微软亚洲研究院院刊

Matrix

NO.62
2022年 7 - 9 月

通用多模态基础模型BEiT-3：引领文本、图像、多模态预训练迈向“大一统”

无限视觉生成模型NUWA-Infinity让视觉艺术创作自由延伸

微软亚洲研究院与CCF携手十六年，打造国际化学术共同体

本期封面图由NUWA基于《清明上河图》学习后生成

01 焦点

02 前沿求索

科研第一线

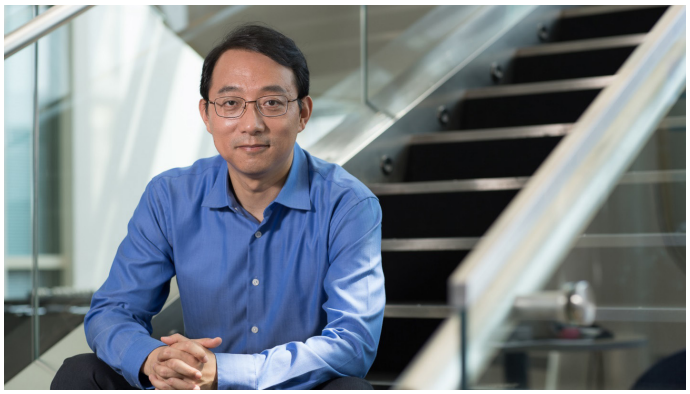
03 文化故事

04 观点

05 媒体报道

微软宣布周礼栋博士升任微软公司全球资深副总裁

2022 年 9 月，微软公司宣布，微软亚洲研究院院长周礼栋博士升任为微软公司全球资深副总裁。



周礼栋博士现任微软亚洲研究院院长，全面负责微软亚洲研究院在中国及亚太地区的研究工作以及与学术界和产业界的合作。

作为系统研究领域首屈一指的专家，多年来他一直专注于推动可靠、可信及可扩展的分布式系统的理论研究和实践探索，并致力于促进中国以及整个亚洲地区的计算机系统研究与合作。作为微软在设计和开发大规模分布式系统方面的重要技术带头人，周礼栋博士主持设计和开发的系统支持着微软从搜索引擎、大数据基础设施、云可靠性和可用性到 AI 基础设施的主要服务。

周礼栋博士是电气电子工程师学会会士（IEEE Fellow）和国际计算机学会会士（ACM Fellow）。他还是 ACM 计算机系统会刊（ACM Transactions on Computer Systems）、ACM 计算机存储会刊（ACM Transactions on Storage）、IEEE 计算机会刊（IEEE Transactions on Computers）的编委会成员， 并担任 ACM 软件系统奖项评选委员会（ACM Software System Award Committee）主席，以及 ACM 操作系统原理大会（SOSP）指导委员会（Steering Committee）成员。

周礼栋博士 2002 年加入微软公司，曾任职微软硅谷研究院研究员、微软雷德蒙研究院系统研究组首席研究员、微软亚洲研究院常务副院长，并于 2021 年升任微软亚洲研究院院长。周礼栋博士毕业于复旦大学，并获得了计算机科学学士学位，之后在康奈尔大学深造，先后获得计算机科学硕士及博士学位。

微软亚洲研究院与 CCF 携手十六年，打造国际化学术共同体

从 1992 年到 2022 年，微软扎根中国 30 年，与中国的信息产业共同发展壮大。而伴随着微软在中国的不断成长，微软亚洲研究院也已发展成为具有世界级影响力的计算机基础和应用研究机构。能够取得这样的成绩，除了微软亚洲研究院一直对科研秉持"长期主义"、持续创新突破外，也离不开与全球顶级高校、科研机构及企业的合作。其中，微软亚洲研究院与中国计算机学会（CCF）的合作已有十六载，双方不断推动领域内的交流合作，营造新型健康的学术生态体系，一起见证了中国科技行业的繁荣。

2022 年是中国计算机学会（CCF）创建 60 周年。8 月 6 日，CCF 举办了 60 周年庆典活动。微软亚洲研究院常务副院长张冬梅代表微软亚洲研究院在庆典活动的重头环节，接受了 "CCF 创建 60 周年杰出贡献奖" 奖杯和证书。这一奖项旨在表彰过去 60 年在 CCF 的创建与发展过程中为 CCF 各级各类工作机构和重要项目及关键事项做出杰出贡献的个人和单位。微软亚洲研究院能够获此殊荣，不仅代表了 CCF 对微软亚洲研究院过去十六年良好合作关系与大力支持的认可，也更加坚定了微软亚洲研究院与 CCF 携手共创计算机事业美好未来的信念与决心。

中国计算机学会理事长梅宏表示，"近 20 年来，CCF 高速发展，在国际化发展战略的指导下，与多家国际知名学术机构建立战略合作伙伴关系，拓宽学会发展空间。其中，非常感谢微软亚洲研究院过去十六年的大力支持，与 CCF 一起探索创新之路，为科技发展贡献力量。"

微软亚洲研究院院长周礼栋表示，"非常荣幸微软亚洲研究院与中国计算机学会基于一致的理念，共同予力计算机领域青年学者的成长与发展。作为一家有着国际化背景的企业研究院，微软亚洲研究院非常期待能够一如既往地与中国计算机学会持续合作，共铸更多荣光！"

回首微软亚洲研究院与中国计算机学会相守相伴的十六年，“互学互鉴，多元创新”是双方携手共进的内在要义。

微软亚洲研究院常务副院长张冬梅（右二）代表微软亚洲研究院在 CCF 创建 60 周年庆典上领奖

微软亚洲研究院与 CCF 相识相知十六年

2006 年，中国计算机学会正处于发展壮大时期，需要学术界与产业界的力量和支持。微软亚洲研究院是国内计算机科学学术领域极具影响力的存在，因此，时任 CCF 秘书长杜子德来到研究院寻求合作机会。然而，昙花一现式的简单合作并不是微软亚洲研究院所期望的，研究院希望可以通过与 CCF 更长远地携手同行，做更多有意义的工作。

双方在思考合作定位之后，共同设立了第一个 CCF 优秀博士学位论文奖，即“中国计算机学会优秀博士学位论文奖”（简称 CCF 优博奖）。彼时，虽然国内有“全国优秀博士学位论文评选”，但每年计算机领域的获奖博士和论文却屈指可数。“CCF 优博奖”则专门表彰博士期间取得优秀成绩的计算机领域学生，以此鼓励计算机科研领域更多新鲜血液进行创新探索，为中国计算机事业的持续发展建立人才储备。

在第一次成功合作的基础之上，微软亚洲研究院也在思考如何与中国计算机学会打造可持续的合作关系。为了不断发现和培养新一代青年学者，微软亚洲研究院设立了“铸星计划”，该计划专门针对“博士毕业五年以内的学者”设立。所以，在支持“CCF 优博奖”的同时，微软亚洲研究院开始与“CCF 优博奖”的获得者建立联系，通过“铸星计划”、为他们提供持续的支持。截至 2022 年，已经有 17 位“CCF 优博奖”获得者通过“铸星计划”与微软亚洲研究院的研究员结成“星搭档”，参与到微软的创新研究项目中。此外，微软亚洲研究院还通过科研项目上的广泛合作，与青年学者们共同探索前沿科技的更多可能。

2010 年度“CCF 优博奖”获得者、2013 年度微软亚洲研究院“铸星计划”学者、清华大学副教授翟季冬表示，“获得‘CCF 优博’不仅是对我博士期间工作的肯定与认可，也为我在高校从事科研工作增强了信心。而参与微软亚洲研究院的‘铸星计划’，则让我能够与业界优秀的学者交流合作，一起探索解决真实挑战的创新性技术，在国际化的平台上拓展自己的视野。对于我的学术成长，CCF 和微软亚洲研究院都助益良多。”

2009 年，中国计算机学会开始设立“CCF 优博论坛”，以加强学术交流，帮助国内优秀青年学者成长。微软亚洲研究院自论坛设立之初便积极参与，13 年来持续为“CCF 优博论坛”提供多方面支持。一年一次的优博论坛，作为优博们见面聚会、加强学术交流的重要活动，微软亚洲研究院与 CCF 都会投入百分百的精力，进行论坛的设计、邀请合适的讲者、设计别出心裁的环节。例如，2021 年优博论坛通过“铸星夜话”学术交流座谈，邀请了微软亚洲研究院的研究员和青年学者们品茗交流，畅谈学术与生活。

“CCF 优博论坛”2021 年在成都召开

创新是推动双方持续合作的内在基因

“创新”是中国计算机学会的关键词之一，这一点与微软亚洲研究院可谓志同道合。微软亚洲研究院希望能在中国、亚洲乃至全球学术界不断进行创新尝试。与 CCF 同行的十六年中，微软亚洲研究院的很多创新举措已经融入到了 CCF，在助力 CCF 活动多元化发展的同时，也促进了 CCF 的创新变革。

2010 年，时任微软亚洲研究院副院长周明与微软亚洲研究院学术合作总监马歆牵头举办了第一期以自然语言为主题的中国计算机学会学科前沿讲习班（ADL）。从内容到讲师，微软亚洲研究院都进行了精心设计，活动时每天人员爆满。这次活动的成功举办，为此后的 ADL 在组织形式和环节设置方面提供了诸多可以借鉴的经验，也成为其他学会争相模仿的举办模式。

2017 年，微软亚洲研究院又和中国计算机学会计算机视觉专委会合作召开了首届 CVPR 论文分享会，这是后来很多顶会论文分享活动的最早探索，也是线上线下结合方式的首次尝试。自第一届举办起，CVPR 论文分享会的参会人数逐年增加，到 2019 年，线下参会人数已由 260 人增加到了 350 人。2020 年，因疫情原因，活动从线下转为线上，在线观众人数更是达到了 5.3 万。6 年间，CVPR 论文分享会共邀请了 123 位来自学术界和产业界的学者前来分享计算机视觉领域最新的前沿技术，有超过 220 位学生在 CVPR 论文分享会的 Poster 环节展示了其研究成果，线上线下共影响近 15 万人，大大促进了计算机视觉领域的交流和发展。

2017 年，马歆加入中国计算机学会女工委，并陆续将微软亚洲研究院的多个项目成功引入 CCF 和女工委。2020 年，马歆当选为 CCF 女工委主任，将“助力下一代女性计算机人才培养”设立为 CCF 女工委未来四年的发展愿景之一。自 2014 年成立以来，CCF 女工委创建了 CNCC IT 女性精英论坛、“计算之美”学术大会、YOCSEF Lady “一方” 沙龙、“计算之美” Ada Workshop 四个品牌活动，通过走进高校、与其他组织和企业合办活动等多样形式，为广大科技女性搭建起交流和展示的平台，探寻女性榜样的力量。

2021“计算之美” Ada Workshop

中国计算机学会女工委还联合中国图象图形学学会女工委和中国人工智能学会女工委，组织了一系列女性学生的成长活动。2021 年 5 月，该活动在线上成功举办第一期，每个女工委都邀请了一名女性演讲人分享其成长历程，并在线上回答参与者的问题。下一步，在中国科协的指导与支持下，CCF 女工委将以科技创新为主题，面向世界科技前沿和国家重大需求组织沙龙，并形成促进女性科技人才发挥更大作用的建议等报告。

马歆强调，“作为中国计算机学会专门服务计算机领域女性工作者的组织，CCF 女工委具有独特的凝聚性、专业性、传承性，不断践行着‘以推动女性工作者在计算机领域内的发展为己任，致力于提升计算机行业的多元与包容性’的使命。越来越多的多元化创新活动将促使着 CCF 持续向前发展。”

合作共赢，不断推动中国计算机事业发展

过去的十六年中，微软亚洲研究院一直鼓励研究员们积极参加中国计算机学会的多项学术活动，沈向洋、洪小文、郭百宁等来自微软的重量级专家们多次成为中国计算机大会的特邀报告嘉宾。微软亚洲研究院首席研究员谢幸加入了 CCF 普适计算专业委员会，现已成为普适计算专业委员会副主任、CCF 会士、CCFADL 工作组成员。此外，还有微软（亚洲）互联网工程院副院长姜大昕、微软亚洲研究院常务副院长张冬梅和微软亚洲研究院副院长刘铁岩等二十余人也都参与到了 CCF 各个工委、专委、活动、大会中。

2021 年 10 月，中国计算机学会成立了计算艺术分会，聚焦于机器学习等人工智能技术对音乐、美术、设计、影视、戏曲等多种艺术学科的融合发展，微软亚洲研究院主管研究员谭旭也在其首届执行委员名单中。2021 年 12 月，CCF 又成立了开源发展委员会，重点打造开源、开放、中立的产学研协同开源创新服务平台，微软亚洲研究院作为开源实践的践行者也将为其开源生态建设提供助力。未来，微软亚洲研究院还将在更多领域探索与 CCF 合作的新机会，深度参与 CCF 的其他活动，包括成为 CCF 首批赞助优博论文集的企业、参与 UR Club 的筹办等等。

走过 60 年风雨的中国计算机学会，如今正在梅宏理事长的带领下探索实现“企业化”、“行业化”和“国际化”的目标，成为具有广泛影响力的计算机领域学术共同体。国际化，任重而道远。微软作为跨国科技企业将最大程度地发挥国际化优势，助力 CCF 成为具有国际影响力的学术社团。与此同时，微软亚洲研究院也将继续与 CCF 携手同行、紧密合作，为推进计算机科学事业的发展持续贡献力量！

通用多模态基础模型 BEiT-3：引领文本、图像、多模态预训练迈向“大一统”

近年来，基础模型（foundation models，也被称为预训练模型）的研究从技术层面逐渐趋向于大一统（the big convergence），不同人工智能领域（例如自然语言处理、计算机视觉、语音处理、多模态等）的基础模型从技术上都依赖三个方面：一是 Transformer 成为不同领域和问题的通用神经网络架构和建模方式，二是生成式预训练（generative pre-training）成为最重要的自监督学习方法和训练目标，三是数据和模型参数的规模化（scaling up）进一步释放基础模型的潜力。

技术和模型的统一将会使得 AI 模型逐步标准化、规模化，从而为大范围产业化提供基础和可能。通过云部署和云端协作，AI 将有可能真正成为像水和电一样的“新基建”赋能各行各业，并进一步催生颠覆性的应用场景和商业模式。

近期，微软亚洲研究院联合微软图灵团队推出了最新升级的 BEiT-3 预训练模型，在广泛的视觉及视觉 - 语言任务上，包括目标检测（COCO）、实例分割（COCO）、语义分割（ADE20K）、图像分类 (ImageNet)、视觉推理 (NLVR2)、视觉问答 (VQAv2)、图片描述生成（COCO）和跨模态检索（Flickr30K，COCO）等，实现了 SOTA 的迁移性能。BEiT-3 创新的设计和出色的表现为多模态研究打开了新思路，也预示着 AI 大一统渐露曙光。

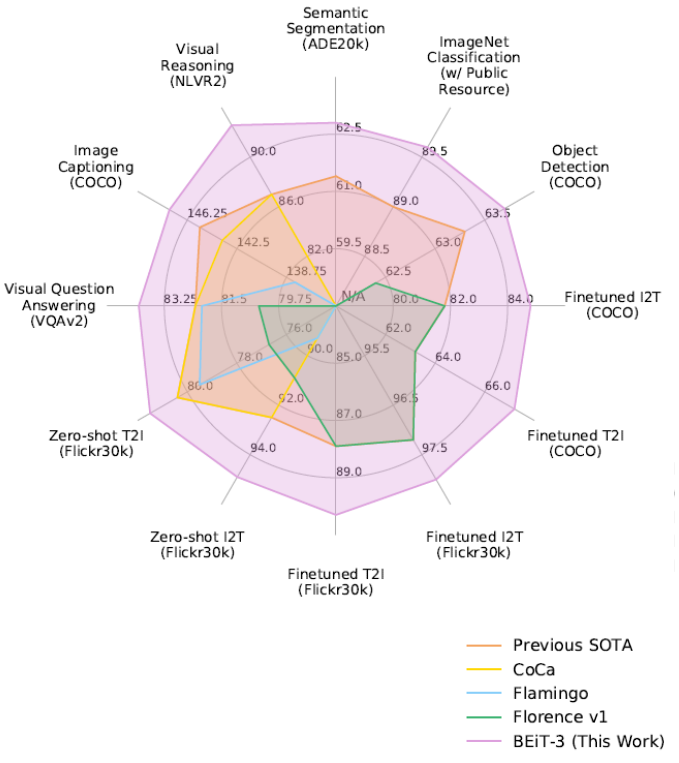


图 1：截至 2022 年 8 月，BEiT-3 在广泛的视觉及视觉 - 语言任务上都实现了 SOTA 的迁移性能

事实上，在早期对于 AI 和深度学习算法的探索中，科研人员都是专注于研究单模态模型，并利用单一模态数据来训练模型。例如，基于文本数据训练自然语言处理 (NLP) 模型，基于图像数据训练计算机视觉 (CV) 模型，使用音频数据训练语音模型等等。然而，在现实世界中，文本、图像、语音、视频等形式很多情况下都不是独立存在的，而是以更复杂的方式融合呈现，因此在人工智能的探索中，跨模态、多模态也成了近几年业界研究的重点。

大规模预训练正在趋向“大一统”

“近年来，语言、视觉和多模态等领域的预训练已经开始呈现出大一统（big convergence）趋势。通过对大量数据的大规模预训练，我们可以更轻松地将模型迁移到多种下游任务上。这种预训练一个通用基础模型来处理多种下游任务的模式已经吸引了越来越多科研人员的关注。” 微软亚洲研究院自然语言计算组主管研究员董力表示。微软亚洲研究院看到，大一统的趋势已经在三个方面逐渐显现，它们分别是骨干网络 (backbone)、预训练任务和规模提升。

首先，骨干网络逐渐统一。模型架构的统一，为预训练的大一统提供了基础。在这个思想指引下，微软亚洲研究院提出了一个统一的骨干网络 Multiway Transformer，可以同时编码多种模态。此外，通过模块化的设计，统一架构可以用于不同的视觉及视觉 - 语言下游任务。受到 UniLM（统一预训练语言模型）的启发，理解和生成任务也可以进行统一建模。

其次，基于掩码数据建模（masked data modeling）的预训练已成功应用于多种模态，如文本和图像。微软亚洲研究院的研究员们将图像看作一种语言，实现了以相同的方式处理文本和图像两种模态任务的目的。自此，图像 - 文本对可以被用作 “平行句

子”来学习模态之间的对齐。通过数据的归一化处理，还可以利用生成式预训练来统一地进行大规模表示学习。BEiT-3 在视觉、视觉-语言任务上达到 SOTA 性能也证明了生成式预训练的优越性。

第三，扩大模型规模和数据大小可提高基础模型的泛化能力，从而提升模型的下游迁移能力。遵循这一理念，科研人员逐渐将模型规模扩大到数十亿个参数，例如在 NLP 领域，Megatron-Turing NLG 模型有 5300 亿参数，这些大模型在语言理解、语言生成等任务上都取得了更好的成效；在 CV 领域，Swin Transformer v2.0 具有 30 亿参数，并在多个基准上刷新了纪录，证明了视觉大模型在广泛视觉任务中的优势。再加之，微软亚洲研究院提出了将图像视为一种语言的方式，可直接复用已有的大规模语言模型的预训练方法，从而更有利于视觉基础模型的扩大。

BEiT 为视觉基础大模型开创新方向

在 CV 领域的模型学习中，通常使用的是有监督预训练，利用有标注的数据。但随着视觉模型的不断扩大，标注数据难以满足模型需求，当模型达到一定规模时，即使模型再扩大，也无法得到更好的结果，这就是所谓的数据饥饿 (data hungry) 。因此，科研人员开始使用无标注数据进行自监督学习，以此预训练大模型参数。以往在 CV 领域，无标注数据的自监督学习常采用对比学习。但对比学习存在一个问题，就是对图像干扰操作过于依赖。当噪声太简单时，模型学习不到有用的知识；而对图像改变过大，甚至面目全非时，模型无法进行有效学习。所以对比学习很难把握这之间的平衡，且需要大批量训练，对显存和工程实现要求很高。

对此，微软亚洲研究院自然语言计算组研究员们提出了掩码图像建模 (Masked Image Modeling, MIM) 预训练任务，推出了 BEiT 模型。与文本不同，图像是连续信号，如何实现掩码训练呢？

为了解决这一问题，研究员们将图片转化成了两种表示视图。一是，通过编码学习 Tokenizer，将图像变成离散的视觉符号 (visual token) ，类似文本；二是，将图像切成多个小“像素块”(patch)，每个像素块相当于一个字符。这样，在用 BEiT 预训练时，模型可以随机遮盖图像的部分像素块，并将其替换为特殊的掩码符号 [M]，然后在骨干网络 ViT 中不断学习、预测实际图片的样子。在 BEiT 预训练后，通过在预训练编码上添加任务层，就可以直接微调下游任务的模型参数。在图像分类和语义分割方面的实验结果表明，与以前的预训练方法相比，BEiT 模型获得了更出色的结果。同时，BEiT 对超大模型（如 1B 或 10B）也更有帮助，特别是当标记数据不足以对大模型进行有监督预训练时。

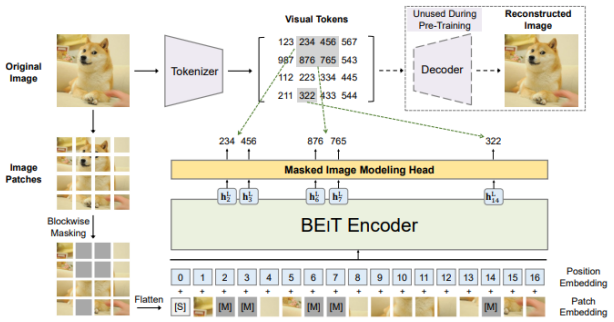


图 2：BEiT 预训练示意图

BEiT 相关论文被 ICLR 2022 大会接收为 Oral Presentation（口头报告论文，54 out of 3391）。ICLR 大会评审委员会认为，BEiT 为视觉大模型预训练的研究开创了一个全新的方向，首次将掩码预训练应用在了 CV 领域非常具有创新性。（了解更多详情，请查看 BEiT 论文原文：https://openreview.net/forum?id=p-BhZSz59o4）

ICLR 2022 Conference Program Chairs
21 Jan 2022 ICLR 2022 Conference Paper2678 Decision Readers: Everyone
Decision: Accept (Oral)
Comment: Inspired by BERT and the corresponding masked language modeling objective, this paper proposes masked image modeling as a pre-training technique for vision transformer. More precisely, the image is tokenized using a pre-trained tokenizer, and the goal is to predict the token indices corresponding to masked patches of the image. As noted by the reviewers, the proposed method is simple, works very well in practice and the paper is well written. Since this work potentially opens a whole new research direction, my recommendation is to accept with oral presentation.

图 3：BEiT 论文在 ICLR 2022 的评审意见

在 BEiT 的基础上，微软亚洲研究院的研究员们在 BEiT-2 中进一步丰富了自监督学习的语义信息（了解更多信息，请查看 BEiT-2 论文原文：https://arxiv.org/abs/2208.06366）。近日，研究员们又将其升级到了 BEiT-3。BEiT-3 利用一个共享的 Multiway Transformer 结构，通过在单模态和多模态数据上进行掩码数据建模完成预训练，并可迁移到各种视觉、视觉 - 语言的下游任务中。

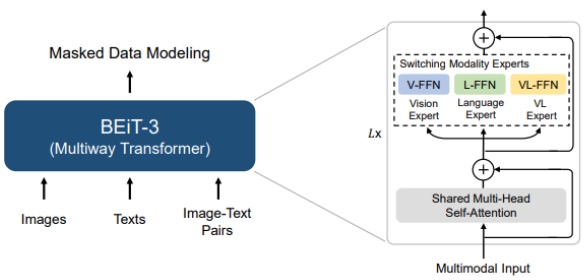


图 4：BEiT-3 预训练示意图

BEiT-3 的创新之处包含三个方面：

骨干网络：Multiway Transformer。研究员们将 Multiway Transformer 作为骨干网络以对不同模态进行编码。每个

Multiway Transformer 由一个共享的自注意力模块(self-attention)和多个模态专家 (modality experts) 组成，每个模态专家都是一个前馈神经网络（feed-forward network）。共享自注意力模块可以有效学习不同模态信息的对齐，并对不同模态信息深度融合编码使其更好地应用在多模态理解任务上。根据当前输入的模态类别，Multiway Transformer 会选择不同模态专家对其进行编码以学习更多模态特定的信息。每层 Multiway Transformer 包含一个视觉专家和一个语言专家，而前三层 Multiway Transformer 拥有为融合编码器设计的视觉 - 语言专家。针对不同模态统一的骨干网络使得 BEiT-3 能够广泛地支持各种下游任务。如图 4 所示，BEiT-3 可以用作各种视觉任务的骨干网络，包括图像分类、目标检测、实例分割和语义分割，还可以微调为双编码器用于图像文本检索，以及用于多模态理解和生成任务的融合编码器。

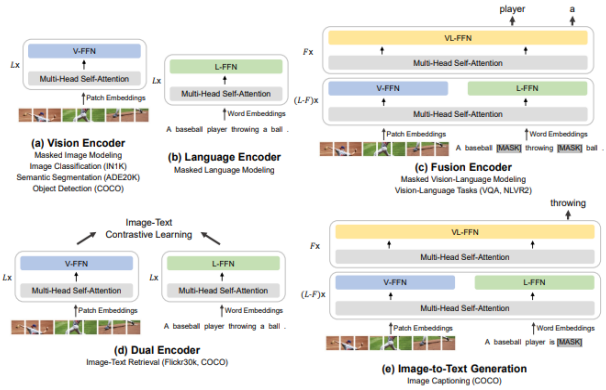


图 5：BEiT-3 可迁移到各种视觉、视觉 - 语言的下游任务

预训练任务：掩码数据建模 （masked data modeling）。研究员们在单模态（即图像与文本）和多模态数据（即图像 - 文本对）上通过统一的掩码 - 预测任务进行 BEiT-3 预训练。预训练期间，会随机掩盖一定百分比的文本字符或像素块，模型通过被训练恢复掩盖的文本字符或其视觉符号，来学习不同模态的表示及不同模态间的对齐。不同于之前的视觉 - 语言模型通常采用多个预训练任务，BEiT-3 仅使用一个统一的预训练任务，这对于更大模型的训练更加友好。由于使用生成式任务进行预训练，BEiT-3 相对于基于对比学习的模型也不需要大批量训练，从而缓解了 GPU 显存占用过大等问题。

扩大模型规模：BEiT-3 由 40 层 Multiway Transformer 组成，模型共包含 19 亿个参数。在预训练数据上，BEiT-3 基于多个单模态和多模态数据进行预训练，多模态数据从五个公开数据集中收集了大约 1,500 万图像和 2,100 万图像 - 文本对；单模态数据使用了 1,400 万图像和 160GB 文本语料。

“BEiT 系列研究有一个一以贯之的思想和原则，就是我们认为从通用技术层面看图像也可视为一种‘语言’（Imglish），从而可以以统一的方式对图像、文本和图像 - 文本对进行建模和学习。如果说 BEiT 引领和推进了生成式自监督预训练从 NLP 到 CV 的统一，那么，BEiT-3 实现了生成式多模态预训练的统一。”微软亚洲研究院自然语言计算组首席研究员韦福如说。

BEiT-3 使用 Multiway Transformer 有效建模不同的视觉、视觉 - 语言任务，并通过统一的 mask data modeling 作为预训练目标，这使得 BEiT-3 成为了通用基础模型的重要基石。“BEiT-3 既简单又有效，为多模态基础模型扩展打开了一个新方向。接下来，我们还将持续进行对 BEiT 的研究，以促进跨语言和跨模态的迁移，推动不同任务、语言和模态的大规模预训练甚至模型的大一统。”

多模态和通用基础模型研究还有更广阔的空间等待探索

人的感知和智能天生就是多模态的，不会局限在文本或图像等单一的模态上。因此，多模态是未来一个重要的研究和应用方向。另外，由于大规模预训练模型的进展，AI 的研究呈现出大学科趋势，不同领域的范式、技术和模型也在趋近大一统。跨学科、跨领域的合作将更加容易和普遍，不同领域的研究进展也更容易相互推进，从而进一步促进人工智能领域的快速发展。

“尤其是通用基础模型和通才模型等领域的研究，将让 AI 研究迎来更加激动人心的机遇和发展。而技术和模型的统一会使得 AI 模型逐步标准化、规模化，进而为大范围产业化提供基础和可能。通过云部署和云端协作，AI 将有可能真正成为像水和电一样的‘新基建’赋能各行各业，并进一步催生颠覆性的应用场景和商业模式。”韦福如表示。

相关链接：

论文链接：

BEiT: BERT Pre-Training of Image Transformers
https://openreview.net/forum?id=p-BhZSz59o4

BEiT-2: Masked Image Modeling with Vector-Quantized Visual Tokenizers
https://arxiv.org/abs/2208.06366

Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks（BEiT-3）
https://arxiv.org/abs/2208.10442

GitHub 链接：
https://github.com/microsoft/unilm

无限视觉生成模型 NUWA-Infinity 让视觉艺术创作自由延伸

此前，微软亚洲研究院提出了多模态模型 NUWA，它可以基于给定的文本、视觉或多模态输入生成图像或视频，并支持多种视觉艺术作品创建任务，包括文本到图像或视频的生成、图像补全、视频预测等。近日，微软亚洲研究院公开发表了新的研究成果：NUWA 的升级版——无限视觉生成模型 NUWA-Infinity，让视觉艺术创作趋于"无限流"，可生成任意大小的高分辨率图像或长时间视频。一起来感受一下 AI 的无限创作力吧！

或许你也曾有过这样的想法——那些 "世界名画" 画框外的景色是怎样的？让 NUWA-Infinity 带我们去 "一探究竟" ！发现梵高《星空》画框外更广阔的风景：

NUWA-Infinity 还能将静态的图像转化成超高清视频，为其带来 "活力"。 此外，NUWA-Infinity 也可以依据文本生成超高清图片，为艺术创作带来更加丰富的想象力。欢迎大家前往 NUWA-Infinity 演示页面，直观感受 NUWA-Infinity 的无限创作能力。

为什么微软亚洲研究院会开发 NUWA-Infinity，背后又用到了哪些新技术？

随着以消费为基础的注意力经济逐渐转为以生产为基础的创意经济，越来越多的人已经成为日常创作者，通过利用各种图片、视频编辑工具，实现艺术作品的创新或再创作。然而，高质量的视觉艺术创作从来都不是一件容易的事，往往需要专业的技能和设备，并花费大量的时间。与此同时，日常的视觉艺术创作对更高分辨率的图像或持续时间更长的视频也有着越来越高的需求。

为此，微软亚洲亚洲研究院 NUWA 团队研发出了无限视觉生成模型 NUWA-Infinity。与同样覆盖图像和视频创作的 NUWA 相比，NUWA-Infinity 在分辨率和可变大小视觉艺术作品生成方面具有更优的性能，并支持五个高分辨率视觉任务的生成，包括无条件图像生成高分辨率图、文本生成高分辨率图像、文本生成高分辨率视频、图像生成高分辨率动画和图像生成高分辨率图像。

在 NUWA-Infinity 模型中，研究员们提出了一种全局自回归嵌套局部自回归的生成机制，通过全局自回归建模视觉块之间的依赖关系和局部自回归建模视觉词之间的依赖关系，让 NUWA-Infinity 能够生成全局一致且局部细节丰富的高质量图像和视频，并提出任意方向控制器（Arbitrary Direction Controller, ADC）来决定合适的生成顺序并学习顺序感知的位置嵌入。相比其他多模态生成模型，NUWA-Infinity 可以从给定的文本、图像或视频生成与之相关的任意形状、任意大小的超高分别率图像，以适配不同设备、平台和场景；更重要的是，NUWA-Infinity 还支持长时间视频的生成，比如图像动画的制作。

此外，NUWA-Infinity 模型还引入了附近上下文池（Nearby Context Pool, NCP）来缓存已经生成的局部图像，作为正在生成的当前图像的上下文，这可以在不牺牲视觉块间依赖性的前提下，显著节省计算成本。NUWA-Infinity 极大地弥补了市场上现有技术仅支持生成大小有限的视觉内容以及视觉内容创作计算成本高昂的不足。

下一步，NUWA 团队将持续推动 NUWA 的演进，并希望研发出能从三个方面为专业和日常艺术创作者赋能的技术：

构思——通过自动快速和多样化的设计生成能力，降低构思门槛，在构思阶段为艺术创作者提供更多信息和灵感。美学——降低创意门槛，支持普通用户以适当的美学 / 设计质量来创作创意作品。效率——通过将 NUWA 的能力集合到一套智能工具中，来提高创作效率，降低创作工作量。

未来，由 AI 生成的高分辨率视觉内容将会更加符合图像设计、广告、动画、游戏等行业的视觉内容创作需求，为创作者提供源源不断的创造灵感。欢迎更多的科研人员、开发者与微软亚洲研究院共同探索 AI 视觉创作领域的广阔未来。

扫描二维码了解更多详细信息

USB：首个将视觉、语言和音频分类任务进行统一的半监督分类学习基准

当前，半监督学习的发展如火如荼。但是现有的半监督学习基准大多局限于计算机视觉分类任务，排除了对自然语言处理、音频处理等分类任务的一致和多样化评估。此外，大部分半监督论文由大型机构发表，学术界的实验室往往由于计算资源的限制而很难参与到推动该领域的发展中。为此，微软亚洲研究院的研究员们联合西湖大学、东京工业大学、卡内基梅隆大学、马克斯 - 普朗克研究所等机构的科研人员提出了 Unified SSL Benchmark（USB）：第一个将视觉、语言和音频分类任务进行统一的半监督分类学习基准。该论文不仅引入了更多样化的应用领域，还首次利用视觉预训练模型大大缩减了半监督算法的验证时间，使得半监督研究对研究者，特别是小研究团体更加友好。相关论文已被国际人工智能领域顶级学术大会 NeurIPS 2022 接收。

监督学习通过构建模型来拟合有标记数据，当使用监督学习 (supervised learning) 对大量高质量的标记数据（labeled data）进行训练时，神经网络模型会产生有竞争力的结果。例如，据 Paperswithcode 网站统计，在 ImageNet 这一百万量级的数据集上，传统的监督学习方法可以达到超过 88% 的准确率。然而，获取大量有标签的数据往往费时费力。

为了缓解对标注数据的依赖，半监督学习（semi-supervised learning/SSL）致力于在仅有少量的标注数据时利用大量无标签数据（unlabeled data）来提升模型的泛化性。半监督学习亦是机器学习的重要主题之一。深度学习之前，这一领域的研究者们提出了诸如半监督支持向量机、熵正则化、协同训练等经典算法。

深度半监督学习

随着深度学习的兴起，深度半监督学习算法也取得了长足的进步。同时，包括微软、谷歌、和 Meta 等在内的科技公司也认识到了半监督学习在实际场景中的巨大潜力。例如，谷歌利用噪声学生训练（noisy student training）这一半监督算法提高了其在搜索方面的性能 [1]。当前最具代表性的半监督算法通常对标注数据使用交叉熵损失进行训练，对无标注数据使用一致性正则技术 (consistency regularization) 鼓励对输入扰动进行不变预测。例如，谷歌在 NeurIPS 2020 提出的 FixMatch[2] 算法，利用增强锚定（augmentation anchoring）和固定阈值（fixed thresholding）技术来增强模型对不同强度增强数据的泛化性和减少噪声伪标签（noisy pseudo labels）的影响。在训练中，FixMatch 过滤了低于用户指定（user-provided / pre-defined）阈值的无标签数据。

微软亚洲研究院与东京工业大学等在 NeurIPS 2021 合作提出的 FlexMatch[3] 则考虑到了不同类之间的学习难度不同，因此提出了课程伪标签（curriculum pseudo labeling）技术，对于不同类应该采用不同的阈值。具体来说，对于容易学习的类别，模型应该设置高阈值以降低噪声伪标签的影响；对于难学习的类别，模型应该设置低阈值鼓励该类的拟合。每个类的学习难度评估取决于落入该类且高于固定值的未标记数据样本的数量。

同时，微软亚洲研究院的研究员们还合作提出了一个统一的基于 Pytorch 的半监督方法代码库 TorchSSL[4]，对该领域的深度方法、常用数据集和基准结果进行了统一的支持。

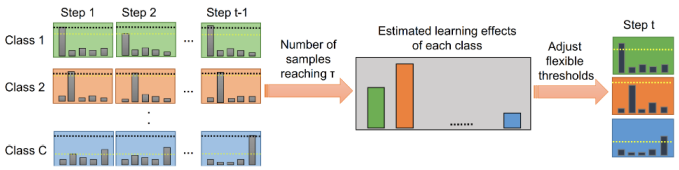


图 1：FlexMatch 算法流程

当前半监督学习代码库存在的问题与挑战

尽管半监督学习的发展如火如荼，但是，研究员们注意到目前大部分半监督方向的论文只关注计算机视觉 (CV) 分类任务，对于其他领域，例如自然语言处理 (NLP)、音频处理 (audio)，研究者无法得知这些在 CV 任务上有效的算法到了不同领域是否依然有效。另外，大部分半监督相关的论文都是由大型机构发表，学术界的实验室往往由于计算资源的限制而很难参与到推动该领域的发展中。总的来说，半监督学习基准目前存在以下两个问题：

（1）多样性不足。现有的半监督学习基准大多局限于 CV 分类任务（即 CIFAR-10/100，SVHN，STL-10 和 ImageNet 分类），排除了对 NLP、audio 等分类任务的一致和多样化评估，而在 NLP 和 audio 中缺乏足够的标记数据也是一个普遍问题。

（2）耗时且对学术界不友好。现有的半监督学习基准（如 TorchSSL）通常是耗时且不环保的，因为它往往需要从头开始训练深度神经网络模型。具体而言，使用 TorchSSL 评估 FixMatch[1]

大约需要 300 个 GPU 日。如此高的训练成本使得许多研究实验室（尤其是学术界的实验室或小研究团体）无法负担得起 SSL 的相关研究，从而阻碍了 SSL 的进展。

USB：任务多样化且对研究者更友好的新基准库

为了解决上述问题，微软亚洲研究院的研究员们联合西湖大学、东京工业大学、卡内基梅隆大学、马克斯 - 普朗克研究所等机构的科研人员提出了 UniPed SSL Benchmark（USB），这是第一个将视觉、语言和音频分类任务进行统一的半监督分类学习基准。相比于之前的半监督学习基准（如 TorchSSL）只关注少量视觉任务，该基准不仅引入了更多样化的应用领域，还首次利用视觉预训练模型（pretrained vision Transformer）大大缩减了半监督算法的验证时间（从 7000 GPU 时缩减至 900 GPU 时），从而使得半监督研究对研究者、特别是小研究团体更为友好。相关论文已被国际人工智能领域的顶级学术大会 NeurIPS 2022 接收。

USB 提供的解决方案：

（1）为增强任务多样性，USB 引入了 5 个 CV 数据集，5 个 NLP 数据集和 5 个 audio 数据集，并提供了一个多样化且具有挑战性的基准，对来自不同领域的多个任务进行一致的评估。表 1 提供了 USB 与 TorchSSL 的任务和训练时间等方面的详细对比。

（2）为了提高训练效率，研究员们将预训练的 vision Transformer 引入 SSL，而不是从头训练 ResNets。具体而言，研究员们发现在不影响性能的情况下使用预训练模型可以大大减少训练迭代次数（例如，将 CV 任务的训练迭代次数从 100 万步减少到 20 万步）。

（3）为了对研究人员更加友好，研究员们开源实现了 14 种 SSL 算法并开源了一个模块化代码库和相关的配置文件以供研究者轻松再现 USB 报告中的结果。为了快速上手，USB 还提供了详细的文档和教程。此外，USB 还提供了 pip 包以供使用者直接调用 SSL 算法。研究员们承诺未来会在 USB 中不断加入新的算法（例如不平衡半监督算法等）和更多更具挑战性的数据集。表 2 展示了 USB 中已支持的算法和模块。

半监督学习通过利用大量无标签数据来训练更精确、更鲁棒的模型，在未来有着重要的研究和应用价值。微软亚洲研究院的研究员们期待通过 USB 这一工作，能够予力学术界和工业界在半监督学习领域取得更大的进展。

Table 1: A summary of datasets and training cost used in (a) the existing popular protocol and (b) USB. USB largely reduces the training cost while providing a diverse, challenging, and comprehensive benchmark covering a wide range of datasets from various domains. Training cost is estimated by FixMatch [20] on a single NVIDIA V100 GPU from Microsoft Azure Machine Learning platform, except for ImageNet where 4 V100s are used. Experiments in (a) follow the settings in [21]. More results with different pre-trained backbones are available in Appendix E.

(a) TorchSSL [21]

Domain & Backbone	Dataset Name	Classification Task	GPU Hours	Total GPU Hours	Total GPU Hours w/o ImageNet
CV, ResNets	CIFAR-10	Natural Image	110 hour × 3 settings × 3 seeds	8031 GPU Hours (335 GPU Days)	6687 GPU Hours (279 GPU Days)
	CIFAR-100	Natural Image	300 hours × 3 settings × 3 seeds		
	SVNH	Digital	108 hours × 3 settings × 3 seeds		
	STL-10	Natural Image	225 hours × 3 settings × 3 seeds		
	ImageNet	Natural Image	336 hours × 4 GPUs		

(b) USB

Domain & Backbone	Dataset Name	Classification Task	GPU Hours	Total GPU Hours
CV, ViTs	CIFAR-100	Natural Image	7 hours × 2 settings × 3 seeds	897 GPU Hours (37 GPU Days)
	STL-10	Natural Image	13 hours × 2 settings × 3 seeds	
	EuroSAT	Satellite Image	10 hours × 2 settings × 3 seeds	
	TissueMNIST	Medical Image	5 hours × 2 settings × 3 seeds	
	Semi-Aves	Fine-grained, Long-tailed Natural Image	13 hours × 1 setting × 3 seeds	
NLP, Bert	IMDB	Movie Review Sentiment	7.5 hour × 2 settings × 3 seed	
	AG News	News Topic	3.5 hours × 2 settings × 3 seeds	
	Amazon Review	Product Review Sentiment	5 hours × 2 settings × 3 seeds	
	Yahoo! Answer	QA Topic	3.5 hours × 2 settings × 3 seeds	
	Yelp Review	Restaurant Review Sentiment	6 hours × 2 settings × 3 seeds	
Audio, Wave2Vec 2.0 and HuBert	GTZAN	Music Genre	12 hours × 2 settings × 3 seeds	
	UrtraSound8k	Urban Sound Event	10 hours × 2 settings × 3 seeds	
	FSDnoisy18k	Sound Event	32 hours × 1 setting × 3 seeds	
	Keyword Spotting	Keyword	10.5 hours × 2 settings × 3 seeds	
	ESC-50	Environmental Sound Event	34 hours × 2 settings × 3 seeds	

表 1：USB 与 TorchSSL 框架的任务和训练时间对比

Algorithm	PL	CR Loss	Thresholding	Dist. Align.	Self-supervised	Mixup	W-S Aug.
Π-Model		MSE					
Pseudo Labeling	✓	CE					
Mean Teacher		MSE					
VAT		CE					
MixMatch		MSE				✓	
ReMixMatch		CE		✓	Rotation	✓	✓
UDA		CE	✓				✓
FixMatch	✓	CE	✓				✓
Dash	✓	CE	✓				✓
CoMatch	✓	CE	✓	✓	Contrastive		✓
CRMatch	✓	CE	✓		Rotation		✓
FlexMatch	✓	CE	✓				✓
AdaMatch	✓	CE	✓	✓			✓
SimMatch	✓	CE	✓	✓	Contrastive		✓

表 2：USB 中已支持的算法和模块

参考文献

[1] https://ai.googleblog.com/2021/07/from-vision-to-language-semi-supervised.html

[2] Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems, 33:596–608, 2020.

[3] Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. Advances in Neural Information Processing Systems, 34, 2021.

[4] TorchSSL: https://github.com/TorchSSL/TorchSSL

相关链接：

文章链接：
https://arxiv.org/pdf/2208.07204.pdf

代码链接：
https://github.com/microsoft/Semi-supervised-learning

如何高效、精准地进行图片搜索？看看轻量化视觉预训练模型

你是否有过图像检索的烦恼？或是难以在海量化的图像中准确地找到所需图像，或是在基于文本的检索中得到差强人意的结果。对于这个难题，微软亚洲研究院和微软云计算与人工智能事业部的研究人员对轻量化视觉模型进行了深入研究，并提出了一系列视觉预训练模型的设计和压缩方法，实现了视觉 Transformer 的轻量化部署需求。目前该方法和模型已成功应用于微软必应搜索引擎，实现了百亿图片的精准、快速推理和检索。本文将深入讲解轻量化视觉预训练模型的发展、关键技术、应用和潜力，以及未来的机遇和挑战，希望大家可以更好地了解轻量化视觉预训练领域，共同推进相关技术的发展。

近来，基于 Transformer 的视觉预训练模型在诸多计算机视觉任务上都取得了优越性能，受到了广泛关注。然而，视觉 Transformer 预训练模型通常参数量大、复杂度高，制约了其在实际应用中的部署和使用，尤其是在资源受限的设备中或者对实时性要求很高的场景中。因此，视觉预训练大模型的"轻量化"研究成为了学术界和工业界关注的新热点。

对此，微软亚洲研究院和微软云计算与人工智能事业部的研究员们在视觉大模型的结构设计和训练推断上进行了深入探索，同时还对大模型的轻量化、实时性以及云端部署也做了创新应用。本文将从轻量化视觉预训练模型的发展谈起，探讨模型轻量化研究中的关键技术，以及轻量化视觉 Transformer 模型在实际产品中的应用和潜力，并展望轻量化视觉模型的未来发展机遇和挑战。

视觉大模型层出不穷，轻量化预训练模型却乏人问津

最近几年，深度学习在 ImageNet 图像分类任务上的进展主要得益于对视觉模型容量的大幅度扩增。如图 1 所示，短短几年时间，视觉预训练模型的容量扩大了 300 多倍，从 4,450 万参数的 ResNet-101 模型，进化到了拥有 150 亿参数的 V-MoE 模型，这些大型视觉预训练模型在图像理解和视觉内容生成等任务上都取得了长足进步。

无论是微软的 30 亿参数 Swin-V2 模型，还是谷歌发布的 18 亿参数 ViT-G/14 模型，视觉大模型在众多任务中都展现了优越的性能，尤其是其强大的小样本 (few-shot) 甚至是零样本 (zero-shot) 的泛化能力，对实现通用智能非常关键。

然而，在很多实际场景中，由于存储、计算资源的限制，大模型难以直接部署或者无法满足实时需求。因此，轻量级的视觉预训练模型研究变得越来越重要，且具有很强的实际应用价值。尽管目前有一些工作在探讨轻量级模型，但是这些方法大多是针对特定任务、特定结构设计的，在设计和训练过程中没有考虑到模型的通用性，存在跨数据域、跨任务的泛化局限性。

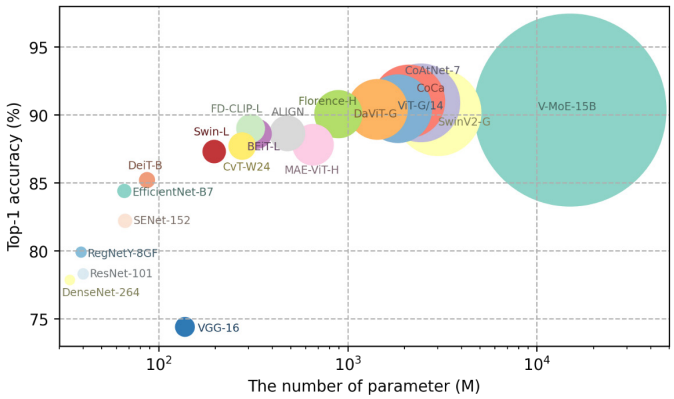


图 1：视觉预训练模型参数量的变化趋势图

轻量化视觉模型的关键技术研究

为了实现轻量化视觉预训练模型，微软的研究员们发现了两大关键问题：1）如何设计出通用性更强的轻量化模型结构？ 2）受制于轻量化视觉预训练模型的有限容量，如何设计高效的预训练方法让小模型也能学习到大规模数据中的有效信息？面对这些难题，研究员们通过坚持不懈的研究和探索，目前取得了一些阶段性成果。

由于提高轻量化预训练模型通用性的核心在于如何在资源受限（参数量，时延等）的情况下强化模型的学习能力，使其能够更好地在大规模数据中学习通用特征，因此，研究员们从以下三个角度进行了深入探索：

1. 轻量化模块设计

轻量、低延时的模块是组成轻量级模型的重要部分。在卷积神经网络中，具有代表性的轻量级模块有 MobileNet 的反向残差

模块（Inverted Residual Block）以及 ShuffleNet 的通道随机交叉单元（Shuffle Unit）。在视觉 Transformer 结构中，由于图像块之间注意力的计算没有很好地考虑相对位置编码信息，因此研究员们设计了即插即用的轻量级二维图像相对位置编码方法 iRPE [1]，它不需要修改任何的训练超参数，就能提高模型的性能。此外，针对视觉 Transformer 参数冗余的问题，研究员们设计了权重多路复用（Weight Multiplexing）模块 [2]。如图 2 所示，该方法通过多层权重复用减少模型参数的冗余性，并且引入不共享的线性变换，提高参数的多样性。

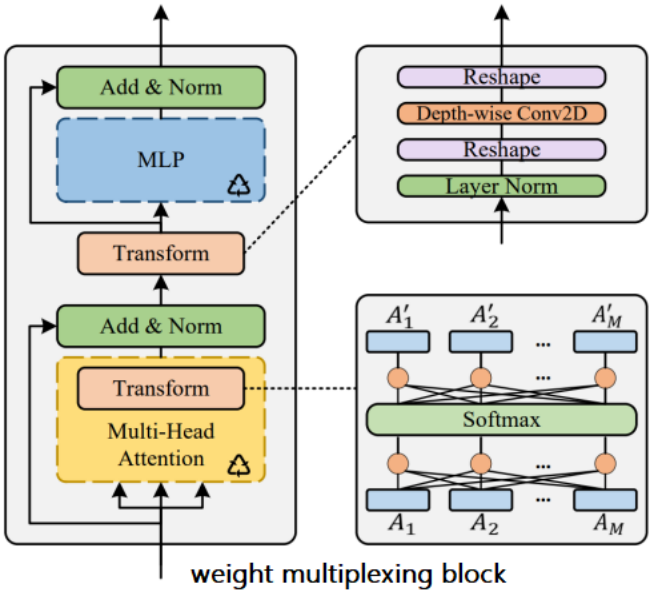


图 2：Transformer 中的权重多路复用模块

2. 轻量化模型搜索

网络结构搜索（Neural Architecture Search）可以从模型设计空间中自动找到更加轻量、性能更加优异的模型结构 [3]。在卷积神经网络中，代表性工作有 NASNet 和 EfficientNet 等。在视觉 Transformer 结构搜索中，针对视觉模型中的通道宽度、网络深度以及 head 数量等多个维度，研究员们先后提出了 AutoFormer [4] 和 S3 [5]，实现了视觉模型的动态可伸缩训练与结构搜索。在同样模型精度的情况下，搜索得到的新模型具有更小的参数量和计算量。值得注意的是，在 S3 中，研究员们利用 E-T Error [5] 以及权重共享超网来指导、改进搜索空间，在得到更高效的模型结构的同时也分析了搜索空间的演进过程，如图 3 所示。与此同时，模型结构搜索的过程为轻量化模型的设计提供了有效的设计经验和参考。

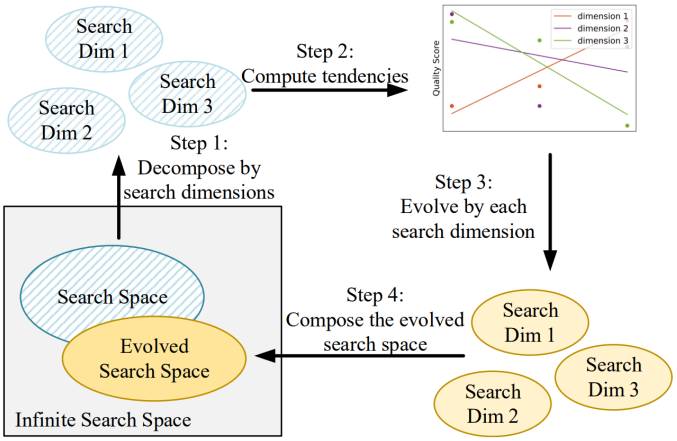


图 3：轻量级模型搜索空间进化过程

3. 视觉大模型压缩与知识迁移

轻量级预训练模型的另一难题在于，由于模型容量有限，难以直接学习大规模数据中包含的丰富信息和知识。为了解决这一问题，研究员们提出了快速预训练蒸馏方案，将大模型的知识迁移到轻量化的小模型中 [6]。如图 4 所示，和传统的单阶段知识蒸馏不同，快速预训练蒸馏分为两个阶段：1）压缩并保存大模型训练过程中使用的数据增广信息和预测信息；2）加载并恢复大模型的预测信息和数据增广后，利用大模型作为教师，通过预训练蒸馏指导轻量化学生模型的学习和训练。不同于剪枝和量化，该方法在权重共享的基础上使用了上文中提到的权重复用 [2]，通过引入轻量级权重变换和蒸馏，成功压缩视觉预训练大模型，得到了通用性更强的轻量级模型。在不牺牲性能的情况下，该方法可以将原有大模型压缩数十倍。

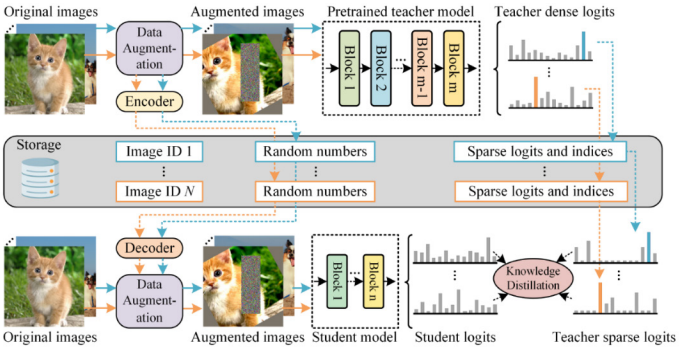


图 4：快速预训练知识蒸馏

这一系列的研究成果，不仅在计算机视觉的顶级学术会议上（CVPR、ICCV、ECCV、NeurIPS 等）发表了多篇论文 [1-6]，也通过和微软必应的合作，成功将轻量化预训练模型应用到了图像搜索产品中，提高了实际业务中图像和视频内容理解的能力。

轻量级视觉预训练模型的应用

轻量级视觉预训练模型在实际中有诸多用途，尤其是在实时性要求高或者资源受限的场景中，例如：云端视频实时渲染和增强、端测图像、视频内容理解。轻量级视觉模型已经在智能零售、先进制造业等领域展现出了广阔的应用前景，将来还会在元宇宙、自动驾驶等新兴行业发挥重要作用。下面以微软必应产品中的图像内容搜索为例，展示轻量化视觉模型的实际应用和部署。

目前，基于内容的图片搜索在图片的类别属性理解上已经比较成熟，但对于复杂场景的内容理解仍有很大的挑战。复杂场景的图片通常具有大景深、背景杂乱、人物多、物体关系复杂等特点，显著地增加了内容理解的难度，因而对预训练模型的鲁棒性和泛化性提出了更高的要求。

举例来说，动漫图片的搜索质量在很长一段时间内无法得到有效提升，其主要的挑战包括：绘画线条和颜色比真实场景图片更加夸张，包含更多动作和场景，不同漫画之间的风格内容差异巨大。图 5 展示了"灌篮高手"的动漫人物和行为，其动作表现和内容差别迥异。如何有效地理解漫画图片内容，对视觉预训练模型提出了较高的要求。

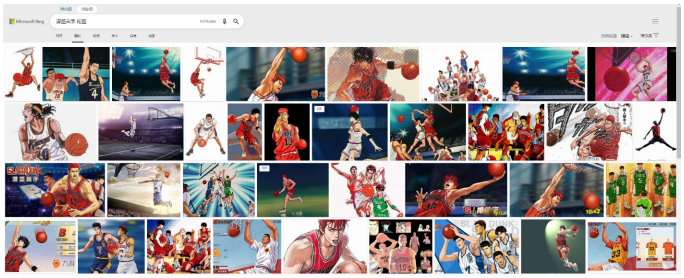


图 5：在微软必应搜索引擎中，对灌篮高手的动作理解包括：扣篮、运球、抢断、投篮等

上文中提到的轻量级视觉通用模型以及快速预训练蒸馏算法目前已成功应用于微软必应搜索引擎中。借助微软亚洲研究院提供的视觉语言多模态预训练模型，微软必应图片搜索功能增强了对漫画内容的理解，可以返回与用户需求更为匹配的图片内容。

与此同时，微软必应搜索引擎庞大的索引库对于检索效率有非常高的要求。微软亚洲研究院提供的快速预训练蒸馏方法有效地将预训练大模型的索引能力迁移到轻量化模型中，在识别准确率上将现有模型提升了 14%，同时极大地优化了模型的计算效率，实现了百亿图片的快速推理。

未来的机遇与挑战

模型轻量化是人工智能未来应用落地的核心。随着视觉技术、算法、算力和数据等不断完善，模型的复杂度急剧攀升，神经网络计算的能耗代价越来越高。轻量化视觉模型高效的计算效率和低廉的部署应用成本，能够在未来更多的实际产品中发挥巨大优势。除此之外，本地化的轻量级预训练视觉模型在支持更多服务的同时，还能够更好地保护用户数据和隐私。用户的数据将不再需要离开设备，即可实现模型服务等功能的远程升级。

当然，研究人员也意识到轻量级预训练视觉模型所面临的挑战：一方面在模型结构设计上，如何在模型参数量和推理延时的限制下达到模型的最优学习能力，一直以来都是学术界和工业界密切关注的问题。虽然目前已经沉淀了不少有效的模型结构，在通用近似定理（UAT）、神经网络结构搜索（NAS）等领域也取得了长足的发展，但是现有的轻量级预训练视觉模型和视觉大模型之间仍有差距，有待进一步优化和提升。另一方面在训练方法上，学术界和工业界针对视觉大模型提出了自监督、图像分类和多模态等多种训练方法，显著提升了模型的通用能力。如何针对容量有限的轻量级模型设计更有效的训练方式，还需要进一步的研究和探索。微软亚洲研究院的研究员们将不断推进轻量级预训练视觉模型的科研进展，也欢迎更多科技同仁共同交流、探索该领域的相关技术。

参考文献

[1] Rethinking and Improving Relative Position Encoding for Vision Transformer, ICCV 2021.

[2] MiniViT: Compressing Vision Transformers with Weight Multiplexing, CVPR 2022.

[3] Cyclic Differentiable Architecture Search, TPAMI 2022.

[4] AutoFormer: Searching Transformers for Visual Recognition, ICCV 2021.

[5] Searching the Search Space of Vision Transformer, NeurIPS 2021.

[6] TinyViT: Fast Pretraining Distillation for Small Vision Transformers, ECCV 2022.

像编辑文本一样编辑语音，可能吗？

如今在各种社交网络平台上发布的视频，因拍摄便捷、可实时分享、互动交流等特点而深受大众喜爱。视频深刻影响和改变了人们观察世界、记录生活和表达情感的方式。然而，现在市面上许多视频或音频剪辑软件为了满足用户需求尽管拥有丰富的功能，但操作却很复杂，很多简单的剪辑任务都还需要在软件中逐帧对照确定剪切时间点。对于以语音为主要背景声音的视频，如线上会议录像、演示视频、Vlog 等，如果我们能通过编辑文本的形式，直接编辑音视频中的语音内容，让音视频的编辑自动根据文本完成，那么将大大降低音视频的编辑难度，提高创作者的效率。为此，微软亚洲研究院的研究员们研发了一个基于文本的语音编辑系统。本文将详细介绍这个基于文本的语音编辑系统和研究员们研发的语音合成及填充词检测技术。

无论是演示视频、教学视频、会议录像还是记录生活片段的 Vlog，在很多实际的应用场景中，人们常常需要重新录制语音（视频）或对语音（视频）进行编辑。因为拍摄的素材中往往会存在大量停顿和脱口而出、词不达意的语句，或者是冗余的内容。但由于声音的特性，我们没有办法在录音底本的基础上去修改字词，只能一帧一帧在剪辑上下功夫，因此声音的剪辑工作繁琐又充满挑战。如果拥有一个基于文本的语音编辑系统，可以通过直接编辑语音对应的文本，完成对语音（视频）的编辑，那么普通用户也能成为一个有创意的剪辑师，把一段冗杂的音视频变得清晰、自然又专业。

市场上现在已经有一些类似的产品或相关的研究工作，但都有一些限制：有的研究工作可以根据文本合成匹配上下文的语音，但是必须是模型训练过程中学习过的音色；有的产品想合成定制化的声音，比如用户自己的音色，但需要用户准备至少 10 分钟的声音，并将声音上传，然后再等待 2-24 小时，通过后台对声音进行训练之后，软件才可以合成定制化的声音。这些限制无疑都给基于文本的语音编辑在现实中的使用带来了极大的不便。为此，微软亚洲研究院的研究员们研发了一个基于文本的语音编辑系统，来解决这些技术难点。

技术难点

在以语音为主的音视频中，语音中的内容和文本有着时间上的一一对应。研究员们发现，若要让基于文本的语音编辑系统可以直接编辑文本，再根据语音和文本的对应关系自动完成语音的编辑，需要着重关注以下技术要点：

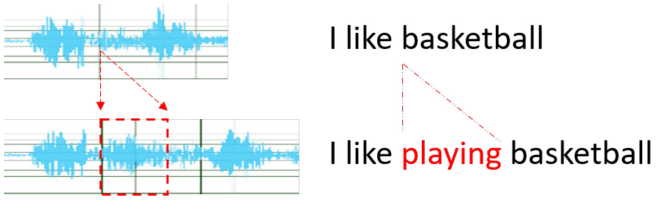


图 1：语音和文本的对应关系

1. 自动语音识别：如果语音不是按照已有脚本读的，那就没有文本信息，需要 ASR（自动语音识别）来识别得到文本，现有的 ASR 系统已经能够准确地识别语音，但受限于训练数据，部分 ASR 系统并不能完全检测出语音中的填充词。

2. 语音和文本对齐：研究中需要一个语音文本对齐（forced alignment）模块来提供准确的文本和语音的对齐结果，以便可以精确定位到要编辑的文本在语音中的时间戳。这是一个非常基础但十分重要的技术点，也是一个传统的研究问题，一般用于语言学研究或发音评估或为 TTS（语音合成）提供对齐的训练数据。现有的强制对齐方法几乎可以满足这些需求，但是对于基于文本的语音编辑，则需要更准确的对齐方式。现有的对齐方法在几十毫秒的误差级别下仍不能做到完美，而一旦出现几十毫秒的误差，比如切割语音的时候多切或少切了几十毫秒，人的听觉会很容易察觉到，并产生不适。

3. 语音合成：当插入或修改文本时，需要语音合成模块来生成新的声音。语音合成的最大挑战是自然和流畅，对于基于文本的语音编辑尤其重要，因为如果只修改语音的一部分，稍有不连贯就会非常明显。而且研究员们期望在使用语音编辑技术的时候，可以随意进行编辑而不需要准备足量的语音数据去微调模型。因此一个零样本上下文感知（zero-shot context-aware）的 TTS 是必不可少的。

4. 填充词检测：填充词检测模块可以自动检测语音中的填充词，用户可以选择手动删除部分或自动全部删除。上文提到，部分 ASR 系统不能检测到全部填充词，此时就没办法通过文本删除填充词来编辑语音。有的词是否是填充词可能取决于语境，比如 “you know” 是英语中常用的填充词，但是在 “Do you know him?” 这句话中它并不是填充词，这就需要一个语言模型来进行判断。

基于文本的语音编辑系统

研究员们首先调用微软云计算平台 Azure 上的 ASR 服务将上传的语音文件转化为文本，同时调用自行研发的填充词检测模型，并将填充词检测结果和 ASR 识别结果合并。然后就可以对文本进行编辑——对于删除操作，系统会根据对齐结果删除对应的语音片段；对于插入操作，系统会调用语音合成模型合成要插入的语音并插入原有语音中。下面是几个通过上述方法完成的基于文本的语音编辑样例：

修改文本样例

原始文本：
understand for the question and answer benchmark we're also the Þrst reach human parity
修改后文本：
understand for the question and answer benchmark we're also the second reach human parity

插入文本样例

原始文本：
The song of the wretched
修改后文本：
The famous song of the wretched

删除文本样例

原始文本：
some have accepted it as a miracle without physical explanation
修改后文本：
some have accepted it without physical explanation

填充词检测和去除样例

原始文本：
We can edit your speech uh by just editing.You know, its transcript.
修改后文本：
We can edit your speech by just editing its transcript.

扫描二维码，听一听这些语音编辑样例

基于文本的语音编辑系统关键技术点一：语音合成

服务于语音编辑的语音合成模型需要做到三点：零样本、自然和流畅。其中自然又可细化为两点子要求：生成与目标说话人相似的音色，以及足够高的音质。经过不断探索，微软亚洲研究院的研究员们达成了以上目标，开发了一个零样本上下文感知 TTS 模型 RetrieverTTS，并在语音领域的顶级学术会议 InterSpeech 2022 上发表了论文"RetrieverTTS: Modeling Decomposed Factors for Text-Based Speech Insertion"（欲了解论文详情，请查看：https://arxiv.org/pdf/2206.13865.pdf）。

设计思路

不同于已有方法中将语音插入任务视为文本 - 语音模态融合的思路 [3,5]，如图 2 所示，研究员们将语音先解耦成文本、韵律（音素序列的音高、音量、时长）、音色、风格四个要素，再在每个要素上进行可控的编辑操作，最后将四部分合成为插入后的语音。一言以蔽之，即"先解耦再编辑"。

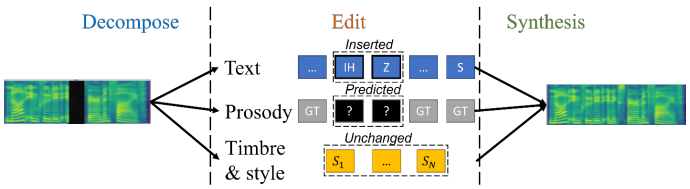


图 2：语音合成设计思路

但是，在执行插入操作时对四种要素的操作是不同的：文本可以直接使用用户编辑后的文本；对于韵律而言，未被编辑的部分无需改变，而插入词的部分需要由模型根据上下文预测得到；由于说话人没有改变，因此音色和风格两个要素保持不变。文本和韵律在一句话的不同时刻是不同的，而音色和风格在一句话甚至连续的几句话中都不会改变，所以前者为局域要素，而后者为全局要素。

模型架构

研究员们使用 Fastpitch [1] 作为语音合成的主干网络。为了准确地以零样本的方式适应到任意说话者的音色，全局要素与局域要素之间的解耦应足够精准，全局要素的表征需足够完备且应泛化至任意说话人。研究员们在 ICLR 2022 发表的论文"Retriever: Learning Content-Style Representation as a Token-Level Bipartite Graph"[2] 中，已经在很大程度上解决了这一问题，并在零样本语音风格转换任务中取得了最先进的性能。在此，研究员们将"Retriever"中的全局要素建模方法引入到了语音插入任务中。

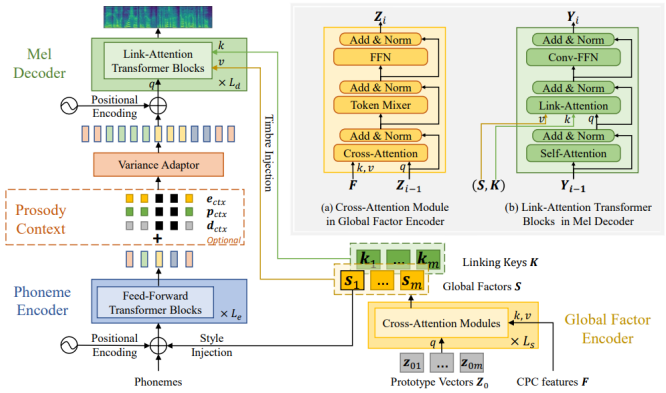


图 3：语音合成模型架构

实验结果

如表 1 所示，RetrieverTTS 的语音插入效果对插入长度并不敏感。对于训练中没见过的说话人，即使插入语音长度超过两秒乃至生成一整句（long insert, full generation），其语音自然度仍然保持在较高水平。在插入少于 6 个词时 （short insert, mid insert），甚至达到与真人录音相近的自然度评分。

	insertion length	MOS
ground-truth	—	4.01 ± 0.17
short insert	~ 0.63 s	3.97 ± 0.17
mid insert	~ 1.28 s	3.98 ± 0.17
long insert	~ 2.18 s	3.83 ± 0.16
full generation	—	3.85 ± 0.19

表 1：语音合成对不同插入长度的鲁棒性测试

在表 2 消融实验中，研究员们分别去掉了对抗训练（- adv），韵律平滑任务（- prosody-smooth），以及 Retriever 的全局要素建模方法 （- retriever），结果均发现明显的性能下降。在三次实验的测试样例中，分别出现了音质差、韵律不连贯以及音色不像的问题。这验证了 RetrieverTTS 三方面设计均达到了设计初衷。

	MOS@mid	SMOS
Ground truth	4.27 ± 0.21	4.45 ± 0.19
Full method	**3.93 ± 0.23**	**3.70 ± 0.25**
- adv	3.53 ± 0.23	—
- prosody-smooth	3.82 ± 0.22	—
- retriever	3.75 ± 0.22	3.60 ± 0.27

表 2：语音合成消融实验

在表 3 中，研究员们与其他方法进行了对比，发现基于模态融合的方法 [3] 在插入较长语音的情形下语音自然度表现较差，而其他的零样本说话人自适应语音合成（zero-shot speaker adaptive TTS）[4] 在音色相似度方面与 RetrieverTTS 的方法有较大差距。上述对比体现出了 RetrieverTTS 方法的优越性。

	MOS@long	SMOS
Ground-Truth	4.20 ± 0.14	4.45 ± 0.19
C.Tang et al. [3]	2.73 ± 0.25	—
Meta-StyleSpeech [4]	3.88 ± 0.20	2.91 ± 0.31
RetrieverTTS (Ours)	**3.95 ± 0.17**	**3.70 ± 0.25**

表 3：语音合成系统对比

基于文本的语音编辑系统关键技术点二：填充词检测

很多 ASR 模型由于受限于训练数据，不能完整的检测到填充词，因此需要一个单独的模块进行填充词的检测。事实上，基于语音的填充词检测技术属于语音关键词检测的一个特例，所以研究员们将语音关键词检测视为目标检测问题，而不是语音分类问题。受计算机视觉中目标检测方法的启发 [6]，研究员们提出了一种名为 AF-KWS（anchor free detector for continuous speech keyword spotting） 的关键词检测方法。

在 AF-KWS 方法中，研究员们通过预测一个关键词热力图，得到每一类关键词在连续语音中的位置，然后通过两个预测模块，分别预测关键词的长度和用于矫正关键词位置误差的位置偏移量。不同于计算机视觉中的目标检测算法 [6][7][8]，研究员们引入了一个 “unknown” 类别，表示非目标关键词的其他词，这种设计将 “unknown” 和语音中的背景噪音和安静片段分开，能够显著提高关键词检测的准确性。该方法的论文 “An Anchor-Free Detector for Continuous Speech Keyword Spotting” 已经被 InterSpeech 2022 接收（更多论文细节，请查看：https://arxiv.org/pdf/2208.04622.pdf ）。

算法框架

如图 4 所示，对于输入语音，研究员们首先提取了语音的 STFT 频谱图，然后使用 ResNet[9] 进一步提取特征，然后将特征输入三个预测模块，分别用于预测关键词的热力图、关键词的长度和关键词位置偏置。在训练阶段，热力图以关键词的位置为中心，使用高斯核函数将关键词的位置扩展。在预测阶段，研究员们取预测得到的热力图的峰值点作为预测得到的关键词的位置，然后提取对应位置的关键词长度和偏置的预测结果，计算得到最终的关键词的位置和类别。

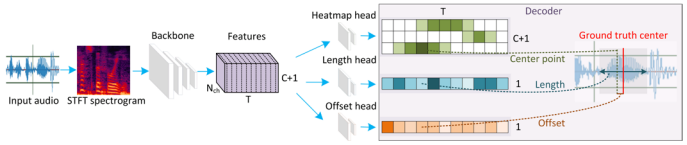


图 4：AF-KWS 算法框架

实验结论

研究员们选用了两个先进的关键词检测模型 [10][11] 做对比，在连续语音关键词检测的实验设定中，AF-KWS 的模型在可比的运行速度下，平均准确率远超其他模型。

Model	AP@5↑	AP@75↑	mAP↑	FRR@5↓	FRR@15↓	FRR@25↓	Classification Accuracy↑	RTF↓
DSTC-ResNet	0.748	0.058	0.398	0.647	0.519	0.402	0.961	0.018
MHAtt-RNN	0.795	0.076	0.426	0.530	0.418	0.374	0.978	0.057
AF-KWS (ours)	0.952	0.886	0.860	0.140	0.074	0.049	N/A	0.031

表 4：填充词检测性能对比

为了验证 AF-KWS 方法的提升不是因为 backbone 更强大，研究员们将三个预测模块替换成了一个分类模块（AF-KWS-cls），发现模型性能明显下降。为了验证引入的 “unknown” 类别的有效性，研究员们去掉了这个类别 (w/o unknown)，发现模型性能也明显下降。

Model	AP@5	AP@75	FRR@5	FRR@25
AF-KWS-cls	0.876	0.062	0.570	0.265
w/o unknown	0.867	0.800	0.234	0.117
AF-KWS (ours)	0.952	0.886	0.140	0.049

表 5：填充词检测消融实验

关键词检测模型用于填充词检测

由于填充词也可以看作一种特殊的关键词，所以研究员们基于 SwitchBoard 数据集 [12]，制作了一个语音填充词检测数据集，并在这个数据集上重新训练关键词检测模型。在真实的测试数据中，AF-KWS 方法得到了与市面上最好的方法几乎相同的性能。针对填充词的特点，比如填充词一般包含的音节较少，更容易与特定类别的词混淆，研究员们会继续改进模型。

未来展望

尽管现有的技术和本文中的语音编辑系统已经实现基于文本的语音编辑的部分功能，但仍有很多研究需要持续探索，包括：开发富文本格式，进行语音解耦，精准控制语音的重度，语气语调和情绪；开发更精确的语音文本对齐算法；在 TTS 中背景噪音或背景音乐进行建模，让合成的语音包含匹配的背景噪声或背景音乐；开发结合语音和文本的多模态填充词检测检测算法等等。

参考文献：

[1] A. Lancucki, “Fastpitch: Parallel text-to-speech with pitch prediction,” in ICASSP, 2021.
[2] Y. Dacheng, R. Xuanchi, L. Chong, W. Yuwang, X. Zhiwei, and Z. Wenjun, “Retriever: Learning content-style representation as a token-level bipartite graph,” in ICLR, 2022.
[3] C. Tang, C. Luo, Z. Zhao, D. Yin, Y. Zhao, and W. Zeng, “Zero-shot text-to-speech for text-based insertion in audio narration,” in Interspeech, 2021.
[4] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, “Meta-stylespeech : Multi-speaker adaptive text-to-speech generation,” in ICML, 2021.
[5] Z. Borsos, M. Sharifi, and M. Tagliasacchi, “Speechpainter: Textconditioned speech inpainting,” arXiv preprint arXiv:2202.07273, 2022.
[6] X. Zhou, D. Wang, and P. Kr¨ahenb¨uhl, “Objects as points,” arXiv preprint arXiv:1904.07850, 2019.
[7] C. Zhu, Y. He, and M. Savvides, “Feature selective anchor-free module for single-shot object detection,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 840–849.
[8] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, “Foveabox: Beyound anchor-based object detection,” IEEE Transactions on Image Processing, vol. 29, pp. 7389–7398, 2020.
[9] B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 466–481.
[10] O. Rybakov, N. Kononenko, N. Subrahmanya, M. Visontai, and S. Laurenzo, “Streaming keyword spotting on mobile devices,” Proc. Interspeech 2020, pp. 2277–2281, 2020.
[11] S. Majumdar and B. Ginsburg, “Matchboxnet: 1d time-channel separable convolutional neural network architecture for speech commands recognition,” Proc. Interspeech 2020, pp. 3356–3360, 2020.
[12] John J Godfrey, Edward C Holliman, and Jane Mc-Daniel, “Switchboard: Telephone speech corpus for research and development,” in Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on. IEEE, 1992, vol. 1, pp. 517–520.

科研第一线

微软亚洲研究院论文荣获“SIGKDD Test of Time Award”（时间检验奖）

8 月 14 日至 18 日，能够反映最前沿数据领域研究风向、被誉为全球数据挖掘最高级别的学术会议 ACM SIGKDD 2022 正式举行。微软亚洲研究院发表于 KDD 2012 的论文《Discovering regions of different functions in a city using human mobility and POIs》荣获 “SIGKDD Test of Time Award”（时间检验奖）。在这篇论文中，作者们基于大规模人群移动数据以及地图兴趣点数据，提出了一种基于数据的城市功能分区自动发现方法，从而帮助人们更深入地理解不断演化的大城市。

扫描二维码了解更多信息

文档智能多模态预训练模型：兼具通用性与优越性

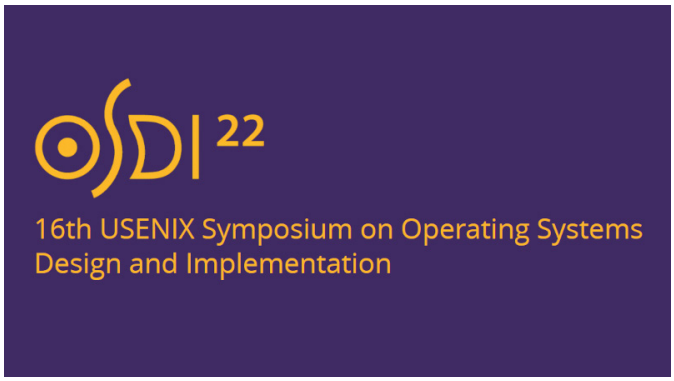
企业数字化转型中，以文档、图像等多模态形式为载体的结构化分析和内容提取是其中的关键一环，快速、自动、精准地处理包括合同、票据、报告等信息，对提升现代企业生产效率至关重要。因此，文档智能技术应运而生。过去几年，微软亚洲研究院推出了通用文档理解预训练 LayoutLM 系列研究成果，并不断优化模型对文档中文本、布局和视觉信息的预训练性能。近期发表的最新的 LayoutLM 3.0 版本，在以文本和图像为中心的任务上有了更加出色的表现，让文档理解模型向跨模态对齐迈出一大步！

扫描二维码了解更多信息

ICML 2022 | 请查收这份机器学习前沿论文精选

ICML 被认为是人工智能、机器学习领域最顶级的国际会议之一，在计算机科学界享有崇高的声望。ICML 2022 于 7 月 17 日 -23 日以线上线下结合的方式举办。我们精选了微软亚洲研究院在此次大会上发表的 7 篇论文，来为大家进行简要介绍，从强化学习、图神经网络、知识图谱表示学习等关键词带你一览机器学习领域的最新成果！

扫描二维码了解更多信息



OSDI 2022 | 微软亚洲研究院计算机系统领域最新论文

OSDI（Operating Systems Design and Implementation）是计算机系统领域最顶级的学术会议之一，汇集了全球计算机科学家们对于计算机系统的前瞻性思考。第 16 届 OSDI 于 2022 年 7 月 11 日至 13 日召开，本次会议共有 253 篇论文投稿，接收 49 篇，接收率为 19.4%。本文中，我们将分享微软亚洲研究院被 OSDI 2022 收录的 3 篇论文，希望可以帮助大家了解计算机系统领域的前沿趋势。

扫描二维码了解更多信息

邓攀的“贪心”算法：从生物跨界到计算机是什么体验?

科研之路并非繁花似锦，很多时候是在一条没有脚印的道路上探索未知。科研之路应该怎么走？如何抓住机遇实现转弯？微软亚洲研究院主管研究员邓攀在以《人生的“贪心”算法》为题的演讲中，分享了自己从本科毕业到现在一路走来的经历与收获。从生物跨界到计算机，邓攀是如何做到“内心有谱，丝毫不慌”的？遇到机会，她又是如何竭尽全力把握住每一个可能的？一起来看邓攀怎样编写了自己人生的“贪心”算法吧！

大家好，我是邓攀。我本科就读于清华大学生命学院，期间进行的是生殖干细胞相关的研究。博士的研究方向是线粒体，研究细胞的能量中心在各种毒性损伤下的应激反应。现在，我是微软亚洲研究院的一名研究员。

有时候我自己都会诧异：我一个学生物的，怎么就来了微软呢？我来了微软，怎么还在做生物呢？我是怎么做到这么酷的事情的呢？话说回来，我现在确实是在做着自己非常喜欢的事情。今天，我将和大家分享我一路走来的经历与收获。我的分享主题是《人生的“贪心”算法》——回顾自己从本科毕业到现在的十年历程，“贪心”的思想真是精确描述了我的每一步选择。

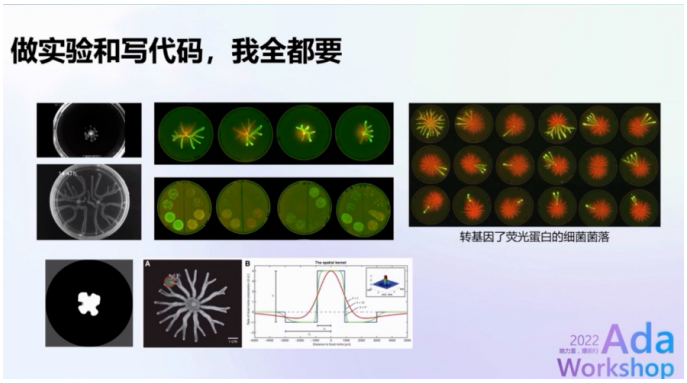
做实验和写代码，我全都要

故事开始于 2012 年，是我抵达纽约、开始博士学习的第一年。

我们研究生院有“轮转”制度，每个新入学的博士生在第一年的时候都可以选择 3-5 个自己感兴趣的实验室，进行 2-3 个月的“短期体验”，再决定自己想要加入哪个实验室进行博士阶段的研究。

那个时候，计算生物学还是一个挺小众的方向。出于好奇，我选择了一个利用计算方法研究微生物群体行为的实验室。

我在那个实验室轮转时进行了一个短期的探索项目。当时，我们发现：一种叫做假单胞杆菌的细菌被添加到培养基上之后，逐渐就会长成这种树杈状的分支结构。如果同时放上两个菌落，狭路相逢的时候，它们还会互不打扰，各退一步。



邓攀在实验室轮转时进行的科研探索项目

我们很好奇，在培养基上，是什么让细菌在扩散时出现这种模式的呢？相隔老远的细菌又是怎么知道彼此的存在的呢？

通过对微生物进行各种定向基因改造，结合很多脑洞大开的实验，我们找到了影响菌落形态和菌落间通讯的分泌因子，用数学模型解释了菌落行为的原因，还在 Matlab 里模拟出了菌落群体的扩散过程。

现在讲起来，我还觉得这项研究非常有趣。当时，我也一度考虑过要加入这个实验室，继续这个方向的研究。

虽然生物和编程我都很感兴趣，但生物实验的训练必须要有场地和环境；而编程学起来则自由很多，我可以利用晚上和周末的时间学习。因此，最终我还是选择了传统生物学研究方向，利用课余时间自学编程。

不得不说，年轻时候的我真是精力旺盛。但也就是这个想法，才让我今天站在这里成为了可能。

不管有没有用，我全都学

说到做到。在接下来的几年里，虽然我主要进行的还是传统生物学研究，可我把研究生院里能选的关于计算生物、生物统计的课程全部上了一遍，还在 Coursera 上学习了很多公开课，并认真做了笔记、完成了作业。当时慕课刚刚兴起，网上有非常多高质量的公开课程。

邓攀的线上学习部分摘录

其实当时我也不知道这些课程对于未来的生物研究有没有帮助，或者对我未来找工作有没有帮助，大多时候我的学习纯粹出于对知识的好奇。

举例来说，听说一门算法课教得很好，但是这门课程的语言是 Java，那我就去学习 Java；听说 C++ 更能培养计算思维，我就去 USACO 上用 C++ 刷题（那时候 LeetCode 刚刚成立）；听说大数据云计算特别 fashion，我就去学习了云计算的课程……

后来，这些"不功利"的学习经历都对我起到了非常大的帮助。比如来微软面试算法时，我能够"内心有谱，丝毫不慌"。这段算法学习经历也帮我获得了第一次计算机方向实习的可能。

遇到机会，就全力以赴

这些年来，我有一个观察：中国学生和欧美学生相比，女生和男生相比，都更容易说一句话：我不行。但真的是自己不行吗？还是你觉得自己没有准备好？因为自己没有十足的把握，害怕失败，所以不敢尝试？

由于专业性质的原因，生物学专业的学生一般是不会在在校时到企业实习的。但博四那年春天，我突然听到两个消息：一是实验室要在那年夏天从纽约搬到麻省的伍斯特，所以暑假的时间几乎进行不了实验；二是有一个名为 Google Summer of Code 的暑期实习项目正在接收申请。

Google Summer of Code 这个项目每年二月份启动，组织方先筛选一批符合标准的开源项目，再开放学生报名，由开源组织筛选他们心仪的学生。入选的学生会在暑期进行三个月的线上全职实习，为开源项目贡献代码，许多学生也会选择在实习结束后成为长期的开源贡献者。

这个项目对于当时只会闷头写代码的我来说是一个好消息，恰好那个暑假我也有充裕的时间。问题只有一个：我听到这个消息的时候，距离学生申请项目的截止日期只剩两周。

在这两个星期的时间里，我需要从一百多个开源项目中选出我想要申请的项目，按项目的要求完成代码任务，并提交申请书。而当时我对于 GitHub 的了解仅限于：在一堂不到 2 个小时的入门课上，我建立了账号，并 fork 了一次同学刚刚建立的 repo。

根据我当时可怜的技能和兴趣，我很快锁定了要申请的项目——一个用 C++ 编写的主要应用于生物医学研究的机器学习库。它们的项目主页上写着：欢迎女生申请。

当时的我想：我一定要抓住这个机会。

"我是女生，我是生物医学专业的，我写过 C++ 。"

靠着这点强行给自己找来的优势，我开始啃起了这个项目的代码任务：在线性代数库里增加均值计算的功能模块。听起来很简单，但对于当时的我来说好像要小学生去解一道微积分的习题。整个代码库有大约 50 万行代码，我需要从里面找到目标路径，理解 dependency，参照其它功能的实现形式来完成这个同时支持 CPU 和 GPU 后端的 feature，写好本地测试，在 Docker 里跑通，再通过 GitHub 提交这个任务——每一步都需要从头学起。

不瞒大家说，那两个星期里，我是边哭边完成这件事情的。看不懂代码、配不好环境、编译会报错、测试通不过，甚至别人的讨论我也看不明白。觉得好难，觉得我什么都不会，全都是问题，觉得没有时间了……

但最后，我还是坚持了下来，尽我所能的完成了 feature，并通过了 Pull Request 的所有测试，如愿以偿拿到了实习的 offer。

那时的我心虚吗？说实话，我是心虚的。申请时我简直是靠着一口气完成了任务，实习期间我又能够应付的过来吗？

但我要退缩吗？我不要。那时候我发了一条朋友圈，到现在我也觉得很有道理：人生的所有机会都是"赶鸭子上架"。想等到什么都准备好，可能就来不及了。

那个暑假，我重构了基于 C++11 特性的线性代数库后端，统一了数据存储和运算的接口，引入了全新的序列化模块，完成了近 4 万行代码的增删。后来，我继续贡献着代码，正式成为了团队的一员，在第二年担任了实习项目的 mentor，并参加了团队组

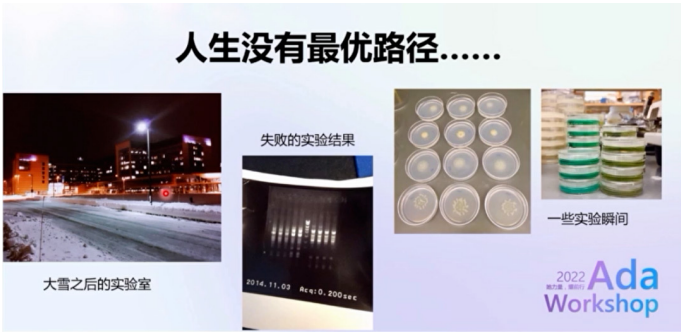
织的布达佩斯线下 hackathon。

这段经历极大地"膨胀"了我的自信。此后，我也曾在全球 C++ 开发者大会上"厚脸皮"地做过闪电演讲，在 San Jose 世界科幻大会上申请担任 coffee talk 活动的志愿者负责人，在刚加入微软亚洲研究院不久时，在全院活动上主动代表讨论小组向全院汇报……我发现，这种经历会不断带来正向的反馈。现在，面对我的确感到有挑战的任务，我不会说 "我不行" ，而是说"我没有把握，但我可以试试 "。

邓攀博士四年级的暑期

人生没有最优路径

你们是不是发现一个问题：以我现在的职业发展方向——计算生物学来看，我其实走了一些弯路。



邓攀演讲时分享的照片

读博中期那几年，我也曾十分困扰。在通宵实验连轴转却始终观察不到我想要的生物现象时；在周日的暴风雪里开车半小时冲到实验室却只看到了阴性结果继而只能打道回府时；在做了三个月的实验不知为何失败却又没有办法设断点排除故障时……我也曾后悔：为什么博士一年级的时候我没有选择计算生物方向呢？如果当初做出不同的选择，是不是就不会这么心力交瘁，也可以有更光明的未来——比如投入毕业就转码的大潮中，或者更早就乘上计算生物的这股东风，职业发展更加风生水起呢？

但现在我不再这么认为了。

首先，受限于我当时的视野与能力，我已经做出了当时最合理的选择。人生没有固定答案，我们也没有办法规划出最优路径。

其次，也许我看上去做了一些"无用功"，但我认真对待过的每一份经历，都会形成我的独特积淀，最终塑造出我的独特人生。

最后，最重要的是，在这个不停尝试和探索的过程中，我找到了自己真正热爱、愿意为之奋斗的领域。

找到自己真正的热爱，勇敢地走下去

科研其实是一件挺痛苦的事情，你太容易感受到失败和挫折。现在大家还一个比着一个的"卷"：今年你发了 5 篇论文，明年我就要发 10 篇，同辈压力令人难以承受。

这个时候，只有找到你自己真正热爱的领域，你才不会轻易被外界的压力裹挟，才能不去追逐热点与 low hanging fruit，而是真正静下心来，去思考、去推敲、去创造一些真正有价值的成果。

而如果你真的找到了自己的真爱，那就勇敢地走下去，不要轻易地放弃。毕竟，坚持和投入才是成功的诀窍。

在我心中，微软亚洲研究院一直是一个学术圣殿。准备面试的时候，我和朋友说：我高考恐怕都没有这么认真过。但当铁岩博士在面试中问我，除了计算生物我还对什么方向感兴趣的时候，我大概给出了一个标准的面试错误答案。我说：如果不是知道微软亚洲研究院在做计算生物，我可能就不投简历了。

在知道自己到底想要什么之后，说话就是这么硬气。

想想我来研究院已经快两年了，感觉我还处在和研究院的"蜜月期"。这里有极大的学术自由，尊重每个人的研究兴趣，还有太多优秀、靠谱的同事，可以进行思维的碰撞与跨领域的交流。虽然做研究依然让我时不时长吁短叹、抓耳挠腮，但我依然感觉这是在做一件让自己快乐的事情，我也总是充满了动力。

最后，用著名英国女作家弗吉尼亚·伍尔夫的一句话作为收尾吧 ： No need to hurry , no need to sparkle. No need to be anybody but oneself .

科研是一个不停求索的过程，人生也是。希望我今天的分享能够为年轻的大家带来一些帮助与启发。

实习派 | 何灏迪：大四一年从初出茅庐到顶会论文作者

2021 年 8 月，来自中国科学技术大学计算机科学与技术专业的何灏迪，通过微软创新人才学院项目来到微软亚洲研究院视觉计算组实习。从没有做过完整的科研项目，到完成一作论文并被 ECCV 2022 接收，何灏迪实现了自己初入微软亚洲研究院时立下的目标：在这里完成自己的科研项目。与此同时，他也收获了自己的第一篇国际顶会论文。

被 mentor 微软亚洲研究院研究员元玉慧评价为 “过去五年中带过最优秀的实习生之一” 的何灏迪，后续将到斯坦福大学进行博士阶段的深造，他也将带着从微软亚洲研究院收获的科研习惯和科研视野继续他的科研探索。

何灏迪

从初出茅庐到以一作收获顶会论文

一年前，看到微软创新人才学院的招生通知，正在读大三的何灏迪就毫不犹豫地投出了申请，“这在中科大是一个特别好的机会，大四如果可以出来实习，微软亚洲研究院是世界上最好的地方之一。” 通过微软中科大创新实践项目以及后续面试选拔，何灏迪顺利进入微软亚洲研究院视觉计算组，开启了他在这里一年的科研探索之旅。

微软创新人才学院旨在帮助有志于从事科研的学生，在本科阶段就能在国际一流的研究氛围中，发掘自己的科研潜力。入选微软创新人才学院的学生将在大四阶段在微软亚洲研究院实习一年，与微软亚洲研究院的研究员一起探索前沿课题。

初出茅庐的何灏迪并不具备丰富的科研经验。本科前三年，在学校以上课为主的环境中，何灏迪并没有很多进行科研项目的机会，只在课余时间参与过学校的一些课题。因此，这时的他也还没有明确的科研目标和未来的发展方向。

来到微软亚洲研究院后，何灏迪开始在 mentor 元玉慧的指导下进行自己的第一次完整科研尝试。

经过 mentor 的指导，何灏迪决定在计算机视觉的领域内选择“图片视频分割”作为自己的研究方向。“每天早上 9 点 30 到公司开始分析之前的实验结果，10 点 30 到 11 点 30 和 mentor 开会讨论研究进展并解决目前遇到的问题，下午开始写代码，做实验的迭代，晚上会针对一天的研究进度进行复盘。”这是何灏迪在微软亚洲研究院的工作日常，也是他为了实现目标做出的努力——在初入研究院时，他就决定要在这里完成自己的科研项目。

他的努力逐渐开花结果。一年的时间里，他以第一作者完成了“基于多标签类别排序的图片视频分割”、“对 DETR 框架的收敛加速研究”这两个项目。第一个项目中，他提出的 RankSeg 算法可应用于提高 MaskFormer、Mask2Former 等最先进的分割模型，该研究成果已被计算机视觉领域顶会 ECCV 2022 接收。

成功并非一蹴而就，这个科研项目也曾有一次被拒稿的经历。最初，何灏迪是基于 Segmenter 等像素分类型的分割模型进行研究，到 2021 年 11 月就有了一些不错的结果，于是他将论文投稿至 CVPR，收到的却是拒稿的消息。没有气馁，何灏迪开始根据审稿意见对研究进行改进。通过增加分析性实验，在不同的模型和数据集上进行测试，他更好地证明了提出方法的有效性和优越性，并收获了自己的第一篇顶会论文。

“在研究院，我完整经历了学术课题从最初寻找 idea 开始，到一步步深入研究、实现代码，进行论文的写作以及分析性实验设计，进行论文 rebuttal，最后发表、开源的整个过程，这个过程对我各个方面的能力都有充分的提升。” 何灏迪说，“之前，我对计算机视觉任务并没有那么全面的理解，只是从一两个子任务上入门，而经过这里的科研锻炼，我对领域内的各种问题都或多或少有了自己的理解。”

与 mentor 亲密无间的合作带他走出瓶颈

何灏迪在科研上的进步离不开 mentor 元玉慧的悉心指导。何灏迪用“紧密”来形容他和元玉慧的合作——每天都会一起讨论实验结果、一起设计新的网络结构、一起检查代码中可能的问题。在过程中，元玉慧提出的很多意见很好地促进了何灏迪对课题的思考，也补全了他忽略的细节问题，对课题研究和他的科研习惯都有很大的帮助。

在项目遇到瓶颈时，何灏迪也遭遇了心态上的焦虑。好在 mentor 元玉慧在背后不断给予支持，一方面他给何灏迪补充了许多在计算机视觉领域的知识，另一方面在他们每天的高频交流中，

元玉慧会在第一时间解决何灏迪遇到且无法解决的困难，并带领何灏迪不断地去做尝试。

“灏迪是我过去 5 年带过的实习生中最优秀的几个人之一。”元玉慧这样评价何灏迪，“因为何灏迪是第一次做计算机视觉研究，而且是在计算视觉竞争最激烈的赛道上，肯定会遇到非常多的困难，比如工程量大、实验迭代周期长等。我们每天都会一起讨论来解决这些困难。实际上，我们从开始一起合作 RankSeg 到第一次投稿仅仅花了 2 个半月，可以看到，我们做科研的节奏强度是相对很高的。”

何灏迪则认为 mentor 元玉慧是“很努力、很努力”的人。这种“努力”也激励何灏迪更主动地去进行科研探索。“因为我们一直在做讨论，所以就会不断有新的科研想法迸发出来。”何灏迪说。同时，在更高科研理念层面，何灏迪也在元玉慧的指导下有了更深的理解：要做有影响力的工作，在具体的科研中，还要注意把控住每一个细节。除了科研本身，元玉慧也很关心何灏迪的未来发展。在何灏迪申请斯坦福大学博士生的过程中，元玉慧也给予了许多帮助。

在微软创新人才学院开拓科研视野

作为微软创新人才学院的一员，除了常规的科研工作，何灏迪还通过学院为学生设计的一系列专属课程全方位地提升能力、开拓科研视野。

前沿研究讲座系列就是其中之一。该讲座由各领域的资深研究员带来关于领域过去发展、当下热点和未来趋势的洞察。因为在计算机视觉领域探索，微软亚洲研究院高级研究员胡瀚的分享给何灏迪带来的收获最大。“胡瀚老师对计算机视觉概念和对未来发展的介绍给了我很大的启发，他还指出了该领域的许多潜在机会。计算机和人类之间仍然存在着很大的差距。我希望自己将来能投身于这个领域，并努力将这种差距降到最低。”

创新人才学院开设的高级软件工程课程，也为何灏迪打开了“新天地”。在来微软亚洲研究院之前，何灏迪并未完整接触过软件工程领域相关的专业知识。学院的课程激发了何灏迪的学习兴趣，每一次课程上布置的项目任务他都和小组成员积极出色地完成。其中一次任务中，他与三位小组成员一起为 VSCO 制作了一个代码管理相关的插件，为了更完美地展示插件效果，灏迪还自学视频剪辑，制作了一个宣传视频，引起了不错的反响。

何灏迪用“自由平等”和“超级努力”两个词概括了微软亚洲研究院的科研氛围，他说：“加入微软亚洲研究院之后最大的感受就是，大家虽然已经取得了一定的成就，但还是依然在科研道路上坚持不懈的努力，比如我的 mentor 在项目的投入度上就非常高，还会竭尽所能地为实习生提供支持和帮助。微软亚洲研究院提供了一个非常平等的学术交流环境，在这里我有机会见到很多非常优秀的研究学者，他们也都非常愿意和同学们交流学术经验。”

微软创新人才学院结课合影

一年的时间，何灏迪在微软亚洲研究院从一个初出茅庐的科研新生，逐渐成长为可以独立进行项目、有自己科研想法的优秀科研力量。他也在这里认识了志同道合的朋友，培养了沟通交流、与人合作的能力。

奔赴斯坦福大学的读博之旅即将开启，何灏迪与微软亚洲研究院的缘分也不会就此终结，他将和 mentor 元玉慧一起针对之前的研究题目开展更加深入的探索，他也将带着微软亚洲研究院给他的科研馈赠不断前行！

何灏迪

Mentor 寄语

元玉慧
微软亚洲研究院
研究员

很高兴能够参与指导灏迪在微软亚洲研究院的这段科研实习，我和组里的同事一起见证了灏迪的成长、蜕变以及阶段性的成功。尤其值得高兴的是，灏迪不仅顺利拿到了斯坦福的博士 offer，也顺利以一作身份发表了他的第一篇顶会文章！希望灏迪未来在美国求学的日子中继续砥砺前行、迎难而上，早日成长为一名杰出的科研工作者！

寻星记｜在实习生里，寻找闪闪发光的你！

“寻星记”的灵感来源于微软亚洲研究院实习生同学的日常：平时在工位上一丝不苟写代码的程序员同学午休时为自己的乐队写歌；有着可爱外表的二次元小姐姐还是个软件开发大佬；工作上认真严谨的项目管理实习生背地里是百变小樱的热衷粉；时间管理能力超强的研发实习生居然已经是孩子他爹了 ……

微软亚洲研究院的每一个同学都集多项技能、多重身份于一身，快来看看身边每一位闪闪发光的同学吧！

2021 年 7 月入选中国科学技术大学 - 微软亚洲研究院联合培养博士生项目的刘逸菲，在微软亚洲研究院度过了充实又难忘的一年——同时探索两个科研方向，极具挑战也收获颇丰。逸菲的 mentor、微软亚洲研究院副院长杨懋认为，大四阶段应该着重拓宽知识广度，在研究生阶段再追求深度。考虑到逸菲的科研兴趣包括两个不同的方向，杨懋邀请了异构计算方向的曹婷和强化学习方向的薛卉这两位研究员来同时指导逸菲的科研工作。同时进行两个方向的科研探索，让刚开始着手做科研的逸菲颇感压力，也因此出现了精力不集中、效率不高等情况。

经过 mentor 们的悉心指导，以及逸菲自己的适应和摸索，她逐渐进入状态，掌握了多任务处理的能力，也学会了梳理待办事宜、排列优先级，从而更合理地安排每天的工作时间。

经过一年的探索，逸菲对自己感兴趣的两大方向逐渐入门。她在曹婷的指导下完成了为 Office 的语法检查模型做剪枝的工程项目，根据实践的启发，针对剪枝算法与稀疏计算实现相割裂的问题，提出了算法和系统实现协同设计的创新方法，从而使模型精度和运行性能都达到最优，该工作论文正在撰写当中。同时，她也在薛卉的指导下探索了以强化学习的复现为主的相关问题。

除了科研工作，逸菲还参加了微软亚洲研究院为入选联合博士生项目的大四学生安排的课程。通过高级软件工程、科技英语和前沿讲座三门课程，逸菲不仅提升了动手实践能力，也了解了各方向的前沿进展，全方位地提升了自己科研能力。

在微软亚洲研究院一整年的时间，逸菲深刻地体会到了这里 Work Hard, Play Harder 的文化，她不仅在学术上从科研小白进阶为有自己科研想法、能独立进行科研探索的小研究员，也在课余积极参加了微软音乐节、实习生圣诞派对等活动。博士阶段的逸菲将面对更多挑战，她期待自己能继续在微软亚洲研究院自由开放的平台下努力做科研，探索更多新颖的领域。

2022 年 5 月，来自加州大学洛杉矶分校的奕欣来到微软亚洲研究院实习，实习期间她努力完成工作、积极参加各种活动，在研究院里绽放着自己的的光芒。

“勇敢争取机会，不收敛自己的光芒”是奕欣一直以来坚持的学习和工作态度。在校园里，奕欣就很积极地为自己争取学术资源和校内外科研合作机会。进入微软亚洲研究院之后，奕欣努力学习自己之前不熟悉的语音信号处理基础知识，甚至每天睡前还会默默回顾自己没有弄懂的内容。与此同时，每周五奕欣会和 mentor 进行一对一会议，除了同步工作进展之外，她还会把自己的想法跟 mentor 讨论，每次都能写满一整个白板，mentor 也赞赏她“积极主动，有想法”。她还主动报名担任了研究院女性实习生活动 Ada Dialogue 的主持人，她希望让更多的女孩「被看见」，勇敢闪耀自己的光芒。

2021 年 8 月，东南大学人工智能专业的张淼森入选了微软亚洲研究院与东南大学的联合培养博士生项目。他说，这满足了他“进入有科研活力的研究院，靠实力去拼未来 ” 的心愿。他也正式开启了自己在研究院的实习。科研中的他专注且上进：每天早上 9 点到公司，吃完早餐开始写代码跑实验，下午会打印论文找个安静的咖啡厅阅读，这是淼森一天的工作日常。下班后，充满活力、热爱街舞是淼森的另一个标签。他从高三开始学习街舞，到现在已经 4 年了。

“科研”和”街舞”像是两个截然相反的世界，但是张淼森却在它们之间找到了共通之处：要想有好的研究成果，就要修炼好代码基础、数学基础，提升读论文、总结论文的能力；街舞中也是一样，不断练习动作、提升力量水平是很枯燥的过程，却也是决定舞台呈现的关键。另一方面，在科研中需要明确自己的研究长处，找到对应合适的研究课题，就像在街舞中需要把动作卡对节拍一样。

对于张淼森来说，他热爱科研中的静心专注，也享受在音乐中肆意舞动。在街舞的世界中，他希望自己沉浸在音乐的世界里找到个人风格；在科研的世界中，他希望自己能摒弃功利和浮躁的心态，日臻至善。

2021 年 8 月，多伦多大学计算机科学与工程专业的赵介宸来到微软亚洲研究院实习。从加拿大到北京，从学校的实验室到企业的研究院，一年的时间里，他在微软亚洲研究院不断蓄力成长。

一年前，博士一年级的介宸被微软亚洲研究院自由开放的科研氛围吸引，加入网络研究组。介宸在 mentor 舒然的指导下，从 0 到 1 探索了一个新颖而艰巨的项目：利用智能网卡 (SmartNIC) 为云中数据管理提供高效可移植的动态热迁移通用解决方案。由于相关研究较少，系统设计复杂性也很高，在开辟新的研究方案的过程中，mentor 耐心指导，提供积极反馈，并在项目暂时遇阻时让介宸重拾希望。在 mentor 舒然的指导下，介宸将设计按照重要性和复杂度分离，为从 0 到 1 的研究探索奠定了基调。经过近一年的研究，这个解决方案在微软云产品部门的 Demo 顺利完成并达到了预期效果，并且介宸完成了一篇顶级学术会议的投稿。在攻克了一个又一个难题之后，他也进一步提出了为智能网卡的硬件资源提供高效可重配置的虚拟化支持，并在 mentor 舒然和其他几位教授的指导下完成了另一篇顶级学术会议的投稿。

除了 mentor 给予的支持之外，介宸还经常主动在研究院寻找学习的机会。研究院每周举办的“Intern Tech Talk”他也积极参与，即使分享的内容不是自己的研究领域，他也乐于接触不一样的科研视角。一年的时间，介宸在微软亚洲研究院不断打磨专业技能，积累科研经验。与微软亚洲研究院的其他实习生一样，作为新生力量，他们借助研究院的平台不断向上生长，厚积薄发。

如果你也是微软亚洲研究院的一员，欢迎你向我们推荐身边有趣、有亮点、有故事的“明日之星”。 如果你还不是微软亚洲研究院的一员，欢迎扫描二维码查看实习生岗位介绍相关推文，向我们投递简历。

相关阅读

扫描二维码查看文章

实习派 | 钟宛君：从“七次拒稿”到“微软学者”，在科研挑战中成长

星跃重洋 | 刘国栋：非典型理工男在微软亚洲研究院的科研“旅”记

对话 | AI 与教育的深度融合，究竟什么是核心问题?

AI+ 教育是近年来教育行业乃至整个社会都非常关注的热点话题。相比于 AI 在其他领域的落地应用，AI+ 教育的进展一直相对缓慢。作为未来教育领域发展的大势所趋，AI+ 教育到底面临怎样艰难的挑战?

此前，微软亚洲研究院与华东师范大学就基于双方在各自领域的领先优势展开战略合作，希望依托计算机技术推进教育与人工智能的深度融合。目前，双方的合作已经形成了在学术界与产业界均具有引导性的创新研究成果，中文写作智能辅导系统“小花狮”即是代表成果之一。

“小花狮”合作项目负责人之一，来自华东师范大学上海智能教育研究院的郑蝉金副教授不久前作为访问学者来到微软亚洲研究院进行访问研究。我们特别邀请微软亚洲研究院首席开发经理夏炎与郑蝉金副教授就 AI 与教育的结合和落地等问题进行了一场深入探讨的对话。作为对 AI 领域有着深入了解的教育学专家，郑教授深度解析了 AI 落地教育领域所面临的困境、教育领域亟待解决的核心痛点、AI 在教育领域可发挥的作用，以及 AI+ 教育的未来发展方向等各界关注的话题。下面就让我们在与郑教授的精彩对话中一探究竟吧!

> **夏炎**：现在各行各业都在积极投身数字化转型，但我们知道教育是一个非常独特的领域。作为教育领域的专家，您认为在教育行业的数字化转型过程中有哪些核心痛点是亟待解决的?

郑蝉金：我认为，教育领域数字化转型的核心痛点至少有两个。第一，大规模个性化，也就是我们经常提到的因材施教。因材施教一直是我们教育界里的最高理念，但是如何实现大规模个性化是我们真正的痛点。第二个痛点是，AI 如何真正进入我们全人发展的过程。简而言之，就是 AI 如何影响我们的个体发展，包括认知发展和非认知发展（如社会情感等）。

AI 要怎样在这两大领域帮助人类去成长呢？我觉得第一个痛点对应的是影响面，就是如何利用 AI 技术低成本、快捷地实现大面积的个性化；而第二个痛点对应的是教育的诉求点。因为教育注重的就是人的全面发展，而 AI 主要就是在实质性教育目标方面进行辅助。我们常说教育包含“学、教、管、评”四个方面，而在此之前信息技术已经助力教育实现了“教”、“管”和“评”的部分功能，例如，信息技术已经被应用在教育管理层面，帮助教师和学校进行各种数据的采集。现阶段 AI 助力教育的核心关切点主要是在“学”这一方面。这也是教育领域的实质性最高目标。

> **夏炎**：没错，我也很认同您的观点。从我们研究院之前与不同行业伙伴的合作来看，AI 在物流、金融等领域的任务会比较清晰，涉及人员也相对较少，更多的是针对已有模块进行优化。但教育领域的任务存在着一定的不确定性，还会涉及到多人间的互动，任务会变得复杂很多。请问郑教授怎么看待这个问题？您觉得 AI 在教育场景的应用，相对于其他行业有什么不同?

郑蝉金：我总结了一下，AI+ 教育相对于其他行业主要有三大不同之处。第一个不同之处是教育行业有着长链条的利益相关方，不仅涉及到多方主体，还会牵扯到多个层面。

首先从管理层来看，最高代表就是国家，这在全球都是一样的，教育行业反映的是国家的意志。国家之下就是具体的管理部门，再到落地，以国内为例依次就是教育局、教研员和校长。单从管理层我们就能看到一条关系链条，而 AI 应用到教育场景时离不开这个链条的要求和管控。

此外，教育领域还有条重要的链条叫做使用者。虽然使用者链条可能不会像管理层的那么长，利益方之间的紧密程度也不太

一样，但该链条包含了教师和学生，以及隐形的使用者——家长。

除了管理层、使用者，我还想强调一个特殊的链条——微观技术指导，主要包括了领域专家、教研员、高级教师和一线教师等。具体来说，领域专家指的是高校中各类顶尖的教育学、心理学系教授和专家。教研员则是负责连接理论研究和一线实践的关键性人物，需要把领域专家提出的理念和教学实际情况相结合，并传达给高级教师。高级教师与一线教师之间则如同师傅和徒弟，高级教师会将有效合理的教学方法推广给各位一线教师们。

细观以上这三个链条可以发现，部分主体的角色是多元交织的，而且教学过程往往需要涉及多方主体。以监督为例，教育领域的监督既包括了国家的法规需求监督，又包含了微观技术指导（督导）。这些都充分展现了教育领域复杂的多方多层利益相关方。

微软亚洲研究院 AI 语言学习项目团队访问华东师范大学（右四：郑蝉金，右二：夏炎）

第二个不同之处就是教育领域缺乏大一统的理念指导。正如刚刚夏炎所说，其他领域的任务往往会对问题有一个相对清晰的界定，即 well-defined problem。但在教育心理学，我们很难对全人发展和培养给出一个清楚的问题界定。目前，在教育和心理学领域存在着形形色色、多种多样的理论思想，仅美国心理学会旗下就有 55 个子分支，美国教育研究协会下有 12 个子分支，可以想见由此叠加会衍生出多少种不同的视角、观点和理论去看待教书育人的问题。而这也使得 AI 应用到教育场景时面临着一个困境：到底要依照哪个理论去设计模型？因此，AI 在应用到教育领域时必须细致地应对每一个具体的教育场景去做独特的开发。尽管 "社会生物学" 奠基人 Edward Wilson 曾提出的知识大融通这个想法很激动人心，但是在教育领域是很难实现的。

第三个不同之处是学习是"反人性"的，需要人全力付出，学习的场景更加复杂。在教育心理学里有一个最近发展区的概念，它由 Lev Vygotsky（维果斯基）提出，指的是学生的学习发展有一个可能的发展水平，但学生需要突破现有水平，通过努力才能达到。那么如何将这个理论应用到 AI+ 教育的产品中呢？其实这就是我们常说的自适应学习系统。该领域最成功的案例之一就是自适应测评，即根据学生的能力为其推荐题目。这个测评系统不会为学生推荐百分百可以答对的题目，而是推荐学生百分之五十可能会答对的题目，以激励学生挖掘潜力。自适应测评之所以相对于其他的教育场景更成功一些是因为它是一个 well-defined problem。但现在 AI+ 教育却非常不同，面对的学习场景比考试场景更加复杂，也更难将最近发展区这些概念应用到 AI 系统中。

夏炎：郑教授总结的非常完善。我也想就这个问题补充一点从计算机领域角度出发的看法。计算机最擅长解决的就是确定性的问题，问题越明确，模型越好进行收敛。但教育和学习的过程都是相对开放的，因此很难去衡量一个学生学习的过程是不是有效。例如，在定义学习发展目标时，家长、老师和学校的想法和理解可能就很不一样，这就会导致计算机很难针对多方给出一个大家都满意的最优解。所以 AI 在教育领域现在面临的问题就相当于是一个多重限制下的优化问题。

> **夏炎**：另外，您刚刚提到 AI+ 教育与其他行业一样都离不开领域专家的支持，但 AI+ 教育涉及到多方多层的利益方，应用场景也会更复杂。那么，AI+ 教育与其他行业相比，所需要的专业知识会否更深入？您认为 AI 与教育结合要解决的最关键问题是什么呢？

郑蝉金：针对这个问题，我的想法是 AI+ 教育所涉及的专业知识未必更深入，但应该是更多面的，而这会造成实施的成本和难度大幅提升。为了破解这些难题，华东师范大学在上海市政府的支持下，特别是上海市教委的指导下，成立了上海智能教育研究院。站在教育本身的立场，我们强调以人为本——是"教育 +AI"而不是以往的"AI+ 教育"。成立一年多以来，我们上海智能教育研究院的不同团队在不同维度上都进行了很多的探索，取得了阶段性成就。我目前对教育 +AI 的认知，很多也是跟着大家一起学习、探索而获取的新知识。回顾这段时间的探索与经验，我认为 AI 与教育的结合目前急需解决的最关键问题是 AI 如何真正地融入教育生态系统。任何一款 AI+ 教育产品从设计开发到最后的落地应用，都需要有一个将技术和教育接轨的角色。在微软亚洲研究院和华东师范大学合作的中文写作智能辅导系统"小花狮"项目中，我就扮演着这个接轨的角色。作为扎根于教育学部教育心理学系的一名教育测量专家，我主要负责教育测量的有关技术，此外我们团队还有来自教师学院专门负责做作文测评的同事。目前我们主要有三个基础团队，之后我们还会考虑引入心理学或语言教育的专家。当然，一款好的 AI+ 教育产品不能只停留在设计开发上，还要考虑到在实际落地应用时教师们的需求以及如何降低他们使用时的学习成本。这也就是我刚才强调的如何融入教育生态系统。

以 "小花狮 " 为例，我们在设计开发环节时就咨询过多个领域专家。其中，大学教授主要为我们提供了理论和逻辑上的指导；高级教师为我们提供了实际应用上的反馈和建议；教研员则充当着桥梁，既兼顾理论又为我们反馈实际情况。这一条多方的长链条中任何一方的缺失都可能导致最后产品在上线和落地时面临质疑和困境。而在考虑到产品定位的前提下，如何将多方意见进行整合处理，是我们 AI+ 教育团队需要关切的核心问题。

中文写作智能辅导系统"小花狮"

近年来，许多 AI 公司想要进入教育领域，但又面临着诸多困难和挫折，发展得也较为缓慢，并不如我们预期的一样蓬勃。究其原因，就是教育领域所需关涉的多方多层相关方和复杂细致的应用场景使得 AI 如何真正融入教育生态系统成为一大难题。

> **夏炎**：那您觉得，我们接下来需要做些什么才能使 AI 真正融入教育生态系统呢？或者说，AI 落地到教育领域需要具备哪些必备条件？

郑蝉金：我认为，想让 AI 落地到教育领域首先需要让产品简单易用。因为产品最后的使用者是一线的教师们。我们只有将产品做得简便易于上手，才能降低教师们在使用时所需付出的学习成本，避免教师们产生隐形抵抗，从而可以更好地推广产品。

其次，我们要真正去解决实质性的堡垒，抓住和攻克刚需。无论是帮助学生提高学业成绩，还是在某个领域帮助学生提高学习能力，产品都需要至少抓住一个痛点和刚需。从教育的整体发展来看，我们应该将关注点放在学习过程和学习能力上。但相比于考试和教育管理，学习过程培养和学习能力提升的场景更加复杂。如何将 AI 结合到学习过程和学习能力这个层面，将是所有 AI+ 教育领域的企业和机构下一步要着重关注和思考的问题。

夏炎：从计算机领域来看，尤其是近年来我们微软亚洲研究院在实际设计、开发 AI 助力语言学习系列项目的过程中发现，算法中的部分成分是可以在教育场景中通用的。以"小花狮"为例，其部分算法就沿用了我们之前在英语写作上的一些框架。我们也由此构建了一个通用的底层框架，上层应用再根据具体的学科场景去搭载不同的专业领域知识。我认为，我们搭建的这类通用底层框架或许反馈给教育领域，也可以帮助教育领域去做出一些改变。

> **夏炎**：提到刚需，我经常听到大家在谈论一个教育理念叫 4C，请您为我们介绍一下什么是 4C？为什么培养 4C 在教育领域中如此重要？

郑蝉金：4C 本质上是二十一世纪教育与人才培养目标的一种表述，定义了在新世纪教育领域要为整个社会培养出什么样的人才。而教育目标的制定是由各大国际组织（如 OECD，经济合作与发展组织）和各个国家根据自身的教育情况所做出的决定。4C 主要指的是批判性思维（Critical Thinking）、创造与创新（Creativity and Innovation）、沟通（Communication）和协作（Collaboration）。需要指出的是，尽管很多国际组织与国家都给出了不同版本的表述，但上述提到的这些 4C 元素在这些方案中都有一定的体现。

教育领域如何达成这些核心素养培养目标是目前我们非常最关切的教育问题。 这一热潮已经体现在目前正在如火如荼开展的课程改革之中，其重要性不言自明。以前大家可能更侧重在认知层面，现在则既需要培养认知方面的能力（如批判性思维）又要培养非认知方面的能力（如沟通、协作等）。所以说，新世纪对人们提出了新的要求，那么教育能不能回应这些要求就成为了教育界需要应对新的挑战。同理，如何助力应对这一挑战自然成为教育 +AI 的主要战场。

> **夏炎**：那以我们合作的"小花狮"为例，您认为 AI 可以在培养核心素养和运用教育心理学方面做哪些辅助性工作呢？

郑蝉金：写作是个高级复杂的任务，它本身就蕴含了 4C 的这些元素。比如，写作中会有递进、转折这些逻辑关系，也就是我们说的批判性思维。此外，书面写作本身也是一种沟通能力。我们现在还会提倡高级写作，希望能够讲述一个可以吸引人阅读的故事，这就是创造性思维。而协作能力则可以在写作这个过程中去培养和训练。在很多的一线教学中，老师们会布置"漂流的作文本"这样的写作任务，去鼓励学生们合作完成写作。所以我们也希望通过"小花狮"里的各种设计来体现这些元素。我们或许暂时未必能做到面面俱到，但是有些基本的东西是一定要做到的，比如，批判性思维和沟通能力就是最基础的设计。

那么 AI 在这个过程中能够做到哪些辅助性工作呢？事实上，AI 应用于写作已经有很长的历史了，其中最成熟的首先还是自动评分。但是"小花狮" 的目标不是简单的打分，而是如何帮助学生去学习，去发展他们的写作能力。也就是说，我们要让 AI 去帮助学生评估他们写作在各个方面的优缺点，并通过最近发展区去提供有效的针对措施来帮助提高写作能力。通过以上描述，大家也可以了解到这个任务会比自动评分要难很多。自动评分其本质就是利用线性回归来排序，但是帮助学生发展写作能力并不能只靠统计模型去实现，必须要借助 AI 提供诊断性信息，提升学生遣词造句、谋篇布局等具体的写作能力。如果我们能再进一步融入协作能力，利用写作活动来实现其他素养的提升，特别是非认知素养的提升，那样将更接近真正意义上的助力全人发展的目标。

> **夏炎**：我们也与您一样期待未来能够设计、开发出真正意义上帮助实现全人发展的产品。您既是教育专家，又很了解计算机科学，可否与我们分享一下您在进行 AI+ 教育的跨界合作中的经验和建议？

郑蝉金：根据我的自身经历，我主要有以下两点建议：第一是绝

对的开放心态，第二是绝对的深度融合。

即便我是学教育的，我在做 AI+ 教育的过程中，还是需要不停地向一线的老师们学习。而只有团队中的每一个人都保持开放的心态，大家都愿意往前多走一步，那么我们才有可能把这个事情真的做好。当然，开放的心态也为第二点——深度融合，奠定了基础。没有深度融合的产品最后呈现出来的效果也是不理想的。

但这两点确实都很困难。第一点更加侧重于心态，第二点则强调了心态加上能力。由于我的个人经历跨越了很多领域，所以我对这两点的体会也更深刻。我本科是英语专业的，后来接受了心理学专业的训练，又刚好是负责最偏技术的方向，所以对计算机和统计学都有所了解，现在我的工作又是在华东师范大学的上海智能教育研究院与教育学部两个不同的二级单位之内。这些经历让我知道每个领域都如同大海一般非常广博，而我们唯一能做的就是保持谦逊的态度。

我们也很荣幸可以与微软亚洲研究院的这些科学家和工程师一起并肩攻克教育 +AI 的难题。虽然教育和计算机这两个领域跨度很大，但大家都非常谦虚、有耐心，并且非常真诚地交流在项目过程中发现的各种问题。很多微软亚洲研究院的同事们还学习了大量教育领域的知识，让我们的交流与合作更加顺畅。因此，我认为每一个参与跨领域合作的人都应该保持开放的心态，聆听和接受来自不同领域的人提供的反馈，这样最后呈现出来的效果才会是真正的深度融合。

夏炎：我十分认同郑老师说的"要保持开放的心态"。在这里也非常感谢与我们合作的各位华师大的老师，正是这些领域内专家们开放的心态给予了我们沟通和合作的机会。这种开放的心态一定是双向的，可以让我们更加积极地去解决问题。

感谢郑教授的分享！期待我们未来继续携手共进！

从 AI 基础设施谈起，展望 AI 产业成熟

2022 年 7 月，IDEA 研究院举办了"RE/IN 科技沙龙"第三期直播。IDEA 研究院认知计算与自然语言研究中心讲席科学家张家兴博士、微软亚洲研究院高级研究员、系统研究组负责人杨凡博士，以及 OneFlow 创始人袁进辉博士，从 AI 基础设施谈起，展望 AI 产业成熟。

以数据库产业的发展路径为例，人们为了不再做重复的文件系统开发，提出了关系型数据库。在关系型数据库成为标准产品后，又逐渐催生了分布式、云原生，并在其基础上构建出各式场景服务。

AI 领域也正经历着这一专业化分工。上有落地应用和方案层，中层做标准化算法研究，底层含通用硬件如 GPU 和 AI 芯片等。在深度学习出现后，随着 AI 框架、工具链的发展，AI 软件层面也正在出现标准化趋势。

通用计算从 CPU 发展到软件层面，走过了若干个十年，而深度学习出现后，其发展速度比通用计算快了有十倍以上。AI 技术栈分层中，起到支撑作用的是各类 AI 通用硬件以及开始标准化的软件。AI 基础设施经历了从以 Caffe 为代表的早期开发工具，到以 TensorFlow 为代表的、借鉴大数据设计理念的开发框架，再到以 PyTorch 为主流的、更加便利的框架发展。今天大模型横空出世，占据了各个领域的主流任务。

杨凡指出，随着通用学习模型越来越大，不可避免地会出现如何在分布式状态下去训练大模型的问题，带来了可用性和性能之间的矛盾。新的生态催生了新的框架需求，"比如 Oneflow 正在做的分布式，比如移动端和云端中间出现的新概念边缘计算。"他认为，"现在是一个 AI 的黄金时代，充满了机会和希望。"

AI 基础设施建设赋能普通人，开源则是实现 AI 技术普惠化的重要手段之一。"无论是模型开源还是框架开源，开源的首要目的都是把新技术用可复现、可检验的方式分享给公众，这是一件非常好的事情，它能让 AI 走进更多人的生活。"杨凡总结道。

扫描二维码查看回放

对话 | AI、机器学习在材料科学研究中能发挥哪些作用?

近年来，越来越多的实践证明，AI 是一项可以用于发现规律的关键技术，除了工程技术领域，AI 也为自然科学提供了新的科学发现工具。科学家们利用 AI 技术、基于大量高通量数据分析，不仅能加速实验进程，甚至还可以从数据中总结和发现尚未被人类知晓的科学规律。微软亚洲研究院很早就看到了这一趋势，并在过去几年中，陆续开展了 AI+ 生物学、AI+ 环境科学、AI+ 物理学等方向的研究。

2022 年 6 月，微软亚洲研究院邀请了中国科学院半导体研究所首席科学家、北京龙讯旷腾公司首席科学顾问汪林望教授，就"高性能大规模原子材料模拟的挑战与机遇"等话题进行了分享，并与微软亚洲研究院副院长、微软研究院科学智能中心亚洲团队负责人、微软杰出首席科学家刘铁岩博士展开了深入对话。汪林望教授在材料科学领域深耕近 30 年，对大规模电子结构计算、密度泛函理论（DFT）、第一性原理计算的研究有着深厚的经验。对话中，汪教授深度解析了当前材料领域研究技术的发展现状、面临的挑战、存在的问题，以及 AI 技术在材料科学中的应用方向和待解决的问题。希望这场与材料科学领域专家的精彩对话，可以为 AI 探索更多自然科学领域带来新的灵感。

刘铁岩：汪教授在材料学领域已深耕近三十年，包括对密度泛函理论 (DFT)、电子结构等问题都有深入的研究。先请您介绍一下当前材料科学领域的发展情况。

汪林望：在材料科学研究领域，理论计算方面多数使用的是量子力学（quantum mechanics），也就是第一性原理计算。如今，第一性原理计算已经渗透到了材料科学的方方面面，它与合成、表征并称为材料研究的三大基石。最近 20 年，中国在第一性原理计算领域的研究论文数量呈指数级增长。

尤其是，从头算材料模拟（ab initiomaterial simulations）在学术界被广泛使用，因为传统的做法需要大量的经验参数，而从头算不依赖于经验参数，只用电子、质子质量、原子的位置等少量数据作为输入，就可以计算出材料性质。这是一种强大的工具，因为它所依赖的薛定谔方程是非常基础的。 可以说，有了这个方程，所有材料科学的问题都可以得到解决。

刘铁岩：第一性原理计算主要被用于基础科学研究，由于算力的提高和算法的更迭，当前产业界也对这个工具产生了浓厚的兴趣，特别是我们注意到诸如锂电池、化工、制药公司纷纷开始加大对这一领域的投入。那么，第一性原理计算与产业结合的前景如何？还有哪些挑战?

汪林望：目前工业界中广泛使用的是 CAD (Computer Aided Design，计算机辅助设计）工业软件，主要依靠牛顿力学，使用连续介质方法，比如有限元方法求解偏微分方程等。这类软件通过一些参数来表征材料的性质，并以这些材料为基础去设计不同的工业器件；它们很大程度上解决了实验成本过高的问题，比如可应用在飞机制造，代替风洞实验等。但这种辅助设计方案并不设计材料本身，仅着眼于宏观层面。

然而，在接下来的二、三十年，除了关注材料宏观结构的设计，我们将更聚焦于新材料本身的研发设计，例如关注研发耐高温的航空材料、坚硬抗腐蚀的海洋材料、电池正负极材料等。而材料的性质决定于原子排布，这就需要借助 Quantum Computer Aided Design，形成一套新的方法，我们称之为 Q-CAD。

但要想真正在工业界应用这些工具还有很多问题需要克服。在学术界一个大的材料项目几乎有 20% 的工作花在计算上，但在工业界这个比例可能少于 1%。当前学术界做的是几百个原子量级的计算，但在工业界，比如模拟电池的正负极材料需要的则是百亿原子的量级，其中的时间尺度很大。如何弥补这些差距？未来就需要人工智能的帮助。

在这个过程中我们要考虑几点：首先是复杂性，比如材料研究求解催化过程常常会从几十种情况演变成几百种，需要高通量计算、遗传算法的支持；其次是准确性，密度泛函理论也存在一些局限，比如 d 电子做催化时计算结果不够准确；还有尺寸标度问题，当前我们只能计算几百或几千个原子，但我们更需要对上亿的原子进行计算；另外还有时间标度，一般的分子动力学也就做到皮秒或者纳秒，但是当我们真正要解决一个工艺问题时，则需要秒甚至小时级别的模拟。

目前，我所参与的 Q-CAD 解决方案是通过自主研发的 PWmat 算法与高性能计算 GPU 加速和人工智能机器学习力场三种技术相结合，力图解决工业界计算所需要的大体系、长时间尺度的问题。当下和未来，人工智能等新技术不失为弥补材料科学研究与工业应用差距的一个好方法。

> **刘铁岩**：当前很多科研人员开始聚焦机器学习和第一性原理计算的结合，例如利用各种机器学习模型去拟合 DFT 的势能面，您是怎么看待这个方向的？未来，人工智能可以在哪些方面帮助到材料科学研究？

汪林望：在材料研发上，有两个方向可以应用机器学习技术。一是在数据挖掘上，比如基于数据库，利用机器学习在海量数据中发现分子结构、属性之间的相互关系，并找出映射（mapping），这是传统材料科学研究最关心的问题。当前利用密度泛函理论、材料基因组、高通量计算都会产生海量数据，与以往实验所得的数据相比更规整，很适合计算机科学和机器学习使用。

二是机器学习可以用来开发经典力场，再用经典力场做分子动力学模拟。机器学习开发经典力场的思路是：当使用中小体系计算密度泛函理论时，会产生大量的数据，而在经典力场中只需要知道原子结构、映射出能量和原子受力，就可以利用机器学习不断迭代、反复学习映射，使之变得更好，这就是机器学习力场。从这个角度来看机器学习力场将会改变当前材料科学的研究现状。目前经典力场模拟分子动力学的思路是：从物理学角度出发设定简洁模型，比如给每个共价键一个能量表达式，然后再调整各影响变量的参数。这对于原子种类相对较少的经典力场较为有用，但材料领域涉及的大部分情况下没有力场，或已有的力场精度很低根本不能用。引入人工智能、机器学习技术，有可能改变这一现状。

汪林望教授（右）与刘铁岩博士（左）
在微软亚洲研究院分享活动现场合影

> **刘铁岩**：在自然科学领域，AI、机器学习开始发挥作用，从事 AI 的学者也希望将自然科学作为一个重要的应用课题去研究，微软近年来也在持续探索计算机科学与自然科学交叉研究的新范式，例如分子动力学模拟等等。在您看来，值得科技工作者投身研究的重大科学问题有哪些？

汪林望：在机器学习力场中还存在着很多待解决的问题。人工智能在材料领域的应用还处于初步探索阶段，我们无法确定它在这个领域能发挥多大的作用。比如，要如何估计一个系统的内在“有效维度”或复杂性，如果维度增加很多，机器学习是否还会有效？现在的神经网络越来越复杂，它的极限在哪里？是否能应用更深的神经网络，而又可以快捷的训练？这些都需要进一步验证。

再如，用于机器学习训练的参数空间（parameter space landscape）随着数据点（data points）是如何变化的？我们优化训练时，其实它是一个空间切分（partition space）问题，那么如果数据量变大，参数空间是怎么改变的？经验表示，应该是数据量越大运行越顺畅，但目前没有定量数据来证明这一点，还需要理论指导。还有，在拟合限制下，或更复杂的网络下，如何平衡网络的更大的表现功能和它的训练可行性？

构建更复杂的网络时，理论上它的能力也会越强，但网络参数空间也会越复杂，训练过程中更容易进入局部最小值（local minimum）。如何在没有过度超参数调整下加速训练？如何跳过局部最小值？如何选择采样点？这些都是值得研究的方向。另外，应该结合基于物理角度出发的模型与基于大数据的模型，而不只是偏于一边。这样的模型也许更有效。

> **刘铁岩**：确实我们需要对 AI 保持客观的态度，AI 不是万能的。如果期望 AI 在材料科学，或者更大范围的自然科学领域产生作用，我们一定要深入理解目标问题。比如汪教授之前提到的， 在小体系中学了某种映射，我们也希望这个映射能够应用在大体系中，但小体系到大体系的过渡，并不符合机器学习的独立同分布假设，需要全新的数学工具来分析它的泛化能力。这些问题都需要 AI 学者与自然科学家一起深入思考， 那么对于跨学科合作您有哪些建议？

汪林望：人工智能、机器学习等技术在材料科学领域的应用才刚刚起步，可以说还处于探索期，如何使用这些技术，我们目前还没有明确的方法，更多的还需要参考一些其它的行业案例，从中借鉴经验然后再不断试错。有人说，在所有方面都可以尝试新技术，但其实并非如此，在尝试前，我们要了解我们真正需要什么、不需要什么。材料科技领域中，在一些函数、经典力场的准确率上，在观察更复杂的化学反应、物理反应等问题上，我们可以尝试引入人工智能技术。事实上，任何一项新技术的结合所产生的结果都有正负两面。

不论是在材料科学领域，还是其它领域，借助人工智能技术都是一个大趋势，它转变了传统领域的观念，让我们意识到大数据、机器学习等技术是可以与材料学的高通量计算相结合的。跨领域的结合也得益于计算机的发展，在产生海量数据后，用类似统一的方法建立模型，这是一种新方法，也将带来新的成果。

以前的第一性原理计算可以让我们计算几百个原子，机器学习技术的引入可能让我们达到一个新高度，实现百万级别原子的模拟。虽然经典力场是材料科学的重要一支，但之前的应用很有限。在以后，它与从头算（ab initio）方法的鸿沟将会变得没有那么难以逾越，从头算与经典力场的结合也会更容易，从而可以解决更高级别的问题。

至于未来十年这些新技术能带来哪些改变？比如只需两天就能找到新的阳极材料？我认为不太可能。但不论到哪个阶段，都需要行业专家利用领域经验知识提炼问题、简化问题、找出关键问题，再进行百万原子模拟，并与工业界的工艺过程结合，我相信这可以推动相关工艺向前迈进一大步。

> **刘铁岩**：最后想问汪教授，当时为什么会选择进入材料领域，并且三十年如一日地坚持下来？另外，作为一位成功的学者，您对从事学术研究的年轻人是否有一些建议可以帮助他们更好地管理自己的职业生涯？

汪林望：对我个人而言，选择材料科学领域是个偶然。我在美国留学时，神经网络研究非常火热，那是真正生物学意义上的神经网络，我也花费了两年时间来学习和研究诸如大脑六层神经网络结构，以及人脑思维的模型等等。不过找博士导师的时候，我发现很多导师已经有博士生了，最后找到了比较务实的电子结构计算方向。那个时候第一性原理计算刚刚开始不久，所以我的学术生涯也正好见证了第一性原理计算的发展与起飞。在这同一时间段，计算机的运行速度增加了几万倍，它成为第一性原理计算背后的强有力的推动力。

我发现这两者的结合开发是一件非常有意思的事。当你在一个方向中深入研究时，你会发现有越来越多的东西值得探索。当计算机速度加快后，算法的相应改变和开发也变得十分重要。总的算力的提升有一半来自于计算机，另一半来自于算法的提升。这也能看出自然科学、材料科学与计算机发展结合是必然趋势。

总之，找到方向深入钻研，然后再慢慢地拓宽范围，这是成功的基石。重要的是，不要被外界浮躁的事物所打扰。

相关阅读

扫描二维码查看文章

科学智能（AI4Science）赋能科学发现的第五范式

未来十年，深度学习注定将会给自然科学带来变革性的影响。其结果具有潜在的深远意义，可能会极大地提高我们在差异巨大的空间和时间尺度上对自然现象进行建模和预测的能力。这种能力是否代表着科学发现新范式的曙光？

计算生物学揭秘奥密克戎强感染性原因

2019 年底开始的新冠疫情席卷全球，影响着我们每一个人。病毒的不断变异使得抗疫变得愈加艰辛和漫长。如何预测病毒的变异和疫情的发展，甚至从根本上遏制病毒的传播，这似乎已经超出了传统生物学的能力范围。

对此，微软亚洲研究院积极展开了一系列针对新冠病毒相关的研究，并取得了一定的阶段性成果。就在近期，微软亚洲研究院和清华大学生命科学学院及医学院在奥密克戎变异株强传染性的机理解释方面又有了新的突破，相关成果已被生物学领域的顶级期刊《Cell Research》接收。想了解奥密克戎为什么感染性这么强么？让我们在今天的文章中一探究竟。

对话 | AI+ 生物医药，如何双向赋能?

近年来，人工智能的深入发展助力生物医学研究取得了重大突破。"AI+ 生物医药"成为了学术界和产业界都非常关注的热门赛道。在后疫情时代，"AI+ 生物医药"能否保持强劲的发展态势，又将面临哪些机遇与挑战?

在世界人工智能大会 2022 的上海生物计算论坛上，微软杰出首席科学家、微软研究院科学智能中心亚洲区负责人、微软亚洲研究院副院长刘铁岩与上海市生物医药促进中心副主任唐军，华深智药创始人兼 CEO 彭健展开了一场精彩的圆桌论坛。三位拥有交叉背景及行业视野的对话嘉宾分别从研究现状、人才吸引、产业落地等多个角度对"AI+ 生物医药"进行了探讨，并展望了该领域未来的发展蓝图，为观众带来了一场赋有启发的讨论。

主持人：药物设计领域已经站在了新一轮爆发的起点，各种技术涌现，我们首先想请教一下几位嘉宾，对于接下来领域内大的技术发展有什么预测，是会呈现出一家独大还是百花齐放的态势?

彭健：我们可以看到包括生物计算在内的许多新技术在最近三到五年已经涌现出来了，我个人判断，未来一定是百花齐放的形式。

生物医药行业和传统互联网以及其他的产业还是有些区别的。生物医药行业是非常长的链条，制药发展的各个环节，从早期靶点到后面的发现，甚至到临床实验，每一个环节都是在整个产业链中非常有价值的部分。不管是从 AI 还是计算，甚至是从实验或其他的角度来讲，都必定在这个产业链的每一个环节中有很多不同的贡献。因此，我认为我们的行业在未来相当长的一段时间内，会呈现出百花齐放的情况。

刘铁岩：我非常同意彭健的说法。其实，药物设计相关的研究本身就是一个非常广谱且丰富的事情。从研究对象来讲，我们有很丰富的药物设计，比如小分子药物、抗体药、基因疗法、PROTAC 等，它们的原理与应用场景都有很大的差异；从制药的流程来看，从前端的靶点发现、先导化合物的筛选优化，到后期 ADMET 的预测、甚至临床效果的预测，各个环节都有各自独特的技术挑战。面对这样非常丰富的研究场景，本来就应该是百花齐放的状态。

但如果我们审视一下今天的 AI 制药领域，就会发现事实上里面还存在着一些问题。例如有一些扎堆的现象，卷到靶标蛋白的结构预测或者结合力预测（binding affinity prediction）这些问题上。之所以会出现扎堆的现象，其中一部分原因是这些领域已经有比较成熟的技术，比较容易获取那些唾手可得的成果。大家没有以一种长期主义的心态来思考如何构建自己的技术壁垒。刚才彭健提到我们微软研究院 2022 年成立了科学智能中心，我们这个中心的目的是以更加长远、更加基础的视角来看待人工智能在整个科学领域的应用，其中就包括 AI 制药，希望能够借由我们的努力引导大家以更长期的心态来看待这个领域的研究工作。

方向上应该百花齐放，不过下沉到技术层面，我们还是能看到一些趋势的。首先，AlphaFold 2 的成功让人们体会到深度学习、大数据、大模型、大计算带来的不同，而这种不同正是近年来人工智能领域发展的某种体现。比如，通过预训练大规模的基础模型来实现 AI 学习的规模效应，为丰富的下游任务提供有力的支撑，像微软投资的 GPT-3 等都是非常优秀的基础模型。我们相信这种趋势未来也会在生物医药领域进一步延展：比如，如何构建更适合小分子通用表示的基础模型，包括其骨干结构设计以及预训练的方法；如何有效地解决模型的泛化性和外推性，从而应对生物医药领域里有效样本不足和目标问题非常复杂之间的矛盾。

其次，强化学习技术在药物设计方面应该会有很大的发挥空间，因为药物设计本质上就是一个搜索问题，各种属性预测的深度学习模型扮演的就是价值函数的角色，而在这些价值函数的指导下，如何在巨大的分子空间中寻找到一个好的原子组合及其三维结构，是需要一些巧妙的策略做支撑的，蛮力搜索是不可取的。

目前在深度学习和强化学习这两个方面，人们还在大量使用着为传统领域发明的人工智能工具，针对制药领域进行的特异化设计还非常不足，所以我个人认为在生物医药领域人工智能要走得路还非常远。而这就需要我们计算机科学家和生物专家、化学制药的专家密切合作。做一个大胆的预测，我们有可能需要 5-10 年的时间才能真正形成比较稳定的技术路线，也可能再需要 5-10 年，我们才能够对制药行业产生本质的颠覆性的影响。

微软杰出首席科学家、微软研究院科学智能中心亚洲区负责人、微软亚洲研究院副院长刘铁岩

唐军：人工智能技术可以运用到整个药物从研发、中试到生产的所有关键技术环节，如人的免疫原性实验，在研发蛋白药物、抗体药物等过程中，作为临床前毒理试验的重要内容，必须完成。传统的药物开发过程是从分子水平到细胞水平，然后到动物实验，再到人体实验。我们不能等到了人体实验才考虑免疫原性，以前的做法是把人的免疫系统通过转基因技术放在小老鼠的模型上筛选药物的免疫原性，但准确率和效率较低，尤其在大量候选药物筛选的时候，工作量巨大，耗费的金钱成本也较多。

自从有了 AI 技术模拟抗原免疫原性筛选系统以后，工作量减少了很多，这样我们的一些判断就可以提前到侯选药物的筛选阶段，这个工作非常有意义，对医药行业也有非常大的支持。因为一个创新药物要开发出来往往需要 10 年的时间，还需要数亿美元的经费。假如能在最前面的环节解决问题，那么花费和时间都会有所节省。这给生物医药行业带来了颠覆性的改变。

另外，关于蛋白质结构的预测，我觉得也是非常有意义的。我们都知道氨基酸序列是肽或蛋白的一级结构，很快就能测出来的，但是它的二级 / 三级结构、空间结构、折叠却很难检测，或者目前的检测精度不够。假如利用了 AI 技术，那么在蛋白药物的设计和筛选方面会有很大的帮助。

上海市生物医药促进中心副主任唐军

下面我从成果转化和产业化的角度来谈一下，AI 和医药结合的产品是怎么从技术发展到新产品上市的，这期间大概会经历哪些比较难跨越的阶段。

首先，在实验室里发现一个技术，或者实验室发明了一个新的检测试剂或药物，要转化到工业化生产条件下进行生产制造，这就是一个难点。因为这里面要考虑质量的控制、成本的控制，以及中式放大产业技术条件的筛选，但是实验室里的科学家对这些是没有概念的，那么就需要工业界的专家进行指导。其次，新药从研发到上市最主要的一环就是需要大量的经费和时间投入，那么长久的、持续的股权融资就非常重要。第三，在新产品注册规划方面也会比较困难。因为药物和医疗器械的监管非常严格，细分领域的技术指导原则都非常细致，所以我们需要有药物注册专家，或者器械注册专家帮助提前规划。尤其是准备开发哪些种类新药、诊断试剂、或者做疫苗，所以一开始就要规划好。再者，最难的是人体实验，人体实验还需要临床资源、GCP 机构和医生大量的配合。临床实验完成以后，上市销售也比较困难，创新药要努力开拓市场，仿制药要想办法抢占原研药的市场，所以还需要销售专家、医保系统、定价系统和政府相关部门协调、合作，完成上市。

刘铁岩：唐主任刚才讲得非常好，向我们阐述了从技术研发到成果落地一系列的环节，也提到了鸿沟的问题，对此，我想稍微做点补充，尤其是从 AI 的角度来说一说我看到的鸿沟是什么样子的。

现在，很多从事 AI 制药的人都在走一条“捷径”，比如锚定已有的基准数据集或者一些公开的比赛（常见的如药物 - 靶标相互作用、药物 - 药物相互作用等），然后在这些任务上进行模型调优，以期获得 SOTA 结果。因为，一旦有了这样的结果就有机会发表论文、进行宣传，甚至获得资本的关注、实现研究的产业化。

但是这样的技术研发路径是正确的吗？是否存在问题呢？

首先，我们注意到这些基准数据集很难反映药物设计的全貌和丰富的场景，它们只覆盖了其中的部分环节。所以如何构建更加多样化、更加可信的数据集，其实非常重要。像蛋白质结构预测领域做得就很好，蛋白质结构一旦解析出来以后，大家都愿意放到 PDB 这样的公开数据库里。但是在医药领域，药物研发过程中的数据多是药厂私有的，是不拿出来公开的，这不利于学术界从事相关的研究工作。

其次，如果 AI 从业者针对有限已知的数据集，面对单一的评价指标只是不断调优，而不是思考模型在目标应用上是否存在着本质的设计缺陷，不去解决深层次的问题，那么我们很难保证学到的 AI 模型可以在新药研发领域有很好的表现。我们绝对不能天真地认为，只要手里拿着 AI 的大锤到处敲一敲就可以颠覆制药行业。想要实现 AI+ 医药就需要跨领域的专家合作，也需要 AI 从业者不断提高自己的修养，把相关领域知识消化吸收。

还有一个观点想跟大家分享：如果我们站在 AI 的角度来看待药物发现，其实药物发现并不是典型的人工智能问题。为什么这么讲？药物发现的目标是所找到的最好的药物要足够有效，而不是要求整个药物筛选流程里所有的候选药物在期望意义上都有效。这一点和我们经典的机器学习是非常不同的。另外，制药问题对 AI 模型的分辨率要求非常高，要细致到能够捕捉到关键蛋白的突变信息，而不是像多数已有的 AI 模型那样有很强的光滑性假设。

最后，刚才唐主任也提到了不管前期做什么样的预研，最后都需要严苛的临床实验过程。目前 AI 制药的研究主要集中在临床之前（preclinical），尚未打通整个药物研发的闭环。当然，一部分原因也是由于临床阶段数据更难获得，问题更加复杂， 更不可控，对已有的人工智能技术会造成非常大的挑战。因此很多用计算方法或者人工智能找到的候选药物，都折戟在了临床的阶段。

华深智药创始人兼 CEO 彭健

彭健：非常感谢刘铁岩老师和唐主任的分享！唐主任主要从产业的角度讲述了技术到底是怎么落地的，这里面有许多的困难需要克服，要促成一个成功的技术转化甚至落地，中间不仅仅需要 AI 和技术层面的提升，也包括政府和产业的联动，才能使真正的技术从早期的研发到最终的成果落地持续贯穿。刚才刘铁岩老师也给了我们很好的建议，特别是面向 AI+ 医药领域的研究者和创业者，怎么能够更好地利用 AI 去解决一些真正的现实问题。

我原来也是在学术界，去年回国创立了华深智药公司，我可以从创业者的身份跟大家分享一下我这一年来的感受。像唐主任说的，我们真正要去做一家公司，想要落地 AI 技术，这不仅仅是学术界要做的事情。前面谈到的免疫原性的例子，不管是做核酸药，还是蛋白药，它的流程都非常长，当我们把早期的工作做了以后，后面还有很多关于生产等各方面的评估。

2012 年以前基本上是以单抗为主，当时做抗体药物的时候，大家其实不考虑这些因素，因为那时的技术还不够发达，大家通常想的是只要能够找到结合的就可以了，后面有什么问题后面再去解决。如果大家去看 2013 年以后上市的药物就会发现，它们和过去的抗体药物有很大的不同——大家会把后面的生产、验证、临床逐渐早期化。这样，很早的时候我们就能够把分子找到，从而满足一些我们想要的性质，制药成功率就会大大提高。这一年我也看了许多的例子，也和很多的专家讨论过，大家现在认为成功率是最为重要的，一旦前面早期的决策做错了，后期的时间成本和资金成本都是不可估量的。当然，提高效率、提高精度是很重要的。但是很多时候我们需要从产业链条的角度思考这个问题，就是能够把重要的信息很早地注入在 AI 算法里，使得 AI 算法在设计做预测的时候就能起到重要的作用。

这一年我从产业界学到了很多，比如看问题更综合。在学术界讨论的问题有时候就只考虑成本、精度、计算速度等等，但后来我们逐渐意识到在药物研发的过程中有非常多的参数要同时考虑。而且从做产业的角度而言也是一个很复杂的过程，包括资本的运作、政策的扶持等等。

唐军：像彭教授这样的科学家，如果想来上海创业，目前是一个非常好的时机。上海关于产业高质量发展方面刚刚出台了一些新的政策。其中，对创新药物、创新器械、国外注册的药物器械等每个环节，都有相应的政策支持。此外，上海还提出了“1+5+X”的产业园区新政策。“1”是张江核心区，“5”是临港、奉贤、宝山、金山、闵行几个大的生物医药基地；“X”指的是很多细分领域的园区，像浦江镇的基因谷、张江细胞治疗产业基地等。很多细分的产业园区我们都制定了相应的产业政策，包括土地规划、资金支持、人才服务、子女教育等保障都做了相关的政策和规定。每个园区也都跟资本联合，搞了产业支持资金。在这里我也呼吁一下，如果一些科学家想创业，我想现在是最好的时机，政府也是大力支持的。特别是 AI 与生物医药的结合，对于生物医药我们专门做了关于数字化转型的细分支持政策。

生物医药和 AI 实际上都是属于比较前沿和比较尖端的交叉学科，都需要顶尖的人才，最后想请问两位科学家，在交叉学科方面有没有什么经验可以分享。

活动现场

刘铁岩：就像唐主任说的，AI 和生物现在可能是整个学术界、产业界发展最快的领域，它们的结合还会涉及到物理、化学、数学等其他的支撑学科。跨学科的交流和融合从来都不是一件容易的事情，甚至不同学科词汇的差异不亚于不同语言之间的差异。

我想从两方面讨论一下跨界合作，或者是跨领域研发这件事。

首先，我们要构建一支高效且多样化的团队。比如我们在微软研究院组建科学智能团队的时候，非常强调要招三类人：第一是顶级的人工智能科学家、第二是一流的自然科学家、第三是有丰富跨界合作经验的人才作为粘合剂。这里，我想着重强调一下顶级人工智能科学家和顶级自然科学家的重要性：没有一流的自然科学家，我们很难提出真正的一流问题；没有世界顶级的 AI 科学家，我们就没有能力和魄力去颠覆性地创造新的人工智能算法和工具，只靠拿来主义和魔改是没有办法构筑真正的技术壁垒的。

另外，如果两个不同学科的团队进行跨界合作时，那么双方都必须要有敬畏之心。AI 科学家和自然科学家不是生产者与消费者的关系、不是运动员与裁判员的关系，而是队友、合伙人，是一个团队。大家要携手共创，有充分的互信，而不是相互揣测、相互试探、甚至相互鄙视。这一点说起来容易做起来难，大家需要突破一定的固有思维模式的，要有成长型思维。在微软，我们非常重视成长型思维，鼓励不断突破自己的知识局限，乐于学习新知识，勇于踏入新领域，不断刷新自己的知识瓶颈。

未来 AI 制药一定是跨界融合、蓬勃发展的领域，也希望在这个过程中，大家能够不断地做探索，通过求同存异让不同背景的人能够在一起共同把这个领域发展好。

彭健：我最后从人才培养的角度来谈一谈。刚才刘铁岩老师也说了，我们有很好的自然科学家和人工智能的专家，也需要位于交叉点的人才，但同时接受两边训练的人才其实是比较稀缺的。他要能够同时理解 AI 技术，同时又对自然科学，像药物发现、生物学、化学有着非常深入的了解，这种人才非常少。之前我们各个高校的学科边界设立得过于明显，但很多世界顶级的学校都是鼓励学生选修其他学科的专业。我想，要从本质上解决交叉学科人才的问题，在教育、人才培养方面也需要更多的创新。

主持人：非常感谢三位嘉宾的分享，作为新兴的科研领域，生物计算需要跨学科、跨行业、跨产业部门的沟通与合作。就像刚才三位嘉宾分享的那样，这也代表了未来科技和产业发展的方向。

相关阅读

扫描二维码查看文章

你真的了解计算生物学和 AI for Science 吗?

近年来，计算生物学无疑是人工智能领域的一大热门话题。但计算生物学究竟是什么？目前进展如何？未来又蕴藏了怎样的机遇?

在量子位对撞派推出的"计算生物学"专题直播中，微软亚洲研究院副院长刘铁岩、首席研究员邵斌和主管研究员王童介绍了微软亚洲研究院计算生物学领域的最新研究，并对未来 AI for Science 的发展和融合进行了分享。

AI 挺进生命科学领域，分子动力学模拟加速新冠病毒致病机理研究进程

近年来大数据、AI 等技术的发展和应用，为生命科学研究开启了新范式。利用新技术，科学家们可以模拟瞬间变化的生命现象、发现生命机理的规律、降低研究新成本、获得更好的研究结果。

不同领域的科学家协同合作的秘籍是什么？如何在 AI for Science 的趋势中拔得头筹？让我们从微软亚洲研究院与清华大学的合作分享中一探究竟吧。

《未来媒体访谈》｜3D 视频系统，轻松与朋友在线“确认眼神”

远程办公的兴起，推动了在线会议系统的普及，什么样的在线会议能让会议场景更加沉浸、更具有交互性？

近日，微软亚洲研究院首席研究员童欣博士接受了新浪新闻、封面新闻联合推出的《未来媒体访谈》节目的采访。在访谈中，童欣博士介绍了微软亚洲研究院在 3D 视频会议系统方面的技术突破和相关技术的未来应用，并展望了 3D 视频会议系统将如何赋能工业界与现实生活，以及图形学的发展趋势。

以下为访谈实录：

主持人：大家好，这里是由新浪新闻、封面新闻共同推出的未来媒体访谈节目，细致入微的表情变化，自然的肌肤纹理没有一丝一毫的违和感。如果我不告诉您，您能看得出刚刚这几位参与者其实他们不在一个办公环境吗？这就是微软亚洲研究院的研究项目之一 ——3D 视频会议系统。今天我们也非常荣幸的邀请到了微软亚洲研究院首席研究员童欣博士，来给我们聊一聊在线会议的未来——3D 视频会议系统。童老师好！

童欣：主持人好。

主持人：刚刚我们从小片里比较粗略地了解到了，3D 视频会议系统它到底这个作用是什么，那么我们这里有一个很尖锐的问题了，在线视频会议其实已经不是一个新鲜的事物了，很多工作软件都带有在线视频会议的系统，那么我们想问的是微软的 3D 视频会议系统和刚刚我提到的这些有什么样的区别？

童欣：谢谢您，您问了一个特别好的问题，我想您看到的视频会议无处不在的事情，也在告诉我们，大家有很强的在远程与不同的人之间进行会议、进行沟通的需求。

我想大家看到目前的会议系统的时候，一方面它给大家提供了很多便利，但如果我们两个人或者多个人真正在同一个环境中开会，大家还是能看到一些区别的，比如最简单的，今天我们两个人坐在这里，我们可以有很自然的眼神交流对吧？我可以看到你很真实的所有身体的动作等等这些东西。

那么在多个人的交流环境中，大家如何切换话题，谁应该讲话，在一个自然的共同环境中，我们都很容易做到。但这在远程的会议系统中或者视频会议系统中，目前都是非常难以做到的，那我们做的这个 3D 视频会议系统，最终想达到的一个目标就是希望我们创建一个这样的计算机环境，让大家在开会的时候，感觉就像在同一个环境中开会一样自然，同时为了达到这样一个目标，我们也希望我们的设备足够简单，然后通过一套设备的设置能够实现不同的会议场景，比如像多人对谈的会议，或者是大家一起工作的时候，我们叫做双边的交互，就是一边看着眼前的屏幕一边交互的这样一个场景。

3D 视频会议系统 VirtualCube

主持人：您跟我解释了以后我大概就明白了，比如说电话会议是 1.0 版本，普通的在线视频会议是 2.0 版本，那么微软研究出来的 3D 视频会议系统就是 3.0 版本，如果我们达到了 3.0 版本的话，这个门槛是不是很高？

童欣：我觉得可以叫做一个 3.0 版本，但同时就是说要达到远程的非常逼真的体现这个人的所有外观动作这样一件事情，其实一直是计算机图形学和计算机视觉的一个挑战。

为了做到这件事情，我们有三个需求，第一个需求是我们需要高保真，因为我们人在日常生活中和人交互时、和人交流时，我们对人脸上所有细微的表情，他的动作什么是真什么是假，我们有非常严苛的标准在我们的意识里，这是第一件事情，所以我们要必须做到能够再现他所有细微的表情动作等等这些事情。

第二件事情，我们是一个实时的会议系统，所以所有的东西我们希望能够达到实时的需求，所有的东西必须实时地呈现给对方，对方的反馈实时呈现给我们，我们才能做很好的沟通，这是第二件事情。第三件事情，为了实现这个目标，我们也希望我们所有的设备和捕捉手段足够的便宜，足够的方便，那么可以说这三个需求要同时达到，一直是一件非常难的事情。

在过去几十年的图形学和视觉的研究中，大家研发了很多的技术，比如在影视业中，通过非实时的大量的技术手段，我们已经实现了可以说和真人没有差别的绘制。但是它没法实时。在游戏中我们可以做到实时，但是这个形象还达不到完全逼真。在视觉中我们有一些捕捉手段，通过一些其他方法，我们可以捕捉非常逼真的人，甚至做到实时，但是捕捉的整个设施是非常昂贵的，所以现在我们需要在这三个方面同时做了突破之后，才能实现现在的这样一个会议的成果。

主持人：我曾经在 2012 年看过一个报道，当时微软就说我们已经开始开发 3D 视频会议的系统了，那么现在是 2022 年，十年磨一剑，那么像您说的基于当时对于图形图像的研究，还没有办法实现这样的一个设想，那么所以 VirtualCube 是如何实现的?

童欣：就像您刚才讲到的一样，3D 视频会议系统实际上在视频会议系统刚刚开始的时候，不论是心理学家还是计算机视觉和计算机图形学研究人员，就一直以此为目标，微软也一直在这方面投入了很多的精力做研究，包括您看到的 2012 年的这个 Viewport 这个系统，还有我们后来做的 Holoportation 都是朝着这个目标前进的，那么到现在为止，我们为了做现在这个系统和已有的系统有什么样的突破呢？在我们的系统中，我们有两个关键的技术：V-Cube Assembly 和 V-Cube View 。

我们先来讲第一件事情，刚才讲到，我们希望每一个人在一个标准设置中，能够实现所有不同的会议场景，这里面有一个关键技术，就是我们需要把每一个人他所在的空间位置和一个虚拟环境的空间位置做很好的映射，有了这个映射之后，我们就可以把空间中不同地方的人通过拍摄的三维视频映射到一个共享的虚拟空间中，那么他们在虚拟空间中互相的位置关系和我们真实想模拟的物理位置关系是完全一致的。在这个情况下，我们通过不同的映射改变，就可以实现不同的会议场景，这是一个关键技术。

这个关键技术有了以后，为了我们实现不同会议者互相之间的沟通，我们就需要从不同的视角让每一个人看到的都非常逼真。这里我们需要一个叫 3D View 的技术。就是说我要显示这个视频，能够自由地切换我们的视点，从各个视点看起来都是非常逼真的。在这个方面要研发的技术，我们利用了传统的一些算法的基础思想，结合我们目前最先进的计算机视觉的技术，以及我们在深度学习方面的一些工作，最后实现了这样一个实时的算法，和已有的算法相比，在保证实时的前提下，该算法很大程度上提高了整个绘制的质量，实现了现在的这个效果。

主持人：在这 10 年计算机图形学这个领域，它还有哪些研究发展帮助了这一设想的实现呢?

童欣：在过去的几年中，我们把图形学的进展叫做智能图形学的发展就是说在传统中我们已经有了一些手段，这些手段通过一些软件，结合艺术家大量的手工工作，是可以产生高质量内容的。但在过去的几年中，图形学会结合硬件上的进展，比如深度摄像头这样的设备，以及已有的大量的高质量数据，和一些深度学习或者机器学习的算法一起工作，从而方便每一个普通的用户能够产生大量高质量的内容，并且是自动、低成本的产生。这些技术的发展或多或少都对我们 VirtualCube 所用到的技术都有所助益。

主持人：3D 在线视频会议系统除了让我们有一种在线的交流感，以及我们在场一对一交流的这种沉浸感，它还能应用在哪些方面?

童欣：我觉得交流这个事情或者会议这个事情，实际上是一个无所不在的场景，如果大家有兴趣的话，你可以用任何搜索引擎在互联网上去搜索会议的图片，大家会发现一个非常有趣的现象，就是你会找到各种各样的场景，远远超出你的想象，除了大家正襟危坐的在会议室的场景，两个人坐在屋子里一边喝咖啡一边聊天，它也是一种会议的场景。

所以可能对 VirtualCube 来讲，一个最重要的应用就是提供给大家一个泛在的或者无所不在的非常自然的互相远程沟通的场景，这是我们的一个目标。再往后面一步，为了达到这个目标，我们所研发的技术，比如我们的捕捉设备的技术，包括我们绘制的技术，我相信对其他的内容生产，如我们的视频产生、高质量逼真的内容，不论是用到影视中还是用在游戏中，我相信这个对他们都会有所助益，将来也都会推动这些技术和这些应用的发展。

主持人：我们通常说一个设备被广泛的应用，甚至普及的一个前提就是说成本的控制。那么我们刚刚讲到 3D 视频会议系统，给我们带来一对一的这种现场交流的沉浸感，达到这样的效果，是不是它的成本是很昂贵的?

童欣：成本我们可以从两方面说，一方面我们在设计 VirtualCube 系统中，很注意的一件事情，就是我们希望在达到效果的同时，探索可能性的同时，尽量地采用商用的硬件（off-the-shelf）。所有这些硬件不是定制的，是从市场上你就可以买到的。

所以在 VirtualCube 的系统中，在捕捉方面我们用了 6 个微软的深度摄像头，Azure Kinect 摄像头，然后同时我们在整个计算上，用了现在比较先进的 GPU 来做这件事情。

6 个 Azure Kinect RGBD 摄像头捕捉人像和眼神等动作

从另一方面讲，目前的所有这些设施，大家要马上用到每个人的普通环境中还是相对来说成本较高的，但是它的好处是所有这些东西都是可以量产的，那么随着硬件生产工艺的进步，这个普及，我相信这个成本会得到很大的下降，未来这条路通向每个人都能使用的程度是可以预见到的。

主持人：微软的创始人比尔盖茨先生曾经公开表示，因为疫情的发展加上现在通信设备的发展，我们有可能以后会改变工作的模式，也许有一天我们都可以到元宇宙里去开会了。我知道任何事物都有它的两面性，有它的优势就有它的劣势。那么我们 3D 视频会议系统有什么局限性，也可以说它的短板是什么？

童欣：你问了一个特别好的问题，也是一个尖锐性的问题。虽然虚拟办公环境或者远程办公变得流行或者变得更加重要，但是我们的理解，它并不是一个替代的关系，换句话说它并不会替代以前这种物理环境中大家的工作，因为在一个物理环境中，我们人的很多交流，是需要见到真实的人的，它的很多便利我觉得是无可替代的。所以到最后无论是 VirtualCube 也好，还是其他技术也好，都给大家提供了更多的可能性。还有一些环境中，我们认为最后会实现混合的办公环境，就是所有的技术手段，技术提供的所有可能性，大家会根据自己所在的情境，选择一个最有效的方法和别人做最有效的交流。

就像您刚才讲到的目前的 VirtualCube，我们专注的是提供一个高质量的、沉浸式的参与感很强的这样一个体验。但为了实现这样的一个体验，你对设备、你对这个环境可能就有一定的要求，如果一个人在车上，他要怎么实现一样的环境？特别是我们 VirtualCube 现在需要一个很大的屏幕，如果你只有一个手机，我们怎么努力可能都没法实现沉浸式的眼神交流这样一个体验。

这个是它的一个限制，但我觉得任何一个技术这样的限制可能都是存在的，最终的目标是说如何把这些技术融合在一起，提供给一个大家，我们叫做无差别的或者具有包容性的解决方案来实现最有效的沟通，我觉得这可能是我们最终的一个目标。

主持人：无论是 3D 的视频会议，还是这种各种跨界空间的交互办公，可以看出来微软一直在试图打破这种真实和虚拟的技术，再追求一个关键的元素，那就是沉浸感，我们不妨天马行空的想一想，除了办公方面的应用，还有哪些可以让这些智能媒体大显身手的地方？

童欣：其实我觉得办公是一个非常重要的事情，但是就像我们讲的，一个人的生活可以分成两部分，一部分是办公，一部分是普通的生活。比如，有两个老人，他们生活在两个城市中，由于各种各样的原因，他们没法互相去旅行了，那么我们也希望用这样的一个系统给他们提供一个沉浸式的、非常逼真的体验。我相信对他们个人生活质量的提高，幸福感的提高都是非常有用的。

那么同时这些技术的发展，大家可以看到在我们的日常的娱乐中，其他的媒体中包括新闻报道中。比如有一天也许真的可以用远程的方式你就可以采访我了，但可能我们没办法大家坐在一个屋子里，我相信对其他很多的应用，很多的我们的媒体也好，或者对生活也好，都能起到很大的作用。

主持人：我们上面讲到的这些 3D 视频会议系统都是在一个显示设备上呈现出来的，比如说大屏幕，未来计算机图形学能否结合虚拟现实的技术，将 3D 这个图像直接投射在我们真实的生活里，而不仅仅是屏幕上。

童欣：是的，这是跟显示技术的发展相关的。按照显示尺寸，我们可以分成两种，一种就像我们现在用的大屏幕这样的东西，它更多的是尺寸比较大，好处就是大家不需要戴任何的眼镜。还有另外一个就是增强现实技术（AR），那么微软也有产品，比如我们的 HoloLens 就是这样一个产品，它通过大家戴一个眼镜，就可以把影像呈现在大家眼前，它的好处是随着人的走动，这个影像可以跟着人去做各种移动。

物理屏幕的缺陷是你的位置比较固定，但是另外一方面你戴着眼镜的缺陷是不太方便，还有很多的限制。其实，即便是在大屏幕的呈现中，有投影的技术或者其他的技术来做这些事情，最后这些技术可能都会并存，融合在一起给大家提供一个无缝的虚拟和现实完全融合的场景或者体验。

就像现在新一代的年轻人，可能他们使用 iPad 这样的电子产品已经习以为常了，我的梦想是也许再过 10 年下一代人对他们来说不太区分什么是现实的，什么是虚拟的，从他们出生那一天起现实和虚拟就是很自然地结合在一起的，这是我们的一个愿景。

主持人：您刚刚提到的智能产品，我们就说现在手机已经是人所必备的一个智能的终端，未来能不能将上述我们提到的这些技术在手机上呈现，比如我想跟朋友分享一个我刚买的一个小物件，我给他拍一张照片发给他，他就能随意地拖拽、360 度的观看物件。

童欣：这方面的技术其实微软在过去有很多的研究，最近一段时间大家可以看到我们有一个叫做 NERF 的捕捉技术，进展非常快，目前已经有一些比较成熟的或者说比较好的应用或产品来帮大家做这些事情了。就像您讲的，通过捕捉一个 360 度的视频，我就可以在里面很自然地实现一些拖拽，看这个物体。

然而目前相关技术的发展还有一些限制，比如说我虽然能看到这个物体了，可是我不好操作这个物体，当我把这个物体放在我的家里的时候，我希望它的光照所有体现的效果跟真实在我家里完全一致，这些方面还有很多的技术有待于大家进一步提高，把它变得更鲁棒（robust）变得更通用。但是我相信这些技术很快就会成熟，大家很快就能把这些技术用到自己的实际生活中。

主持人：您认为智能媒体和对其起到支撑帮助作用的图形学未来的发展趋势是什么？

童欣：从我们的角度来看，我觉得未来图形学的发展，我把它总结为几个趋势，第一个趋势叫智能化。在过去二三十年的图形学发展中，如果和人类做一个类比的话，可以说我们终于实现了农业时代，什么意思？我们发明了锄头，发明了镰刀，艺术家通过学会怎么用锄头镰刀终于能把粮食种出来了，但是普通人你是种不出来的。那么我们认为智能提供了什么，通过人工智能的技术，我们可以说实现了机械化，让普通人也能利用智能技术通过简单的交互就能把他心中想的东西创作出来，包括您说的看到的东西能够数字化成一个三维模型放到计算机里，这个趋势我觉得是非常明显的。在未来几年中大家能看到很多技术的突破，甚至一些实用的应用产生，我们把它叫作智能化。

第二个趋势是综合化或者叫集成化。就是说你去看很多的东西，除了我们做游戏等等这样一个三维的形体，它其实不光光有它三维几何或者外观存在。我们在游戏中要和它交互，每个人这样交互，比如刚才您讲到说扫描了一个物体，我要各个角度看，但是对大家来讲，我买一个东西除了看，我们还有别的需求，比如我想摸一摸它的质感是怎么样的，我想操作一下。所以每一个物体除了它的几何外观，还有很多的属性，比如它的物理学属性，材质是什么样的，甚至我想知道它的温度是暖的还是冷的。

所以这些属性其实在图形学或者其他的学科中，目前都是被单独处理的，每个学科每个领域只负责其中一小块，如果想得到一个统一的计算表达，满足所有的需求，那就需要这些学科的人坐在一起。同时通过各种技术的集成，包括打通各个领域的东西，真正提供一个物体的全表达，就是既有它的几何属性、物理属性、材质属性等等，各种属性都有，那么就真正可以做到我们在它里面可以进行各种操作了。这个我把它叫做集成化或综合化。

人工智能技术会推动综合化的发展，因为大家可以看到深度学习技术提供了一种跨领域的方法论，一种统一的能力，那么最后一个我们可以叫做泛带化或者叫做平民化。以前，对于图形学技术大家觉得离我们非常远，只有专业人士拿到了，然后创造一些电影、游戏，我们只是消费者，从来不会去创作图形内容。我们希望，以后每一个无论是个人想创作他脑海里想象的东西，还是企业想用图形学的技术来模拟真实的世界做一些预测、规划的时候，这些图形学的技术能变成水和电一样的一种资源或者服务无所不在，每个人都可以经过简单的学习就能使用，能够在日常生活和工业应用中无所不在地起到它的作用。这是我们对未来的一个期望或者我们的一个愿景。

扫描二维码查看完整视频访谈回放

相关阅读

扫描二维码查看文章

3D 视频会议系统 VirtualCube：相隔万里也如近在咫尺般身临其境

常言道："眼睛是心灵的窗户"，眼神交流所传达的信息也可以进一步提升人们的沟通效果。然而，随着视频聊天、视频会议逐渐成为常态，大家不禁要问，我们有多久没有与同事、朋友、家人确认过眼神了？

而微软亚洲研究院的研究项目 3D 视频会议系统 VirtualCube，可以让在线会议的与会者建立自然的眼神交互，沉浸式的体验就像在同一个房间内面对面交流一样。该技术的相关论文被全球虚拟现实学术会议 IEEE Virtual Reality 2022 接收并获得了大会的最佳论文奖（Best Paper Award – Journal Papers Track）。

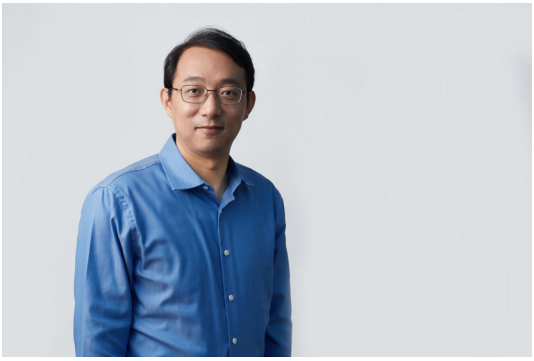
VirtualCube 相关链接：

论文链接：

https：//arxiv.org/abs/2112.06730

项目页面：

https://www.microsoft.com/en-us/research/project/virtualcube/



周礼栋
微软亚洲研究院院长

“二十多年来，微软亚洲研究院始终秉承开放、积极的心态，致力于打造自由、平等、可持续的科研协作环境，让分工、协调、合作链环上的每个人都成为新的发现与贡献的核心主体，为各种创造性想法的星星之火提供形成燎原之势的催化剂。

一个创新型组织的成长是不断拓展视野并承担更大社会责任的过程。微软亚洲研究院从创立伊始就持续与国内外计算机科研机构展开深度合作，携手进步，共同发展。在面对当下可持续发展、碳中和、医疗健康等人类社会亟待解决的关键问题时，微软亚洲研究院将守正创新，践行所有有利于激发创新力的原则，大胆接受和改造各种新的范式，与各界伙伴共同推动计算技术的跨界融合发展。”

关于微软亚洲研究院

微软亚洲研究院成立于 1998 年，在北京和上海拥有 300 多位科学家和工程师，是微软公司在亚太地区设立的、美国本土以外最大的研究机构。通过来自世界各地不同学科和背景的专家学者们的鼎力合作，微软亚洲研究院已经发展成为世界一流的计算机基础及应用研究机构，致力于推动整个计算机科学领域的前沿技术发展，将最新研究成果快速转化到微软的关键产品中，并且着眼于下一代革命性技术的研究，助力公司实现长远发展战略和对未来计算的美好构想。

作为微软研究院全球体系的一员，微软亚洲研究院拥有广阔的国际视野，同时扎根中国，辐射亚洲，通过融合东西方创新文化的精髓，以高度的社会责任感，持续开展有影响力、有温度、面向未来的基础科学研究和技术创新。微软亚洲研究院始终秉持相互信赖、相互尊重以及开放合作的理念，承诺与高校和科研机构开展持久而有效的合作，激发创新潜力、推进行业发展。

微软亚洲研究院倡导对技术进步怀有远大抱负，推崇富于冒险的极客创新精神，鼓励研究人员拓展研究的深度与广度，跨越计算机领域的界限，把视野拓展到解决具有广泛社会意义的问题上：提高人类的知识水平，推动基础研究的发展；增强人类的创造力和成就；培育有韧性、可持续的社会；支持健康的全球社会；确保技术值得信赖，让每个人都可以受益。

微软研究院全球布局

微信

知乎

电话：86-10-59178888
网址：http://www.msra.cn/
微博：http://t.sina.com.cn/msra