顶会观察

CVPR 2024

南方科技大学 叶顶强 于仕琪

国际计算机视觉与模式识别会议（CVF/IEEE Conference on Computer Vision and Pattern Recognition，CVPR）是计算机视觉和模式识别领域最重要的会议之一。CVPR 于 1983 年在美国华盛顿特区举办，每年举办一次，一般在美国举办。CVPR 2024 于 6 月 17 日至 21 日在美国西雅图举办。

一、 会议概况

CVPR 2024 收到 11,532 篇投稿论文，经过评审后接收了 2,719 篇。投稿论文数比上一年度增加 26%，创历史新高。投稿论文的作者总数是 35,691。组织者对投稿作者的email域名进行了统计，其中cn域名占39%，edu 域名 17%，com 域名 15%，kr 域名 4%，de 域名 3%，其他的少于 3%。

这届会议的参会人数也创了新纪录，共有 12,000 人注册，其中线下注册约 9000 人。来自美国的注册人数 5071 位居第一，其次是中国大陆 1511 人，排第三的是韩国775人，其后德国、加拿大和日本分别是377、352 和 347 人。

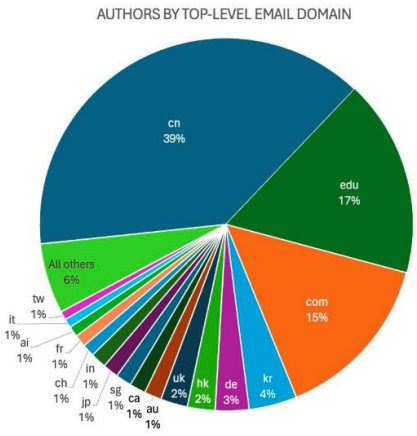


图 1　所有投稿作者的电子邮箱域名统计

会议的前两天 6 月 17 和 18 日是 Workshops 和 Tutorials 时间，CVPR 2024 共组织了 123 个 Workshops 和 24 个 Tutorials。主会论文展示有口头报告和墙报两种形式，论文的口头报告被限制在 8 分钟，且只有较少数量的论文通过口头报告展示。所有主会论文，包括口头报告论文，都会通过墙报方式展示。每场墙报论文展示 400-500 篇论文，历时一个半小时。

二、 参会感受

6 月中下旬是西雅图最好的季节，气温 20 多度，温暖舒适。同时因为临近夏至日，西雅图的白天特别长，当地时间晚上 9:30 天才黑下来，这也为参加会议组织的社交活动提供了方便。毕竟大部分人不愿意在有很多流浪汉的街上夜行。会场是在西雅图市中心的会议中心，会议中心有两栋建筑，相距 5 分钟步行的距离。参加不同的活动，参会人员需要频繁的往返于两栋建筑之间。

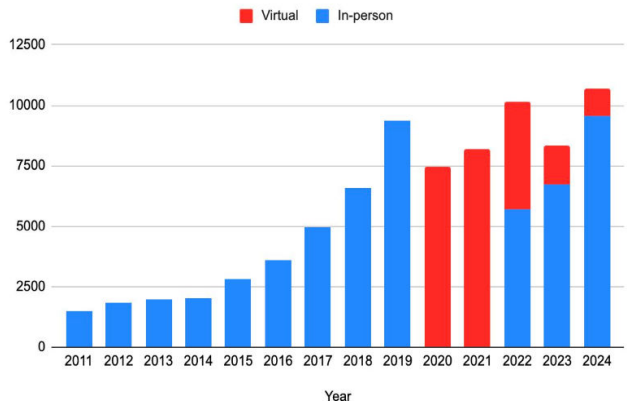


图 2　历年注册人数

参会的最大感受是人多。会场人山人海，无论口头报告会场还是墙报会场，以及展示厅，甚至会场周围的街道上都是戴着 CVPR 会议胸牌的参会者。本文作者居住的 Airbnb 公寓里也张贴了“预警”告示，提示居民旁边的西雅图会议中心有近万人的活动。附近的海底捞火锅店因为招待来自世界各地的华人学者，也是不停地翻台。味道正宗的海底捞，满足了无数久居异乡的华人学者品尝国内美食的愿望。

墙报论文大厅是讨论气氛最热烈的区域。几乎每个展板前都站满了人，在热烈地跟论文作者讨论交流。墙报大厅同时有 400 多篇论文展示，粗粗浏览一遍都会让脚走疼，逐一仔细观看和讨论更是不可能，只能快速定位感兴趣的论文，然后跟作者交流。

因为签证问题，来自中国内地的教授比较少，内地参会人员以学生为主。即便如此，参会学者中华人学者的比例依然非常高，粗略观察约有一半。本文作者于仕琪老师还担任了 OpenCV Foundation 的志愿者，在会议展厅介绍和推广 OpenCV，为 OpenCV Foundation 募捐。在展台讲解中，约 1/3 时间是使用普通话讲解，足见华人在此领域比例之高。

今年CVPR的热点话题毫无疑问的是“内容生成”。只要是涉及内容生成的 Workshop 或 Tutorial，会场都会爆满。个别会场不得不安排工作人员站在门口阻止进入，以防人数过多导致不安全。

在学术交流之外，学术会议还可以具有社交属性。在会议上可以遇到很多朋友和同行，会议结束后晚上还可以在餐馆小聚，大家聊聊近况，探讨一下可能的合作等。对于参会的学生来说，可以了解毕业去向等信息，

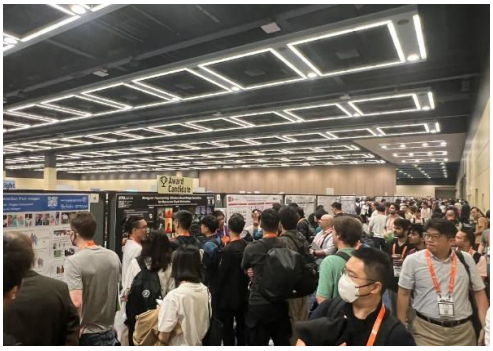


图 3　墙报展区一角

图 4　多模态基础模型研讨会的会场内人山人海

认识未来可以一起共事的人，也为未来的发展收集信息。会议期间除了会议组织的各种社交活动外，一些公司和组织也安排了社交活动，并邀请参会人员参加。本文作者于仕琪老师参加了 Pattern Recognition 期刊的编委会午餐会，蹭得一顿精美午餐；也参加了 Intel 公司组织的 Reception 风格的 Intel Networking Event。西方的 Reception 宴会无固定座位，也无正餐，仅提供食品和饮料，是一种适合自由交流的宴会，这种方式在西方流行但在中国较少。因为美国的食品价格比较贵，Reception 虽不是大吃大喝，但对学生们还是很有吸引力的。

三、 大会获奖论文

会议共选出了 10 篇获奖论文：2 篇最佳论文，2 篇最佳论文候选，2 篇最佳学生论文和 4 篇最佳学生论文候选。其中的 2 篇最佳论文和 2 篇最佳学生论文如下。

Best Paper 1：Generative Image Dynamics[1]：作者是来自谷歌研究院的团队。自然界总是处于运动之中。即使是看似静止不动的物体，其实也存在轻微摆动，如微风抚树，湖光粼波等。人们对物体的真实运动十分敏感。如何用神经网络模型模拟出逼真的物体运动是一大难题。因为这些运动受到不同物体各自独特物理属性的影响，如质量，弹性等。幸运的是，测量这些物理属性不是必须的。比如，只需要分析一些可观察到的二维运动，就能模拟真实场景中的可信运动。在本篇工作中，作者通过从大量真实视频序列中自动提取运动轨迹的方式，为图像空间场景运动（即单张图片中所有像素的运动）建立了一个生成式运动先验模型。具体来说，作者利用扩散式生成模型从每个训练视频中预测出物体

的频谱体积特征。频谱体积特征是一种密集型长距离像素的频域表征，能直接转化为图片的运动纹理，也可以解释为用于模拟动态图像空间的基模态。每次推理扩散模型都从图片中预测某个特定频率的频谱图，不同频谱之间通过共享参数的注意力模块进行协调。这些频谱体积表征可以用来合成未来运动帧，将静态图片变成逼真动画。与传统的基于 RGB 像素的运动先验表征相比，频谱体积表征能够捕捉更细粒度的运动，进而解释长距离的像素变化，生成更连贯和精细的动画。该工作提出的方法可以应用于生成逼真的动画，还能支持多种下游应用，如创建无缝循环，交互式动画等，为自然图像的运动建模提供了有力支持。

Best Paper 2：Rich Human Feedback for Text-to-Image Generation[2]：第一作者是来自加利福尼亚大学圣迭戈分校和谷歌研究院的团队。利用文本指导图片的生成是一个十分有前景的研究方向。它能把人们内心幻想的画面通过文本生成式模型还原出来，对娱乐，艺术，设计和广告等各领域的创作都有巨大帮助。尽管现在的生成式模型取得了很大的进展，但现有方法仍存在一些问题，如扭曲的物体，异常的手指个数，与文本描述不符等。同时已有的评估方式往往是根据图像分布计算，无法反映细粒度的差异。在本篇工作中，作者提出了一个包含丰富人为标注的 RichHF-18K 数据集，其中包含一万八千张生成的图片，图片失真区域的点状标注，与生成图片不对齐的关键词，以及四种细粒度的图片分数（包含可读性，文本图像对齐性，美观度和总体得分）。在此数据集的基础上，作者设计了一个多模态的 Transformer 模型在生成的图片中去预测这些失真区域，错误关键词和四种得分。通过这种巧妙的有监督学习方式，让模型模拟人类的感受，从而对生成的虚拟图片进行评测和优化。该模型可以应用与辅助下游生成式模型，为生成的图片提供可解释的评测标准，从而帮助生成式模型生成更逼真的，与文本相符的虚拟图片。

Best Student Paper 1：Mip-Splanting: Alias-free 3D Gaussian Splatting[3]：作者是来自图宾根大学、上海科技大学等单位的团队。新视角图片合成技术在计算机图形学和计算机视觉中发挥着至关重要的作用，其应用涵盖虚拟现实，电影拍摄，机器人等。除了热门的基于多层感知机（MLP）表征物体形状和独立视角的 NeRF 技术外，3D 高斯溅射（3DGS）技术因其可在高分辨率下实时渲染的优点在最近收获到很多关注。3D 高斯溅射技术是指利用一组 3 维空间中的高斯云来代表物体，通过基于溅射的光栅化渲染方法将高斯云投影到 2 维屏幕空间以便合成新视角图片。现有的 3D 高斯溅射方法在图片进行放大和缩小时会出现伪影问题。作者提出该问题的根源是缺少对 3D 频率的约束以及使用 2D 膨胀滤波。具体来说，拉远镜头会导致投射到屏幕上的 2D 高斯变小，如果还使用相同的膨胀量，就会导致伪影。拉进镜头则相反，投射的 2D 高斯会膨胀导致生成的图片变形。为了解决以上问题，作者提出了对 3 维空间中 3D 特征的正则化方法。首先引入 3D 平滑滤波来正则化 3D 表征的最高频率从而去除拉近镜头产生的伪影。原理是生成图片的最高频率是继承于训练数据，是满足 Nyquist-Shannon 采样定律的。同时这个平滑滤波器将成为场景特征的固有一部分。其次作者通过使用 2D Mix 滤波器替换 2D 膨胀滤波来解决镜头拉远导致的混叠和扩张伪影。该方法使得 3D 高斯溅射成像技术可以在改变成像采样率，镜头焦距和相机距离的情况下仍生成逼真的图片。

Best Student Paper 2：BioCLIP: A Vision Foundation Model for the Tree of Life[4]：作者是来自俄亥俄州立大学的团队。现如今利用计算机视觉来回答生物问题仍然是一个艰巨的任务，因为在训练模型前需要依赖专业的生物学家对其感兴趣的特定任务数据进行昂贵的手工标注。现有的 CLIP 和 GPT-3 等基础模型展现出强大的零样本 (zero-shot) 泛化能力，对样本外的数据有很强的适应力。因此作者提出一个设想，借鉴上述模型的设计也构建一个基于自然生物界的视觉基础模型，来大大降低人工智能应用于生物学的门槛。为了满足实际生物学任务的需求，作者认为模型需要符合以下几点需求。首先应该尽可能的泛化到整个生命树，确保能支持不同的生物工作者的研究。同时除了已知的训练类群，模型还能够泛化到未知生物类群中。其次模型应该学习生物图片的细粒度表征，因为生物学常常会区分外观相似的生物，如同属中的近亲，物种的伪装色等等。最后由于自然生物数据的收集以及标注都是十分

昂贵的，因此模型的少样本泛化能力十分关键。基于此，作者提出了 TREEOFLIFE-10M 数据集，其中包含 1 千万张图片，涵盖生命树中的 45 万个类群。每一张训练图片都被尽可能的细分类别等级，以及生命树中的更高分类等级。作者还提出了 BIOCLIP 模型，一个借鉴了 CLIP 对比学习方式的专门用于生物学的基础模型。通过对比学习方法，能够让模型学习到复杂的，多层级的生物分类方式。作者将该模型在域外的数据集上进行零样本迁移测试，均显著超过以往模型。该工作的提出，将应用在帮助生物工作者们更好的进行科研研究，如物种划分，个体识别，性状检测，种群结构测定以及生物多样性保护等。

四、 总结展望

人工智能领域热度持续提升，ChatGPT 在文字方面的成功极大地刺激了生成式智能的发展，特别是图像生成和视频生成方面的研究热情。这一点在 CVPR 2024 会议得到了充分体现。因为图像格式与文字不同，在图像生成特别是视频生成方面，尚没有出现一个“杀手锏”式的高效率框架。图像和视频的数据规模远超过文字，数据的标签更加模糊，科研的成本也更高；同时科研领域的“马太效应”也愈加明显，赢家通吃。在这一趋势下，科研模式必然发生变化，如何开展科研也许是我们需要思考的问题。

责任编委 王金甲

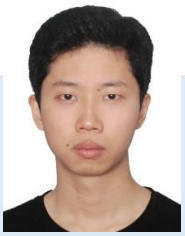
参考文献

[1] Zhenggqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative Image Dynamics. In Computer Vision and Pattern Recognition (CVPR), pages 24142-24153, 2024.

[2] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, el at. Rich Human Feedback for Text-to-Image Generation. In Computer Vision and Pattern Recognition (CVPR), pages 19401-19411, 2024.

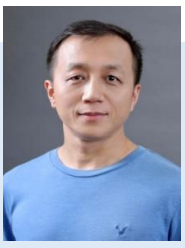
[3] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andres Geiger. Mip-Splatting: Alias-free 3D Gaussian Splatting. In Computer Vision and Pattern Recognition (CVPR), pages 19447-19456, 2024.

[4] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Caryln, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BIOCLIP: A Vision Foundation Model for the Tree of Life. In Computer Vision and Patttern Recognition (CVPR), pages 19412-19424, 2024.



叶顶强

南方科技大学计算机科学与工程系硕士生，主要研究方向步态识别。CVPR 2024 录用论文为 BigGait: Learning Gait Representation You Want by Large Vision Models，该论文尝试使用视觉大模型来提升步态识别，取得了显著效果。
Email: 11810121@mail.sustech.edu.cn



于仕琪

博士生导师，南方科技大学计算机科学与工程系副教授，研究方向为步态识别和视觉目标检测。在步态识别方面，创建的 CASIA-B 步态数据库目前被作为本领域的评估标准，是使用最广泛的评估库之一；所创建的 OpenGait 开源项目已经成为步态识别领域主要的算法评估框架。在目标检测方面，人脸检测算法被世界排名前 100 的多家上市公司采用，同时也被众多的中小企业广泛使用。在遥感图像处理方面获 2021 年度广东省科学技术奖自然科学奖二等奖。现担任中国图形图像学学会监事，Board Member of OpenCV Foundation，Vice President on Education of IEEE Biometrics Council，Vice Chair of IAPR TC4。
Email：yusq@sustech.edu.cn