CAICT 中国信通院　京东探索研究院 JD EXPLORE ACADEMY

集智白皮书 No.202212

人工智能生成内容（AIGC）白皮书

（2022 年）

中国信息通信研究院

京东探索研究院

2022年9月

前　言

习近平总书记曾指出，“数字技术正以新理念、新业态、新模式全面融入人类经济、政治、文化、社会、生态文明建设各领域和全过程”。在当前数字世界和物理世界加速融合的大背景下，人工智能生成内容（Artificial Intelligence Generated Content，简称 AIGC）正在悄然引导着一场深刻的变革，重塑甚至颠覆数字内容的生产方式和消费模式，将极大地丰富人们的数字生活，是未来全面迈向数字文明新时代不可或缺的支撑力量。

本白皮书重点从 AIGC 技术、应用和治理等维度进行了阐述。**在技术层面**，梳理提出了 AIGC 技术体系，既涵盖了对现实世界各种内容的数字化呈现和增强，也包括了基于人工智能的自主内容创作。**在应用层面**，重点分析了 AIGC 在传媒、电商、影视等行业和场景的应用情况，探讨了以虚拟数字人、写作机器人等为代表的新业态和新应用。**在治理层面**，从政策监管、技术能力、企业应用等视角，分析了 AIGC 所暴露出的版权纠纷、虚假信息传播等各种问题。最后，从政府、行业、企业、社会等层面，给出了 AIGC 发展和治理建议。由于人工智能仍处于飞速发展阶段，我们对 AIGC 的认识还有待进一步深化，白皮书中存在不足之处，敬请大家批评指正。

目 录

图 目 录

一、人工智能生成内容的发展历程与概念

1950 年，艾伦·图灵（Alan Turing）在其论文《计算机器与智能（Computing Machinery and Intelligence）》中提出了著名的“图灵测试”，给出了判定机器是否具有“智能”的试验方法，即机器是否能够模仿人类的思维方式来“生成”内容继而与人交互。某种程度上来说，人工智能从那时起就被寄予了用于内容创造的期许。经过半个多世纪的发展，随着数据快速积累、算力性能提升和算法效力增强，今天的人工智能不仅能够与人类进行互动，还可以进行写作、编曲、绘画、视频制作等创意工作。2018 年，人工智能生成的画作在佳士得拍卖行以 43.25 万美元成交，成为世界上首个出售的人工智能艺术品，引发各界关注。随着人工智能越来越多地被应用于内容创作，人工智能生成内容（Artificial Intelligence Generated Content，简称 AIGC）的概念悄然兴起。

（一）AIGC 历史沿革

结合人工智能的演进历程，AIGC 的发展大致可以分为三个阶段，即：早期萌芽阶段（20 世纪 50 年代至 90 年代中期）、沉淀积累阶段（20 世纪 90 年代中期至 21 世纪 10 年代中期），以及快速发展阶段（21 世纪 10 年代中期至今）。

早期萌芽阶段（1950s-1990s），受限于当时的科技水平，AIGC 仅限于小范围实验。1957 年，莱杰伦·希勒（Lejaren Hiller）和伦纳

德·艾萨克森（Leonard Isaacson）通过将计算机程序中的控制变量换成音符完成了历史上第一支由计算机创作的音乐作品——弦乐四重奏《依利亚克组曲（Illiac Suite）》。1966 年，约瑟夫·魏岑鲍姆（Joseph Weizenbaum）和肯尼斯·科尔比（Kenneth Colby）共同开发了世界第一款可人机对话的机器人“伊莉莎（Eliza）”，其通过关键字扫描和重组完成交互任务。80 年代中期，IBM 基于隐形马尔科夫链模型（Hidden Markov Model, HMM）创造了语音控制打字机“坦戈拉（Tangora）”，能够处理约 20000 个单词。80 年代末至 90 年代中，由于高昂的系统成本无法带来可观的商业变现，各国政府纷纷减少了在人工智能领域的投入，AIGC 没有取得重大突破。

沉淀积累阶段（1990s-2010s），AIGC 从实验性向实用性逐渐转变。2006 年，深度学习算法取得重大突破，同时期图形处理器（Graphics Processing Unit, GPU）、张量处理器（Tensor Processing Unit, TPU）等算力设备性能不断提升，互联网使数据规模快速膨胀并为各类人工智能算法提供了海量训练数据，使人工智能发展取得了显著的进步。但是 AIGC 依然受限于算法瓶颈，无法较好地完成创作任务，应用仍然有限，效果有待提升。2007 年，纽约大学人工智能研究员罗斯·古德温装配的人工智能系统通过对公路旅行中的一切所见所闻进行记录和感知，撰写出小说《1 The Road》。作为世界第一部完全由人工智能创作的小说，其象征意义远大于实际意义，整体可读性不强，拼写错误、辞藻空洞、缺乏逻辑等缺点明显。2012 年，微软公开展示

了一个全自动同声传译系统，基于深层神经网络（Deep Neural Network, DNN）可以自动将英文演讲者的内容通过语音识别、语言翻译、语音合成等技术生成中文语音。

快速发展阶段（2010s-至今），自 2014 年起，随着以生成式对抗网络（Generative Adversarial Network, GAN）为代表的深度学习算法的提出和迭代更新，AIGC 迎来了新时代，生成内容百花齐放，效果逐渐逼真直至人类难以分辨。2017 年，微软人工智能少女“小冰”推出了世界首部 100%由人工智能创作的诗集《阳光失了玻璃窗》。2018 年，英伟达发布的 StyleGAN 模型可以自动生成图片，目前已升级到第四代模型 StyleGAN-XL，其生成的高分辨率图片人眼难以分辨真假。2019 年，DeepMind 发布了 DVD-GAN 模型用以生成连续视频，在草地、广场等明确场景下表现突出。2021 年，OpenAI 推出了 DALL-E 并于一年后推出了升级版本 DALL-E-2,主要应用于文本与图像的交互生成内容，用户只需输入简短的描述性文字，DALL-E-2 即可创作出相应极高质量的卡通、写实、抽象等风格的绘画作品。

来源：中国信息通信研究院

图 1 AIGC 发展历程

（二）AIGC 的概念与内涵

目前，对 AIGC 这一概念的界定，尚无统一规范的定义。国内产学研各界对于 AIGC 的理解是“继专业生成内容（Professional Generated Content, PGC）和用户生成内容（User Generated Content, UGC）之后，利用人工智能技术自动生成内容的新型生产方式”。在国际上对应的术语是“人工智能合成媒体（AI-generated Media 或 Synthetic Media）”[1]，其定义是“通过人工智能算法对数据或媒体进行生产、操控和修改的统称”。**综上所述，我们认为 AIGC 既是从内容生产者视角进行分类的一类内容，又是一种内容生产方式，还是用于内容自动化生成的一类技术集合。本白皮书主要聚焦于 AIGC 含义**

[1] 维基百科：“人工智能合成媒体（AI-generated Media 或 Synthetic Media）”
https://en.wikipedia.org/wiki/Synthetic_media

中的技术部分。

为了帮助不同领域的受众群体更好的理解 AIGC，我们从发展背景、技术能力、应用价值三个方面对其概念进行深入剖析。

从发展背景方面来看，AIGC **的兴起源于深度学习技术的快速突破和日益增长的数字内容供给需求。一方面，**技术进步驱动 AIGC 可用性不断增强。在人工智能发展初期，虽然对 AIGC 进行了一些初步尝试，但受限各种因素，相关算法多基于预先定义的规则或者模板，还远远算不上是智能创作内容的程度。近年来，基于深度学习算法的 AIGC 技术快速迭代，彻底打破了原先模板化、公式化、小范围的局限，可以快速、灵活地生成不同模态的数据内容。**另一方面，**海量需求牵引 AIGC 应用落地。随着数字经济与实体经济融合程度不断加深，以及 Meta、微软、字节跳动等平台型巨头的数字化场景向元宇宙转型，人类对数字内容总量和丰富程度的整体需求不断提高。数字内容的生产取决于想象能力、制造能力和知识水平；传统内容生产手段受限于人力有限的制造能力，逐渐无法满足消费者对于数字内容的消费需求，供给侧产能瓶颈日益凸显。基于以上原因，AIGC 在各行业中得到越来越广泛的应用，市场潜力逐渐显现。

从技术能力方面来看，AIGC **根据面向对象、实现功能的不同可分为三个层次。一是智能数字内容孪生**，其主要目标是建立现实世界到数字世界的映射，将现实世界中的物理属性（如物体的大小、纹理、颜色等）和社会属性（如主体行为、主体关系等）高效、可感知地进

行数字化。**二是智能数字内容编辑**，其主要目的是建立数字世界与现实世界的双向交互。在数字内容孪生的基础上，从现实世界实现对虚拟数字世界中内容的控制和修改，同时利用数字世界高效率仿真和低成本试错的优势，为现实世界的应用提供快速迭代能力。**三是智能数字内容创作**，其主要目标是让人工智能算法具备内容创作和自我演化的能力，形成的 AIGC 产品具备类似甚至超越人的创作能力。以上三个层面的能力共同构成 AIGC 的能力闭环。

从应用价值方面来看，AIGC **将有望成为数字内容创新发展的新引擎，为数字经济发展注入全新动能。一方面，**AIGC 能够以优于人类的制造能力和知识水平承担信息挖掘、素材调用、复刻编辑等基础性机械劳动，从技术层面实现以低边际成本、高效率的方式满足海量个性化需求；同时能够创新内容生产的流程和范式，为更具想象力的内容、更加多样化的传播方式提供可能性，推动内容生产向更有创造力的方向发展。**另一方面，**AIGC 能够通过支持数字内容与其他产业的多维互动、融合渗透从而孕育新业态新模式，打造经济发展新增长点，为千行百业发展提供新动能。此外，2021 年以来，“元宇宙”呈现出超出想象的发展爆发力；作为数实融合的“终极”数字载体，元宇宙将具备持续性、实时性、可创造性等特征，也将通过 AIGC 加速复刻物理世界、进行无限内容创作，从而实现自发有机生长。

二、人工智能生成内容的技术体系及其演进方向

AIGC 作为人工智能技术和产业应用的要素之一，随着技术能力的不断迭代升级，正在降低内容创作门槛、释放创作能力，未来将推动数实融合趋势下内容创作的范式转变。探讨其能力体系的构成，即赋能内容创作的技术路径，对制定领域内标准、建立行业生态、争取更加广泛的开发者和应用场景具有十分重要的意义。

本部分从技术驱动的视角出发，对 AIGC 的能力体系进行归纳和推理，展示现有技术应用和其背后技术演化整体进程。第一节首先从技术趋势的角度，提出 AIGC 的技术创新已经完成由传统方法向深度学习过渡的应用创新阶段，并逐步深化到学习范式和网络结构方面的理论创新阶段。第二节则重点分析前沿理论多模态大模型方面的突破，让 AIGC 进行跨模态融合性创新成为可能，也给予了 AIGC 前所未有的产业空间与实践潜力。第三节进一步归纳总结在前沿技术驱动下，AIGC 赋能内容创作的三大能力，并对三大能力的技术演化路径进行展望。

（一）AIGC 技术升级步入深化阶段

人工智能算法的不断迭代是 AIGC 发展进步的源动力，从技术演进的角度出发，可将 AIGC 技术可大致划分为传统基于模板或规则的前深度学习阶段和深度神经网络快速发展的深度学习阶段。

早期的 AIGC 技术主要依据事先指定的模板或者规则，进行简单的内容制作与输出，与灵活且真实的内容生成有较大的差距。该时期

的人工智能算法并不具备强大的学习能力，而是大多依赖于预先定义的统计模型或专家系统执行特定的任务。通过巧妙地规则设计，早期AIGC 技术可以完成简单线条、文本和旋律的生成。例如，通过定义复杂的函数方程组，计算机所绘出的函数曲线具备某种美学图样；通过记录大量的问答文本，在面对新的问题时，计算机可以通过检索和匹配的方式生成简单的答案，甚至于改写故事。但是由于缺乏对客观世界的深入感知和对人类语言文字等知识的认知能力，早期的 AIGC 技术普遍面临所生成的内容空洞、刻板、文不对题等问题。参考人类的内容创作过程，研究人员们提出，理想的 AIGC 算法需要具备对数据内容的学习能力，在理解数据的基础上进行知识与分布的学习，最终实现高质量的内容创作。

深度神经网络在学习范式[2]和网络结构上的不断迭代极大的提升了人工智能算法的学习能力，从而推动了 AIGC 技术的快速发展。不同于传统人工智能算法，深度学习中的损失函数和梯度下降算法可以灵活快速的调整深度神经网络中的参数，从而实现从数据中进行学习功能。2012 年，卷积神经网络 AlexNet[1]凭借优秀的学习能力，在当年的 ImageNet 大规模视觉识别挑战赛中一举夺魁，比第二名传统机器学习算法的错误率提升 10.8 个百分点，开启了深度学习时代的序幕。就在紧随其后的 2013 年，深度变分自编码器[2]的提出让 AIGC 技术能力有了极大的进步。对于给定的神经网络，深度变分自编码器要

[2] 人工智能的学习范式是指人工智能模型从数据中进行学习的方法。

求网络的输出是对于输入内容的重建，通过重参数化等技巧，网络在重建过程中学习训练数据的统计分布。在测试阶段，变分自编码器通过在学习到的统计分布中进行采样，首次能比稳定的生成从未观测过的低分辨率图像。2014 年，一种新的博弈学习范式伴随着生成对抗网络[3]被提出。生成对抗网络由一个生成器和一个判别器组成，判别器致力于不断寻找生成数据和真实数据间的不同，生成器根据判别器的反馈不断完善自身，以求生成真假难辨的内容。得益于双方博弈的学习策略，生成内容的真实性和清晰度都得到了极大的提升，生成对抗网络也被应用于很多内容生成的具体应用。除了变分自编码器和生成对抗网络，强化学习[4]、流模型[5]、扩散模型[6]等学习范式均取得了喜人的进展，这些模型范式在不同场景中各有优势，让 AIGC 技术可以快速地应用到不同的场景和任务中。

深度神经网络的结构升级是推动 AIGC 快速发展的另一主要因素。一方面，实验证明，深度神经网络的学习能力和模型大小呈正相关，伴随着模型参数量的增加，相对应深度神经网络的能力一般会取得大幅提升。但是，随意地增加神经网络规模是行不通的，越大规模神经网络往往意味着更大的训练难度，因此深度神经网络的结构设计显得尤为关键。从早期的玻尔兹曼机，多层感知机，卷积神经网络，到深度残差网络和 Transformer 大模型，网络结构进化带来了深度学习模型参数量从几万到数千亿跃升，模型层数也从开始的个位数逐步发展到成百上千。深度学习模型规模上的量变引起了 AIGC 技术能力

的质变，在新型网络结构的加持下，上述的生成对抗网络等算法开始能生成超高清晰度的视频，高质量的文本段落和优美灵动的乐曲旋律。另一方面，研究者们在深度神经网络结构的设计中引入包含语义的隐式表达和物理知识，以降低模型的训练难度、增强生成内容的丰富程度。例如，研究者发现通过在神经网络的每一层引入隐式表达，能够极大地提升内容生成算法的可控性和生成效果[7]。另外，在三维数据的生成任务中，神经辐射场[8]在网络结构设计时充分考虑了物理世界的固有约束，极大提升了三维渲染效率和效果。

AIGC 要真正发挥对不同行业的驱动作用，需要与各行各业的特异性场景深度融合。在处理这些实际应用中，深度学习算法在感知、认知、模仿、生成等方向的基础能力决定了 AIGC 技术所能创作的生产力。近些年中，这些算法技术齐头并进、百花齐放，并最终形成了 AIGC 应用于不同场景的底层支撑。通过人工智能支撑技术的不断升级，AIGC 技术将持续赋能各类文化创意、生产生活、科学发现[9,10]等各种场景。

（二）AIGC 大模型架构潜力凸显

超级深度学习近年来的快速发展带来了深度神经网络技术在大模型和多模态两个方向上的不断突破，并为 AIGC 技术能力的升级提供了强力的支撑和全新的可能性。当前 AIGC 技术已经从最初追求生成内容的真实性的基本要求，发展到满足生成内容多样性、可控性的进阶需求，并开始追求生成内容的组合性。数字内容的组合性一方面

关注复杂场景、长文本等内容中各个元素的组合，例如虚拟数字世界中人、物和环境间的交互并组合生成为整体场景；长篇文字内容用词、语句、段落间的相互呼应和组合。另一方面，组合性追求概念、规则等抽象表达的组合，以此完成更加丰富和生动的数字内容生成，这些新出现的需求对传统单一模态的人工智能算法框架提出了新的挑战。近年来，研究界在大规模深度网络、多模态人工智能方面的探索表明大模型具备易扩展性，能够实现跨模态的知识沉淀，以大模型为基础模型，通过大模型小型化技术使得人工智能在小数据集场景下也能具备优秀的理解、生成和泛化能力，**具有超大规模、超多参数量的多模态大型神经网络将引领 AIGC 技术升级正在成为学界、产业界共识[3]。**

1. 视觉大模型提升 AIGC 感知能力

以图像、视频为代表的视觉数据是互联网时代信息的主要载体之一，这些视觉信息时刻记录着物理世界的状态，并在不断传播和再创作的过程中，反映人的想法、观念和价值主张。赋以人工智能模型感知并理解这些海量的视觉数据的能力[11]，是实现人工智能生成数字内容、数字孪生的基础；感知能力的提升，是实现生成视觉内容语义明确、内涵丰富、效果逼真的前提。

针对视觉信息的感知研究，在传统机器学习时代主要基于科研人员手动建模的特征和基于统计学习理论构建的朴素分类器，例如支持

[3] 百度文心大模型：https://wenxin.baidu.com/；OpenAI DALL·E 2 大模型：https://openai.com/dall-e-2/；智源研究院大模型：https://mp.weixin.qq.com/s/j8q018Lck1TWHO3NxQDiJQ

向量机模型（SVM），其能完成的任务类型和感知能力都非常有限；在深度学习时代，主要基于深度神经网络模型，例如深度残差网络（ResNet），其数据驱动的端到端学习范式使得模型的感知能力有了显著提升，在工业界也得到广泛的应用。但是，这类模型往往针对单一感知任务进行设计，很难同时完成多种视觉感知任务。如何解决不同场景、环境和条件下的视觉感知问题，并实现鲁棒、准确、高效的视觉理解，是 AIGC 技术必须要解决的挑战。

以视觉 Transformer （ViT，一种神经网络模型）[12] 为代表的新型神经网络，因其优异的性能、模型的易扩展性、计算的高并行性，正在成为视觉领域的基础网络架构，并且逐渐发展出来十亿甚至百亿参数规模的模型。在过去的 2-3 年间，视觉感知和理解技术正迎来突飞猛进的发展。无监督学习技术，包括对比式自监督学习（例如 SimCLR 和 MoCo 系列技术）和生成式自监督学习（例如 MAE 技术），能够大幅降低训练模型所需的有标注数据的数量。经过无监督预训练的深度神经网络模型，仅需要在少量的有标注样本上经过微调学习，即可在多种场景，线上线下均取得优异的性能。近年来基于 Transformer 衍生出来一系列网络结构，例如 Swin Transformer[13]、ViTAE Transformer[14,15]。通过将人类先验知识引入网络结构设计，使得这些模型具有了更快的收敛速度、更低的计算代价、更多的特征尺度、更强的泛化能力，从而能更好地学习和编码海量数据中蕴含的知识。这些新型的大模型架构，通过无监督预训练和微调学习的范式，

在图像分类、目标检测、语义分割、姿态估计、图像编辑以及遥感图像解译等多个感知任务上取得了相比于过去精心设计的多种算法模型更加优异的性能和表现[16,17]，有望成为基础视觉模型（Foundation Vision Model），显著提升场景感知能力，助力 AIGC 领域的发展。

基于视觉 Transformer 完成多种感知任务的联合学习是目前的研究热点。通过探索不同任务关联关系，挖掘丰富的监督信号，能够促使模型学习到更具泛化能力和可被理解的特征表示。此外，联合文本、语音等不同模态数据进行联合学习，探索不同模态数据的语义关联和信息互补，也是训练视觉大模型的重要路径。由此得到的视觉基础大模型在环境感知、内容检索、语义理解、模态对齐等任务上具备先天的优势，对于提升 AIGC 基础环境孪生能力、丰富 AIGC 应用场景具有重要价值。

2. 语言大模型增强 AIGC 认知能力

作为人类文明的重要记录方式，语言和文字记录了人类社会的历史变迁、科学技术和知识文化等。利用人工智能技术对海量语言、文本数据进行信息挖掘和内容理解是 AIGC 技术的关键一环。一方面，语言模型的训练和学习是进行文本生成的核心基础；另一方面，学习并理解人类语言将大幅丰富数字内容的生产能力，创新、丰富数字内容的生产方式，例如构建低门槛创作工具，使用户通过语言描述就能完成例如语言定位、语言编辑等高阶编辑操作。

在如今信息复杂的场景中，数据质量参差不齐、任务种类多，导

致数据孤岛和模型孤岛的存在，传统自然语言处理技术的不足尤为明显：模型设计、部署困难；数据难以复用；难以学习海量无标签数据挖掘、知识提取的共性能力。

对于传统自然语言处理技术的普遍问题，基于语言的大模型技术可以充分利用海量无标注文本进行预训练，从而赋予文本大模型在小数据集、零数据集场景下的理解和生成能力。基于大规模预训练的语言模型不仅能够在情感分析、语音识别、信息抽取、阅读理解等文本理解场景中表现出色，而且同样适用于图片描述生成、广告生成、书稿生成、对话生成等文本生成场景。这些复杂的功能往往只需要通过简单的无标注文本数据收集，训练部署一个通用的大规模预训练模型即可实现。研究者们相信基于语言的认知智能可以更快的加速通用人工智能的到来。例如，谷歌和 OpenAI 分别提出大规模预训练模型 BERT[18] 和 GPT[19]，在诸多自然语言理解和生成任务上取得了突破性的性能提升，验证了大模型在零资源、小样本、中低资源场景的优越性。紧随其后，国内外知名企业和高校均投入非常大的人力、算力、数据于自然语言处理大模型的研发，包括谷歌、微软、Meta、清华大学、斯坦福大学、华盛顿大学、卡内基·梅隆大学、京东、华为、百度等等。模型参数量也从最初的千万级发展到了千亿级别[20]，训练代价也从数十天增长到了不容忽略的几十万天（按在单张 V100 上计算）。

显然，指数级增长的成本换取的微弱增益让人们意识到，如何设计更高效率的自监督学习方法、更高参数效用比的模型架构、更绿色

节能的训练框架成为了大模型未来方向之一。在这个方向上，诸多机构开始了高效绿色的大模型探索之路，并且取得了显著的效果，如通用语言理解评估基准（GLUE）目前（2022 年 6 月）在榜第一名的是由京东探索研究院研发的 Vega v1 织女模型[4]，依托于预训练阶段多种文本粒度、语种类型、负采样方式上的自监督学习创新，实现了高效的数据知识提取，并采用了有理论支撑的更快捷的分布式优化器。此外，超级深度学习模型可以通过非常低成本的微调快速适应新的产业、领域、行业，实现跨模态、全链路的知识积累、沉淀、传播、复用。

基于语言的超级深度学习技术的发展趋势主要体现在训练模型的数据量日益增大、数据种类也更加丰富，模型规模增大、参数量以指数倍增加。通过不断构建语义理解能力增强、逻辑知识可抽象学习、同时适用于多种任务的语言大模型，将会对 AIGC 场景中的各项认知应用产生极大价值。

3. 多模态大模型升级 AIGC 内容创作能力

在日常生活中，视觉和语言是最常见且重要的两种模态[21]，上述的视觉大模型可以构建出人工智能更加强大的环境感知能力，而语言大模型则可以学习到人类文明的抽象概念以及认知的能力。然而 AIGC 技术如果只能生成单一模态的内容，那么 AIGC 的应用场景将极为有限、不足以推动内容生产方式的革新。**多模态大模型的出现，**

[4] https://gluebenchmark.com/leaderboard

则让融合性创新成为可能，极大丰富了 AIGC 技术可应用的广度。对于包含多个模态的信息，多模态大模型则致力于处理不同模态、不同来源、不同任务的数据和信息，从而满足 AIGC 场景下新的创作需求和应用场景。

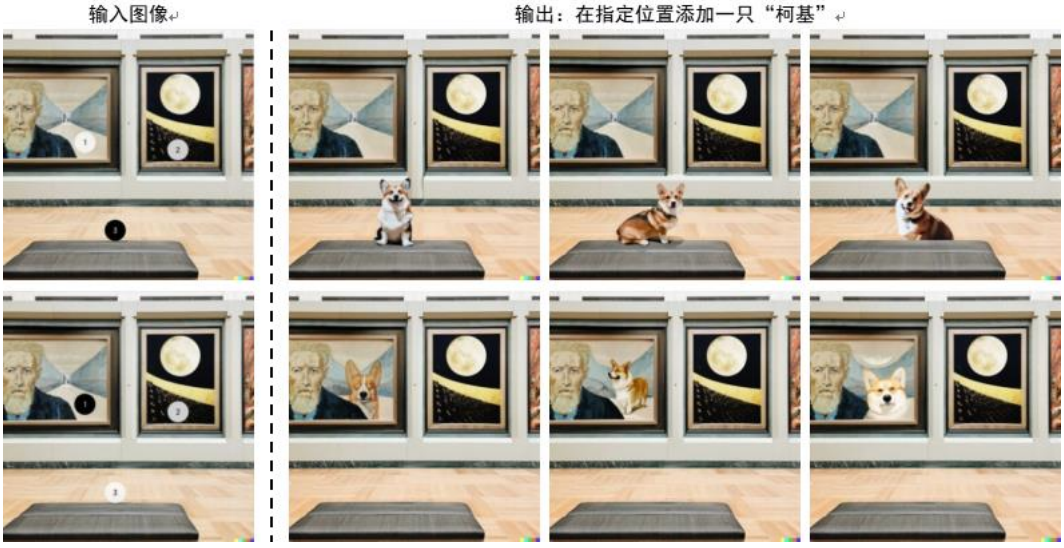
多模态大模型拥有两种能力，一个是寻找到不同模态数据之间的对应关系，例如将一段文本和与之对应的图片联系起来；另一个是实现不同模态数据间的相互转化与生成，比如根据一张图片生成对应的语言描述。为了寻找到不同模态数据之间的对应关系，多模态大模型将不同模态的原始数据映射到统一或相似语义空间当中，从而实现不同模态的信号之间的相互理解与对齐，这一能力最常见的例子就是互联网中使用文字搜索与之相关图片的图文搜索引擎。在此基础上，多模态大模型可以进一步实现不同模态数据间的相互转化与生成，这一能力是进行 AIGC 原生创作的关键。

来源：京东探索研究院

图 2　AIGC 多模态大模型生成结果图

如图 2 所示，只需给定用户简单手绘的语义图或是素描图，多模

态大模型学习模型便能够创作出逼真的风景图像，同时，当给定具体文本语义时，图像中的内容也将随之改变，展现出不同的季节亦或是“黄昏时河道干涸”的场景。再以 OpenAI 最新提出的多模态大模型 DALL-E 2 为例，给定一个已有的场景图像，该模型能够在指定位置添加指定的目标主体，如图 3 所示，当要求在沙发上（位置 3 处）添加一只柯基狗时，算法可以在指定位置添加不同形态的真实的柯基；当要求在左侧画框中（位置 1 处）添加一只柯基时，算法先是成功的识别出该位置是一幅画，并创作了符合相应画风的柯基狗[5]。基于多模态大模型，AIGC 具备了更加接近于人类的创作能力，并真正的开始展示出代替人类进行内容创作，进一步解放生产力的潜力。



来源：OpenAI

图 3　OpenAI AIGC 多模态大模型 DALL E 2 生成结果图

对于人工智能而言，能够高质量的完成多模态数据的对齐、转换

[5] https://openai.com/dall-e-2/

和生成任务意味着模型对物理世界具备了极为深刻的理解。从某种程度而言，基于多模态大模型的 AIGC 是人工智能算法迈向通用人工智能的重要一步。就好像人类通过不断的对比试错、总结归纳来了解我们身处的物理世界一样，多模态 AIGC 大模型也有希望能够自行总结客观规律，发展出认知与常识，进而帮助人类创造出新的数字世界。

（三）AIGC 技术演化出三大前沿能力

AIGC 技术被广泛应用于音频、文本、视觉等不同模态数据，并构成了丰富多样的技术应用。**本节归纳 AIGC 变革内容创作方式的三大前沿能力（如图 4 所示），分别是智能数字内容孪生能力，智能数字内容编辑能力和智能数字内容创作能力。**

来源：京东探索研究院

图 4 AIGC 的三大前沿能力

1. 增强与转译构建数字内容孪生能力

内容数字化是现今所有数字系统得以存在和运转的前提，其过程

是指将视觉、声音、文本等信息转化为数字格式。传统的数字化主要关注对传感器所采集数据的客观记录和储存，但容易忽略所记录的内容本身的完整性和相关语义。相比于传统的内容数字化，智能数字内容孪生技术致力于进一步挖掘数据中的有效信息，在深入理解数据内容的基础上，实现一系列高效、准确、智能的数字内容孪生任务。作为传统数字化的扩充和升级，数字内容的孪生技术受到了持续且广泛的研究。

智能数字内容孪生可大致分为智能增强技术和智能转译技术两个主要分支。考虑现实场景中数据采集、传输和储存中可能遇到的多种限制，原始的数字内容经常会存在缺失或者损坏等问题。智能增强技术旨在消除上述过程中的干扰和缺失问题，根据给定的低质量原始数据生成经过增强后的高质量数字内容，力求在数字世界中孪生并重构完整逼真的客观世界。在计算机视觉任务中，智能增强技术多被用于修复并增强由采集设备或环境因素引起的视觉内容受损，例如低分辨率、模糊、像素缺失等。同理，对于有缺陷的文本和音频数据，相关的智能增强技术被用于解决片段缺失、脉冲干扰和音频失真等问题，在实际生产生活中为相关应用生成复原高质量的数字内容。

除了对各种模态数据内容的修复和增强，近年间，数字内容孪生中智能增强技术在三维视觉领域取得了快速地发展。具体来说，数字图像是三维世界在摄影设备上的二维投影，传统的数字化记录了拍摄影像的色彩信息，但却无法保留三维世界中的深度、材质和光照等信

息。现有的数字孪生技术，可以利用对同一场景拍摄的多张照片，重构并生成相应的三维内容。最近，谷歌等多家国内外科技公司正探索使用互联网上商家和用户上传的照片，生成并渲染不同餐厅、街道和景点的三维全景。通过数字内容孪生中的智能增强技术，算法可以过滤剔除不同照片中天气、时间、行人等扰动信息，专注于生成并渲染不同场所的全时间段三维全景[22]。

数字内容孪生中的智能转译技术是建立在对客观世界内容感知的基础上，进一步理解孪生后的数字内容，从而实现多样化的内容呈现的一类技术集合。现阶段比较成熟的智能转译技术包括给定语音信号进行字幕合成，依据文字进行语音生成等。对于智能转译技术，放在第一位的是生成内容的准确性，无论是语音到文本还是文本生成语音，准确地呈现原始信息是该类技术走向实际应用的基础。在准确的基础上，为应对不同的使用场景，相关算法、工程人员还在不断地提高转译算法的实时性和生成语音的真实性。近些年间，智能转译技术已被越来越多地应用于社交、传媒、协同办公、残疾人辅助等实际场景中，为人们的生成生活带来更多的便利。

相比于较为成熟的语音/字幕合成，视觉内容描述[23]是近年间学术领域的热点研究课题之一。视觉描述技术致力于生成能够准确描述给定视觉内容（例如图像、视频等）的文本和语音。视觉内容描述技术可以被广泛地应用于赛事转播、智慧交通、影视娱乐等各类应用场景中。虽然现阶段的智能转译技术已经可以初步的描述图像(或视频)

中的人物、物体和环境信息，但如何能够准确地生成有关人物行为和主体关系的描述仍是现有技术亟需突破的问题。相比于智能增强技术，智能转译技术更加关注数字世界中不同模态的数字内容间相互理解、融合和转换的能力，从而丰富智能数字内容孪生技术的应用范围和灵活性。

数字内容孪生技术通过对真实世界中内容的智能增强和转译，将现实世界的物理属性（如物体的大小、纹理、颜色等）和社会属性（如主体行为、主体关系等）高效、可感知地进行数字化，实现现实世界到数字世界的映射，构建了在数字世界中重现现实场景的能力。通过数字内容孪生技术，不同行业的从业者可以更好地在数字世界中进行内容的组织和展示。

2. 理解与控制组成内容编辑能力

在数字内容孪生技术的基础上，智能数字内容编辑的相关技术构建了虚拟数字世界与现实物理世界间的交互通道。一方面，对数字内容的编辑和控制，例如数字人技术，可以直接作用于物理世界，实现实时的反馈和互动，起到对现实世界中主体陪伴或服务等功能；另一方面，数字内容编辑技术是实现数字仿真的基础。例如在自动驾驶仿真场景中，通过智能编辑，可以实现对同一道路上不同车况和天气状况的控制。基于数字内容仿真，算法模型可以在数字世界中学习到相

应的知识和技能，这些知识可以被用来反哺解决现实世界中的问题[6]。

从技术角度看，智能数字内容编辑主要通过数字内容的语义理解和属性控制两类技术来实现对内容的修改和控制。首先，理解数字内容是对其进行编辑和修改的必要前提。例如，在处理音频数据进行人声分离时，算法模型需要先理解输入的原始声音信号，才能进一步分离其中的人声信号和背景音，生成两段独立的音频内容。同理，对于计算机视觉中的图片、视频剪辑和自然语言处理中的摘要生成任务，都需要数字内容的语义理解技术进行相关语义的理解和概括，继而修改输入的原始数据以得到最终的生成结果。

值得注意的是，现实世界中的内容大多是由多种不同的语义信息组成的。例如，一张人脸照片实际上是由人物的身份信息、面部动作、拍摄视角、摄影设备和光照条件等许多语义信息一同决定的。早期的语义理解技术更多的是将某个内容当做一个整体进行理解，在学习到的数字表征中不同类别的语义信息往往是纠缠在一起的。虽然可以应用于解决某些数字内容编辑任务，但却难以对不同的语义进行精确的理解和修改。基于生成模型的可解耦语义学习技术是解决语义纠缠问题的可行解决方案之一，并在近些年间取得了快速的发展。通过理解并学习不同语义成分的变化，可解耦语义学习技术对数据内容具有更深刻的理解，并逐渐开始服务于人工智能试妆、试衣、生成同一个人

6 https://www.nvidia.cn/omniverse/media-entertainment/

不同年龄照片等新兴应用程序。

在充分理解数字内容语义的基础上，属性控制技术构成了数字内容编辑的另一主要分支。在语义理解的基础上，数字内容的智能属性控制技术将直接根据用户指定的属性，对原有的内容进行精确地修改、编辑和二次生成。常用的属性控制技术已经广泛地应用于智能图像编辑、文本情感改写和智能调音等多项应用中，并潜移默化地服务人们的生活，作为辅助功能提升内容创作者的效率。此外，先进的智能内容编辑技术结合了语义理解技术和属性控制技术，在处理三维动画内容时，在学习可解耦的视角、光照和角色等语义特征的基础上，智能属性控制技术以比传统算法更加高效且稳定的方式完成虚拟现实、游戏、电影中的渲染和操控[24]；在构造数字人时，属性控制能力可以根据实际需要快速地编辑数字人的外貌、音色、感情、表情等属性，以完成数字人技术在不同场合环境中的应用。

数字内容编辑技术在内容孪生技术的基础上，具备了对现实世界内容进行语义理解和属性操控的能力，从而构建了数字世界对现实世界内容的影响和反馈。在数字世界中的操作和尝试将不受限于场地、成本、资源消耗等客观约束，所得到的经验知识也能够更好地反馈给现实世界，提升生产生活的效率。

3. 模仿与概念学习造就内容创作能力

上述的数字内容的孪生和编辑能力主要面向客观世界中的真实内容，通过对现实内容的智能孪生、理解、控制和编辑，AIGC 算法

可以快速准确地将现实世界的内容映射到虚拟世界中，并通过控制仿真等方法，对现实世界产生正向的反馈和帮助。更进一步，**数字内容的智能创作旨在让人工智能算法具备类似甚至超越人的创作能力。**1968 年，毕加索曾这样评价计算机技术：“它们是没用的，只能简单的给出答案。”但在 54 年后的今天，百度已经可以通过人工智能模型进行绘画创作，并被西安美院的教授评价为具有“美院毕业生水平”，在短短 24 小时内就售出了 8700 多份，销售额超过 17 万元[7]。无需基于任何现实世界中存在的内容主体，基于人工智能算法的内容创作能力有望生成海量的原创数字内容。

按照技术的发展进程和实际应用的形态，数字内容的创作能力可划分为基于模仿的创作和基于概念的创作两类。基于模仿的创作需要人工智能模型首先观察人类的作品，通过学习某一类作品的分布特性，人工智能生成模型可以进行模仿式的新创作。以前文中提到的佳士得拍卖的肖像画为例，人工智能算法利用大约 15000 张创作于 14 世纪到 20 世纪的肖像画，从中学习作画的笔法、内容、艺术风格等。最终，人工智能内容生成模型所创作的肖像画通过了视觉图灵测试，让绝大部分人类都难以区分这幅画是艺术家创作的，还是人工智能的作品。不仅仅局限于智能作画，基于模仿的人工智能生成模型在旋律创作、文本写作和诗词创作等具体任务中都取得了不错的表现。对于某一类具体的内容，例如人物画像、押韵诗歌或乐曲旋律，现有的人工

[7] https://www.sohu.com/a/557118794_362042

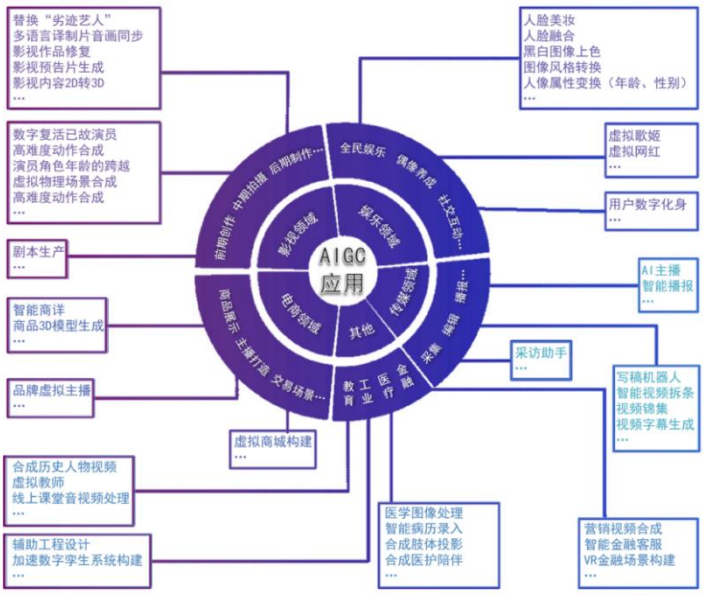
智能技术基本可以创作出让人真假难辨的数字内容。但同时，面对更加复杂的数据内容，例如三维数据、视频数据等，现有的技术所创作的内容相比于真实内容仍有一定差距，需要算法模型的不断完善来缩小这些内容的创作难度。

基于概念的创作不再简单的对固定种类的数据进行观察和模仿，而是致力于在海量的数据中学习抽象的概念，进而通过对不同概念的组合进行全新的创作。以文本到图像的生成为例，给定的文本不仅可以描述生成内容中需要包含的主体内容、数量和关系，还可以指定生成图像的风格、年代等属性。在现实世界中，人们可能只能见到“木头制作的椅子”，“狮子在捕猎獾鼠”等内容，但是通过文本描述，基于概念的创作技术可以创作出“牛油果制作的椅子”，“在猎捕狮子的獾鼠”等视觉内容[25]。在更进一步理解不同主体间动作、行为、和关系基础上，已经有相关的前沿研究开始尝试通过故事或者剧本描述，创作影视短片。总体来说，基于概念的智能创作与上述智能孪生中的转译技术不同，智能转译更关注对已有内容的精确表达和转换，而基于概念的智能创作是在给定模糊概念的基础上，进行自由生成和创作。数字内容基于概念的创作很大程度上依赖于算法模型对多模态数据的理解、对齐、融合和生成，依赖于人类社会中海量的数据以及相关的描述。基于概念的创作摆脱了对简单学习纹理、形状、颜色的模仿，进一步像人类一样开始学习和总结创作中包含的概念元素，实现更通用、更高效、更智能的 AIGC 应用。

伴随着深度神经网络的快速发展，人工智能模型的规模和能力都在不断被刷新，凭借着数据内容的快速增长，算力的爆发以及算法模型的不断迭代，数字内容创作技术突破到了一个新的高度，规模上不断变大，逐步趋近并开始超过人脑的神经元个数，能力上不断增强，展现出强大的多模态理解和生成能力。

三、人工智能生成内容的应用场景

在全球新冠肺炎疫情延宕反复的背景下，各行业对于数字内容的需求呈现井喷态势，数字世界内容消耗与供给的缺口亟待弥合。AIGC以其真实性、多样性、可控性、组合性的特征，有望帮助企业提高内容生产的效率，以及为其提供更加丰富多元、动态且可交互的内容，或将率先在传媒、电商、影视、娱乐等数字化程度高、内容需求丰富的行业取得重大创新发展。



来源：中国信息通信研究院

图 2 AIGC 应用视图

（一）AIGC+传媒：人机协同生产，推动媒体融合

近年来，随着全球信息化水平的加速提升，人工智能与传媒业的融合发展不断升级。AIGC 作为当前新型的内容生产方式，为媒体的内容生产全面赋能。写稿机器人、采访助手、视频字幕生成、语音播报、视频锦集、人工智能合成主播等相关应用不断涌现，并渗透到采集、编辑、传播等各个环节，深刻地改变了媒体的内容生产模式，成为推动媒体融合发展的重要力量。

在采编环节，一是实现采访录音语音转写，提升传媒工作者的工作体验。借助语音识别技术将录音语音转写成文字，有效压缩稿件生产过程中录音整理方面的重复工作，进一步保障了新闻的时效性。2022 年冬奥会期间，科大讯飞的智能录音笔通过跨语种的语音转写助力记者 2 分钟快速出稿。**二是实现智能新闻写作，提升新闻资讯的时效。**基于算法自动编写新闻，将部分劳动性的采编工作自动化，帮助媒体更快、更准、更智能化地生产内容。比如 2014 年 3 月，美国洛杉矶时报网站的机器人记者 Quakebot，在洛杉矶地震发生后仅 3 分钟，就写出相关消息并进行发布；美联社使用的智能写稿平台 Wordsmith 可以每秒写 2000 篇报道；中国地震台网的写稿机器人在九寨沟地震发生后 7 秒内就完成了相关消息的编发；第一财经“DT 稿王”一分钟可写出 1680 字[26]。**三是实现智能视频剪辑，提升视频内容的价值。**通过使用视频字幕生成、视频锦集、视频拆条、视频超分等视频智能化剪辑工具，高效节省人力时间成本，最大化版权内容