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

03 2024

总第41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 03 期

总第 41 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

CCF-CV 走进高校系列报告会

第 138 期 宁夏大学

2024 年 6 月 30 日，由中国计算机学会、中国电信股份有限公司宁夏分公司、宁夏科学技术协会主办，中国计算机学会计算机视觉专委会（CCF-CV）、宁夏大学信息工程学院、CCF 银川联合承办，宁夏 “东数西算”人工智能与信息安全重点实验室协办的第 138 期 CCF-CV 走进高校系列报告会——“智算融合•新质青年论坛”在宁夏大学未来教室报告厅成功举行。本期活动执行主席是宁夏大学信息工程学院院长冯锋教授、宁夏大学信息工程学院副院长刘昊教授和宁夏大学人工智能与信息安全实验室副主任李振东副教授。邀请了北京理工大学邬霞教授、浙江大学李玺教授、西安电子科技大学王楠楠教授、哈尔滨工业大学洪晓鹏教授、上海交通大学戴文睿副教授等 5 名国家级人才计划入选者作特邀报告。来自宁夏大学、北方民族大学、宁夏医科大学等高校的师生聆听了五位专家的精彩报告。

会议首先由北京理工大学邬霞教授、浙江大学李玺教授、西安电子科技大学王楠楠教授、哈尔滨工业大学洪晓鹏教授、上海交通大学戴文睿副教授做主题报告。随后，五位专家与师生互动，共同探讨、交流，并对师生提出的问题做出详尽的回答，提出了许多有价值的学术见解，论坛现场气氛热烈。

本期 CCF-CV 走进高校系列报告会从多尺度认知机制启发的智能感知与决策、多模态视觉结构学习、人脸隐私保护与伪造检测、增量学习到通用视觉大模型的 Transformer 优化设计等方面进行了分享，报告覆盖了计算机视觉的多个方面，充分展示了前沿理论与实践应用紧密结合。极大地激发了师生们的科研热情，为师生们提供了一次宝贵的学习交流机会。

责任编委 毋立芳

CCF-CV 计算机视觉前沿讲习班

2024 年 8 月 1 日-2 日，第三届 CCF 计算机视觉前沿讲习班在甘肃兰州成功举办，吸引了计算机视觉领域的高校教师、研究生、企业技术人员等 300 余人报名参加。本次活动由中国计算机学会（CCF）主办，中国计算机学会计算机视觉专委会（CCF-CV）、兰州理工大学与兰州城市学院联合承办，中国科学院计算技术研究所王瑞平研究员与兰州理工大学李晓旭教授担任执行主席。讲习班旨在促进计算机视觉领域的学术交流与高级人才培养，帮助该领域青年从业者提升技术水平，开拓实践视野，掌握最前沿的理论成果和创新应用。本届讲习班共邀请 9 位知名专家报告前沿学术进展，帮助学员全面学习并系统掌握计算机视觉前沿理论、方法与技术。

讲习班开班仪式由执行主席、CCF 计算机视觉专委会秘书长王瑞平研究员主持，介绍了本次活动的目的、组织和报名情况。

讲习班开幕式首先由 CCF 计算机视觉专委会副主任王亮研究员致辞，介绍了专委会的宗旨和讲习班的意义，对参加讲习班的讲者和学员们表示热烈欢迎。

兰州理工大学副校长郑小平教授发表感谢致辞，作为本次活动的承办方介绍了兰州理工大学的基本情况，对参加讲习班的专家和学员表达了热烈欢迎和诚挚谢意。

随后，讲习班全体成员进行了合影留念。

浙江大学求是特聘教授李玺教授为学员讲授第一课。李老师以“多模态视觉结构学习”为题，从一个新视角对多模态视觉结构学习的研究内容进行了讲解，围绕数据驱动的人工智能方法，进行大规模图像/视频数据的视觉特征学习，从目标视觉感知特性、视觉特征表达、深度学习器构建机制、高层语义理解等多维度视角进行了深入剖析，并引出了大规模多模态特征学习所涉及的主要研究问题和技术方法。最后，李老师系统地回顾了多模态特征表达和学习领域的不同发展阶段，介绍了近年来团队围绕视觉语义分析和理解所做的一系列代表性的研究工作及其实际应用。

清华大学代季峰副教授以“多模态基础模型研究”为题，为学员带来了精彩讲解。代老师表示大模型已经给多方面领域研究带来显著变革，在迅速发展的数字世界中，机器理解和创造的能力是一个引人入胜的关键主题。代老师认为我们正在见证一个非凡的时代，大型基础模型不仅仅进行信息处理，它们正在学习理解和生成具有惊人创造力的复杂语言和图像内容。多模态基础模型无缝集成了多种形式的数据，如文本和图像/视频，它们不仅仅是工具，而是合作伙伴，增强人类的创造力，重塑我们对人工智能能力的理解。在这次报告中，代老师详细地讲解了大基础模型的复杂工作原理，并分享了研究团队的最新研究进展，引导学员们纵览语言和图像研究领域，理解这些模型如何看待人类的世界。

哈尔滨工业大学左旺孟教授以“视觉生成与编辑技术研究”为题，介绍了课题组近年来在文生图像、视频和 3D 领域的研究进展与成果。文生图模型作为当下研究热点，是多模态技术发展的重要方向。事实上，不同于通用图像生成，许多应用需求中都会涉及特定个体或特定布局的生成。基于此，左老师讲解了课题组的主要研究工作，包括：(1) 图像生成方面，介绍了图像定制化和特定布局引导的图像生成及其应用；(2) 视频生成方面，讲解了基于引导信息的视频生成、文生视频和长视频方面的工作；(3) 3D 生成方面，从数据驱动和模型

驱动两个角度展开了机制的探索。最后，整合文生图像、视频和 3D 领域的工作，提出了一种物理机制引导的视频生成方法。

国防科技大学徐凯教授以“融合三维感知与多模态大模型的具身智能”为题，为学员讲解了具身智能的前沿知识与技术。视觉感知是机器人探索、感知和理解未知环境的重要方式。随着三维传感和重建技术的飞速发展，三维图形学正与机器人视觉深度融合，产生了基于三维几何引导具身感知与交互的新途径，结合多模态大模型的强大知识推理和任务规划能力，实现支持机器人在三维世界中交互的具身智能。徐老师围绕机器人的主动重建、主动理解、以及任务驱动的交互，介绍了团队近年来的研究工作。其中，在任务驱动的交互方面，重点介绍在线语义理解驱动的物体目标导航机器人，以及多模态大模型驱动的室内物品整理机器人。最后探讨了三维感知与交互的世界模型的构建，及其对具身智能发展的推动作用。

清华大学黄高副教授以“面向视觉 Transformer 的高效注意力机制”为题，指出视觉 Transformer 模型在最近几年得到了极大的发展，相关工作在分类、分割、检测等视觉任务上都取得了很好的效果。基于 Transformer 的基础模型在计算机视觉和多模态学习领域展现了巨大的潜力，然而 Transformer 中自注意力机制的计算复杂度与输入序列长度呈平方关系，导致训练时间长、显存开销大、容易过拟合等问题，尤其是在高清图像和长视频处理方面存在明显的计算瓶颈。黄老师详细讲解了如何利用空间动态计算实现稀疏高效的自注意力计算，以及如何利用线性自注意力实现长序列建模，为设计与训练高效的视觉和多模态大模型提供了新的思路与方法。

中国科学院自动化研究所王亮研究员以“多模态认知计算”为题，为学员讲解了多模态认知计算方面的相关知识。王老师表示多模态认知计算旨在模拟人类联觉，强化机器对多模态输入的感知与理解。虽然最近多模态预训练模型异军突起，在众多基础任务上取得了很好的结果，但是仍存在模型架构同质化、类人认知能力低等方面的局限性，无法较好地满足公共安全、智能制造等重点领域对多模态认知技术的迫切需求。本报告通过分析现有多模态模型在认知能力方面的不足，从类人记忆、推理等认知机制计算建模的角度，介绍了如何通过语义记忆、逻辑推理、混合地图等方式来提升模型在图文匹配、视觉语言导航等典型多模态任务上的认知能力。

清华大学朱军教授以“扩散模型：不止于高维数据生成”为题，深入探讨了扩散概率模型的发展与应用。生成式模型(AIGC)发展迅速，作为 AIGC 的关键技术之一，扩散概率模型在跨模态的文图生成、3D 生成、视频生成等方面取得显著进展。朱老师详细讲解了基于扩散模型的高维数据生成过程，包括扩散概率模型的基础理论和高效算法、大规模多模态扩散模型架构、3D 生成和视频生成等内容。最后，分析了预训练扩散模型的鲁棒分类以及扩散模型的离线强化学习等内容。报告在总结部分指出，当前广泛关注的扩散概率模型，其实质可以描述为一个前向的扩散过程，该过程可视为逐渐增加噪声的过程，扩散模型在许多场景下的生成效果非常突出，这也为其带来了广泛的应用。

中国科学院计算技术研究所蒋树强研究员以“具身智能中的视觉导航：机理、方法与实践”为题，指出具身智能是真实物理场景下人工智能的重要表现形态，在动态开放环境下的无人系统与人机协同系统中具有重要应用价值。视觉导航是具身智能的一项重要任务，是智能体应用在现实世界中的一项重要能力，在动态且未知环境下，现有技术缺乏准确的地图且无法进行高效的导航。相比于机器，人类自身的行为机理可以在未知动态环境中高效地导航至目标物体，因此如何将人类的视觉导航能力赋予机器是视觉导航任务的关键问题。蒋老师的报告系统地分析了模拟环境与真实环境的差异与关联，探讨了具身导航从模拟环境到现实环境的迁移与实践。

北京大学王鹤博士以“面向通用机器人的具身多模态大模型系统”为题，详细解读了具身智能大模型的定义、范围和关键技术。王鹤博士认为具身多模态大模型系统可以直接对高度泛化的物理任务输出动作。通过对比本体和数据，对人形机器人的形态进行探讨，提供硬件的发展思路和泛化训练数据的获取途径。对于能力层，通过合成数据和 Sim2Real 实现了多个泛化的移动和操作技能，包括二指灵巧抓取、铰接类物体操作、柔性物体操作、端到端视觉语言大模型等，这些内容构成了机器人的小脑模型。而对于大脑模型，王鹤博士展示了 GPT-4V 为代表的非具身多模态大模型进行视觉感知、任务规划和调用中层的三维视觉过程，实现从家用电器泛化操作到开放指令物体摆放的功能。

在 2 天的课程中，9 位顶尖学者为学员们奉上了精彩纷呈的学术盛宴。为表达对讲者的诚挚谢意，CCF-CV 专委会与承办单位精心设计制作了感谢牌并在报告现场进行颁发。

第三届 CCF 计算机视觉前沿讲习班为计算机视觉领域的广大师生和工程师提供了一个与专家学者近距离

交流学习的宝贵机会，大家对专家所讲的课程内容产生了极大兴趣，现场互动非常活跃，学术氛围浓厚。

在结业典礼上，执行主席王瑞平研究员对讲习班进行了总结。首先，祝贺本次活动在兰州理工大学举办取得了圆满成功，并对各位讲者、学员、承办单位表达了衷心感谢。然后，征集学员意见反馈，请大家为下一届讲习班的组织工作提供宝贵建议。最后，全体学员领取了结业证书并合影留念，记录下了第三届 CCF 计算机视觉前沿讲习班的难忘瞬间。

责任编委 潘金山

CCF-CV 计算机视觉前沿进展研讨会

2024 年 8 月 3 日，中国计算机学会计算机视觉专委会（CCF-CV）年度学术研讨会 RACV（Recent Advances on Computer Vision）在兰州成功召开。RACV 定位为国内计算机视觉领域的小规模精品研讨会，通过定向邀请方式汇集领域专家，深度研讨计算机视觉领域中的若干核心问题并形成进展报告。研讨会试图通过务实、开放与平等的对话与讨论，深入发掘相关研究领域潜在的问题，为广大的科研人员提供观察问题的新视角与新观点。

研讨会开幕式由专委会秘书长、中国科学院计算技术研究所王瑞平研究员主持，中国科学院院士、南京大学党委书记谭铁牛教授，专委会主任、中国科学院计算技术研究所陈熙霖研究员，兰州城市学院党委书记曹洁教授，兰州理工大学副校长郑小平教授致开幕辞。

本次会议设置了 4 项研讨主题。每项主题首先由引导发言嘉宾进行主题发言，之后所有与会人员进行自由讨论。

8 月 3 日上午首先进行了专题一“具身智能中的视觉挑战与机遇”的研讨。该专题由专委会常委、清华大学鲁继文教授，专委会执行委员、清华大学苏航副研究员，上海科技大学马月昕助理教授组织，邀请了中国科学院计算技术研究所蒋树强研究员、国防科技大学徐凯教授、北京大学王鹤助理教授 3 位嘉宾进行主题发言。与会嘉宾围绕具身智能给视觉和机器学习研究带来哪些新的挑战、大模型对于具身智能的推动作用是什么、

目前具身智能的数据收集和获取方式有哪些瓶颈、视觉对于具身智能研究的意义是什么、人形机器人在具身智能研究中的作用是什么、具身智能近期和远期的应用场景有哪些等问题进行了精彩的讨论和观点分享。

专题二“视觉世界模型”由专委会常委、南开大学程明明教授，专委会执行委员、中山大学林倞教授，专委会执行委员、清华大学黄高副教授组织，邀请了东南大学魏秀参教授、北京大学袁粒助理教授、上海人工智能实验室李弘扬博士 3 位嘉宾进行主题发言。嘉宾们围绕视觉世界模型与强化学习中的世界模型有何区别与联系、训练通用视觉世界模型的数据、SORA 等视频生成模型是否能成为世界模型、大语言模型的世界常识如何迁移至视觉世界模型、如何强化视觉世界模型的空间感知能力以及视觉世界模型如何服务于具身智能等场景的决策等议题展开了深入探讨。

8 月 3 日下午继续进行了专题三“视觉研究范式：数据+知识”的研讨。该专题由专委会常委、百度计算机视觉首席科学家王井东博士，专委会常委、北京工业大学毋立芳教授，专委会执行委员、香港中文大学（深圳）吴保元副教授组织，邀请了中国科学院自动化研究所张兆翔研究员、中国科学院自动化研究所刘静研究员、中国人民大学赵鑫教授 3 位嘉宾进行主题发言。嘉宾们围绕视觉的研究范式是否开始从以模型为中心转移到以数据为中心、合成数据是否会成为计算机视觉发展的核心、真实数据和合成数据在模型训练中将分别扮演什么样的角色、高质量真实&合成数据是否可以满足视觉发展的大部分需求、数据驱动的方式存在的固有问题如何解决、“知识+数据”协同是否是一种可行的研究范式以及面临哪些挑战等议题展开了深入探讨。

专题四“计算机视觉前沿创新战略”由专委会副主任、中国科学院自动化研究所王亮研究员，专委会常委、北京邮电大学马占宇教授，专委会秘书长、中国科学院计算技术研究所王瑞平研究员组织，特别邀请了中国科学院院士、南京大学党委书记谭铁牛教授作专题指导发言，邀请了北京大学查红彬教授、上海科技大学虞晶怡教授、中山大学郑伟诗教授 3 位嘉宾进行主题发言。与会嘉宾们围绕计算机视觉的战略价值、计算机视觉的核心科学和技术问题、我国及国际视觉研究的基础与条件、未来发展趋势及政策建议等议题展开了深入探讨。

本次研讨会深入探讨了本领域最前沿研究问题，主题发言视角广阔，自由讨论热情激烈，参会嘉宾们纷纷

表示本次会议内容丰富，收获良多。按照计划，组委会后续将整理相关主题的发言与讨论文稿，形成观点性文档进行发布，把讨论从线下延伸到线上，欢迎更多专家学者积极参与。

本次研讨会由兰州理工大学和兰州城市学院承办，滴滴出行提供独家赞助。会议最后安排了特别致谢环节，专委会陈熙霖主任、王亮副主任、刘青山副主任分别为承办单位和赞助单位颁发感谢牌。

责任编委 黄岩

专题综述

多模态生成式大模型的自回归式建模

金阳 孙至诚 穆亚东
北京大学

本文主要介绍北京大学与快手科技公司合作研究的成果，发表在ICLR 2024的工作LaVIT[1]和ICML 2024的工作Video-LaVIT[2]，代表了基于自回归架构的多模态生成式大模型的最新研究进展。

一、研究背景

生成式人工智能是近年来学术界和工业界的研究热点。美国人工智能公司 OpenAI 在 2024 年初发布文生视频模型 Sora。该模型只需要读入一段提示文本（即所谓的“prompt”），即能自动生成包含提示所描述的复杂场景背景、多角色交互、动作语义的高清视频。在不少示例中，生成视频的质量会随着提示文本中的描述性细节的丰富而得到提升。与现有的文生视频模型如 Runway Gen-2、Pika 等相比较，Sora 在多个样例视频中展示了长达一分钟的连续、稳定、高品质视频，而现有其他模型通常仅能产生几秒钟连贯性的视频输出。Sora 系统的发布，普遍被认为是继 ChatGPT 之后生成式人工智能发展上的又一个重要技术突破。其背后的核心技术，包括视频时空表征学习和基于 Transformer 的扩散模型等，不仅在文生视频这一任务中取得当前最佳性能，也可以被预见将被迁移至更多模态的复合型内容生成任务中，将迅速地改变当前广告、互动娱乐、影视制作和媒体宣传等行业的技术生态，成为推动社会经济发展的重要技术之一。

近年来扩散模型和自回归模型逐渐成为相关任务的主流技术。针对文生图任务，OpenAI 公司在 2021 年发布了基于 VQ-VAE 和 GPT 的 DALL-E[3]，之后又发布了改进版本 DALL-E2[4]和 DALL-E3[5]，国内的清华大学等单位提出了 CogView[6]模型，通过对文本和图像同时进行大规模的协同预训练来实现更精确的文生图。隐扩散模型（latent diffusion）[7]和 GLIDE[8]提出了基于文本指导的图像扩散过程，经验性对比了 CLIP 指导

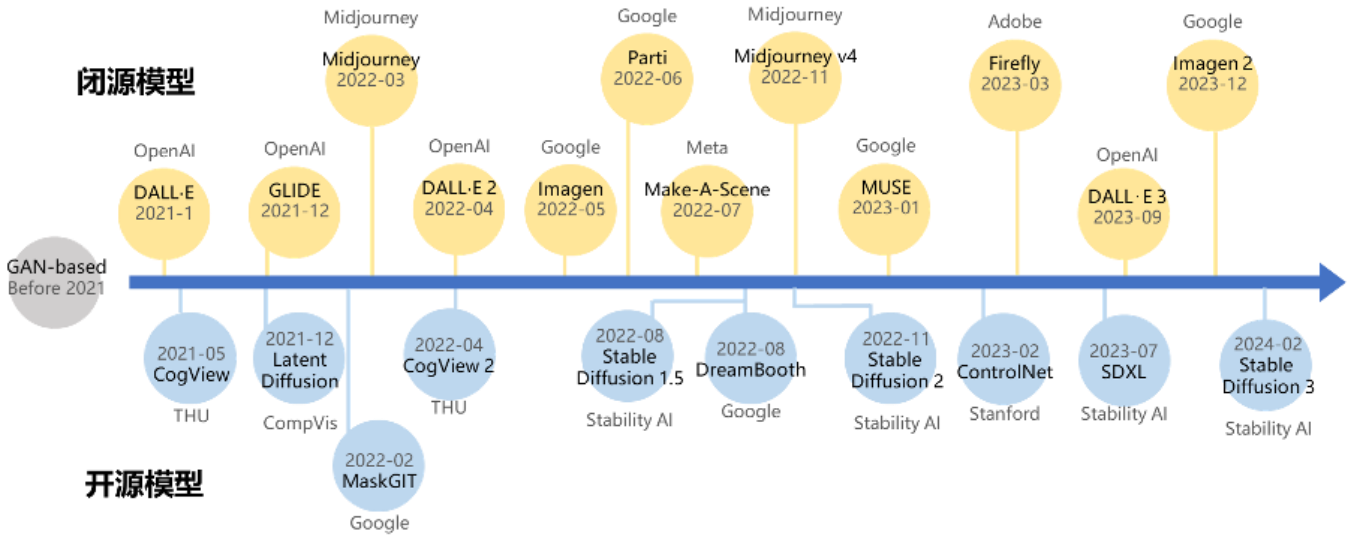


图 1　文本生成图像模型的技术发展时间线

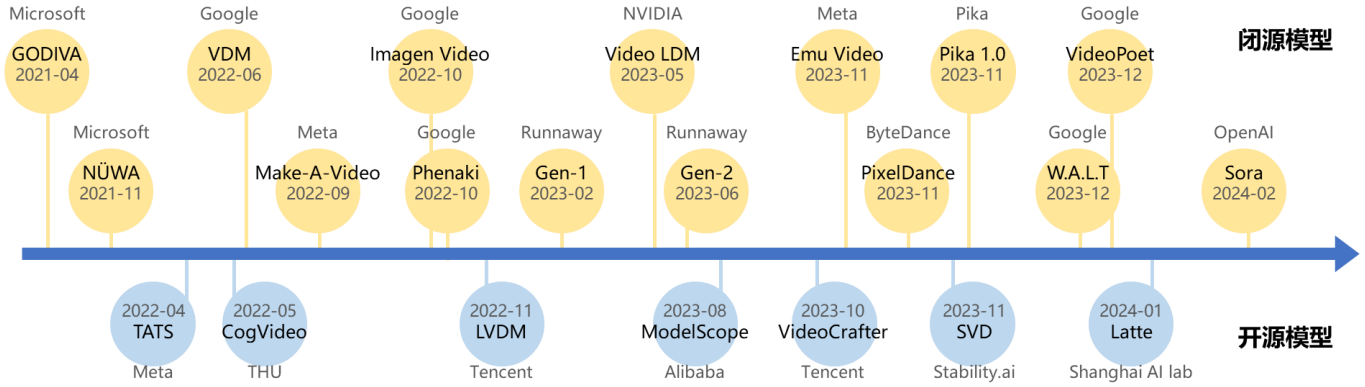


图 2　视频生成式模型的技术发展时间线

和无分类器指导（classifier-free guidance）两种策略的效果，验证了后者在图片真实性和主题相似方面效果更好。其他模型还包括 MaskGIT[9]，改进模型 CogView2[10]，Stable Diffusion 及其改进版本，Imagen[11]及其改进版本 Imagen2，闭源模型 Parti[12]、Make-A-Scene[13]、DreamBooth[14]、MUSE[15]、Midjourney、Firefly 等。ControlNet[16]通过添加额外控制条件，来引导 Stable Diffusion 按照创作者的创作思路生成图像，从而提升图像生成的可控性和精度。该方向的技术发展回顾见图 1。

针对文生视频任务，主流方法主要基于扩散模型，如 LVDM[17]、Stable Video Diffusion[18]等。视频生成式模型可以选择不同的提示信息作为输入，如文本描述或单幅图像。其中，[18]旨在利用静态图像作为条件帧，从而实现基于此单一图像输入的视频生成。在 2021 年提出的多任务模型 “女娲” NÜWA[19]以文本或视觉草图作为输入的自适应编码器和由 8 个视觉合成任务共享的解码器组成，可以支持涂鸦生成图像、图像填充、涂鸦转视频等 8 种模式。为了更好地支持多任务的统一建模，自回归模型（如 EMU-Video[20]、VideoPoet[21]等）自 2023 年也逐渐受到关注，这些方法首先对不同模态各自进行离散标记化，继而送入 LLM 风格的自回归模型进行多任务的统一训练。图 2 展示了相关技术近来的发展时间线。

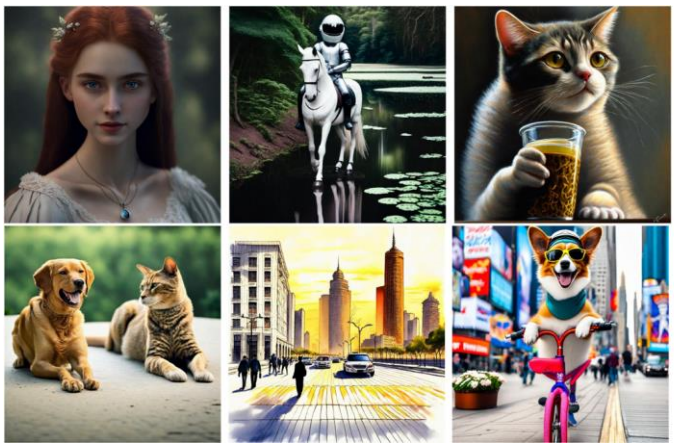


图 3　LaVIT 模型的文生图结果示例

图 4　LaVIT 模型的多模态生成示例

二、自回归文生图大模型LaVIT

1．模型简介

LaVIT 作为一个新型的通用多模态基础模型，可以

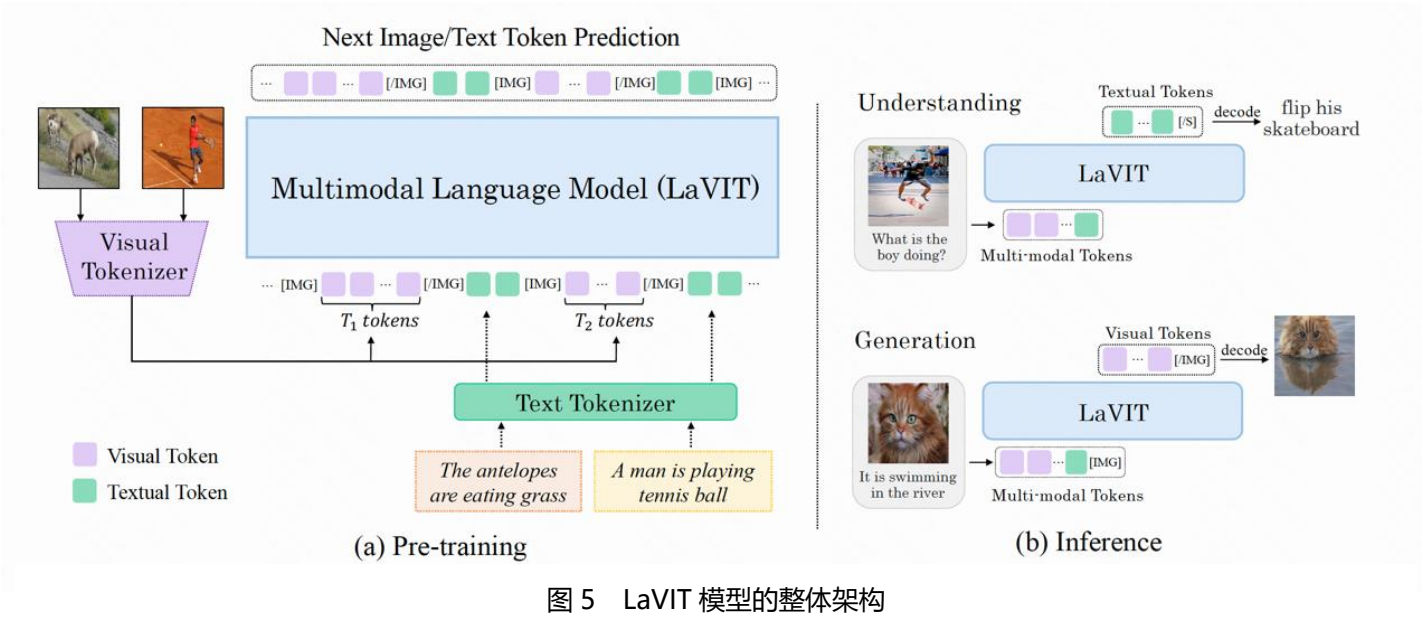


图 5　LaVIT 模型的整体架构

像语言那样，既能够理解也能生成视觉内容。LaVIT 继承了大语言模型成功的训练范式，即以自回归的方式预测下一个图像或文本 token。在训练完成后，其可以充当一个多模态通用接口，无需进一步的微调，就能执行多模态理解和生成任务。例如，LaVIT 具有以下的能力：

（1）实现高质量文本到图像的生成：LaVIT 能够根据给定的文本提示生成高质量、多种纵横比和高美感的图像（如图 3）。其图像生成能力与最先进的图像生成模型（如 Parti、SDXL 和 DALL-E3）相媲美；

（2）根据多模态提示进行图像生成：由于在 LaVIT 中，图像和文本都被统一表示为离散化的 token，因此其可以接受多种模态组合（例如图 4 中的文本、图像+文本、图像+图像）作为提示，生成相应的图像，而无需进行任何微调。

（3）理解图像内容并回答问题：在给定输入图像的情况下，LaVIT 能够阅读图像内容并理解其语义。例如，模型可以为输入的图像提供 caption 并回答相应的问题。

2．模型架构

LaVIT 模型的整体架构如图 5 所示，其优化过程包括两个阶段：

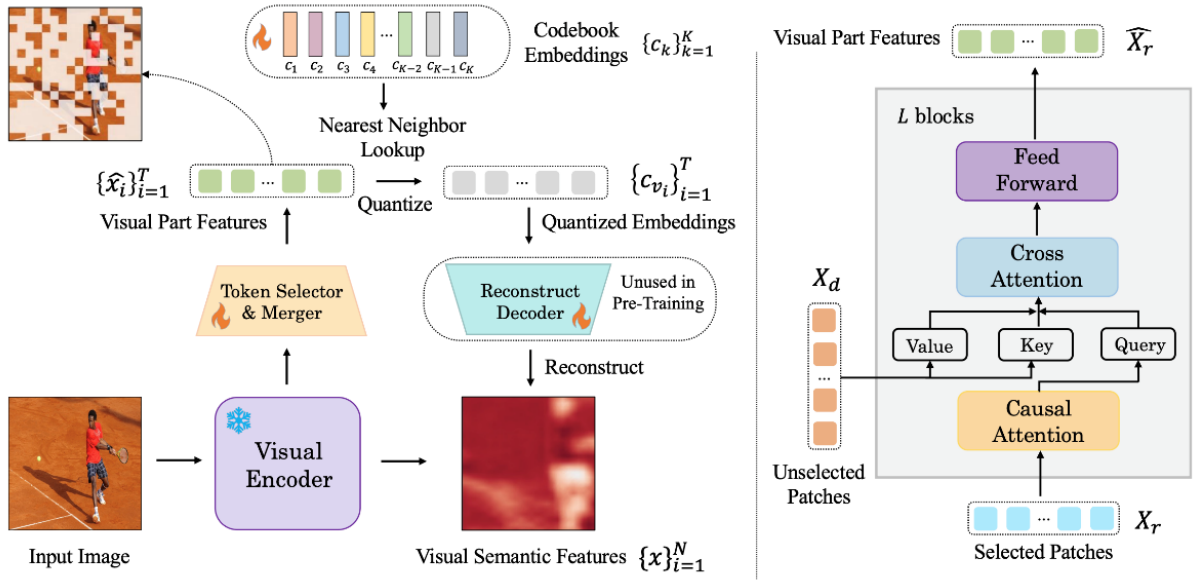
阶段 1：动态视觉分词器

为了能够像自然语言一样理解和生成视觉内容，LaVIT 引入了一个设计良好的视觉分词器，用于将视觉内容（连续信号）转换为像文本一样的 token 序列，就像 LLM 能够理解的外语一样。作者认为，为了实现统一视觉和语言的建模，该视觉分词器（Tokenizer）应该具有以下两个特性：

离散化：视觉 token 应该被表示为像文本一样的离散化形式。这样对于两种模态采用统一的表示形式，有利于 LaVIT 在一个统一的自回归生成式训练框架下，使用相同的分类损失进行多模态建模优化。

动态化：与文本 token 的不同的是，图像 patch 之间有着显著的相互依赖性，这使得从其他图像 patch 中推断另一个 patch 相对简单。因此，这种依赖性会降低原本 LLM 的 next-token prediction 优化目标的有效性。LaVIT 提出通过使用 token merging 来降低视觉 patch 之间的冗余性，其根据不同图像语义复杂度的不同，编码出动态的视觉 token 数量。这样对于复杂程度不同的图像，采用动态的 token 编码也进一步提高了预训练的效率，避免了冗余的 token 计算。

图 6 展示了 LaVIT 所提出的视觉分词器结构。该动态视觉分词器包括 token 选择器和 token 合并器。如图 6 所示，token 选择器用来选择最具信息的图像区块，而 token 合并器则将那些 uninformative 的视觉块的信息压缩到保留下的 token 上，实现对冗余 token 的 merging。整个动态视觉分词器则通过最大限度地重构输入图像的语义进行训练。



(a) Dynamic Visual Tokenizer　　(b) The Token Merger

图 6　LaVIT 模型的动态视觉 token 生成器（左）和 token 合并器（右）

Token 选择器：Token 选择器接收 N 个图像区块级的特征作为输入，其目标是评估每个图像区块的重要性并选择信息量最高的区块，以充分代表整个图像的语义。为实现这一目标，采用轻量级模块，由多个 MLP 层组成，用于预测分布π。通过从分布 π 中采样，生成一个二进制决策 mask，用于指示是否保留相应的图像区块。

Token 合并器：Token 合并器据生成的决策掩码，将 N 个图像区块划分为保留X_r和舍弃X_d两组。与直接丢弃X_d不同，token 合并器可以最大限度地保留输入图像的详细语义。token 合并器由 L 个堆叠的块组成，每个块包括因果自注意力层、交叉注意力层和前馈层。因果自注意力层中，X_r中的每个 token 只关注其前面的 token，以确保与 LLM 中的文本 token 形式一致。与双向自注意相比，这种策略表现更好。交叉注意力层将保留的 token X_r作为 query，并根据它们在语义上的相似性合并X_d中的 token。

阶段 2：统一的生成式预训练

经过视觉分词器处理后的视觉 token 与文本 token 相连接形成多模态序列作为训练时的输入。为了区分两种模态，作者在图像 token 序列的开头和结尾插入了特殊 token：[IMG] 和 [/IMG]，用于表示视觉内容的开始和结束。为了能够生成文本和图像，LaVIT 采用两种图文连接形式：[image, text] 和 [text; image]。

对于这些多模态输入序列，LaVIT 采用统一的、自回归方式来直接最大化每个多模态序列的似然性进行预训练。这样在表示空间和训练方式上的完全统一，有助于 LLM 更好地学习多模态交互和对齐。在预训练完成后，LaVIT 具有感知图像的能力，可以像处理文本一样理解和生成图像。

3．实验结果

零样本多模态理解：如表 1 所示，LaVIT 在图像字幕生成（NoCaps、Flickr30k）和视觉问答（VQAv2、OKVQA、GQA、VizWiz）等零样本多模态理解任务上取得了领先的性能。

零样本多模态生成：在这个实验中，由于所提出的视觉 tokenizer 能够将图像表示为离散化 token，LaVIT 具有通过自回归生成类似文本的视觉 token 来合成图像的能力。作者对模型进行了零样本文本条件下的图像合成性能的定量评估，比较结果如表 2 所示，从表中可以看出，LaVIT 的表现优于所有其他多模态语言模型。与 Emu 相比，LaVIT 在更小的 LLM 模型上取得了进一步改进，展现了出色的视觉-语言对齐能力。此外，LaVIT 在使用更少的训练数据的情况下，实现了与最先进的文本到图像专家 Parti 可比的性能。

Method	Image Captioning		Visual Question Answering			
	Nocaps	Flickr	VQAv2	OKVQA	GQA	VizWiz
Flamingo-3B (Alayrac et al., 2022)	-	60.6	49.2	41.2	-	28.9
Flamingo-9B (Alayrac et al., 2022)	-	61.5	51.8	44.7	-	28.8
OpenFlamingo-9B (Awadalla et al., 2023)	-	59.5	52.7	37.8	-	27.5
MetaLM (Hao et al., 2022)	-	43.4	41.1	11.4	-	-
Kosmos-1 (Huang et al., 2023)	-	67.1	51.0	-	-	29.2
Kosmos-2 (Peng et al., 2023)	-	80.5	51.1	-	-	-
BLIP-2 (Vicuna-7B) (Li et al., 2023)	107.5	74.9	-	-	41.3	25.3
BLIP-2 (Vicuna-13B) (Li et al., 2023)	103.9	71.6	-	-	32.3	19.6
CM3Leon-7B (Yu et al., 2023)	-	-	47.6	-	-	37.6
Emu (LLaMA-13B) (Sun et al., 2023)	-	-	52.0	38.2	-	34.2
Ours (LLaMA-7B)	**114.2**	**83.0**	**66.0**	**54.6**	**46.8**	**38.5**

表 1　LaVIT 模型的零样本多模态理解任务评估

Method	Model Type	FID(↓)
Text2Image Specialist:		
DALL-E (Ramesh et al., 2021)	Autoregressive	28.0
CogView (Ding et al., 2021)	Autoregressive	27.1
SD (Rombach et al., 2022)	Diffusion	12.6
GLIDE (Nichol et al., 2021)	Diffusion	12.2
DALL-E2 (Ramesh et al., 2022)	Diffusion	10.4
Make-A-Scene (Gafni et al., 2022)	Autoregressive	11.8
MUSE-7.6B (Chang et al., 2023)	Non-Autoregressive	7.9
Imagen-3.4B (Saharia et al., 2022)	Diffusion	7.3
Parti-20B (Yu et al., 2022b)	Autoregressive	**7.2**
Multimodal Large Langauge Model:		
GILL (OPT-6.7B) (Koh et al., 2023)	LLM	12.2
Emu (LLaMA-13B) (Sun et al., 2023)	LLM	11.7
CM3Leon-7B (Yu et al., 2023)	LLM	10.8
Ours (LLaMA-7B)	LLM	**7.4**

表 2　LaVIT 模型的零样本文本到图像生成性能

图 7　LaVIT 模型的多模态图像生成结果示例

多模态提示图像生成：LaVIT 能够在无需进行任何微调的情况下，无缝地接受多种模态组合作为提示，生成相应的图像，而无需进行任何微调，如图 7。LaVIT 生成的图像能够准确反映给定多模态提示的风格和语义。而且它可以通过输入的多模态提示修改原始输入图像。在没有额外微调的下游数据的情况下，传统的图像生成模型如 Stable Diffusion 无法达到这种能力。

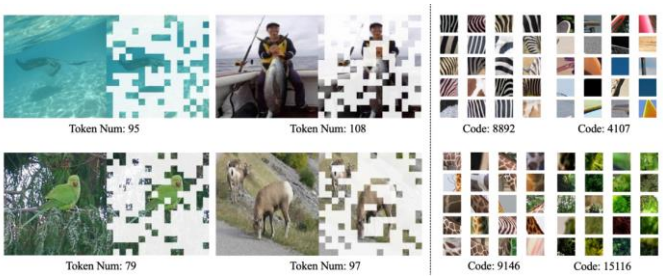


图 8　LaVIT 模型动态视觉分词器（左）和学习到的 codebook（右）的可视化示例

定性分析：如图 8 所示，LaVIT 的动态分词器可以根据图像内容动态选择最具信息量的图像块，学习到的代码本可以产生具有高层语义的视觉编码。

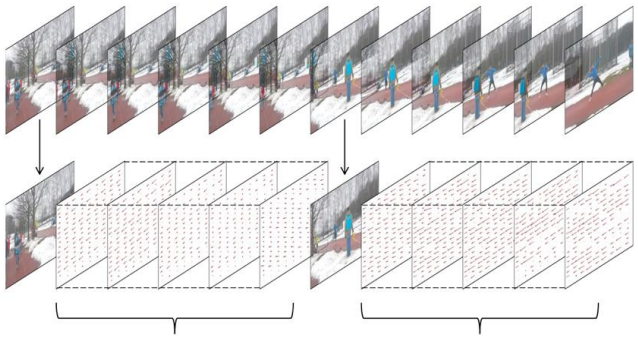


图 9　基于视觉运动解耦的视频表示

三、自回归文生视频大模型 Video-LaVIT

1．背景介绍

最近，大语言模型（LLMs）的重大突破引发了研究者们开发多模态大语言模型的热潮，已经出现了像GPT-4V，Gemini这样的多模态智能体，可以准确地理解图像中的内容。尽管取得了一定的成功，这些多模态大语言模型仍主要集中在图像-文本数据上，对于视频模态的探索则相对较少。与静态图像相比，视频作为一种动态的媒体形式，其更符合人类的视觉感知。因此，从视频数据中学习对于帮助智能体理解现实世界尤为重要。

视频理解的关键挑战在于：如何有效地对时空动态信息进行建模，例如随着时间变化的动作和场景等。目前，已经有一些方法尝试去利用语言模型(LLM)的强大推理能力来处理视频数据，它们将不同的视频帧当作不同的图像分别进行独立编码。然而，这种编码方式无法很好地捕捉时序信息。尽管最近的研究VideoPoet[21]尝试通过3D视频编码器来处理视频生成，但其适用性受限于短视频片段，因为其产生的长token序列（例如，VideoPoet对于一个2.2秒的视频片段需要使用1280个token进行编码）会导致计算资源的巨大消耗。

那么，如何以更加高效的方式在语言模型中编码视频呢？可以观察到，同一个视频镜头中的不同视频帧之间通常存在较多的时间冗余，没有必要将所有的帧都编码为输入到语言模型的token。因此，本文旨在寻找一种更高效的方法来编码视频中的时间运动信息，无需一次编码所有帧。如图9所示，我们将一个视频片段分解为交替的关键帧和运动向量，并在语言模型中进行分别编码。关键帧表示主要的视觉语义，而运动向量则表示基于关键帧的时间演变。这种解耦的视觉运动表示具有两个主要优势。首先，我们不需要编码所有视频帧来建模时间信息，只需编码相邻帧之间的差异。由于运动向量比密集的像素值更为稀疏，因此编码运动向量所需的token数量可以大大减少，这可以提高大规模视频预训练的效率。此外，解耦表示使得我们能够分别建模视觉和运动信息。关键帧就类似一张图像，所以我们可以自然地继承基于图像的模型的视觉知识，而不需要从头开始训练。

基于这种视觉运动解耦的表示，本文提出的Video-LaVIT模型将得到的关键帧和运动向量都分词为离散化的token，并以自回归的方式预测下一个图像、运动、或文本token，因此实现对视频、图像和文本的统一生成预训练。在训练完成后，Video-LaVIT具有对图像和视频内容的理解和生成能力。

2．模型架构

视频编码：Video-LaVIT 模型的核心在于将视频分解为关键帧和运动向量，关键帧捕捉主要的视觉语义，而运动向量描述其对应的关键帧随时间的动态演变。具体来说，在像 MPEG-4 这样的视频编码协议中，主要有两种帧类型：I 帧和 P 帧。I 帧是关键帧传达主要的视觉内容，并作为后续 P 帧的参考。P 帧仅保留与其前一帧

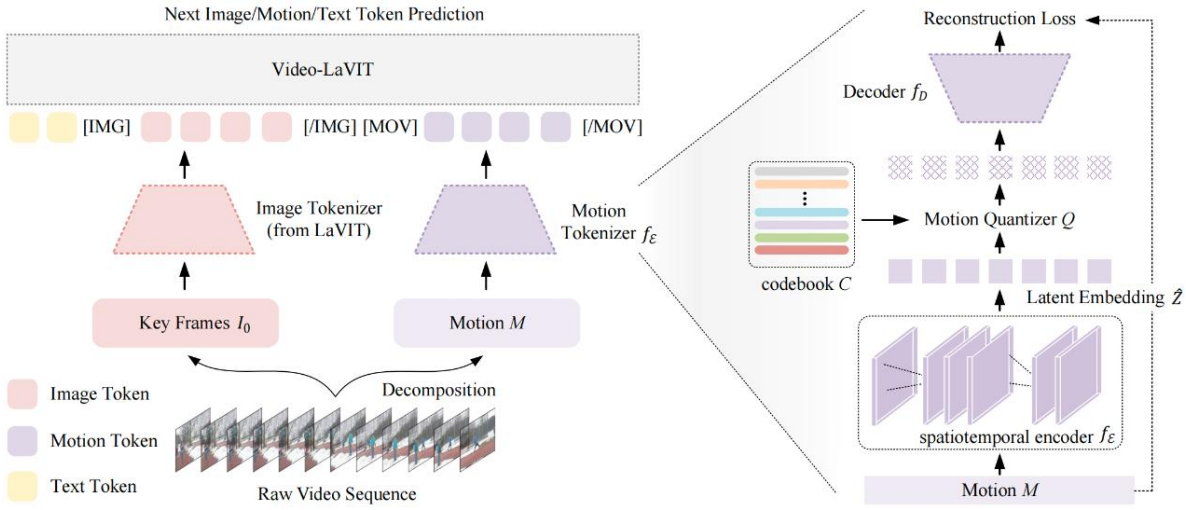


图 10　Video-LaVIT 模型的整体架构

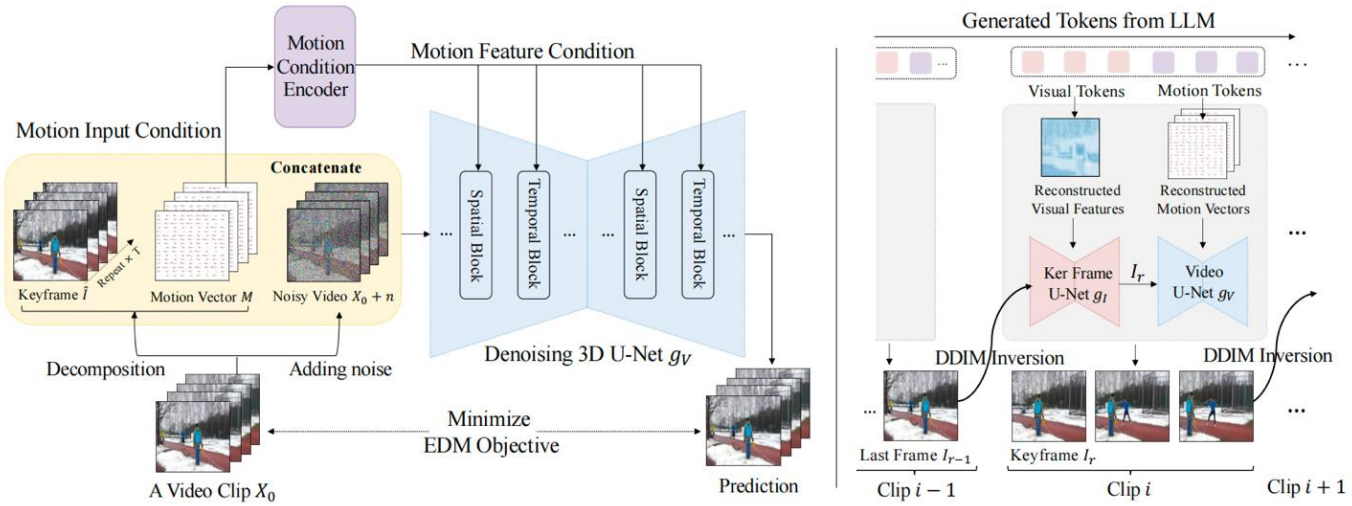


图 11　Video-LaVIT 模型的视频解码

的差异，而这些差异通过运动向量进行编码。具体来说，每个 P 帧被划分为 16×16 的不重叠宏块，其运动向量则是通过找到与相邻帧之间最佳块对应关系来估计的。简而言之，运动向量是视频中时序动作的高效表示。与每个像素位置的密集光流相比，它计算的是块之间的运动，并且可以直接从视频编解码器中高速提取。如图 10 所示，Video-LaVIT 将得到的关键帧和运动向量都分词为离散化的 token，和文本一起放在语言模型中进行统一建模。

视频解码：相对于上一节的编码，视频解码则是一个逆过程：将由多模态语言模型生成的离散化token映射回其原始的连续像素空间。考虑到学习从离散标记到高维视频空间直接映射的挑战，我们采用了顺序解码策略。如图11所示，首先解码关键帧，随后的其他帧则通过视频解码器恢复。视频解码器采用去噪U-Net架构，它通过以关键帧和运动向量作为条件，重建原始视频片段的所有视频帧。为了确保重建的视频帧严格遵循原本视频的运动信息，我们还设计了一种增强的运动编码策略。除了将运动向量作为解码器的直接输入外，我们还在解码器的block中增加了空间和时间交叉注意力层，加强对运动信号的进一步编码。整个视频解码器的训练则只需视频数据，不需要任何文本描述，因此可以扩展到更多无监督的视频数据。

由于视频被表示为多个交替的⟨visual, motion⟩序列，Video-LaVIT可以自然地通过自回归的方式生成多个视频片段，支持创建更长的视频。值得注意的一点是，如果分别对不同的视频片段进行解码，不同的视频片段之间会出现一些细粒度的视觉细节不一致的情况。为了解决这个问题，我们在解码视频片段时加入了一个显式的噪声约束。如图11所示，我们将最后一个片段的结束帧反转为中间噪声状态，然后将这个噪声状态作为下一个视频片段解码过程中的初始噪声。这样，相邻的视频片段就可以明确地相互关联起来，这对于长视频的生成至关重要。

多模态内容的联合预训练：基于本文提出的视觉-运动解构的离散化分词策略，我们可以不加区分地将所有模态（视频、图像和文本）视为输入到语言模型中的一系列离散token。这使得模型能够继承大语言模型成功的训练范式，以自回归的方式直接最大化每个标记的似然性。经过预训练后，Video-LaVIT能够生成不同模态的token，实现多模态的理解和生成。

3．实验结果

图像和视频理解：在11个常用的图像和视频基准测试中，Video-LaVIT展示了其在多模态理解上的能力。在图像理解上，其在八个广泛使用的图像问答和多模态基准测试中都取得了最佳的性能。例如，在SQA上，它比具有更高输入分辨率的LLaVA-1.5高出3.2%。在三个常见的视频基准测试中，Video-LaVIT与多个最近的视频-语言模型进行了比较，在这三个基准测试中均取得了最先进的性能。例如，在MSVD-QA上超过了之前领

Method	LLM size	Image Question Answering				Multimodal			
		VQA^{v2}	GQA	VizWiz	SQA^{I}	MME	MMB	SEED	MM-Vet
Flamingo (Alayrac et al., 2022)	9B	51.8	-	28.8	-	-	-	-	-
BLIP-2 (Li et al., 2023b)	13B	41.0	41.0	19.6	61.0	1293.8	-	46.4	22.4
InstructBLIP (Dai et al., 2023)	13B	-	49.5	34.3	63.1	1212.8	44.0	-	25.6
CM3Leon (Yu et al., 2023a)	7B	47.6	-	37.6	-	-	-	-	-
Emu (Sun et al., 2024)	13B	52.0	-	34.2	-	-	-	-	36.3
DreamLLM (Dong et al., 2024)	7B	72.9*	-	49.3	-	-	58.2	-	**36.6**
Video-LLaVA (Lin et al., 2023)	7B	74.7*	60.3*	48.1	66.4	-	60.9	-	32.0
LLaMA-VID (Li et al., 2023f)	7B	78.3*	63.0*	52.5	67.7	1405.6	65.3	59.7	-
LLaVA-1.5 (Liu et al., 2023a)	7B	78.5*	62.0*	50.0	66.8	1510.7	64.3	58.6	30.5
Video-LaVIT	7B	**80.3***	**64.4***	**56.0**	**70.0**	**1551.8**	**67.3**	**64.0**	33.2

Method	LLM size	MSVD-QA		MSRVTT-QA		ActivityNet-QA	
		Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM (Yang et al., 2022)	1B	32.2	-	16.8	-	24.7	-
Video-LLaMA (Zhang et al., 2023)	7B	51.6	2.5	29.6	1.8	12.4	1.1
VideoChat (Li et al., 2023d)	7B	56.3	2.8	45.0	2.5	26.5	2.2
Video-ChatGPT (Maaz et al., 2023)	7B	64.9	3.3	49.3	2.8	35.2	2.7
LLaMA-VID (Li et al., 2023f)	7B	69.7	3.7	57.7	3.2	47.4	3.3
Video-LLaVA (Lin et al., 2023)	7B	70.7	3.9	59.2	**3.5**	45.3	3.3

表 3　Video-LaVIT 模型的图像和视频理解性能

先的模型Video-LLaVA 2.5%

文本和图像生成视频：通过在大规模图像视频数据上进行预训练，Video-LaVIT能够根据人类指令，以自回归的形式生成多个不同的视频片段。如图12所示，与商用的生成模型Gen-2相比，Video-LaVIT能够生成更复杂的物体运动，同时不违反物理规则的视频内容。

除了文本提示，Video-LaVIT还支持图像到视频的生成。给定一张图像输入，其可以将图像处理为离散的视觉标记，并输入到多模态语言模型中生成运动token。生成的运动信息与输入图像结合后，可以解码成一个视频片段。Video-LaVIT可以通过生成不同的合理运动来为输入图像添加动画效果。

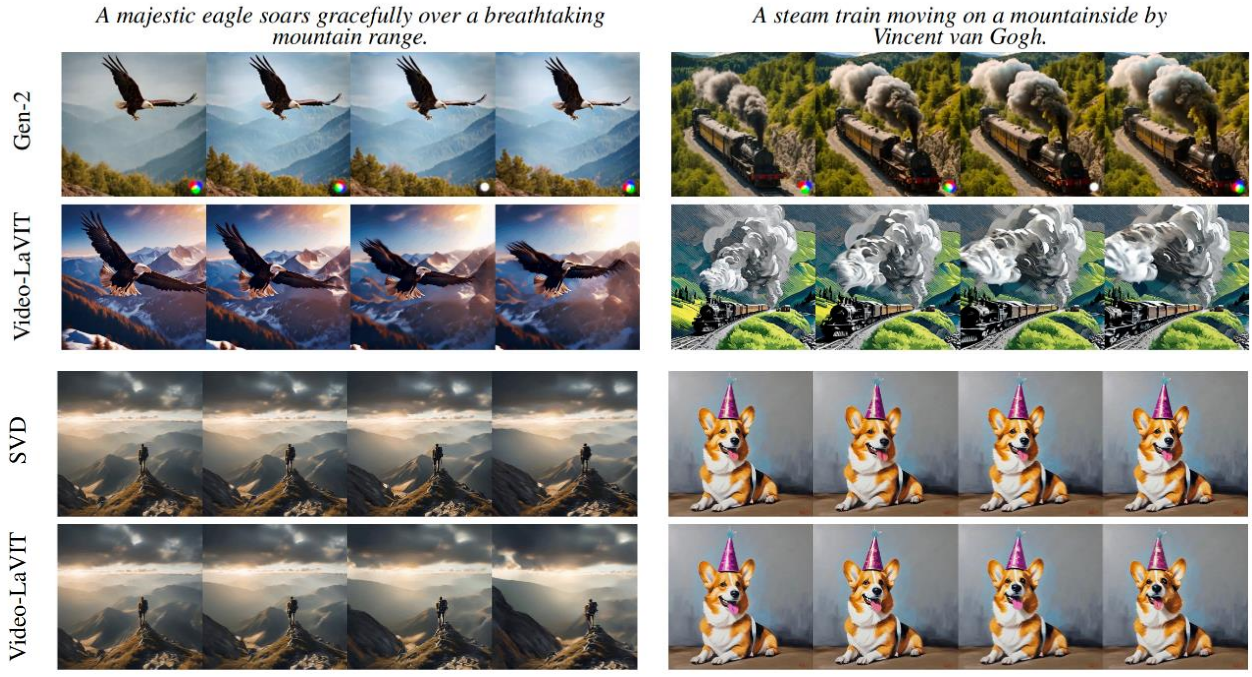


图 12　Video-LaVIT 模型根据文本或图像指令生成视频，与 Gen-2 和 SVD 的对比

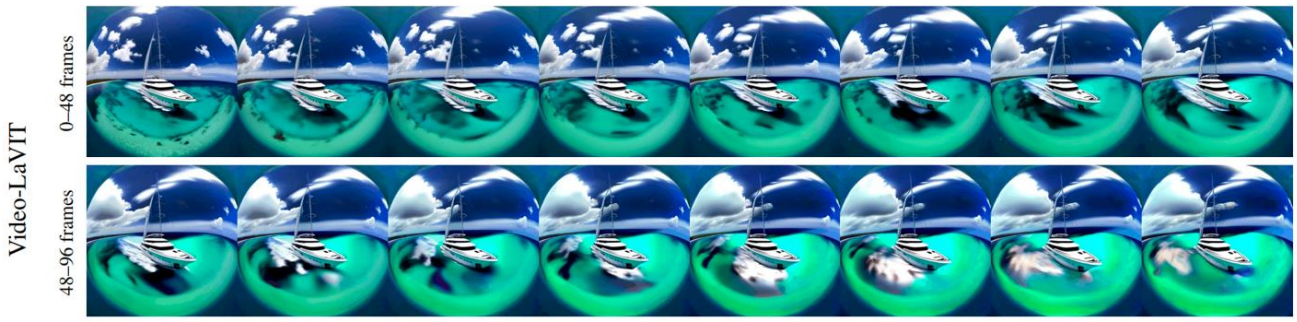


图 13　Video-LaVIT 模型的长视频生成效果

长视频生成：通过在解码连续视频片段时显式地约束噪声，Video-LaVIT能够在长视频生成中保持高度的时间一致性。如图13所示，所生成的视频片段之间的运动和视觉信息都具有高度的一致性。

四、总结

该论文提出了基于自回归架构的多模态生成式大模型 LaVIT 与 Video-LaVIT。其中 LaVIT 的提出为多模态任务的处理又提供了一种创新范式，通过使用动态视觉分词器将视觉和语言表示为统一的离散 token 表示，继承了 LLM 成功的自回归生成学习范式。通过在统一生成目标下进行优化，LaVIT 可以将图像视为一种外语，像文本一样理解和生成它们。Video-LaVIT 通过引入视觉与运动解耦，将该思路推广至文生视频任务。上述方法的成功为未来多模态研究的发展方向提供了新的启示，利用 LLM 强大的推理能力，实现更智能、更全面的多模态理解，并为生成打开新的可能性。

最后，本文介绍的 LaVIT、Video-LaVIT 的代码和模型均发布于：https://github.com/jy0205/LaVIT。

责任编委　**崔海楠**

参考文献

[1] Jin, Yang, et al. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. In ICLR 2024.
[2] Jin, Yang, et al. Video-LaVIT: Unified video-language pre-training with decoupled visual-motional tokenization. In ICML 2024.
[3] Ramesh, Aditya, et al. Zero-shot text-to-image generation. In ICML 2021.
[4] Ramesh, Aditya, et al. Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125, 2022.
[5] Betker, James, et al. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3, 2023.
[6] Ding, Ming, et al. CogView: Mastering text-to-image generation via transformers. In NeurIPS 2021.
[7] Rombach, Robin, et al. High-resolution image synthesis with latent diffusion models. In CVPR 2022.
[8] Nichol, Alex, et al. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In ICML 2022.
[9] Chang, Huiwen, et al. MaskGIT: Masked generative image transformer. In CVPR 2022.
[10] Ding, Ming, et al. CogView2: Faster and better text-to-image generation via hierarchical transformers. In NeurIPS 2022.
[11] Saharia, Chitwan, et al. Photorealistic text-to-image diffusion models with deep language understanding. In NeurIPS 2022.
[12] Yu, Jiahui, et al. Scaling autoregressive models for content-rich text-to-image generation. In TMLR 2022.
[13] Gafni, Oran, et al. Make-A-Scene: Scene-based text-to-image generation with human priors. In ECCV 2022.
[14] Ruiz, Nataniel, et al. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In CVPR 2023.
[15] Chang, Huiwen, et al. Muse: Text-to-image generation via masked generative transformers. In ICML 2023.

[16] Zhang, Lvmin, et al. Adding conditional control to text-to-image diffusion models. In ICCV 2023.
[17] Blattmann, Andreas, et al. Align your latents: High-resolution video synthesis with latent diffusion models. In CVPR 2023.
[18] Blattmann, Andreas, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127. 2023.
[19] Wu, Chenfei, et al. NÜWA: Visual synthesis pre-training for neural visual world creation. In ECCV 2022.
[20] Girdhar, Rohit, et al. Emu video: Factorizing text-to-video generation by explicit image conditioning. In ECCV 2024.
[21] Kondratyuk, Dan, et al. VideoPoet: A large language model for zero-shot video generation. In ICML 2024

金 阳

金阳，北京大学前沿交叉学科研究院 2022 级博士研究生，导师为穆亚东研究员，主要研究方向为多模态大语言模型,视频理解。
Email: jiny@stu.pku.edu.cn

孙至诚

孙至诚，北京大学前沿交叉学科研究院 2020 级博士研究生，导师为穆亚东研究员，主要研究方向为持续学习，多模态生成。
Email: sunzc@pku.edu.cn

穆亚东

穆亚东，北京大学研究员、长聘副教授、博士生导师、北大博雅青年学者，先后在北京大学获得理学学士和理学博士学位。曾在新加坡国立大学、美国哥伦比亚大学、华为香港诺亚方舟实验室、美国电话电报公司研究院（AT&T Labs）担任研究职位，入选国家级人才计划，在国际主流会议和期刊发表论文 120 余篇，其中 CCF 推荐 A 类会议和 ACM/IEEE 汇刊论文 80 余篇，申请国内外专利 30 余项。获得陕西省自然科学一等奖和国际会议 SIGIR 最佳论文提名奖。担任多媒体领域旗舰期刊 IEEE Transactions on Multimedia 的编委，多次担任计算机视觉领域顶级会议（如 CVPR、ACM Multimedia）的领域主席。
Email: myd@pku.edu.cn

热点追踪

神经场的网格模型正切核理论（GTK）

赵泽林[1] 范凤磊[2] 廖文龙[3] 严骏驰[4]
[1]上海交通大学 [2]香港中文大学 [3]上海交通大学 酷哇科技 [4]上海交通大学

本文获得CVPR24最佳论文提名，在盲审(pre-rebuttal)阶段获得三个审稿人的一致满分意见(5/5/5)。许多当代研究利用基于网格的模型来表示神经场，但对这些模型的系统分析仍然缺失，阻碍了这些模型的改进。因此，本文引入了一个基于网格模型的理论框架。该框架指出，这些模型的逼近和泛化行为由网格切线核（GTK）决定，GTK 是基于网格模型的内在属性。所提出的框架促进了对各种基于网格模型的一致和系统的分析。此外，该框架还激发了一个名为乘法傅里叶自适应网格（MulFAGrid）的新型基于网格模型的发展。数值分析表明，MulFAGrid 相较于其前身具有更低的泛化界限，表明其具有强大的泛化性能。如图1所示，MulFAGrid 在包括二维图像拟合、三维符号距离场（SDF）重建和新视图合成在内的各种任务中实现了最先进的性能，展示了卓越的表示能力。本工作也即将在Jittor深度学习框架平台进行实现和开源。

一、研究背景

首先，我想介绍一下神经场及其广泛的应用。神经场是基于坐标的网络，表示一个场，实质上是一种连续参数化，代表一个物体或场景的物理量。神经场在计算机视觉和其他研究领域的各种任务中显示出了显著的进展和潜力。

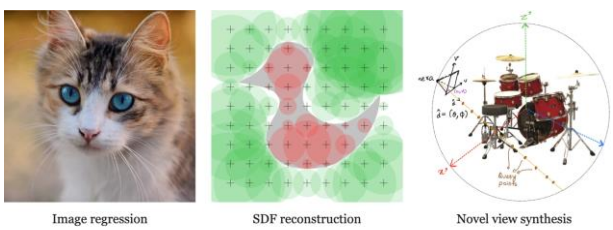


图 1　神经场的应用场景举例

我们的工作主要针对网格模型（grid-based models），这是一类主要的神经场模型。网格模型在参数化和功能上与传统的神经网络(如 MLP)有根本不同。主要的不同包括 MLP 往往包含多层非线性神经网络，并且 MLP 没有显示查询的过程，同时 MLP 的输入也不一定非是位置坐标。如图 2 所示，网格模型以查询坐标为输入，该坐标被发送到下标函数以从网格中获取一组特征向量。然后，模型输出核函数和这些特征向量的加权平均值。该模型需要学习的主要是特征向量。最简单的核函数是不含参数的插值算法（如最近邻算法或者双线性插值算法）。核函数里面也可以包含可学习的参数。

图 3 是英文版的网格模型的示意图，选择不同的下标函数可以让我们的算法适配不同的网格模型。

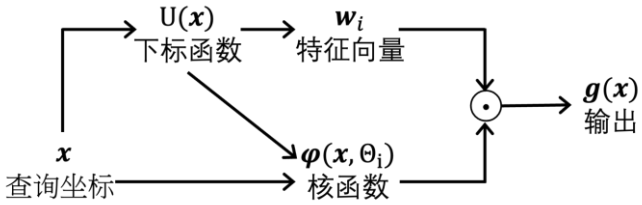


图 2　网格模型的通用示意图

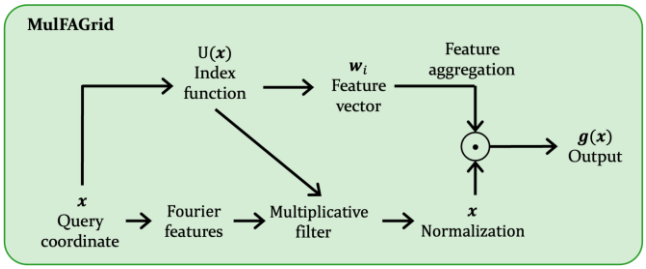


图 3　我们提出的新的网格模型 MulFAGrid 的示意图

二、理论介绍

1. 主要研究问题

我们的理论旨在通过三个主要问题来理解和增强网格模型：

1、我们如何理解网格模型的训练动态?

2、我们如何衡量网格模型的泛化性能?

3、我们如何设计一个更好的网格模型?

为了解决这些问题，我们提出了一个基于正切核的理论框架。

这里要先介绍一下什么是正切核。正切核这一概念来自于著名的深度学习理论文章神经正切核[1]（Neural Tangent Kernels，NTK）。NTK 是一种核函数，最初由研究者在研究神经网络的训练过程时提出的。当神经网络在参数空间中靠近其初始值时，通过对神经网络梯度下降优化过程的分析，发现网络的行为可以用一个固定的核函数来描述，这个核函数就是神经正切核。比如神经网络的输出可以用其参数的梯度来表示。在训练过程中，网络参数的更新会导致输出的变化，而这种变化在参数空间中的变化速率可以用梯度来表示。NTK 定义了两个输入数据点的输出变化之间的相似度。形式上，对于输入数据点 x_i 和x_j，神经正切核$\Theta(x_i, x_j)$可以定义为网络输出对参数的梯度的内积：

$$\Theta(x_i, x_j) = \langle \frac{\partial f(x_i, \theta)}{\partial \theta}, \frac{\partial f(x_j, \theta)}{\partial \theta} \rangle$$

其中，$f(x, \theta)$是神经网络的输出，θ是网络的参数。

2. 理论成果

我们的理论结果表明，网格模型的近似和泛化性能与网格切线核（GTK）有关。GTK 被定义为一个正半定矩阵，它测量梯度空间中两个数据点之间的距离。这里我们展示了 GTK 的定义：g 是由 w(t)参数化的网格模型，X 是一个数据集，其中 Xi 是第 i 个数据。GTK 可以这样表示：

$$[G_g(t)]_{i,j} = \langle \frac{\partial g(X_i, w(t))}{\partial w}, \frac{\partial g(X_j, w(t))}{\partial w} \rangle,$$

注意这个形式跟神经正切核（NTK）的形式是吻合的，因为他们都是正切核，他们的主要区别是适用的模型不同，GTK 主要适用于网格模型。后面可以看出，因为网格模型本质上比较简单纯粹，所以 GTK 的理论基本不需要近似，但是 NTK 的理论需要网络无穷宽的假设才能成立。

我们的定理一（网格模型优化定理）说明，网格模型的模型参数根据微分方程（如下面方程所示）演化。

$$\frac{dO(t)}{dt} = -G_g(t)(O(t) - Y)$$

这里 O(t)表示网格模型的输出，G(t)表示网格模型的 GTK，而这里的 Y 表示数据集的标签（向量化，Yi 表示第 i 个数据的标签），这个定理有什么意义呢？直观地讲，有了这个定理我们就可以预测模型的效果（不用亲自炼丹即可确定模型的好坏），这一定理在理论上是有很高价值的，也是朴素成立的，不依赖于具体的模型细节。

接下来，我们提出了另一个定理（GTK 不变定理）。定理 2 指出，网格模型的 GTK 在训练期间保持不变。(这是一个非正式的说法，如果想了解正式版本的话请参考我们的论文，相关证明在论文的附录中有，论文以及附录在 arxiv 可以下载)这意味着无论网格模型的大小如何，初始 GTK 在整个训练过程中保持恒定。这一定理揭示了 GTK 是由模型和数据集决定的一个内在特性，与模型的训练过程无关，有了这个定理，自然也不难理解网格模型的很多性质都与 GTK 有关了。

定理三，描述的是网格模型的泛化性能。在理论深度学习中，泛化性能的好坏通常由泛化界（generalization bound）来刻画。该定理揭示了网格模型的泛化界由一个特定的度量Δ决定，而$\Delta = Y^T G^{-1} Y$，与网格模型的 GTK 和数据集的标注有关。形式化的说，该泛化界提供了模型性能的概率保证。该定理说明了模型的泛化性能既与 GTK 有关，也与数据集的结构有关。结合该定理与 GTK 的特征值，我们能获得更多关于泛化性能的信息。

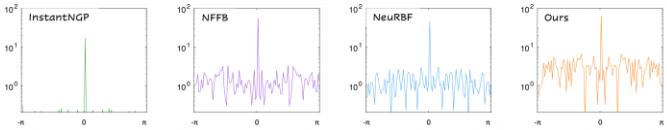


图 4　傅立叶频谱分析结果

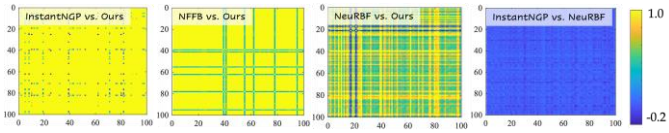


图 5　泛化性能对比可视化

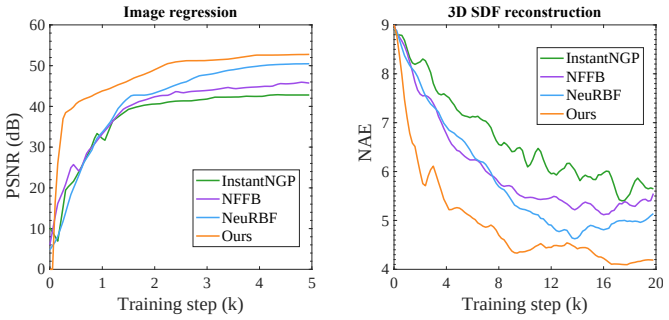


图 6　2D 神经场和 3D 神经场不同方法训练曲线

3. 新的网格模型

我们的 GTK 理论可以激发未来设计具有更好训练和泛化性能的网格模型。我们介绍了一种新的网格模型，名为 MulFAGrid 如图 3 所示。该模型使用傅里叶特征来提升高频信号的学习，并采用乘法滤波器来为模型提供节点信息。我们的模型示意图如图 3 所示。

三、实验分析与结果

然后，我们基于 GTK 理论展示了一组数值实验。我们对比了比较常见的网格模型如图 4，包括 InstantNGP[2]，NFFB[3]，NeuRBF[4]等等。首先，在频谱分析中，MulFAGrid 显示了比较宽的频谱，特别是在高频域。这一特性导致它的高频成分的收敛速度更快。

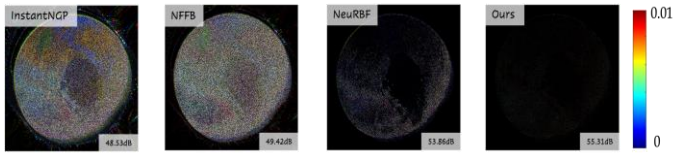


图 7　图片拟合结果对比

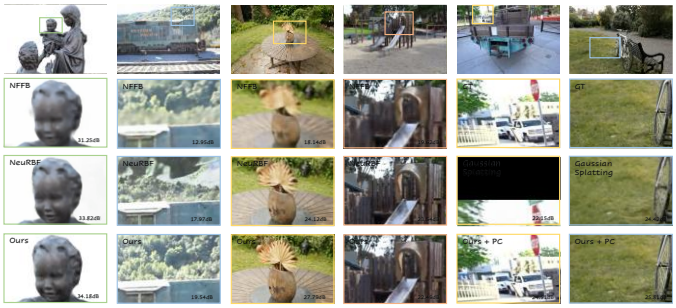


图 8　新视角生成的结果对比

在下一个实验中，我们构建了一个包含两个数据点及其对应标签的数据集。如图 5 所示 MulFAGrid 对于大多数标签值表现出更紧的泛化界表明其性能更好。

图 6 展示了各种基线方法和我们的误差图。误差图衡量预测图像与真实图像的差异。MulFAGrid 提供了更准确的拟合，展示了其优越的性能。

图 7 我们展示了我们的模型在拟合二维图像和三维符号距离函数（SDF）方面的性能。结果突出了 MulFAGrid 的准确性和效率。

最后，我们在图 8 展示了我们在新视角合成方面的结果。详细结果表明，MulFAGrid 在生成高质量的新视角方面表现出色，突显了其实际应用性。

我们论文的 Jittor 版本会在 Project Page 里放出，链接为 https://sites.google.com/view/cvpr24-2034-submission/home。敬请关注。

责任编委 **魏秀参**

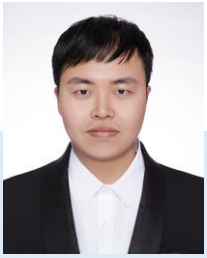
参考文献

[1] Jacot, Arthur, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks." Advances in neural information processing systems 31 (2018).

[2] Müller, Thomas, et al. "Instant neural graphics primitives with a multiresolution hash encoding." ACM transactions on graphics (TOG) 41.4 (2022): 1-15.

[3] Wu, Zhijie, Yuhe Jin, and Kwang Moo Yi. "Neural fourier filter bank." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[4] Chen, Zhang, et al. "Neurbf: A neural fields representation with adaptive radial basis functions." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.



赵泽林

赵泽林，获得上海交通大学计算机系本科学位，即将加入佐治亚理工学院攻读博士学位。上海交通大学本科毕业，本科专业排名 2/150，本科期间获得国家奖学金，上海市奖学金，曾在 NeuRIPS，ECCV，CVPR，AAAI 等顶会发表四篇一作论文，引用数超过 600。

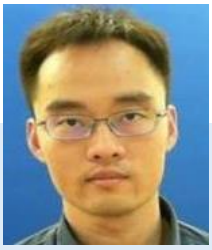
Email: sjtuytc@gmail.com

范凤磊

本文通讯作者来自香港中文大学数学系研究助理教授范凤磊博士，他所在的 Center for Mathematical AI 由曾铁勇教授创立。中心自 2018 年成立以来，在中心主任曾铁勇教授的带领下，先后承担科技部国家重点研发计划项目等一系列关键项目。

范凤磊博士于美国伦斯勒理工学院（Rensselaer Polytechnic Institute）获得博士学位，导师为国际知名影像专家王革教授，主要研究方向是脑启发智能以及神经网络的数学理论，在 JMLR，TMI，TNNLS，TCI 等杂志发表论文二十余篇，引用数过千。曾获得 IBM AI Horizon Scholarship 和国际神经网络协会（INNS）2021 年杰出博士论文奖。

Email: flfan@math.cuhk.edu.hk



严骏驰

严骏驰教授带领实验室发表第一/通讯作者 CCF-A 类论文超百篇，谷歌引用过万次，获 PaperDigest 评选的最具影响力 AAAI21、IJCAI23 论文榜首。

严骏驰教授长期任机器学习三大会议 ICML/NeurIPS/ICLR 领域主席，模式识别旗舰期刊 TPAMI、PRJ 编委。实验室学生获得挑战杯特等奖、CCF 优博/CV 新锐奖、交大学术之星等荣誉和本科生自然科学基金。

Email: yanjunchi@sjtu.edu.cn

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。

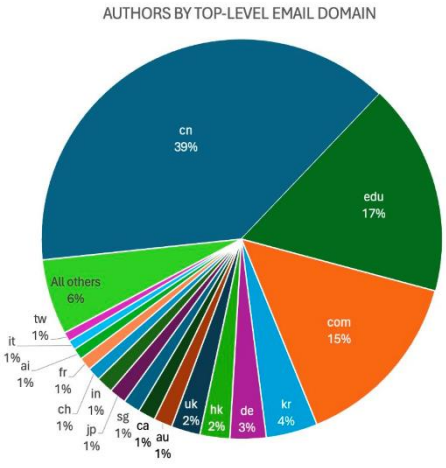


图 1　所有投稿作者的电子邮箱域名统计

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。

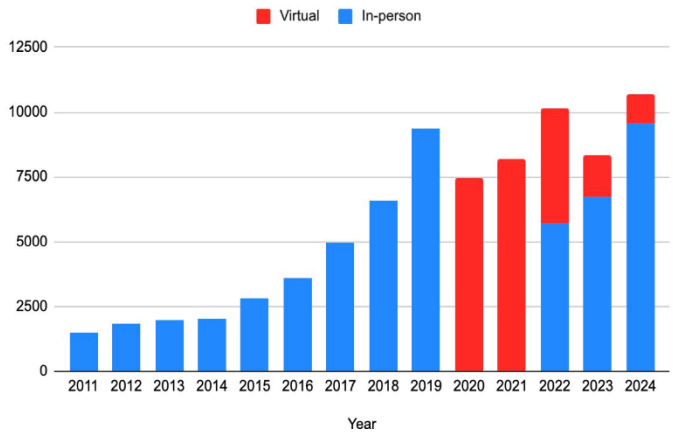


图 2　历年注册人数

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，

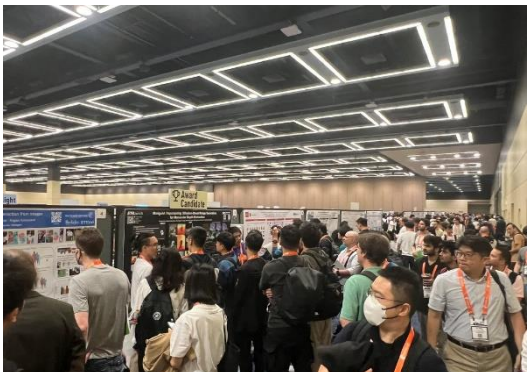


图 3　墙报展区一角

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

福州大学赵铁松教授访谈

2024 年 7 月 12 日，《CCF-CV 专委简报》在线采访了福州大学博士生导师赵铁松教授。下面是采访实录。

问题 1：赵老师，您好！首先，请您分享一下您的个人学习和研究经历。

您好！我于 2006 年毕业于中国科学技术大学，获得电子信息工程专业本科学位；2012 年毕业于香港城市大学计算机科学专业，取得博士学位；此后在加拿大滑铁卢大学、美国纽约州立大学水牛城分校从事博后研究工作。2015 年下半年回国加入福州大学，有幸入选了国家青年人才计划及若干福建省人才项目，后面一直在这边从事研究工作。2019 年我和同事们一起，组建并获批了福建省媒体信息智能处理与无线传输重点实验室，并参与建设福州大学人工智能研究院，期间一直得到学术委员会中一些相关领域的知名前辈的热心指导。

问题 2：您的研究工作深耕于计算机视觉与触觉的交互融合、多媒体通信技术的革新以及虚拟现实与具身智能的交叉领域。请问，在这些多维度的探索中，您认为自己的研究工作具备哪些独到的特色与优势？

国内外做计算机视觉相关领域的专家非常多，我们作为地方高校，不敢说有什么优势。与专委会内各位优秀专家不同的是，我的工作不仅关注计算机视觉本身，更关注其在大规模系统及具身智能的交叉应用，比如考虑智能物联网系统的“感知-传输-计算”协同优化，及面向具身智能应用的计算机“视觉-触觉”交互与融合等，相关领域涉及到计算机视觉与通信、物联网等领域的交叉研究。面向我校卓越工程师培养目标，我们也关注计算机视觉的边缘端落地应用，特别是其智慧海洋应用。由于我相对比较“杂”的经历，在电子信息、通信、物联网、计算机专业都有求学或任教，因此在这些交叉领域研究中具有一定的基础。

问题 3：在您的众多研究成果之中，请问哪一项是您个人最引以为傲的成就？同时，您又认为哪一项研究成果在未来具备最大的产业化潜力与价值，能够为社会带来深远的影响？

引以为傲谈不上。最近几年，我与各位同事、福建省本地企业一起，在视频计算与通信、物联网的融合任务中，研发相关算法与平台，包括一些国产化程度较高的设备与设施，并在省内外取得了广泛应用。但这些技术已经相对成熟，放眼未来的话，我认为视觉与触觉的深度融合，并在具身智能取得应用，是实现更加智能的未来产线的一种重要方式，值得我们更加深入研究。这也是响应总书记所提的构建新质生产力的重要途经。

问题 4：您在计算机视觉的边缘计算应用及具身智能领域的探索上取得了成就，请问对于这两个前沿方向，您是如何看待其未来的发展趋势？

个人浅见，边缘计算应用目前已经在广泛的产业化阶段，后续发展主要在于其和硬件、通信技术等的结合。而具身智能仍处在研究的萌发期，其理论、方法与应用都有太多未知数，比如智能体通过与环境的互动以第一

视角主动取得数据，并推理得到人工具象，从而解决当前人工智能算法对数据的依赖，是一个非常有意义的课题，也因此对我们更有吸引力。

问题 5：从中国科技大学的本科启航，经香港城市大学硕士与博士的连贯深造，再到加拿大滑铁卢大学博士后的学术精进，随后作为研究科学家在美国纽约州立大学水牛城分校深耕，最终扎根于福州大学，您的学术之旅跨越了众多顶尖学府，请问这一路走来，您有哪些宝贵的经历或感悟可以分享？同时，这些顶尖学府中，哪些元素给您留下了最为深刻的印象？

求学多年，我越发深刻理解这句话“所谓大学者，非谓有大楼之谓也，有大师之谓也”。我所经历的几所学校，或许没有华美的校园，但一直有各位德高望重的老师的谆谆教导，有各位优秀学长作为榜样，对于他们的指导和帮助我一直深怀感激。也是因为这些老师和学长的引领，我自己后来也放弃了企业界的机会，走上了教师的道路。

另外比较凑巧的是，我所就读的中国科学技术大学及工作的福州大学都是 1958 年建校，两校前辈筚路蓝缕以启山林的开拓精神，鞠躬尽瘁孜孜为国的奉献精神，都给我们青年教师树立了光辉的榜样。虽然很惭愧我们远未达到他们的研究水平，但一直致力于跟随前辈们的道路，严于治学，刻苦求索，争取为我们国家的科技事业尽绵薄之力。

问题 6：您不仅在学术领域深耕细作，还荣任了国际知名期刊如 IET Electronics Letters 与 IEEE SMC 的编委，以及 Sensors、IEEE Network、中兴通讯英文版等期刊的客座编辑，同时担任了 IEEE HAVE 2018 的宣传主席、IEEE MMTC Reviewer Board 成员，并频繁出任多个国际会议的 PC/AC 成员及论坛主席。请问，从这些丰富多样的职务与活动中，您积累了哪些宝贵的经验？又有哪些深刻的感悟，能够激励和启迪学术界同仁？

参与科研界的义务服务是我们大家的荣幸，也给了我们和更多优秀专家学者交流学习的机会。此外，这些工作给了我另外一个维度去思考科研问题。通常我们关注的科研问题是散点式的，作为普通科研人员，除非在比较大的团队或者从事科技规划，很少有机会从线、面层次去观察，去考量当前领域所面临的前沿问题、研究现状和未来发展。而去从事这些期刊、会议的组织和编辑等活动，给了我们这样的一个机会，通过这样的思考让我们保持科研不落伍。我相信专委会的各位同仁也有类似感悟。

问题 7：您获得过教育部高等学校科学研究优秀成果二等奖、福建省科技进步三等奖及福建青年科技奖等荣誉，同时也在福州大学荣获了杰出青年教师励志奖和十佳青年教工等荣誉。请问，能否详细谈谈这些获奖背后的故事，以及您是如何通过不懈努力与创新思维，在科研领域取得这些显著成就的？

这些成果和努力程度和我们领域各位杰出的专家相比还有一定距离。上面很多成果的取得，都离不开和各位团队成员、各位合作者的长期紧密合作，在此表示真诚的感谢。特别是跨领域的合作，我们也经历了很长的磨合期。例如，在前期与一位通信领域学者的合作中，我们就一直努力学习对方领域知识，力求在通信理论和人工智能方法方面找到较好的平衡点，相关合作持续两三年才有初步成果。但也得益于这种较长阵痛期的前期磨合，我们目前思考问题都有了一定的跨领域思维，并且期望后期基于此取得更多成果。

问题 8：作为所在学校的学位点领航者，您不仅开设了《机器学习与计算机视觉导论》、《计算机视觉》及《智能媒体与人机交互》等前沿课程，还成功将《计算机视觉》课程打造为校级一流线上线下融合课程与思政示范课程，同时引领了省级教学改革项目，发表教改论文并被收录，更编著了专业教材。在科研任务繁重的背景下，您仍能在教学领域取得显著突破，请问这是如何做到的？您秉持着怎样的教学理念，以激发学生的潜能，促进知识的有效传播与融合？

教学才是我们教授的本职，过去七八年来，我一直

担任部分本、硕、博学位点和博后流动站负责人，并结合学科培养体系与人工智能前沿知识，不断更新课程培养体系。首先，坚持思政引领，不断将中国制造 2025、发展新质生产力等思政要素融入培养计划和教学大纲，率先建设思政示范课程。其次，不断通过培养计划修订督促图像处理等基础课程改革，通过引导、鼓励方法培养前沿课程，率先将机器学习、计算机视觉等课程引入本学科的培养体系，提升学生对前沿知识的把握及方法的运用。再次，通过导师团队建设，培育了福建省电子信息专业学位硕士导师团队，通过研学相长改善研究生培养水平。最后，通过与产业的结合，推动了“亿联-福大”联合研究生培养专班和产学联合课程，从而保障了学生学有所成，学以致用。

问题 9：作为福建省“媒体信息智能处理与无线传输”重点实验室的掌舵人，您是如何精心构建并引领这样一支高素质科研团队的？在青年教师与研究生的培养与管理上，您采取了哪些独到而高效的策略？能否分享一些您在实践中总结出的优秀做法，以资同行借鉴与学习？

我们从以下几个方面着手，不断凝聚实验室研究与开拓方向。首先，跨领域融合研究，通过将人工智能、数字媒体、通信、物联网相结合，在交叉领域启发创新点，激发新活力，产生新成果。其次，产学研协同，通过研究生培养专班和产学联合课程锻炼学生基础能力，通过企业联合的科研开发，锻炼培养卓越工程师的深度探索能力。再次，本硕博贯通培养，通过模糊本科阶段培养基础知识、硕士阶段开拓研究视野、博士阶段锻炼科研能力等时间界限，制定个性化的学生培养方案，使得不同学生能够分别发挥自己的优势，提升科研兴趣与能力。最后，对于青年教师在科研和教学方面有老教师传帮带，尽快实现从博士生到导师身份的转变，完成更多高水平成果。

问题 10：在繁忙的工作之余，您有哪些爱好，以给自己放松和充电呢？同时，您又是如何平衡工作与个人家庭生活，确保两者和谐共生的？

我个人比较喜欢阅读，特别是科幻小说阅读，对于国外科幻黄金时代的著作和国内刘慈欣先生等的图书都读过比较多，这些书本除了放松之外，也有助于让我们保持对科研的兴趣。至于工作和家庭的平衡，因为我太太也是科研工作者，我们有一定的互相理解和支持。当然要达到和谐共生的境界，还需要持之以恒的努力。

问题 11：如果吐露研究工作者的心声，您最想说的是什么？

博士生和青年教师是科研生力军，但当前他们的求职和考核都还处于数论文、数项目的阶段，不利于完成真正原创的成果。原创成果需要静心做冷板凳，但他们恰恰坐不得冷板凳。就这个问题，我其实没有很好的解决方案，也真诚向科研界同仁请教。

责任编委 余烨 赵振兵

赵铁松

赵铁松，福州大学物理与信息工程学院教授、博士生导师，福州大学人工智能研究院副院长，福建省“媒体信息智能处理与无线传输”重点实验室主任。曾入选国家青年人才计划（2016），福建省百人计划（2018），获得或合作获得教育部高等学校科学研究优秀成果二等奖、福建省通信学会科学技术奖二等奖、福建省科技进步三等奖及福建青年科技奖等荣誉。担任 IEEE 高级会员，IEEE Electronics Letters、IEEE SMC、中国图象图形学学会通讯编委，ACM MM 等若干国内外会议的领域主席、论坛主席等。任中国青年科技工作者协会第六届理事，中国计算机学会计算机视觉专委会、多媒体专委会委员等。

2006 年毕业于中国科学技术大学电子工程与信息科学系，获学士学位；2011 年毕业于香港城市大学电脑科学系，获博士学位（导师 Sam Kwong 教授）。此后于香港城市大学、加拿大滑铁卢大学及美国纽约州立大学水牛城分校做博后和后续研究工作（合作导师 Zhou Wang 教授，Chang Wen Chen 教授）。2015 年底回国，任福州大学物理与信息工程学院教授，博士生导师。

曾担任福州大学物联网工程本科学位点、信号与信息处理硕士学位点、电子信息工程博士学位点视频通信与智慧显示方向、信息与通信工程博士后流动站负责人。先后主讲《图象处理与分析》、《计算机视觉》、《智能媒体与人机交互》等 5 门课程。主持获批物联网工程国家一流本科专业，“智能媒体与网络”福建省电子信息专业学位研究生导师团队，福建省“媒体信息智能处理与无线传输”重点实验室等。

委员好消息

✪ 2024 年 8 月 12 日，2023 年度江苏省科学技术奖综合评审结果公示，CCF-CV 专委会执行委员、南京理工大学**唐金辉**拟授青年科技杰出贡献奖。

✪ 2024 年 8 月 14 日，ACL 2024 最佳论文奖公布，CCF-CV 专委会 6 位委员指导的 1 篇论文获奖：华中科技大学**白翔**、**刘禹良**和华南理工大学**金连文**指导的论文 Deciphering Oracle Bone Language with Diffusion Models 获最佳论文奖，上海人工智能实验室**乔宇**、大连理工大学**王立君**和**卢湖川**指导的论文 PsySafe: A Comprehensive Framework for Psychological-based Attack, Defense, and Evaluation of Multi-agent System Safety 获杰出论文奖。

✪ 2024 年 9 月 1 日，2024 年度湖北省科学技术奖拟获奖项目公示，拟授 CCF-CV 专委会常委委员、华中科技大学**白翔**青年创新奖，拟授 CCF-CV 专委会执行委员、华中科技大学**王兴刚**等完成的“高效率弱监督视觉理解方法研究”自然科学二等奖。

✪ 2024 年 9 月 23 日，2024 年 CCF 新晋杰出会员名单公布，CCF-CV 专委会 20 位执行委员获评杰出会员，他们是：天津理工大学**高赞**、苏州科技大学**胡伏原**、厦门大学**纪荣嵘**、北京工业大学**贾熹滨**、兰州理工大学**李策**、浙江大学**李玺**、南京理工大学**李泽超**、北京大学**连宙辉**、大连理工大学**刘日升**、中国科学院计算技术研究所**闵巍庆**、湘潭大学**欧阳建权**、华中科技大学**桑农**、北京大学**施柏鑫**、常州大学**王洪元**、西北工业大学**王鹏**、中山大学**杨猛**、重庆大学**张磊**、东南大学**郑文明**、中国海洋大学**仲国强**、南京邮电大学**周全**。

责任编委 刘海波

神经渲染应用开源代码

西北工业大学 张鼎文 李昊

三维重建作为计算机连接、感知、理解世界的桥梁，已广泛应用于自动驾驶，物理仿真器，态势感知，具身智能等传统和新兴领域。随着大模型网络和神经渲染算法的发展，基于隐式表征的神经辐射场（Neural Radiance Field，NeRF）和基于显式表征的高斯泼溅（3D Gaussian Splatting，3D-GS）在场景重建之外还能提供极具真实感的视角生成。

1. GP-NeRF: Generalized Perception NeRF for Context-Aware 3D Scene Understanding

将神经辐射场 (NeRF) 应用于场景理解和表示的下游感知任务是一个越来越受欢迎的研究方向。大多数现有方法通过将语义预测视为“标签渲染”任务来构建语义 NeRF。然而，由于没有充分考虑渲染图像的上下文信息，这些方法通常存在物体边界分割不精准和物体内部像素分割异常的问题。为了解决这个问题，我们提出了一种新的 3D 场景理解方法——广义感知 NeRF（GP-NeRF），从而使广泛使用的分割模型和 NeRF 能够在统一框架下兼容工作，以实现具备上下文感知能力的 3D 场景感知。我们在两个感知任务（即语义和实例分割）下进行了实验比较，结果表明我们的方法在广义语义分割、微调语义分割和实例分割任务中分别比 SOTA 方法提高了 6.94%、11.76%和 8.47%。

如图 1 所示，GP-NeRF 利用“场聚合”Transformer 来聚合辐射场和语义嵌入场，并使用 “射线聚合” Transformer 在新视角中联合渲染它们。这两个过程在联合优化框架下执行。具体来说，我们在新视角中渲染丰富的语义特征而不是标签，并将它们输入到一个强大的 2D 分割模块中进行上下文感知的语义感知。为了使我们的框架能够兼容工作，我们进一步引入了两种新的自蒸馏损失：1) 语义蒸馏损失，它增强了语义场的区分度和质量，从而通过感知头实现更准确的语义预测性能；2) 深度引导的语义蒸馏损失，旨在监督语义场内每个点的语义表示并保持几何一致性。在这些机制下，我们的方法弥合了强大的 2D 分割模块和 NeRF 方法之间的差距，提供了一种与现有下游感知头集成的可行方案。

论文地址：

https://arxiv.org/abs/2311.11863

开源代码：

https://lifuguan.github.io/gpnerf-pages/

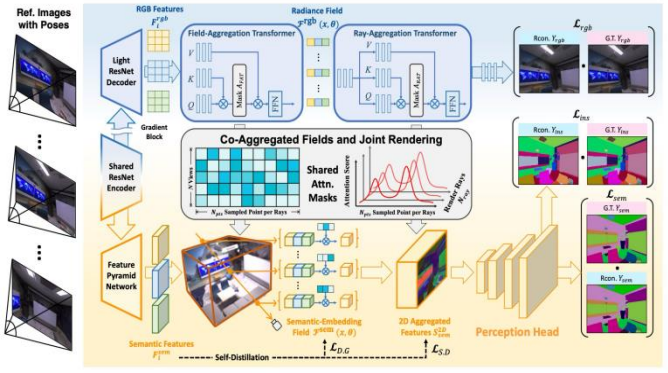


图 1 GP-NeRF 的整体框架

2. VDG: Vision-Only Dynamic Gaussian for Driving Simulation

动态高斯技术通过新视角合成推动了场景重建和图像渲染的显著进步。然而，现有方法严重依赖预先计算的位姿和通过运动结构（Structure from Motion）算法或昂贵的传感器进行的高斯初始化。我们的工作旨在通过提出仅依赖视觉信息的动态高斯（VDG）来解决上述问题，它可以仅使用视觉输入（即无位姿）构建由3DGS 表示的整个动态驾驶场景。

如图 2 所示，我们的方法引入了一种自监督视觉里程计方法，可以实现精确的相机位姿估计。利用其自监督特性，我们的方法还提供密集深度估计，这种密集深度估计在高斯初始化中起着至关重要的作用，提高了我们方法的有效性。此外，我们在框架中提出了一个场景分解步骤，将实际场景解耦为静态场景和动态物体。为了进一步提升效果，我们对位置进行了相应的参数化和迭代更新。由于我们的框架采用的是一种自分解的策略，因此在没有真实位姿的场景下会变得较为困难。为了解决这个问题，我们利用视觉里程计生成的运动掩码并引入运动掩码监督机制。该机制增强了网络识别场景中动态物体的能力。

论文地址： https://arxiv.org/abs/2406.18198

开源代码： https://3d-aigc.github.io/VDG/

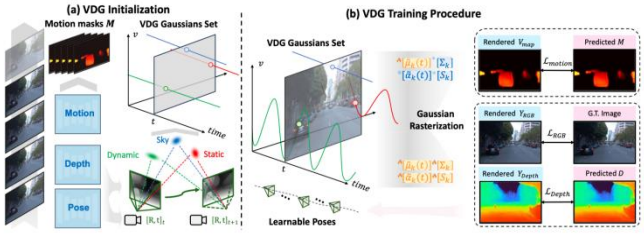


图 2 VDG 的整体框架

3. GGRt: Towards Pose-free Generalizable 3D Gaussian Splatting in Real-time

本文介绍了一种可推广的新视图渲染方法 GGRt，减少相机位姿需求，简化高分辨率图像处理和优化过程，增强 3D 高斯（3D-Gaussian Splatting)动静态解耦在现实场景的适用性。如图 3 所示，该框架由迭代位姿优化网络（Iterative Pose Optimization Network)和泛化 3D 高斯((3D-Gaussian Splatting)模型组成。通过联合学习机制，框架能从图像中估计稳健的相对位姿信息，降低对真实相机位姿的依赖。此外，我们实现了延迟反向传播机制，以支持高分辨率训练和推理，克服了以往渲染方法的分辨率限制。

为了提高渲染速度和效率，我们进一步引入了一个渐进式高斯缓存模块，该模块可在训练和推理过程中动态调整。作为第一个无位姿泛化 3D-GS 框架，GGRt 实现了 5 FPS 的推理和 100 FPS 的实时渲染。大量实验证明了我们的方法在推理速度和有效性方面优于现有的基于 NeRF 的无位姿技术。它还可以接近真实的基于位姿的 3D-GS 方法。

论文地址： https://arxiv.org/pdf/2403.10147.pdf

开源代码： https://3d-aigc.github.io/GGRt/

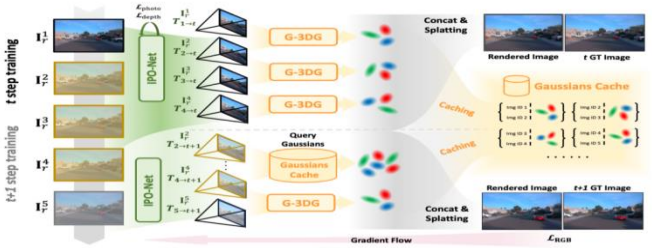


图 3 GGRt 的整体框架

张鼎文

西北工业大学自动化学院教授、博士生导师，致力于建立面向开放环境下、具备动态学习能力的视觉感知框架。电子邮箱：zdw2006yyy@nwpu.edu.cn

李 昊

博士研究生，西北工业大学自动化学院，研究方向为三维重建、场景理解。
电子邮箱：ifugan_10027@outlook.com

目标计数数据集

香港城市大学 付陈平 大连理工大学 樊鑫

目标计数（Object Counting）旨在估计图像中目标实例的数量，是计算机视觉中一个重要的研究方向。目标计数任务可推动实例分割，动作定位，行人检测等其他视觉任务的发展与完善。此外，目标计数还通常应用于智能环境导航，养殖业，视频监控，人群计数，野生动物保护，饮食模式理解和细胞群分析等行业领域。

目标计数主要包括两个研究子方向：通用目标计数（Generic/Common Object Counting，GOC）和密集目标计数（DenseObject counting, DOC）。这两个研究子方向的主要区别在于计数场景的不同。通用目标计数常用于在自然场景下的目标计数，例如 PASCAL VOC 和 COCO 数据场景。一张图像中待计数目标数量往往少于 10 个。而密集目标计数则主要面向拥挤场景中某一类目标类别的计数。一张图像中通常具有成百上千个待计数的目标数量，其中被计数的类别主要有人类，汽车或植物。得益于大规模相关数据集的提出与发布，通用目标计数和密集目标计数均得到了极大的研究与发展。

下面，本文将介绍 3 个代表性的大规模目标计数数据集，分别是 MALL 数据集，UCF-QNRF 数据集和 IOCfish5K 数据集。

1、MALL 数据集

公共场所的人群计数在人群控制、公共空间设计、行人行为分析等方面具有广泛应用。在某些简单的人群计数应用场景中，例如，对火车站台上的人群计数，一般只需要估计整个场景中的全局人数。而对于那些复杂的人群计数应用场景中，还需要估计不同空间位置的计数。例如，在购物中心的人群计数，不仅需要知道购物场景中有多少人，还需要知道他分布在哪里，进而可分析哪个商店更受欢迎。

针对上述大型购物中心的人群计数需求，本文提出了一个新的公共场景数据集（MALL）。MALL 包含 60000 多个行人实例，用于人群分析。MALL 数据集中的图像来自于某购物中心的公共可访问的监控视频。该数据涵盖从稀疏到拥挤的多样化人群密度，以及在一天中不同时间大范围的光照条件下的不同活动模式（即静态人群和移动人群）。此外，MALL 数据集具有严重的透视失真，因此一张图像场景中的人群目标具有丰富的深度，尺寸以及外观变化。同时，由于图像场景视野宽阔，室内植物和摊位等物体对人群目标产生一定程度的遮挡，进一步增加了 MALL 数据集的使用难度。MALL 数据集的图像示例见图 1。更多有关该数据集的详细情况可参考发布该数据集的论文“Feature mining for localised crowd counting”。

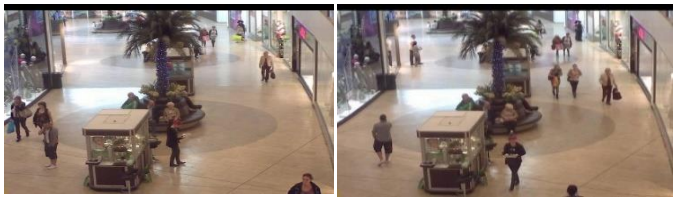


图 1 MALL 图片示例

数据集地址：
http://personal.ie.cuhk.edu.hk/~ccloy/downloads_mall_dataset.html

2、UCF-QNRF 数据集

从朝圣到抗议，音乐会到马拉松，节日到葬礼，每年都有数百万人聚集在一起。在高度密集的人群中进行计数在人群安全和管理，衡量抗议和示威活动的政治意义等方面具有广泛适用性。

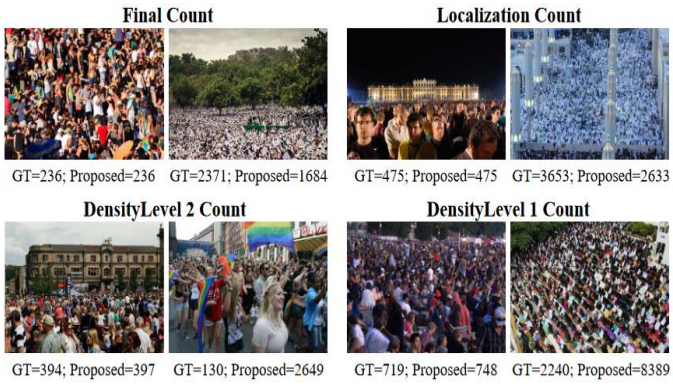


图 2 UCF-QNRF 图片示例

由于人群计数需要高质量的图像和标注，本文提出了 UCF-QNRF 数据集。该数据集具有 1535 张图像，同时包含 1251642 个人工标注点。UCF-QNRF 数据集中的图像来自 3 个数据源，分别是： Flickr，网络搜索和朝圣摄像。其中，朝圣摄像中的图片包含多样捕捉地点，视角，透视效果和时间。对于 Flickr 和网页搜索，研究人员通过 Flickr 和谷歌图像搜索的 API 手动输入查询词获得图像。这些查询词包括：人群，朝觐，观众人群，朝圣，抗议人群和音乐会人群。研究人员为每个查询网站设置了具体的图像数量，Flickr 为 2000 张，谷歌为 200 张。而后，研究人员将所有搜索结果按标题和标签进行相关性排序。同时，研究人员只下载具有原始分辨率的图像。针对所有查询词条，研究人员提取并保存所有图像的静态链接，然后使用相应的 API 下载这些链接。随后，研究人员通过计算图像相似度来检查图像是否重复，手动验证并丢弃重复的图像。UCF-QNRF 数据集的图像示例见图 2。更多有关该数据集的详细情况可参考发布该数据集的论文“Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds”。

数据集地址：

https://www.crcv.ucf.edu/data/ucf-qnrf/

3、IOCfish5K 数据集

不可分辨场景理解在计算机视觉领域引起广泛关注。研究人员通过对不可分辨物体计数（Indiscernible Object Counting, IOC）的系统研究，进一步推进了这一领域的前沿发展。不可分辨物体计数旨在相对于周围环境混合的目标进行计数。由于缺乏合适的不可分辨物体计数数据集，本文提出了一个大规模应用数据集，即 IOCfish5K。该数据集包含 5637 张高分辨率图像和 659024 个标注中心点。该数据集由水下场景中大量不可分辨的物体（主要是鱼）组成。相较于先前的计数数据集，标记过程更具挑战性。由于 IOCfish5K 的规模更大，图像分辨率更高，注释更多，场景密度更大，因此优于现有的面向不可分辨场景的计数数据集。图 3 展示了 IOCfish5K 数据集中的图像示例。

图 3 IOCfish5K 图片示例

水下场景中存在着许多难以分辨的物体，例如海马、礁石鱼、狮子鱼、叶海龙等。这是因为水下场景的可视性有限，而且海洋动物会主动模仿水下植被外貌。因此，本文专注于收集水下场景的图像。在 Youtube 网站上，研究人员使用水下场景，海洋潜水，深海场景等通用关键词搜索水下场景视频。同时，研究人员还使用切鱼，模仿章鱼，琵琶鱼，石头鱼等特定类别的关键词进一步搜索水下场景视频。最终，研究人员一共收集了 135 个高质量的视频，长度从几十秒到几个小时。接下来，研

究人员保持每 100 帧（3.3 秒）抽取一幅图像来避免图像重复采集。然而，这种做法依然无法完全避免大量低质且类似的图像。因此，在图像收集的最后一步，6 名专业标注者仔细审查了整个数据集并删除了那些不符合要求的图像。图片采集共计花费 200 个小时。

在 IOCfish5K 的所有图像中，957 张图像具有中等到高等的对象密度，即一张图像具有 101 到 200 个实例。此外，1017 张图像具有非常密集的场景，每张图像具有超过 200 个实例的目标。为了标准化 IOCfish5K 的基准测试，研究人员将其随机分为三个不重叠的部分，分别是：训练集（3,137 张图像）、验证集（500 张图像）和测试集（2,000 张图像）。

与现有数据集相比，IOCfish5K 具有 4 大优势：（1）IOCfish5K 是不可分辨场景中规模最大的目标计数数据集。（2）IOCfish5K 的图像具有更密集的计数场景，这使其成为目前最具挑战性的不可分辨物体计数的基准。（3）尽管 IOCfish5K 是专门为不可分辨物体计数任务提出的，但与现有的密集目标计数数据集相比，它有一些优势。例如，与规模最大的密集目标计数基准之一的 JHU-CROWD++相比，IOCfish5K 数据集包含更多的图像和更高的分辨率。（4）IOCfish5K 侧重于具有海洋动物信息的水下场景，这使得它显著不同于现有的目标计数数据集。更多有关该数据集的详细情况可参考发布该数据集的论文“Indiscernible Object Counting in Underwater Scene”。

数据集地址:

https://github.com/GuoleiSun/Indiscernible-Object-Counting

责任编委 贾同 王田

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，水下成像。

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，担任中日国际信息与软件学院院长。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

好 文 推 荐

浦项科技大学“Generalizable Novel-View Synthesis using a Stereo Camera”的最新成果被CVPR-2024收录。

论文：Lee, Haechan, et al. "Generalizable Novel-View Synthesis using a Stereo Camera." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2024.

新视角合成是计算机视觉和图形学中一个长期存在的病态问题。因其需要从目标场景的图像中同时预测几何和纹理，导致具有极大的挑战性。最近，神经辐射场（NeRF）通过使用基于坐标的网络联合优化几何和辐射场，取得了照片级逼真的效果。然而，每个场景需要单独进行优化，这限制了其应用性。近期的一些方法，如MVSNeRF、GeoNeRF解决了这一问题，能够在不进行逐场景优化的情况下，即时生成新视角图像，但是它们仍然难以预测准确的几何，并且合成精度有限。

为了解决这一挑战，文章提出了首个利用立体图像的可泛化NeRF方法，称为StereoNeRF。StereoNeRF架构如图1所示。StereoNeRF将立体匹配引入基于NeRF的可泛化视角合成方法中，其中立体匹配提供了重要的几何信息。首先，文章引入了立体特征提取器，通过相关联的立体图像中的水平对极线提取几何感知特征。此外，立体特征提取器利用来自现成的立体估计网络的立体相关特征，将丰富的几何知识传递给模型。然后，文章借助从立体估计网络中估算出的可靠深度，通过深度引导的平面扫描技术聚合多视角特征。该技术确保在代价体构建过程中，对几何体周围的对应关系进行匹配。同时文章利用估计出的立体深度构建了一种立体深度损失。这些来自立体匹配的附加几何线索有效缓解了可泛化视角合成中的病态问题。最后文章提出了StereoNVS数据集，这是首个用于立体图像训练和评估新视角合成的数据集。StereoNVS数据集提供了真实世界和合成的立体图像。

文章在StereoNVS数据集上的广泛评估表明，立体图像输入可以有效提高新视角合成的质量，并表明StereoNeRF在合成图像和形状质量方面优于之前的可泛化新视角合成方法。

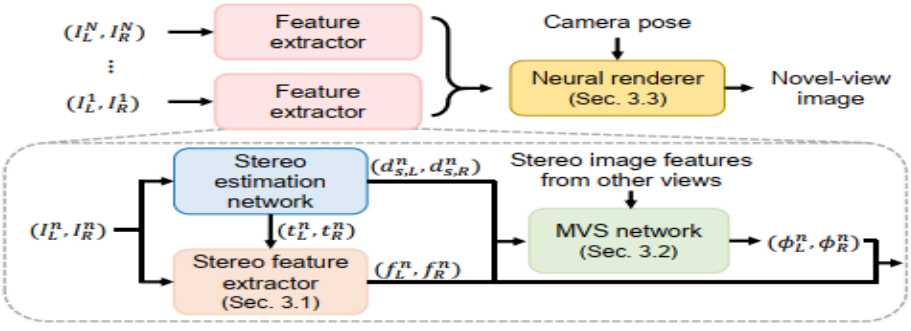


图1 提出的StereoNeRF架构图

责任编委 樊鑫 王田

好 文 推 荐

中国科学院智能信息处理重点实验室“Interpretable Object Recognition by Semantic Prototype Analysis”的最新成果被 WACV-2024 收录。

论文：Wan Q, Wang R, Chen X. Interpretable object recognition by semantic prototype analysis[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,2024.

近年来，越来越多的研究人员将精力集中在可解释 AI (XAI)上，以便将 AI 模型部署到具有高可靠性要求的真实应用场景中，如自动驾驶、医疗、交通、安全等领域。在现有的 XAI 方法中，用户更倾向于选择概念和原型作为解释形式。基于概念的解释使用与类别无关的文本概念来解构模型的分类，而基于原型的解释则通过可视化样本来解释模型，通常集中在图像的局部区域。然而，由于在构建两种解释形式时存在显著差异，目前几乎所有的可解释识别方法都只专注于部分解释或语义解释中的一个方面。因此，文章提出了一种语义原型分析网络 SPANet，能够通过“指出应该关注的位置”和“解释为什么是这样”这两种方式，提供更加清晰和易懂的决策过程解释。

文章提出的模型包括原型识别、语义附加和重建三个部分，其中语义附加模块能将图像和文本映射到共同的语义空间。原型识别模块旨在比较图像特征与存储原型，分配类别标签，并提供语义标签以生成分类解释。重建模块虽然不是分类工作流程的一部分，但对于模型的可解释性非常重要，它可以将语义原型恢复到视觉空间，从而实现对人类友好的解释。输入图像首先通过图像编码器被编码为特征图，然后计算特征图中部分特征与学习到的语义原型之间的相似性，以进行最终预测。语义原型通过语义标签和图块可视化进行解释，其中语义标签由经过微调的视觉-语言模型标注，而原型的图块可视化则来自基于检索的、无参数的重建方法。SPANet 方法体系结构如图 1 所示。

文章在公开数据集 Caltech-UCSD Birds200-2011（CUB）和 Stanford Cars 上评估了提出的方法。在不同设置下的广泛实验表明，SPANet 在识别方面的性能几乎与不可解释的模型相当，同时还能为其决策过程生成清晰易懂的解释。

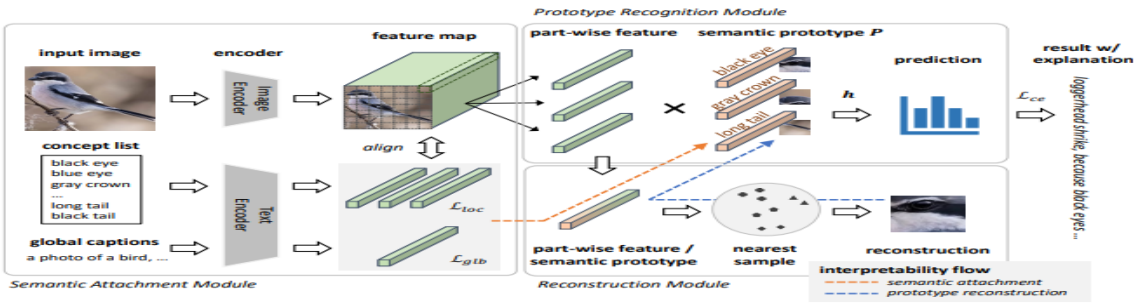


图 1 提出的 SPANet 示意图

责任编委 李策 樊鑫

好 文 推 荐

北京理工大学的最新研究成果“STQD-Det: Spatio-Temporal Quantum Diffusion Model for Real-time Coronary Stenosis Detection in X-ray Angiography”发表在 IEEE TPAMI 2024。

论文：Xinyu. Li, Danni AI, Hong. Song, et al. Hu, STQD-Det: Spatio-Temporal Quantum Diffusion Model for Real-time Coronary Stenosis Detection in X-ray Angiography, IEEE TPAMI, 2024.

近年来，X 线血管造影（X-ray angiography，XRA）作为冠状动脉疾病（Coronary Artery Disease，CAD）诊断的金标准，因其高分辨率使得动脉狭窄的定位变得更加精确。然而，医生在审查 XRA 图像时，常依赖个人经验，主观差异会影响 CAD 的诊断和治疗决策。因此，自动化狭窄检测显得尤为重要。但，快速、准确地实现冠状动脉狭窄的自动识别面临多个挑战，主要挑战包括血管在 XRA 图像中重叠，以及由呼吸和心跳引起的动脉移动和变形，容易导致狭窄的检测性能不足。同时，造影剂的不均匀分布和前景与背景之间的低对比度，存在量子噪声，也会导致狭窄的过检测。随着深度学习技术的发展，越来越多基于直接检测狭窄的算法被提出，尝试解决这些问题，但计算效率仍面临挑战。

扩散模型作为一种新兴的检测方法，在图像目标检测中展现了巨大潜力，因此该工作提出了基于量子噪声扩散的时空特征共享模型，用于实时检测 XRA 序列中的冠状动脉狭窄，所提出的模型如图所示。在该模型中将冠状动脉狭窄检测视为从噪声到检测框的去噪过程，通过前向扩散生成量子噪声框，然后通过反向扩散过程还原图像中的狭窄区域的检测框，从而完成冠状动脉狭窄检测。为了解决由于呼吸、心跳或图像质量差异导致的误检和漏检问题，时空特征共享模块通过在同一序列的不同帧之间共享正确的检测特征来提高检测精度和一致性，它将正确检测到的帧的特征传递给误检帧，从而重新检测这些帧。该方法的反向扩散过程分为三个阶段：首先，由狭窄检测解码器生成初步检测结果；然后，通过时空特征共享模块修正误检帧；最后，再次通过检测器重新检测误检帧进一步提升检测精度。该项工作提出的模型具有较高的检测速度，在实验中达到了每秒 25.08 帧的处理速度，能够实现实时冠状动脉狭窄检测。

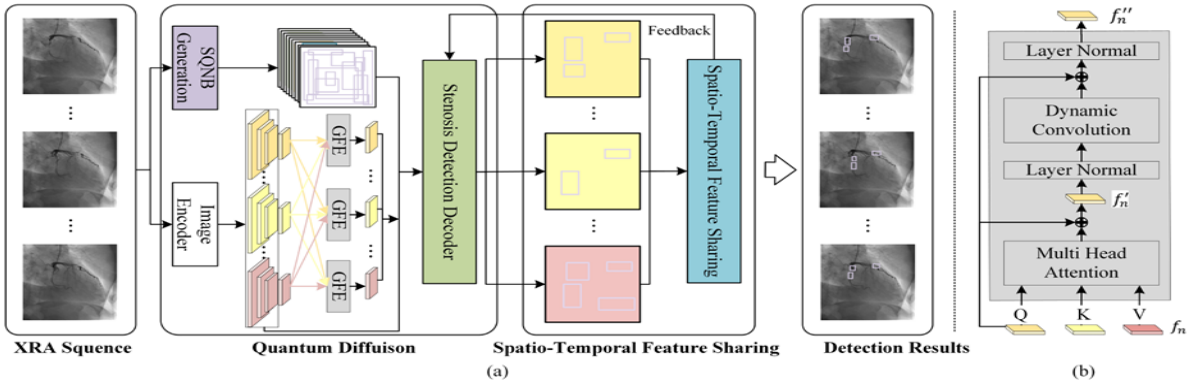


图 1 所提方法模型

责任编委 贾同 王田

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, Pattern Recognition 和 Image and Vision Computing。

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
ICLR 2025	2025.08.24-28	Singapore	2024.10.02	https://iclr.cc/Conferences/2025
NAACL 2025	2025.04.29-05.04	Albuquerque, New Mexico	2024.10.15	https://2025.naacl.org/
ECIR 2025	2025.04.06-10	Lucca, Tuscany, Italy	2024.10.10	https://ecir2025.eu/
AAMAS 2025	2025.05.19-23	Detroit, Michigan, USA	2024.10.17	https://aamas2025.org/
CVPR 2025	2025.06.10-17	Nashville, USA	2024.11.15	2025 Conference (thecvf.com)

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
IVC	Recent Advances in Computer Vision for Assisted Living	https://www.sciencedirect.com/journal/image-and-vision-computing/about/call-for-papers	2024.10.30
JBHI	Internet of Medical Things (IoMT) Based Healthcare Informatics Systems: Emerging Techniques, Challenges, and Future Directions	https://www.embs.org/jbhi/wp-content/uploads/sites/18/2024/08/JBHI_IoMT_Healthcare_SI-final2.pdf	2024.10.31
PR	Celebrating the Life and Research Contributions of Edwin Hancock	https://www.sciencedirect.com/journal/pattern-recognition/about/call-for-papers#special-edition-of-pattern-recognition-celebrating-the-life-and-research-contributions-of-edwin-hancock	2024.11.30
JBHI	Autonomous AI for Smart Healthcare	https://www.embs.org/jbhi/wp-content/uploads/sites/18/2024/08/CFP.pdf	2024.12.31

COMPUTER VISION NEWSLETTER

03 2024
总第 41 期

计算机视觉专委会简报

CCF 计算机视觉专委会