工程科学学报
Chinese Journal of Engineering

大模型及其在材料科学中的应用与展望

李长泰 韩旭 蒋若辉 员培文 胡鹏飞 班晓娟

Application and prospects of large models in materials science

LI Changtai, HAN Xu, JIANG Ruohui, YUN Peiwen, HU Pengfei, BAN Xiaojuan

引用本文:
李长泰, 韩旭, 蒋若辉, 员培文, 胡鹏飞, 班晓娟. 大模型及其在材料科学中的应用与展望[J]. 工程科学学报, 2024, 46(2): 290–305. doi: 10.13374/j.issn2095–9389.2023.09.20.002
LI Changtai, HAN Xu, JIANG Ruohui, YUN Peiwen, HU Pengfei, BAN Xiaojuan. Application and prospects of large models in materials science[J]. *Chinese Journal of Engineering*, 2024, 46(2): 290–305. doi: 10.13374/j.issn2095–9389.2023.09.20.002

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2023.09.20.002

工程科学学报, 第 46 卷, 第 2 期: 290−305, 2024 年 2 月
Chinese Journal of Engineering, Vol. 46, No. 2: 290−305, February 2024
https://doi.org/10.13374/j.issn2095-9389.2023.09.20.002; http://cje.ustb.edu.cn

大模型及其在材料科学中的应用与展望

李长泰[1,2)], 韩　旭[2)], 蒋若辉[2)], 贠培文[3,4)], 胡鹏飞[5)], 班晓娟[1,2,6,7)]✉

1) 北京科技大学北京材料基因工程高精尖创新中心, 北京 100083　2) 北京科技大学智能科学与技术学院, 北京 100083　3) 北京科技大学新材料技术研究院材料先进制备技术教育部重点实验室, 北京 100083　4) 北京科技大学新材料技术研究院现代交通金属材料与加工技术北京实验室, 北京 100083　5) 北京科技大学新材料技术研究院, 北京 100083　6) 北京科技大学智能仿生无人系统教育部重点实验室, 北京 100083　7) 辽宁材料实验室材料智能技术研究所, 沈阳 110004

✉通信作者, E-mail: banxj@ustb.edu.cn

摘　要　以大模型在材料科学中的应用为着眼点, 首先综述了大模型, 介绍了大模型的基本概念、发展过程、技术分类与特点等内容; 其次从通用领域大模型和垂直领域大模型两个角度, 总结了大模型的应用, 列举分析了不同种类大模型的应用场景和功能. 再次, 结合材料科学领域中的具体需求研究现状, 调研并综述了语言大模型、视觉大模型和多模态大模型在材料科学中的应用情况, 以自然语言处理和计算机视觉中的具体任务为切入, 参考典型应用案例, 综合提示工程策略和零样本知识迁移学习, 厘清了当前将大模型应用至材料科学的研究范式和制约因素, 并利用改进 SAM 视觉大模型在四种材料显微图像数据上进行了验证性图像分割与关键结构提取实验, 结果表明 SAM 带来的零样本分割能力对于材料微结构的精准高效表征具有巨大应用潜力. 最后, 提出了大模型相关技术、方法在材料科学中的未来研究机遇, 从单模态到综合性多模态的大模型研发与调优, 评估了可行性及技术难点.

关键词　大模型; 深度学习; ChatGPT; SAM; 材料科学; 多模态

分类号　TP391

Application and prospects of large models in materials science

LI Changtai[1,2)], *HAN Xu*[2)], *JIANG Ruohui*[2)], *YUN Peiwen*[3,4)], *HU Pengfei*[5)], *BAN Xiaojuan*[1,2,6,7)]✉

1) Beijing Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology Beijing, Beijing 100083, China

2) School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China

3) Key Laboratory for Advanced Materials Processing (MOE), University of Science and Technology Beijing, Beijing 100083, China

4) Beijing Laboratory of Metallic Materials and Processing for Modern Transportation, Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing 100083, China

5) Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing 100083, China

6) Key Laboratory of Intelligent Bionic Unmanned Systems, Ministry of Education, University of Science and Technology Beijing, Beijing 100083, China

7) Institute of Materials Intelligent Technology, Liaoning Academy of Materials, Shenyang 110004, China

✉Corresponding author, E-mail: banxj@ustb.edu.cn

ABSTRACT　Representative large models and their related applications, such as Bidirectional encoder representations from transformers (BERT), Generative pretrained transformer (GPT), Segment anything model (SAM), ChatGPT, DALL-E, Wenxin, and Pangu, have made astounding strides and exerted considerable influence across various fields domestically and abroad. They constantly attract the attention and follow-up of diverse societal sectors, including enterprises, universities, and research institutions. Large model applications have been successfully applied in scenarios such as biology, medicine, law, and social governance. Designing, modifying,

收稿日期: 2023−09−20

基金项目: 国家自然科学基金资助项目(U22A2022); 科技创新 2030-重大项目(2022ZD0118001)

and constructing domain-specific large models are crucial for truly harnessing their application value. Therefore, this paper provides inspiration for the application of large models in materials science. First, it provides an overview of large models, introducing their basic concepts, developmental process, technical classification, and features. Second, from the perspectives of the general domain and specific large models, this paper summarizes the applications of large models and analyzes the application scenarios and functions of various types of large models. Subsequently, considering the specific needs and current state of research in the field of materials science, this paper reviews the application of large language models, large visual models, and large multimodal models. It integrates engineering strategies and zero-shot knowledge transfer learning from specific tasks in natural language processing and computer vision and referencing typical application cases, clarifying current research paradigms and limiting factors for applying large models to materials science. To verify the effectiveness and potential of the visual large model, basal experiments of image segmentation and key structure extraction are performed on the microscopic image data of four types of materials using improved SAM, including Ni-superalloy, superalloy, polycrystalline pure iron grain, and Inconel 939. The experimental results reveal that the zero-shot segmentation capability of SAM has enormous potential for accurate and efficient representation of material microstructures. With the help of tailored prompt engineering, precise masks of the precipitated phase, grain boundaries, and cracks can be outputted without any label. Finally, this paper proposes future research opportunities for technologies and methods related to large models in materials science. This paper assesses the feasibility and technical challenges for the development and tuning of unimodal to comprehensive multimodal large models. With continuous innovations and collaborations, the horizon for large models in materials science seems boundlessly promising. The integration of these models can produce a new era of advanced research, leading to advancements that were previously considered unattainable. The symbiosis between materials science and large models can pave the way for unforeseen discoveries, enriching our scientific prowess.

KEY WORDS large models; deep learning; ChatGPT; SAM; materials science; multi-modality

人工智能(Artificial intelligence, AI)在各领域中的广泛应用从科研热点、社会关切、政策支持等维度都体现出极大的研究与应用价值[1]. 随着人工智能的土壤——数据的指数级增长以及计算能力的跃升, 以深度学习为代表的突破性人工智能算法不断涌现[2], 逐渐代替传统的机器学习和基于规则的方法, 并在众多场景下得以大范围实际应用[3−4], 如人脸识别[5]、自动驾驶[6]、文本生成[7]等. 2022 年底, OpenAI 公司发布 ChatGPT 应用并迅速进入大众的视野[8], 推出仅两个月后月活跃用户就已超一亿, 成为历史上用户群增长最快的消费应用. 基于语言大模型开发的人工智能产品 ChatGPT 被认为是人工智能技术的新突破, 吸引了社会各界的重点关注, 引发了国内外新一轮人工智能产品应用落地. 可以这样说, 以 ChatGPT 为时间起点, 人工智能正式进入“大模型时代”, 大模型也正在重塑各种任务并在众多复杂的下游任务中取得了不俗的成绩[9−11].

1 大模型概述

1.1 大模型

大模型(Large models, LMs)通常指具有数十亿、百亿甚至更多参数级别的深度神经网络模型[12], 其训练所需数据量远大于一般的深度学习算法模型(图 1). 大模型也可称为大规模预训练模型(Pre-trained models, PMs)或基础模型(Foundation models, FMs). 通常而言, 这种参数规模大、训练成本高的模型采用自监督学习范式(Self-supervised learning, SSL)获取强大且通用的数据表示, 其本身并不针对特定的下游任务, 而是获得对于训练数据的“理解”与“掌握”[13]. “大模型应用”表示将预训练得到的大模型通过迁移学习将获得的知识整合、迁移到各个下游具体任务, 并根据业务需求集成封装后的整体解决方案. 以 ChatGPT 为例, 其本身应被定义为基于语言大模型的生成式聊天应用, 它是在 GPT(Generative pre-trained transformers)系列预训练语言大模型的基础上经过复杂精调得到的商业化落地产品[14].

1.2 大模型相关技术

1.2.1 深度无监督表征学习

大模型的建立事实上是大规模深度无监督表征学习的结果[15]. 通过大量数据预训练后, 将模型参数作为下游任务的初始化参数并在相应任务的目标数据上进行微调训练的策略称之为预训练–微调策略[16]. 这种学习策略遵循着迁移学习的思想[17], 在自然语言处理任务中首先获得成功并逐渐影响计算机视觉任务相关方法的设计, 视觉自注意力模型(Vision transformer, ViT)[18]及相关变

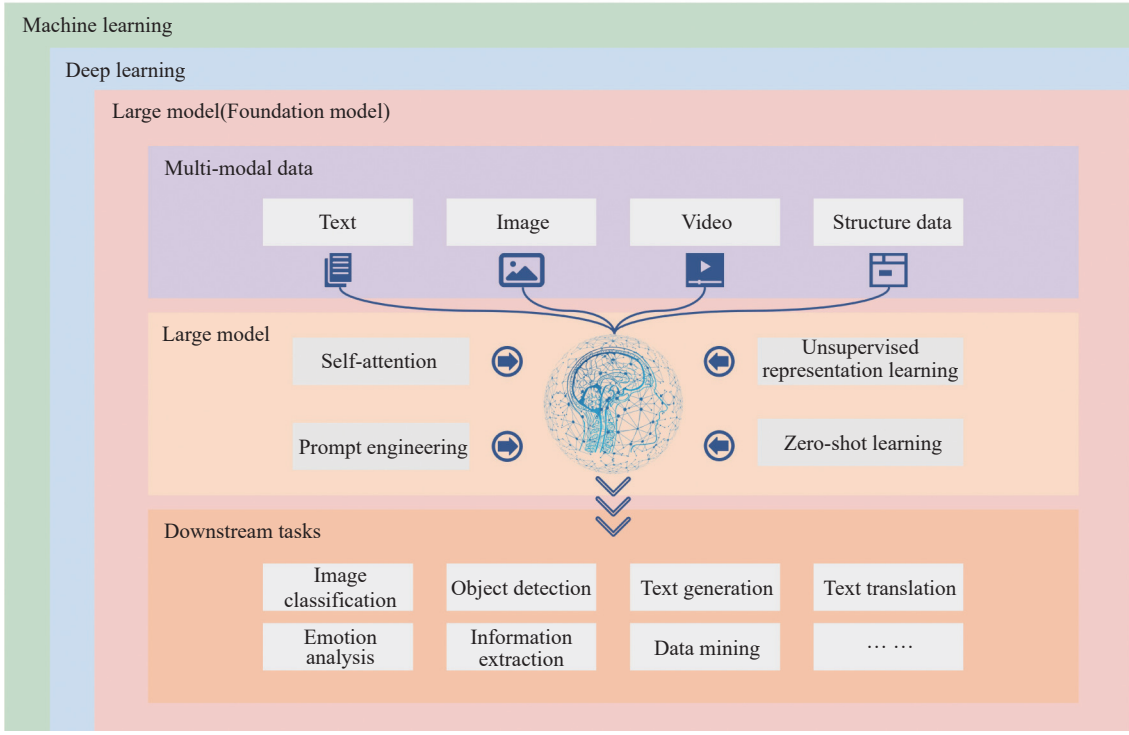

图 1　人工智能发展: 从机器学习到大模型

Fig.1　Development of artificial intelligence: from machine learning to large models

体模型[19]也凭借迁移学习的思想将图像分类[20]、目标检测[21]、语义分割[22]等视觉任务的成绩提升到史无前例的高度. 从预训练阶段数据的标注有无出发, 预训练可以分为有监督预训练、半监督预训练和无监督预训练[23]. 为实现无监督预训练, 解决模型训练时的标注受限问题, 自监督学习方法通过无监督代理任务预训练和有监督下游任务微调两阶段的结合, 平衡标注数量和模型精度, 取得了匹敌有监督预训练模型的效果[24]. 在大模型建立之后, 其下游应用以自监督学习范式为技术基础. 自监督学习旨在使深度神经网络具备感知数据自身的某些属性, 在预训练阶段, 模型不需要标注数据进行训练, 而是通过先前设计的监督信号和相应的网络结构开展学习, 这种学习模式避免了有监督预训练中仅仅拟合数据与简单且高度抽象的标注之间的映射关系, 能让模型学习到语义表达能力丰富、泛化能力强、更加稳健的数据表示, 从而更好地将学习到的“知识”迁移到具体的下游任务中, 产生正面的影响. 进一步地, 作为深度无监督表征学习的重要组成, 同样面向无监督预训练的对比学习方法[25](Contrastive learning, CL)正主导该领域的发展并且在以 CLIP(Contrastive language image pre-training)为代表的多模态特征对齐任务中发挥了关键的作用[26]. 由此可见, 深度无监督表征学习、自监督学习和对比学习在当前以及未来仍然是深度学习领域中的研究重点.

1.2.2　自注意力模型

大模型的建立过程离不开自注意力模型(Transformer)的作用[27]. Transformer 作为深度学习发展历史中的里程碑工作, 于 2017 年被谷歌公司的研究团队提出, 用于机器翻译任务并取得了良好的效果. 相较于其他网络模型如卷积神经网络(Convolutional neural network, CNN)[28], 自注意力(Self-attention, SA)机制能够有效地对数据(文本、图像等)中的长距离上下文关系进行建模从而获取更丰富、有效的数据表征. BERT(Bidirectional encoder representations from transformers)[29]、GPT[30]、T5[15]、XLNet[31]等语言大模型均采用基于预训练–微调策略的 Transformer 架构进行训练.

由于 Transformer 架构在自然语言处理领域体现出来的强大的数据表征能力和可扩展性, 研究者尝试将这些特性迁移应用到视觉任务中, 产生了一系列里程碑工作[32], 如视觉自注意力模型[18]和基于滑动窗口的分层视觉自注意力模型(Swin transformer)[33]等. Transformer 与自监督学习的结合同时为语言、视觉、多模态等大模型应用的发展奠定了坚实的技术基础. Caron 等[24]针对采用自监督学习方法训练的 ViT 模型所呈现出的数据表征特点进行了深入的研究, 探寻自监督训练中的监督信号对于整体训练结果的影响, 发现: (1)将自

监督学习与 ViT 结合, 模型能够学习到图像中明显的语义特征, 自注意力权重图中能够较为清晰地呈现出场景布局、目标边界等信息, 而这些信息是有监督 ViT 和 CNN 无法直接获取到的; (2)在无需微调训练的情况下, 自监督预训练得到的 ViT 能够在 k-近邻(k-Nearest neighbor, k-NN)分类任务上取得较好的结果. 结果表明自监督学习与 ViT 相结合有相当的潜力构造出如同自然语言处理领域 BERT 的视觉大模型. Oquab 等[34] 延续了该工作的思路, 将不同的训练技术结合并扩展了预训练阶段的数据量和模型规模, 通过提出的自动化流程建立多样精良的图像数据集, 在十亿级别参数量的 ViT 上进行训练, 并最终取得了比 OpenCLIP 更好的通用特征表示.

1.3 大模型的特点

大模型一般是在基于扩展性强的 Transformer 架构上经过大规模数据集训练的产物. 由于网络架构庞大、参数量很大, 大模型会具备一定的涌现(Emergence)能力[2]. 涌现能力是大模型的一个重要特征, 指经过训练后的模型出现了起初并未预期出现的能力, 这种能力在训练阶段没有被显式地引入, 而在各种下游任务中以各种形式展现出来. 以语言大模型 GPT-3[35] 为例, 在仅通过给其提供提示词(Pompt)的情况下, 它能够完成若干自然语言处理任务, 比如文本生成、段落总结等实现没有经过训练的任务, 且效果能够达到一定的水平. 此外, 在视觉大模型中, 能力较强的模型在经过大规模预训练后, 可以结合具体的提示词, 以各种形式(文本、点、框等)使模型具备零样本分析能力, 即对于之前未见过的样本产生一定程度上可以接受甚至出乎预期的结果[36].

提示词这一概念源于提示工程(Prompt engineering), 最初在语言大模型中使用, 提示词作为输入的一部分作用于大模型, 从而完成一系列特定语言任务[37]. 此外, 提示词的种类不局限于单种模态, 从最初的文本(Text)到 GPT-4[38] 模型的图像等都可以经过神经网络的编码, 作为辅助的信息与大模型融合. 随着自然语言处理领域中的技术逐渐成为视觉、通用大模型的研究重点, 提示工程也成为了大模型及其应用中必不可少的一环. 在视觉大模型中, 提示词不同, 得到的结果也不尽相同甚至大相径庭, 想要得到符合预期的图像任务结果, 需要针对相应数据和需求设计提示词. 因此深入探索提示工程相关方法有助于更好地了解大模型的能力和局限性.

除了设计和研发提示词以外, 提示工程还包含与大模型交互的思想和技术实现, 可以与大模型以及下游应用和系统进行交互对接, 来提高语言大模型的安全性[39], 也可以赋能语言大模型, 比如借助专业领域知识和外部工具来增强语言大模型能力. 将提示工程比喻为大模型应用过程中的催化剂, 按照设定好的提示策略, 引导大模型一步步给出想要的答案, 同时也可以将其视作不同大模型之间的黏合剂, 针对复杂任务场景(如多场景、多任务、多模态等)时, 能够起到有效衔接、信息交互的作用. 需要指出的是, 大模型相较普通深度学习模型具备的涌现能力可能会导致大模型给出远超预定范围的结果, 因而大模型及其应用的安全性方面应当引起足够的重视[40].

1.4 大模型的分类

依据训练数据和输出的模态种类, 将大模型主要分为语言大模型(Large language models, LLMs)、视觉大模型(Vision foundation models, VFMs)和多模态大模型(Large multi-modal models, LMMs). 表 1 总结了 16 个典型的大模型及其发布时间和参数量.

1.4.1 语言大模型

语言大模型也称为预训练语言模型(Pre-trained language models, PLMs), 从大规模文本数据中以自监督方法学习通用数据表征[47]. 自 2018 年自然语言处理领域涌现出 BERT[29]、GPT[30] 等代表性工作, 自监督预训练–微调的深度学习范式也随之成为领域内的主流路线.

BERT[29] 作为该领域内公认的开山之作, 突破了以往预训练语言模型的单向表征特性, 采用双向 Transformer 架构, 在计算单个词元(token)的自注意力权重时, 兼顾该 token 前向和后向所有的 token 的语义信息, 从而完成更合理有效的特征学习, 在 11 个自然语言处理任务中取得了良好的效果. BERT 构建了两种训练方式, 其一是利用掩码思想, 将文本中某个单词移除并使模型预测当前文本缺失的单词, 通过该简洁的分类任务学习数据表征, 但该方式更多注重在独立 token 上的表征, 在很大程度上制约了其在理解多个句子关联和整体层次结构的能力, 而这种能力对于段落理解、知识问答等任务是至关重要的. 另一种预训练方式是让模型判断两个子句是否有前后关系, 该任务称之为下一句预测(Next sentence prediction, NSP). 这两种任务以多任务形式同时训练, 均不需要人工标注, 只需要对收集到的句子进行基于规则的变换即可.

GPT 模型家族为自回归式语言模型(Autoreg-

表 1　大模型代表性工作

Table 1　Representative works on large models

Model	Release date	Provider	Modality	Parameters/billion	Ref
GPT-1	2018.06	Open AI	NLP	1.17	[30]
BERT	2018.10	Google	NLP	3.4	[29]
GPT-2	2019.02	Open AI	NLP	15	[41]
XLNet	2019.06	Google	NLP	3.4	[31]
T5	2019.10	Google	NLP	110	[15]
GPT-3	2020.05	Open AI	NLP	1750	[35]
DALL-E	2021.02	Open AI	Multi-modal	120	[10]
PALM	2022.04	Google	NLP	5400	[42]
DALL-E 2	2022.04	Open AI	Multi-modal	35	[43]
MOSS	2023.02	FDU	NLP	160	—
LLaMA	2023.02	Meta AI	NLP	650	[44]
GPT-4	2023.03	Open AI	Multi-modal	18000	[38]
SAM	2023.04	Meta AI	CV	10	[36]
DINOv2	2023.04	Meta AI	CV	11	[34]
PALM2	2023.05	Google	Multi-modal	3400	[45]
LLaMA2	2023.07	Meta AI	NLP	700	[46]

ressive LM), 经过了从 GPT-1 到 GPT-4 的迭代更新, 其训练任务可以概括为给定一句话, 预测下一个词, 直到句子结束. 这种训练方式再现了文本生成的过程, 使 GPT 系列模型在生成式任务中大放异彩, 也为其展现出的涌现能力和应用至 ChatGPT 提供了可能. GPT-1 在 2018 年被提出, 首次致力于通过无监督学习在 Transformer 架构上训练生成式语言模型, 预训练模型随后在下游任务上进行了微调[30]. 2019 年开发的 GPT-2 主要引入了多任务学习, 使用比初代模型更多的网络参数和数据进行训练, 使得预训练的生成式语言模型可以在大多数有监督下游任务上实现泛化, 无需进一步微调[39]. 为了进一步提高模型在少样本或零样本情况下的性能, GPT-3 将元学习与上下文学习相结合, 极大地提高了模型的泛化能力, 在各种下游任务上超越了大多数现有方法[35]. 此外, GPT-3 的参数规模比 GPT-2 增加了 100 倍, 它也是第一个超过 1000 亿参数规模的语言模型. 当前性能最强大的 GPT-4 模型可以接受图像和文本输入, 并产生文本输出, 在各种专业和学术基准测试中展现出与人类水平相当的能力[38].

1.4.2　视觉大模型

视觉模型的发展同样也遵循着预训练–微调范式, 在自监督学习席卷计算机视觉领域之前, 往往在大规模图像分类数据集 ImageNet[48] 上做有监督预训练[16], 在其他公开或私有的数据集上做下游任务(目标检测、语义分割等)的微调训练[23,49]. 下游任务表现的好坏通常会受到预训练学习特征的可迁移性、泛化性强弱的影响. 就像语言大模型一样, 视觉大模型预训练的数据、模型、任务选择不当也会对下游任务造成负优化, 降低下游模型的性能. 随着视觉模型的预训练–微调范式逐渐采用自监督学习模式, 由于代理任务(Proxy task)基于无标注数据开展训练, 其任务种类也得到了进一步丰富, 对比学习策略能够在大范围内获得较好的结果, 甚至超越了有监督的预训练策略[13], 因此在后续出现的重要视觉表征以及多模态特征对齐融合工作中大多采用了对比学习策略. 在这些模型中, 图像编码器可以被视作视觉大模型[50−51], 但其需要进一步微调训练, 无法直接利用视觉大模型产生预期的结果. 分割一切模型(Segment anything model, SAM)[36] 的横空出世改变了这一现状, 它率先将语言大模型中的提示工程策略有效移植到视觉领域, 大体上有三个特点: (1)由 SAM 产生的分割结果仅为无语义信息的目标掩码; (2)SAM 在图像分割的任务中需要定义提示词; (3)SAM 具备大模型的涌现能力, 具体体现在零样本分割的惊人结果.

1.4.3 多模态大模型

在以文本、图像、语音等单模态数据为基础的大模型发展的同时，多模态大模型也自然地获得研究人员的关注[52−53]. GPT 模型在第四代版本中引入了文本–图像两种模态输入的机制，旨在实现基于图文对齐的图像描述、论文总结等功能. 由于生成式 AI(Artificial intelligence generated content, AIGC)、大模型等概念的实践与落地，人们开始对通用人工智能(Artificial general intelligence, AGI)[54]的实现予以期待，而多模态大模型或通用大模型被视为通往 AGI 的必由之路. 不同于单模态大模型仅在特定类型的数据集上进行训练，多模态大模型主要解决的问题是如何将两种以上类型的数据表征有效融合形成一个整合的单模型以适配不同数据模态的输入和输出[55−56]. 在大量的多模态数据上训练意味着需要更多更强大的计算资源，而随着相关硬件的不断迭代增强，训练真正意义上的通用大模型成为现实可能.

2021 年初，CLIP[26] 与 DALL-E[10] 相继被推出，CLIP 能够基于文本实现判别式任务——图像分类；DALL-E 能够根据文本实现生成式任务——图像生成. 这标志着无监督预训练模型正式在多模态领域开始发力. CLIP 在新构建的 WIT(WebImageText)数据集[26](通过网络爬虫获取的四亿个图像–文本对)上分别设置文本编码器(Text encoder)和图像编码器(Image encoder)提取相应的隐空间向量，并使用对比学习策略进行训练优化，判断文本和图像是否属于同一对. CLIP 通过无监督的训练方式，打破了以往固定数据量、数据标注的限制，计算图文之间的相似度，通过对比学习进行两者之间的语义对齐，为少样本甚至零样本的任务提供可行的解决方案. 紧随其后，在计算机视觉领域，零样本图像分类、零样本目标检测、零样本语义分割等任务相继被定义且发展. 在生成任务方面，DALL-E 利用 2.5 亿余个图文对进行训练，最终实现根据给定的文本描述生成相应的图像. 时隔一年，其优化版本 DALL-E 2 进一步优化了生成图像的质量，能够产生原创、真实和绘画等风格的图像，且生成图像的分辨率能够达到初代版本的四倍之多，由其产生的 AI 绘图工具在绘画、设计等领域引起强烈反响. 此外，为了综合利用各种模型的优势，Zhang 等[57] 提出了一种级联大模型的策略，可以概括为“提示、生成并级联”，利用 CLIP 进行文本对比学习、DINO 进行图像对比学习、GPT-3 进行文本生成学习以及 DALL-E 进行图像生成学习，并将这四种预训练方法对图文多模态数据进行知识理解与融合，形成 CaFo(Cascade of foundation models)大模型并在小样本分类任务上取得最佳的结果. 值得说明的是，当前的众多多模态大模型仍存在模态数量少、数据对齐难、场景难以设定等问题. 可喜的是，在生物医疗[58]、科学发现[59]、社会治理[60] 等多模态数据广泛大量存在的领域，已经有相应的大模型解决方案.

2 大模型的应用

与其他深度学习技术相比，大模型展现出强大的涌现能力和零样本知识迁移的潜力，大模型在各个领域取得了引人瞩目的成绩，如表 2 所示，对建立材料科学领域内专用的包括文本、视觉和通用大模型具有较大的指导与借鉴意义.

2.1 通用领域大模型的应用

2.1.1 语言大模型应用: ChatGPT

ChatGPT 是以 GPT3.5 语言大模型为基础的，因此其具备 GPT-3.5 的技术特征，能够实现面向开放域的多轮对话和可用的、符合一定逻辑的生成式文本. 不同于之前的聊天对话机器人只支持在单轮对话中正常响应，ChatGPT 可以不断回溯上下文内容，将每一轮历史对话的信息和当前用户追问的信息同时纳入模型，学习并整合用户多轮对话信息，逐轮聚焦、精准理解用户需求，自动生成新的预测序列，并进一步结合已习得的海量数据、具体对话语境，逐步预测回复文本的各个字词，并生成新的回复文本，以提供多类型、多领域的准确响应，展现出其具备概念理解、推理决策、拟人交流等“类人”现象. 正是 ChatGPT 的这种特性颠覆了人们对传统的聊天机器人“答非所问”“记忆消失”等印象，开始对 ChatGPT 的智能人机交互场景有了新的想象.

OpenAI 并未公布 ChatGPT 的精调细节，当前对 ChatGPT 精调过程的认知基本上来源于先于 ChatGPT 发布并披露技术细节的 InstructGPT[61]，该工作被视为 ChatGPT 的“姊妹工作”. 目前，学界通常认为从 GPT-3.5 语言大模型到 ChatGPT 的诞生经历了两个重要的精调过程: (1)上下文学习(In-context learning, ICL). 在 GPT-3.5 模型的训练中，作为一种包含内部循环的元学习(Meta learning)方法，ICL 可以对更多的上下文信息进行建模来解决特定任务，不仅可以提高各种任务的效果，还可以更好地应对零样本和少样本的学习场景. 在 ICL 的支持下，GPT-3.5 系列模型无需对 NLP 任务进行任

表 2　国内外大模型代表性应用

Table 2　Representative application of large model

Application	Release date	Provider	Domain	Website
Pangu-Weather	2021.04	HUAWEI	Weather	https://github.com/198808xc/Pangu-Weather
ChatGPT	2022.11	OpenAI	Chatbot	https://chat.openai.com/
Bard	2023.02	Google	Chatbot	https://bard.google.com/
NewBing	2023.02	Microsoft	Chatbot	https://www.bing.com/
ERNIE Bot	2023.03	Baidu	Chatbot	https://yiyan.baidu.com/
Tongyi Qianwen	2023.04	Alibaba	Chatbot	https://qianwen.aliyun.com/
Pangu	2023.04	HUAWEI	Versatile	https://www.huaweicloud.com/product/pangu.html
Xinghuo	2023.06	iFLYTEK	Chatbot	https://xinghuo.xfyun.cn/
Wudao3.0	2023.06	BAAI	Versatile	https://www.baai.ac.cn/
360 ZhiNao	2023.06	360	Versatile	https://ai.360.cn/
Zidong Taichu	2023.06	CAS	Versatile	http://taichu.ia.ac.cn/
Med-PALM M	2023.07	Google	Medicine	https://sites.research.google/med-palm/
Bai Yulan	2023.07	SJTU	Chemistry	http://www.baiyulan.org.cn/
TransGPT	2023.07	BJTU	Transportation	https://github.com/DUOMO/TransGPT
Hunyuan	2023.08	Tencent	Chatbot	https://hunyuan.tencent.com/
Xinghuo v2	2023.08	iFLYTEK	Versatile	https://xinghuo.xfyun.cn/

何训练和微调, 就可以取得很好的效果, 在文章生成和代码编写等需要具备创意和逻辑思维的任务重中取得了惊人的效果. (2)基于人类反馈的强化学习(Reinforcement learning from human feedback, RLHF). 强化学习(RL)在游戏[62]、控制[63]、决策[64]等领域取得了一定的效果, 其主要思想为使智能体与环境的不断交互中最大化其能获取的奖励. 对于语言大模型, 需要对其生成内容进行“好坏判断”以符合健康、安全等客观使用要求. 对于 ChatGPT 生成的内容, 无法使用预先定义的规则判断其好坏, 而人工评价所有的生成内容代价极其昂贵, 因此专门训练一个奖励函数, 根据部分人工标注的数据, 给出对于生成内容的定性评价, 从而完成整个强化学习的最优化.

2.1.2　视觉大模型应用: Segment anything model

SAM(Segment anything model)是计算机视觉领域首个通用图像分割视觉大模型并正在成为大模型应用新的研究增长点[11]. Meta AI 于 2023 年 4 月公开了该模型的技术细节, 如图 2 所示, 介绍了提示分割(Promptable segmentation)的图像分割新任务及与该任务对应的分割模型和一种数据收集引擎. SAM 使用超大数据集在海量带标注图像数据(十亿级)的训练下, 凭借简洁的模型设计和提示工程的构建与引入, 能够实现比肩有监督学习的效果, 并且涌现出强大的零样本分割能力. 实验表明, 其预训练模型能够普遍适用于多种下游任务, 并为其提供零样本分割能力, 对全新的图像实现精准的物体、边界等目标的像素级识别. SAM 主要基于一种交互式分割任务设计, 对于一张图像, 通过提供分割提示(如一个或多个点、框)或由算法自动生成提示, 使用预训练的模型进行零样本图像分割. 给定一组提示, 若待分割对象存在歧义, SAM 会自动生成最多三个逐步细化的有效分割结果. 得益于其强大的零样本分割能力与交互式分割原理, SAM 能够在不进行额外训练的情况下, 对未见过类型的对象保持较好的分割性能, 使得用户可以按需分割感兴趣的对象和区域.

2.2　垂直领域大模型的应用

2.2.1　SAM 在视觉领域中的迁移应用

在计算机视觉领域, 基于 SAM 的探索性工作不断产生, 以检验其在不同任务中的适用效果, 部分工作深入探讨了 SAM 在不同图像分割任务、数据集上效果优劣的原因与可能的改进方向. 本节主要介绍 SAM 在医学图像分割[65]、精细化图像分割[66]、遥感图像分割[67]不同领域工作中的发展情况和典型工作.

医学图像分割是计算机视觉中非常重要的领域, 对现代医学数据分析与处理有重要意义. 在医

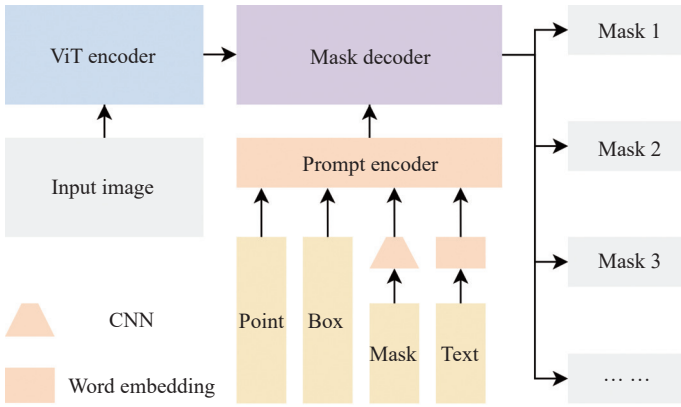

图 2 SAM 结构图

Fig.2 Schematic of the SAM

学图像分割任务中, 模型需要为医生提供目标病灶区域的病理结构或分割结果, 以辅助计算机诊断分析或智能临床手术. 随着基于深度学习的医学图像分割方法发展, 其准确度和速度都相对于传统方法有了巨大进步. 但目前基于卷积神经网络或 Transformer 的方法都只能针对特定任务, 缺乏对其他任务的泛化能力. 在此背景下, SAM 的提出提供了一种通用的分割框架成功扩展到了医学图像分割任务中并得到全面的利用与改进. 其包括三种类型: (1)SAM 辅助标注工作. 医学图像标注需要理解专业知识的专家花费大量时间, 而 SAM 提供了两种方便、快捷的标注方式, 提升医学图像处理效率. 一是用户手动为 SAM 提供提示, SAM 生成分割结果供用户选择或修改; 二是算法自动生成提示并分割出多个待选择对象, 用户在其中挑选并进行修改. 两种方式都可以在算法或流程方面进行优化. (2)SAM 与其他分割模型协同. SAM 具有根据提示零样本分割多个结果的能力, 基于此特性可以使用其他模型作为先验方法, 为 SAM 提供对象点或边界框提示; 此外, SAM 还能够为其他专用有监督模型的训练提供精确的分割掩码, 以低成本扩充标注数据集, 将其直接或经过修改后进行迭代训练. (3)特定任务的微调. SAM 的预训练模型经过大量数据集训练, 但在特定的任务或数据集上表现仍有待提升. 可以通过使用少量特定任务中的标注数据, 微调预训练模型, 使 SAM 保持原有零样本分割能力的同时对特定领域的医学图像分割效果更好. 大量实验结果表明, 医学图像分割任务中, 手动给予模型提示进行分割的效果由于全自动分割, 即 SAM 在依赖人类先验知识的情况下能显著提升模型性能, 且 SAM 在分割边界清晰的器官或病灶区域方面效果良好, 但在准确识别无定型病变区域方面仍存在困难.

精细化图像分割同样是受到重点关注的领域, 对于细节部分的分割效果和模型可解释性都是重点研究方向. Ke 等[66] 提出了 HQ-SAM 方法, 在原 SAM 模型基础上添加了一层并行编解码结构, 仅引入极少参数量的情况下, 显著提高了 SAM 对复杂结构、多对象目标输出的分割效果, 并且在微调过程中, HQ-SAM 学习到的 token 和 MLP 层不会过拟合特定数据集的标注偏差, 避免了微调后 SAM 灾难性遗忘的问题, 提升了 SAM 在新数据集上的效果. 类似这样的探索不仅证明了 SAM 在千万量级数据上预训练的效果具有较好的零样本分割效果, 而且也说明大规模预训练得到的模型需要通过合适的方法进一步挖掘(结构修改或微调)才能在全新数据集上实现对未知对象的准确、精细化分割.

SAM 在其他图像分割领域同样有不俗的表现. Chen 等[67] 从 SAM 的提示工程任务出发, 分析遥感图像本身与遥感图像分割任务的特点, 引入了附加的提示生成器, 为 SAM 提供了更良好的先验提示, 从而提升了 SAM 模型在遥感图像实例分割任务中的能力. 该工作在大量卫星遥感数据上进行了实验, 并证明在给予合适提示的情况下, SAM 面对大分辨率、多目标的图像仍可以保持很高的分割精度. Wang 等[68] 在声呐数据上进行实验, 此类数据往往成像模糊、包含较多杂质, 属于困难图像分割任务, 该工作等在水下声呐图像、导弹声呐图像等数据集上对 SAM 进行微调, 结果相对于 SAM 本身和其他深度学习方法都有较大进步, 但仍有提升空间. 该工作证明了 SAM 在声呐成像场景中的应用潜力, 也为其他低分辨率、模糊场景图像分割任务提供了有价值的参考.

综上所述, 视觉大模型 SAM 开创了图像分割领域的全新思路, 在多种应用和场景中展现出巨大的优势. 大量工作证明 SAM 在不同领域的图像分割任务中有良好的泛化性, 也为包括材料显微图像分割在内的领域视觉数据集处理与分析任务提供了新的解决思路: 零样本全自动分割、辅助标注等任务存在可行性, 并且针对领域任务中数据量小的特点执行模型微调、结合专家知识提供高质量的先验提示, 都能够进一步提升 SAM 的效果.

2.2.2　其他大模型应用

2023 年 7 月 5 日, 《Nature》在线发表中国华为公司主导的盘古大模型(Pangu-Weather)用于精准中期全球天气预报, 提出了嵌入地球特定先验知识的三维深度神经网络, 利用地球特定位置偏差捕捉绝对位置对天气的影响[11]. 盘古大模型在训练过程中使用了大量的气象数据, 包括 1979—2017 年的 ERA-5 数据和 2017—2021 年的 TIGGE 档案数据. 训练完成后, 盘古大模型能够进行 1、3、6 和 24 h 的全球天气预报. 对于再分析数据的验证, 盘古大模型在所有的测试变量中, 均取得了比来自欧洲中期天气预报中心(European centre for medium-range weather forecasts, ECMWF)更好的精确预报, 并成功跟踪预测了极端天气飓风的路径, ECMWF 也上线了人工智能天气预测系统, 将 Pangu-Weather 模型投入到实际的预报工作中.

在交通领域, 北京交通大学联合多家机构发布国内首款开源可商用的交通大模型 TransGPT·致远, 分为单模态和多模态两种, 能够实现交通情况预测、智能咨询助手、公共交通服务、交通规划设计、交通安全教育、协助管理、交通事故报告和分析、自动驾驶辅助系统等功能, 致力于在真实交通行业中发挥实际价值. TransGPT 的训练数据主要包含通用预训练数据集合交通领域数据集(非对话式的领域预训练数据集合用于 RLHF 精调的领域微调数据集), 经过训练后, 能够具备行业常识, 为道路、桥梁、隧道等工程和公路、水路、公共交通等运输方面提供知识与经验, 并且能够以此为基础, 在特定的交通应用场景中落地使用.

在医学领域, 海量的复杂多模态数据和专业医生资源分布不均等因素在客观上为发展领域专有的人工智能算法和大模型开发提供了现实需求和一定的数据基础. Huang 等[69] 针对公开的医疗图像缺乏有效的标注从而限制数据驱动方法研究与创新的问题, 介绍了面向病理图像分析的视觉–语言大模型尝试收集在网络社交媒体中大量由专业医生分享的病理图像及其文本评论, 构造了一套包含 208414 对病理图像和自然语言描述的数据集 OpenPath, 并提出了病理图像–文本特征对齐的预训练策略 PLIP 分别在病理图像零样本分类下游任务和线性分类器训练任务上取得了相较通用预训练模型更好的性能. 此外, 基于 PLIP 预训练策略, 开发了相似病例检索系统, 用户可以通过病理图像或描述语言获取到相应的病例资料. 该工作对病理诊断的知识分享和科研教学具有重要意义, 同时在医疗图像领域也验证了利用公开分享数据开发领域专用的大模型的可行性与有效性. 谷歌公司的 DeepMind 团队提出了多模态生成式通才生物医学大模型[70], 旨在将广泛存在的包括图像、文本、组学等多模态数据进行大规模的编码、整合与解释, 构建新的多模态生物医疗测试基准 MultiMedBench, 囊括医疗问答、细胞图像解释、病历生成、基因组变异检测等任务. 该团队基于通用领域的 PalM 大模型[40], 研发了 Med-PalM M 通才生物医疗大模型, 实现单个模型的多模态任务输出, 且测试结果比对应的单模态模型更优. 此外, 报道了对于生物医疗的新概念、新任务、迁移学习和因果推断等涌现性案例.

3　大模型在材料科学中的应用现状与机遇

大模型的应用结果和价值取决于具体应用场景的设计是否合乎大模型本身的技术特点, 且是否具备足够的可用数据或能够结合相应场景开发稳定、自动化的数据生产机制. 从通用领域大模型到材料科学领域大模型的迁移, 需要综合考量材料科学中的常见任务场景、数据类型、应用需求等方面, 目的是改进材料研究的效率和精度. 因而大模型在材料科学中的应用关键点包括: (1)材料信息的自动提取与挖掘. 大模型在材料科学中的一个重要应用是自动提取和挖掘文献中的关键信息. 通过训练大模型, 研究人员可以将其应用于海量文献数据, 快速识别出材料性质、制备方法、性能评价等关键信息. 这不仅节省了研究人员大量的时间, 还有助于建立更全面的材料数据库, 为材料设计提供更多的参考信息. (2)材料性能预测与优化. 大模型可以将已有的实验数据输入模型进行训练, 学习材料性质与结构之间的关联, 从而预测新材料的性能. 这种方法有助于加速材料设计过程, 减少试验和错误的成本, 并推动新材料的发现和开发. (3)新材料的发现. 研究人员可以设计特定的搜索任务, 使模型自动生成具有特定性质

的新材料结构. 这种基于生成的方法为材料科学带来了全新的思路, 有望推动材料领域的突破性进展. (4)材料知识图谱的构建. 大模型不仅可以处理文本信息, 还可以通过分析文献中的关系和信息构建材料领域的知识图谱, 揭示材料之间的关联和属性. 这有助于更深入地理解材料科学领域的知识体系.

大模型在材料科学领域的应用正方兴未艾, 应用场景主要分布在相对成熟的语言大模型、初见效果的视觉大模型和难度较大的多模态通用大模型中.

3.1 材料科学领域语言大模型的应用

材料科学领域是一个多细分领域且各领域相互关联影响的综合性学科[71]. 海量的科技论文是宝贵的信息载体与数据来源, 这些大量数据势必为材料发现提供新的研究思路. 由于通用领域语言大模型 BERT 在命名实体识别、问答、关联分类等任务中取得了优异的成绩[45], 材料科学领域基于 BERT 修改的领域语言大模型也基本遵循 BERT 的设计思想, 利用可大量获取的科技论文预训练之后在具体的任务上进行微调, 结合自然语言处理技术, 构建材料科学领域专用的语言大模型.

Trewartha 等[72] 将双向长短期记忆神经网络(Bidirectional long short-term memory, BiLSTM)模型与三种具备材料科学知识的 BERT(具有通用知识)、SciBERT(具有一般科学知识)[59] 和 MatBERT(具有材料学知识)进行比较. 实验结果如同预期, MatBERT 总体表现最佳, 意味着具有更多材料科学知识的语言模型能够在材料科学相关任务中表现更好, 甚至比 BERT 更加出色. 将 MatBERT 应用于相关信息提取任务中, 其高质量的结果可以有效地提高材料科学文献中收集信息的效率. Gupta 等[73] 指出通用领域的语言大模型无法在材料科学领域取得良好结果的主要原因是训练它们的数据缺乏材料科学专有的术语和表达方式, 因此提出并开源了 MatSciBERT 模型, 基于改进的 BERT 语言大模型, 在大量经过同行评议的材料科学文献数据集上进行训练, 在命名实体识别、关系分类和摘要分类三个主要任务上, 取得了相较于科学领域语言大模型 SciBERT 更高的准确度.

需要指出, 在材料科学、生物、化学[74] 等领域, 描述符(Descriptor)作为代表数据特征的结构化数据, 对基于学习的方法实现至关重要. 如何从强领域知识的原始的文本信息中提取、分析并转换得到准确的描述符是一个需要重点关注的问题. 目前, 通常使用命名实体识别的方法进行描述符的抽取, 基于 BERT 的各种大语言模型都将该任务作为考察模型性能的重要因素. 来自加拿大蒙特利尔大学米拉分校与英特尔公司的研究人员建立了针对材料科学的自然语言处理任务测试基准 MatSci-NLP[75], 详细描述了材料科学中的自然语言处理任务, 包括: (1)命名实体识别(从材料科学文本中识别材料、描述符、属性等关键信息); (2)关系分类(判别两个文本之间的逻辑关系); (3)事件参数提取(提取材料科学中的事件参数与参数角色); (4)段落分类(判别某个段落是否属于某种材料科学细分领域); (5)合成动作检索(识别材料科学中合成动作的种类并形成完整的材料合成过程); (6)句子分类(判断某个句子的相关实验事实); (7)槽填充(材料科学实体集中从特定句子中提取槽填充符, 并在实验过程的文本中预测并填充槽内容). MatSci-NLP 使用高质量且公开可用的材料科学文本数据, 并结合自然语言处理任务的特点提出能够公平、规范、有效地评测世界范围内面向材料科学设计开发的语言模型, 从而使相关科研人员将更多的精力投入到语言模型本身的开发中, 更加直接地探究语言大模型对材料服役行为、性能提升等关键问题的影响, 助力新材料体系的构建与发展.

3.2 材料科学领域视觉大模型的应用

材料显微图像的处理与分析对样品的表征起到重要的基础作用[76]. 材料显微图像能够直观展示材料内部结构的形貌特征与空间分布, 建立材料微观结构与宏观性能的关联关系一直是材料科学的研究重点[77]. 随着实验仪器与成像技术的更新迭代, 不同种类、不同模态的材料显微图像数量呈几何倍数增长[78], 精准、高效地提取显微图像中的关键组织与结构特征往往需要投入大量的人力, 且受人工操作的主观因素影响. 近年来, 由于深度学习与计算机视觉技术的快速发展, 研究人员针对材料显微图像的本征特点研发出一系列的方法技术、相应的软件平台也成功集成了这些算法, 供材料科研人员使用[79−80].

目前尚未建立针对材料显微图像的视觉大模型, 但一些工作尝试借助通用大模型的预训练–微调策略, 使用大规模数据集建立材料显微图像专有的视觉预训练模型以实现微结构精准分割. Stuckner 等[81] 提出了材料显微图像大规模数据集 MicroNet 包含多个公开及私有数据集中 54 种材料的 110861 张显微图像, 每张图像的分辨率为 1048×741 像素,

远大于自然场景图像分类数据集 ImageNet 的 469×387 像素. 基于此数据集和相应的测试基准, 设计了材料显微图像类别判断的预训练策略并在多种编码器架构上开展预训练, 通过迁移学习, 将预训练的编码器作为下游图像分割任务中的特征提取器进行微调, 在包括 U-Net、DeepLabV3+等卷积神经网络模型上均取得良好的测试结果. 重要的是, 相比在通用数据集 ImageNet 上的预训练–微调结果, 该方法展现出较强的模型泛化能力, 在与训练阶段不同的成像条件和样品上, 即数据分布与训练数据不一致时, 图像分割的准确率更高. 此外, 在仅使用单张带标注图像进行微调训练时, 经过 MicroNet 数据集预训练的模型仍然能输出质量较高且完全可用的分割结果. 该工作使用的网络模型为 CNN 架构, 仅关注局部的空间关系, 并没有采用 ViT 架构学习长距离的空间关系. Alrfou 等[82]在该工作的基础上做了进一步的验证和探索. 首先, 网络模型由 CNN 架构改变为 CNN+ViT 架构, 预训练阶段使用卷积块和 Swin transformer 块共同学习数据特征, 下游图像分割阶段将两种网络共同作为初始编码器并将各自输出的特征进行深度融合, 同时 Swin transformer 网络单独作为初始解码器, 这种将 CNN 和 ViT 同时用在模型学习、迁移阶段的策略能够有效结合各自网络学习的侧重点, 互为增益. 其次, 类似实验证明了使用 Swin transformer 网络的模型在构造的数据集中(约五万张材料显微图像)能够取得更好的分割效果, 且在少样本甚至零样本的极端条件下, 输出较为精确的分割结果.

然而这两种均属于传统的视觉预训练模式, 需要进一步微调训练才能达到图像语义分割的效果. 而类似 SAM 从超大规模数据集构造、大模型结构设计到提示工程策略应用的模型尚未在材料科学领域发现, 大视觉模型在材料科学领域中的应用相较于语言大模型处于起步阶段, 与通用领域的发展差异相似. 分析原因主要有: (1)科技论文、专利、技术报告等训练资料更易获得并利用; (2)语言大模型的研究社区建立更加完善; (3)材料科学领域的公开图像数据集匮乏; (4)材料科学领域有关视觉任务的性能测试基准尚未健全. 文本作为最直接的信息传递载体, 获取成本较低, 有成熟的文本提取、处理、解析与进一步分析的工具基础, 而且文本低维、离散的数据表示相较于高维、连续的图像数据对于信息语义表达能力更强, 且局部与全局的建模难度较低, 在相同计算资源条件下, 对于文本的内在规律的挖掘难度更低. 然而, 随着技术的不断迭代, 强泛化正在成为基于人工智能的材料显微图像分析领域的研究重点与难点. SAM 视觉大模型的出现, 为该领域开辟了新的研究方向, 研究者可借鉴 SAM 的思想, 从专用视觉大模型构建、领域数据集微调、针对领域先验的提示工程策略构建与零样本知识迁移等路径解决相应的问题.

为验证视觉大模型在材料显微图像中应用的可行性, 本文使用 SAM 对多种材料显微图像的成像特点测试并修改了基于点的提示词, 经过 SAM 的特征提取器和解码器, 得到图像中所有的封闭区域, 并经过连通分析与区域合并得到最终的微结构二值图. 图 3 为改进 SAM 在不同显微图像数据集的零样本分割表现, 第一行图像为使用扫描或光学电子显微镜采集到的不同种类的材料微观结构图; 第二行为 SAM 输出的不同分割区域的掩码结果, 每个掩码区域被赋予了一种颜色以示区分; 第三行为经过连通域处理后得到的二值掩码图(仅将感兴趣的区域用白色凸显出来). 可以发现, 在四种材料显微图像数据上, 经过定制的 SAM 能够将在通用领域数据集获得的对目标边界的知识应用到材料显微组织、晶粒等图像上, 实现零标注条件下的区域分割和关键结构提取, 这对材料微结构的快速定量表征从而加速材料构效关系的挖掘具有关键意义.

3.3　材料科学领域通用大模型的应用

在材料科学领域中, 想要实现通用大模型势必要尽可能多的使该大模型能够整合、理解关于不同材料的不同种表示[83]. 虽然某种材料遵循如薛定谔方程等控制方程, 机器学习模型也应该在不同的材料领域之间有效地共享方程背后的共同原则, 但由于很难求解到正确的数值解, 研究人员往往将材料以多种方式表示为可解释的形式, 如分子图结构、显微图像、性能数值、光谱、原子三维构象等. 材料科学通用大模型不只需要文本、图像两种模态的数据融合, 更需要考虑图、数值、光谱信号等多种数据模态——随着模态数量的上升, 模型构建的难度也随之急剧增加. 因此, 材料科学领域通用大模型(或多模态大模型)的基本问题是将多形式的材料表示进行合理的融合与对齐, 实现一套大模型参数能够有效地将材料领域知识迁移到各种下游任务中. 在材料特性预测中, 多模态通用大模型可以接受文本描述和图像数据, 从中提取关键特性并预测材料性能. 例如, 研

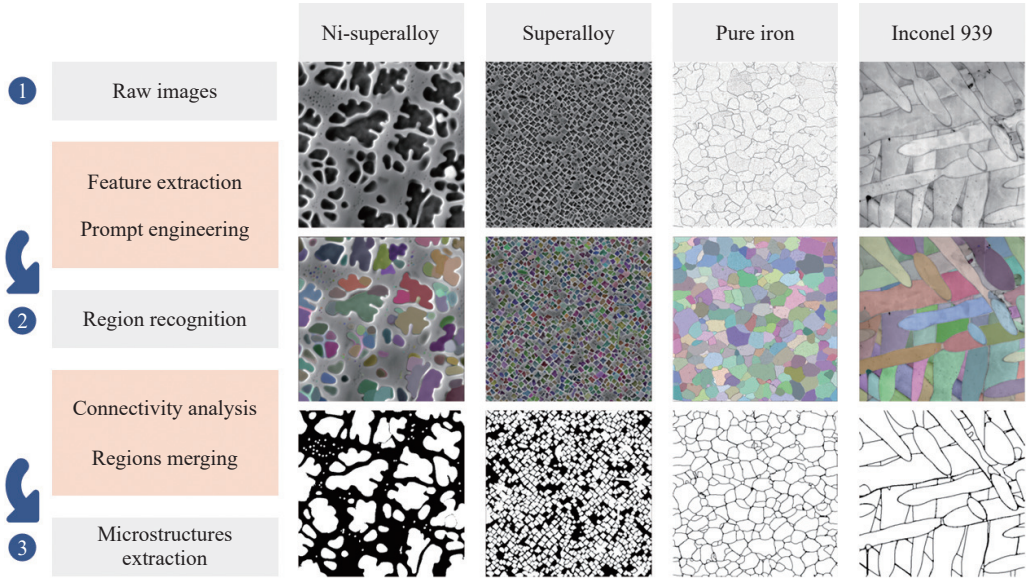

图 3 改进 SAM 在四种材料显微图像数据集上的零样本分割表现

Fig.3 Improving zero-shot segmentation performance of SAM on four different material microscopic image datasets

究人员提供材料的文字描述以及相关的图像，模型可以根据这些信息预测材料的电导率、热导率等性能指标. 这种多模态分析可以更全面地捕捉材料的性质，有助于提高预测的准确性. 另一方面，多模态通用大模型可以用于材料结构的快速识别和优化. 通过输入文本描述和图像，模型可以生成符合指定性能要求的材料结构. 这种方法有助于加速新材料的设计和发现过程，尤其在需要考虑多种性能因素和约束条件时具有显著的优势. 此外，多模态通用大模型还可以用于材料科学中的知识图谱构建. 通过分析文本和图像数据，模型可以自动构建材料的多模态知识图谱，将文字描述、图像信息和材料性质相互关联，为材料科学领域的知识管理提供新的方式. 其中，大规模多模态数据集构建的挑战包括：(1)材料科学通用大模型需要大规模可用无缺失的多模态数据，需要复杂异构的材料数据集成平台予以支持；(2)材料科学通用大模型构建多模态数据时可能会遗漏蕴含在图像中的关键信息，如曲线图、表格数据等.

针对第一个挑战，需从材料科学数据库基础设施平台的建设考虑. 由于材料基因工程(Materials genome engineering, MGE)相关研究[84]的数据要求和共享共治特征，现代数据基础设施成为其迫切需求. Liu 等[85]构建了材料基因工程数据库(Materials genome engineering databases, MGED). MGED 基于云计算服务提供了一套集异构数据集多源上传收集、标准化统一、个性化表示及材料数据检索、处理、分享的全流程集成平台. 值得说明的是，其设计的无模式存储方法能够让数据上传者根据所拥有的材料样本特性和数据结构规划相应的存储字段和表现形式，包含文本、数值、图像等种类. 而统一结构、符合标准的大批量多源异构数据的采集与处理是综合性多模态材料科学大模型构建的先决条件. 黄鹏儒等[86]通过对公开文献的文本挖掘和现有的公开数据库，收集含氢材料的物理化学性质数据，结合第一性原理构建性能数据集并整合为基于材料基因工程的储氢材料数据库. 这种构造增量数据并结合过去已有数据的数据整合方式特别需要注重数据标准是否统一、数据采集获取时的条件是否明确及数据是否可转换等问题.

针对第二个挑战，需要从基于图像识别的材料科学关键数据提取任务发力. 由于大量的重要数据被记录在以图、表等格式的信息载体中，全面、精准地提取这些数据是构建数据集、训练大模型的重要基础. Zhang 等[87]针对材料科学领域文献中的文字与表格提取与解析问题提出了相应的文献挖掘方法，不同于以往单纯的表格识别+后处理的解决方案，其同时融合了表格上下文中的文本语义信息，以期挖掘到深入、相关性强的关键信息，提取化学元素与占比大小等内容. 此外，在文献中曲线图往往体现着某种材料性能与温度、时间、作用参数等变量的关系，这部分关键数据的识别、提取与解析对数据集的补充至关重要，目前可用的相关工具有 Origin 软件的曲线识别插件、WebPlotDigitizer 等. 但是这些解决方案都是基于半自动交

互式操作, 操作人员需要设定坐标轴、起始点等才能获取全部信息, 无法实现大批量全自动曲线识别与数值提取等功能, 无法有效降低海量文献图、表、曲线信息的处理成本. 从大模型对大规模数据量的需求来看, 针对上述难点, 亟需设计并研发相应的非结构化图表数据的信息提取与知识挖掘的解决方案.

多模态通用大模型可以将自然语言处理和视觉处理的能力整合起来, 为材料科学研究提供多模态分析、性能预测、材料结构优化和知识图谱构建等丰富的工具和视角, 在材料科学中具有广泛的应用潜力. 随着多模态通用大模型技术的不断发展, 预计其在材料科学中的应用将继续拓展, 为材料研究带来更多的创新和可能性.

4　总结与展望

以大模型概念为起点, 首先概述了大模型的发展背景、相关技术、特点以及分类, 简要介绍了语言大模型、视觉大模型和多模态大模型三类大模型的基本特征和典型工作, 指出大模型的建立本质上是大规模深度无监督表征学习的结果, 其自监督式的训练方法和自注意力机制从自然语言处理任务开始, 深刻影响着视觉、多模态乃至通用大模型的发展和进步. 其次, 介绍了通用和垂直两类大模型的应用然后, 针对大模型在材料科学中的应用现状与挑战, 从语言大模型、视觉大模型和通用大模型三个角度论述了相关工作和各自的技术特点, 并在视觉大模型 SAM 的基础上, 实践了基于关键点的提示工程策略, 给出了初步的实验结果. 最后, 总结了大模型在材料科学领域的挑战并对未来工作进行了展望: (1)材料科学语言大模型发展较早并已经出现如 MatBERT 等相对成熟可用的模型和相关应用, 未来应在更多材料科学研究场景中加以验证与实践; (2)材料科学视觉大模型尚未出现针对性的领域实践, 但已有针对多种材料、多种成像方式显微图像的预训练模型, 未来应结合基于生成方法的显微图像数据扩增进行大规模数据集的构建和设计多种符合材料显微图像专有特征的提示工程策略, 支持领域视觉大模型的发展; (3)材料科学通用大模型发展难度较大, 其难点集中在复杂异构数据的海量收集、合理存储与正确使用, 未来应基于视觉识别发展文献关键信息抽取技术以及相应多模态数据的深度语义对齐, 落实材料科学通用大模型的实际应用.

参　考　文　献

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521(7553): 436

[2] Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models. *Mach Learn*, https://doi.org/10.48550/arXiv.2108.07258

[3] Li X R, Ban X J, Yuan Z L, et al. Review on deep learning models for time series forecasting in industry. *Chin J Eng*, 2022, 44(4): 757
(李潇睿, 班晓娟, 袁兆麟, 等. 工业场景下基于深度学习的时序预测方法及应用. 工程科学学报, 2022, 44(4): 757)

[4] Yao C, Zhao J H, Ma B Y, et al. Fast detection method for cervical cancer abnormal cells based on deep learning. *Chin J Eng*, 2021, 43(9): 1140
(姚超, 赵基淮, 马博渊, 等. 基于深度学习的宫颈癌异常细胞快速检测方法. 工程科学学报, 2021, 43(9): 1140)

[5] Taigman Y, Yang M, Ranzato M A, et al. DeepFace: closing the gap to human-level performance in face verification // 2014 *IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, 2014: 1701

[6] Bojarski M, Del Testa D, Dworakowski D, et al. End to end learning for self-driving cars. *Compu Vision Pattern Recogn*, https://arxiv.org/abs/1604.07316

[7] Ji Z W, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*, 2023, 55(12): 1

[8] Liu T. A look into large language models and its applications from the perspective of ChatGPT. *Chin J Lang Policy Plan*, 2023, 8(5): 14
(刘挺. 从 ChatGPT 谈大语言模型及其应用. 语言战略研究, 2023, 8(5): 14)

[9] Liu S L, Zeng Z Y, Ren T H, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *Compu Vision Pattern Recogn*, https://doi.org/10.48550/arXiv.2303.05499

[10] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation // *Proceedings of the* 38*th International Conference on Machine Learning*. Vienna, 2021: 8821

[11] Bi K F, Xie L X, Zhang H H, et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 2023, 619(7970): 533

[12] Zhang Q J. Survey on the development of AI large model. *Commun Technol*, 2023, 56(3): 255
(张乾君. AI 大模型发展综述. 通信技术, 2023, 56(3): 255)

[13] Jing L L, Tian Y L. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans Pattern Anal Mach Intell*, 2021, 43(11): 4037

[14] Wu T Y, He S Z, Liu J P, et al. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA J Autom Sin*, 2023, 10(5): 1122

[15] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of

transfer learning with a unified text-to-text transformer. *J Mach Learn Res*, 2020, 21(140): 1

[16] He K, Girshick R, Dollár P. Rethinking imagenet pre-training // *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul, 2019: 4918

[17] Neyshabur B, Sedghi H, Zhang C Y. What is being transferred in transfer learning? // 34*th Conference on Neural Information Processing Systems*. Vancouver, 2020

[18] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale. *Compu Vision Pattern Recogn*, https://doi.org/10.48550/arXiv.2010.11929

[19] Khan S, Naseer M, Hayat M, et al. Transformers in vision: A survey. *ACM Comput Surv*, 2022, 54(10): 1

[20] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, 2016: 770

[21] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, 2014: 580

[22] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, 2015: 3431

[23] Li Z Z, Hoiem D. Learning without forgetting. *IEEE Trans Pattern Anal Mach Intell*, 2018, 40(12): 2935

[24] Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers // *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal, 2021: 9650

[25] Jaiswal A, Babu A R, Zadeh M Z, et al. A survey on contrastive self-supervised learning. *Technologies*, 2020, 9(1): 2

[26] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision // *Proceedings of the* 38*th International Conference on Machine Learning*. Vienna, 2021: 8748

[27] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need // 31*st Conference on Neural Information Processing Systems*. Long Beach, 2017: 30

[28] Liu Z, Mao H Z, Wu C Y, et al. A ConvNet for the 2020s // *Proceedings of the* 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, 2022: 11966

[29] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *Comput Lang*, https://doi.org/10.48550/arXiv.1810.04805

[30] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training [J/OL]. *OpenAI* (2018–04–11) [2023–09–18]. https://openai.com/research/language-unsupervised

[31] Yang Z L, Dai Z H, Yang Y M, et al. XLnet: Generalized autoregressive pretraining for language understanding // 33*rd Conference on Neural Information Processing Systems*. 2019: 32

[32] Han K, Wang Y H, Chen H T, et al. A survey on vision transformer. *IEEE Trans Pattern Anal Mach Intell*, 2023, 45(1): 87

[33] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows // *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Kuala Lumpur, 2021: 10012

[34] Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning robust visual features without supervision. *Comput Vision Pattern Recogn*, https://doi.org/10.48550/arXiv.2304.07193

[35] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners // 34*th Conference on Neural Information Processing Systems*. Vancouver, 2020: 1877

[36] Kirillov A, Mintun E, Ravi N, et al. Segment anything. *Comput Vision Pattern Recogn*, https://doi.org/10.48550/arXiv.2304.02643

[37] Wang J Q, Liu Z L, Zhao L, et al. Review of large vision models and visual prompt engineering. *Comput Vision Pattern Recogn*, https://doi.org/10.48550/arXiv.2307.00855

[38] OpenAI. GPT-4 technical report. *Comput Lang*, https://doi.org/10.48550/arXiv.2303.08774

[39] Shao Z W, Yu Z, Wang M, et al. Prompting large language models with answer heuristics for knowledge-based visual question answering // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, 2023: 14974

[40] Xu Y M, Hu L, Zhao J Y, et al. Technology application prospects and risk challenges of large language model. *J Comput Appl*, https://kns.cnki.net/kcms/detail/51.1307.TP.20230911.1048.006.html
(徐月梅, 胡玲, 赵佳艺, 等. 大语言模型的技术应用前景与风险挑战. 计算机应用, http://kns.cnki.net/kcms/detail/51.1307.TP.20230911.1048.006.html)

[41] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners [J/OL]. *OpenAI* (2019–02–14) [2023–09–18]. https://openai.com/research/better-language-models

[42] Sharan N, Aakanksha C. Pathways language model (PaLM): Scaling to 540 billion parameters for breakthrough performance. *Comput Lang*, https://arxiv.org/abs/2204.02311

[43] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents. *Comput Vision Pattern Recogn*, https://arxiv.org/abs/2204.06125

[44] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models. *Comput Lang*, https://doi.org/10.48550/arXiv.2302.13971

[45] Google AI. Google AI PaLM 2 [J/OL]. *Google AI* (2023-08-03) [2023-09-18]. https://ai.google/discover/palm2

[46] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models. *Comput Lang*, https://arxiv.org/abs/2307.09288

[47] Korbak T, Shi K J, Chen A, et al. Pretraining language models with human preferences // *Proceedings of the* 40*th International*

Conference on Machine Learning. Hawaii, 2023: 17506

[48] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database // 2009 *IEEE Conference on Computer Vision and Pattern Recognition*. Miami, 2009: 248

[49] Dehghani M, Djolonga J, Mustafa B, et al. Scaling vision transformers to 22 billion parameters // *Proceedings of the* 40*th International Conference on Machine Learning*. Hawaii, 2023: 7480

[50] Riquelme C, Puigcerver J, Mustafa B, et al. Scaling vision with sparse mixture of experts // 35*th Conference on Neural Information Processing Systems*, 2021: 8583

[51] Wang Z Y, Li Y L, Chen X, et al. Detecting everything in the open world: Towards universal object detection // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, 2023: 11433

[52] Lin J Y, Men R, Yang A, et al. M6: Multi-modality-to-multi-modality multitask mega-transformer for unified pretraining // *Proceedings of the* 27*th ACM SIGKDD Conference on Knowledge Discovery Data Mining*. Singapore, 2021: 3251

[53] Wang X, Chen G Y, Qian G W, et al. Large-scale multi-modal pre-trained models: A comprehensive survey. *Mach Intell Res*, 2023, 20(4): 447

[54] Blum L, Blum M. A theoretical computer science perspective on consciousness and artificial general intelligence. *Engineering*, 2023, 25: 12

[55] Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision // *Proceedings of the* 38*th International Conference on Machine Learning*. London, 2021: 4904

[56] Zhong Y W, Yang J W, Zhang P C, et al. Regionclip: Region-based language-image pretraining // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, 2022: 16793

[57] Zhang R R, Hu X F, Li B H, et al. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, 2023: 15211

[58] Li Y X, Li Z H, Zhang K, et al. ChatDoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *Comput Lang*, https://arxiv.org/abs/2303.14070

[59] Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. *Comput Lang*, https://doi.org/10.48550/arXiv.1903.10676

[60] He Z, Zeng R X, Qin W, et al. Social impact and governance of ChatGPT and other new generation artificial intelligence technologies. *E-Government*, 2023(4): 2
(何哲, 曾润喜, 秦维, 等. ChatGPT 等新一代人工智能技术的社会影响及其治理. 电子政务, 2023(4): 2)

[61] Chan A. GPT-3 and InstructGPT: Technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. *AI Ethics*, 2023, 3(1): 53

[62] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning. *Mach Learn*, https://arxiv.org/abs/1312.5602

[63] Khan S G, Herrmann G, Lewis F L, et al. Reinforcement learning and optimal adaptive control: An overview and implementation examples. *Annu Rev Contr*, 2012, 36(1): 42

[64] Yang S, Nachum O, Du Y L, et al. Foundation models for decision making: Problems, methods, and opportunities . *Artl Intell*, https://doi.org/10.48550/arXiv.2303.04129

[65] Huang Y H, Yang X, Liu L, et al. Segment anything model for medical images? *Image Video Process*, https://arxiv.org/abs/2304.14660

[66] Ke L, Ye M Q, Danelljan M, et al. Segment anything in high quality. *Comput Vision Pattern Recogn*, https://arxiv.org/abs/2306.01567

[67] Chen K, Liu C, Chen H, et al. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *Comput Vision Pattern Recogn*, https://arxiv.org/abs/2306.16269

[68] Wang L, Ye X F, Zhu L Q, et al. When SAM meets sonar images [J/OL]. *Comput Vision Pattern Recogn*, https://arxiv.org/abs/2306.14109

[69] Huang Z, Bianchi F, Yuksekgonul M, et al. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat Med*, 2023, 29(9): 2307

[70] Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI. *Comput Lang*, https://doi.org/10.48550/arXiv.2307.14334

[71] Takeda S, Kishimoto A, Hamada L, et al. Foundation model for material science // *Proceedings of the AAAI Conference on Artificial Intelligence*. Washington, 2023: 15376

[72] Trewartha A, Walker N, Huo H Y, et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns*, 2022, 3(4): 100488

[73] Gupta T, Zaki M, Anoop Krishnan N M, et al. MatSciBERT: A materials domain language model for text mining and information extraction. *NPJ Comput Mater*, 2022, 8: 102

[74] Zhou G, Gao Z, Ding Q, et al. Uni-Mol: a universal 3D molecular representation learning framework // *The Eleventh International Conference on Learning Representations*. Kigali, 2023: 1

[75] Song Y, Miret S, Liu B. MatSci-NLP: Evaluating scientific language models on materials science language tasks using text-to-schema modeling // *Proceedings of the* 61*st Annual Meeting of the Association for Computational Linguistics*. Toronto, 2023: 3621

[76] Fu J, Li H, Song X, et al. Multi-scale defects in powder-based additively manufactured metals and alloys. *J Mater Sci Technol*, 2022, 122: 165

[77] Yang Z Z, Buehler M J. Linking atomic structural defects to mesoscale properties in crystalline solids using graph neural networks. *NPJ Comput Mater*, 2022, 8: 198

[78] Ju Y W, Li S A, Yuan X F, et al. A macro-nano-atomic–scale high-throughput approach for material research. *Sci Adv*, 2021,

7(49): eabj8804

[79] Durmaz A R, Müller M, Lei B, et al. A deep learning approach for complex microstructure inference. *Nat Commun*, 2021, 12: 6272

[80] Ma B Y, Jiang S F, Yin D, et al. Image segmentation metric and its application in the analysis of microscopic image. *Chin J Eng*, 2021, 43(1): 137
(马博渊, 姜淑芳, 尹豆, 等. 图像分割评估方法在显微图像分析中的应用. 工程科学学报, 2021, 43(1): 137)

[81] Stuckner J, Harder B, Smith T M. Microstructure segmentation with deep learning encoders pre-trained on a large microscopy dataset. *NPJ Comput Mater*, 2022, 8: 200

[82] Alrfou K, Zhao T, Kordijazi A. Transfer learning for microstructure segmentation with CS-UNet: A hybrid algorithm with transformer and CNN encoders. *Comput Vision Pattern Recogn*, https://doi.org/10.48550/arXiv.2308.13917

[83] Ren D, Wang C C, Wei X L, et al. Building a quantitative composition-microstructure-property relationship of dual-phase steels via multimodal data mining. *Acta Mater*, 2023, 252: 118954

[84] Li Z X, Zhang N, Xiong B, et al. Materials science database in material research and development: Recent applications and prospects. *Front Data Comput*, 2020, 2(2): 78
(李姿昕, 张能, 熊斌, 等. 材料科学数据库在材料研发中的应用与展望. 数据与计算发展前沿, 2020, 2(2): 78)

[85] Liu S L, Su Y J, Yin H Q, et al. An infrastructure with user-centered presentation data model for integrated management of materials data and services. *NPJ Comput Mater*, 2021, 7: 88

[86] Huang P R, Cai D, Lin H Z, et al. Materials genome engineering-based hydrogen storage materials database and its applications. *Sci Sin Chem*, 2022, 52(10): 1863
(黄鹏儒, 蔡丹, 林怀周, 等. 基于材料基因工程的储氢材料数据库构建及其应用. 中国科学:化学, 2022, 52(10): 1863)

[87] Zhang R, Zhang J W, Chen Q C, et al. A literature-mining method of integrating text and table extraction for materials science publications. *Comput Mater Sci*, 2023, 230: 112441