顶会观察

ICCV 2023

西澳大学 武子杰
湖南大学 王耀南

国际计算机视觉大会（IEEE/CVF International Conference on Computer Vision，ICCV）是计算机视觉和模式识别的顶级会议之一，与 CVPR 和 ECCV 并称为计算机视觉领域三大顶会。ICCV 会议不仅是中国计算机学会推荐的人工智能领域 A 类国际学术会议，还位列 Core Conference Ranking 的 A*类推荐会议，H5-index 高达 228，Impact Score 高达 32.51，录用率在 20%-30%之间，在 CV 界具有极高的评价。ICCV 每两年召开一次，与 ECCV 穿插进行。不同于在美国每年召开一次的 CVPR 和只在欧洲召开的 ECCV，ICCV 在世界范围内选址开会。本届 ICCV 大会于 2023 年 10 月 2 日至 2023 年 10 月 6 日在法国巴黎国际会展中心举行，其中前两天为 workshop 和 tutorial，由各发起方自行组织，主会在后三天举办。在主会举办期间，参会者可以面对面的参与被接收论文工作的线下讨论交流。下面本文分别从大会概况、论文录用情况、主题报告、获奖论文的研究工作介绍以及热点报告讲演进行详细的介绍。

一、大会概况

线下线上混合形式：不同于上一次 ICCV2021 的线上虚拟参会，本次 ICCV2023 恢复了线下参会模式，无法线下参会人员仍可选择线上参会，但每个工作至少需要一个作者注册方式为亲自参会。1、线下海报展示：为了方便线下的快速交流，本次会议优先考虑以海报会议为中心的面对面互动讨论的方式。海报形式：对于线下参会的作者，与 2019 年以前一样海报张贴；对于线上参会的作者，主办方提供了官方海报打印合作服务方便参会者们灵活选择。2、线上论文会议：除了线下面对面沟通渠道外，大会还允许作者准备一个五分钟的预录制视频和一张海报的文件，在会议平台上展示工作。为了方便交流，主办方还为每篇论文安排了线上、线下、同步和异步方式与参会者进行充分的交流。

严格的评审机制：今年会议组织方邀请了 311 位专家作为领域主席(area chair)。同时，组织方还邀请了 6990 位从业者作为审稿人参与论文评审，包括 1320+紧急审稿人。更重要的是，每篇论文会由 3 位 AC。本次会议收到了 25558 份审稿意见，平均每篇论文 3.16 份，其中 175 篇论文收到了五个意见以及 1 篇论文十个审稿意见。所有的最终决定由 AC 共同负责主持审稿线上会议， AC 之间相互审核报告，核查错误，并详细讨论审稿意见，经过多次复核后，最终向作者发出通知。

与会帮助：ICCV2023 致力于通过注册和差旅支持，为来自传统上不参加 ICCV/ECCV 的社区学生提供支持。资金分配将根据需求、对会议的贡献、旅行目的地、自我认同的社区以及顾问支持等因素综合考虑。差旅费资助将根据可用资金和旅行距离以固定金额发放。

二、论文录用情况

ICCV2023 收到的有效投稿和录用数量都有显著提高，大会共收到了 8260 篇有效投稿，最终接收了 2161 篇论文，接收率约为 26.2%。相较于 2021 年，今年 ICCV 的投稿量提升 34.3%（2108 篇），录用率基本保持一致，录用论文数量提升约为 34.1%（549 篇）。其中，有 195 篇论文录用为 Oral Presentations，比去年减少 15 篇，Oral 率约为 9.0%（较去年减少 4.0%）。在被接受论文

中，数量最多的研究领域包括：3D from multi-view and sensors，Image and video synthesis，Transfer/low-shot/continual/long-tail learning，low-level and physical-based vision，vision and language 等。这五个研究领域都有超过 100 篇被录用的论文，其中关于 3D from multi-view and sensors 的论文录用数量接近 175 篇论文。数量最少的研究领域包括 First person (egocentric) vision 和 Optimization methods (other than deep learning) 等。大会也分别统计了各个研究领域的接收率，其中六个领域接受率高于 25%，最高的三个领域依次为 Navigation and autonomous driving，vision and graphics 以及 vision and language，而 Optimization methods (other than deep learning)的接受率仅约 0.6%。整体上，三维视觉以及 AIGC 领域今年收到越来越多的关注。

三、主题报告

本次 ICCV2023 会议邀请了两位 Keynote 演讲者，报告内容围绕于大模型下的互动学习、人工智能对科学发展推动的潜力展开探讨。

Interactive Learning in the Era of Large Models. 斯坦福大学计算机科学系教授 Dorsa Sadigh 报告并讨论了大模型时代机器人系统的交互学习。基础表征在学习人机交互过程中有着至关重要的作用，语言指令和潜在动作能够对机器人操作问题的共享自主权进行赋能，其在辅助机器人领域有着深远的影响，如何在大模型时代，引入对当今机器人系统的交互学习推理成为了一个重要的问题。报告者对此提出了两个关键论点：1）为下游机器人任务引入预训练大模型；2）探索挖掘大模型的丰富语义内容的创造性方法以使能更加匹配的具身 AI 智能体。特别对于预训练方面，报告者介绍了一种以语言为基础的视觉表征学习方法（Voltron），利用语言为机器人预训练视觉表征提供了坚实基础。此外，报告者还介绍了一些关于如何利用大语言模型以及视觉大模型学习人类偏好的实例；其实现了有准确的社会推理，使得机器人系统能够利用纠正反馈机制教导人类。最后，报告者就大模型如何成为有效的模式机器系统话题进行了深入讨论；讲解大模型通过鉴别不变的表示风格，实现模式转换、外推；展示一些关键性的解决控制问题的模式优化证据。

The potential of AI in advancing science and the importance of ensuring AI's responsible use。

谷歌 DeepMind 的研究副总裁、人工智能科学项目的领导者 Pushmeet Kohli 描述了人工智能在推动科学发展方面的潜力以及确保负责任地使用人工智能的重要性。过去几个世纪的科学进步提高了全球许多人的生活水平，然而气候变暖以及新冠大流行带来的巨大挑战证明，还有大量未知领域有待我们去了解。本次演讲中，报告者讨论了人工智能（机器学习）在推动科学发展、提高我们对世界的理解以及预测干预结果的能力方面的潜力。最后，Pushmeet Kohli 还强调了以负责任的方式使用人工智能的重要性，并说明人工智能本身可以帮助实现这一点。

四、最佳论文

大会程序主席团成员逐个宣布了 ICCV2023 的颁奖信息，宣布了今年的马尔奖（Marr prize）的评委成员以及评审过程。本年度大会共评选出了 2 篇论文同时获得最佳论文，1 篇论文获得最佳论文荣誉提名，1 篇最佳学生论文。

最佳论文：Adding Conditional Control to Text-to-Image Diffusion Models[1]，来自斯坦福大学。大型预训练模型的可控生成是该论文的主要研究的问题。本文提出了一种用于为大型预训练文本到图像扩散模型添加空间条件控制的神经网络架构 ControlNet。ControlNet 建立在锁定的训练完成的大型扩散模型之上，重新使用在数十亿幅图片上预训练的大模型作为一个强力的骨干网络去获得多种类条件控制的能力。作者们还提出了一种零卷积层，链接骨干网络，从零开始逐步增加参数，确保没有有害噪声影响微调；引入了多种条件控制实现稳定扩撒生成，如边缘、深度、分割、人体姿势等的有效性。

最佳论文：Passive Ultra-Wideband Single-Photon Imaging[2]，来自多伦多大学。高速成像的一

个基本法则是：高速成像与光密切相关，场景变化越快，则需要越多的光来准确成像，从而不会产生过多的噪点或运动模糊。本文考虑的问题是同时对一个动态场景进行从几秒到几皮秒的极端时间尺度范围内的成像，并且由于是被动成像，没有太多光线供应，也没有来自光源的任何定时信号。由于现有的单光子照相机通量估算技术在这种情况下会出现问题，因此开发了一种通量探测理论，该理论建立在随机微积分之上，能够从单调递增的光子探测时间戳流中重建像素的时变通量。该工作通过被动超宽带单光子成像一次被动捕获动态场景，并允许在 9 个以上数量级的时间范围内重新渲染视频，为从单光子相机中被动采集和处理时间戳流开辟了动态成像的新方向。

最佳论文荣誉提名：Segment Anything[3]，来自 Meta AI 研究院。在网络尺度规模的数据集上预训练的大型语言模型具有强大的零样本和少样本泛化能力，这些基础模型能够泛化到超越可见训练的任务和数据分布本身。因此，本文提出了分割一切（SAM）模型，包含全新任务、模型以及数据集，建立了一个图像分割的基础模型，寻求开发一个可接受提示的模型，并使用一个能够强大泛化的任务在一个广泛的数据集上对其进行预训练。使用提示工程化技术解决新数据分布上的一系列下游分割问题。SAM 将图像分割方向扩展到基础大模型尺度，引领了提示分割新任务、新模型(SAM)以及新数据集(SA-1B)，包含 10 亿个掩码和 1100 万个图像，以促进计算机视觉基础模型的研究。

最佳学生论文：Tracking Everything Everywhere All at Once[4]，来自康奈尔大学、谷歌研究院和加州大学伯克利分校。当前运动估计遵循稀疏特征跟踪和密集光流的方法，虽然都被证明对各自的应用是有效的，但并不能完全模拟视频的运动：成对光流无法捕获长时间窗口内的运动轨迹，稀疏跟踪不能模拟所有像素的运动。本文提出了一种新的测试时间优化方法，用于从视频序列中估计密集和远距离的运动。先前的光流或粒子视频跟踪算法通常在有限的时间窗口内运行，难以通过遮挡进行跟踪并保持估计运动轨迹的全局一致性。因为作者提出了一种完整的、全局一致的运动表示，称为 OmniMotion，它允许对视频中的每个像素进行准确的、全长的运动估计。OmniMotion 使用准 3d 规范体积对视频进行统一表征，通过本地和规范空间之间的双射执行逐像素跟踪，使得模型能够确保全局一致性，克服遮挡跟踪，并对相机和物体运动的任何组合进行建模。该方法能够可以对视频中的每个像素进行准确、全长的运动估计，实现高效的逐像素跟踪。

此外大会还给十年前的 Action recognition with improved trajectories 工作颁发了 Helmholtz 奖项。PAMI Everingham 奖被颁发给了 The Ceres Solver Open Source Nonlinear Optimization Software Library 团队和 The Common Object in Context (COCO) dataset 团队。来自马克斯·普朗克智能系统研究所的 Michael Black 和来自约翰·霍普金斯大学的 Ramalingam Chellappa 荣获了 2023 PAMI Distinguished Researcher Award，来自麻省理工学院的 Ted Adelson 获得了 2023 PAMI Azriel Rosenfeld Lifetime Achievement Award。

另外，还有 13 篇论文入选最佳论文入围名单，其中华人学者为第一作者的论文数量超过半数，并且多篇论文也引起了广泛的讨论。

五、大会奖项

Marr Prize. 该奖项因计算机视觉之父、计算机视觉的先驱、计算神经科学的创始人 David Courtenay Marr 而得名。今年获奖论文由斯坦福大学的论文 Adding Conditional Control to Text-to-Image Diffusion Models 和多伦多大学的论文 Passive Ultra-Wideband Single-Photon Imaging。

Helmholtz Prize. 该奖项以 19 世纪医师和物理学家 Hermann von Helmholtz 命名。该奖项又被称为“时间考验奖”，ICCV 每隔一年颁发一次，旨在表彰十年或更早之前对计算机视觉研究产生重大影响的 ICCV 论文。获奖者由 IEEE 计算机协会模式分析和机器智能技术委员会选出。

PAMI Everingham Prize.该奖项由 IEEE 计算机学会模式分析和机器智能技术委员会每年在国际计算

机视觉会议上颁发，以纪念已故的 Mark Everingham 以及其学术生涯中的杰出表现，并鼓励其他人追随他的脚步，采取行动推动整个计算机视觉社区的进一步发展。该奖项通常颁发给为计算机视觉社区的其他成员做出了无私贡献并带来重大利益的研究人员或研究团队。The Ceres Solver Open Source Nonlinear Optimization Software Library 团队的杰出软件为视觉领域内外许多知名算法提供了支持和 The Common Object in Context (COCO) dataset 团队提供了广泛支持各大计算机任务的数据集，由此，两个团队共同获得该奖项。

PAMI Distinguished Researcher Award. 该奖项被授予其研究项目对计算机视觉进步做出重大贡献的候选人。奖项是根据主要研究贡献以及这些贡献在影响和启发其他研究中的作用而颁发的。候选人由计算机视觉社区提名。

PAMI Azriel Rosenfeld Lifetime Achievement Award. PAMI 终身成就奖旨在表彰在其职业生涯中为计算机视觉领域做出重大贡献的研究人员，以此为了纪念计算机科学家和数学家 Azriel Rosenfeld。

六、精彩报告选介

本次大会精彩纷呈，共有 56 场 workshops，10 场 tutorials。由于篇幅所限，这里仅仅选取最具有代表性的几个精彩分享为例作详细地介绍。

Tutorial on Self-Supervised Learning of Visual Representations. 本场 tutorial 由来自 Meta 的 Xinlei Chen， Kaiming He 以及 Christoph Feichtenhofer 所组织，覆盖了自监督视觉表征学习领域的常用方法和最新进展。同时，掩码自动编码器和对比学习等热门主题也被深入分析讲解。讲演者展示了这些框架如何成功地从二维静态图像和动态视频信息中学习，从机器学习的角度讨论自监督学习。本场 tutorial 展示了自监督学习不同技术之间的联系和区别，并提供有关对计算机社区广受欢迎方法的见解。

Workshop on AI for 3D Content Creation. 如何开发能够大规模生成真实、高质量 3D 数据的算法一直是计算机视觉和图形领域长期存在的问题。能够可靠地合成有意义的 3D 内容的生成模型将彻底改变艺术家和内容创作者的工作流程，并且还将通过“生成艺术”实现新的创造力水平。本场 workshop 汇聚了致力于 3D 形状、人类和场景生成模型的多个研究人员，研究探讨了多个 3D 领域的有趣主题：1.为生成有现实意义的具有纹理和高质量细节 3D 对象，最佳表示是什么？2.对生成的对象进行直观控制的最佳表示是什么？3.如何合成真实的人类执行看似合理的动作？4.如何生成完全可控的 3D 环境，从而可以操纵场景元素的外观及其空间结构？5.生成人与人之间或人与物体之间合理的动态和交互的最佳表示是什么？6.人工生成的 3D 内容会产生哪些道德影响以及我们如何解决这些问题。在最新的研究进展中，独立的 3D 实例生成以及取得了重大进展，最新研究表明有意义的、能与人类动态交互的 3D 生成是未来一段时间的研究重点。

Workshop on what is Next in Multimodal Foundation Models? 当今风靡的大模型已经成为了多种任务的基础模型骨干，其一般指代在大规模数据集上预先训练好的大规模模型 (例如，拥有数十亿个参数)，这些模型可以在很少或没有监督的情况下进一步适应各种下游任务，拥有优秀的泛化能力，极大地推动了计算机视觉、自然语言处理、语音分析等领域的技术发展。特别是多模态基础模型，这种模型同时使用多种模态进行训练，在文本到图像/视频/三维生成、零镜头分类、跨模态检索等广泛应用中取得了显著成功。本次 workshop 讨论了多模态基础模型的下一步发展，研究这一新兴研究领域的前进方向和仍需解决的基本问题。Trevor Darrell 回顾了视觉语言大模型的最新进展，Kristen Grauman 介绍了基于大尺度叙事视频上的多模态 “视频-语言 “学习，Vincent Sitzmann 和 Chuang Gan 总结了大语言基础模型在视觉表示和推理中的关键技术。本场 workshop 对多模态基础模型各个方面都展开讨论包括但不限于模型的设计、泛化特性、效率、伦理、公平性、规模和开放性。

七、总结展望

本年度 ICCV 大会中识别、3D 视觉、图像与视频的生成成为主流，迁移学习、底层视觉等热度保持回升。相比于 2021 年，Transformer 不再作为一个单独的研究重点，而是作为基础骨干网络融入到生成、检测、分割等各个任务当中。ICCV 越来越注重任务驱动、解决真实大场景下的视觉问题，从 2D 到 3D，从语言到实体，从简单到智能，从惊艳到生产力。发展的视觉理论把已经探索认知世界的能力交予智能机器人系统，专注于研究更加注重自驱动的具身智能系统，探索人类未曾甚至难以发现的物理世界规律，这是一场充分解放生产力的新机遇，这些命题帮助深度学习理论迈向更深层次的智能，深入世界的底层逻辑之中，帮助我们更好的完成掌握物质规律、获取世界运转信息的职能。

责任编委 **魏秀参**

参考文献

[1] Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. ICCV2023.

[2] Mian Wei, Sotiris Nousias, Rahul Gulve, David B. Lindell, Kiriakos N. Kutulakos. Passive Ultra-Wideband Single-Photon Imaging, ICCV2023.

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, Ross Girshick. Segment Anything. ICCV2023.

[4] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, Noah Snavely. Tracking Everything Everywhere All at Once. ICCV2023.

武子杰

湖南大学电气与信息学院博士，师从王耀南院士。现为 The University of Western Australia 计算机学院 Research Fellow。主要研究方向为机器人视觉、多模态三维视觉等。
Email: wuzijieeee@hnu.edu.cn

王耀南

中国工程院院士，湖南大学教授、博士生导师、机器人视觉感知与控制技术国家研究中心主任、中国图象图形学学会理事长。主要研究方向为智能机器人技术及应用、控制理论与应用等。
Email: yaonan@hnu.edu.cn