association for Computational Linguistics: 9119-9130 [DOI: 10.18653/v1/2020.emnlp-main.733]

Li H and Wu X J. 2019. DenseFuse: a fusion approach to infrared and visible images. IEEE Transactions on Image Processing, 28(5): 2614-2623 [DOI: 10.1109/TIP.2018.2887342]

Li H, Ye M and Du B. 2021a. WePerson: learning a generalized re-identification model from all-weather virtual data//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: Association for Computing Machinery: 3115-3123 [DOI: 10.1145/3474085.3475455]

Li J H, Zhang J W, Bai X, Zheng J, Ning X, Zhou J and Gu L. 2024b. DNGaussian: optimizing sparse-view 3D Gaussian radiance fields with global-local depth normalization//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 20775-20785 [DOI: 10.1109/CVPR52733.2024.01963]

Li K L, Wang J B, Yang L X, Lu C W and Dai B. 2024c. SemGrasp: semantic grasp generation via language aligned discretization [EB/OL]. [2025-04-22]. https://arxiv.org/pdf/2404.03590.pdf

Li L Y, Tang J S, Shao Z W, Tan X and Ma L Z. 2022. Sketch-to-photo face generation based on semantic consistency preserving and similar connected component refinement. The Visual Computer, 38(11): 3577-3594 [DOI: 10.1007/s00371-021-02188-1]

Li P X, Chen K, Liu Z L, Gao R Y, Hong L Q, Zhou G, Yao H, Yeung D Y, Lu H C and Jia X. 2024a. TrackDiffusion: tracklet-conditioned video generation via diffusion models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2312.00651.pdf

Li R T, Cheong L F and Tan R T. 2019a. Heavy rain image restoration: integrating physics model and conditional adversarial learning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1633-1642 [DOI: 10.1109/CVPR.2019.00173]

Li T H, Tian Y L, Li H, Deng M Y and He K M. 2024d. Autoregressive image generation without vector quantization [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2406.11838.pdf

Li X, Chu W Q, Wu Y, Yuan W H, Liu F L, Zhang Q, Li F, Feng H C, Ding E R and Wang J D. 2023b. Videogen: a reference-guided latent diffusion approach for high definition text-to-video generation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2309.00398.pdf

Li Y, Tan R T, Guo X J Lu J B and Brown M S. 2016. Rain streak removal using layer priors//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2736-2744 [DOI: 10.1109/CVPR.2016.299]

Li Y H, Chen X J, Wu F and Zha Z J. 2019b. LinesToFacePhoto: face photo generation from lines with conditional self-attention generative adversarial networks//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: Association for Computing Machinery: 2323-2331 [DOI: 10.1145/3343031.3350854]

Li Y H, Liu H T, Wu Q Y, Mu F Z, Yang J W, Gao J F, Li C Y and Lee Y J. 2023c. GLIGEN: open-set grounded text-to-image generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 22511-22521 [DOI: 10.1109/CVPR52729.2023.02156]

Li Y X, Jiang L H, Xu L N, Xiangli Y B, Wang Z Z, Lin D H and Dai B. 2023d. MatrixCity: a large-scale city dataset for city-scale neural rendering and beyond//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3182-3192 [DOI: 10.1109/ICCV51070.2023.00297]

Li Z, Li Y X, Zhao P H, Song R J, Li X and Yang J. 2023e. Is synthetic data from diffusion models ready for knowledge distillation? [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2305.12954.pdf

Li Z M, Cheng T H, Chen S F, Sun P Z, Shen H C, Ran L J, Chen X X, Liu W Y and Wang X G. 2024e. ControlAR: controllable image generation with autoregressive models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2410.02705.pdf

Li Z Q, Yu T W, Sang S, Wang S, Song M, Liu Y H, Yeh Y Y, Zhu R, Gundavarapu N, Shi J, Bi S, Xu Z X, Yu H X, Sunkavalli K, Hašan M, Ramamoorthi R and Chandraker M. 2021b. OpenRooms: an end-to-end open framework for photorealistic indoor scene datasets [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2007.12868.pdf

Li Z Y, Zhou Q Y, Zhang X Y, Zhang Y, Wang Y F and Xie W D. 2023f. Open-vocabulary object segmentation with diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 7633-7642 [DOI: 10.1109/ICCV51070.2023.00705]

Lian L, Shi B F, Yala A, Darrell T and Li B Y. 2024. LLM-grounded video diffusion models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2309.17444.pdf

Liang J Y, Zhang K, Gu S H, van Gool L and Timofte R. 2021. Flow-based kernel prior with application to blind super-resolution//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10596-10605 [DOI: 10.1109/CVPR46437.2021.01046]

Liang P W, Jiang J J, Liu X M and Ma J Y. 2022. Fusion from decomposition: a self-supervised decomposition approach for image fusion//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 719-735 [DOI: 10.1007/978-3-031-19797-0_41]

Liang W Q, Wang G C, Lai J H and Zhu J Y. 2018. M2M-GAN: many-to-many generative adversarial transfer learning for person re-identification [EB/OL]. [2025-02-28]. http://arxiv.org/pdf/1811.03768.pdf

Liang Y X, Yang X, Lin J T, Li H D, Xu X G and Chen Y C. 2024a. LucidDreamer: towards high-fidelity text-to-3D generation via interval score matching//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, Washington, USA: IEEE: 6517-6526 [DOI: 10.1109/CVPR52733.2024.00623]

Liang Y X, Yang X, Lin J T, Li H D, Xu X G and Chen Y C. 2024b. LucidDreamer: towards high-fidelity text-to-3D generation via interval score matching//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6517-6526 [DOI: 10.1109/CVPR52733.2024.00623]

Liao T T, Yi H W, Xiu Y L, Tang J X, Huang Y Y, Thies J and Black M J. 2024. TADA! Text to animatable digital avatars//Proceedings of 2024 International Conference on 3D Vision. Davos, Switzerland: IEEE: 1508-1519 [DOI: 10.1109/3DV62453.2024.00150]

Lin B, Ge Y Y, Cheng X H, Li Z J, Zhu B, Wang S D, He X Y, Ye Y, Yuan S H, Chen L H, Jia T H, Zhang J W, Tang Z Y, Pang Y T, She B, Yan C, Hu Z H, Dong X Y, Chen L, Pan Z, Zhou X, Dong S L, Tian Y H and Yuan L. 2024a. Open-sora plan: open-source large video generation model [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2412.00131.pdf

Lin C H, Gao J, Tang L M, Takikawa T, Zeng X H, Huang X, Kreis K, Fidler S, Liu M Y and Lin T Y. 2023a. Magic3D: high-

resolution text-to-3D content creation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 300-309 [DOI: 10.1109/CVPR52729.2023.00037]

Lin H, Zala A, Cho J and Bansal M. 2024c. VideoDirectorGPT: consistent multi-scene video generation via LLM-guided planning [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2309.15091.pdf

Lin K E, Lin Y C, Lai W S, Lin T Y, Shih Y C and Ramamoorthi R. 2023c. Vision transformer for NeRF-based view synthesis from a single input image//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 806-815 [DOI: 10.1109/WACV56688.2023.00087]

Lin W, Gao J Y, Wang Q and Li X L. 2021a. Learning to detect anomaly events in crowd scenes from synthetic data. Neurocomputing, 436: 248-259 [DOI: 10.1016/j.neucom.2021.01.031]

Lin X M, Li Y K, Hsiao J, Ho C and Kong Y. 2023b. Catch missing details: image reconstruction with frequency augmented variational autoencoder//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1736-1745 [DOI: 10.1109/CVPR52729.2023.00173]

Lin X Q, He J W, Chen Z Y, Lyu Z Y, Dai B, Yu F H, Ouyang W L, Qiao Y and Dong C. 2024b. DiffBIR: towards blind image restoration with generative diffusion prior [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2308.15070.pdf

Lin Z C, Liu C Y, Qi W B and Chan S C. 2021b. A color/illuminance aware data augmentation and style adaptation approach to person re-identification. IEEE Access, 9: 115826-115838 [DOI: 10.1109/ACCESS.2021.3100571]

Liu J, Rahmani H, Akhtar N and Mian A. 2019a. Learning human pose models from synthesized data for robust RGB-D action recognition. International Journal of Computer Vision, 127(10): 1545-1564 [DOI: 10.1007/s11263-019-01192-2]

Liu J W, Wang Q, Fan H J, Wang Y N, Tang Y D and Qu L Q. 2024a. Residual denoising diffusion models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2773-2783 [DOI: 10.1109/CVPR52733.2024.00268]

Liu R S, Ma L, Zhang J A, Fan X and Luo Z X. 2021. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10556-10565 [DOI: 10.1109/CVPR46437.2021.01042]

Liu R S, Wu R D, van Hoorick B, Tokmakov P, Zakharov S and Vondrick C. 2023. Zero-1-to-3: zero-shot one image to 3D object//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 9264-9275 [DOI: 10.1109/ICCV51070.2023.00853]

Liu S L, Zeng Z Y, Ren T H, Li F, Zhang H, Yang J, Jiang Q, Li C Y, Yang J W, Su H, Zhu J and Zhang L. 2024e. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 38-55 [DOI: 10.1007/978-3-031-72970-6_3]

Liu T Q, Wang G C, Hu S K, Shen L, Ye X Y, Zang Y H, Cao Z G, Li W and Liu Z W. 2025. MVSGaussian: fast generalizable Gaussian splatting reconstruction from multi-view stereo//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 37-53 [DOI: 10.1007/978-3-031-72649-1_3]

Liu W, Piao Z X, Min J, Luo W H, Ma L and Gao S H. 2019b. Liquid warping GAN: a unified framework for human motion imitation, appearance transfer and novel view synthesis//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5903-5912 [DOI: 10.1109/ICCV.2019.00600]

Liu X, Ren J, Siarohin A, Skorokhodov I, Li Y Y, Lin D H, Liu X H, Liu Z W and Tulyakov S. 2024d. HyperHuman: hyper-realistic human generation with latent structural diffusion [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2310.08579.pdf

Liu X C, Gong C Y and Liu Q. 2022. Flow straight and fast: learning to generate and transfer data with rectified flow [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2209.03003.pdf

Liu Y, Lin C, Zeng Z J, Long X X, Liu L J, Komura T and Wang W P. 2024c. Syncdreamer: generating multiview-consistent images from a single-view image [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2309.03453.pdf

Liu Y H, Ke Z H, Liu F, Zhao N X and Lau R W H. 2024b. Diff-Plugin: revitalizing details for diffusion-based low-level tasks//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4197-4208 [DOI: 10.1109/CVPR52733.2024.00402]

Liu Y X, Gao C, Zhang Z L, Wu Y H, Liang M X, Tao L and Lu Y X. 2017. A new multi-agent system to simulate the foraging behaviors of Physarum. Natural Computing, 16(1): 15-29 [DOI: 10.1007/s11047-015-9530-5]

Long X X, Guo Y C, Lin C, Liu Y, Dou Z Y, Liu L J, Ma Y X, Zhang S H, Habermann M, Theobalt C and Wang W P. 2024. Wonder3D: single image to 3D using cross-domain diffusion//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9970-9980 [DOI: 10.1109/CVPR52733.2024.00951]

Loper M, Mahmood N and Black M J. 2014. MoSh: motion and shape capture from sparse markers. ACM Transactions on Graphics, 33(6): #220 [DOI: 10.1145/2661229.2661273]

Lore K G, Akintayo A and Sarkar S. 2017. LLNet: a deep autoencoder approach to natural low-light image enhancement. Pattern Recognition, 61: 650-662 [DOI: 10.1016/j.patcog.2016.06.008]

Lu C and Song Y. 2025. Simplifying, stabilizing and scaling continuous-time consistency models [EB/OL]. [2025-02-28].

https://arxiv.org/pdf/2410.11081.pdf

Lu C, Zhou Y H, Bao F, Chen J F, Li C X and Zhu J. 2022. DPM-Solver: a fast ODE solver for diffusion probabilistic model sampling in around 10 steps//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 5775-5787

Lu T, Yu M L, Xu L N, Xiangli Y B, Wang L M, Lin D H and Dai B. 2024. Scaffold-GS: structured 3D Gaussians for view-adaptive rendering//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 20654-20664 [DOI: 10.1109/CVPR52733.2024.01952]

Lugmayr A, Danelljan M, van Gool L and Timofte R. 2020. SRFlow: learning the super-resolution space with normalizing flow//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 715-732 [DOI: 10.1007/978-3-030-58558-7_42]

Luo Y, Xu Y and Ji H. 2015. Removing rain from a single image via discriminative sparse coding//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 3397-3405 [DOI: 10.1109/ICCV.2015.388]

Luo Z X, Chen D Y, Zhang Y Y, Huang Y, Wang L, Shen Y J, Zhao D L, Zhou J R and Tan T N. 2023. Notice of removal: VideoFusion: decomposed diffusion models for high-quality video generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 10209-10218 [DOI: 10.1109/CVPR52729.2023.00984]

Lyu F J and Nevatia R. 2007. Single view human action recognition using key pose matching and Viterbi path searching//Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE: 1-8 [DOI: 10.1109/CVPR.2007.383131]

Ma X, Wang Y H, Jia G Y, Chen X Y, Liu Z W, Li Y F, Chen C J and Qiao Y. 2024. Latte: latent diffusion transformer for video generation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2401.03048.pdf

Maeda S. 2020. Unpaired image super-resolution using pseudo-supervision//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 288-297 [DOI: 10.1109/CVPR42600.2020.00037]

Mai A, Hedman P, Kopanas G, Verbin D, Futschik D, Xu Q G, Kuester F, Barron J T and Zhang Y D. 2024. EVER: exact volumetric ellipsoid rendering for real-time view synthesis [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2410.01804.pdf

Makthal S and Ross A. 2005. Synthesis of iris images using Markov random fields//Proceedings of the 13th European Signal Processing Conference. Antalya, Turkey: IEEE: 1-4

Mallick S S, Goel R, Kerbl B, Steinberger M, Carrasco F V and De La Torre F. 2024. Taming 3DGS: high-quality radiance fields with limited resources//Proceedings of 2024 SIGGRAPH Asia Conference Papers. Tokyo, Japan: Association for Computing Machinery: #2 [DOI: 10.1145/3680528.3687694]

Maltoni D, Maio D, Jain A K and Prabhakar S. 2009. Synthetic fingerprint generation//Handbook of Fingerprint Recognition. 2nd ed. London: Springer: 271-302 [DOI: 10.1007/978-1-84882-254-2_6]

Matas J, James S and Davison A J. 2018. Sim-to-real reinforcement learning for deformable object manipulation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1806.07851.pdf

McLaughlin N, Del Rincon J M and Miller P. 2015. Data-augmentation for reducing dataset bias in person re-identification//Proceedings of the 12th IEEE International Conference on Advanced Video and Signal Based Surveillance. Karlsruhe, Germany: IEEE: 1-6 [DOI: 10.1109/AVSS.2015.7301739]

Menapace W, Siarohin A, Skorokhodov I, Deyneka E, Chen T S, Kag A, Fang Y W, Stoliar A, Ricci E, Ren J and Tulyakov S. 2024. Snap video: scaled spatiotemporal transformers for text-to-video synthesis//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7038-7048 [DOI: 10.1109/CVPR52733.2024.00672]

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2021. NeRF: representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1): 99-106 [DOI: 10.1145/3503250]

Minaee S and Abdolrashidi A. 2018a. Iris-GAN: learning to generate realistic iris images using convolutional GAN [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1812.04822.pdf

Minaee S and Abdolrashidi A. 2018b. Finger-GAN: generating realistic fingerprint images using connectivity imposed GAN [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1812.10482.pdf

Minaee S, Minaei M and Abdolrashidi A. 2020. Palm-GAN: generating realistic palmprint images using total-variation regularized GAN [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2003.10834.pdf

Mirza M and Osindero S. 2014. Conditional generative adversarial nets [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1411.1784.pdf

Mistry V, Engelsma J J and Jain A K. 2020. Fingerprint synthesis: search with 100 million prints//Proceedings of 2020 IEEE International Joint Conference on Biometrics. Houston, USA: IEEE: 1-10 [DOI: 10.1109/IJCB48548.2020.9304885]

Mittal G, Marwah T and Balasubramanian V N. 2017. Sync-DRAW: automatic video generation using deep recurrent attentive architectures//Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, USA: ACM: 1096-1104 [DOI: 10.1145/3123266.3123309]

Mo K C, Zhu S L, Chang A X, Yi L, Tripathi S, Guibas L J and Su H. 2019. PartNet: a large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 909-918 [DOI: 10.1109/CVPR.2019.00100]

Montulet R and Briassouli A. 2021. Densely annotated photorealistic virtual dataset generation for abnormal event detection//Proceedings of the Pattern Recognition. ICPR International Workshops and Challenges. Virtual Event: Springer: 5-19 [DOI: 10.1007/978-3-030-68799-1_1]

Mou C, Wang X T, Xie L B, Wu Y Z, Zhang J, Qi Z A and Shan Y. 2024. T2I-Adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 4296-4304 [DOI: 10.1609/aaai.v38i5.28226]

Müller N, Siddiqui Y, Porzi L, BulòuS R, Kontschieder P and Nießner M. 2023. DiffRF: rendering-guided 3D radiance field diffusion//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 4328-4338 [DOI: 10.1109/CVPR52729.2023.00421]

Müller T, Evans A, Schied C and Keller A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics, 41(4): #102 [DOI: 10.1145/3528223.3530127]

Navaneet K L, Pourahmadi Meibodi K, Abbasi Koohpayegani S and Pirsiavash H. 2025. CompGS: smaller and faster Gaussian splatting with vector quantization//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 330-349 [DOI: 10.1007/978-3-031-73411-3_19]

Nguyen T D, Le T, Vu H and Phung D. 2017. Dual discriminator generative adversarial nets//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 2667-2677

Niedermayr S, Stumpfegger J and Westermann R. 2024. Compressed 3D Gaussian splatting for accelerated novel view synthesis//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10349-10358 [DOI: 10.1109/CVPR52733.2024.00985]

Niemeyer M and Geiger A. 2021. GIRAFFE: representing scenes as compositional generative neural feature fields//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 11448-11459 [DOI: 10.1109/CVPR46437.2021.01129]

Niu K, Yu H Y, Qian X L, Fu T, Li B and Xue X Y. 2024. Synthesizing efficient data with diffusion models for person re-identification pre-training [EB/OL]. [2025-02-28]. http://arxiv.org/pdf/2406.06045.pdf

Noghre G A, Danesh Pazho A, Sanchez J, Hewitt N, Neff C and Tabkhi H. 2022. ADG-Pose: automated dataset generation for real-world human pose estimation//Proceedings of the 3rd International Conference on Pattern Recognition and Artificial Intelligence. Paris, France: Springer-Verlag: 258-270 [DOI: 10.1007/978-3-031-09282-4_22]

Noguchi A, Sun X, Lin S and Harada T. 2022. Unsupervised learning of efficient geometry-aware neural articulated representations//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 597-614 [DOI: 10.1007/978-3-031-19790-1_36]

Ostrek M, O'Sullivan C, Black M J and Thies J. 2024. Synthesizing environment-specific people in photographs [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2312.14579.pdf

Ou W F, Po L M, Zhou C, Xian P F and Xiong J J. 2022. GAN-based inter-class sample generation for contrastive learning of vein image representations. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(2): 249-262 [DOI: 10.1109/TBIOM.2022.3152345]

Özdenizci O and Legenstein R. 2023. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(8): 10346-10357 [DOI: 10.1109/TPAMI.2023.3238179]

Pan X G, Tewari A, Leimkühler T, Liu L J, Meka A and Theobalt C. 2023. Drag your GAN: interactive point-based manipulation on the generative image manifold//Proceedings of 2023 ACM SIGGRAPH Conference. Los Angeles, USA: ACM: #78 [DOI: 10.1145/3588432.3591500]

Panev S, Kim E, Si Namburu S A, Nikolova D, De Melo C, de la Torre F and Hodgins J. 2024. Exploring the impact of rendering method and motion quality on model performance when using multi-view synthetic data for action recognition//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 4580-4590 [DOI: 10.1109/WACV57701.2024.00453]

Pang Z Q, Guo J F, Sun W B, Xiao Y B and Yu M. 2022. Cross-domain person re-identification by hybrid supervised and unsupervised learning. Applied Intelligence, 52(3): 2987-3001 [DOI: 10.1007/s10489-021-02551-8]

Papamakarios G, Pavlakou T and Murray I. 2017. Masked autoregressive flow for density estimation//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 2335-2344

Patel P, Huang C H P, Tesch J, Hoffmann D T, Tripathi S and Black M J. 2021. AGORA: avatars in geography optimized for regression analysis//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13463-13473 [DOI: 10.1109/CVPR46437.2021.01326]

Peng D, Hu P, Ke Q H and Liu J. 2023. Diffusion-based image translation with label guidance for domain adaptive semantic segmentation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 808-820 [DOI: 10.1109/ICCV51070.2023.00081]

Perarnau G, van de Weijer J, Raducanu B and Álvarez J M. 2016. Invertible conditional GANs for image editing [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1611.06355.pdf

Perlin K. 1985. An image synthesizer. ACM SIGGRAPH Computer

Graphics, 19(3): 287-296 [DOI: 10.1145/325165.325247]

Petrovich M, Black M J and Varol G. 2022. TEMOS: generating diverse human motions from textual descriptions//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 480-497 [DOI: 10.1007/978-3-031-20047-2_28]

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]. [2025-02-28]. https://arxiv.org/pdf/2307.01952.pdf

Polyak A, Zohar A, Brown A, Tjandra A, Sinha A, Lee A, Vyas A, Shi B W, Ma C Y, Chuang C Y, Yan D, Choudhary D, Wang D K, Sethi G, Pang G, Ma H Y, Misra I, Hou J, Wang J L, Jagadeesh K, Li K P, Zhang L X, Singh M, Williamson M, Le M, Yu M, Singh M K, Zhang P Z, Vajda P, Duval Q, Girdhar R, Sumbaly R, Rambhatla S S, Tsai S, Azadi S, Datta S, Chen S Y, Bell S, Ramaswamy S, Sheynin S, Bhattacharya S, Motwani S, Xu T, Li T H, Hou T B, Hsu W N, Yin X, Dai X L, Taigman Y, Luo Y Q, Liu Y C, Wu Y C, Zhao Y, Kirstain Y, He Z C, He Z J, Pumarola A, Thabet A, Sanakoyeu A, Mallya A, Guo B S, Araya B, Kerr B, Wood C, Liu C, Peng C, Vengertsev D, Schonfeld E, Blanchard E, Juefei-Xu F, Nord F, Liang J, Hoffman J, Kohler J, Fire K, Sivakumar K, Chen L, Yu L C, Gao L Y, Georgopoulos M, Moritz R, Sampson S K, Li S K, Parmeggiani S, Fine S, Fowler T, Petrovic V and Du Y M. 2025. Movie gen: a cast of media foundation models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2410.13720.pdf

Poole B, Jain A, Barron J T and Mildenhall B. 2022. DreamFusion: text-to-3D using 2D diffusion [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2209.14988.pdf

Postels J, Danelljan M, Van Gool L and Tombari F. 2022. ManiFlow: implicitly representing manifolds with normalizing flows [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2208.08932.pdf

Prabhakar K R, Srikar V S and Babu R V. 2017. DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4724-4732 [DOI: 10.1109/ICCV.2017.505]

Priesnitz J, Rathgeb C, Buchmann N and Busch C. 2022. SynCoLFinGer: synthetic contactless fingerprint generator. Pattern Recognition Letters, 157: 127-134 [DOI: 10.1016/j.patrec.2022.04.003]

Puig X, Ra K, Boben M, Li J M, Wang T W, Fidler S and Torralba A. 2018. VirtualHome: simulating household activities via programs//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8494-8502 [DOI: 10.1109/CVPR.2018.00886]

Qian G C, Mai J J, Hamdi A, Ren J, Siarohin A, Li B, Lee H Y, Skorokhodov I, Wonka P, Tulyakov S and Ghanem B. 2023. Magic123: one image to high-quality 3D object generation using both 2D and 3D diffusion priors [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2306.17843.pdf

Qian R, Tan R T, Yang W H, Su J J and Liu J Y. 2018a. Attentive generative adversarial network for raindrop removal from a single image//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2482-2491 [DOI: 10.1109/CVPR.2018.00263]

Qian X L, Fu Y W, Xiang T, Wang W X, Qiu J, Wu Y, Jiang Y G and Xue X Y. 2018b. Pose-normalized image generation for person re-identification//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer International Publishing: 661-678 [DOI: 10.1007/978-3-030-01240-3_40]

Qiu L T, Chen G Y, Gu X D, Zuo Q, Xu M T, Wu Y S, Yuan W H, Dong Z L, Bo L F and Han X G. 2024. RichDreamer: a generalizable normal-depth diffusion model for detail richness in text-to-3D//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9914-9925 [DOI: 10.1109/CVPR52733.2024.00946]

Qiu W C, Zhong F W, Zhang Y, Qiao S Y, Xiao Z H, Kim T S and Wang Y Z. 2017. UnrealCV: virtual worlds for computer vision//Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, USA: Association for Computing Machinery: 1221-1224 [DOI: 10.1145/3123266.3129396]

Qu L H, Liu S L, Wang M N and Song Z J. 2022. TransMEF: a transformer-based multi-exposure image fusion framework using self-supervised multi-task learning//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtually: AAAI: 2126-2134 [DOI: 10.1609/aaai.v36i2.20109]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual: OpenReview.net: 8748-8763

Radford A, Metz L and Chintala S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks [EB/OL]. [2025-04-22]. https://arxiv.org/pdf/1511.06434.pdf

Radl L, Steiner M, Parger M, Weinrauch A, Kerbl B and Steinberger M. 2024. StopThePop: sorted Gaussian splatting for view-consistent real-time rendering. ACM Transactions on Graphics, 43(4): #64 [DOI: 10.1145/3658187]

Rahmani H and Mian A. 2015. Learning a non-linear knowledge transfer model for cross-view action recognition//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 2458-2466 [DOI: 10.1109/CVPR.2015.7298860]

Rahmani H and Mian A. 2016. 3D action recognition from novel viewpoints//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1506-1515 [DOI: 10.1109/CVPR.2016.167]

Rahmani H, Mian A and Shah M. 2018. Learning a deep model for human action recognition from novel viewpoints. IEEE Transactions

on Pattern Analysis and Machine Intelligence, 40(3): 667-681 [DOI: 10.1109/TPAMI.2017.2691768]

Raistrick A, Lipson L, Ma Z Y, Mei L J, Wang M Z, Zuo Y M, Kayan K, Wen H Y, Han B N, Wang Y H, Newell A, Law H, Goyal A, Yang K Y and Deng J. 2023. Infinite photorealistic worlds using procedural generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 12630-12641 [DOI: 10.1109/CVPR52729.2023.01215]

Raistrick A, Mei L J, Kayan K, Yan D, Zuo Y M, Han B N, Wen H Y, Parakh M, Alexandropoulos S, Lipson L, Ma Z Y and Deng J. 2024. Infinigen indoors: photorealistic indoor scenes using procedural generation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 21783-21794 [DOI: 10.1109/CVPR52733.2024.02058]

Ramesh A, Dhariwal P, Nichol A, Chu C and Chen M. 2022. Hierarchical text-conditional image generation with CLIP latents [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2204.06125.pdf

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M and Sutskever I. 2021. Zero-shot text-to-image generation//Proceedings of the 38th International Conference on Machine Learning. Virtual: OpenReview.net: 8821-8831

Reimers N and Gurevych I. 2019. Sentence-BERT: sentence embeddings using siamese BERT-networks//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: Association for Computational Linguistics: 3982-3992 [DOI: 10.18653/v1/D19-1410]

Rempe D, Philion J, Guibas L J, Fidler S and Litany O. 2022. Generating useful accident-prone driving scenarios via a learned traffic prior//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 17284-17294 [DOI: 10.1109/CVPR52688.2022.01679]

Rezende D J and Mohamed S. 2015. Variational inference with normalizing flows//Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR.org: 1530-1538

Riazi M S, Chavoshian S M and Koushanfar F. 2020. SynFi: automatic synthetic fingerprint generation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2002.08900.pdf

Richardson E, Metzer G, Alaluf Y, Giryes R and Cohen-Or D. 2023. TEXTure: text-guided texturing of 3D shapes//Proceedings of 2023 SIGGRAPH'23: ACM SIGGRAPH Conference Proceedings. Los Angeles, USA: ACM: #54 [DOI: 10.1145/3592410]

Richter S R, Vineet V, Roth S and Koltun V. 2016. Playing for data: ground truth from computer games//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer International Publishing: 102-118 [DOI: 10.1007/978-3-319-46475-6_7]

Roberts M, Ramapuram J, Ranjan A, Kumar A, Bautista M A, Paczan N, Webb R and Susskind J M. 2021. Hypersim: a photorealistic synthetic dataset for holistic indoor scene understanding//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10892-10902 [DOI: 10.1109/ICCV48922.2021.01073]

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]

Ronneberger O, Fischer P and Brox T. 2015. U-net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Ros G, Sellart L, Materzynska J, Vazquez D and Lopez A M. 2016. The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3234-3243 [DOI: 10.1109/CVPR.2016.352]

Rumelhart D E, Hinton G E and Williams R J. 1986. Learning representations by back-propagating errors. Nature, 323(6088): 533-536 [DOI:10.1038/323533a0]

Runions A, Fuhrer M, Lane B, Federl P, Rolland-Lagan A G and Prusinkiewicz P. 2005. Modeling and visualization of leaf venation patterns. ACM Transactions on Graph, 24(3): 702-711 [DOI: 10.1145/1073204.1073251]

Saharia C, Chan W, Saxena S, Li L L, Whang J, Denton E, Ghasemipour S K S, Ayan B K, Mahdavi S S, Gontijo-Lopes R, Salimans T, Ho J, Fleet D J and Norouzi M. 2022. Photorealistic text-to-image diffusion models with deep language understanding//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 36479-36494 [DOI: 10.5555/3600270.3602913]

Salazar E, Hernández-García R, Barrientos R J, Vilches K, Mora M and Vásquez A. 2021a. Automatic generation of synthetic palm vein images: a nature-based approach//Proceedings of the 11th International Conference of Pattern Recognition Systems. Online: IEEE: 38-43 [DOI: 10.1049/icp.2021.1452]

Salazar E, Hernández-García R, Barrientos R J, Vilches K, Mora M and Vásquez A. 2021b. Generating style-based palm vein synthetic images for the creation of large-scale datasets//Proceedings of the 11th International Conference of Pattern Recognition Systems. Online: IEEE: 182-187 [DOI: 10.1049/icp.2021.1451]

Salem Hussin S H and Yildirim R. 2021. StyleGAN-LSRO method for person re-identification. IEEE Access, 9: 13857-13869 [DOI: 10.1109/ACCESS.2021.3051723]

Sams A, Shomee H H and Rahman S M M. 2022. HQ-finGAN: high-quality synthetic fingerprint generation using GANs. Circuits, Systems, and Signal Processing, 41(11): 6354-6369 [DOI: 10.1007/

s00034-022-02089-1］

Sarıyıldız M B, Alahari K, Larlus D and Kalantidis Y. 2023. Fake it till you make it: learning transferable representations from synthetic ImageNet clones//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 8011-8021［DOI: 10.1109/CVPR52729.2023.00774］

Sauer A, Karras T, Laine S, Geiger A and Aila T. 2023. StyleGAN-T: unlocking the power of gans for fast large-scale text-to-image synthesis//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR. org: 30105-30118［DOI: 10.5555/3618408.3619658］

Sauer A, Schwarz K and Geiger A. 2022. StyleGAN-XL: scaling stylegan to large diverse datasets//Proceedings of 2022 ACM SIGGRAPH Conference Proceedings. Vancouver, Canada: ACM: #49［DOI: 10.1145/3528233.3530738］

Savva M, Kadian A, Maksymets O, Zhao Y L, Wijmans E, Jain B, Straub J, Liu J, Koltun V, Malik J, Parikh D and Batra D. 2019. Habitat: a platform for embodied AI research//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9338-9346［DOI: 10.1109/ICCV.2019.00943］

Schröder G, Senst T, Bochinski E and Sikora T. 2018. Optical flow dataset and benchmark for visual crowd analysis//Proceedings of the 15th IEEE International Conference on Advanced Video and Signal Based Surveillance. Auckland, New Zealand: IEEE: 1-6［DOI: 10.1109/AVSS.2018.8639113］

Schuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M, Coombes T, Katta A, Mullis C, Wortsman M, Schramowski P, Kundurthy S, Crowson K, Schmidt L, Kaczmarczyk R and Jitsev J. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 25278-25294

Schwarz K, Liao Y Y, Niemeyer M and Geiger A. 2020. GRAF: generative radiance fields for 3D-aware image synthesis//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 20154-20166［DOI: 10.48550/arXiv.2007.02442］

Shah S, Dey D, Lovett C and Kapoor A. 2018. AirSim: high-fidelity visual and physical simulation for autonomous vehicles//Proceedings of the Field and Service Robotics: Results of the 11th International Conference. Zurich, Switzerland: Springer International Publishing: 621-635［DOI: 10.1007/978-3-319-67361-5_40］

Shang S, Zhao C L, Zhang R X, Jin J L, Zhang J Y, Guo R Z, Ding S H, Wu Y S, Zhao Y and Jia W. 2025. PVTree: realistic and controllable palm vein generation for recognition tasks［EB/OL］.［2025-02-28］. https://arxiv.org/pdf/2503.02547.pdf

Shen L, Jin J L, Zhang R X, Li H E, Zhao K, Zhang Y Y, Zhang J Y, Ding S H, Zhao Y and Jia W. 2023. RPG-palm: realistic pseudo-data generation for palmprint recognition//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 19548-19559［DOI: 10.1109/ICCV51070.2023.01796］

Shen Y J, Luo P, Yan J J, Wang X G and Tang X O. 2018a. FaceID-GAN: learning a symmetry three-player GAN for identity-preserving face synthesis//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 821-830［DOI: 10.1109/CVPR.2018.00092］

Shen Y J, Zhou B L, Luo P and Tang X O. 2018b. FaceFeat-GAN: a two-stage approach for identity-preserving face synthesis［EB/OL］.［2025-02-28］. https://arxiv.org/pdf/1812.01288.pdf

Sherlock B G and Monro D M. 1993. A model for interpreting fingerprint topology. Pattern Recognition, 26(7): 1047-1055［DOI: 10.1016/0031-3203(93)90006-I］

Shi Y C, Wang P, Ye J L, Mai L, Li K J and Yang X. 2024. MVDream: multi-view diffusion for 3D generation［EB/OL］.［2025-02-28］. https://arxiv.org/pdf/2308.16512.pdf

Shinn N, Labash B and Gopinath A. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection［EB/OL］.［2025-02-28］. http://export.arxiv.org/pdf/2303.11366v1

Shridhar M, Manuelli L and Fox D. 2022. CLIPort: what and where pathways for robotic manipulation//Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR: 894-906

Singer U, Polyak A, Hayes T, Yin X, An J, Zhang S Y, Hu Q Y, Yang H, Ashual O, Gafni O, Parikh D, Gupta S and Taigman Y. 2022. Make-a-video: text-to-video generation without text-video data［EB/OL］.［2025-02-28］. https://arxiv.org/pdf/2209.14792.pdf

Sitzmann V, Martel J N P, Bergman A W, Lindell D B and Wetzstein G. 2020. Implicit neural representations with periodic activation functions//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 7462-7473

Sohn K, Yan X C and Lee H. 2015. Learning structured output representation using deep conditional generative models//Proceedings of the 29th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 3483-3491

Solera-Rico A, Sanmiguel Vila C, Gómez-López M, Wang Y N, Almashjary A, Dawson S T M and Vinuesa R. 2024. β-variational autoencoders and transformers for reduced-order modelling of fluid flows. Nature Communications, 15(1): #1361［DOI: 10.1038/s41467-024-45578-4］

Song J M, Meng C L and Ermon S. 2022. Denoising diffusion implicit models［EB/OL］.［2025-02-28］. https://arxiv.org/pdf/2010.02502.pdf

Song S R, Yu F, Zeng A, Chang A X, Savva M and Funkhouser T. 2017. Semantic scene completion from a single depth image//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 190-198［DOI: 10.1109/

CVPR.2017.28]

Song X W, Zheng J, Yuan S R, Gao H A, Zhao J W, He X, Gu W H and Zhao H. 2024. SA-GS: scale-adaptive Gaussian splatting for training-free anti-aliasing [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2403.19615.pdf

Song Y, Dhariwal P, Chen M and Sutskever I. 2023. Consistency models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 32211-32252

Song Y and Ermon S. 2020. Generative modeling by estimating gradients of the data distribution [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1907.05600.pdf

Song Y, Sohl-Dickstein J, Kingma D P, Kumar A, Ermon S and Poole B. 2021. Score-based generative modeling through stochastic differential equations [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2011.13456.pdf

Straub J, Whelan T, Ma L N, Chen Y F, Wijmans E, Green S, Engel J J, Mur-Artal R, Ren C, Verma S, Clarkson A, Yan M F, Budge B, Yan Y J, Pan X Q, Yon J, Zou Y Y, Leon K, Carter N, Briales J, Gillingham T, Mueggler E, Pesqueira L, Savva M, Batra D, Strasdat H M, De Nardi R, Goesele M, Lovegrove S and Newcombe R. 2019. The replica dataset: a digital replica of indoor spaces [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1906.05797.pdf

Striuk O and Kondratenko Y. 2021. Adaptive deep convolutional GAN for fingerprint sample synthesis//Proceedings of the 4th IEEE International Conference on Advanced Information and Communication Technologies. Lviv, Ukraine: IEEE: 193-196 [DOI: 10.1109/AICT52120.2021.9628978]

Sun C Y, Han J L, Deng W J, Wang X L, Qin Z S and Gould S. 2024a. 3proceduralD-GPT: 3D modeling with large language models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2310.12945.pdf

Sun P Z, Jiang Y, Chen S F, Zhang S L, Peng B Y, Luo P and Yuan Z H. 2024b. Autoregressive model beats diffusion: llama for scalable image generation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2406.06525.pdf

Sun W Q, Chen S, Liu F F, Chen Z L, Duan Y Q, Zhang J and Wang Y K. 2024c. DimensionX: create any 3D and 4D scenes from a single image with controllable video diffusion [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2411.04928.pdf

Sun W, Zhang X, Zhang X R, Zhang G C and Ge N N. 2020. Triplet erasing-based data augmentation for person re-identification. International Journal of Sensor Networks, 34(4): 226-235 [DOI: 10.1504/IJSNET.2020.111782]

Sun X X and Zheng L. 2019. Dissecting person re-identification from the viewpoint of viewpoint//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 608-617 [DOI: 10.1109/CVPR.2019.00070]

Swerdlow A, Xu R S and Zhou B L. 2024. Street-view image generation from a bird's-eye view layout. IEEE Robotics and Automation Letters, 9(4): 3578-3585 [DOI: 10.1109/LRA.2024.3368234]

Tabak E G and Vanden-Eijnden E. 2010. Density estimation by dual ascent of the log-likelihood. Communications in Mathematical Sciences, 8(1): 217-233 [DOI: 10.4310/CMS.2010.v8.n1.a11]

Takahashi R, Matsubara T and Uehara K. 2020. Data augmentation using random image cropping and patching for deep CNNs. IEEE Transactions on Circuits and Systems for Video Technology, 30(9): 2917-2931 [DOI: 10.1109/TCSVT.2019.2935128]

Tang J S, Wang T F, Zhang B, Zhang T, Yi R, Ma L Z and Chen D. 2023. Make-it-3D: high-fidelity 3D creation from a single image with diffusion prior//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22762-22772 [DOI: 10.1109/ICCV51070.2023.02086]

Tang J X, Ren J W, Zhou H, Liu Z W and Zeng G. 2024. DreamGaussian: generative Gaussian splatting for efficient 3D content creation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2309.16653.pdf

Tevet G, Gordon B, Hertz A, Bermano A H and Cohen-Or D. 2022a. MotionCLIP: exposing human motion generation to CLIP space//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 358-374 [DOI: 10.1007/978-3-031-20047-2_21]

Tevet G, Raab S, Gordon B, Shafir Y, Cohen-Or D and Bermano A H. 2022b. Human motion diffusion model [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2209.14916.pdf

Tian K, Jiang Y, Yuan Z, Peng B and Wang L. 2024. Visual autoregressive modeling: scalable image generation via next-scale prediction [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2404.02905.pdf

Tobin J, Fong R, Ray A, Schneider J, Zaremba W and Abbeel P. 2017. Domain randomization for transferring deep neural networks from simulation to the real world//Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver, Canada: IEEE: 23-30 [DOI: 10.1109/IROS.2017.8202133]

Tomczak J M and Welling M. 2016. Improving variational auto-encoders using householder flow [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1611.09630.pdf

Tong T, Li G, Liu X J and Gao Q Q. 2017. Image super-resolution using dense skip connections//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4809-4817 [DOI: 10.1109/ICCV.2017.514]

Torne M, Simeonov A, Li Z C, Chan A, Chen T, Gupta A and Agrawal P. 2024. Reconciling reality through simulation: a real-to-sim-to-real approach for robust manipulation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2403.03949.pdf

Trabucco B, Doherty K, Gurinas M and Salakhutdinov R. 2023. Effective data augmentation with diffusion models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2302.07944.pdf

Tran L, Yin X and Liu X M. 2017. Disentangled representation learning GAN for pose-invariant face recognition//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hono-

lulu, USA: 1283-1292 [DOI: 10.1109/CVPR.2017.141]

Tsao L Y, Lo Y C, Chang C C, Chen H W, Tseng R, Feng C E and Lee C Y. 2024. Boosting flow-based generative super-resolution models via learned prior//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 26005-26015 [DOI: 10.1109/CVPR52733.2024.02457]

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

Vahdat A, Kreis K and Kautz J. 2021. Score-based generative modeling in latent space//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual: Curran Associates Inc.: 11287-11302

van den Berg R , Hasenclever L, Tomczak J M and Welling M. 2018. Sylvester normalizing flows for variational inference//Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence. California, USA: AUAI Press: 393-402

van den Oord A, Kalchbrenner N and Kavukcuoglu K. 2016b. Pixel recurrent neural networks//Proceedings of the 33rd International Conference on Machine Learning. New York, USA: JMLR.org: 1747-1756

van den Oord A, Kalchbrenner N, Vinyals O, Espeholt L, Graves A and Kavukcuoglu K. 2016a. Conditional image generation with PixelCNN decoders//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 4797-4805

van den Oord A, Vinyals O and Kavukcuoglu K. 2017. Neural discrete representation learning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6309-6318

Varol G, Laptev I, Schmid C and Zisserman A. 2021. Synthetic humans for action recognition from unseen viewpoints. International Journal of Computer Vision, 129(7): 2264-2287 [DOI: 10.1007/s11263-021-01467-7]

Varol G, Romero J, Martin X, Mahmood N, Black M J, Laptev I and Schmid C. 2017. Learning from synthetic humans//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4627-4635 [DOI: 10.1109/CVPR.2017.492]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Vincent P, Larochelle H, Lajoie I, Bengio Y and Manzagol P A. 2010. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 11: 3371-3408

Vizcaya P R and Gerhardt L A. 1996. A nonlinear orientation model for global description of fingerprints. Pattern Recognition, 29(7): 1221-1231 [DOI: 10.1016/0031-3203(95)00154-9]

Voleti V, Yao C H, Boss M, Letts A, Pankratz D, Tochilkin D, Laforte C, Rombach R and Jampani V. 2025. SV3D: novel multi-view synthesis and 3D generation from a single image using latent video diffusion//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 439-457 [DOI: 10.1007/978-3-031-73232-4_25]

Von Marcard T, Henschel R, Black M J, Rosenhahn B and Pons-Moll G. 2018. Recovering accurate 3D human pose in the wild using IMUs and a moving camera//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 614-631 [DOI: 10.1007/978-3-030-01249-6_37]

Wang C, Chai M L, He M M, Chen D D and Liao J. 2022a. CLIP-NeRF: text-and-image driven manipulation of neural radiance fields//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 3825-3834 [DOI: 10.1109/CVPR52688.2022.00381]

Wang C, He Z F, Wang C Y and Tian Q. 2022b. Generating intra- and inter-class iris images by identity contrast//Proceedings of 2022 IEEE International Joint Conference on Biometrics. Abu Dhabi, United Arab Emirates: IEEE: 1-7 [DOI: 10.1109/IJCB54206.2022.10007974]

Wang H C, Du X D, Li J H, Yeh R A and Shakhnarovich G. 2023a. Score Jacobian chaining: lifting pretrained 2D diffusion models for 3D generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 12619-12629 [DOI: 10.1109/CVPR52729.2023.01214]

Wang J H, Jin S, Liu W T, Liu W Z, Qian C and Luo P. 2021a. When human pose estimation meets robustness: adversarial algorithms and benchmarks//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 11850-11859 [DOI: 10.1109/CVPR46437.2021.01168]

Wang J N, Yuan H J, Chen D Y, Zhang Y Y, Wang X and Zhang S W. 2023b. Modelscope text-to-video technical report [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2308.06571.pdf

Wang J Y, Yue Z S, Zhou S C, Chan K C K and Loy C C. 2024b. Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision, 132(12): 5929-5949 [DOI: 10.1007/s11263-024-02168-7]

Wang L D, Sindagi V and Patel V. 2018. High-quality facial photo-sketch synthesis using multi-adversarial networks//Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition. Xi'an, China: IEEE: 83-90 [DOI: 10.1109/FG.2018.00022]

Wang Q, Gao J Y, Lin W and Yuan Y. 2019a. Learning from synthetic data for crowd counting in the wild//Proceedings of 2019 IEEE/CVF

Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 8190-8199 [DOI: 10.1109/CVPR. 2019. 00839]

Wang S Y, Du Y Q, Guo X J, Pan B, Qin Z H and Zhao L. 2024c. Controllable data generation by deep learning: a review. ACM Computing Surveys, 56(9): #228 [DOI: 10.1145/3648609]

Wang T F, Zhang B, Zhang T, Gu S Y, Bao J M, Baltrusaitis T, Shen J J, Chen D, Wen F, Chen Q F and Guo B N. 2023c. RODIN: a generative model for sculpting 3D digital avatars using diffusion// Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 4563-4573 [DOI: 10.1109/CVPR52729.2023.00443]

Wang W J, Yang H, Fu J L and Liu J Y. 2024h. Zero-reference low-light enhancement via physical quadruple priors//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 26057-26066 [DOI: 10.1109/CVPR52733.2024.02462]

Wang W J, Yang H, Tuo Z X, He H G, Zhu J C, Fu J L and Liu J Y. 2024d. Swap attention in spatiotemporal diffusions for text-to-video generation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2305.10874.pdf

Wang X D, Darrell T, Rambhatla S S, Girdhar R and Misra I. 2024e. InstanceDiffusion: instance-level control for image generation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6232-6242 [DOI: 10.1109/CVPR52733.2024.00596]

Wang X F, Zhu Z, Huang G, Chen X Z, Zhu J G and Lu J W. 2023e. DriveDreamer: towards real-world-driven world models for autonomous driving [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2309.09777.pdf

Wang X T, Xie L B, Dong C and Shan Y. 2021b. Real-ESRGAN: training real-world blind super-resolution with pure synthetic data//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1905-1914 [DOI: 10.1109/ICCVW54120.2021.00217]

Wang X T, Yu K, Wu S X, Gu J J, Liu Y H, Dong C, Qiao Y and Loy C C. 2019b. ESRGAN: enhanced super-resolution generative adversarial networks//Proceedings of the Computer Vision — ECCV 2018 Workshops. Munich, Germany: Springer: 63-79 [DOI: 10.1007/978-3-030-11021-5_5]

Wang Y and Hu J K. 2011. Global ridge orientation modeling for partial fingerprint identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(1): 72-87 [DOI: 10.1109/TPAMI.2010.73]

Wang Y F, Wan R J, Yang W H, Li H L, Chau L P and Kot A. 2022c. Low-light image enhancement with normalizing flow//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtually: AAAI: 2604-2612 [DOI: 10.1609/aaai.v36i3.20162]

Wang Y F, Xian Z, Chen F, Wang T H, Wang Y, Fragkiadaki K, Erickson Z, Held D and Gan C. 2024g. RoboGen: towards unleashing infinite data for automated robot learning via generative simulation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2311.01455.pdf

Wang Y F, Yu Y, Yang W H, Guo L Q, Chau L P, Kot A C and Wen B H. 2023f. ExposureDiffusion: learning to expose for low-light image enhancement//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 12404-12414 [DOI: 10.1109/ICCV51070.2023.01143]

Wang Y H, Chen X Y, Ma X, Zhou S C, Huang Z Q, Wang Y, Yang C Y, He Y N, Yu J S, Yang P Q, Guo Y W, Wu T X, Si C Y, Jiang Y M, Chen C J, Loy C C, Dai B, Lin D H, Qiao Y and Liu Z W. 2023g. LAVIE: high-quality video generation with cascaded latent diffusion models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2309.15103.pdf

Wang Y N, Liang X Z and Liao S C. 2022d. Cloning outfits from real-world images to 3D characters for generalizable person re-identification// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4890-4899 [DOI: 10.1109/CVPR52688.2022.00485]

Wang Y N, Liao S C and Shao L. 2020. Surpassing real-world source training data: random 3D characters for generalizable person re-identification//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 3422-3430 [DOI: 10.1145/3394171.3413815]

Wang Y Q, He J W, Fan L, Li H X, Chen Y T and Zhang Z X. 2024f. Driving into the future: multiview visual forecasting and planning with world model for autonomous driving//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14749-14759 [DOI: 10.1109/CVPR52733.2024.01397]

Wang Z Y, Lu C, Wang Y K, Bao F, Li C X, Su H and Zhu J. 2023h. ProlificDreamer: high-fidelity and diverse text-to-3D generation with variational score distillation//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 8406-8441

Wecker L, Samavati F and Gavrilova M. 2010. A multiresolution approach to iris synthesis. Computers and Graphics, 34(4): 468-478 [DOI: 10.1016/j.cag.2010.05.012]

Wei Z S, Han Y F, Sun Z N and Tan T N. 2008a. Palmprint image synthesis: a preliminary study//Proceedings of the 15th IEEE International Conference on Image Processing. San Diego, USA: IEEE: 285-288 [DOI: 10.1109/ICIP.2008.4711747]

Wei Z S, Tan T N and Sun Z N. 2007. Nonlinear iris deformation correction based on Gaussian model//Proceedings of 2007 Advances in Biometrics, International Conference. Seoul, Korea (South): Springer: 780-789 [DOI: 10.1007/978-3-540-74549-5_82]

Wei Z S, Tan T N and Sun Z N. 2008b. Synthesis of large realistic iris databases using patch-based sampling//Proceedings of the 19th International Conference on Pattern Recognition. Tampa, USA:

IEEE: 1-4 [DOI: 10.1109/ICPR.2008.4761674]

Wu C F, Huang L, Zhang Q X, Li B Y, Ji L, Yang F, Sapiro G and Duan N. 2021. GODIVA: generating open-domain videos from natural descriptions [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2104.14806.pdf

Wu C F, Liang J, Ji L, Yang F, Fang Y J, Jiang D X and Duan N. 2022. NÜWA: visual synthesis pre-training for neural visual world creation//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 720-736 [DOI: 10.1007/978-3-031-19787-1_41]

Wu G J, Yi T R, Fang J M, Xie L X, Zhang X P, Wei W, Liu W Y, Tian Q and Wang X G. 2024a. 4D Gaussian splatting for real-time dynamic scene rendering//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 20310-20320 [DOI: 10.1109/CVPR52733.2024.01920]

Wu J Z, Ge Y X, Wang X T, Lei S W, Gu Y C, Shi Y F, Hsu W, Shan Y, Qie X H and Shou M Z. 2023a. Tune-a-video: one-shot tuning of image diffusion models for text-to-video generation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 7589-7599 [DOI: 10.1109/ICCV51070.2023.00701]

Wu W J, Zhao Y Z, Chen H, Gu Y C, Zhao R, He Y F, Zhou H, Shou M Z and Shen C H. 2023b. DatasetDM: synthesizing data with perception annotations using diffusion models//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 54683-54695

Wu W J, Zhao Y Z, Shou M Z, Zhou H and Shen C H. 2023c. DiffuMask: synthesizing images with pixel-level annotations for semantic segmentation using diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 1206-1217 [DOI: 10.1109/ICCV51070.2023.00117]

Wu Y S, Shi L Y, Liu H L, Liao H J, Qiu L T, Yuan W H, Gu X D, Dong Z L, Cui S G and Han X G. 2024b. MVImgNet2.0: a larger-scale dataset of multi-view images. ACM Transactions on Graphics, 43(6): #173 [DOI: 10.1145/3687973]

Wyzykowski A B V, Segundo M P and de Paula Lemes R. 2021. Level three synthetic fingerprint generation//Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy: IEEE: 9250-9257 [DOI: 10.1109/ICPR48806.2021.9412304]

Xia B, Zhang Y L, Wang S Y, Wang Y T, Wu X L, Tian Y P, Yang W M and van Gool L. 2023. DiffIR: efficient diffusion model for image restoration//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 13049-13059 [DOI: 10.1109/ICCV51070.2023.01204]

Xia F, Shen W B, Li C S, Kasimbeg P, Tchapmi M E, Toshev A, Martín-Martín R and Savarese S. 2020. Interactive gibson benchmark: a benchmark for interactive navigation in cluttered environments. IEEE Robotics and Automation Letters, 5(2): 713-720 [DOI: 10.1109/LRA.2020.2965104]

Xia F, Zamir A R, He Z Y, Sax A, Malik J and Savarese S. 2018. Gibson env: real-world perception for embodied agents//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9068-9079 [DOI: 10.1109/CVPR.2018.00945]

Xiang F B, Qin Y Z, Mo K C, Xia Y K, Zhu H, Liu F C, Liu M H, Jiang H X, Yuan Y F, Wang H, Yi L, Chang A X, Guibas L J and Su H. 2020. SAPIEN: a simulated part-based interactive environment//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11094-11104 [DOI: 10.1109/CVPR42600.2020.01111]

Xiang J F, Yang J L, Deng Y and Tong X. 2023. GRAM-HD: 3D-consistent image generation at high resolution with generative radiance manifolds//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 2195-2205 [DOI: 10.1109/ICCV51070.2023.00209]

Xiao C W, Li B, Zhu J Y, He W, Liu M Y and Song D. 2018. Generating adversarial examples with adversarial networks//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: AAAI Press: 3905-3911

Xie T Y, Zong Z S, Qiu Y X, Li X, Feng Y T, Yang Y and Jiang C F F. 2024. PhysGaussian: physics-integrated 3D Gaussians for generative dynamics//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4389-4398 [DOI: 10.1109/CVPR52733.2024.00420]

Xing J B, Xia M H, Zhang Y, Chen H X, Yu W B, Liu H Y, Liu G Y, Wang X T, Shan Y and Wong T T. 2025. DynamiCrafter: animating open-domain images with video diffusion priors//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 399-417 [DOI: 10.1007/978-3-031-72952-2_23]

Xu D, Ouyang W L, Ricci E, Wang X G and Sebe N. 2017a. Learning cross-modal deep representations for robust pedestrian detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4236-4244 [DOI: 10.1109/CVPR.2017.451]

Xu H F, Peng S Y, Wang F J H, Blum H, Barath D, Geiger A and Pollefeys M. 2025. DepthSplat: connecting Gaussian splatting and depth [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2410.13862.pdf

Xu J Q, Zou X Y, Huang K Z, Chen Y K, Liu B, Cheng M L, Shi X and Huang J. 2024a. EasyAnimate: a high-performance long video generation method based on transformer architecture [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2405.18991.pdf

Xu J R, Liu S F, Vahdat A, Byeon W, Wang X L and De Mello S. 2023. Open-vocabulary panoptic segmentation with text-to-image diffusion models//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 2955-2966 [DOI: 10.1109/CVPR52729.2023.00289]

Xu J W, Fan Z X, Yang J and Xie J. 2024b. Grid4D: 4D decomposed

hash encoding for high-fidelity dynamic Gaussian splatting [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2410.20815.pdf

Xu W D, Sun H Z, Deng C and Tan Y. 2017b. Variational autoencoder for semi-supervised text classification//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI: 3358-3364 [DOI: 10.1609/aaai.v31i1.10966]

Xu Y W, Zhao Y, Xiao Z S and Hou T B. 2024c. UFOGen: you forward once large scale text-to-image generation via diffusion GANs//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8196-8206 [DOI: 10.1109/CVPR52733.2024.00783]

Yadav S, Chen C J and Ross A. 2019. Synthesizing iris images using RaSGAN with application in presentation attack detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA: IEEE: 2422-2430 [DOI: 10.1109/CVPRW.2019.00297]

Yadav S and Ross A. 2021. CIT-GAN: cyclic image translation generative adversarial network with application in iris presentation attack detection//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2411-2420 [DOI: 10.1109/WACV48630.2021.00246]

Yan C, Misra D, Bennett A, Walsman A, Bisk Y and Artzi Y. 2019. CHALET: cornell house agent learning environment [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/1801.07357.pdf

Yan H, Liu Y L, Jin L W and Bai X. 2023. The development, application, and future of LLM similar to ChatGPT. Journal of Image and Graphics, 28(9): 2749-2762 (严昊, 刘禹良, 金连文, 白翔. 2023. 类ChatGPT大模型发展、应用和前景. 中国图象图形学报, 28(9): 2749-2762) [DOI: 10.11834/jig.230536]

Yan Y Z, Lin H T, Zhou C X, Wang W J, Sun H Y, Zhan K, Lang X P, Zhou X W and Peng S D. 2024a. Street Gaussians: modeling dynamic urban scenes with Gaussian splatting [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2401.01339.pdf

Yan Z W, Low W F, Chen Y and Lee G H. 2024b. Multi-scale 3D Gaussian splatting for anti-aliased rendering//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 20923-20931 [DOI: 10.1109/CVPR52733.2024.01977]

Yang G K. 2023. Fingerprint image generation method based on diffusion models. Journal of the Hebei Academy of Sciences, 40(1): 13-18, 66 (杨光锴. 2023. 基于扩散模型的指纹图像生成方法. 河北省科学院学报, 40(1): 13-18, 66) [DOI: 10.16191/j.cnki.hbkx.2023.01.009]

Yang H W, Fang P Y and Hao Z A. 2021. A GAN-based method for generating finger vein dataset//Proceedings of the 3rd International Conference on Algorithms, Computing and Artificial Intelligence. Sanya, China: Association for Computing Machinery: #18 [DOI: 10.1145/3446132.3446150]

Yang L H, Xu X G, Kang B Y, Shi Y H and Zhao H S. 2023b. FreeMask: synthetic images with dense annotations make stronger segmentation models//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 18659-18675

Yang W H, Tan R T, Feng J S, Liu J Y, Guo Z M and Yan S C. 2017. Deep joint rain detection and removal from a single image//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1685-1694 [DOI: 10.1109/CVPR.2017.183]

Yang W H, Wang S Q, Fang Y M, Wang Y and Liu J Y. 2020. From fidelity to perceptual quality: a semi-supervised approach for low-light image enhancement//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3060-3069 [DOI: 10.1109/CVPR42600.2020.00313]

Yang Z Q, Li S K, Wu W and Dai B. 2023a. 3DHumanGAN: 3D-aware human image generation with 3D pose mapping//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22951-22962 [DOI: 10.1109/ICCV51070.2023.02103]

Yang Z T, Cai Z A, Mei H Y, Liu S, Chen Z X, Xiao W Y, Wei Y K, Qing Z F, Wei C, Dai B, Wu W, Qian C, Lin D H, Liu Z W and Yang L. 2023c. SynBody: synthetic dataset with layered human models for 3D human perception and modeling//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 20225-20235 [DOI: 10.1109/ICCV51070.2023.01855]

Yang Z Y, Gao X Y, Zhou W, Jiao S H, Zhang Y Q and Jin X G. 2024a. Deformable 3D Gaussians for high-fidelity monocular dynamic scene reconstruction//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 20331-20341 [DOI: 10.1109/CVPR52733.2024.01922]

Yang Z Y, Teng J Y, Zheng W D, Ding M, Huang S Y, Xu J Z, Yang Y M, Hong W Y, Zhang X H, Feng G Y, Yin D, Zhang Y X, Wang W H, Cheng Y A, Xu B, Gu X T, Dong Y X and Tang J. 2025. CogVideoX: text-to-video diffusion models with an expert transformer [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2408.06072.pdf

Yang Z Y, Wang J F, Gan Z, Li L J, Lin K, Wu C F, Duan N, Liu Z C, Liu C, Zeng M and Wang L J. 2023d. ReCo: region-controlled text-to-image generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 14246-14255 [DOI: 10.1109/CVPR52729.2023.01369]

Yang Z Y, Yang H Y, Pan Z J and Zhang L. 2024b. Real-time photorealistic dynamic scene representation and rendering with 4D Gaussian splatting [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2310.10642.pdf

Yao K, Gao P L, Yang X, Huang K Z, Sun J and Zhang R. 2022. Out-

painting by queries [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2207.05312.pdf

Ye K Y, Hou Q M and Zhou K. 2024. 3D Gaussian splatting with deferred reflection//Proceedings of 2024 ACM SIGGRAPH Conference Papers. Denver, USA: Association for Computing Machinery: #40 [DOI: 10.1145/3641519.3657456]

Ye Y T, Chang Y, Zhou H Y and Yan L X. 2021. Closing the loop: joint rain generation and removal via disentangled image translation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2053-2062 [DOI: 10.1109/CVPR46437.2021.00209]

Yi T R, Fang J M, Wang J J, Wu G J, Xie L X, Zhang X P, Liu W Y, Tian Q and Wang X G. 2024. GaussianDreamer: fast generation from text to 3D Gaussians by bridging 2D and 3D diffusion models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6796-6807 [DOI: 10.1109/CVPR52733.2024.00649]

Yin X, Yu X, Sohn K, Liu X M and Chandraker M. 2017. Towards large-pose face frontalization in the wild//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4010-4019 [DOI: 10.1109/ICCV.2017.430]

Youwang K, Ji-Yeon K and Oh T H. 2022. CLIP-Actor: text-driven recommendation and stylization for animating human meshes//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 173-191 [DOI: 10.1007/978-3-031-20062-5_11]

Yu A, Foote A, Mooney R and Martín-Martín R. 2024a. Natural language can help bridge the Sim2Real gap [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2405.10020.pdf

Yu A, Ye V, Tancik M and Kanazawa A. 2021. pixelNeRF: neural radiance fields from one or few images//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4576-4585 [DOI: 10.1109/CVPR46437.2021.00455]

Yu S Y, Nie W L, Huang D A, Li B Y, Shin J and Anandkumar A. 2024b. Efficient video diffusion models via content-frame motion-latent decomposition [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2403.14148.pdf

Yu W B, Xing J B, Yuan L, Hu W B, Li X Y, Huang Z P, Gao X J, Wong T T, Shan Y and Tian Y H. 2024c. ViewCrafter: taming video diffusion models for high-fidelity novel view synthesis [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2409.02048.pdf

Yu W H, Tan J, Liu C K and Turk G. 2017. Preparing for the unknown: learning a universal policy with online system identification//Proceedings of Robotics: Science and Systems. Cambridge, USA [DOI: 10.15607/RSS.2017.XIII.048]

Yu X, Guo Y C, Li Y G, Liang D, Zhang S H and Qi X J. 2023b. Text-to-3D with classifier score distillation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2310.19415.pdf

Yu X G, Xu M T, Zhang Y D, Liu H L, Ye C J, Wu Y S, Yan Z Z, Zhu C M, Xiong Z Y, Liang T Y, Chen G Y, Cui S G and Han X G. 2023a. MVImgNet: a large-scale dataset of multi-view images//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 9150-9161 [DOI: 10.1109/CVPR52729.2023.00883]

Yu Z H, Chen A P, Huang B B, Sattler T and Geiger A. 2024d. Mip-splatting: alias-free 3D Gaussian splatting//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 19447-19456 [DOI: 10.1109/CVPR52733.2024.01839]

Yun S, Oh S J, Heo B, Han D and Kim J. 2020. VideoMix: rethinking data augmentation for video classification [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2012.03457.pdf

Yurtsever E, Yang D F, Koc I M and Redmill K A. 2022. Photorealism in driving simulations: blending generative adversarial image synthesis with rendering. IEEE Transactions on Intelligent Transportation Systems, 23(12): 23114-23123 [DOI: 10.1109/TITS.2022.3193347]

Zamir S W, Arora A, Khan S, Hayat M, Khan F S and Yang M H. 2022. Restormer: efficient transformer for high-resolution image restoration//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5718-5729 [DOI: 10.1109/CVPR52688.2022.00564]

Zeng A L, Yang Y H, Chen W D and Liu W. 2024a. The dawn of video generation: preliminary explorations with SORA-like models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2410.05227.pdf

Zeng Y, Wei G Q, Zheng J N, Zou J X, Wei Y, Zhang Y C and Li H. 2024b. Make pixels dance: high-dynamic video generation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8850-8860 [DOI: 10.1109/CVPR52733.2024.00845]

Zhang C Y, Zhu L and Zhang S C. 2019a. PAC-GAN: an effective pose augmentation scheme for unsupervised cross-view person re-identification [EB/OL]. [2025-02-28]. http://arxiv.org/pdf/1906.01792.pdf

Zhang D J, Wu J Z, Liu J W, Zhao R, Ran L M, Gu Y C, Gao D F and Shou M Z. 2025a. Show-1: marrying pixel and latent diffusion models for text-to-video generation. International Journal of Computer Vision, 133(4): 1879-1893 [DOI: 10.1007/s11263-024-02271-9]

Zhang H, Goodfellow I, Metaxas D and Odena A. 2019b. Self-attention generative adversarial networks//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 7354-7363

Zhang H, Sindagi V and Patel V M. 2020a. Image de-raining using a conditional generative adversarial network. IEEE Transactions on Circuits and Systems for Video Technology, 30(11): 3943-3956 [DOI: 10.1109/TCSVT.2019.2920407]

Zhang H, Xu H, Xiao Y, Guo X J and Ma J Y. 2020b. Rethinking the

image fusion: a fast unified image fusion network based on proportional maintenance of gradient and intensity//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 12797-12804 [DOI: 10.1609/aaai.v34i07.6975]

Zhang H, Xu T, Li H S, Zhang S T, Wang X G, Huang X L and Metaxas D. 2017. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5908-5916 [DOI: 10.1109/iccv.2017.629]

Zhang J F, Jiang Z H, Yang D D, Xu H Y, Shi Y C, Song G X, Xu Z C, Wang X C and Feng J S. 2022. Avatargen: a 3D generative model for animatable human avatars. //Proceedings of 2022 European Conference on Computer Vision. Israel.: Springer:, 668-685. [DOI:10.1007/978-3-031-25066-8_39]

Zhang J R, Zhang Y S, Cun X D, Zhang Y, Zhao H W, Lu H T, Shen X and Ying S. 2023a. Generating human motion from textual descriptions with discrete representations//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 14730-14740 [DOI: 10.1109/CVPR52729.2023.01415]

Zhang J T, Sun F W, Song J, Von Ancken A and Zhai R. 2018. Fine-grained image classification via spatial saliency extraction//Proceedings of the 17th IEEE International Conference on Machine Learning and Applications. Orlando, USA: IEEE: 249-255 [DOI: 10.1109/ICMLA.2018.00044]

Zhang J T, Shum H P H, Han J G and Shao L. 2018. Action recognition from arbitrary views using transferable dictionary learning. IEEE Transactions on Image Processing, 27(10): 4709-4723 [DOI: 10.1109/TIP.2018.2836323]

Zhang L M, Rao A Y and Agrawala M. 2023b. Adding conditional control to text-to-image diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3813-3824 [DOI: 10.1109/ICCV51070.2023.00355]

Zhang M L, Wu J, Ren Y X, Li M, Qin J, Xiao X F, Liu W, Wang R, Zheng M and Ma A J. 2023c. DiffusionEngine: diffusion model is scalable data engine for object detection [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2309.03893.pdf

Zhang Q, Lin W and Chan A B. 2021a. Cross-view cross-scene multi-view crowd counting//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 557-567 [DOI: 10.1109/CVPR46437.2021.00062]

Zhang S G, Zhou M Q, Wang Y X, Luo C C, Wang R Y, Li Y W, Yin X C, Zhang Z X and Peng J R. 2024b. CityX: controllable procedural content generation for unbounded 3D cities [EB/OL]. [2025-02-28]. https://export.arxiv.org/pdf/2407.17572.pdf

Zhang S W, Wang J Y, Zhang Y Y, Zhao K, Yuan H J, Qin Z W, Wang X, Zhao D L and Zhou J R. 2023d. I2VGen-XL: high-quality image-to-video synthesis via cascaded diffusion models [EB/OL]. [2025-04-22]. https://arxiv.org/pdf/2311.04145.pdf

Zhang T Y, Wang L, Li H N, Xiao Y S, Liang S Y, Liu A S, Liu X L and Tao D C. 2024a. LanEvil: benchmarking the robustness of lane detection to environmental illusions//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM: 5403-5412 [DOI: 10.1145/3664647.3680761]

Zhang T Y, Xie L X, Wei L H, Zhuang Z J, Zhang Y F, Li B and Tian Q. 2021b. UnrealPerson: an adaptive pipeline towards costless person re-identification//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 11501-11510 [DOI: 10.1109/CVPR46437.2021.01134]

Zhang Y H, Zhang J W and Guo X J. 2019c. Kindling the darkness: a practical low-light image enhancer//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 1632-1640 [DOI: 10.1145/3343031.3350926]

Zhang Y M, Jia G G, Chen L, Zhang M R and Yong J H. 2020c. Self-paced video data augmentation by generative adversarial networks with insufficient samples//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 1652-1660 [DOI: 10.1145/3394171.3414003]

Zhang Y X, Ling H, Gao J, Yin K X, Lafleche J F, Barriuso A, Torralba A and Fidler S. 2021c. DatasetGAN: efficient labeled data factory with minimal human effort//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10140-10150 [DOI: 10.1109/CVPR46437.2021.01001]

Zhang Z, Hu W B, Lao Y X, He T and Zhao H S. 2025b. Pixel-GS: density control with pixel-aware gradient for 3D Gaussian splatting//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 326-342 [DOI: 10.1007/978-3-031-72655-2_19]

Zhao G S, Wang X F, Zhu Z, Chen X Z, Huang G, Bao X Y and Wang X G. 2024. Drivedreamer-2: LLM-enhanced world models for diverse driving video generation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2403.06845.pdf

Zhao H Q, Sheng D M, Bao J M, Chen D D, Chen D, Wen F, Yuan L, Liu C, Zhou W B, Chu Q, Zhang W M and Yu N H. 2023. X-paste: revisiting scalable copy-paste for instance segmentation using CLIP and stablediffusion//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 42098-42109

Zhao H S, Jiang L, Jia J Y, Torr P and Koltun V. 2021. Point transformer//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 16239-16248 [DOI: 10.1109/ICCV48922.2021.01595]

Zhao K, Shen L, Zhang Y Y, Zhou C H, Wang T, Zhang R X, Ding S H, Jia W and Shen W. 2022. BézierPalm: a free lunch for palmprint recognition//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 19-36 [DOI: 10.1007/978-3-031-19778-9_2]

Zhao Q J, Jain A K, Paulter N G and Taylor M. 2012. Fingerprint image synthesis based on statistical feature models//Proceedings of the 5th IEEE International Conference on Biometrics: Theory, Applications and Systems. Arlington, USA: IEEE: 23-30 [DOI: 10.1109/BTAS.2012.6374554]

Zheng D H, Zou Y H, Zhang X W and Bao C L. 2024a. SeNM-VAE: semi-supervised noise modeling with hierarchical variational autoencoder//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 25889-25899 [DOI: 10.1109/CVPR52733.2024.02446]

Zheng G C, Zhou X P, Li X W, Qi Z A, Shan Y and Li X. 2023. LayoutDiffusion: controllable diffusion model for layout-to-image generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 22490-22499 [DOI: 10.1109/CVPR52729.2023.02154]

Zheng Z D, Yang X D, Yu Z D, Zheng L, Yang Y and Kautz J. 2019. Joint discriminative and generative learning for person re-identification//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2133-2142 [DOI: 10.1109/CVPR.2019.00224]

Zheng Z D, Zheng L and Yang Y. 2017. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 3774-3782 [DOI: 10.1109/ICCV.2017.405]

Zheng Z W, Peng X Y, Yang T J, Shen C H, Li S G, Liu H X, Zhou Y K, Li T Y and You Y. 2024b. Open-sora: democratizing efficient video production for all [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2412.20404.pdf

Zhong C H, Xu P X and Zhu L S. 2021. A deep convolutional generative adversarial network-based fake fingerprint generation method//Proceedings of 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology. Fuzhou, China: IEEE: 63-67 [DOI: 10.1109/CEI52496.2021.9574508]

Zhong Z, Zheng L, Kang G L, Li S Z and Yang Y. 2020. Random erasing data augmentation//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 13001-13008 [DOI: 10.1609/aaai.v34i07.7000]

Zhong Z, Zheng L, Li S Z and Yang Y. 2018. Generalizing a person retrieval model hetero- and homogeneously//Proceedings of the 15th European Conference Computer Vision. Munich, Germany: Springer International Publishing: 176-192 [DOI: 10.1007/978-3-030-01261-8_11]

Zhou D Q, Wang W M, Yan H S, Lv W W, Zhu Y Z and Feng J S. 2023a. Magicvideo: efficient video generation with latent diffusion models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2211.11018.pdf

Zhou D W, Li Y, Ma F, Zhang X T and Yang Y. 2024a. MIGC: multi-instance generation controller for text-to-image synthesis//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6818-6828 [DOI: 10.1109/CVPR52733.2024.00651]

Zhou D W, Yang Z X and Yang Y. 2023b. Pyramid diffusion models for low-light image enhancement//Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: ACM: 1795-1803 [DOI: 10.24963/ijcai.2023/199]

Zhou M Q, Wang Y X, Hou J, Zhang S G, Li Y W, Luo C C, Peng J R and Zhang Z X. 2024b. SceneX: procedural controllable large-scale scene generation [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2403.15698.pdf

Zhou S, Zhang J Q, Jiang H, Lundh T and Ng A Y. 2021. Data augmentation with Mobius transformations. Machine Learning: Science and Technology, 2(2): #025016 [DOI: 10.1088/2632-2153/abd615]

Zhou X Y, Lin Z W, Shan X J, Wang Y T, Sun D Q and Yang M H. 2024c. DrivingGaussian: composite Gaussian splatting for surrounding dynamic autonomous driving scenes//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: 21634-21643 [DOI: 10.1109/CVPR52733.2024.02044]

Zhou X Y, Ran X J, Xiong Y J, He J L, Lin Z W, Wang Y T, Sun D Q and Yang M H. 2024d. GALA3D: towards text-to-3D complex scene generation via layout-guided generative Gaussian splatting//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR: 62108-62118

Zhou Y, Wang Q Y, Cai Y X and Yang H. 2024e. Allegro: open the black box of commercial-level video generation model [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2410.15458.pdf

Zhu J Y, Krähenbühl P, Shechtman E and Efros A A. 2016. Generative visual manipulation on the natural image manifold//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 597-613 [DOI: 10.1007/978-3-319-46454-1_36]

Zhu J Y, Li S Y, Liu Y X, Huang P, Shan J L, Ma H M and Yuan J. 2024b. ODGen: domain-specific object detection data generation with diffusion models [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2405.15199.pdf

Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2242-2251 [DOI: 10.1109/ICCV.2017.244]

Zhu J Z, Zhuang P Y and Koyejo S. 2024a. HiFA: high-fidelity text-to-3D generation with advanced diffusion guidance [EB/OL]. [2025-02-28]. https://arxiv.org/pdf/2305.18766.pdf

Zhu Z H, Fan Z W, Jiang Y F and Wang Z Y. 2025. FSGS: real-time few-shot view synthesis using Gaussian splatting//Proceedings of

the 18th European Conference on Computer Vision-ECCV 2024. Milan, Italy: Springer-Verlag: 145-163 [DOI: 10.1007/978-3-031-72933-1_9]

Zou H, Zhang H, Li X G, Liu J and He Z F. 2018. Generation textured contact lenses iris images based on 4DCycle-GAN//Proceedings of the 24th International Conference on Pattern Recognition. Beijing, China: IEEE: 3561-3566 [DOI: 10.1109/ICPR.2018.8546154]

Zou Y L, Choi J, Wang Q T and Huang J B. 2023. Learning representational invariances for data-efficient action recognition. Computer Vision and Image Understanding, 227: #103597 [DOI: 10.1016/j.cviu.2022.103597]

Zuo J Y, Schmid N A and Chen X H. 2007. On generation and analysis of synthetic iris images. IEEE Transactions on Information Forensics and Security, 2(1): 77-90 [DOI: 10.1109/TIFS.2006.890305]

作者简介

马愈卓，男，博士研究生，主要研究方向为计算机视觉。
E-mail: mayuzhuo@buaa.edu.cn

张永飞，通信作者，男，教授，主要研究方向为计算机视觉。
E-mail: yfzhang@buaa.edu.cn

贾伟，男，教授，主要研究方向为计算机视觉。
E-mail: jiawei@hfut.edu.cn

刘家瑛，女，副教授，主要研究方向为多媒体计算。
E-mail: liujiaying@pku.edu.cn

甘甜，女，教授，主要研究方向为多媒体计算。
E-mail: gantian@sdu.edu.cn

杨文瀚，男，副研究员，主要研究方向为底层视觉增强与高效特征适配计算。E-mail: yangwh@pcl.ac.cn

卓君宝，男，副教授，主要研究方向为计算机视觉。
E-mail: junbaozhuo@ustb.edu.cn

刘武，男，特任教授，主要研究方向为多模态生成、多智能体。
E-mail: liuwu@ustc.edu.cn

马惠敏，女，教授，主要研究方向为计算机视觉。
E-mail: mhmpub@ustb.edu.cn