主办 CCF计算机视觉专业委员会

COMPUTER VISION NEWSLETTER

CCF 计算机视觉专委会简报

01 2024

总第 39 期

CCF 计算机视觉专委会

COMPUTER VISION NEWSLETTER

计算机视觉专委会简报

2024 年第 01 期

总第 39 期

CCF 计算机视觉专委会

主　办　CCF 计算机视觉专业委员会

编委会

CONTENTS
简报目录

CCF 计算机视觉专委会

CCFCV.CCF.ORG.CN

CCFCVN@GMail.com

第 19 期

CCF-CV 视界无限系列研讨会

2023 年 12 月 22 日，由中国计算机学会计算机视觉专委会主办、上海交通大学人工智能研究院承办的 CCF-CV 视界无限系列研讨会——“Vision for Science 前沿论坛”在上海圆满举办。会议邀请了专委会主任、北京大学信息科学技术学院教授、机器感知与智能教育部重点实验室主任查红彬教授，以及上海交通大学人工智能研究院常务副院长、人工智能教育部重点实验室主任杨小康教授致辞。

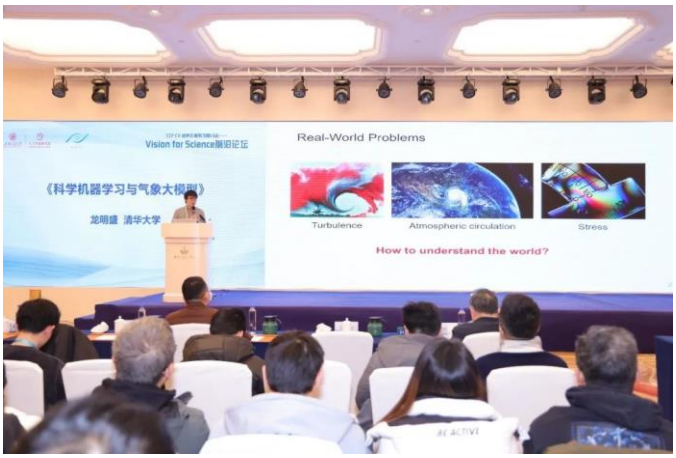
北京大学彭宇新教授、上海科技大学虞晶怡教授、北京师范大学邬霞教授、同济大学苗夺谦教授、清华大学龙明盛副教授、湖南大学刘敏教授、复旦大学姜育刚教授、华中科技大学白翔教授、浙江大学周晓巍研究员、上海交通大学王韫博助理教授、晏轶超助理教授做主题报告。复旦大学付彦伟研究员、同济大学史淼晶教授、德州大学奥斯汀分校黄其兴副教授、上海交通大学高岳副教授及以上多位讲者进行了深度讨论。会议由上海交通大学马超副教授等共同主持。

查红彬教授在致辞中介绍了 CCF-CV 视界无限系列活动的基本情况，对本次论坛关于视觉与科学交叉问题的讨论表示期待并预祝活动取得圆满成功。

杨小康教授在致辞中表示上海交大人工智能研究院与 CCF-CV 有着长期紧密的交流合作，希望通过本次活动，交流碰撞出更多、更新的思想火花，推动视觉与科学前沿交叉问题研究。

彭宇新教授分享了题为“细粒度视觉分析及其在生物医学中的应用”的报告,首先介绍了团队在细粒度图像分类方向取得的进展，特别是在鸟类识别、干细胞分类、智慧医疗以及肺部疾病诊断等领域做出的研究，不仅体现出细粒度图像分类在生物医学领域的重要应用价值，同时也为相关研究和临床诊断提供了强有力的技术支持；同时，围绕细粒度视觉生成方向，探讨了如何提高模型的泛化性以及如何提高扩散模型的效率等技术细节问题，为该领域的发展提供了宝贵的经验。此外，彭教授从细粒度视觉分析技术与大模型、知识图谱等技术结合的角度，分享了团队在海报生成、视频生成、自动驾驶、老人摔倒检测等方面的创新应用，展现了细粒度视觉分析在生物医学等领域的广阔前景。

邬霞教授分享了题为“类脑视觉智能：原理与方法”的报告，涵盖人脑视觉信息加工基本原理、类脑视觉智能方法和多模态识别的脑机制三个方面。她首先介绍了人脑视觉处理通路的腹侧通路和背侧通路，腹侧通路主要负责视觉物体的识别和决策，背侧通路主要负责感知物体的空间位置和运动信息，并在受到人脑视觉感知处理机制的启发下，团队探索出类脑视觉智能方法：视觉机制启发的强化学习和类脑强化学习。由于人脑在多模态整合方面有显著优势，通过模态整合、模态交互及神经同步等机制充分融合多个模态的信息，从而促进跨模态信息的交互与增强。邬教授团队受人脑神经元对不同模态差异化响应特征的启发，构建神经网络的多样性神经元，已实现较高精度的类脑多模态情绪识别，证明了神经元多样性启发的方法在多模态交互中起着重要作用。



龙明盛副教授分享了题为“科学机器学习与气象大模型”的报告，他将科学机器学习与气象模型等流体相关领域结合，构建了物理知识嵌入、观测数据驱动的科学机器学习气象大模型，解决了临近预报场景中多尺度多物理问题导致的预测结果难以把控的问题。团队将物理方程转化为适用于机器学习的网络算子，通过物理建模提升深度学习模型在大气演变规律上的知识。通过处理原始观测数据，摆脱对再分析数据 ERA5 的依赖，实现了模型的物理规律约束与数据驱动观测结果的解耦。这一方法在极端天气预报上的应用甚至取得了比人类专家更接近真实观测的结果。龙教授同时分享了目前气象大模型面临的挑战，如边界问题及极端区域拟合等，并表示未来将继续深耕这些问题，进一步推动科学机器学习在气象领域的应用与发展。

苗夺谦教授分享了题为“行人搜索算法研究进展”的报告。苗教授首先系统地介绍了行人搜索算法的发展历程及应用价值，以及行人搜索算法由两阶段算法向“端到端”算法的演进历程，详细地介绍了团队提出的

队序列化模型--SeqNet 模型，一方面通过网络结构的创新将回归任务与 Re-ID 任务由原来的并行改为串行，另一方面通过在人与人的相似度计算上的创新，使模型学习到更多的上下文信息，大幅度提升了现有端到端模型的性能。随后，苗教授介绍了基于二分图匹配的上下文信息建模（CBGM）算法，针对上下文信息不足的问题，将行人搜索抽象成二分图匹配问题，为行人搜索领域的后续探索奠定了基础。最后，苗教授从提高搜索精度、研究轻量化技术、采用粒计算思想三个方面进行了总结，并对行人搜索领域的未来研究方向进行了展望。

刘敏教授分享了题为“基于视觉感知的表面缺陷检测方法研究”的报告，他指出表面缺陷的快速精准检测是工业质检领域的一项重要任务，在高端装备制造、工业机械生产、精密电子加工等方面都有着重要意义。他表示，团队针对复杂场景下的工业缺陷检测难题开展研究，重点钻研基于视觉感知的复杂外形产品表面缺陷检测的理论方法及应用，并以航空发动机叶片表面缺陷检测项目为例，从表面缺陷数据获取、表面缺陷数据增广以及高精度缺陷检测定位三个方面，分享了相关的技术路线，以及团队有针对性的开发出的用于成像、数据增强以及缺陷识别的一系列关键核心技术，包括重叠区域感知的自适应成像视点轨迹生成技术、知识驱动的特征自适应视觉数据自动增广方法、基于模型轻量化及软硬件加速的高效目标识别技术等，未来还将可能进一步应用于高端装备成像与检测机器人的研发、工程机械目标识别定位系统的形成等。

虞晶怡教授以视频的方式分享了题为“Modeling Heterogenous Protein Structures: From Imaging to Generative AI”的报告。他首先系统地介绍了计算机图形学中三维重建的发展历程以及相应技术的优缺点，并强调深度学习方法所带来的巨大变革，与传统方法相比，在速度、清晰度和可拓展性等方面都有了质的飞跃。接着分析了蛋白质结构重建中存在的挑战以及现有方法的局限性，并详细介绍了团队在使用基于深度学习重建蛋白质结构中的进展，不仅速度更快、结果更清晰，还能进一步拓展到动态蛋白质的结构重建，为相关领域的研究提供了强有力的工具和技术支撑。虞教授表示未来将继续进行探索，结合现有技术以及新技术，如引入人类的先验知识、大模型的应用或使用定制化的芯片等，来加速蛋白质的三维重建等。

姜育刚教授分享了题为“视频内容智能分析”的报告。姜教授首先回顾了视频内容分析的主流技术路线，并指出该领域当下的发展趋势包括：训练数据逐步从单一拓展为多元、网络架构从分散逐渐趋于统一、训练模式从强监督发展为自监督。同时，模型的部署推理则会

从普适向专用发展，智能分析应该既能对不同场景进行动态调整的分析，同时也能实现计算资源的动态调整。最后，姜教授指出，该领域依然存在诸多问题，如需要进一步探索模型的可解释性、公平性、鲁棒性等。

白翔教授分享了题为“多模态大模型的细节描述能力提升方法”的报告，他指出要利用好大语言模型强大的理解和对话能力，利用好丰富的预训练知识来进一步为多模态大模型提供场景理解能力，并介绍了团队提出的 Monkey 多模态大模型，针对分辨率小导致细粒度信息抽取能力受限的问题，采取了复制微调方法的策略，实现了无需大量训练资源就能提升输入分辨率的目标；而针对场景细粒度对齐数据匮乏的问题，采取了多层级详细描述生成的方法，利用现有的多模态大模型、图像处理技术来产生高质量描述。通过与现有其他多模态大模型的对比，Monkey 在不需从头开始预训练的情况下，显著提升了细粒度图像信息的描述能力。最后，白教授从多模态大模型在未来会有更轻量化的设计、会更依赖于视觉基础模型、能有更好的跨模态协同学习能力和持续学习能力等方面进行了展望。

周晓巍教授分享了题为“无检测器的特征匹配方法及其在三维视觉中的应用”的报告。他指出重建和定位是计算机 3D 视觉领域最基础的问题，并从特征匹配的角度出发，回顾了特征匹配的发展历史，引出了传统特征匹配算法所面临的特征点检测可重复性问题，同时介绍了团队的 LoFTR 模型，创造性地提出了一种不需要检测器的特征匹配方法，通过先进行一个较为粗糙的稠密匹配，再对匹配的结果进一步改善，实现了高效准确

的特征匹配。此外，周教授团队进一步优化了 LoFTR 模型的效率，在保证性能的同时，加快了模型的推理速度，并进一步运用在多视图匹配中，实现了在多种复杂环境下的精确三维重建。周教授在分享中表示道无检测器的特征匹配方法在多种任务下具有非常大的潜力，未来将继续探索特征匹配任务研究，推动无检测器的特征匹配方法得到进一步应用。

王韫博助理教授分享了题为“物理世界的视觉直觉学习”的报告，介绍了团队在流体模型的视觉反向推演和物理世界解耦与强化学习这两方面的成果。团队创新性地提出“神经流体模型-NeuroFluid”，旨在将物理模拟建模为神经渲染的反问题。NeuroFluid 基于先进的生成式人工智能技术--流体粒子的神经辐射场，将物理启发的机器学习模型用于流体粒子模型的计算模拟与观测反演,通过流体表面多视图 3D 重构,推断流体内部动力学模型；并搭建了流体视觉直觉实验平台，以供真实流体场景下的图像数据采集和验证。另一方面，团队通过物理世界解耦助力强化学习，通过状态解耦可以

基于对自然状态的独立推演，生成更具有“预判能力”的策略，并且提升了世界模型在开放视觉环境中的泛化能力。王教授表示，视觉直觉学习通过视觉感知、环境交互理解现实世界中的难以用方程刻画的物理规律，有能力实现高速的物理模拟和反问题求解，同时有助于鲁棒高效的视觉环境决策。

晏轶超助理教授分享了题为“三维数字人的重建、编辑与驱动研究”的报告，介绍了团队在三维数字人领域的最新进展。他指出，利用机器学习来实现三维数字人的重建、驱动与编辑是未来发展的重要方向。在三维数据的重建任务中，团队提出了通过单图估计高质量、无偏差的人脸反照率重建方法，减轻了反照率的模糊性，并生成高保真反照率图以进行真实渲染。在三维数字人的编辑任务中，团队在三维数字人风格化领域创造性地提出了利用 hypernetwork 直接预测预训练模型参数偏移的方式，实现了文本引导的多风格、可叠加的风格化编辑，绕开了微调算法耗时的方式，解决了其模型受限于单一风格的问题。在三维数字人人脸编辑领域，团队则提出了通过直接操纵三维关键点的方法，实现了更符合用户习惯、直观高效的人脸编辑。晏教授表示，三维数字人的发展潜力巨大同时也面临着许多挑战，未来将致力于解决相关问题，实现高保真数字人的重建、驱动与编辑。

研讨会的最后是 Panel 与交流环节，与会专家与老师同学们进行了深入交流与探讨，最后论坛在大家热烈的思想碰撞中落下帷幕。

Panel 实录：

本次研讨会邀请了白翔教授、付彦伟研究员、史淼晶教授、黄其兴副教授、高岳副教授、晏轶超助理教授、王韫博助理教授进行圆桌对话。对话由晏轶超主持，围绕“随着 AI for Science 快速兴起，语言模型和视觉模型都在科学任务上展现出了巨大潜力，这两类模型各自更有可能在哪些领域实现科学突破？”、“在科学研究中，如何看待 CV 模型在鲁棒性、可控性、可解释性方面的问题？”、“CV 任务往往需要大量的训练数据，相关科学任务是否能提供足够的数据支撑？如何突破数据瓶颈？”、“很多传统的视觉任务都已经进入瓶颈期，CV 未来的研究重点是否会转向科学领域？”以及“对学生来说，想要从事 Vision for Science 的研究，需要做好哪几件事？对学校来说，如何为交叉学科的学生培养创造条件？”等五个议题开展了热烈讨论。

晏轶超（主持人）：现在火热发展的语言大模型、视觉大模型以及多模态大模型，您认为他们在哪些科学领域更有可能取得突破？我们先请白翔老师谈一谈。

白翔：科学问题的广度和复杂性使得我难以提供一个简单的答案。我们需要学科交叉并同时在多个领域中寻找问题的解决方案。要取得真正的突破，需要与专业领域的重大需求密切结合，学者之间需要坐在一起，相互学习，共同探讨问题。比如古希腊文铭文的恢复涉及到历史学家和人工智能科学家的合作，因为除了视觉处理的问题，还需要考虑古籍传承的规则、习惯和历史背景。同样，甲骨文的破译也需要语言学专家和历史学家等多个领域的专业知识，才能设计出相关的算法。年轻学者可能更关注一些热门的问题和新的模型，但更为重要的是要考虑人工智能是否真的解决了实际的问题。在解决问题的过程中，学者可以更加长远和开阔地看待问题，从交叉学科的角度出发，将精力放在深入研究问题本身上，而不是过度关注模型的调优。

晏轶超（主持人）：白老师给我们提出了殷切的期望，希望年轻学者能够真正做一些原创性的科研工作，感谢白老师。付老师现在也做比较多医学相关的工作，能从您的角度谈谈您的观点么？

付彦伟：白老师已经提到了一些很好的观点，我想进一步补充一下。最近，我们也在尝试零样本甲骨文的研究，采用零样本深度学习的方法来解码甲骨文。我们已经进行过一次投稿，但遇到了一些问题。跨学科进行这类研究确实颇具挑战，对于这个问题而言，我们发现语言模型和视觉模型在各个领域都能取得科学突破，几乎找不到哪个领域用不上这些技术。目前，甚至在数学领域也可以应用这些模型，只是问题的规模和是否为国家紧急需求等因素在影响着研究的方向。最近我们涉足一些有趣的研究方向，比如尝试通过解码信号，利用视觉和大型模型来还原人类所能看到和思考的内容。同时，我们也在应用视觉技术解决一些医学问题。

晏轶超（主持人）：白老师和付老师对话很有启发性，还有其他老师想对这个问题做补充吗？

黄其兴：这个东西怎么用，有两种情况，一种是本质上是视觉和语言的问题。我认为这是一个相对狭窄的领域，更为重要的是，如果有一个在视觉上成功应用的模型架构，然后将其应用于其他领域，通常情况下，这些视觉领域的架构在语言领域的架构上具有更广泛的适用性。特别是在当前自然语言模型发展迅猛的情况下，我们需要深入了解为什么它们能够取得如此快的发展，并且在泛化性方面表现得相当不错，这可能会带来更广泛的应用。

晏轶超（主持人）：谢谢黄老师。我们进入下一个话题，在科学研究当中怎么样看待 CV 模型鲁棒性、可控性和可解释性的问题，因为 CV 模型存在这些问题，但是科学研究需要一个很严谨的结论，或者说过程，不能够当做一个黑箱来处理，我们怎么样来看这个问题，还是请白老师先来谈一谈。

白翔：CV 模型在鲁棒性和可解释性方面的问题确实非常重要，但实际上，有影响力的工作相对较少，这在全球范围内都是一个普遍现象。然而，这个问题对于未来视觉应用的落地具有深远的影响。在很多应用场景中，如果没有解释性，存在巨大的风险，因为无法追溯或还原错误的过程。这实际上会导致视觉应用存在着很大的局限性。如果缺乏可解释性和可控性，基本上什么都无法实现。尽管大家通过可视化、数学模型和各种物理模型以及对神经元的思考来解释深度学习的黑盒问题，但目前来看，尚未实现可控性和鲁棒性。因此，我的观点是，这个问题确实非常重要，但也相当困难。如果要处理这类问题，我们也可以采用一种不必涉及通用模型的研究方法，比如专注于某一个场景，在这个场景中减少异常情况，或者至少降低风险,这也是一种可接受的研究方法和思路。

晏轶超（主持人）：感谢白老师带来的精彩观点，那高老师从您研究的角度来分享一下。

高岳：我提一个不一样的角度，对于可解释性的诉求，或许本身就存在一些问题。人类在很多情况下都有不同的看法和答案，比如对于一个简单的物体的定义，不同人可能给出不同的回答。因此，期望计算机有一致的答案可能是一个值得反思的问题。这或许反映了人类对许多事物的不信任感，体现了人类对解释性的一种心理诉求。另一个问题是，是否所有模型在采用深度学习或大型模型的情况下都无法实现可解释性。这种观点似乎忽略了一种平衡，不能期望马儿跑得快同时却不提供足够的养料，这只是我的观点。

晏轶超（主持人）：高老师的观点很有意思，能请史老师谈谈对这个问题的看法吗？

史淼晶：白老师和高老师刚才的讲解都很深刻，特别是在可解释性方面，随着深度学习和大模型的发展，这一问题变得越来越复杂。在研究可解释性时，我们面临着数据可解释性和网络结构可解释性两个方面的挑战，这在大规模数据和网络的情况下变得更加困难。我们的团队在尝试解决这个问题时发现，在处理自然图像时，公平性问题尤为突出。我们努力通过调整算法来保证在不同人种中不出现过度差异的表现，以确保公正性。然而，在医学领域，我们面临着不同的挑战。在医学应用中，模型在特定种群上表现良好是至关重要的，因为医学是一门应用科学，对精度要求极高。这也引发了对于公平性的不同看法，尤其是在特定领域。如何平衡模型在不同种群中的表现，以及如何解释模型的决策，是一个需要深入思考的问题。正如高老师刚才所说，从不同角度出发，我们对同一问题可能会有不同的看法。

晏轶超（主持人）：谢谢史老师，还有其他老师对这个问题分享观点吗？

王韫博：可解释性问题可以从两个方面来看。首先，神经网络模型，包括语言模型和计算机视觉模型，一直是一个具有挑战性的长期话题。尤其在科学研究中，解释神经网络模型的决策是一个复杂而困难的任务。另一方面，科学现象本身的可解释性也是一个更高层次的挑战。在过去的研究中，例如在气象预报系统的实践中，虽然模型取得了良好的效果，但科学领域的专家经常要求对模型结果进行解释。这挑战在追求科学的可解释性时变得尤为明显。科学研究者希望了解模型为什么产生特定的结果，而不仅仅是依赖于数据驱动。解决这一问题的方法之一是在完全数据驱动的基础上保留实体网络,采用符号回归等方法，或使用符号回归来挖掘潜在的方程，可以提供一些可解释性。通过这种方式，模型的决策过程可以更容易地被专业人士理解和接受，即使这些解释可能不完全准确。在科研领域，不仅要关注结果的优越性，还要考虑模型背后的机制，以便获得真正的科学认可。

晏轶超（主持人）：我们进入下一个问题，就是数据的问题，刚才史老师也提到了，现在采数据有很多问题，那么多少数据才真正的够用，特别是在 3D vision，或者说对于流体来说，都可能会存在这种问题，黄老师您是做 3D 重建资深的研究员，对 3D 视觉大模型，您认为在数据层面我们怎么样来看待这个问题？

黄其兴：这个问题实际上很有深度。首先，我看到过一种观点，认为我们目前训练的大型模型，特别是自然语言模型，所处理的语言量基本上相当于一个正常人需要几千年甚至几万年的阅读量。然而，这样的模型在训练后可能只达到了一个两三岁孩子的水平。对此，我认为发展大模型是必然的，但未来的关键在于如何改进架构。目前语言模型的训练架构是非符号化的，即依赖于大规模的数据，而进一步的发展需要在这一架构上取得突破。对于 3D 方面，考虑到我自己从事 3D 领域，过去十年中，大家在这个领域主要关注三维数据的表示，因为只有解决了这个问题，才能进行机器学习。然而，训练数据一直是一个瓶颈，而不同的表示方式会对结果产生本质的影响。我认为解决序列数据的问题时，不能仅仅关注单一模态。应该将所有的模态结合在一起，共同解决数据的问题，甚至在一个模态内部存在不同的领域时，也不能仅仅使用该领域的信息。

晏轶超（主持人）：我认为刚才黄老师意思是说我们不仅仅只需要关注一个模态的数据，我们要从全局的角度出发来看待这件事情，我们有不同的模态，不同的领域知识，我们要关注他们之间的关系，对此我非常认同，也请高老师分享观点。

高岳：我也提一个问题，因为我自己是做机器人的，大家可能也知道还有一个方向很火，叫具身智能，大模型现在的这套数据采集方法，一般我们使用的还是第三视角，但是具身智能是第一视角，是一种不同的数据采集视角。之前的数据采集是在没有考虑机器人的坐标系下进行的。然而，如果我们假设所有的数据都来自于人形机器人，整体统一到一个坐标系上采集，那么在数据采集时就会自动为每个数据打上时间戳，并将坐标系对齐。这样一来，很多任务的对齐工作就变得更加容易。有没有可能，通过具身智能的这种新的数据采集方式可

以有效的降低大模型所需要的多模态的对接，有效降低整个数据的需求量，听听各位怎么说。

晏轶超（主持人）：付老师的研究方向也跟这些方面相关，请谈谈您的看法。

付彦伟：在 CV 领域，embodied 这个概念在 2017 年 CVPR 引起了广泛关注和讨论。传统的 CV pipeline 通常涉及数据采集、模型训练和应用。而具身智能的思想则是利用各种环境、通过获取多样化的数据来训练模型，使算法更加容易应用。在 2017 年以后，相关研究逐渐增多，近年来更加热门。在 CV 领域本身，已有许多与具身智能相关的工作。举例来说，在 3D 领域存在多种环境，通过在这些环境中训练模型，可以预测出在 CV 中有用的结果。这类研究结果以前可能并不引起太多注意，但实际上在 CV 领域有着很好的预测效果。将这种方法结合到机器人领域是可行的，通过大量数据集训练机械臂等部件的模型。反过来，如果在实践中遇到困难，可以通过增加真实数据进行摸索和改进。总体而言，这是一个既有理论深度又有实践可行性的问题。

晏轶超（主持人）：谢谢付老师，我们进入下一个问题。现在很多 CV 任务的精度已经非常高了，比如说传统的检测识别跟踪分割，所以现在很多人开始研究一些新的方向，包括 AI for science 也是其中的一个，各位专家你们觉得是否会有更多的 CV 或者 AI 的研究者转向 science 方向，或者说其他方向的研究，而传统的 CV 是不是会逐渐消失？这当然是一个开放性的问题，史老师您怎么看？

史淼晶：在实际操作中，尤其是随着大模型的引入和对算力要求的增加，我们在第一线不断寻找出路。在处理这个问题时，可以从不同角度着手，工业界可以从大模型的训练上游着手，而我们也可以对下游任务进行一些有趣的尝试。在图像领域，我们发现有许多细分领域，例如不同的图像模态，包括 X 光、滤镜以及各种形式的影像传达方式，每个分支都有很多可以探索的领域。尽管目前大家普遍都在将 AI 应用于科学研究，我们也在考虑朝着科学方向发展，但在研究中找到自己感兴趣的点，专注投入其中也是可以的。回顾神经网络在起初并不火爆的时候，一些研究者在领域内耕耘十几年，甚至在神经网络还未广泛应用的时候就开始涉足，最终成为备受瞩目的人物。

晏轶超（主持人）：黄老师有什么好的建议吗？

黄其兴：一个领域的发展是持续不断的。以传统视觉任务为例，如分类、分割、检测以及分割，这些任务一旦有新的模型出现，研究者们就会开始针对这些问题展开研究。尽管一些问题尚未完全解决，比如检测问题，最近几年仍然有高质量的文章涌现，表明这些问题仍然备受关注。这并不意味着领域已经达到了瓶颈，发表文章变得更加困难，而是大家仍然在不断取得进展。此外，一个学科要不断发展，它的范围会逐渐扩大。一开始可能是一些核心问题，但随着时间的推移，像 CVPR 这样的会议变得非常包容。例如，投稿医学领域的文章也是可以的，只要与计算机视觉相关。在学科发展的过程中，一些传统方法可以在其他领域应用，比如三维视觉、视频、机器人学、自然语言处理等新兴领域。尽管最传统的问题仍然是分类和检测等，但学科的发展使得研究者可以在计算机视觉的范畴内研究更广泛的主题，这是一个学科前进的自然趋势。

责任编委　潘金山　杨巨峰

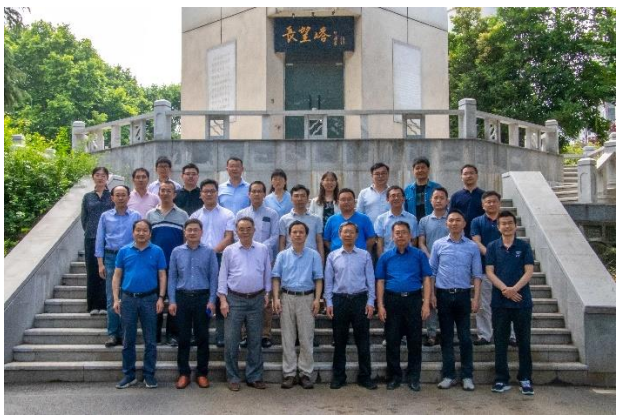
CCF-CV 换届 第四届专委会正式上任

中国计算机学会计算机视觉专委会（CCF-CV）是直属于中国计算机学会的计算机视觉领域的专业分支机构，也是国内唯一以“计算机视觉”命名的专业委员会。2024 年 1 月 28 日，中国计算机学会完成了换届仪式，这标志着第三届专委会任期已满，第四届专委会正式上任。接下来，将简要回顾第三届专委会的精彩瞬间，并介绍第四届专委会的组织结构！

第三届精彩回顾（2020 年-2023 年）

2019 年 11 月，CCF-CV 全体委员工作会议在西安举行，会上选举了第三届领导机构，于 2020 年正式上任。

2020 年至 2023 年期间，专委会持续开展系列活动并做好常规工作，共开展走进高校系列报告会 45 期，累计报告近 200 场，举办线上活动 10+场，B 站人气峰值最高 2.3 万，部分活动被科学网等媒体报道。2022 年 5 月，在南京举办走进高校第 100 期活动，历届活动主席、特邀讲者齐聚一堂。



专委会围绕计算机视觉领域的热点主题，举办“视界无限”系列研讨会 14 期，累计报告 80+场，共走进国内外 10+座城市。专委会还举办走进企业系列交流会 8 期，疫情期间举办 3 次线上活动，参加人员峰值 5000+。

2020 年开始，专委会每年举办一次 RACV 前沿进展研讨会，每次会议实录文章阅读量 2 万+。学会秘书长唐卫清研究员全程参加 RACV2022，并在 CCF 通讯 2022 年第 10 期报道。专委会还每年举办一次 PRCV 中国模式识别与计算机视觉大会，每次线下参会人数 2000+，2022 年入选中国科协《重要学术会议指南（2022）》以及 CCF-C 类国际学术会议。

在这些活动的基础上，专委会守正创新，在 2021 年开始组织“计算机视觉前沿讲习班”，旨在促进计算机视觉领域的学术交流与青年人才培养。截至目前，共举办 2 届，吸引 300+学员报名参加。

2023 年 11 月，专委会在南京举办 CCF-CV 十周年纪念活动，旨在回顾专委会成立十年的发展历程，分享取得的重要成果，并探讨未来的发展方向。各级领导、委员代表等约 100 人参会，学会秘书长唐卫清研究员全程参与。活动期间，正式发布了《计算机视觉十讲》，是 CCF 首批发布的 8 本十讲系列丛书之一。

专委会宣传平台稳固建设，专委会公众号共发布文章 200+篇，总阅读量 27 万+次，关注总人数达到 1.5 万人。B 站账号共发布 63 个视频，总计 4 万+次播放量，关注总人数达到 8000+人。网站平均每年新增独立访客 2.6 万+个，新增浏览量 3.8 万+次。出版专委会简报 16 期，并增设“视界专访”板块。

截至 2023 年，委员们累计获得国内外各类科技奖励和荣誉称号 100+项，其中谭铁牛院士荣获国际模式识别学会 KING-SUN FU 奖，王蕴红教授荣获国际模式识别学会 MARIA PETROU 奖。2020 年，专委会与计算机视觉领域华人学者一起合作，组织申办 ICCV2025。

在疫情期间，专委会委员们积极助力抗击新冠疫情，获得多项奖励荣誉。2022 年 10 月，由委员们牵头制定的国家标准 GB/T 41864-2022 《信息技术　计算机视觉　术语》正式发布。

在计算机学会年度评估中，专委会获得 2020 年的“优秀专委奖”、2022 年的“特色活动奖”以及 2023 年的”优秀专委奖“。

第三届专委会领导机构：

主任：**查红彬**

副主任：**刘青山**

副主任：**王亮**

副主任：**虞晶怡**

秘书长：马占宇

常务委员：白翔、程明明、耿新、纪荣嵘、赖剑煌、林宙辰、卢湖川、鲁继文、山世光、唐金辉、王井东、王涛、毋立芳

第四届扬帆起航（2024 年-2027 年）

2023 年 10 月，CCF-CV 全体委员工作会议在厦门举行，会上选举了第四届领导机构，目前已经正式上任。

接下来，专委会将继续开展丰富多彩的学术活动，用心服务各位委员以及计算机视觉相关领域的专家学者。也希望更多志同道合的朋友加入专委会，为进一步推动我国计算机视觉事业的发展而努力！

第四届专委会领导机构：

主任：**陈熙霖**

副主任：**刘青山**

副主任：**王亮**

秘书长：王瑞平

常务委员：白翔、程明明、耿新、纪荣嵘、赖剑煌、林宙辰、卢湖川、鲁继文、马占宇、施柏鑫、山世光、王井东、王涛、毋立芳、杨健、虞晶怡

责任编委 黄岩

CCF-CV 常务委员会 2024 年度第一次工作会议顺利召开

2024 年 2 月 3 日于北京召开中国计算机学会计算机视觉专委会 (CCF-CV) 常务委员会 2024 年度第一次工作会议，本次常委会工作会议是第四届专委会正式上任后的第一次会议，由专委会主任陈熙霖研究员主持，常务委员会委员参会，秘书处成员列席。

首先，专委会党的工作小组组长陈熙霖研究员带领大家进行了理论学习，学习的主题是《习近平在“国家工程师奖”首次评选表彰之际作出重要指示》。

随后，专委会秘书长王瑞平研究员针对专委会常规工作、特色品牌活动、秘书处后续工作规划等向常委会作了汇报。

接下来，常委会委员们围绕专委会执行委员发展现状、特色活动创新思路、委员参与专委建设等议题展开了热烈讨论，形成了具体可行的指导性建议。

2024 年度 CCF-CV 秘书处第一次工作会议召开

之后，陈熙霖主任针对新一届专委会发展阐述了工作设想，继续坚持以委员为本的发展主基调，着力从深化国际交流合作、强化“CV+”服务产业、助力青年教师成长等方面推进各项工作。

会议在热烈的交流讨论氛围中结束，会后全体成员合影留念。

责任编委 毋立芳

专题综述

从因果视角量化和评估多模态大模型中的单模态偏见

北京大学 陈美琪 张岩　新加坡管理大学 曹艺馨　上海人工智能实验室 陆超超

本文是北京大学、新加坡管理大学、与上海人工智能实验室团队合作研究的成果。论文探讨了多模态大模型（MLLMs）过度依赖单一模态偏见（biases），如语言偏见和视觉偏见，而在复杂的多模态任务中给出错误的答案的问题。针对这一问题，论文提出了一个因果框架来解释视觉问答（VQA）问题中的偏见。通过该框架，论文设计了一个因果图来阐明论文探讨了多模态大模型在VQA问题上的预测，并通过深入的因果分析评估偏见的因果效应。基于因果图，论文介绍了一个新的数据集MORE，包含12000个VQA实例，这些实例设计用来挑战多模态大模型的能力，需要它们进行多跳推理和克服单一模态偏见。大量的定量和定性实验为未来的研究提供了有价值的见解。论文[1]的项目网页公开在 https://opencausalab.github.io/MORE.

一、研究背景

继大语言模型（LLMs）[2,3]的成功之后，多模态大模型（MLLMs）[4, 5] 已被提出用于各种视觉-语言任务[6, 7]。尽管取得了令人鼓舞的结果，但它们是否真正理解图像和文本在多模态推理上下文中的含义仍然不清楚。如图 1 所示的基于知识的视觉问题回答（VQA）问题中，当被问到“哪个国家将在这个场馆之后举办下一届世界杯？”时，多模态大模型，比如 GPT-4V 和 LLaVA，可能会捕捉到“下一届世界杯”的语言偏见，并认为下一届世界杯将是“在卡塔尔举行的 2022 年世界杯”（这也是过时的知识），同时忽略了图像中呈现的确切场馆。同样，当呈现出伦敦的“碎片大厦”图像时，受到视觉偏见的影响，多模态大模型直接识别出“代表性建筑是碎片大厦”，而忽略了问题中提到的特定限制“在柏林”。这些固有的问题对多模态大模型的推理能力提出了重大挑战，尤其是面对更复杂的问题时。

图 1　单一模态偏见过度依赖的例子。多模态大模型由于语言偏见（左侧图像下划线的文本所示）和视觉偏见（右侧图像）错误生成了答案。

为了调查多模态大模型对这种单一模态偏见的过度依赖问题，我们提出了一个因果框架来解释和量化语言和视觉偏见。具体来说，我们首先定义了多模态大模型在 VQA 问题上预测的因果图。因果图是基于预测过程中的各种因果因素构建的，如图像和问题文本。然后，我们在 VQA 问题的背景下识别了一系列干预，从而通过do-演算(do-calculus)[8]来确定单一模态偏见对多模态大模型预测能力的因果效应。通过量化这些因果效应，我们可以评估多模态大模型对单一模态偏见的敏感性和鲁棒性。

Datasets	Knowledge-based	Multi-hop Reasoning	Answer Type	Unimodal Biases Evaluation	Rationale	#Size
Visual7W[9]	✗	✗	Open-ended	✗	✗	327.9K
VQA(v2)[10]	✗	✗	Open-ended	✗	✗	1.1M
FVQA[11]	✓	✗	Open-ended	✗	✓	5.8K
OKVQA[12]	✓	✗	Open-ended	✗	✗	14K
S3VQA[13]	✓	✗	Open-ended	✗	✗	7.5K
A-OKVQA[14]	✓	✗	Open-ended	✗	✓	23.7K
INFOSEEK[15]	✓	✗	Open-ended	✗	✗	1.4M
MORE(Ours)	✓	✓	Open-ended	✓	✓	12K

表 1　MORE 和现有 VQA 数据集的对比

基于上述因果分析，我们创建了一个名为 MORE 的新数据集，包含 12,000 个 VQA 实例。该数据集通过引入专门的单一模态偏见评估，提升了现有 VQA 数据集。为了便于评估，我们采用多项选择题（MCQ）格式，每个实例由一张图像、一个问题和四个候选选项组成。图像来源于现有的 VQA 数据集。为了问题和选项的策划，我们纳入了一个知识图谱（KG），允许我们更好地模拟多模态大模型在因果图中导航对应的伪路径。具体而言，选项由一个正确答案和三个分别针对语言偏见、视觉偏见和多跳推理的干扰项组成。我们还为每个实例提供了 KG 中的推理路径，称为因果推理，为评估提供了可解释性。总得来说，与现有的 VQA 数据集相比，MORE 具有外部知识、多跳推理、单一模态偏见评估和推理路径，展现了更好的全面性。在六个领先的多模态大模型上的实验结果显示：1）大多数多模态大模型在 MORE 上的表现较差，明显倾向于依赖单一模态偏见。2）当处理多模态推理时，多模态大模型仍然难以实现精确的语义理解。

二、因果框架介绍

在本节中，受 Stolfo 等人工作[16]的启发我们首先介绍多模态大模型在 VQA 问题上预测的因果图。然后，我们利用因果图阐明 VQA 中固有的偏见，特别是视觉和语言偏见。最后，我们通过执行受控干预[8]来评估这些偏见对多模态大模型预测的因果效应。

1. 问题设置

我们考虑一个以实体为中心的 VQA 问题，记为M，由一个问题Q和一个图像I组成。图像描绘了一个特定的实体，问题与该实体相关。问题Q由两个不同的元素组成：核心语义内容S，传达问题的真实意图；和文本表面形式T，与问题的核心含义无关。模型的最终答案/预测由A表示。在本文中，我们使用小写字母表示其对应的大写变量的一个实例。

2. 多模态大模型预测的因果图

受人类认知中观察到的直观推理机制的启发[16,17]，我们在 VQA 问题m中制定了人类问题解决的因果机制：

$$s = f_c(q)\ ,\ g = f_s(i)$$

其中，认知过程f_c被用来解析问题q中的核心语义含义s。然后，函数f_s将s与图像i相关联，产生最终答案g。我们在图 2 的绿色子图G_h中展示了这些机制。

与此相反，模型解决同一 VQA 问题 m 的可能的因果机制如下：

$$a = f_b(i, q)$$

其中，f_b作为一个黑匣子，使得模型考虑q的哪些方面以及它如何与图像i交互是不确定的。

为了进一步分析，我们在图 2 中绘制了完整的因果图中可能发生的所有可能的因果机制。值得注意的因果机制包括：

- 视觉偏见：模型可能通过因果路径$I \rightarrow A$直接关注图像I，导致视觉偏见的出现。
- 语言偏见：模型可能直接以两种方式处理问题Q：通过因果路径$Q \rightarrow S \rightarrow A$关注核心语义$S$，或

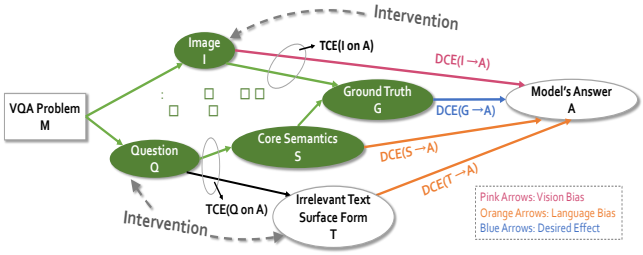


图 2　多模态大模型对 VQA 问题预测的因果图。

通过因果路径$Q \rightarrow T \rightarrow A$关注无关部分$T$。这两条路径都会导致语言偏见。

- 期望的因果机制
- 正确推理的本质在于模型对解决 VQA 问题所需的基本因果机制的把握。如图 2 的绿色子图G_h所示，它应该理解图像和问题如何共同贡献于正确答案G（通过$I \rightarrow G$和$S \rightarrow G$）。因此，模型的预测应该对正确答案的变化显示出敏感性和鲁棒性，即$G \rightarrow A$。没有任何假相关路径可以通过中介G而直接影响A。

基于上述分析，我们阐述模型在 VQA 问题上的敏感性和鲁棒性的概念：

- 敏感性：评估模型在正确答案变化时是否适当调整其预测，即A对G的变化做出反应。
- 鲁棒性：评估单模态偏见的直接因果效应，例如$I \rightarrow A$，$T \rightarrow A$，其中较低的效应表示对不改变正确答案的输入变化具有更好的鲁棒性。

3. VQA 偏见的因果分析

在定义了所期望的因果机制和单模态偏见的路径后，我们可以通过执行受控干预[8]来量化每个因子对另一个因子的因果效应。

因果干预 在 VQA 的上下文中，我们采用以下干预措施来量化图像和问题对模型预测的因果效应：

- 直接对图像I进行干预，将其替换为另一个图像I'。
- 对Q进行部分可控干预。问题Q可以通过两种方式修改：(i) 同时修改S和T，或 (ii) 修改T但保持S不变。

因果效应的计算 接下来，我们解释如何从干预中获得因果效应。考虑一个干预 $do(X: x \rightarrow x')$，其中 $X \in \{I, Q, T\}$，并且 VQA 问题 $M = \{I, Q\}$。我们将干预前的分布 $\mathbb{P}(A \mid I, Q)$ 表示为 P，干预后的分布表示为 P'。

遵循 Pearl (1995) [8]提出的分布式因果效应定义，我们使用距离度量 δ 量化因子 X 在我们的因果图中的效应，即 $CE = \delta(P, P')$，其中 CE 表示因果效应，并且可以进一步细分为总因果效应（TCE，即通过所有从一个变量到另一个变量的定向因果路径的联合效应）或直接因果效应（DCE，即从一个变量到另一个变量的直接因果路径的效应，不经过任何中介变量）。

遵循 Stolfo (2022) [16] 的方法，我们通过评估预测结果的变化来量化因子X对模型答案A的因果效应，即

$$\delta_{cp}(P, P') := \mathbb{I}(a \neq a')$$

其中$a = arg\ max_x P(x)$，$a' = arg\ max_x P'(x)$，$\mathbb{I}$ 表示“答案改变”事件的指示器。

图像的因果效应 当对图像 I 进行干预时，我们可以得到 I 对 A 的因果效应大小，即：

$TCE(I\ on\ A) := E_{i' \sim P(I)}[\delta(P, P')]$，其中 $P' = \mathbb{P}(A \mid Q, do(I = i'))$。

注意，这个 TCE 包含了两条不同的路径，说明了 I 如何影响 A，如图 2 中所示：

- 路径 $I \rightarrow G \rightarrow A$ 代表我们希望模型采用的理想决策路线，其中它响应于正确答案的变化。
- 路径 $I \rightarrow A$ 描述了模型可能学习到的一种虚假关联，其中它依赖于某些可能与训练语料库中的普遍存在相关的视觉上下文。

我们可以量化 I 对 A 的 DCE，即路径 $I \rightarrow A$的强度，通过在每次对 I 进行干预时保持G不变来实现，即：

$DCE(I \rightarrow A) := E_{i' \sim P(I \mid G)}[\delta(P, P')]$，其中 $P' = \mathbb{P}(A \mid Q, do(I = i'))$。

问题的因果效应 对于问题，通过对 Q 进行干预，我们可以计算 Q 对 A 的总因果效应，即：

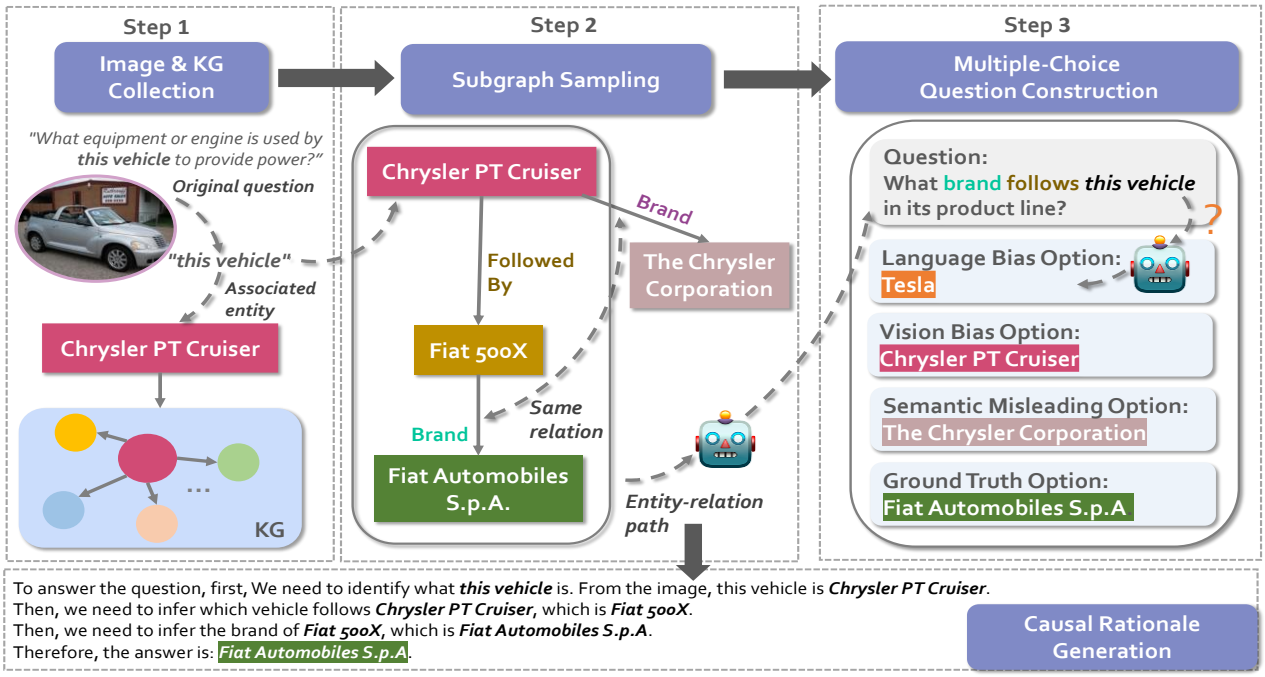


图 3　MORE 数据集的构建流程。

$TCE(Q\ on\ A) := E_{q' \sim P(Q)}[\delta(P, P')]$，其中 $P' = \mathbb{P}(A \mid I, do(Q = q'))$。

控制核心语义意义 S 将允许我们获得文本表面形式 T 对 A 的 DCE，即：

$DCE(T \rightarrow A) := E_{q' \sim P(Q \mid S)}[\delta(P, P')]$，其中 $P' = \mathbb{P}(A \mid I, do(Q = q'))$。

请注意，由于T和A之间没有中介，所以$DCE(T -> A)$也是T对A的TCE。因为枚举T的所有可能扰动通常不可行，我们可以通过干预 Q 而不影响 S，在 T 的某个子集上获得实际结果[16]。此外，在 VQA 问题的上下文中，我们不能只干预 S 而不影响文本表面T。然而，通过比较我们已经知道的两个量，即$TCE(Q\ on\ A)$和$DCE(T -> A)$，可以帮助我们理解S对A的因果影响。

总的来说，计算TCE帮助我们评估模型的敏感性（对正确答案变化的反应），而 DCE 评估其鲁棒性（对固定正确答案时假相关性预测的稳定性）。

三、构建MORE数据集

本章构建了一个新颖 MORE 数据集，要求多模态大模型超越单模态偏见，并从文本和图像中彻底整合信息以选择正确答案。数据生成过程如图 3 所示。

1. 图像和知识图谱收集

我们从一个现有的视觉问答（VQA）数据集 INFOSEEK[15]开始，该数据集将图像中描绘的实体与从 Wikipedia 来源的信息链接起来，要求 VQA 模型回答有关关联实体的问题。基于图像和对应的实体信息，我们在知识图谱（KG）- Wikidata5M[3]中识别所有与关联实体相关的n阶邻居（$n \in \{1,2\}$）。

2. 子图采样

受到因果分析的启发，我们旨在构造需要克服单模态偏见才能正确回答的多跳查询。为此，我们首先识别实体及其在 KG 中的n阶邻居的子图。然后，过滤满足两个标准的路径：1）路径的唯一性：从关联实体到选定邻居的路径是唯一的；2）共享类型关系：它们共享指向唯一实体的相同类型关系，这两个指向的实体不相同。

3. 多项选择题构造

在此小节，我们详细说明构造四个候选选项的多项选择题的过程。

问题生成 获取满足标准的子图后，我们使用子图中的实体-关系路径生成问题。为了获得流畅且连贯的问题，我们将路径输入到一个大语言模型中产生目标问题文本。我们采用了上下文学习（ICL）[3]技术，并为大语言

Model	MORE (Two-hop, acc (%))		MORE (Three-hop, acc (%))		MORE (Overall, acc (%))	
	Open-ended	Multi-choice	Open-ended	Multi-choice	Open-ended	Multi-choice
Random	/	25.0	/	25.0	/	25.0
BLIP2	4.0	16.4	1.4	15.4	2.7	15.9
InstructBlip	3.0	17.0	1.6	16.2	2.3	16.6
mPLUG-Owl	4.0	12.4	8.2	11.4	6.1	11.9
LLaVA	8.0	20.8	6.8	13.6	7.4	17.5
GPT-4V	15.8	25.6	15.3	23.2	15.6	24.4
Gemini Pro	14.2	33.5	10.1	24.4	12.2	28.9

表 2　多模态大模型在 MORE 上的结果对比。

模型提供了几个例子。在比较了不同的大语言模型并调整指令后，我们发现 ChatGPT 生成的多跳问题质量最高，因此选择其结果进行后续的评估。最后，为了防止信息泄露，问题中的实体名称被替换为 “this <OBJECT_NAME>”。

语言偏见选项 如前所述，语言偏见指的是模型过度关注问题文本中的信息。为了模拟这种情况，我们在纯文本设置下使用生成的问题测试多模态大模型。为确保所有多模态大模型的最终选项相同，我们统一使用 GPT-4V 生成得到的答案。

视觉偏见选项 为了探索沿着$I \rightarrow A$路径的视觉偏见，我们将与视觉相关的实体名称（例如，“Chrysler PT Cruiser”）作为一个选项。这允许我们观察模型在遇到与视觉信息对齐的选项时是否直接选择它。

语义误导选项 此外，我们引入了一个语义误导选项，例如“The Chrysler Corporation”，挑战多模态大模型在 KG 中的多跳推理。这个选项指的是被两个关联实体和它们的采样邻居共同拥有的关系所指向的实体。

正确答案选项 与通过$I \rightarrow G$和$S \rightarrow G$的因果路径相对应，这个选项是实体-关系路径的最终实体（例如，“Fiat Automobiles S.p.A.”）。最后，我们检查并确保每个选项与其余三个选项不同，以消除重叠样本。

因果推理路径生成 此外，实体-关系路径可以帮助生成针对当前问题的推理过程，被称为因果推理路径。在这一背景下，我们采用了一种启发式规则基础方法，从关联实体开始，逐步生成因果理由，直至达到正确答案。这些生成的因果理由可以用来验证多模态大模型的推理过程是否正确，从而提供可解释性。它们也可以用于微调专门的多模态大模型，以增强其多跳推理能力。因果理由还可以通过如 ChatGPT 这样的大型语言模型（LLMs）进一步打磨和精炼，我们将这一点留给未来的工作。

四、在MORE上评估多模态大模型

1. 实验设置

数据集 我们使用MORE的所有测试数据进行评估。我们采用了两种不同的设置：

- 开放式。要求MLLM基于输入的图像和问题生成答案。
- 多选。为MLLM提供四个选项，让它从中选择正确答案。后一种设置有一个随机基线（准确率为25%）。

基准 我们以零样本（zero-shot）的方式在我们的MORE数据集上评估各种领先的多模态大模型，包括两个有限访问的多模态大模型：GPT-4V和Gemini Pro，以及四个开源的多模态大模型：BLIP-2（6.7B），InstructBLIP (13B)，mPLUG-Owl (7B)，和LLaVA (v1.5, 13B)。就评估指标而言，我们采用VQA准确率对所有模型进行公平比较。

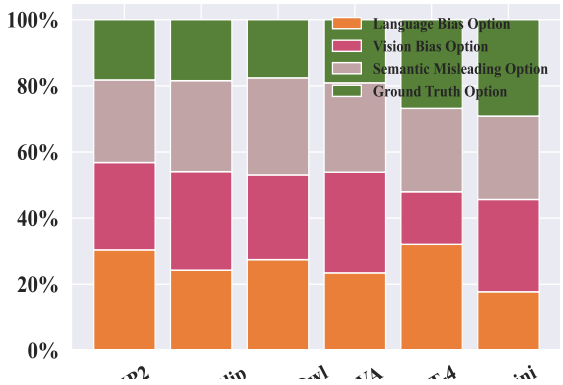


图 4　多模态模型的选项分布

2. 评估结果

我们在表2中分别展示了多模态大模型在MORE数据集的两跳、三跳和所有数据上的结果。我们观察到：1）所有基线在MORE上的性能都较差（例如，在``多选''设置下，只有Gemini Pro超过了随机基线，准确率为28.9%），这表明MLLM对语言和视觉偏见的脆弱性。2）在MORE上，开源模型与有限访问模型之间仍存在差距，尤其是在“开放式”设置下。3）大多数模型在两跳数据上的表现优于三跳数据（Gemini Pro在两跳数据上表现尤为出色，准确率达到33.5%），这表明当问题变得更加复杂时，多模态大模型的推理能力受到挑战。4）GPT-4V在开放式''设置下表现最佳，但在多选''设置下相比Gemini Pro略显不足，可能是因为在构造语言偏见选项时，我们使用了同源的ChatGPT生成的干扰项，这对GPT-4V的判断构成了更大的挑战。这一点也在后续的分析中得到验证。

3. 对 VQA 偏见的因果分析

在本小节中，我们通过因果视角分析多模态大模型的性能。

选项分布 图4显示了在“多选”设置下各种MLLM的选项分布。我们观察到：（1）BLIP2和GPT-4V经常错误选择表明语言偏见的选项，这与我们对GPT-4V的先前分析一致。（2）在所有模型中，语言或视觉偏见的比例超过了40%，显示了单模态偏见对它们预测的显著影响。（3）在一定程度上，模型选择语义上误导的选项表明了一些结合视觉和文本信息的能力，尽管并未完全掌握问题。这突出了我们的MORE数据集对当前MLLM所提出的挑战。请注意，这里呈现的正确选项的比例与表2中报告的准确率值之间可能存在差异，因为某些模型（例如，mPLUG-Owl）的输出可能不符合提供的选项，这影响了有效答案的计数。

图像和问题的因果效应 为了进一步分析视觉偏见和语言偏见对模型预测的影响，我们根据第二章提供的定义评估了因果效应。具体而言，我们随机选择100个样本进行干预，然后测量所有实例的效果平均值，以计算TCE（对应于模型的敏感性）和DCE（对应于模型的鲁棒性）。总的来说，较高的TCE是可取的，表示更好的敏感性，而较低的DCE表示更好的鲁棒性。

从图5可以看出：1）当前的多模态大模型展现出高敏感性（高TCE），一个可能的原因是指令调整使得模型对输入的变化更为敏感。2）然而，鲁棒性相对较低（高DCE），显示出即使在固定的答案值下，预测也会随着输

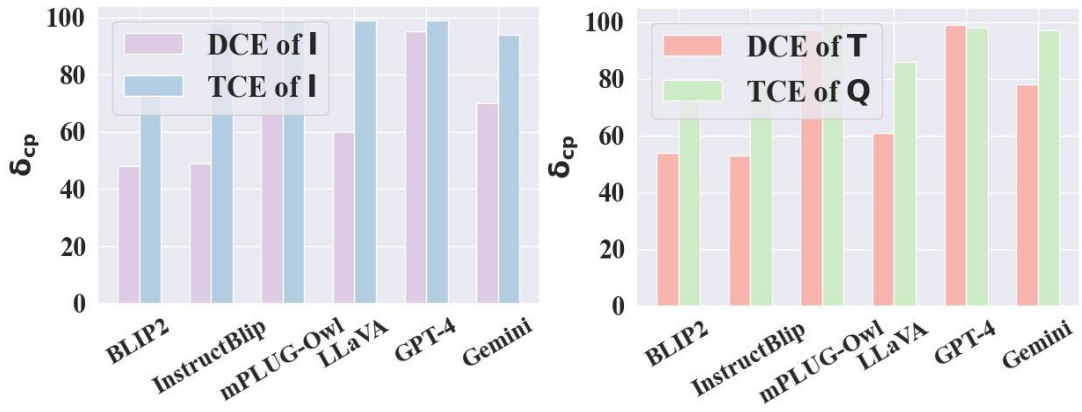


图 5　比较图像和问题对多模态大模型预测的因果效应

入的变化而改变，这表明依赖于伪造路径而非真实的因果特征。

五、总结

本文提出了一种全面的方法来量化和评估多模态大模型中的单模态偏见。通过我们的因果推理框架，我们深入分析了这些偏见对 VQA 问题中模型预测的因果效应。我们引入的 MORE 数据集要求多模态大模型进行多跳推理，并克服语言和视觉偏见，从而扩展了它们的推理能力边界。一系列定量与定性的分析实验为未来的工作提供了见解。

责任编委 **魏秀参**

参考文献

[1] Chen, M., Cao, Y., Zhang, Y., Lu, C., Quantifying and Mitigating Unimodal Biases in Multimodal Large Language Models: A Causal Perspective. *arXiv:2403.18346.*

[2] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *NeurIPS*, *35*, 27730-27744.

[3] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv:2302.13971*.

[4] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv:2303.08774*.

[5] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J. B., Yu, J., ... & Ahn, J. (2023). Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*.

[6] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2023). A survey on multimodal large language models. *arXiv:2306.13549*.

[7] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. *NeurIPS*, *36*.

[8] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669-688.

[9] Zhu, Y., Groth, O., Bernstein, M., & Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. *CVPR*, pp. 4995-5004.

[10] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *CVPR*, pp. 6904-6913.

[11] Wang, P., Wu, Q., Shen, C., Dick, A., & Van Den Hengel, A. (2017). Fvqa: Fact-based visual question answering. *IEEE T-PAMI*, *40*(10), 2413-2427.

[12] Marino, K., Rastegari, M., Farhadi, A., & Mottaghi, R. (2019). Ok-vqa: A visual question answering benchmark requiring external knowledge. *CVPR*, pp. 3195-3204.

[13] Jain, A., Kothyari, M., Kumar, V., Jyothi, P., Ramakrishnan, G., & Chakrabarti, S. (2021, July). Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. *ACM SIGIR*, pp. 2491-2498.

[14] Schwenk, D., Khandelwal, A., Clark, C., Marino, K., & Mottaghi, R. (2022, October). A-okvqa: A benchmark for visual question answering using world knowledge. *ECCV*, pp. 146-162.

[15] Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., & Chang, M. W. (2023). Can pre-trained vision and language models answer visual information-seeking questions? *arXiv:2302.11713*.

[16] Stolfo, A., Jin, Z., Shridhar, K., Schölkopf, B., & Sachan, M. (2022). A causal framework to quantify the robustness of mathematical reasoning with language models. *arXiv:2210.12023*.

[17] Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., ... & Wen, J. R. (2022). Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, *13*(1), 3094.

[18] Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., & Tang, J. (2021). KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, *9*, 176-194.

陈美琪

北京大学智能学院 2020 级博士研究生，导师为张岩教授，主要研究方向为自然语言处理和大模型。

Email:meiqichen@stu.pku.edu.cn

曹艺馨

新加坡管理大学助理教授，清华大学计算机科学博士，曾担任南洋理工大学的研究助理教授和新加坡国立大学 NExT++研究中心的研究员。研究领域涵盖自然语言处理、知识图谱和推荐系统，各项工作已发表在包括 ACL、EMNLP、COLING 和 WWW 在内的顶级会议上，共获得超过 4,000 次引用。

Email: caoyixin2011@gmail.com

张岩

张岩，北京大学智能学院教授，博士生导师。研究兴趣是信息检索、自然语言处理、数据挖掘、网络科学和大数据分析，近年来在这些领域的国际期刊和会议上发表论文 100 多篇，作为项目负责人和技术骨干承担和参与国家自然科学基金项目、核高基重大专项、科技支撑计划重点项目、973 项目、北京市基金、教育部项目、粤港合作项目等二十余项。

Email:zhyzhy001@pku.edu.cn

陆超超

上海人工智能实验室青年科学家，博士生导师，因果智能团队负责人，主要从事因果推理方面的理论研究及其在机器学习等相关领域的应用。

Email: luchaochao@pjlab.org.cn

热点追踪

基于自先验的盲图像修复方法

中国科学院自动化研究所　王隽

一、引言

图像修复是计算机视觉领域的重要研究问题。根据受损区域位置是否已知，图像修复可分为非盲图像修复和盲图像修复。非盲图像修复的前提是需要事先给定一个二进制的掩膜用于标注受损区域的位置。盲图像修复则是在掩膜未给定的条件下对受损图像进行修复，因此盲图像修复任务更具挑战性。

图像修复是一个病态问题，即修复的结果有无数可能的解。为了得到更加准确的修复结果，需借助一定的先验信息。传统意义上的“先验“通常人为进行定义，在本工作中，我们提出挖掘和利用图像自身的先验信息来指导修复过程。具体地，我们从盲图像修复的两个关键问题入手：一是“修复何处”，二是“如何修复”。我们将两个问题分别定义为受损区域检测问题和图像全局语义结构预测问题。相应地，获得一个掩膜图和一个布局图，两者被称为“自先验”，用于指导模型在合适的位置、以更有效的策略完成修复任务。

二、基于自先验的盲图像修复方法总览

总体框架如图 1 所示，主要由三个关键部分组成：

自先验学习网络：包括一个语义连续性检测网络（semantic-discontinuity detection network，SDN）和一个布局预测网络（layout prediction network，LPN），两者将用于估计受损图像的自先验信息，即掩膜（mask）和布局图（layout）。

自先验引导的修复网络：在所估计的掩膜的引导下，提取有用的上下文信息；同时，在所估计的布局图的引导下，自动合成逼真的纹理细节。

基于自先验的损失函数：对两个自先验学习网络再利用，衡量修复图像和参考图像的语义连续性和感知一致性，加入到损失函数中进一步促进修复图像的质量。

（1）语义连续性检测网络

该网络的提出是为了解决盲图像修复任务中“修复何处”的问题。污损信号破坏了图像的语义结构性，因此我们将受损区域定位建模为语义连续性检测问题，通过区分图像信号和污损信号的模式差异，来估计输入图像的掩膜图。语义连续性检测网络（SDN）采用 UNet 网络结构，由一个编码器和一个解码器组成。编码器以受损图像作为输入，输出图像的隐式表征；解码器从该隐式表征中发现语义不连续的区域，输出一幅掩膜图，每一个位置的数值表示图像中对应位置像素值属于未受损区域的概率。

（2）布局预测网络

尽管基于 CNN 的图像修复方法在性能方面大幅超过传统方法，但是当受损区域较大时，生成的细节仍然不够逼真。因此我们将“如何修复”问题分解为全局语义结构推断和局部细节合成两个步骤。我们首先构建了一个布局预测网络（LPN）来实现全局语义结构的推断，生成一幅描绘了图像中各类别物体轮廓的布局图，其类别和形状信息将用于指导修复图像的细节合成。我们采用 Deeplab V3+来实现 LPN。由于输入图像中存在的污损信号具有复杂、多样化的特点，将导致模型估预测的准确率显著下降。我们将第一步预测的掩膜二进制化后与输入图像相乘，来消除污损信号的干扰。同时 LPN 输出的布局图也会经过形态学算子（如腐蚀、膨胀等）以消除噪声、平滑边缘细节。

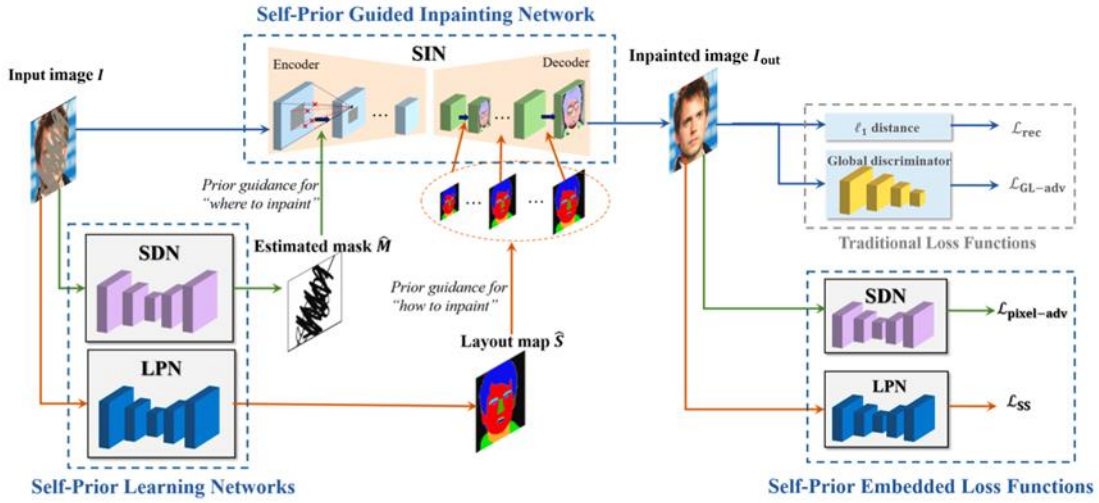


图 1　基于自先验的盲图像修复方法框架总览

（3）自先验引导的修复网络

前两步分别得到了两项重要的自先验信息，即掩膜和布局图。本方法的核心网络是自先验引导的修复网络(Self-prior guided inpainting network，SIN)，它将在两项自先验信息的引导下生成一幅完整的图像。SIN 采用基于编码器-解码器的 UNet 结构，详细的网络结构设计如图 2 所示。为了将两项自先验信息有效嵌入到 SIN 中来指导网络“修复何处”以及“如何修复”，我们设计了两个新的网络结构：局部软卷积（soft-partial convolution，SPConv）和语义结构嵌入单元（semantic structure embedding unit, SSEU）。

SPConv：传统卷积操作对滑动窗内的所有像素值赋予相同的权重，而在修复任务中，输入图像中的受损区域与未受损区域携带的图像信息量不同，不应采用相同的权重。因此，我们提出 SPConv，在卷积计算过程中根据估计的掩膜图对图像/特征图中每个像素值赋予一个权重，卷积的结果为滑动窗内所有像素值的加权和。给定一幅输入图像x和掩膜x，令x_{win}和m_{win}分别表示当前卷积核在输入图像x和掩膜x上覆盖的区域，经过 SPConv 操作后的输出x_{out}在位置(u,v)处的值可由下式进行计算：

$$x_{out}(u,v) = W^T(x_{win} \odot m_{win}) \frac{\sum_{(i,j)\in R_{win}}(\mathbf{1}(i,j))}{\sum_{(i,j)\in R_{win}}(m_{win}(i,j))} + b,$$

其中R_{win}表示当前区域，b为偏置，$\mathbf{1}$为单位矩阵，和W具有相同的形状。

SSEU：目的是将布局图中的类别和形状信息嵌入到 SIN 中，以指导 SIN 在特定类别的区域内生成语义一致的纹理细节。SSEU 的核心技术为自适应仿射变换，使用估计的布局图中的语义编码对 SIN 中间生成的特征图的均值和方差进行调制，将布局图的语义嵌入到 SIN 的特征图中，从而实现全局语义结构信息的迁移。输入从受损图像中提取的特征图f和估计的布局图a，SSEU 需要首先对a学习两个仿射参数γ和β，然后对l进行如下操作得到输出特征图：

$$\hat{f} = \gamma \frac{f-\mu}{\sigma} + \beta,$$

其中μ和σ分别表示特征图f的均值和方差。

（4）基于自先验的损失函数

图像重建任务普遍采用基于像素域的距离损失作为损失函数。但该损失函数往往会导致重建模糊问题。近年来，感知损失和对抗损失的使用极大提高了重建图像的质量，促进了边缘的锐利化和细节的丰富度。但是，在图像修复任务中，这两项损失函数往往会引发语义不连续性、受损边界重影等问题。为了缓解该问题，我们提出了两项新的损失函数，即像素对抗损失函数和语义结构损失函数。

像素对抗损失函数：生成对抗网络的判别器对整幅图像判别真假，但该方式过于粗糙，往往会忽略图像中的虚假细节，如修复图像中的受损边界重影。对此，该

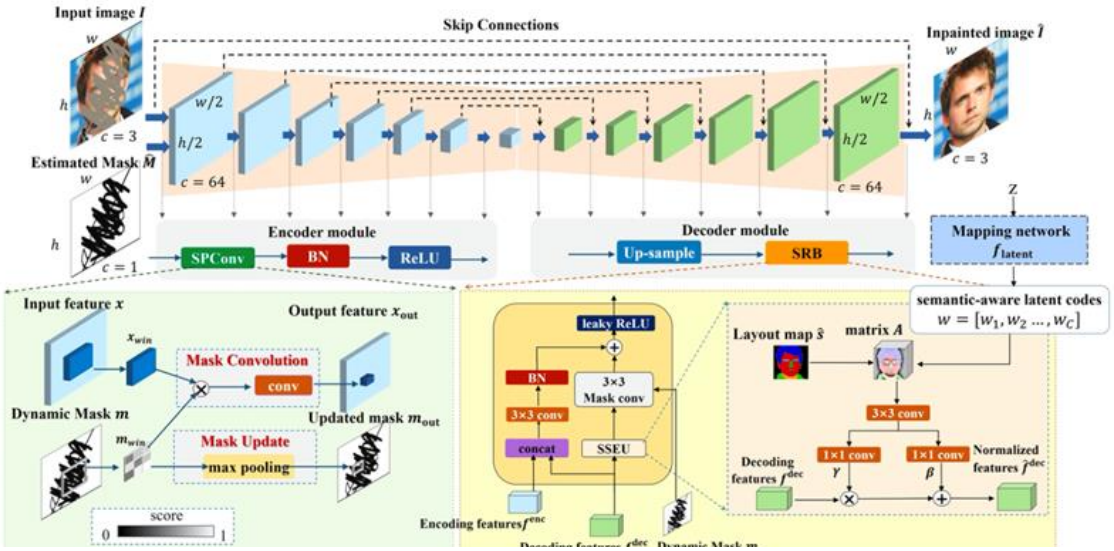


图 2　自先验引导的修复网络结构设计

工作提出了一种基于密集预测的判别器。它根据输入图像中每个像素值与周围区域的语义连续性，判别每个像素值的真假，为生成器提供像素级的反馈。这里对 SDN 进行重利用，将它作为判别器，将 SIN 作为生成器，两者构成一对生成对抗网络。令$\hat{I}$表示修复图像，D 表示对抗网络，像素对抗损失函数定义如下：

$$L_{pix-adv}(\hat{I}) = -\frac{1}{\Phi}\sum_{(i,j)} \log\left[D(\hat{I}(i,j)\right],$$

其中Φ表示图像$\hat{I}$中所有像素的数目。

语义结构损失函数：目前流行的感知损失是通过预训练的神经网络 (如 imagenet 上预训练的 VGG) 计算两幅图像在特征域的相似性。然而，如果目标数据集与预训练数据集的分布差异过大，将导致语义鸿沟问题，造成感知损失计算的准确性下降。对此，该工作使用在目标数据集上预训练的 LPN 来代替 VGG 计算感知损失，改进的损失函数称为语义结构损失函数。令f_{LPN}表示 LPN 网络，I_{gt}表示参考图像，语义结构损失函数定义如下：

$$L_{ss}(\hat{I}, I_{gt}) = \sum_{t=1}^{T} \frac{1}{\Theta_t} ||f_{LPN}^{t}(\hat{I}) - f_{LPN}^{t}(I_{gt})||_1,$$

其中Θ_t表示 LPN 网络第 t 层输出特征图上像素的总数，T 为总的层数。这里我们取 LPN 网络的第 2,3 和 4 个池化层。

最后，总的损失函数是基于像素域的距离损失（ℓ_1 损失）、全局对抗损失，以及所提出的像素对抗损失和语义结构损失，共四项损失的加权和。

三、实验结果

为验证所提方法的有效性，我们在多个数据集上进行了实验验证，包括修复图像的主客观质量对比、消融实验以及应用扩展等。

指标	掩膜比例	CelebA					
		CA	PC	GC	EC	PDGAN	Ours
ℓ_1	10-20%	2.48	0.83	0.88	0.85	0.90	0.79
	20-30%	3.98	1.46	1.54	1.54	1.45	1.43
	30-40%	5.64	2.36	2.33	2.37	2.68	2.09
	40-50%	7.35	4.01	3.29	3.37	3.02	2.99
PSNR	10-20%	25.32	33.05	32.69	32.53	32.40	33.13
	20-30%	22.09	29.10	29.45	29.19	29.64	29.78
	30-40%	19.94	27.24	27.01	26.72	27.23	27.77
	40-50%	18.41	23.46	24.98	24.67	24.39	25.51
SSIM	10-20%	0.888	0.978	0.976	0.978	0.847	0.986
	20-30%	0.819	0.956	0.954	0.957	0.808	0.971
	30-40%	0.750	0.926	0.927	0.928	0.748	0.942
	40-50%	0.678	0.839	0.892	0.891	0.687	0.909
FID	10-20%	13.74	7.63	25.49	5.03	18.32	4.72
	20-30%	17.78	8.74	29.68	7.37	21.02	5.69
	30-40%	24.07	19.21	70.37	10.66	24.63	10.27
	40-50%	37.35	27.11	75.04	14.53	27.12	13.13
IS	10-20%	3.43	3.44	3.32	3.42	3.09	3.97
	20-30%	3.34	3.42	3.17	3.34	3.02	3.84
	30-40%	3.10	3.25	3.06	3.13	2.90	3.61
	40-50%	3.01	3.06	2.88	2.94	2.70	3.35
LPIPS	10-20%	0.043	0.037	0.080	0.028	0.070	0.028
	20-30%	0.078	0.056	0.116	0.051	0.094	0.043
	30-40%	0.127	0.101	0.192	0.074	0.129	0.070
	40-50%	0.195	0.143	0.261	0.106	0.173	0.098

表 1　修复方法在 CelebA 数据集的客观质量对比

图 3　各方法修复图像的主观质量对比，从左到右分别为(a) 输入图像; (b) ground truth; (c) CA[2]; (d) PC[3]; (e) EC[1]; (f) PIC[4]; (g) SN[5]; (h) PHD[6]; (i) 该工作生成的布局图; (j) 该工作的修复结果。

指标	掩膜比例	Paris Street View				
		PC	GC	EC	PRVS	Ours
ℓ_1	10-20%	1.23	1.26	1.11	1.05	1.12
	20-30%	2.12	2.07	1.95	1.82	1.97
	30-40%	3.09	3.00	2.87	2.66	2.61
	40-50%	4.21	4.06	3.93	3.63	3.57
PSNR	10-20%	30.76	31.42	31.05	32.00	31.14
	20-30%	27.62	28.12	28.05	28.79	28.65
	30-40%	25.51	25.80	25.98	26.62	26.79
	40-50%	23.81	23.93	24.29	24.87	24.99
SSIM	10-20%	0.953	0.959	0.956	0.964	0.962
	20-30%	0.910	0.920	0.917	0.928	0.928
	30-40%	0.858	0.873	0.869	0.885	0.887
	40-50%	0.780	0.815	0.811	0.832	0.834
FID	10-20%	27.21	53.68	27.66	21.04	20.11
	20-30%	41.97	82.41	41.72	25.32	23.17
	30-40%	55.79	137.49	60.28	40.51	32.56
	40-50%	66.72	182.02	86.52	46.70	41.19
IS	10-20%	3.02	2.88	2.97	3.03	3.11
	20-30%	3.01	2.82	2.94	3.00	3.07
	30-40%	2.95	2.56	2.92	2.99	3.01
	40-50%	2.87	2.48	2.91	2.81	2.97
LPIPS	10-20%	0.045	0.112	0.047	0.056	0.044
	20-30%	0.067	0.172	0.071	0.123	0.062
	30-40%	0.107	0.267	0.118	0.268	0.102
	40-50%	0.159	0.384	0.173	0.305	0.134

表 2　修复方法在 Paris Street View 数据集的客观质量对比

各方法修复图像的客观质量对比如表 1 和表 2 所示。结果表明，当受损区域较大时，即 mask 在 30%-50%时，本方法相比现有方法的优势更加明显，各项指标均超过了非盲修复方法。本方法为盲图像修复方法，不需要事先给定一个准确的掩膜图。各方法修复图像的主观质量对比如图 3 所示。可以看到，相比现有方法，本方法能够更有效地去除污损，并生成与周围区域语义连续性与感知一致性的修复结果。

为验证本方法中各个关键部分的有效性，训练了多个变体模型。各个变体模型的组成和消融实验结果如表 3 所示。结果表明，所有变体模型在两个数据集上的所有指标都相比完整模型有所下降，从而证明本框架中的各个关键部分对本模型的性能均具有正向促进作用。

Model	PSNR	SSIM	LPIPS
w/o mask	26.96	0.921	0.102
w/o layout	27.81	0.933	0.091
w/o SPConv	28.92	0.939	0.069
w/o $L_{pix-adv}$	28.21	0.923	0.089
w/o L_{SS}	28.02	0.921	0.098
Full	28.97	0.942	0.067

表 3　本方法在 CelebA 数据集上的消融实验

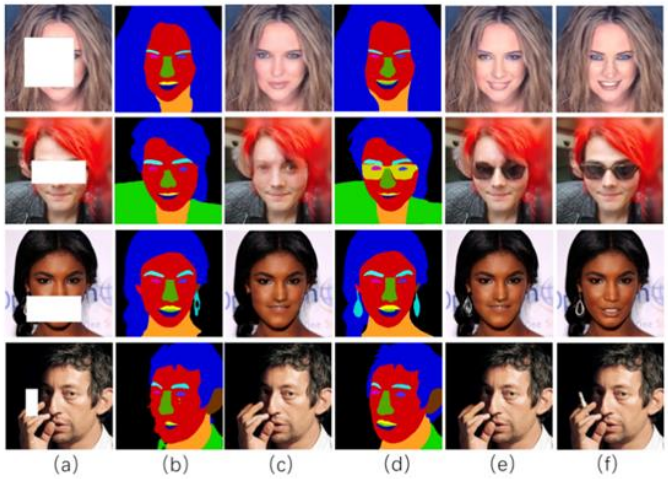


图 4　使用不同布局图修复图像的结果。从左到右分别为 (a) 输入图像; (b) 该工作生成的布局图; (c) 使用(b)的修复结果; (d) ground truth 布局图; (e) 使用(c)的修复结果; (f) ground truth。

图 4 展示了使用不同布局图生成修复图像的结果对比。可以看到，使用 ground truth 布局图可以正确修复原始图像中的人物表情 (例如微笑) 和丢失的属性 (例如墨镜和耳环)。

图 5 展示了该工作在物体去除任务上的结果。从前两行的对比结果可以看到，对比方法 PHD 的修复图像中仍然有墨镜和胡子的影子未有效消除，而该工作可以产生更加干净的修复结果。从后两行的对比结果可以看到，该工作产生更加逼且令人眼舒适的修复结果。

图 6 展示了在真实老照片上的修复结果。可以看到，两个对比方法的修复结果均表现出不同程度的模糊，以及有未消除的折痕。相比之下，该工作的修复结果在视觉效果方面更加令人满意。该实验结果表明该工作具备从合成失真分布到真实失真分布的泛化能力。

五、总结和展望

本文针对盲图像修复“修复何处”和“如何修复”两个关键问题提出了一种基于自先验引导的盲图像修复方法。在该工作中，构建了一个语义连续性检测网络和一个布局预测网络，分别用于估计受损区域的位置和全局语义结构，作为两项自先验信息。在自先验的引导下，修复网络自适应调整卷积权重以提取有效的图像信息，以及在类别和形状信息的提示下合成丰富的图像细节。大量的实验结果表明，所提方法能够生成比现有模型客观指标上更加准确、主观视觉效果上更令人满意的修复结果。

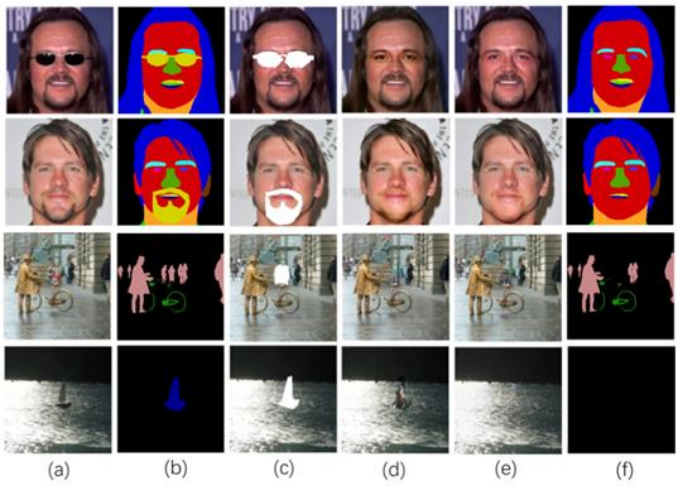


图 5　物体去除结果展示。从左到右分别为(a) ground truth; (b) ground truth 布局图; (c) 遮挡图像; (d) PHD[6]的修复结果; (e) 该工作的修复结果; (f) 该工作生成的布局图。

图 6　老照片修复结果对比。从左到右分别为(a) 输入图像; (b) 方法[7]的修复结果; (c) PHD 的修复结果[6]; (d) 该工作的修复结果。

该成果已发表在国际顶级期刊 IEEE Transactions on Pattern analysis and machine intelligence (TPAMI)。

责任编委　储珺

参考文献

[1] K. Nazeri, E. Ng, T. Joseph, F. Qureshi and M. Ebrahimi, “Edge- connect: Generative image inpainting with adversarial edge learn- ing,” arXiv preprint arXiv:1901.00212, 2019.

[2] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu and T. Huang, “Generative image inpainting with contextual attention,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5505–5514.

[3] G. Liu, F. Reda, K. Shih, T. Wang, A. Tao and B. Catanzaro, “Image inpainting for irregular holes using partial convolutions,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 85–100.

[4] C. Zheng, T. Cham and J. Cai, “Pluralistic image completion,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 1438–1447.

[5] Z. Yan, X. Li, M. Li, W. Zuo and S. Shan, “Shift-net: Image inpainting via deep feature rearrangement,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 1–17.

[6] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz and B. Catanzaro, “High- resolution image synthesis and semantic manipulation with con- ditional gans,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8798–8807.

[7] D. Chen P. Zhang D. Chen J. Liao Z. Wan, B. Zhang and F. Wen, “Bringing old photos back to life,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 2744–2754.

王雋

中国科学院自动化研究所，多模态人工智能系统全国重点实验室副研究员，主要研究方向为深度学习、计算机视觉、图像修复、图像质量评估。

Email:jun_wang@ia.ac.cn

热点追踪

双鉴别器多模态医学图像融合网络

山东财经大学 刘慧

多模态图像融合是计算机视觉领域中的一项重要任务，其目的是结合各个模态图像的特点，如有物理含义的高亮区域和纹理细节。在医学影像辅助诊断领域，近年来多种模态的医学图像在临床诊断和疾病分析中发挥着越来越重要的作用，由于各类设备的成像原理不同，其产生的医学图像所反映的病理特征各异，提取不同模态医学图像中的关键信息和互补特征，并将其融合到一幅图像中，这一任务即多模态医学图像融合。Ma 等[1]提出了基于生成对抗网络的 FusionGAN 来融合红外和可见光图像，该方法可以通过生成器和鉴别器之间的对抗学习自动提取和融合图像的关键特征。然而，FusionGAN 只专注于针对一种模态的特征进行对抗性训练，却忽略了对其他模态有用信息的保留。随后，Ma 等[2]又提出了一种双鉴别器条件生成对抗网络（DDcGAN）。该网络包含一个发生器和两个鉴别器，分别对不同模态的特征进行对抗，弥补了 FusionGAN 的不足。实验结果表明，DDcGAN 在 MR-T1 和 PET 图像融合任务中表现良好。

鉴于此，本文提出了一种双鉴别器多模态医学图像融合网络（DDIFN），该网络由一个生成器和两个对称鉴别器组成。其中，生成器主要负责特征提取、特征融合和图像重建。对称双鉴别器分别用于根据不同模态的特征与生成器进行对抗训练。与 DDcGAN 不同的是，为了降低神经网络反向传播的计算复杂以及避免模式崩塌，DDIFN 采用最小二乘损失函数作为对抗损失。此外，DDIFN 还定义了针对多模态医学图像融合任务的内容损失函数，以充分保留不同模态图像的细微特征及显著信息，如 MR-T1 的梯度特征以及 MR-T2 和 PET 的像素活动。为了进一步避免融合过程中特征丢失，DDIFN 将 U-Net 作为整个网络的生成器，利用其跳跃连接将下采样过程中丢失的特征进行及时补充。然而，传统 U-Net 通常采用反卷积进行上采样，由于卷积核和步长之间的关系，容易出现卷积核重叠不均匀的情况，这会导致融合结果中出现棋盘状伪影（如图 1）。因此，DDIFN 用双线性插值代替了反卷积的上采样方法，这不仅能够有效避免融合图像出现棋盘效应，而且减少了模型参数量，进而有效避免因样本数量少而导致模型过拟合。最后，多模态特征保留程度不均衡会严重影响融合图像的视觉效果，所以 DDIFN 中的两个鉴别器具有相同的结构，以确保来自不同源图像的特征可以被均衡地保留。DDIFN 的具体结构如图 2 所示。

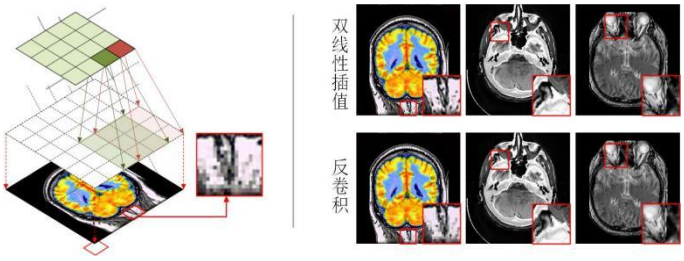


图 1　反卷积过程及不同上采样方式结果对比

一、生成器

我们选择 U-Net 作为 DDIFN 的生成器，利用其对源图像进行特征提取和融合。如图 3 所示，该生成器由 7 个 block 组成，相较于原始的 U-Net，结构得到了简化，目的是在不影响融合效果的前提下尽可能减少网络参数量并降低计算过程复杂程度。在 DDIFN 的生成器中，block1-block3 构成编码器，block5-block7 构成

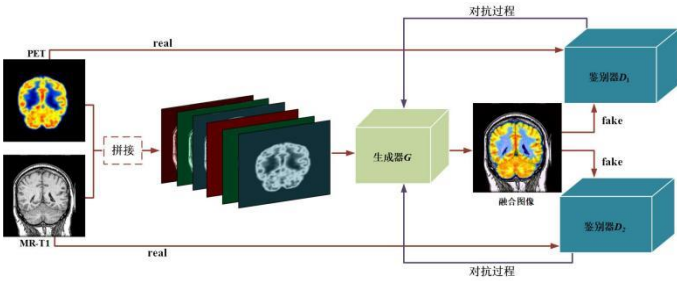


图 2　DDIFN 结构图

解码器，block4 用于连接编码器和解码器。此外，为了避免梯度消失，DDIFN 中的每一层均采用批归一化对卷积后的特征进行进一步调整，并同时采用 ReLU 作为激活函数，以保证特征稀疏性，避免过拟合。

在编码器中，每个 block 均包含两个卷积层和一个下采样层。卷积的目的是通过将输入数据映射到不同特征空间来提取不同特征，是模型自适应提取显著特征的关键。下采样用于获取不同尺度的特征张量，同时扩展特征映射的感受域，使模型能够提取到更加抽象的语义信息。在解码器中，block5 和 block6 分别包含一个上采样层和两个卷积层。原始 U-Net 的上采样操作一般通过反卷积运算进行实现。但是对于医学图像融合任务而言，反卷积操作很容易导致融合结果中出现棋盘状伪影。因此，DDIFN 中生成器的上采样通过双线性插值来实现。相比于最近邻插值和双三次插值，双线性插值可以在不增加反向传播计算复杂度的情况下获得理想的上采样效果。另外，此改进措施还在很大程度上减少了网络参数量，不仅降低了模型的训练复杂度，还能有效避免因样本数量少而导致的模型过拟合。

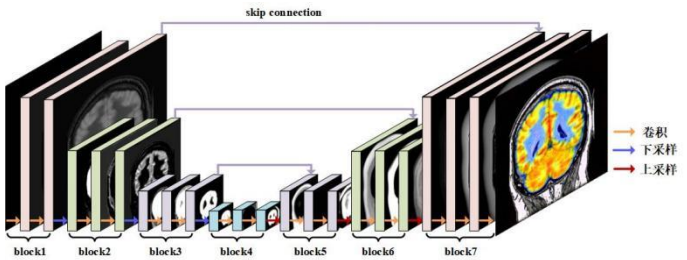


图 3　生成器网络结构

二、鉴别器

为了避免融合模型只着重保留某一种模态的特征而弱化其他特征，DDIFN 设计了两个对称的鉴别器 D_1 和 D_2，这能够在很大程度上平衡不同模态特征的提取及融合程度。具体而言，D_1 的目标是通过不断地训练优化，能够将源图像 a_i 和各阶段的融合图像 f_i 进行准确分类。与 D_1 相同，D_2 的任务是尽可能地将源图像 b_i 和不同阶段的融合图像 f_i 进行区分。如图 4 所示，鉴别器 D_1 和 D_2 均由三个卷积层和一个全连接层组成，且卷积核大小均为 3×3。为了避免较高神经网络层在反向传播过程中出现梯度消失，鉴别器同样在每次卷积后对特征张量进行批归一化操作。对于整个 DDIFN 来说鉴别网络属于高层网络，无需进一步强调特征的稀疏性，所以鉴别网络应用激活函数 LeakyReLU 进行非线性映射。

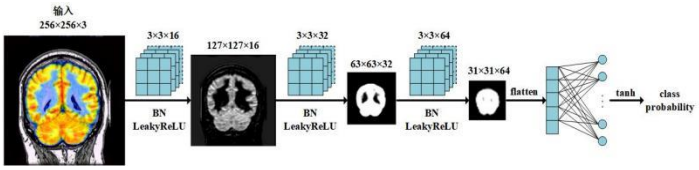


图 4　鉴别器网络结构

三、实验结果

DDIFN 模型在 MR-T1/PET 图像、MR-T1/MR-T2 图像、MR-T2/CT 图像融合结果在纹理细节和像素活动信息保留方面取得了较其他对比方法更好的结果，如图 5 所示。

本文[3]于 2023 年 3 月发表在多媒体权威期刊 TOMM。

责任编委　崔海楠

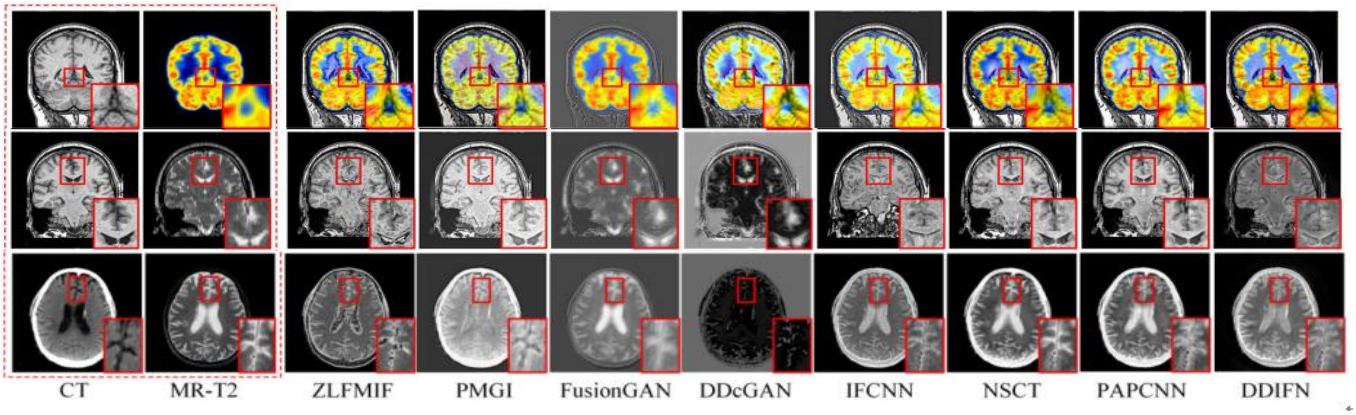


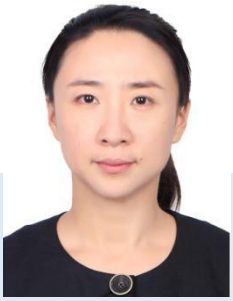
图 5　本文方法与对比方法的融合效果

参考文献

[1] Ma J., Yu W., Liang P., etc. FusionGAN: A generative adversarial network for infrared and visible image fusion[J]. Information Fusion, 2019, 48: 11-26

[2] Ma J., Xu H., Jiang J., etc. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion[J]. IEEE Transactions on Image Processing, 2020, 29: 4980-4995

[3] Liu H., Li S. Zhu J., etc. DDIFN: A dual-discriminator multi-modal medical image fusion network. ACM Transactions on Multimedia Computing, Communications and Applications. 2023,19(4):145.



刘慧

山东财经大学计算机科学与技术学院，二级教授、博士生导师、泰山学者特聘专家

Email:liuh_lh@sdufe.edu.cn

顶会观察

NeurIPS 2023

上海交通大学　陈思衡　雷梓行　魏思哲

神经信息处理系统大会（Conference on Neural Information Processing Systems, NeurIPS）是机器学习领域的顶级会议，是神经计算方面最好的会议之一，在中国计算机学会推荐国际学术会议中被评为人工智能领域的 A 类会议。在 Google Scholar 发布的 2020 年学术指标中，h5 指数高达 309，位于计算机领域的第 2 位、位列所有领域出版物的第 9 位。

今年第 37 届 NeurIPS 于 12 月 10 日到 16 日在美国路易斯安娜州的新奥尔良举办。

一、NeurIPS 2023 的亮点

今年的 NeurIPS 是新冠疫情流行结束后第一届 NeurIPS 大会，参会人数大量增加，大会在巨大的新奥尔良 Ernest N. Morial 会议中心举办，有着巨大的 Poster 展区和多个报告厅，能够容纳数千名参加者。大会采用 Whova App 进行组织，参会者能够从手机实时查询各个报告、研讨会、海报的所在位置和参与嘉宾，同时可以非常轻松地将相关日程加入日历提醒，方便安排会议行程。同时，大会主办方也鼓励大家通过 App 进行社交活动，自由地创建群组，并寻找潜在的合作者。

除了论文成果展示之外，NeurIPS 的一大特色是精心组织的社交活动。这些社交活动有着丰富多彩的目标人群，例如女性研究者、交叉背景的研究者、医疗健康从业人群等等。这些社交活动提供了一些非正式的场合让各种身份的研究者们互相了解，分享经验，增进友谊。

二、论文录用情况

NeurIPS 2023 共有 13330 篇论文被提交，创历史新高。最终有 3540 篇论文入选，录取率不到 26.6%，略高于去年的 26.1%。其中 77 篇被录用为 Oral，400 篇被录用为 Spotlight，Oral 和 Spotlight 的录取率仅有 0.578%和 3.00%。

在 NeurIPS 2023 ，据不完全统计，谷歌位于所有高校/机构的榜首，入选论文高达 180 篇。斯坦福大学与麻省理工学院并列排名第二，共有 130 篇论文入选。卡耐基梅隆大学位居第三，共有 112 篇论文入选。国内高校/机构中，清华大学排名第一，共有 111 篇论文入选，排在国内外所有高校/机构的第 4 位；北京大学第二，共有 98 篇论文入选，排在国内外所有高校/机构的第 5 位。入选论文较多的国内机构还有中国科学院、上海交通大学、香港中文大学、浙江大学和中国科学技术大学。

NeurIPS 作为机器学习领域的顶级会议，收录的论文包罗了人工智能领域的各种主题，包括大语言模型，人工智能生成内容（AIGC），深度学习及其应用，强化学习和规划，计算机视觉，纯理论研究、概率方法、优化，机器学习和社会，神经科学和认知科学等方方面面。特别是在今年生成式人工智能的浪潮下，大语言模型以及相关工作受到了广泛的关注。

三、邀请报告

NeurIPS 2023 共有七个邀请报告 (Invited Talk)，主题丰富，涵盖了火热的生成式人工智能，负责任的人工智能，神经认知科学，机器学习系统，大语言模型和强化学习与数字医疗等，反映了机器学习正与其他学科不断相互影响并相互交融；同时探讨了人工智能大潮下的多个社会问题，展现了科研人员的社会责任感。慕尼黑大学的正教授 Björn Ommer 介绍了他对规模扩张的

错觉与生成式人工智能的未来的看法。Google Research 的研究科学家 Lora Aroyo 呼吁机器学习系统应当考虑内容的固有模糊性和人类观点的自然多样性，应当建立对文化意识和以社会为中心的研究，关注数据质量和数据多样性的影响，用于训练和评估机器学习模型，并在不同的社会文化环境中促进负责任的人工智能部署。印第安纳大学布卢明顿分校的杰出教授，认知科学和认知发展领域国际公认的领导者 Linda Smith 报告了她采用复杂系统的视角，旨在理解感知、运动和认知发展在产后前三年间的相互依赖性的研究。加州大学伯克利分校电气工程与计算机科学系的教授 Jelani Nelson 介绍了数据草图这一概念。数据草图对内存进行压缩的总结，但仍然允许回答有用的查询；作为一种工具，在算法设计、优化、机器学习等领域都有所应用。斯坦福大学人工智能实验室的副教授 Christopher Ré 介绍了他关于基础模型的研究，着重描述了一种基于经典信号处理的新型架构。哈佛大学统计与计算机系的教授 Susan Murphy 介绍了她的小组在开发在线强化学习（RL）算法以应用于数字健康干预中所面临的一些挑战及初步解决方案，用于帮助那些正在努力应对诸如物质滥用、高血压和骨髓移植等健康问题的患者。在“Beyond Scaling Panel”，来自 Google Brain 的 Aakanksha Chowdhery、康奈尔大学的 Alexander Rush、Meta AI 的 Angela Fan、清华大学的唐杰教授以及斯坦福大学的 Percy Liang 分享了他们对大模型的看法和思考。

四、会议热点论文

本次会议涌现了许多优秀的工作，它们具有非常高的学术价值与应用价值。NeurIPS 2023 共有 6 篇获奖论文，其中包含杰出论文（2 篇）、杰出论文亚军（2 篇）和杰出数据集和基准论文（2 篇）。值得一提的是，六篇获奖论文中有四篇是关于大语言模型，这也凸显了研究人员们对这一领域的重视。荣获杰出论文大奖的两篇论文分别是自 Google Deepmind 团队的 Privacy Auditing with One (1) Training Run、斯坦福的 Are Emergent Abilities of Large Language Models a Mirage?。

Privacy Auditing with One (1) Training Run. 这项工作提出了一种只需要单次训练来检查隐私机器学习系统的方案。相比现有方法动辄需要数百个训练模型，论文中提出的方案仅需单次训练，其基于差分隐私机器学习系统能够独立添加或删除多个训练示例的并行性的特点，分析了差分隐私和统计泛化的联系，避免了群体隐私的成本。该方法对算法的假设要求很低，可以应用于黑盒或者白盒环境。

Are Emergent Abilities of Large Language Models a Mirage? 斯坦福的研究人员们在这篇工作中针对大语言模型的“涌现”能力提出了一种新的解释。针对“涌现”能力的突现性和不可预测性，该文认为在特定任务和模型中研究者选择了特定的度量标准。具体来说，非线性或者不连续的度量会产生明显的“涌现”能力，而线性或者连续度量则会产生平滑、连续、可预测的模型性能变化。该文通过三种互补的方式针对包括 GPT-3 系列中声称具有“涌现”能力的任务进行了检验，以及展示了如何选择度量标准从而在视觉任务中创造“涌现”能力。该文证明“涌现”能力与度量或者统计标准有关，而不是人工智能的基本属性得到了扩展。

杰出数据集论文奖由 ClimSim: A large Multi-scale Dataset for Hybrid Physics-ML Climate Emulation 获得。该数据集由来自于加州大学尔湾分校、哥伦比亚大学、英伟达等 20 个机构的气候科学家和机器学习研究人员共同发布。文章中发布的数据集 ClimSim 是一个用于混合物理-ML 气候仿真的大型多尺度数据集，将物理学与机器学习相结合引入了新一代更高保真度的气候模拟器，通过机器学习模拟器执行计算密集、短时、高分辨率的任务，从而绕开摩尔定律。借助该数据集，有望提高诸如风暴等气象预测的准度与精度，造福人类社会。

杰出基准论文奖颁给了 DECODINGTRUST: A Comprehensive Assessment of Trustworthiness in GPT Models。该基准由伊利诺伊大学香槟分校、斯坦福大学、加州大学伯克利分校等机构共同发布，其提出了一套针对大语言模型的全面可信度评估框架，包括毒性、刻板印象、对抗性攻击鲁棒性、分布外鲁棒性、对抗性示范鲁棒性、隐私、机器伦理和公平性等维度，并

发现了之前未被公开的有关信任度的漏洞。该基准框架的发布有助于 GPT 模型在更多领域的应用。

杰出论文亚军分别是来自 Hugging Face、哈佛大学、图尔库大学的 Scaling Data-Constrained Language Models 和来自斯坦福大学与 CZ Biohub 的 Direct Preference Optimization: Your Language Model is Secretly a Reward Model。前者探讨了在数据有限的情况下大语言模型的扩展，后者介绍了一种改进大语言模型行为以符合人类偏好的新方法——直接偏好优化。

除了六篇获奖论文之外，今年的时间检验奖颁给了 Distributed Representations of Words and Phrases and their Compositionality。该论文由十年前还在 Google 从事研究的 Tomas Mikolov，Ilya Sutskever，Kai Chen，Greg Corrad，Jeffrey Dean 等人撰写，被引量已经超过了 4 万次。论文作者团队中 Greg 与 Jeffrey 也来到现场做了演讲，并分享了这篇工作诞生的经历。官方给出的颁奖理由是：这项工作引入了开创性的词嵌入技术 word2vec，展示了从大量非结构化文本中学习的能力，推动了自然语言处理新时代的到来。可谓实至名归。

NeurIPS 2023 一共收录了 77 篇 Oral 论文和 400 篇 Spotlight 论文，大会将 Oral 论文分成了 22 场 talk，包括：RL, Datasets & Benchmarks, Tractable models, DL Theory, Efficient Learning, Objects/Neuroscience/Vision, Causality, Privacy, Neuro, NLP / Tools, Diffusion Models, Optimization, COT / reasoning, GNNs / Invariance, Privacy / Fairness, Probability / Sampling, Vision, LLMs, Theory 等方面。

Oral 论文仅仅是录用论文的冰山一角。其他录用论文中还有大量值得探索和讨论的内容，毕竟 word2vec 并不是那一年的 Oral 或者 Spotlight，甚至其原始论文还曾遭遇 ICLR 拒稿，可见非 oral 的录用论文中也是藏龙卧虎。

海报展览在应用、深度学习、机器学习、优化、概率方法、强化学习、社会影响、理论等八大领域中展开，分为六个时间段进行。每个时间段仅有两小时展示时间，让我们难以全面、细致地了解每一篇研究论文。虽然 NeurIPS 偏重理论研究，但人工智能应用领域的研究同样激发了研究人员的浓厚兴趣。像 ChatGPT 这样的成功案例更是吸引了众多研究者，促使他们思考如何将尖端技术应用于实际。本次大会中两个尤为引人瞩目的方向是 AI Safety 和 AI Agent。在主旨演讲以及多个 workshop 上呼吁更多的研究者能够关注到大语言模型时代愈发突出的 AI 安全性问题，主题包括如何避免偏见、与人类对齐、数据安全和攻击防护等诸多方向。AI agent 则是另外一个引人注目的方向，年初斯坦福大学的 AI 小镇风靡全球，也吸引了更多的研究者关注到这个方兴未艾的领域。在 Foundation Models for Decision Making Workshop 中，来自全球各地的研究者分享了他们在 AI agent 方面的令人激动的研究成果。例如，Percy Liang 教授从 AI 小镇开始，描述了在多智能体社会中信息“Diffusion”，给我们多智能体交互的研究甚至是社会学研究提供了新的思路和方向。Liang 教授研究组关于 research AI 的研究也非常有趣，在大语言模型时代，我们也许应该将我们的思维从“如何做一件事”转向“如何让 AI agent 学会帮我做这件事”。CMU 的 Ruslan Salakhutdinov 教授介绍了他的研究组关于 Web AI agent 的研究，通过增加模型多模态能力和思维链能力，AI agent 可以不断进化，帮助人类完成许多日常的任务，例如订机票，交电费，搜索汇总资料等等。AI agent 很有可能在未来几年内深刻影响人类的数字生活，为每个人都配备善解人意、随叫随到的“贾维斯”。学科交叉也是 NeurIPS 的鲜明特征，这次的 NeurIPS 上有许许多多的交叉学科研究。有很多的研究者创造性地将大模型应用在各个领域，包括用于自动驾驶的 DriveGPT，用于地理领域的 GPT4GEO 等等。神经科学与 AI 的交叉也受到了广泛的关注，大会中有数篇论文关注到了通过 Diffusion Model 将 MRI 信号解码成视觉图像，帮助我们更深刻地理解人类大脑。如同 AlphaFold 深刻地影响了结构生物学，随着 AIGC 日新月异的发展，越来越多的领域有可能借助 AI 的力量，找到新的方向和机遇。

除此之外，大会还举办了 14 个教程（Tutorials）和 58 个研讨会（Workshops），包括语言模型、世界模型、Diffusion Model 等丰富的领域。这些教程和研讨会的安排非常紧凑，在同一时间段内甚至有多场供参会人员选择，为科研人员扩展视野和交流互动提供了便利的渠道和平台，利于机器学习领域的长期发展。

五、总结与展望

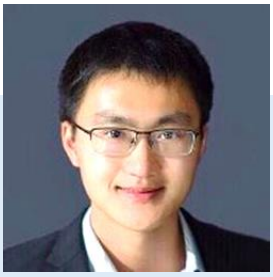
在 NeurIPS 2023 录用的 3584 篇论文关键词中，“Bayes”（174 次）、“Gaussian”（195 次）出现的频次仍然很高，说明经典的方法与理论研究仍然在该会议中受到青睐。与此同时，一些时下热门研究方向在 NeurIPS 上也能够大放异彩。据不完全统计，“Generative”、“Transformer”、“Agent”、“Zero/Few-shot”等词语分别出现在 508、271、280、211 篇接收论文关键词中，在会场中也能感受到 LLMs、Diffusion Models 等领域受到了广泛的关注。此外，尽管是一场机器学习领域的会议，视觉领域和机器人领域的工作也非常多，在教程和研讨会上有相当数量的参与者，共同讨论着各个应用领域的热门话题。

参与疫情后首个全面面对面的会议，与来自世界各地的研究者进行直接的沟通与交流，这种体验远胜于线上虚拟会议。面对面的对话和休闲闲谈在会场中激发了更多的科研创意，也促成了与全球顶尖科研人才的合作。因此，尽管线上会议有其成本效益，但面对面会议带来的益处是不言而喻的。

责任编委 王金甲

参考文献

[1] https://neurips.cc/virtual/2023/papers.html?mode=detail&filter=titles
[2] https://blog.neurips.cc/
[3] https://voxel51.com/blog/neurips-2023-and-the-state-of-ai-research/



陈思衡

上海交通大学未来媒体网络协同创新中心长聘副教授，国家高层次人才青年项目。研究方向为多智能体协作学习。团队信息见网站 https://siheng-chen.github.io。
Email: sihengc@sjtu.edu.cn

雷梓行

上海交通大学硕士生，师从陈思衡教授。研究兴趣包括智能体交互，具身智能，更多信息详见：https://chezacar.github.io/。
Email: chezacarss@sjtu.edu.cn

魏思哲

上海交通大学硕士生，师从陈思衡教授、张娅教授。研究方向为自动驾驶、协作感知。个人网站：https://sizhewei.github.io。
Email: sizhewei@sjtu.edu.cn

西安电子科技大学邓成教授访谈

2024 年 3 月 5 日，《CCF-CV 专委简报》在线采访了西安电子科技大学博士生导师邓成教授。下面是采访实录。

邓老师，您好！您的研究主要集中在多模态数据感知计算与分析推理等领域，能否介绍一下您在这些领域中最突出的几项研究成果？

无论是从学术界还是产业界，目前多模态感知计算与分析推理都是大家关注的一个热点内容，尤其近年来，以 ChatGPT、文心一言为代表的多模态大模型极大地推动了多模态学习相关技术的发展。我们的研究也是受到了领域内许多前辈和同行的启发和帮助。具体来说，关于多模态信息检索，我们设计了基于样本关系约束以及自监督学习的系列多模态检索算法，应该算是较早地利用深度学习进行多模态检索算法开发的团队之一；关于多模态内容生成，我们利用条件对抗生成网络，设计了基于草图的分阶段自然图像生成算法以及多句辅助对抗的细粒度文本到图像生成；关于多模态内容理解，我们设计了基于分层语义关联的视频时序定位算法以及基于非对称跨注意力引导的演员及其在视频中行为的像素分割框架等系列算法；针对多模态网络安全，我们设计了领域内首个针对汉明空间检索的对抗样本生成算法以及针对跨模态学习的对抗样本生成算法等。基于以上研究内容，我们也和阿里等高新头部企业合作，陆续完成了一些应用项目。

作为知名中青年学者，您获得了国家级高层次人才、国家百千万人才工程入选者和国家有突出贡献中青年专家等，您认为这些荣誉对您的学术研究有什么样的影响？

这些荣誉提醒我要肩负起更大的责任：要持续探索新的研究方向，要勇于挑战更加前沿的科学问题；吸引和培养更优秀的年轻学者和研究生，推动学术领域的发展；将研究成果用于解决“卡脖子”核心技术难题，以科学研究服务于更广泛的社会需求。

近年您发表了多篇顶刊或顶会学术论文，而且您的论文具有很高的引用次数，能跟大家分享一下您是如何做到持续产出高水平论文的呢？您在学术影响力方面做了哪些努力呢？

国内外有很多团队都比我们做得要好，我们也一直在摸索前行。针对不同的学生，我倾向于给学生较大的选择空间，在有较好的研究意义和应用前景的前提下，我会尽可能尊重他们的研究兴趣，并鼓励他们做出有价值和特色的工作。在实际过程中，尤其是针对低年级的同学，我会最大程度地参与进去，在论文的研究动机、解决框架、写作逻辑等方面尽可能地提供建议和意见，帮助学生能够快速成长。至于论文的引用和影响可以看作是研究质量的一个反馈。我们能做的就是关注那些相对比较重要的问题，提高研究的质量。

可否请您谈一下在第三代人工智能时代，计算机视觉将如何发展？面临哪些挑战？哪些研究方向会特别有价值呢？

第三代人工智能强调了数据和知识的双轮驱动，为人工智能的发展带来新的思路和方案。目前，自然语言领域的大模型取得了极大成功，在未来，我觉得属于计算机视觉领域的大模型也会取得突破性进展。同时计算机视觉领域的研究可能会在数据的隐私和安全、算法的可解释性、开放世界的泛化能力、资源和能耗绿色化等方面面临挑战。相对应地，我认为多模态学习、小样本和迁移学习、自适应和在线学习、能效优化、可信机器学习等将会是比较有价值的研究方向。

这些年，您的研究方向也一直都在与时俱进，请问您是如何发展和开拓新的研究方向的呢？

我们正处在一个飞速发展的时代，科技发展日新月异，作为一线的科研工作者一定要学会去拥抱变化，敢于尝试和迎接新的挑战，同时也要时刻关注国家的重大发展战略和产业变革趋势，根据国家的发展需求开展有组织的科研。我最初读博士期间的研究方向是图像水印，后来参加工作后，面对大数据时代海量多模态数据的挑战，积极转型主攻多模态智能信息处理。后来，又从数据驱动的人工智能转向认知驱动的人工智能，探索更加前沿、更加具有挑战的研究课题。现在，面向国家“AI+”的战略需求，我们也开始尝试人工智能与多学科间交叉的“AI+”研究。如果我们抱着一个研究方向固步自封，很容易就被时代所抛弃，成为时代的一粒沙，随风飘散。

作为教育部电子信息类教学指导委员会秘书长，您是如何充分发挥教学指导委员会对高等教育改革的研究、咨询、指导作用的呢？

在教育部高教司和郝跃院士的指导下，我主要做好服务，搭好平台，让各路教育专家能够聚在一起，共同探讨和推动教育改革。我们经常组织一些讨论会，邀请不同高校的老师们分享他们的经验和见解，特别是关于如何使用新技术新方法来提升教学质量。同时，我也很注重收集和分析数据，确保我们的建议和决策都是基于实际的研究成果。此外，国内各高校积极承办各种研讨会和论坛，这不仅有助一线教师了解最新的教育趋势，也为大家提供了和教育教学大咖交流和学习的机会。当然，我们也很看重教师和学生的反馈，因为他们是教育改革的直接受益者。通过这些努力，希望能够为中国高校的电子信息类人才培养和教育改革做出实质贡献。

您获评第十三届陕西省“教学名师”，这一荣誉是您实干苦干，甘为人梯的结果，可否分享您在教书和育人方面的心得？

在教书育人方面，我的原则是始终坚持学生为中心。我相信，每个学生都是独特的，所以我会尽力了解他们的兴趣和需求，然后根据这些信息来调整教学方法。同时，我也很重视培养学生的批判性思维和独立解决问题的能力，而不仅仅是传授知识。作为一名教师，不仅要教书，更要以身作则，传递正确的价值观和人生观。我总是鼓励学生们追求卓越，但同时也教导他们要有责任感，将社会需要和自我发展结合起来。简而言之，我认为教育不仅是知识的传递，更是人格的塑造，这一直是我教书育人的核心理念。

根据您一路走来的经历，对于高校青年教师，您有没有好的成长经验给大家分享一下呢？

新入职的青年教师一定要走好自己工作的前五年，规划好前十年。刚参加工作的前五年对青年教师来讲至关重要，一定要走稳、走扎实。要将个人的发展与学院、学校的发展以及学科的建设相融合，抓住新机遇，积极主动融入到学院和学校的事业发展中去，为个人、为学院和学校谋求更好更快发展。既得做好科学研究不断取得新的突破，还得努力站稳讲台，使自己快速成长为一名真正的全面发展的人民教师。在脚踏实地的同时，我们还得学会仰望星空，思考十年之后要成为一个怎样的

人，在科研上取得什么成果，在教学上达到什么水平。唯有仰望星空，方能笃定前行。

如果吐露研究工作者的心声，您最想说的是什么？

希望科研工作者们能够关注人工智能领域的核心和基础问题，针对国家和社会的重大需求开展原创性研究，共同推进我国的科技创新和产业发展。

责任编委 **赵振兵 余烨**

邓成

西安电子科技大学教授、博士生导师。国家级高层次人才，国家百千万人才工程入选者，国家有突出贡献中青年专家，陕西省重点科技创新团队负责人，陕西省青年科技奖获得者(标兵称号)。担任 2018-2022 教育部电子信息类教学指导委员会秘书长。中国计算机学会杰出会员、IEEE 高级会员、中国图象图形学会高级会员。长期从事多模态感知计算与认知推理的研究工作。主持包括国家重点研发计划重点专项、国家自然科学基金重点项目近 30 项。近 5 年，在本领域国际一流期刊和 CCF A 类会议上发表论文 200 余篇。现担任领域国际著名期刊 Pattern Recognition、Neurocomputing 等副编辑，以及几个刊物专辑的客座编辑；担任国际顶级会议 CVPR、ICCV、NeuIPS 领域主席。连续多年入选爱思唯尔中国高被引学者榜单，入选斯坦福全球 2%顶尖科学家榜单。研究成果获陕西省自然科学一等奖 2 项、国家自然科学二等奖 1 项。

委员好消息

✪ 2024 年 1 月 9 日，2023 年 CCF 夏培肃奖评奖结果发布，CCF-CV 专委会顾问委员会委员、北京航空航天大学**王蕴红**被授予 CCF 夏培肃奖。

✪ 2024 年 1 月 30 日，首届上海杰出人才名单发布，CCF-CV 专委会执行委员、复旦大学**姜育刚**入选。

✪ 2024 年 1 月 25 日，2023 年度 CAAI 教学成果激励计划入选项目公示，CCF-CV 专委会 10 位执行委员的成果入选：北京理工大学**付莹**“红色基因传承、多维融合创新：新工科人工智能卓越人才培养模式探索与实践”、上海大学**曾丹**“面向智能时代的电子信息类人才培养模式创新与实践”、北京工业大学**毋立芳**、**马伟**、**简萌**“目标导向、教研融合，以学生为中心的数字图像处理（双语）课程改革与实践”、北京航空航天大学**王蕴红**、**张永飞**“融通式人工智能创新型人才本研一体化培养模式探索与实践”、四川大学**胡鹏**、**彭玺**“师生二元长效机制驱动的人工智能类 人才培养模式改革与实践”入选一类成果，东北大学**贾同**“人工智能驱动的自动化专业新工科改革探索与实践”入选一类成果。

✪ 2024 年 1 月 27 日，2023 CCF 颁奖大会上，**CCF-CV 专委会**获优秀专业委员会奖。

✪ 2024 年 1 月 27 日，2023 年度湖北省科学技术奖拟奖项目公示，CCF-CV 专委会常务委员、华中科技大学**白翔**等完成的“面向场景文字检测与识别的深度学习方法研究”拟授自然科学一等奖。

✪ 2024 年 3 月 7 日，2023 年度吴文俊人工智能科学技术奖名单公布，CCF-CV 专委会 15 位执行委员的项目获奖：北京大学**林宙辰**“深度学习网络设计与优化算法”、上海交通大学**卢策吾**“行为视觉理解”、北京航空航天大学**徐迈**“视觉感知启发的视频质量优化理论与方法”获自然科学一等奖，山东师范大学**朱磊**“领域自适应迁移学习”获自然科学二等奖，中国科学院自动化研究所**赫然**、**董晶**、**谭铁牛**“网络虚假视觉信息智能感知技术及应用”获技术发明一等奖，浙江大学**章国锋**、**周晓巍**“基础模型与知识融合的复杂电力巡检视觉智能分析关键技术及应用”、北京理工大学**杨健**“多场景图像智能分析关键技术及应用”获科技进步一等奖，北京交通大学**朱振峰**“场景适配的一站式数智化服务系统关键技术及应用”获科技进步二等奖，中国人民解放军军事科学院**赵健**获优秀青年奖，清华大学**鲁继文**指导的《结构信息引导的图像超分辨率重建方法研究》、西安电子科技大学**邓成**指导的《基于特征关系挖掘的度量学习算法研究》获优秀博士学位论文，**林宙辰**等指导的《自监督对比学习的理论与算法研究》、东南大学**耿新**指导的《基于标记分布学习的 K 近邻分类》获优秀博士学位论文提名。

✪ 2024 年 3 月 22 日，2023 中国电子学会科学技术奖授奖名单发布，CCF-CV 专委会 5 位执行委员的项目获奖：西北工业大学**韩军伟**、**程塨**、**张鼎文**“面向不完备图像数据的模式挖掘理论与方法”获自然科学一等奖，北京科技大学**殷绪成**“面向大规模数据的 Angel 机器学习平台关键技术及应用”获科技进步一等奖，上海交通大学**林巍峣**“网络适配的视频智能传输及调度优化关键技术与应用”获科技进步二等奖。

✪ 2024 年 3 月 26 日，美国国家人工智能科学院终身院士名单公布，CCF-CV 专委会常务委员、百度计算机视觉首席架构师**王井东**入选。

责任编委 刘海波

基于深度学习的去雨方法及其开源代码

香港城市大学 付陈平 大连理工大学 樊鑫

单幅图像去雨任务是典型的底层视觉问题，该任务旨在从雨图中恢复出干净的图像。由于干净图像和雨条纹是不可知的，因此单幅图像去雨任务是不适定的逆问题。早期传统方法通常设计、施加各类雨条纹和清晰图像先验解决该问题。然而，这类手工设计先验的方式难以适用于复杂多变的真实降雨场景。近年来，相较于传统方法，基于深度学习的去雨方法取得了更好的泛化性，引起了研究人员的广泛关注。鉴于此，本文总结并介绍了最近几年流行的去雨方法及其开源代码。

1、DRSformer 方法

介绍: 本文提出一种基于 Transformer 网络的去雨方法，称为 DeRaining network, Sparse Transformer (DRSformer)。DRSformer 可自适应地保持特征聚合中最有用的自关注值，使聚合的特征可更好地从雨图中恢复出高质量的清晰图像。接下来，对 DRSformer 去雨方法进行简单介绍。

如图 1 所示，本文开发了一个可学习的 top-k 选择算子，自适应地从每个查询的键中保留最重要的关注分数，以更好进行特征聚合。此外，由于基于 Transformer 的前向网络不能模拟对潜在清晰图像恢复至关重要的多尺度信息，本文进一步开发一种有效的混合尺度的前向网络，以更好恢复图像去雨特征。结合卷积神经网络算子的局部上下文信息，本文在模型中加入混合专家特征补偿器，并提出一种合作细化训练方案，以学习一组丰富的混合特征集。

图 2 展示了本文所提方法与其他去雨方法的部分定量结果。可以看出，相较于现存的去雨方法，DRSformer 可以在去除雨线的同时更好保留了图像细节，展示出较为优越的去雨性能。

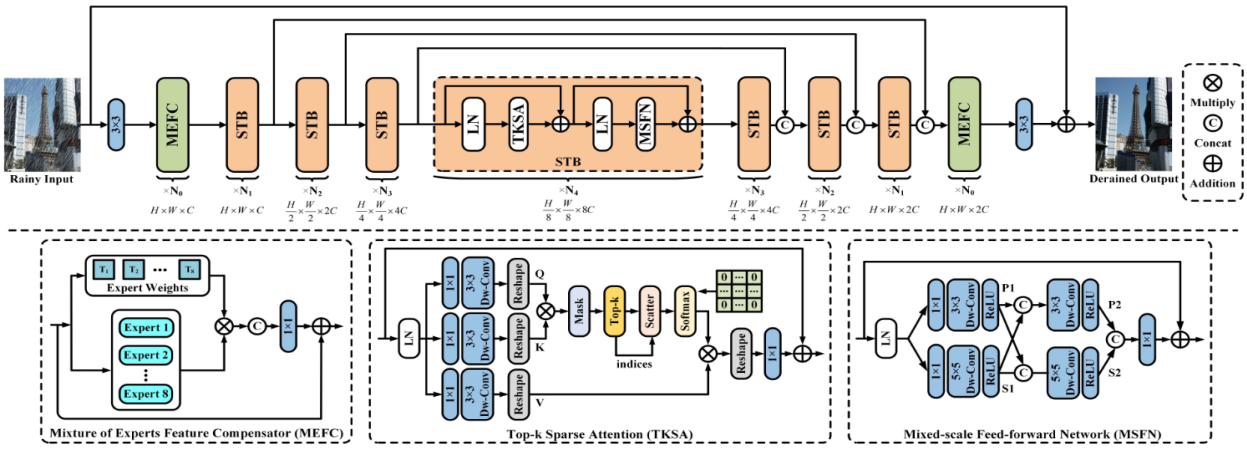


图 1 DRSformer 去雨方法框架图

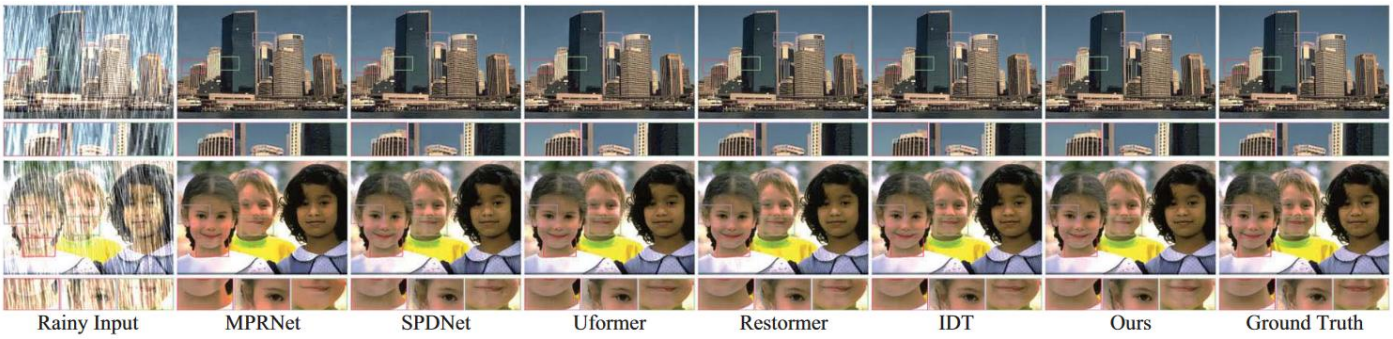


图 2 DRSformer 去雨方法部分定量结果展示

论文地址：
https://ieeexplore.ieee.org/document/10204775
代码地址：
https://github.com/cschenxiang/DRSformer

2、HCT-FFN 方法

介绍：本文提出一种轻量级基于卷积神经网络和 Transformer 网络的特征融合去雨方法，称为 HCT-FFN。所提方法 HCT-FFN 利用卷积神经网络和 Transformer 网络各自的学习优势，通过协调学习这两类网络，将雨图像恢复为干净图像。接下来，对 HCT-FFN 去雨方法进行简单介绍。

如图 3 所示，本文在基于卷积神经网络的阶段堆叠一系列退化感知混合专家（DaMoE）模块，其中适当的局部专家自适应地使 HCT-FFN 模型强调空间变化的降雨分布特征。在基于 Transformer 网络的阶段，采用背景感知视觉 Transformer（BaViT）模块对图像的空间长特征依赖进行补充，在保留所需结构的同时实现全局纹理地恢复。此外，由于卷积神经网络特征和 Transformer 网络特征间存在不确定的知识差异，本文在相邻阶段引入一个交互融合分支，从而有利于重构高质量的训练结果。

图 4 展示了本文所提方法与其他去雨方法的部分定量结果。可以看出，相较于现存的去雨方法，HCT-FFN 展示出较为优越的去雨性能。

论文地址：
https://ojs.aaai.org/index.php/AAAI/article/view/25111
代码地址：
https://github.com/cschenxiang/HCT-FFN

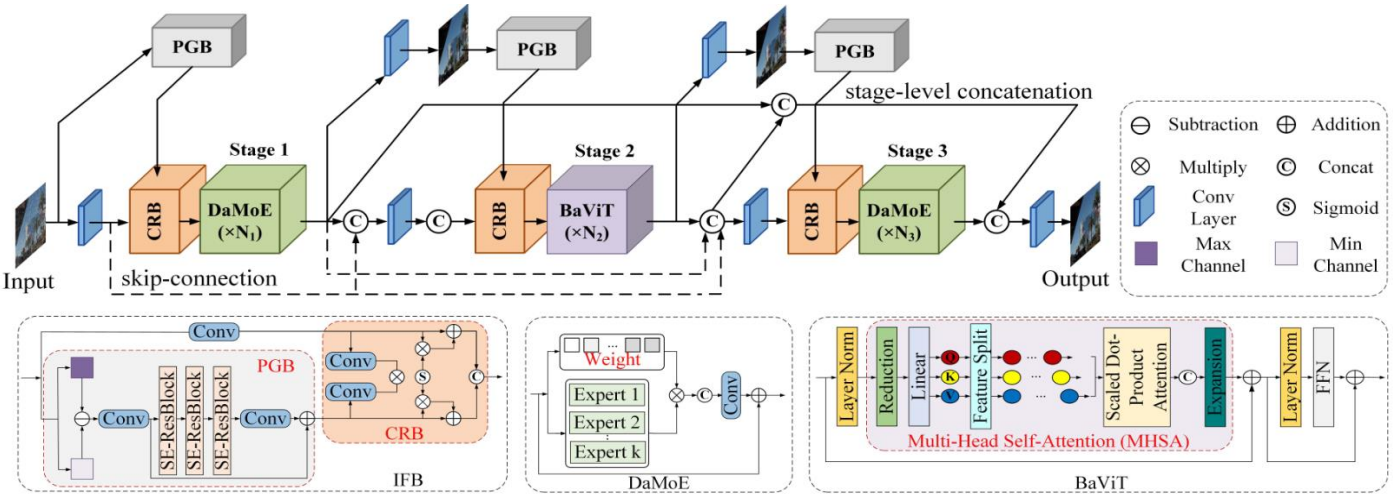


图 3 HCT-FFN 去雨方法框架图

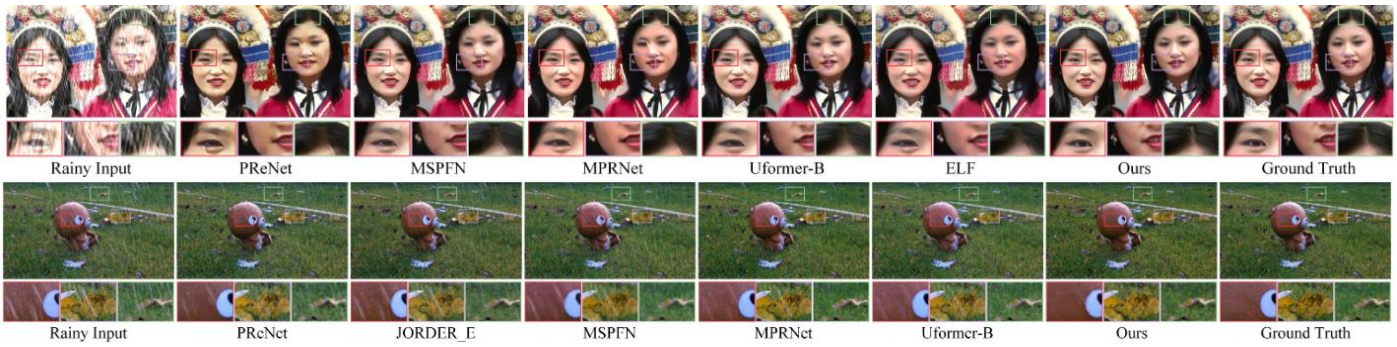


图 4 HCT-FFN 去雨方法部分定量结果展示

3、DCD-GAN 方法

介绍：本文提出一种非配对的单幅图像去雨对抗方法，称为 DCD-GAN。该方法通过深度特征空间中的双重对比学习方式探索未配对示例的相互属性，主要包括两个协同分支，即双向翻译分支(BTB)和对比引导分(GCB)。接下来，对 DCD-GAN 去雨方法进行简单介绍。

如图 5 所示，BTB 充分利用对抗一致性的循环架构，生成丰富的样本对，并通过双向映射挖掘两个主体之间的潜在特征分布。同时，CGB 隐式约束不同样例在深度特征空间的嵌入，使相似的特征分布更接近，而不相似的特征分布更远离，从而更好促进雨线的去除和图像的恢复。

大量实验表明，DCD-GAN 方法在合成和真实数据集上都优于现有的非配对去雨方法。此外，相较于全监督或半监督的去雨模型，所提方法同样显示出较好的去雨性能。

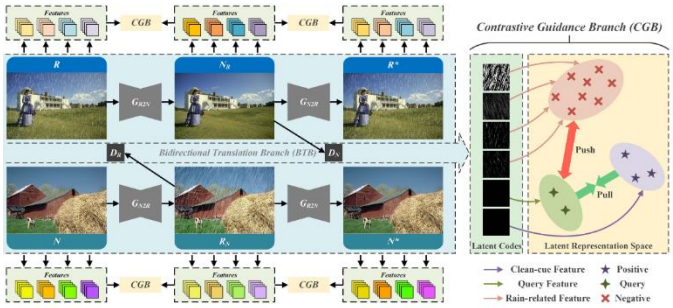


图 5 DCD-GAN 去雨方法框架图

论文地址：
https://ieeexplore.ieee.org/document/9880261
论文代码：https://cxtalk.github.io/projects/DCD-GAN.html

责任编委 李策 贾同

付陈平

博士后，香港城市大学计算机科学学院，研究方向为计算机视觉，目标检测，图像增强。

樊 鑫

博士生导师，大连理工大学国际信息与软件学院从事教学与科研工作，担任软件学院院长。研究方向为计算机视觉与图像处理、医学影像分析。

个人主页：http://faculty.dlut.edu.cn/Xin_Fan/zh_CN/index.htm

无人机集群数据集

沈阳航空航天大学 王传云 苏阳 高赛

无人机集群具有数量多、成本低、尺寸小、灵活性强、协同工作等特点，无人机集群作战正朝着智能化、实战化迅猛发展，逐渐成为未来战场上重要的作战样式。

针对无人机集群的反制理论、技术与系统的研究与开发逐渐受到学术界、工业界和军方的高度关注。利用计算机视觉探测与跟踪来袭的无人机集群，成为反制无人机集群的基础理论和技术支撑之一。然而，作为研究基准的无人机集群数据集尚未见公开报道，制约了无人机集群反制技术创新与系统研发进展。

本文重点介绍了一个无人机集群数据集-UAVSwarm Dataset。

1、UAVSwarm Dataset 简介

UAVSwarm Dataset 提供了地对空无人机集群和空对地无人机集群两种拍摄视角，并且摄像机状态包含静止状态和运动状态，这使得背景具有复杂而动态的变化。每个图像序列中的无人机都包含不同的运动模式，例如：离开视野、进入视野、形态变换和快速运动等。

UAVSwarm Dataset 包含 72 个图像序列，共 12598 张图像，包含 13 种不同真实世界场景和 19 种不同型号的无人机，每个图像序列中的无人机个数由 3 至 23 架不等。

图 1 UAVSwarm Dataset 中部分无人机示例

图 2 UAVSwarm Dataset 中部分图像示例

2、UAVSwarm Dataset 注释规则

为了确保 UAVSwarm Dataset 中图像注释的一致性，本数据集使用边界框尽可能准确地标记每个图像序列中的无人机，并严格遵循以下图像注释规则：

(1) 在每个视频序列中，无人机目标都会尽早被标记，并尽可能晚地结束。换言之，如果无人机目标在视野内，并且可以清楚地确定其路径，则可以保留 ID。

(2) 在每一帧中，标记所有型号和所有姿态的无人机目标。

(3) 在每一帧中，目标的边界框应包含属于该无人机目标的所有像素，边界框应尽可能靠近无人机目标。

(4) 如果可以指定无人机目标的确切位置，则始终按顺序注释。如果遮挡很长，并且无法使用简单的推理(例如，恒定速度假设)来确定无人机目标的路径，则在无人机目标再次出现后，为之分配一个新的 ID。

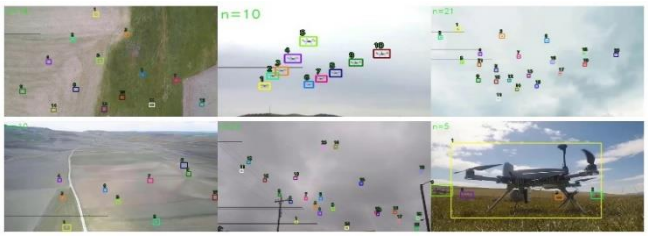


图 3 UAVSwarm Dataset 中部分注释图像示例

3、UAVSwarm Dataset 数据格式

为了更好的评估多目标跟踪算法，本文构建的 UAVSwarm Dataset 的数据格式和 MOT16 的数据格式保持一致。同时，所有图像均为 JPEG 格式，并以 6 位数的文件名命名(例如：000001.jpg)。

(1) 检测文件数据格式

为了更加关注多目标跟踪的跟踪性能，本文将 YOLOX 算法的测试结果用作检测文件。检测文件是逗号分隔值文件(Comma-Separated Values，CSV)，每行表示一个无人机对象，每行包含 10 个值。第 1 个数字表示对象出现在第几帧中；第 2 个数字表示该对象的 ID(因为 ID 尚未指定，所以在检测文件中 ID 都设置为‘-1’)；第 3、4、5、6 个数字表示无人机边界框在二维图像坐标中的位置，分别表示左上角坐标 x 值、y 值、边界框的宽度和高度；第 7 个数字表示对象的置信度分数；第 8、9、10 个数字表示忽略状态(‘1’表示忽略此对象；‘-1’表示不忽略此对象。本文设置为‘-1’)。检测文件内容示例如下：

1，-1，174，243，12，12，1，-1，-1，-1

1，-1，215，326，13，14，1，-1，-1，-1

1，-1，235，167，13，14，1，-1，-1，-1

1，-1，273，250，11，12，1，-1，-1，-1

(2) 注释文件数据格式

注释文件是 CSV 文件，每行表示一个无人机对象，每行包含 9 个值。前 6 个数字含义与测试文件的前 6 个数字含义相同。第 7 个数字表示对象的置信度分数，并用作是否考虑输入的符号(‘0’表示在计算中忽略此对象；‘1’表示此对象被标记为活动。在本文的注释文件中都设置为‘1’)。第 8 个数字表示对象的类别(由于在 UAVSwarm Dataset 中只识别无人机，因此都设置为‘1’)。第 9 个数字表示每个边界框的可见性比率(范围从 0 到 1，这取决于另一个静态或移动目标的遮挡程度或图像边界的剪裁，本文设置为‘1’)。注释文件示例如下：

1，1，352，20，11，11，1，1，1

2，1，352，20，11，11，1，1，1

3，1，352，20，11，11，1，1，1

4，1，352，20，11，11，1，1，1

此外，为了获得整个基准测试的有效结果，本文为遵循上述格式的每个图像序列创建了一个单独的 CSV 文件，命名方式为‘Sequence-Name.txt’。

4、多目标连续鲁棒跟踪算法

利用 UAVSwarm Dataset，对新提出的一种无人机集群多目标 (UAVS-MOT) 连续鲁棒跟踪算法进行了验证，该算法基于 FairMOT 模型的多分支无锚框预测结构，将坐标注意力模块与 DLA-34 网络相结合，构建了全新的主干特征提取网络以提升特征信息的表达能力。此外，引入全新的 ArcFace Loss 损失函数进行训练以提高模型的收敛速度，并利用 BYTE 数据关联方法以降低目标漏检率和提高轨迹的连贯性。

无人机集群多目标跟踪算法 UAVS-MOT 在 UAVSwarm Dataset 上的多目标跟踪准确度 (MOTA) 和目标识别准确度 (IDF1) 分别为 73.4%与 76.1%，相比原有 FairMOT 算法分别提升 5.7%与 2.9%，可以解决目标的漏检、误检和跟踪精度低的问题，鲁棒性好。

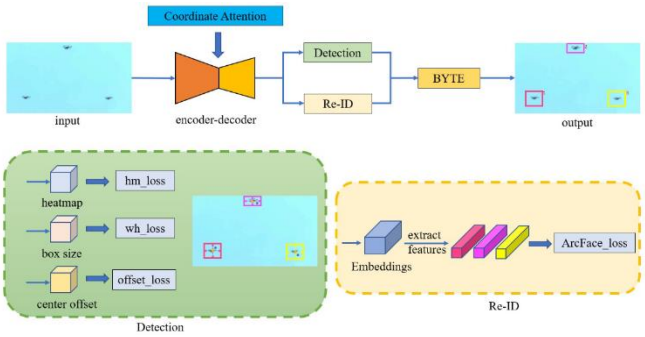


图 4 面向反制无人机集群的多目标连续鲁棒跟踪算法

5、UAVSwarm Dataset 相关资源

数据集下载地址：https://github.com/UAVSwarm/UAVSwarm-dataset/tree/master

相关论文链接：(1) UAVSwarm Dataset: An Unmanned Aerial Vehicle Swarm Dataset for Multiple Object Tracking，https://www.mdpi.com/2072-4292/14/11/2601

(2) 面向反制无人机集群的多目标连续鲁棒跟踪算法，https://hkxb.buaa.edu.cn/CN/10.7527/S1000-6893.2023.29017

责任编委 王田 李策

王传云

沈阳航空航天大学副教授，主要研究方向为模式识别、智能博弈、反无人机等。Email：wangcy0301@sau.edu.cn

苏 阳

沈阳航空航天大学硕士研究生，主要研究方向为无人机集群、多目标跟踪、轨迹分析等。Email：suy0970@163.com

高 骞

沈阳航空航天大学讲师，主要研究方向为快速三维重建、深度学习几何学等。Email：gaoqian@buaa.edu.cn

好 文 推 荐

清华大学自动化系和新畅元科技公司“Animatable Gaussians: Learning Pose-dependent Gaussian Maps for High-fidelity Human Avatar Modeling”最新成果发表在 CVPR-2024。

论文：Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. "Animatable Gaussians: Learning Pose-dependent Gaussian Maps for High-fidelity Human Avatar Modeling." CVPR 2024.

该文提出了一种利用 2D CNNs 和 3D Gaussian splatting 来创建高保真人体模型的方法。该方法通过学习从输入视频中得到的参数化模板，将 3D 高斯与可动画 avatar 相关联。该模板对穿着的衣服具有自适应性，可用于建模例如连衣裙等较宽松的衣服。这种基于模板的 2D 参数化是一种新的虚拟形象表示方法，将显式的 3D 高斯溅射引入虚拟形象建模中，能够使用基于 StyleGAN 的 CNN 来学习姿态相关的高斯映射以建模动态外观。

该论文的方法包含两个主要步骤，算法的具体框架如图 1:

1.学习参数化模板。从输入视频中选择一个近似 A 姿势的帧，通过 SMPL 蒙皮和基于 SDF 的体积渲染，优化规范的 SDF 和颜色场以适配多视角图像。然后，提取模板网格。随后将 SMPL 顶点的蒙皮权重扩散到模板表面，获得一个可变形的参数化模板。

2.学习依赖于姿势的高斯图。首先通过线性混合蒙皮(LBS)将模板变形到姿势空间，并获得两个位置图。位置图作为姿势条件，通过 StyleUNet 转换为前后高斯图。然后提取有效的 3D 高斯体，将规范的 3D 高斯体变形到姿势空间。最终，我们通过基于溅射的可微分光栅化将姿势 3D 高斯体渲染到给定的摄像机视角。

论文使用 THuman4.0、ActorsH 和 AvatarReX 数据集进行实验。结果表明，该论文提出的方法与隐式 NeRF-based 方法相比具有更高度动态、逼真和通用的外观。基于提出的模板引导参数化和姿势投影策略，使该论文的方法不仅可以重建详细的人类外貌，还可以为新姿势合成生成逼真的服装动态。总体而言，目前该论文的方法优于其他先进的虚拟形象建模方法，同时也相信该论文提出的基于 3D 高斯分布的虚拟形象表示方法将推动 3D 人体表示领域取得更大进展。

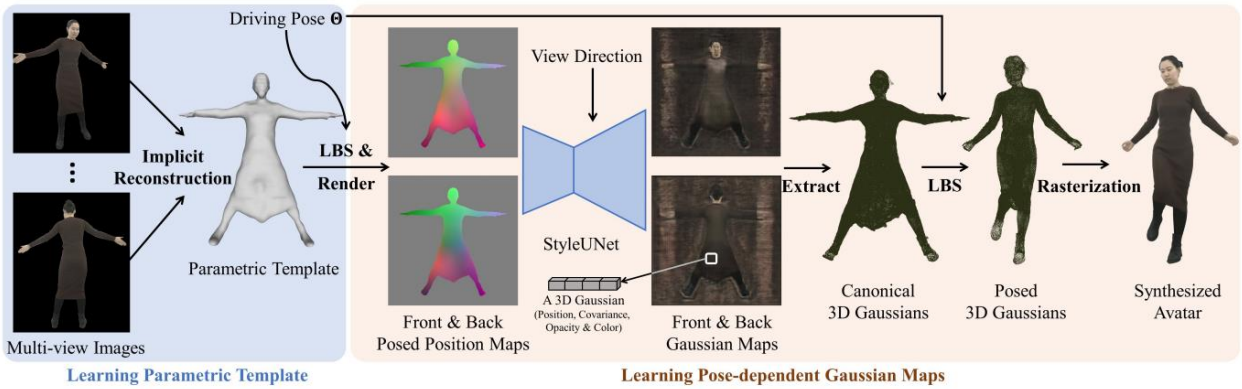


图 1 Animatable Gaussians 方法的具体框架图示

责任编委 樊鑫 贾同

好 文 推 荐

清华大学的最新成果“Probabilistic Contrastive Learning for Long-Tailed Visual Recognition”发表在 IEEE TPAMI 2024。

论文：Chaoqun Du, Yulin Wang, Shiji Song, Gao Huang. Probabilistic Contrastive Learning for Long-Tailed Visual Recognition, IEEE TPAMI, 2024

在真实世界的数据中，长尾分布是十分常见的现象。这意味着数据集中少数类别所拥有的样本数量远远超过了大多数类别。以生态系统分类任务为例，一些常见物种（如鸽子或麻雀）的观察实例可能多达成千上万，而一些稀有物种（如某些特定种类的猫头鹰或蝴蝶）的观察实例可能不足十个。这种类别数据不平衡问题显著地影响了标准监督学习算法的性能，因为这些算法通常是针对相对平衡的训练集设计的。它们在学习过程中往往会优化对多数类的预测性能，而忽视对少数类的准确识别。因此，针对不平衡数据，需要一种能够有效处理的学习算法，以确保对少数类的重视程度与其在实际问题中的重要性相匹配。

最近的研究揭示，自我监督对比学习在缓解数据不平衡方面展现出了巨大潜力。尽管自我监督对比学习有很多优势，但它的性能却存在一个固有的局限性：自我监督对比学习需要足够大的训练数据批次来构造覆盖所有类别的对比对，但在类别数据不平衡的环境下满足此要求是非常困难的。为了克服这一障碍，文中提出了一种新颖的概率对比（ProCo，probabilistic contrastive）学习算法，如图 1 所示。该算法旨在估计特征空间中每个类别样本的数据分布，并据此抽样对比对。然而，在实际应用中，对于不平衡数据，使用小批量特征来估计所有类别的分布是不可行的。文中的关键思想是引入一个合理且简单的假设，即在对比学习中的归一化特征遵循单位空间上的冯·米塞斯-费希尔（vMF，von Mises-Fisher）混合分布，这带来了双重好处。首先，分布参数可以仅使用第一个样本的矩估计，这使得可以通过不同批次的在线方式高效计算；其次，基于估计的分布，vMF 分布允许抽样无限数量的对比对，并推导出期望对比损失的封闭形式，以便高效优化。除了应对长尾问题外，ProCo 还可以直接应用于半监督学习，通过为未标记数据生成伪标签，随后可以用来逆向估计样本的分布。理论上，文中分析了 ProCo 的误差界。实证上，广泛的实验结果在监督/半监督视觉识别和对象检测任务上表明，ProCo 一致地超越了现有方法，在不同数据集上都取得了优异的表现。

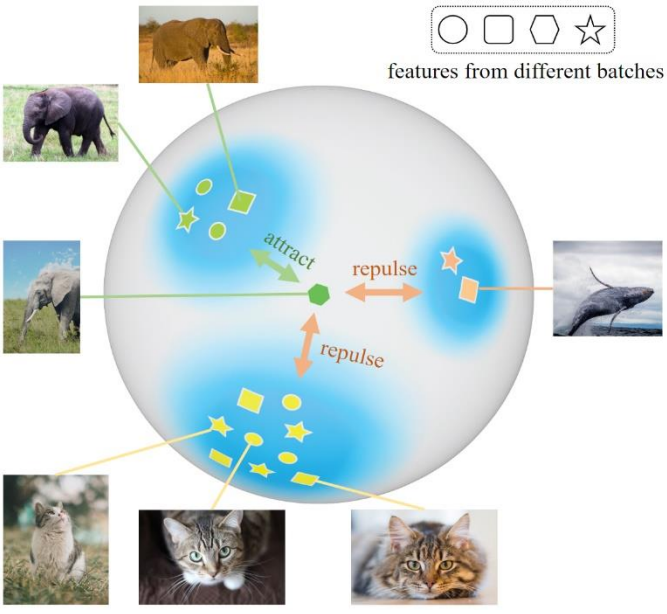


图 1 ProCo 总体框架示意图

责任编委 贾同 王田

好 文 推 荐

国防科技大学的最新成果“Few-Shot Domain-Adaptive Anomaly Detection for Cross-Site Brain Images”发表在 IEEE TPAMI 2024。

论文：Jianpo Su, Hui Shen, Limin Peng and Dewen Hu, Few-Shot Domain-Adaptive Anomaly Detection for Cross-Site Brain Images, IEEE TPAMI, 2024

近年来，静息态功能磁共振成像（rs-fMRI，Resting-State Functional Magnetic Resonance Imaging）成为精神科医生诊断复杂精神障碍的重要工具。功能连接（FC，Functional Connectivity）是根据 rs-fMRI 数据中自发性血氧水平依赖（BOLD，Blood Oxygen Level Dependent）信号波动的时间相关性而提取的，能够描述大脑内不同区域的功能交互模式。基于机器学习的研究表明，FC 可以有效地区分精神病患者和健康对照组。大多数基于 FC 的精神障碍诊断分类使用有监督机器学习算法，但在实际应用中，很难获得足够数量的正确标记的患者样本。理想情况下，研究者期望以一种无监督的方式完成患者的诊断和亚型分类。然而，FC 具有较高的特征维度，单个站点（地点或机构）中样本数量有限，和 FC 中显著的个体间差异，会显著增加过度拟合噪音的风险。

异常检测可作为一个替代方法，虑将健康对照者的公共数据集作为源领域，将每个站点中的数据作为目标域，使用少样本学习来解决不同标签空间的问题。由于源域和目标域的成像数据来自不同的人群、不同的 MRI 机器和不同的扫描参数，因此所采集的 fMRI 图像之间存在一定的内在差异。此外，源域和目标域的标签空间不相同，阻碍了传统域自适应方法。文中提出了 Few-shot 域自适应异常检测（FAAD）算法，如图 1 所示，利用来自多站点 fMRI 图像的 FC 作为特征，将患者从健康对照者中筛选出来。1）提取不同脑区的 BOLD 时间序列，计算 Pearson 相关系数作为 FC 并向量化。2）使用源 fMRI 数据的 FC 进行网络的预训练，采用三层自编码器，根据重构损失进行训练。3）对每个站点的数据采用三次重复的三次验证策略。4）添加残差校正块和条件对抗域自适应来补偿源数据和目标数据之间的域差异。总损失由异常检测损失 L_{ad} 和域自适应损失 L_{da} 组成。实验证明该方法基源数据和目标域数据的少量标记样本能进行有效的检测，并展现出算法的优越性和鲁棒性，在可视化分析中也体现出了算法的生物学真实性。

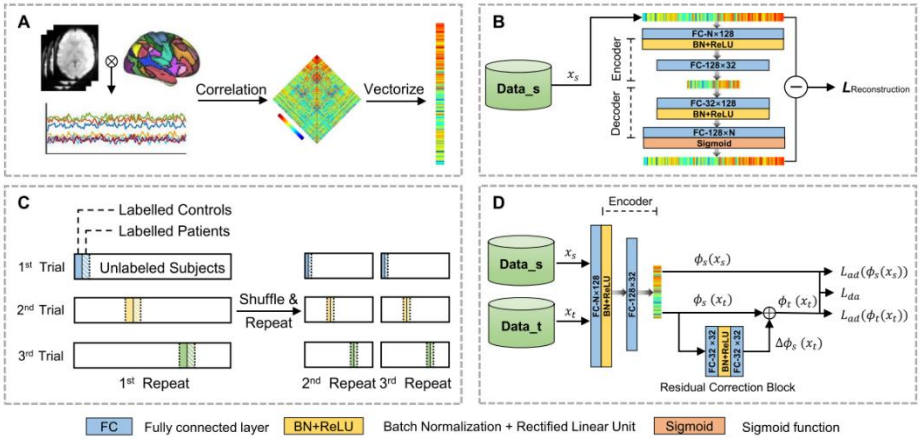


图 1 所提方法模型

责任编委 李策 樊鑫

征文通知

1 会议征文

计算机视觉领域相关国内外会议的征文通知如表 1 所示。同时，可继续关注每个会议举办的 workshop 或 special session。

2 期刊征文

计算机视觉领域近期相关期刊专刊的征文通知如表 2 所示，包括 IEEE Journal of Biomedical and Health Informatics, Neural Network 和 Computer Vision and Image Understanding。

3 会议简介

中国模式识别与计算机视觉学术会议 PRCV (Chinese Conference on Pattern Recognition and Computer Vision)，由中国计算机学会（CCF）、中国自动化学会（CAA）、中国图象图形学学会（CSIG）和中国人工智能学会（CAAI）联合主办，定位国内顶级的模式识别和计算机视觉领域学术盛会。

第七届 PRCV 将于 2024 年 10 月 18 日至 10 月 20 日在乌鲁木齐举办，由新疆大学承办。本届会议旨在汇聚国际国内模式识别和计算机视觉领域的广大科研工作者及工业界同行，分享最新理论研究进展和技术研发成果。通过此次会议，能加强本领域学术界和企业界进行深入的“产学研”交流与合作，从而进一步促进模式识别与计算机视觉领域的协同创新。

责任编辑 **刘帅奇**

表 1 计算机视觉领域相关国内外会议

会议名称	会议时间	会议地点	截稿日期	会议网站
MM 2024	2024.10.28-11.01	Melbourne, Australia	2024.04.13	https://2024.acmmm.org/
ECAI 2024	2024.10.19-24	Santiago Compostela, Spain	2024.04.26	https://www.ecai2024.eu/
Siggraph Asia 2024	2024.12.03-06	Tokyo, Japan	2024.05.20	https://asia.siggraph.org/2024/
CoRL 2024	2024.11.06-09	Munich, Germany	2024.06.07	https://www.corl.org/

表 2 计算机视觉领域相关国内外期刊专刊

期刊名称	专刊题目	投稿网址	截稿日期
CVIU	Advances in Deep Learning for Human-Centric Visual Understanding	https://www.sciencedirect.com/journal/computer-vision-and-image-understanding	2024.04.30
CVIU	Trustworthy Cross-Modal Reasoning for Video-Language Understanding	https://www.sciencedirect.com/journal/computer-vision-and-image-understanding	2024.04.30
NN	Brain-inspired Neural Networks for Biomedical Signal Processing	https://www.editorialmanager.com/neunet	2024.04.15
JBHI	Domain Adaptation and Generalization for Biomedical and Health Informatics	https://www.embs.org/jbhi/wp-content/uploads/sites/18/2024/02/JBHI_Domain_Adaptation_SI.pdf	2024.05.31

COMPUTER VISION NEWSLETTER

01 2024
总第 39 期

计算机视觉专委会简报

CCF 计算机视觉专委会