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

02 2023

总第36期

CCF 计算机视觉专委会

COMPUTER VISION NEWSLETTER

计算机视觉专委会 简报

2023 年第 02 期

总第 36 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

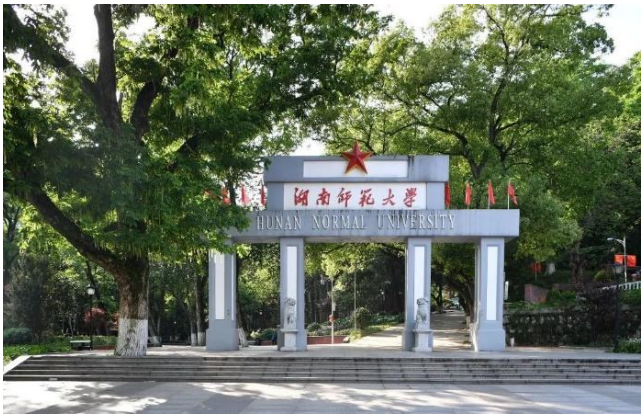
CCF-CV 走进高校系列报告会

第 123 期 武汉理工大学

2023 年 4 月 15 日上午，由中国计算机学会计算机视觉专委会 (CCF-CV) 主办、武汉理工大学承办的走进高校系列报告会第 123 期活动在武汉理工大学马房山校区会议中心成功举办。本次活动由武汉理工大学计算机与人工智能学院熊盛武院长和朱安娜副教授担任执行主席。

随后，北京大学查红彬教授、北京航空航天大学徐迈教授、大连理工大学樊鑫教授、中科院计算技术研究所王瑞平研究员做主题报告。最后，中科院计算技术研究所陈熙霖研究员及四位讲者参与了 panel 讨论环节。Panel 环节结束后，武汉理工大学计算机与人工智能学院向剑文副院长围绕本次系列活动“123”期的期数进行了活动总结，首先感谢了五位专家的精彩报告与学术交流分享，然后感谢 CCF-CV 专委会和来参加本次活动的校内外师生的大力支持，并欢迎各位专家学者能再次来武汉理工大学进行学术指导和交流，最后祝贺本次活动取得了圆满成功。

第 124 期 湖南师范大学



2023 年 5 月 28 日下午，由中国计算机学会计算机视觉专委会 (CCF-CV) 主办、湖南师范大学承办的走进高校系列报告会第 124 期活动在湖南师范大学中和楼 342 报告厅成功举办。此次活动由湖南师范大学信息科学与工程学院院长代建华和人工智能系主任王润民担任执行主席，副院长肖林、人工智能系主任王润民担任主持。

王润民副教授首先介绍了出席此次活动的嘉宾并对参会的专家及师生表示欢迎。随后，代建华院长致辞，他对中国计算机学会计算机视觉专委会专家们的到来表示诚挚的欢迎，简要汇报了学院的基本情况及发展状况，并预祝此次活动圆满成功。CCF 计算机视觉专委会常务委员白翔教授代表专委会致辞，他对主办方的精心筹备表示感谢，并向大家介绍了专委会的基本情况与计算机视觉研究的发展趋势，同时表示希望能够增强有关科研合作。

第 125 期 中原工学院

2023 年 6 月 3 日上午，由中国计算机学会计算机视觉专委会主办，中原工学院电子信息学院承办的“CCF-CV 走进高校”学术交流活动在 3 号组团楼学术报告厅隆重举行。本次活动的主题是“计算机视觉前沿技术及应用”，并邀请了厦门大学纪荣嵘教授、北京航空航天大学黄迪教授、上海交通大学严骏驰副教授等专家进行学术报告。中原工学院瞿博阳副校长、近二百名研究生导师及学生现场聆听了报告。本次报告由中原工学院科技处处长孙玉周主持。

中国计算机学会计算机视觉专委会主任、北京大学教授查红彬在致辞中表示，CCF 计算机视觉专委会的走进高校系列报告会，已经举办 124 届，在很多高校和研究所开展活动，受众遍及祖国大江南北，在学术界产生了积极反响，对于宣传计算机视觉前沿研究成果，推动相关领域的发展起到了重要的作用。感谢各位讲者对活动的大力支持，也对中原工学院为筹办这次活动所付出的努力表示感谢，并祝贺活动圆满成功。

责任编委 朱安娜

CCF-CV 走进企业系列交流会

第 26 期 清博智能

为了深入推动国内多媒体与计算机视觉领域的技术发展和产学研合作，中国计算机学会多媒体技术专委会（CCF-MM）、计算机视觉专委会（CCF-CV）携手清博智能，于 2023 年 5 月 18 日举办了“AIGC 赋能元宇宙”研讨会。中国计算机学会多媒体技术专委会秘书长、计算机视觉专委会秘书长等领导，以及来自中科院计算所、中国传媒大学、中国人民大学、北京清博智能智能科技有限公司等 20 余位高校和企业专家参加了本次活动，共同探讨多媒体、人工智能、计算机视觉等技术在元宇宙领域的“产学研”创新应用。

最后各位嘉宾就视频文字相结合的通用大模型的前景、难点进行了热烈讨论，同时提到了北京政府致力于整合算力打造超大规模 AI 模型训练平台的公示，对 AIGC 赋能元宇宙的未来发展充满了信心。

责任编委 潘金山

CCF-CV 走进河南省图书馆科普活动

2023 年 6 月 3 日下午，由中国计算机学会计算机视觉专委会（CCF-CV）主办、河南省图书馆和中原工学院联合承办的“人工智能与计算机视觉前沿技术与展望”科普活动暨“CCF-CV 走进河南省图书馆科普报告会”在河南省图书馆馆童梦剧场举行。

河南省图书馆副馆长申丽平致欢迎辞：本次活动是中国计算机学会计算机视觉专委会举办的首场科普活动，围绕“人工智能与计算机视觉前沿技术与展望”主题，分享和探讨这两个领域的最新研究成果、应用场景和未来发展方向，活动的成功举办，将促进河南省图书馆科普工作建设，推进科技资源科普化，拉近前沿科学与公众的距离。她向报告讲者专家们表示感谢，也感谢 CCF-CV 专委会、中原工学院在活动筹办过程中提供的大力支持！

本次科普活动搭建了读者和专家交流的平台，让参与活动的读者对人工智能和计算机视觉前沿技术有了更深入的了解，在开拓视野的同时，以提升专业能力来应对更多科技创新带来的机遇和挑战。

责任编委 毋立芳

第 16 期 认知启发的新一代人工智能技术

CCF-CV 视界无限系列研讨会

2023 年 3 月 26 日，由中国计算机学会计算机视觉专委会主办、西北工业大学承办的第 16 期 CCF-CV“视界无限”系列活动——“认知启发的新一代人工智能技术”研讨会在西北工业大学友谊校区国际会议中心举行。研讨会在西北工业大学自动化学院韩军伟院长的主持之下，由北京大学查红彬教授与国家自然科学基金委信息学部原常务副主任张兆田致辞。随后北京大学查红彬教授、南京信息工程大学刘青山教授、中国科学院自动化研究所张兆翔研究员、浙江大学李玺教授、西安交通大学陈霸东教授、天津大学朱鹏飞副教授、山东师范大学朱磊教授做主题报告。

查红彬教授的报告题目是“动态视觉与 SLAM：在线学习方法”。动态视觉目前还存在许多挑战，针对 SLAM 任务，查红彬教授介绍了 SLAM 的相关进展并且总结了其中存在的两个主要的挑战，包括缺少对时序连续性的关注和对 SLAM 问题的系统化定义。为了解决这两个挑战，首先通过 Flow-Based 方法对 SLAM 进行了系统性的定义，然后介绍了采用 Memory and Refinement 机制与在线自适应方案来解决 SLAM 系统在动态开放环境中的长时记忆问题与跨场景的适应性问题。在此基础之上，进一步介绍了利用概率图对建图结果进行表示与采用增量贝叶斯框架对概率化建图结果进行更新的理论与方法。

刘青山教授的报告题目是“基于视觉的情感计算”。视觉是情绪表达中的重要载体，占据交流的总情绪的 55%。在视觉情感计算任务中，刘青山教授首先介绍了三维卷积存在的全局时序信息缺失的问题，并受混沌理论中相空间重构方法启发，提出了相空间驱动时空表情特征学习。随后介绍了自适应多视角时空 Transformer 的 3D 人体姿态估计方法，该方法可以兼容图像与视频、

单目与多目的 3D 人体姿态估计任务，并且可以学习不同视角间的相互关系。

张兆翔研究员的报告题目是“类脑机器感知与学习”，他从神经结构建模、多路信息融合以及知识的归纳推理等方面介绍了团队中的一系列工作。针对神经结构建模，从神经元多样性建模和神经环路自适应选择两方面对网络结构进行改进，通过多样性建模和目标数据的通路调整可以在多种数据上提升泛化性能。在多路信息融合方面，他从多通路多模态信息协同和前向反向多路信息协同两方面介绍了团队的工作。针对知识的归纳推理，他介绍了点云引导的无监督物体发现和从运动信息中归纳发现物理概念知识两个工作，在物理规律归纳中可以从无标注的数据归纳出速度、加速度等物理概念。

李玺教授的报告题目是“视觉结构建模和特征学习”。对视觉的建模和特征学习是计算机视觉的重要研究内容。李玺教授从几何结构建模、动态推理结构、结构层面的图像生成等方面介绍了团队的相关工作。针对车道线检测任务，抛弃以往的密集预测方式，直接基于 anchor 的离散预测，极大地提高了准确率与实时性。然后，介绍了利用动态路由实现样本自适应推断的新方法和基于自然语言文本引导的视觉模型权重生成框架。最后，还介绍了在图像生成任务中的相关工作，提出了专为布局图输入设计的扩散模型图像可控生成算法 Layout Diffusion。

陈霸东教授的报告题目是“信息论学习：面向脑机接口和脑启发智能”。信息论是数学中的一个重要分支，在机器学习领域有广泛的应用。陈霸东教授首先介绍了信息论的基础知识，并且总结了信息论在学习算法中的两种应用范式：作为优化函数或者用于描述学习过程与学习机制。随后分别介绍了所提出的信息论的改进，包括最小误差熵准则等，与信息论在脑启发智能、人机交互等领域中的应用，包括点云配准任务、基于信息论学习的神经解码任务、脑信号的因果性分析任务等。

朱鹏飞副教授的报告题目是“认知启发的智能无人机系统协同进化 “。在过去的几年无人机数量迅速增加，无人机在日常生活、军事国防等领域都有重要应用。朱鹏飞副教授介绍了面向低空无人机的环境感知数据缺失问题、全天候鲁棒感知问题、多无人机协同感知问题与轻量级网络实时感知问题，并介绍了所搭建的开源生

态：一系列多源、多模态、多任务的大规模无人机视觉数据集。然后，针对数据算力受限、多机多传感器难配准和新场景泛化困难等难点，介绍了时空邻域近邻网络、不确定性感知的双光融合、跨机 Transformer 跟踪等工作，最后重点阐述了在视觉大模型方面的探索与思考。

朱磊教授的报告题目是“高效能跨模态检索“。人类的认知很多是通过多模态的方式实现，但海量的跨模态数据给检索系统的效率带来重大挑战。朱磊教授针对跨模态检索中信息损失大、多模态语义建模中跨模态关联缺失等研究挑战，介绍了面向无监督跨模态检索的关联-本体重构跨模态哈希框架用以缩小二值转化过程中的语义损失。并且介绍了位感知语义 Transformer 哈希方法用于建模多模态细粒度语义，并针对鲁棒语义建模问题提出了基于近邻感知补全学习的多模态哈希方法。最后，介绍基于查询的黑盒攻击方法用以生成对抗样本。

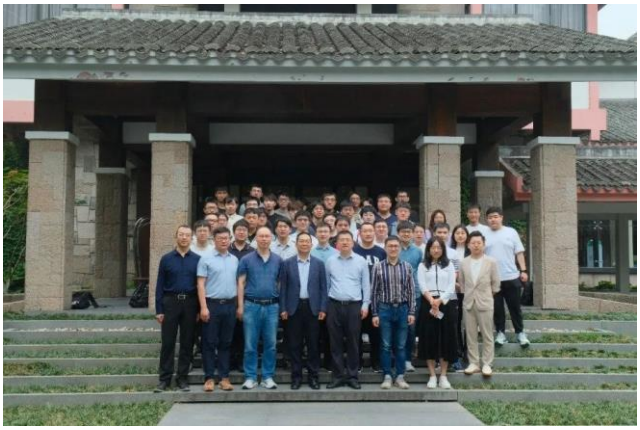
Panel 环节由西北工业大学张鼎文教授主持，参与嘉宾包括北京大学查红彬教授、南京信息工程大学刘青

山教授、中国科学院自动化研究所张兆翔研究员、浙江大学李玺教授、西安交通大学陈霸东教授、天津大学朱鹏飞副教授、山东师范大学朱磊教授。大家就“ChatGPT 系列是否具有类脑智能？类脑研究和 ChatGPT 系列未来的结合点和研究趋势是什么样的？”、“当前需要解决的基础大模型的应用瓶颈有哪些？面向边端小模型的研究工作应如何开展？”、“计算机视觉领域当前亟待解决的问题是什么，与国家战略密切关联的研究方向有哪些？”、“近年来人工智能领域许多突破性的成果出自工业界，如何更好的发挥高校团队在领域内的引领作用？”四个问题展开了全方位、多角度的深入讨论，与参加会议的老师同学们分享了广阔的研究视角与深刻的思想理念。

责任编委 杨巨峰

第 17 期 知识驱动的感知与生成技术

CCF-CV 视界无限系列研讨会



2023 年 5 月 21 日，由中国计算机学会计算机视觉专委会主办、浙江大学人工智能省部共建协同创新中心承办的第 17 期 CCF-CV“视界无限”系列活动——“知识驱动的感知与生成技术”研讨会在杭州市金溪山庄举行。研讨会在浙江大学杨易教授和吴飞教授的主持之下，由浙江大学庄越挺教授致辞。随后浙江大学庄越挺教授，北京大学彭宇新教授，大连理工大学卢湖川教授，西安电子科技大学邓成教授，清华大学黄高副教授，北京航空航天大学刘偲教授做主题报告。

庄越挺教授以“跨媒体智能的突破机遇”为题，生动地讲述了跨媒体智能从提出到现在的研究进展。他的报告从跨媒体智能的起源说起，梳理了上世纪 90 年代以来，从单媒体检索到跨媒体融合的发展历程。他先介绍了 Transformer 模型在跨媒体智能中的重要作用，以及基于 Transformer 的大规模多模态自监督预训练方法和 CLIP 模型如何实现视觉语言的细粒度语义对齐。他重点介绍了 HuggingGPT 项目，这是一个将 ChatGPT 从语言模型扩展到跨媒体智能领域的创新项目。HuggingGPT 项目依托 Huggingface 这一 AI 社区，通过 ChatGPT 作为控制器，Huggingface 作为执行者，能够完成复杂场景下的各种 AI 任务。最后结合 ChatGPT 和 AIGC 领域的最新进展，展示跨媒体智能从提出到现在所取得的巨大成就，并介绍了基于 Segment Anything Model 的通用人工智能近期的突破。

彭宇新教授的报告题目是“数据-知识协同驱动的细粒度多模态感知与生成”。他从人脑对世界的认知出发，回顾了跨媒体智能的由来，并结合跨媒体智能发展

的七大任务，介绍了关于细粒度多模态感知与生成的近期研究进展。细粒度图像分类任务存在诸多挑战，如精细的数据标注需要专业知识和大量人力，类内差异大和类间差异小等。他介绍了在该任务中的新算法，利用数据-知识协同驱动的方法，提升了分类的精度和鲁棒性，降低了对精细标注数据的需求。在细粒度视频检索领域，他提出一种基于自监督学习的视频片段指纹技术，以实现存储低、速度快、精度高的细粒度视频检索。最后重点介绍了在图文展示布局生成方面的新进展，提出了 PKU PosterLayout 数据集，设计序列生成网络，目前已被微软等多个高校和研究机构使用。

卢湖川教授报告的主题为“一网通吃 V2：视觉通用小模型”。通用语言模型（GLM）已经统一了几乎所有的 NLP 任务，而通用视觉模型（GVM）的研究在最近两三年才刚刚有起色。卢教授从 NLP 领域的大模型说起，介绍了如何用一个通用的视觉神经网络在不改变任何网络结构和参数的情况下实现视觉任务的大一统。他从视觉的通用表征学习和任务架构统一两个方面介绍目前通用视觉模型的最新研究进展，最后，讲述了视觉通用小模型，Unicorn 和 UniNext 在四个视频跟踪任务（SOT、MOT、VOS、MOTS）与十个实例级感知任务（Object Detection、Instance Segmentation、REC、RES、SOT、MOT、MOTS、VIS、SVOS、RVOS）上的大一统，实现一网通吃。

邓成教授报告的主题为“在线增量学习”。他的报告从生物智能和机器智能的差异出发，指出人是不断学习的，具有终身学习的能力，可以不断接纳新知识，并对以往知识进行整合。传统深度学习模型在增量学习场景

下面临巨大的困难和挑战，如灾难性遗忘、数据分布不平衡等。受到人类学习机制的启发，针对遗忘问题，他总结了目前增量学习的主要研究方向和方法，包括数据重放、正则化、参数隔离和基于大模型的提示学习，分别介绍了各自的优缺点和使用场景。他介绍了在增量学习方面的工作，针对在线增量学习的背景下，如何从数据流中挖掘有效的语义信息并选取重要样本进行数据重放，他提出补偿特征空间的语义差距，捕捉新旧类之间的关系以缓解遗忘，以上方法在多个主流数据集上都表现出了明显的性能提升。

黄高副教授报告的主题为“视觉大模型的高效训练”。他从训练大模型的挑战出发，就大模型训练过程存在收敛缓慢，优化不稳定，过拟合问题严重，计算开销巨大，传统训练方法难以并行化等困难，提出了分而治之和循序渐进两种高效的视觉大模型训练方法。分而治之的训练方法针对计算开销巨大和难以并行化等困难，提出将传统大模型端到端的训练拆分为数个小模型，实现在多个计算单元上并行训练。他分析了基于局部监督的训练方法的优势和不足，指出传统贪婪监督学习方法会导致中间特征可分性变好而最终结果性能下降等问

题，针对这一问题，他提出了一种保留中间特征和模型输入互信息，同时过滤任务无关的互信息的方法，在局部监督的训练方式上实现性能提升，达到和端到端训练可比的实验结果，同时大幅减低了计算开销。另外它展示了一种完全解耦的训练框架，局部模型训练过程无需交互，共用一个元模型，利用小模型孵化大模型，实现高效训练，还能提升优化稳定性，缓解过拟合问题，增加模型深度。循序渐进方法介绍了一种基于课程学习的训练方法，展示了两种课程设计思路，利用频域分析和数据增强的方式区分简单 pattern 和复杂 pattern，从简单的 pattern 开始学习，动态调整训练难度，有效提升模型泛化能力和鲁棒性。

刘偲教授报告的主题为“知识引导的跨模态感知与生成”。她结合了潘云鹤院士的多重知识理论，展示了在跨模态感知与生成方面的三项进展。她介绍了基于视觉知识的开放词表检测任务方面的工作，对于无标注的类别如何根据视觉特征进行分类。她提出使用预训练模型提取目标类别，引入了目标区域自适应变形机制，实现对未训练类别的有效标注。另外她介绍了基于知识图谱的视觉-语言导航系统，任务希望实现根据自然语言指令，机器人在位置环境中自主导航找到特定目标。她提出利用额外知识，结合物体类别之间的语义关系，物体之间的空间共现关系，构建、推理知识图谱，实现对特定目标精准有效的定位。最后她介绍了基于符号知识的视频背景音乐的生成，提出结合符号化乐理知识设计新的音乐表示形式，相比传统的音乐生成方法，有效缩短了建模长度，将复杂的音乐结构解耦成和弦、伴奏、旋律进行层次化建模，用结构化的知识形成音乐。

Panel 环节由浙江大学朱霖潮研究员主持，参与嘉宾包括北京大学彭宇新教授、西安电子科技大学邓成教授、北京航空航天大学刘偲教授、清华大学黄高副教授、北京交通大学魏云超教授、浙江大学王文冠研究员。大家就“数据和知识双轮驱动的技术中，知识通常包括哪些类型，有怎样的表现形式？针对不同任务，我们要如何去构建和选择所需要的知识？”、“在大模型时代，结合了知识的人工智能技术相较于纯数据驱动的方法，有哪些方面的优势和劣势？随着大模型的不断发展，知识嵌入会不会失去其必要性？”、“通用知识是指能被广泛应用于多个领域和问题的知识；专用知识是指针对特定领域和问题的知识。在当前的人工智能发展中，通用知识和专用知识的哪一方更具有发展潜力？如何实现通用知识和专用知识的结合？如何平衡通用知识与专用知识的应用？”、“大模型时代下，高校研究者应该如何更好地拥抱大模型和适应越来越快的研究节奏？高校研究的特色如何凸显？”四个问题展开了全方位、多角度的深入讨论，与参加会议的老师同学们分享了广阔的研究视角与深刻的思想理念。

责任编委 杨巨峰

CCF-CV 常务委员会 2023 年度第一次工作会议顺利召开

2023 年 4 月 15 日于武汉召开中国计算机学会计算机视觉专委会 (CCF-CV) 常务委员会 2023 年度第一次工作会议，本次常委会工作会议由专委会主任查红彬教授主持，专委会顾问委员会委员陈熙霖研究员和常委会委员参会，秘书处全体成员列席。

首先，专委会党小组组长、副主任刘青山教授组织了党小组学习。

接下来，专委会主任查红彬教授带领常委会委员，就专委会领导机构换届相关规定和流程、专委会十周年纪念活动策划、PRCV2023 会议筹备情况、RACV2023 会议和第二届计算机前沿讲习班筹备情况，以及第一届中国工业视觉大会 (CIVC2023) 筹备情况等议题进行了讨论，形成了具体可行的指导性建议。

最后，查红彬主任作了总结发言。会议在紧张而有序的热烈讨论氛围中结束。

责任编委 黄岩

专题综述

自动驾驶场景多模态融合感知

上海交通大学　马超

近年来，自动驾驶相关技术逐渐成为了当前的研究热点，其目的是辅助或者代替人类进行交通工具的操控。相比于传统的人工驾驶，自动驾驶在安全性、便利性、高效性等诸多方面具有显著的优势，这对交通运输和生产安全有着重大意义。自动驾驶的技术栈包括软硬件的结合，其中软件部分包括： 1）环境定位模块； 2）环境感知模块； 3）决策规划模块。作为自动驾驶系统中至关重要的一环，环境感知算法对后续的决策和路径规划至关重要，吸引了大量研究人员的关注。在自动驾驶场景中，激光雷达和相机传感器是两种常用的环境感知传感器。一方面，激光雷达通过获取 3D 空间中的激光点云，提供高精度的定位，但是稀疏无序的点云缺少表观纹理特征。另一方面，相机提供的 RGB 图像能为目标检测提供丰富的语义信息，但是由于缺乏深度估计，很难对物体的 3D 位置进行准确的预测。

基于上述背景，在业界，不同公司针对自己的业务特点和技术理解，面向不同的自动驾驶级别选择了不同的传感器配置和算法方案，其主要分为两派：纯视觉感知方案以及多模态融合方案。以 Tesla 公司为代表的一派主张采用纯视觉感知方案，仅依靠来自多个摄像头的图像进行环境感知与决策。纯视觉方案更接近人类驾驶直觉，成本相对较低但容易受到环境中光照等因素的影响；另一派以谷歌 Waymo 为代表采用多传感器融合的方案，通过对摄像头、激光雷达等传感器数据融合进行环境感知，这使得车辆对物体的位置、距离和大小的感知更加准确，但由于多传感器融合方案要求配备更多的传感器设备并对计算芯片有更高的算力要求，因此成本较高。然而，随着技术的发展，激光雷达成本不断降低，十年间车用激光雷达价格从 10 万美元价格区间已经下降到 100 美元区间，这使得多模态融合感知算法逐渐显示出其优势。相比于纯视觉方案，使用多模态融合方案的优势主要体现在三点：

- 安全优势：人类司机驾驶仅仅依靠双眼而无需使用雷达进行辅助是各公司主张纯视觉感知方案的主要依据。然而，纯视觉方案虽然更接近人眼感知直觉，但在当前，人工智能与人类感知仍相差巨大，且正如人眼会存在误判所导致的每年有大量交通事故的发生，纯视觉算法也在许多场景下存在误判，而安全性对自动驾驶车辆至关重要。因此，自动驾驶技术首要任务是实现更安全的驾驶而非仅仅是模拟人类司机驾驶习惯，多模态方案中激光雷达的加入可以大幅度提高感知算法对障碍物的检出能力，保证更为安全可靠的自动驾驶技术的实现。
- 数据需求优势：自动驾驶场景是一高度复杂的开放环境，纯视觉感知算法通常需要使用大规模的不同种类和场景的真实驾驶数据进行训练，以使模型应对多变的驾驶场景。而激光雷达通过发射多种激光并接收反射光的方式检测障碍物，这减轻了对数据的依赖，更具可靠性与可解释性。同时，采用多模态融合方案一定程度上规避了大型公司的数据垄断带来的技术垄断，能有效促进自动驾驶技术相关研究更为快速的发展。
- 鲁棒性优势：不同传感器具有不同的优点和局限性。例如，激光雷达可以提供准确的距离信息，但在雨雪等恶劣天气下可能会受到影响。而摄像头则可以提供更丰富的视觉信息，但在低光等条件下可能会

受到影响。通过将不同传感器的信息融合起来，可以弥补单一传感器的局限性，不同传感器采集的信息可以互相补充，从而形成一个更全面、更准确的场景理解。同时，激光雷达对目标距离和速度的准确感知能力有助于自动驾驶车辆做出更为准确的驾驶决策，提高场景感知的鲁棒性。

总之，多模态融合方案可以利用不同传感器间的互补性，弥补单一传感器的局限性，提高场景感知的准确性和可靠性，并且更好地适应复杂场景，因此在自动驾驶等场景感知应用中具有重要的优势。 3D 目标检测和跟踪算法能够对环境中的物体进行分类定位和关联，是环境感知的核心环节。目前自动驾驶场景下的多模态融合感知算法主要集中在研究相机与激光雷达两种模态下的环境感知。有效地融合这两种互补的模态，不仅能提升检测任务的精度，也能够丰富跟踪任务中的匹配线索。

一、多模态融合3D检测

1.1. 多模态数据融合

如图 1 所示，目前已有的基于多模态 3D 检测方法，根据两种传感器的融合阶段可以分为：1）结果层级的融合，2）候选框层级的融合和 3）点云层级的融合：

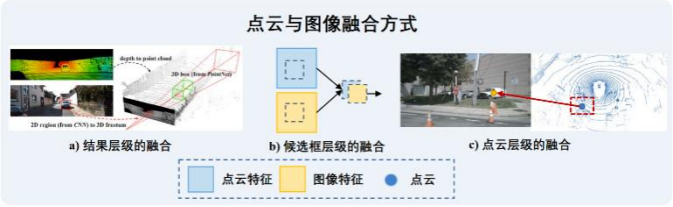


图 1　三种点云与图像融合方式

结果层级的融合方法直接利用 2D 图像上所获取的二维候选框区域与对应的点云特征进行融合。2018 年，Qi 等人提出了 Frustum PointNets[1]。该方法首先直接利用已有的 2D 目标检测器在图像上获取 2D 候选框，并将对应到 2D 候选框内的点云视锥（frustum）裁剪出来。接着将视锥的点云经过一个类似 PointNet 的模块获取分割信息并过滤前景点以进一步用于 3D 候选框的回归。F-ConvNet[2]在此基础上进行了改进，实现了更细粒度和多层级的点云特征提取，从而获得更优的检测性能。Frustum PointNets 系列方法借助成熟的 2D 检测算法提供一定程度上的先验知识，从而减小了 3D 搜索空间。然而，级联方法的缺点是它严重依赖于 2D 检测器提供候选框从而导致漏检情况。

基于候选框的方法在候选框层面上以 ROI pooling 的形式融合点云与图像特征，从而进一步的优化候选框。MV3D[3]和 AVOD[4]是基于候选框融合的典型方法。为了更好的对特征进行融合，如图 2 所示，工作[5]提出一两阶段融合框架，在第一阶段通过分别提取图像与点云特征并利用一种联合 anchor 机制生成区域候选框，在第二阶段进一步融合生成的 2D 与 3D 候选框中的密集特征。尽管基于候选框的融合方法更好的利用了不同模态的信息，但这类方法往往存在速度慢且计算量大的问题。

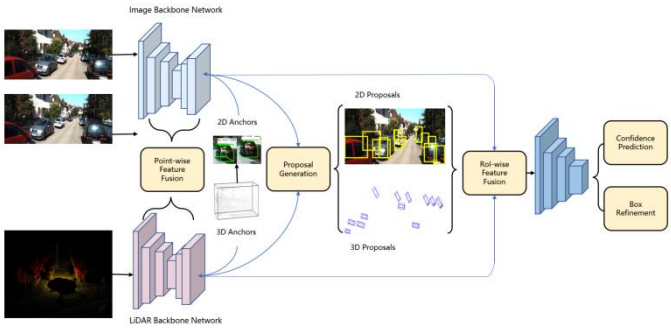


图 2　相机与激光点云两阶段融合框架[5]

点云层级的融合则是直接建立点与图像的关联，将图像信息增强到点云上，作为点云的额外输入特征。其中一部分方法[13-16]的目标是构建一个 BEV 的图像特征，在和点云特征结合后，在鸟瞰图上做 3D 检测，这种直接构建 BEV 特征的思路常常会引入特征模糊。目前在多模态融合方法上，采用更多的是点到点的思路，即直接将点云投影到图像上并抓取相应的图像特征后以点对点的方式直接增强到点云上。这种方式直接建立了点云与图像的联系，减少了信息的损失。PointPainting[6]利用来自图像的语义分割信息来增强点云的输入。如图 3 所示，该方法首先利用 2D 分割网络对每个图像像素进行分类，然后将点云直接投影到分割掩码中，将对应的分割分数作为图像特征附加到每个点云上。最后，将“绘制”后的点云部署到任意纯点云 3D 检测器中用于定位和分类。

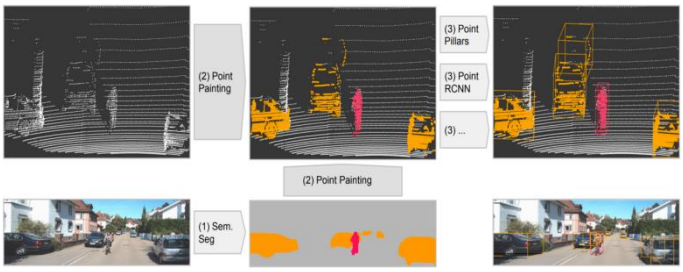


图 3　PointPainting 算法框架图[6]

尽管 PointPainting 取得了明显的改进，但是分割信息只为点云提供了类别，并没有充分利用图像信息，可以认为是一种紧致但次优的图像表征。直观上，图像的高维 CNN 特征包含了更丰富的外观纹理信息和更大的感受野，应该更适合与点云融合。因此，PointAugmenting[7]方法选择图像特征作为增强点云的输入，模型的网络结构如图 4 所示。该方法在点云检测器 Centerpoint[17]的基础上，主要做了两点改进用于融合图像和点云信息：首先，基于先前的结论，将点云投影到图像上，抓取点云相应的图像特征并增强到点云上；此外，考虑到点云与图像特征之间存在数据特性与数据分布的差异，在 3D backbone 中额外添加了一个图像分支，用于处理图像特征，最后两个模态在 BEV 特征上进行融合。

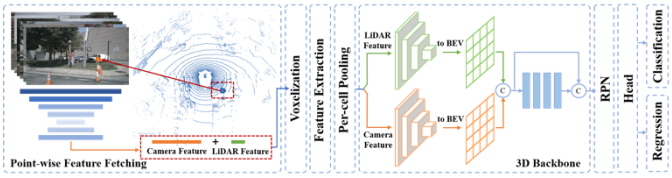


图 4　PointAugmenting 算法框架图[7]

1.2. 多模态数据增广

另一方面，数据增强是视觉感知领域提升感知精度的最为重要的手段，而由于模态间数据的差异，大部分数据增强方法无法直接迁移到多模态场景。例如，在点云检测器的训练中，常常采用 GT-Paste[8]增强方法将其他场景中的物体粘贴到当前训练场景。GT-Paste 可以缓解数据集的类别不平衡问题，并加速模型的收敛。但是这种有效的数据增强方式并不能直接迁移到多模态的场景中，因为这种粘贴方式会破坏点云和图像之间的一致性关联。因此 PointAugmenting 同时提出了一种数据增强方式，能够使得 GT-Paste 适用于多模态的检测器。如图 5(a)和(d)所示，汽车存在于原始场景中，它的点云由黄色表示；骑行者是当前想要粘贴到当前场景的虚拟物体，它的点云由绿色表示。从观察者的角度来看，由于粘贴的自行车在原始 3D 场景中被汽车部分遮挡，导致在图像视角上，两个物体的图像块存在部分重叠。如果直接将虚拟物体的图像块粘贴到图像上，如图 5(b)和(e)所示，则在图像块的重叠区域中，投影的物体点云可能会获取不匹配的图像信息。为解决此问题，PointAugmenting 保持物体之间的遮挡关系，并从观察者的角度过滤那些被遮挡的点云。对于图像数据，所有虚拟物体和原始物体将按照由远到近的顺序，将其对应的图像块粘贴到原始图像上，从而与点云的遮挡关系保持一致。

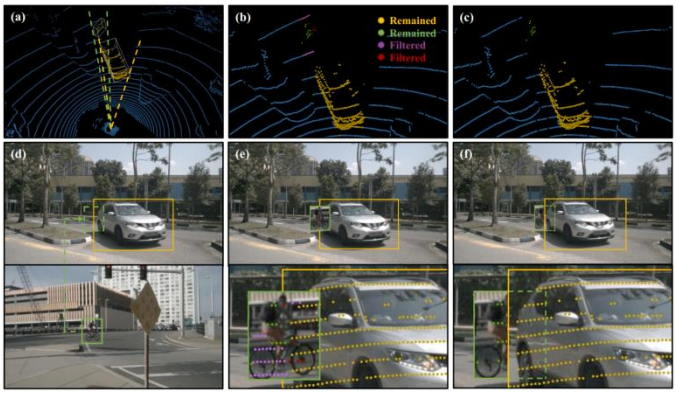


图 5　多模态数据增强示例

基于上述信息融合与增强方法，PointAugmenting 在 nuScenes 和 waymo 数据上都取得了明显的精度提升，在 nuScenes 上达到了同期最好的检测结果，比纯点云的基线算法 CenterPoint 精度提升了 6.5 个点。

1.3. 时序多模态融合

时序多帧数据为多模态融合感知提供多视角、运动特征等额外信息，能够进一步提升物体检测精度。目前大多数方法对时序的利用是将来自不同帧的输入投影到同一关键帧中，以起到增强点云稠密度的作用。但是简单的投影合并在运动场景中存在明显的拖尾效应，因此只适合融合时间跨度小的非关键帧。为了获取更长时间序列中更丰富的信息，设计合理的时序数据的融合方

案是非常有意义的。尽管最近的工作[9]对学习时序多模态模型进行了早期尝试，但实际上，它是使用点连接的预处理方案进行时间融合，也就是把时序信息和多模态信息融合的建模视为独立的两个部分。相比之下，对时序多模态信息进行显式的融合建模的方法更有利于充分利用未对齐的互补信息。

为了解决上述问题，工作[10]提出了一种时序多模态融合模型(LiDAR Image Fusion Transformer，LIFT)，可以直接学习四维时序多模态信息的相互对齐。具体来说，如图 6 所示，所提出的模型包含一个格状特征编码器和一个跨模态垮时间注意力模块。首先，在格状特征编码器中，LIFT 获取对应点的相机特征并进行柱状特征提取以将激光雷达点和逐点相机特征投影到鸟瞰图表示上。通过保持相对较少数量的网格，LIFT 能够有效地计算网格间的相互交互和网格内的细粒度注意力。同时，该项工作通过设计一种 4D 位置编码来进行时序多模态数据的四维位置定位。进一步地，为了减少自注意力模块的计算开销，LIFT 还设计了稀疏的窗口分区机制进一步舍弃不含激光点的空窗口区域来降低计算量，并构建了金字塔上下文结构以扩大特征感受野。

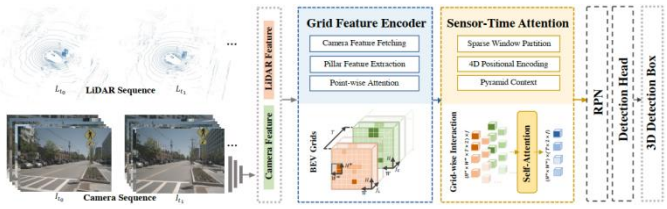


图 6　时序多模态融合 3D 目标检测 LIFT 框架[10]

在之前单帧的点云检测模型中，一种基于随机物体粘贴的数据增强方案因为能够增强训练数据的多样性，被验证对模型训练很有帮助。但是在时序多模态数据中，普通的粘贴方案无法保持跨模态和跨时序的一致性，因此如图 7 所示，LIFT 扩展了传统数据增强方案，把单个点云物体的粘贴扩展为序列点云和对应的序列图片块的粘贴，从而保证了粘贴在时序多模态训练数据中的增强物体保持了时空一致性。在 nuScenes 数据集上进行的大量定性与定量实验表明时序信息对多模态检测的促进作用。

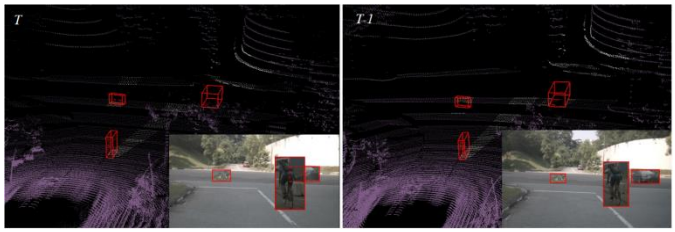


图 7　跨模态垮时序数据增强可视化

二、多模态融合3D跟踪

通常在检测跟踪框架中，检测结果与轨迹被表示成和位置有关的点，并假设连续帧之间的位置变化在局部区域内，进而用点位置距离进行检测结果和轨迹的关联计算。尽管取得了不错的效果，由于缺乏丰富的表观信息，这种仅依赖位置进行匹配的方法在较大运动变化和存在噪声检测结果的情况下往往会失败。多模态 3D 目标检测与跟踪算法 AlphaTrack[11]同时考虑 3D 物体位置和表观变化。如图 8 所示，该方法用独立的 3D 卷积网络分别提取两个模态的特征，并在点级别结合激光雷达点云和对应的图像特征；其次，为了获取 3D 对象的表观信息，以在跟踪算法中能够显式地利用表观线索，该方法为模型附加了一个表观分支以学习实例级的表观特征。最后，该工作进一步提出了一个三阶段跟踪算法以在算法中同时利用位置线索和表观线索进行匹配关联。第一阶段是基于位置的匹配，以中心点距离作为位置相似度初步匹配检测和跟踪轨迹；然后 AlphaTrack 以表观特征的余弦距离作为表观相似度对第一阶段的结果进行合理性筛选，即将相似度排序靠后的匹配对予以删除，最终在第三阶段中用表观特征对剩余未匹配对进行重匹配。通过以上三个步骤，实现了在跟踪关联中显式地利用位置和表观两种信息。AlphaTrack 在 nuScenes 数据集 3D 跟踪上取得同期最佳结果（在排行榜上占据榜首位置长达一年）。

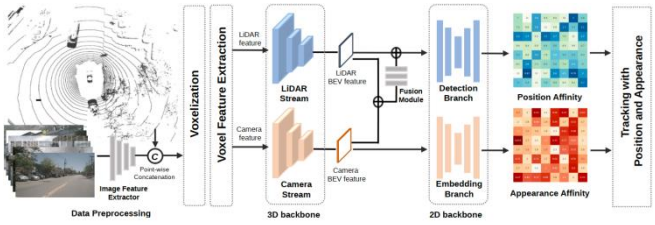


图 8　AlphaTrack 整体架构[11]

三、通用跨模态知识蒸馏

基于上述相关研究工作背景可以发现，在 3D 目标检测器中，单模态检测器模型简单但检测精度较低，而多模态检测器检测性能好但系统复杂度高。跨模态知识蒸馏方法可以根据已有模型进行知识迁移以提高目标模型检测精度，然而现有的跨模态知识蒸馏框架限定了教师与学生检测器的模态，即限定了教师和学生的模态为 LiDAR 和 camera 以及 Fusion 和 LiDAR，而现实中不同检测器都有广泛应用场景，因此应用范围受到了限制。从以上问题出发，构建统一的知识蒸馏框架，对教师和学生的任意模态组合均适用变得越来越重要。通过对不同模态检测器的结构进行分解寻找一致的中间特征表示，可以发现现有的表现较好的检测器通常会在 BEV 视角下进行检测，并且具有统一的流程。如图 9 所示，检测器首先对输入数据提取特征，并将特征投影到 BEV 视角下，得到 low-level 的 BEV 特征，然后对该特征进一步处理，得到 high-level 的 BEV 特征，最后由一个检测头进行预测，输出 response 特征用于生成预测结果。基于以上的统一分解形式，对于不同模态的教师和学生检测器，就可以在这些一致的中间特征表示上进行知识蒸馏。

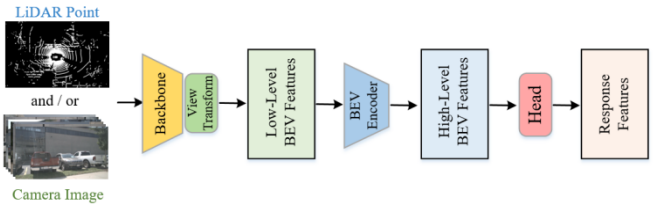


图 9　3D 目标检测器通常遵循统一的流程

针对上述背景，工作[12]提出一通用跨模态知识蒸馏框架 UniDistill。如图 10 所示，该方法提出三种蒸馏损失，分别为特征蒸馏(feature distillation)、关系蒸馏(relation distillition) 以及响应蒸馏(response distillation) 模块。首先，为消除背景影响以及平衡不同尺度的检测框对蒸馏损失的占比，在特征蒸馏与关系蒸馏中，UniDistill 对每一个 GT 检测框只选取了关键的 9 个点以分别对 low-level 的 BEV 特征以及余弦相似度进行对齐。同时，由于在检测框中心的预测值较为准确，因此对于响应蒸馏部分，UniDistill 针对每一个 GT 检测框中间的一个高斯掩膜区域的特征进行了对齐。

在 nuScenes 数据集对四种教师与学生的模态组合方式 (Fusion→LiDAR/Camera, LiDAR→Camera, Camera→LiDAR)进行验证，实验结果表明 UniDistill 对提高学生检测器性能的优异表现。

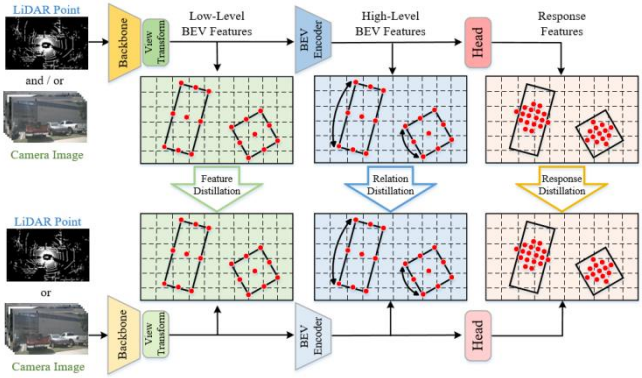


图 10　UniDistil 算法整体框架[12]

四、总结与展望

本文首先从当前自动驾驶的感知模块选用纯视觉感知方案还是多模态融合方案这一热点问题出发，对视觉感知中多模态融合方案的优势进行了探讨。围绕自动驾驶场景的多模态融合感知，分别介绍了多模态融合 3D 检测与跟踪领域的当前研究进展，并介绍了跨模态知识蒸馏方法帮助检测模型提高精度的进展。基于目前已取得的实验结果和结论，笔者认为多模态融合感知领域还存在较大的深入研究空间，比如以下四个方向：

- 数据融合算法的优化：现有的跨模态融合方案大都是基于投影关系进行的，这样的融合方案虽然在实际应用中是计算简单、并验证有效的，但考虑到不同传感器数据之间存在一定的标定误差，投影方案不可避免地也会引入一定的误差，此外不同模态之间的信息可靠度在不同的应用环境下也是不稳定的。因此如何进一步优化跨模态数据的融合是非常有价值的，具体来说，可以考虑以下两种优化方向：1) 优化投影关系的准确度。传感器间的投影误差是有可能通过深度学习方案缓解的；2) 优化模态间的信息衡量标准。不同模态的特征可靠性是可以通过一定的衡量标准进行评估的，以此提供在不同环境下更合理融合的参考。

- 基于主动学习、迁移学习、自监督学习等方案的数据标注优化研究：3D 数据的标注对于感知模型的训练是至关重要的，然而精确的 3D 数据标注需要高昂的人力成本。因此存在两种降低标注成本的方向，一是降低标注的数目，即在无标注或部分标注的情况下也能取得性能优秀的感知模型；二是改善粗糙标注的质量，即通过深度学习方案自主提升低质量标注的精度，以降低精确标注的成本。其中主动学习作为一种自动挖掘错误标注、难标注的方案，在辅助进行高质量标注的过程中能够提供一种低成本的技术支持。此外迁移学习能够帮助模型提升在无标注数据域上的泛化性能，更轻便简洁的迁移方案对实际应用具有重要研究价值。
- 感知与决策的耦合：在自动驾驶系统中，感知和决策是紧密耦合的。多模态感知提供了丰富的环境信息，但同时也增加了感知与决策之间的耦合程度。为了实现高效的自动驾驶，需要在感知和决策之间建立有效的信息传递机制。使得感知与决策模块相互配合，这方面的研究可分为两个方向：1) 研究如何将多模态感知的结果进行有效的融合和分析，以支持决策模块的实时决策，这需要形成对道路拓扑结构、高精度地图以及动态障碍物的联合感知与理解；2) 研究统一的端到端感知与决策联合预测模型。对上述方向的推进有助于实现简洁高效的自动驾驶统一框架。
- 多模态感知的安全性和可靠性：自动驾驶技术是一个非常复杂的系统，需要确保其在各种情况下都能够安全、可靠地运行。因此，在设计和优化多模态感知算法时，必须考虑到各种异常情况和边界条件，并采取相应的措施以确保系统的安全性和可靠性。在自动驾驶中，不确定性来源于多个方面，包括传感器的噪声、遮挡、动态环境等，这可能会导致自动驾驶系统出现错误的决策，甚至造成事故。为了实现稳定可靠的自动驾驶，需要研究如何在多模态感知中有效地管理不确定性。因此，需要采用多种冗余机制来确保系统的可靠性和安全性。这方面的研究可以从两方面进行：1) 研究如何在数据融合过程中考虑不确定性；2) 研究如何在感知和决策之间传递不确定性信息，以支持决策模块在不确定性条件下进行决策。

随着自动驾驶技术的不断发展和应用，多模态感知技术作为其中最核心的一部分，也将不断得到完善和优化。目前，该领域仍然存在的亟待解决关键问题具有重要研究价值。通过不断的创新和发展，自动驾驶系统将实现更高的环境感知精度和决策能力，进而实现高效、精准、可靠和安全的环境感知技术。

责任编委 王金甲

参考文献

[1] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, Leonidas J. Guibas. Frustum pointnets for 3d object detection from rgb-d data, CVPR 2018.

[2] Wang, Zhixin, and Kui Jia. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection, arXiv:1903.01864, 2019.

[3] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, Tian Xia. Multi-view 3d object detection network for autonomous driving, CVPR 2017.

[4] Ku J, Mozifian M, Lee J, et al. Joint 3d proposal generation and object detection from view aggregation, IROS 2018.

[5] Ming Zhu, Chao Ma, Pan Ji, Xiaokang Yang. Cross-Modality 3D Object Detection, WACV 2021.

[6] Vora S, Lang A H, Helou B, et al. Pointpainting: Sequential fusion for 3d object detection, CVPR 2020.

[7] Chunwei Wang, Chao Ma, Ming Zhu, Xiaokang Yang. PointAugmenting: Cross-Modal Augmentation for 3D Object Detection, CVPR 2021.

[8] Yan, Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. Sensors, 18(10):3337, 2018.

[9] Piergiovanni A J, Casser V, Ryoo M S, et al. 4D-Net for Learned Multi-Modal Alignment, ICCV 2021.

[10] Yihan Zeng, Da Zhang, Chunwei Wang, Chao Ma, et al. LIFT: Learning 4D LiDAR Image Fusion Transformer for 3D Object Detection, CVPR 2022.

[11] Yihan Zeng, Chao Ma, Ming Zhu, Zhiming Fan, and Xiaokang Yang, Cross-Modal 3D Object Detection and Tracking for Auto-Driving, in IROS 2011

[12] Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, Chao Ma, UniDistill: A Universal Cross-Modality Knowledge Distillation Framework for 3D Object Detection in Bird's-Eye View, CVPR 2023.

[13] Ming Liang, Bin Yang, Shenlong Wang, Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection, ECCV 2018.

[14] Ming Liang, Bin Yang, Yun Chen, Rui Hu, Raquel Urtasun. Multi-task multi-sensor fusion for 3d objectdetection, CVPR 2019.

[15] Yoo J H, Kim Y, Kim J, et al. 3D-CVF: Generating Joint Camera and LiDAR Features Using Cross-View Spatial Feature Fusion for 3D Object Detection, arXiv:2004.12636, 2020.

[16] Tianwei Yin, Xingyi Zhou, Philipp Krahenbuhl. Center-based 3d object detection and tracking, CVPR 2021.

马超

上海交通大学人工智能研究院长聘轨副教授，博士生导师。上海交通大学与加州大学默塞德分校联合培养博士。2016 至 2018 年澳大利亚机器人视觉研究中心(阿德莱德大学)博士后研究员。中国图象图形学学会优博，上海市浦江人才。主要研究方向是多模态物件检测与跟踪。担任中国图象图形学学会青年工作委员会副秘书长、中国图象图形学学会优博俱乐部轮值主席。谷歌学术总引用近 1 万次，自 2020 年起连续入选爱思唯尔中国高被引学者。研究成果应用于华为达芬奇芯片及其无人驾驶 MDC 平台，获华为技术合作领域 2021 年度优秀技术成果奖。

Email: chaoma@sjtu.edu.cn

专题综述

重新思考点云配准中的生成和选择过程

陈志[1]，孙琨[2]，杨帆[1]，郭琳[1]，陶文兵[1]
[1] 华中科技大学 [2] 中国地质大学（武汉）

一、摘要

误匹配消除是基于特征的点云配准方法的重要步骤。本论文重新思考了基于传统的 RANSAC 方法中的模型生成和模型选择过程。具体来说，对于模型生成，本文提出了一种新的二阶兼容性度量 (SC^2) 来计算匹配对之间的相似性。本文从概率的角度证明了该度量能够极大地降低模型生成中采样一致性集合时采样到误匹配的概率，从而提升模型生成过程中的采样稳定性。对于模型选择，本文提出一种特征和空间一致性引导的截断倒角距离 (FS-TCD) 来评估生成模型的质量。它综合地考虑了全局对齐质量、几何信息和特征信息，缓解了现有度量过度依赖特征匹配准确性的问题。所提出的方法在室内以及室外数据集上取得了当前最好的配准性能。论文已被 TPAMI 2023 收录，代码已开源：https://github.com/ZhiChen902/SC2-PCR-plusplus。

二、引言

三维刚体点云配准旨在恢复两个具有重叠区域的点云之间的刚体变换。常用的基于特征的方法先为点云中的每个点提取局部特征描述子并建立粗略的点云对应关系，然后通过稳健的模型估计算法，从含有误匹配（外点）的粗匹配关系中寻找正确匹配（内点）并估计刚体变换。本研究主要关注如何在粗匹配中含有大量错误匹配的情况下进行点云配准。

RANSAC[1]最早采用迭代的策略来进行模型估计。然而，它需要大量的采样来保证算法的收敛，并且在内点率过低的情况下并不能保证一定能找到正确解。一些方法通过空间兼容性来解决 RANSAC 的问题，它们利用变换的刚体属性，即：空间中任意两个点经过刚体变换之后长度不变。因此，一阶空间兼容性度量认为如果两个匹配对之间的空间距离差越小，他们的相似度越大。由于正确匹配对（内点）之间的空间距离差理论上应为 0，这样两个内点的相似性就会比较大，从而在内点之间形成聚类效应来方便采样。

然而一阶空间兼容性存在模糊性，即外点也有可能与内点有很高的相似度，如图 1(b)中黄色底纹的方格。当前许多研究利用传统方法[2,3]或深度学习框架[4,5]来减轻模糊性带来的问题，虽然一定程度上缓解了这一问题，但是在内点率很低的情况下有时也会失效。为了解决这一问题，我们提出一个二阶兼容性测度来度量两个匹配对之间的相似性。具体来说，我们首先二值化一阶空间

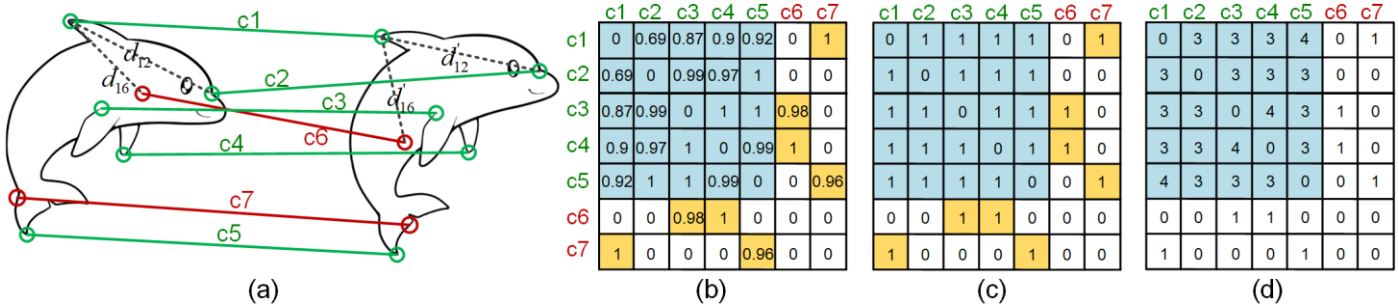


图 1　二阶空间兼容性动机及示意图

兼容性测度（1 代表两个匹配对是兼容的，0 代表不兼容），如图 1（c）。然后对于任意两个兼容的匹配对，我们计算与他们共同兼容的匹配对的个数（可以表示为兼容性的二阶形式）作为他们的相似性。由于内点之间都是相互兼容的，因此任意两个内点之间的相似度至少为所有匹配对中内点的个数（除去这两个匹配对自身），而内点和外点之间并没有这样的性质。如图 1(d)所示，在所提出的二阶兼容性度量中，内点和外点之间的高相似性被抑制了。我们从概率的角度证明了二阶兼容性度量发生模糊性事件的概率远小于一阶度量，证明了它能够更稳定地用于内点采样，从而提升模型生成的效率和鲁棒性。

除了模型生成，另一个问题是如何从生成的多个模型中选择出最好的模型。当前的方法通常采用内点计数（IC）的方式，即利用生成的模型对齐两个点云，然后统计对齐误差小于某一阈值的匹配点对个数作为 IC 值，并选择 IC 值最大的模型作为最终结果。然而，这种选择方式依赖初始匹配对的准确性。当初始匹配中准确匹配对个数较小时，即使是估计出了正确的模型，它对应的 IC 值可能也比较小，使得它无法被选择为最终的结果。为了解决这个问题，本文提出了一种特征和空间一致性引导的截断倒角距离 (FS-TCD) 作为模型选择的度量。它从倒角距离这种全局度量出发，通过引入特征信息和空间一致性信息的引导，解决了直接将倒角距离作为模型选择度量的低效和不稳定因素，同时保留了倒角距离的全局性。由于全局信息的引入，该度量也缓解了 IC 度量对于初始匹配对的依赖性。

基于所提出的二阶兼容性 (SC^2) 和特征和空间一致性引导的截断倒角距离（FS-TCD），本文重新构建了一个点云稳健姿态估计方法。

三、二阶兼容性分析

为了分析用于采样的度量的有效性，我们定义了一个模糊性事件的概率：

$$\mathrm{P}_{am}(M) = \mathrm{P}\left(M_{in,out} > M_{in,in}\right)$$

其中M是一个具体的度量。$M_{in,out}$是内点和外点之间的相似度，而$M_{in,in}$是两个内点之间的相似度。P()表示一个事件发生的概率。当$M_{in,out} > M_{in,in}$ 时，外点就会变

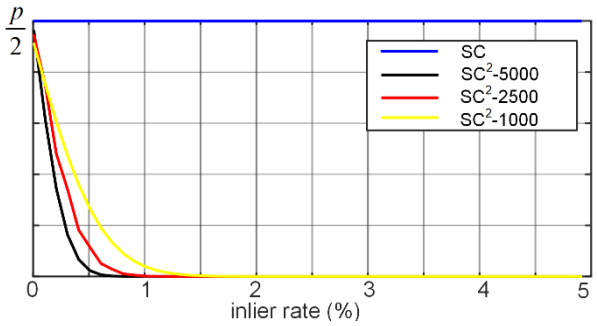


图 2　一阶和二阶度量模糊性概率对比（SC^2-N,N=5000,2500,1000 为匹配对个数）

成内点在该度量下的近邻，从而基于度量的采样就越不稳定。因此，概率值越小，基于该度量的采样越稳定。

传统的一阶兼容性度量 SC 通常被定义为如下形式：

$$SC_{ij} = \phi\left(d_{ij}\right), d_{ij} = \left| d\left(x_i, x_j\right) - d\left(y_i, y_j\right)\right.$$

其中ø为一个单调递减函数，d_{ij}为两个匹配对之间的距离差。本文提出的二阶兼容性 (SC^2) 具有如下形式：

$$C_{ij} = \begin{cases} 1, d_{ij} \leq d_{thr} \\ 0, d_{ij} > d_{thr} \end{cases},$$

$$SC_{ij}^2 = C_{ij} \cdot \sum_{k=1}^{N} C_{ik} \cdot C_{kj}.$$

其中C_{ij}是将d_{ij}经过阈值化过后的{0,1}值。二阶空间兼容性统计了任意两个兼容的匹配的共同兼容匹配对个数。

为了对比传统的一阶兼容性和提出的二阶兼容性用于模型生成时的鲁棒性，我们分别推导了它们的模糊性概率公式，并做出了他们的分布图，如图 2 所示。从图中可以看出，所提出的方法可以大大降低模糊性概率值，从而提升采样的稳定性。

四、特征和空间一致性引导的截断倒角距离

在 RANSAC 中，在生成大量模型后，通常通过内点计数（IC）的方式选择最好的模型。IC 度量的计算方式为：通过估计的模型对齐初始匹配对，并统计成功对齐的匹配对个数。然而，这种方法受限于初始匹配对的准确率。也就是说，即使是正确的模型，它的 IC 值最大为初始匹配对中正确匹配对的个数。图 3（a）展示了两对待匹配点云以及正确匹配对的个数，图（b）是一个错误估计的模型，而图（c）是一个正确估计的模型。然而，由于初始匹配对中正确匹配的数量过少，正确模型的 IC

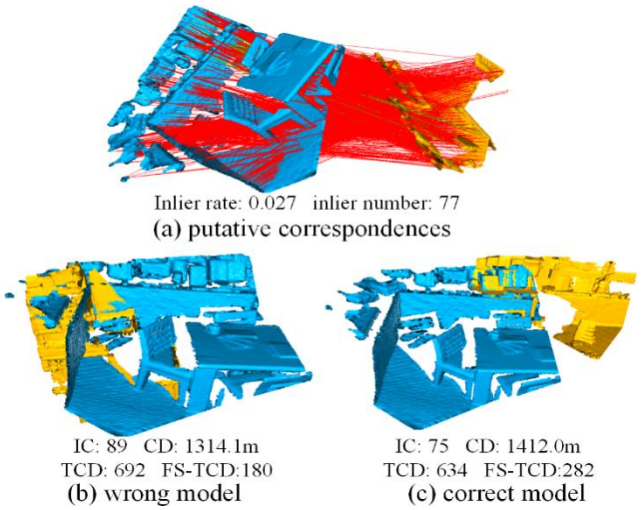


图 3　不同的模型选择度量对比（IC、TCD 和 FS-TCD 越高越好，CD 越低越好）

值只有 75，而错误模型的 IC 值反而偶然地大于正确模型的 IC 值，使得正确模型无法被选择出来。

因此，我们考虑利用倒角距离（CD）这种全局度量作为模型选择依据。然而，直接应用 CD 是不行的，因为它需要在全局范围内搜索近邻，使得它用于大量的模型选择时比较低效。同时，它没有考虑点云的部分重叠问题。为了解决这一问题，我们首先将 CD 改写成截断的形式（TCD），使得它不考虑非重叠区域的对齐质量。然后，我们用特征和空间一致性来引导倒角距离的搜索空间，这种方式既能缩短 CD 的搜索时间从而提升效率，又能通过特征和空间一致性的引导来减少 CD 的偶然误差。从图 3 中可以看出，所提出的 FS-TCD 对正确模型和错误模型有更好的区分度。

五、总体方法

基于所提出的二阶兼容性（SC^2）和特征和空间一致性引导的截断倒角距离（FS-TCD）度量，我们设计了一个高效的配准算法，名为 SC^2-PCR++。方法流程图如图 4 所示。具体来说，方法的输入是待匹配的点云对以及为它们分别提取的描述子。在模型生成步骤中，我们首先通过一对一匹配建立初始匹配集合，然后为它们构建逐匹配的 SC^2 矩阵。然后，我们通过谱分解的方式结合非极大值抑制选择一些可靠的匹配作为种子点，目的是减少采样次数来提升算法效率。在这之后，我们通过一个两阶段的采样方式为每个种子点构造一个一致性集合。在第一阶段，我们选择与每个种子点相似度得分最高的 K1 个匹配构建一个局部相似度矩阵。在第二个阶段中在局部集合中做进一步筛选来剔除潜在的误匹配并保留 K2 个匹配对。最后，我们通过局部谱匹配的方式结合可微的奇异值分解来为每一个种子点的一致性集合求一个刚体变换。

在模型选择阶段，我们首先构建一对多的匹配关系矩阵用作后续计算 FS-TCD 的引导。尽管 FS-TCD 相比于 CD 十分高效，但是如果为每个种子点生成的刚体变换都计算一个 FS-TCD 度量仍然十分耗时。因此，我们采用 IC 度量首先对生成的刚体变化做一个过滤。我们首先为每个变换计算 IC 值，并选择出 IC 值最高的N_S个作为候选模型，最后通过计算 FS-TCD 并选择得分最高的最终的结果。

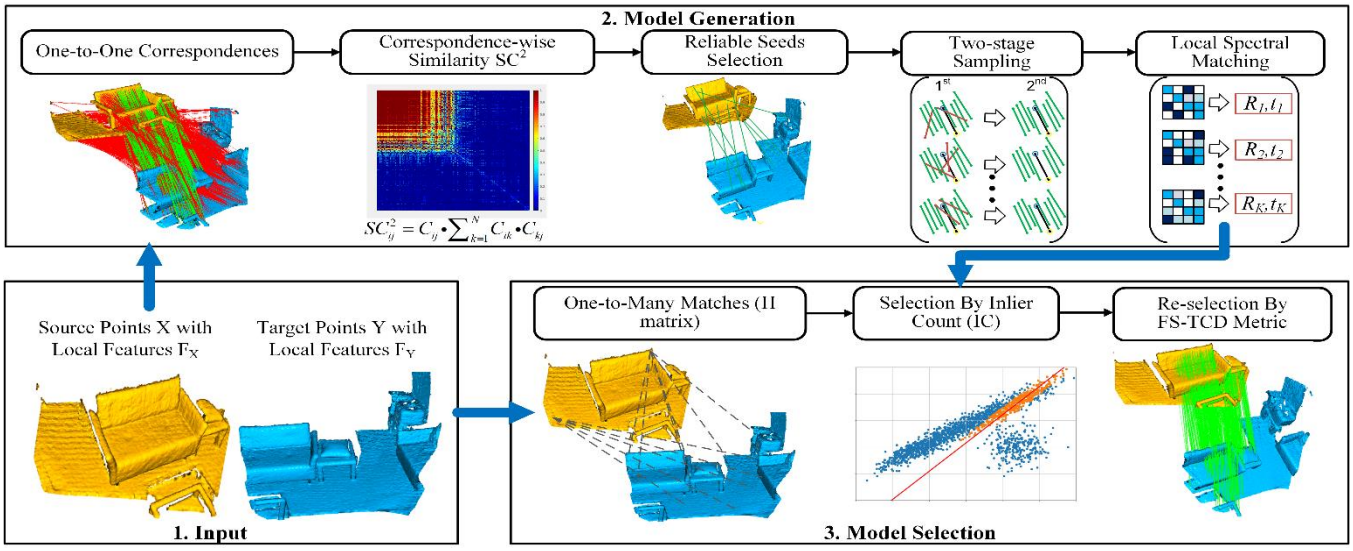


图 4　方法总体流程

表 1　在室内配准数据集 3DMatch 的结果

	FPFH [7] (traditional descriptor)						FCGF [8] (learning-based descriptor)						
	RR(%) ↑	RE(deg) ↓	TE(cm) ↓	IP(%)↑	IR(%)↑	F1(%) ↑	RR(%) ↑	RE(deg) ↓	TE(cm) ↓	IP(%)↑	IR(%)↑	F1(%) ↑	Time (s)
OM-Net* [12]	-	-	-	-	-	-	35.90	4.16	10.50	-	-	-	**0.08**
RegTR* [13]	-	-	-	-	-	-	92.00	**1.57**	**4.90**	-	-	-	0.18
DGR [14]	32.84	2.45	7.53	29.51	16.78	21.35	88.85	2.28	7.02	68.51	79.92	73.15	1.53
DHVR [5]	67.10	2.78	7.84	60.19	64.90	62.11	91.93	2.25	7.08	80.20	78.15	78.98	3.92
PointDSC [4]	77.57	**2.03**	**6.38**	68.45	71.56	69.75	92.85	2.08	6.51	78.91	86.23	82.12	0.10
FGR [2]	40.91	4.96	10.25	6.84	38.90	11.23	78.93	2.90	8.41	25.63	53.90	33.58	0.89
TEASER [15]	75.48	2.48	7.31	73.01	62.63	66.93	85.77	2.73	8.66	**82.43**	68.08	73.96	0.07
GC-RANSAC [16]	67.65	2.33	6.87	48.55	69.38	56.78	92.05	2.33	7.11	64.46	**93.39**	75.69	0.55
RANSAC-4M [1]	66.10	3.95	11.03	64.27	59.10	61.02	91.44	2.69	8.38	78.88	83.88	81.04	2.86
CG-SAC [3]	78.00	2.40	6.89	68.07	67.32	67.52	87.52	2.42	7.66	75.32	84.61	79.90	0.27
SC^2-PCR [9]	83.98	2.18	6.56	72.48	78.33	75.10	93.28	2.08	6.55	78.94	86.39	82.20	0.11
SC^2-PCR++	**87.18**	2.10	6.64	**76.49**	**81.72**	**78.82**	**94.15**	2.04	6.50	80.57	87.69	**83.71**	0.28

六、实验结果

表 1 展示了所提出的方法与其他传统方法和深度学习方法在室内数据集 3DMatch[6]上的对比结果。我们报告了配准召回率（RR）、平均旋转误差（RE）、平均平移误差（TE）、匹配准确率（IP）、匹配召回率（IR）和匹配 F1 分数等评价指标。其中配准召回率为最重要的指标，因为它直观地反应了姿态估计正确的点云对的比例。为了更加全面地对比误匹配消除性能，我们分别将所有的误匹配消除方法与传统描述子 FPFH[7]和深度学习描述子 FCGF[8]结合。除了对比了一些误匹配消除方法，我们也对比了一些端到端的点云配准方法（带*号的方法）。由于这些方法不需要匹配，因此对于这些方法我们没有报告与匹配准确性相关的指标（IP，IR，F1）。为了更清晰地展示所提出方法的性能，我们还对比了我们方法之前的版本（SC^2-PCR[9]），该方法对应的工作被 CVPR2022 录用。从表中的结果可以看出，SC^2-PCR 和 SC^2-PCR++取得了当前最好的效果，无论是与传统的点云描述子 FPFH 结合还是与深度学习描述子 FCGF 结合。SC^2-PCR++相比于 SC^2-PCR 也取得了明显的性能提升。效率方面，SC^2-PCR 取得了和深度学习方法相当的效率。SC^2-PCR++效率低于 SC^2-PCR，这是由于 SC^2-PCR++引入了更加精细的模型选择策略，增加了一定的时间消耗。考虑到性能的显著提升，增加的时间消耗也是可以接受的。

表 2 展示了所提出的方法与其他传统方法和深度学习方法在 3DLoMatch[10]的对比结果。3DLoMatch 由重叠率较低的待匹配点云对组成，比较具有挑战性。我们分别选取了当前三种比较有代表性的基于深度学习的描述子与误匹配消除方法结合，包括 FCGF[8]，

表 2　在低重叠率配准数据集 3DLoMatch 上的结果

	FCGF [8]						
	RR↑	RE↓	TE↓	IP↑	IR↑	F1↑	Time(s)
DHVR [5]	54.41	4.14	12.56	41.96	38.60	39.22	3.55
DGR [14]	43.80	4.17	10.82	42.22	38.96	39.05	1.48
PointDSC [4]	56.09	3.87	**10.39**	44.51	52.38	47.57	**0.10**
FGR [2]	19.99	5.28	12.98	27.63	19.16	19.98	1.32
RANSAC [1]	46.38	5.00	13.11	40.70	44.61	42.02	2.86
CG-SAC [3]	52.31	3.84	10.55	42.16	47.02	44.61	0.25
SC^2-PCR [9]	57.83	3.77	10.46	44.87	53.69	48.38	0.11
SC^2-PCR++	**61.15**	**3.72**	10.56	**47.12**	**56.52**	**50.85**	0.26
	Predator [10]						
	RR↑	RE↓	TE↓	IP↑	IR↑	F1↑	Time(s)
DHVR [5]	65.41	4.97	12.33	54.75	54.66	53.70	3.55
DGR [14]	59.46	**3.19**	10.01	51.38	54.24	51.62	1.48
PointDSC [4]	68.89	3.43	9.60	56.55	67.52	60.82	**0.10**
FGR [2]	35.99	4.77	11.64	47.18	38.76	39.10	1.32
RANSAC [1]	64.85	4.28	11.04	56.44	65.68	60.01	2.86
CG-SAC [3]	64.01	3.86	10.94	56.88	64.12	59.25	0.25
SC^2-PCR [9]	69.46	3.46	**9.58**	56.98	67.47	61.08	0.11
SC^2-PCR++	**71.59**	3.45	9.61	**59.61**	**70.17**	**63.73**	0.26
	GeoTransformer [11]						
	RR↑	RE↓	TE↓	IP↑	IR↑	F1↑	Time(s)
DHVR [5]	73.83	4.49	10.21	61.06	71.85	64.21	2.71
PointDSC [2]	77.82	3.00	8.71	63.65	76.87	68.39	0.09
RANSAC [1]	77.48	3.37	9.69	**64.91**	73.98	68.68	2.03
CG-SAC [3]	76.92	3.34	9.81	62.10	75.27	67.05	0.22
LGR [11]	77.20	2.99	8.58	64.47	76.04	68.86	**0.05**
SC^2-PCR [9]	78.33	3.04	8.81	64.63	76.67	69.19	0.08
SC^2-PCR++	**78.72**	**2.96**	**8.56**	64.80	**77.02**	**69.55**	0.24

Predator[10]和 GeoTransformer[11]。从表 2 中展示的结果可以看出，无论与哪种描述子结合，SC^2-PCR++都取得了当前最好的性能。尤其是在特征描述子的表达能力相对较弱时（如 FCGF），SC^2-PCR++取得的性能提升尤为明显。这是由于 SC^2-PCR++对低内点率的情况有较好的鲁棒性。

为了更全面地评价方法的性能，我们在表 3 中展示了我们的方法在室外点云配准数据集 KITTI 上和其他方法的对比结果。

为了展示配准在下游任务中的应用，我们将配准方法用于室内地图重建并在 ICL_NUIM 数据集上对比了一些经典的 SLAM 方法。从表 4 的结果中可以看出，SC^2-PCR++在其中两个场景上取得了最低的轨迹误差，并且平均轨迹误差为所有方法中最小的。

表 3　在室外数据集 KITTI 上的结果

	FPFH [7]						
	RR↑	RE↓	TE↓	IP↑	IR↑	F1↑	Time(s)
DHVR [5]	-	-	-	-	-	-	-
DGR [14]	77.12	1.64	33.10	78.39	54.12	62.15	2.29
PointDSC [4]	98.20	0.35	8.13	92.85	93.87	93.11	0.45
FGR [2]	5.23	0.86	43.84	4.93	0.05	0.10	3.88
RANSAC [1]	74.41	1.55	30.20	78.50	52.66	60.72	5.43
CG-SAC [3]	74.23	0.73	14.02	78.64	60.82	67.11	0.73
SC^2-PCR [9]	**99.64**	**0.32**	7.23	93.63	95.89	94.63	**0.31**
SC^2-PCR++	**99.64**	**0.32**	**7.19**	**94.07**	**96.19**	**95.00**	0.86
	FCGF [8]						
	RR↑	RE↓	TE↓	IP↑	IR↑	F1↑	Time(s)
DHVR [5]	**99.10**	**0.29**	**19.80**	-	-	-	0.83
DGR [14]	98.20	0.34	21.70	72.19	78.06	75.13	2.29
PointDSC [4]	98.02	0.33	21.03	82.00	90.84	85.83	0.45
FGR [2]	89.54	0.46	25.72	**95.13**	4.25	8.18	3.88
RANSAC [1]	98.02	0.39	23.17	81.89	90.36	85.52	5.43
CG-SAC [3]	97.84	0.37	22.91	81.85	90.84	85.74	0.73
SC^2-PCR [9]	98.20	0.33	20.95	82.01	91.03	85.90	**0.31**
SC^2-PCR++	98.56	0.32	20.61	82.17	**91.23**	**86.09**	0.86

表 4　多路配准数据集 ICL_NUIM 上的轨迹误差

	Living1	Living2	Office1	Office2	AVG
ElasticFusion	66.61	24.33	13.04	35.02	34.75
InfiniTAM	46.07	73.64	113.8	105.2	85.68
BAD-SLAM	fail	40.41	18.53	26.34	-
Multiway + DGR	21.06	21.88	15.76	11.56	17.57
Multiway + PointDSC	20.25	15.58	13.56	11.30	15.18
Multiway + DHVR	22.91	16.37	**12.58**	10.90	15.69
Multiway+ FGR	78.97	24.91	14.96	21.05	34.98
Multiway + RANSAC	110.9	19.33	14.42	17.31	40.49
Multiway + SC^2-PCR	18.68	**14.31**	14.63	11.95	14.90
Multiway + SC^2-PCR++	**17.56**	14.37	13.24	**9.49**	**13.67**

责任编委　崔海楠

参考文献

[1] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM.

[2] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In European Conference on Computer Vision, pages 766–782. Springer, 2016.

[3] Siwen Quan and Jiaqi Yang. Compatibility-guided sampling consensus for 3-d point cloud registration. IEEE Transactions on Geoscience and Remote Sensing, 58(10):7380–7392, 2020.

[4] Bai X, Luo Z, Zhou L, et al. Pointdsc: Robust point cloud registration using deep spatial consistency[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15859-15869.

[5] Junha Lee, Seungwook Kim, Minsu Cho, and Jaesik Park. Deep hough voting for robust global registration. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15994–16003, 2021.

[6] Zeng A, Song S, et al. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1802-1811.

[7] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In 2009 IEEE international conference on robotics and automation, pages 3212–3217. IEEE, 2009.

[8] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In Proceedings of the IEEE International Conference on Computer Vision, pages 8958–8966, 2019.

[9] Chen Z, Sun K, Yang F, et al. Sc2-pcr: A second order spatial compatibility for efficient and robust point cloud registration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 13221-13231.

[10] Huang S, Gojcic Z, Usvyatsov M, et al. Predator: Registration of 3d point clouds with low overlap[C]//Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2021: 4267-4276.

[11] Qin Z, Yu H, Wang C, et al. Geometric transformer for fast and robust point cloud registration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11143-11152.

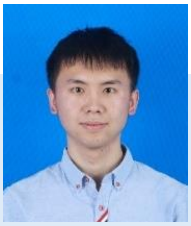
[12] Xu H, Liu S, Wang G, et al. Omnet: Learning overlapping mask for partial-to-partial point cloud registration[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 3132-3141.

[13] Yew Z J, Lee G H. Regtr: End-to-end point cloud correspondences with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 6677-6686.

[14] Choy C, Dong W, Koltun V. Deep global registration[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 2514-2523.

[15] Yang H, Shi J, Carlone L. Teaser: Fast and certifiable point cloud registration[J]. IEEE Transactions on Robotics, 2020, 37(2): 314-333.

[16] Barath D, Matas J. Graph-cut RANSAC[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6733-6741.



陈志

华中科技大学在读博士生。主要研究方向为图像匹配、三维点云配准等。
Email: z_chen@hust.edu.cn

孙琨

中国地质大学（武汉）副教授。主要研究方向为多视图图像匹配、三维重建和点云处理。
Email: sunkun@cug.edu.cn

杨帆

华中科技大学在读博士生。主要研究方向为三维点云配准和深度学习几何学。
Email: fanyang@hust.edu.cn

郭琳

华中科技大学在读硕士生。主要研究方向为点云配准和图像匹配等。
Email: linguo@hust.edu.cn

陶文兵

华中科技大学教授。主要研究方向为三维配准、多视图立体几何、表面重建、图像分割等。
Email: wenbingtao@hust.edu.cn

热点追踪

基于布朗桥扩散模型的图像翻译

南昌航空大学　李波 薛凯韬 刘彬
Cardiff University　Yu-Kun Lai

一、引言

图像到图像的翻译问题是计算机视觉（Computer Vision，CV）领域的一个重要问题，它是指在两个不同的图像域之间建立映射关系。许多计算机视觉领域的问题都可以看作是图像翻译问题，比如图像去雾去噪，灰度图像着色，图像补全等。

现有的图像翻译方法通常基于生成对抗网络（Generative Adversarial Network，GAN）。但是基于 GAN 的方法存在训练不稳定和模式坍塌的问题。其他的方法诸如基于变分自编码器（Variational Autoencoder，VAE）的方法，基于自回归模型（Autoregressive Models，AM）的方法并不能达到和基于 GAN 的方法相同的质量和泛化能力。近年来扩散概率模型（Diffusion Probabilistic Models，DPM）在图像生成上已经可以达到与 GAN 方法相当的质量。并且有一些条件扩散模型被用于图像翻译任务。但是这些条件扩散模型将条件信息直接加入到 U-Net 中，这种加入条件信息的机制缺乏清晰的理论基础保证。

我们提出了一种基于布朗桥扩散模型（Brownian Bridge Diffusion Models，BBDM）的图像翻译方法。如图 1 所示，传统的条件扩散模型从高斯噪声出发，通过直接将条件信息与每一步的噪声图像拼接之后输入神经网络来达到条件生成的目的。布朗桥扩散模型采用了截然不同的条件机制。本文首次提出利用布朗桥随机过程直接建模两个不同图像域之间的映射，可以从理论上很好的保证扩散过程可以稳定的从条件域出发，最终收敛到目标图像域的分布。在不同的图像翻译，诸如语义图像生成真实图像，边界图像生成真实图像，风格迁移，图像补全，灰度图像着色任务上的实验证明我们的布朗桥扩散模型可以生成高质量和多样的结果。

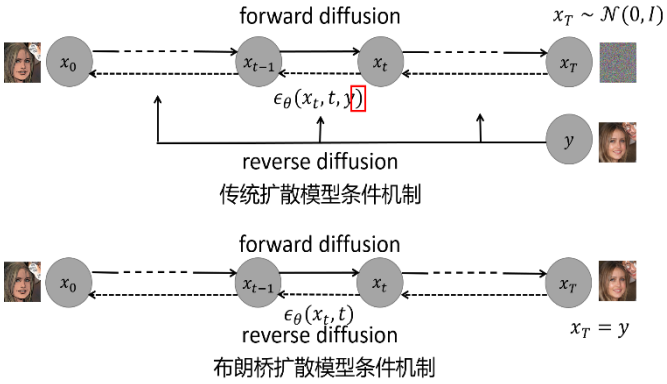


图 1　传统扩散模型和布朗桥扩散模型

二、基于布朗桥扩散模型的图像翻译总览

布朗桥（Brownian Bridge）是一种连续时间上的随机过程，它的分布为在条件$B_0 = a, B_1 = b$下的维纳过程。在基于布朗桥的图像翻译任务中假设条件图像的域表示为$\boldsymbol{A}$，目标图像域表示为$\boldsymbol{B}$，布朗桥随机过程在初始时刻的分布视为条件图像域$\boldsymbol{A}$，在最终时刻的分布视为目标图像域$\boldsymbol{B}$，这样就可以利用布朗桥随机过程实现图像域之间的转换。

基于布朗桥的图像翻译的整体过程如图 2 所示。

为了降低计算复杂度，提升模型的泛化性能，我们采用了隐式扩散模型的思想。对于一张条件图像$\boldsymbol{I_A} \in \boldsymbol{A}$，用预训练好的 VQGAN 的编码器提取出条件图像的隐式特征$\boldsymbol{L_A}$，以压缩图像的冗余信息，然后在隐空间将条件图像的隐式特征通过布朗桥扩散模型转换到目标图像的隐式特征$\boldsymbol{L_B}$，最后利用预训练好 VQGAN 的解码器从目标图像的隐式特征中解码出目标域的图像 $\mathbf{I_B} \in \mathbf{B}$。

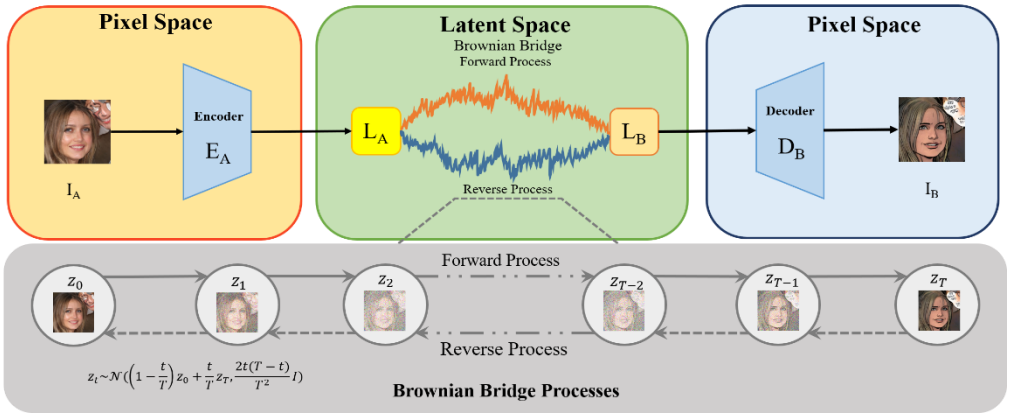


图 2　基于布朗桥扩散模型的图像翻译框架图

三、布朗桥扩散模型原理

我们利用布朗桥扩散模型实现域之间的转换。为了更一般化的表示布朗桥扩散模型原理，首先定义纯净的数据分布为$x_0 \in q(x_0)$，条件信息的分布为$y \in q(y)$，布朗桥随机过程则需要在给定y的情况下得到其对应的x_0。为了达到这个目的，与其他扩散模型类似，布朗桥扩散模型也分为前向过程（Froward Process）和反向过程（Reverse Process）。

前向过程中每一个时刻的边缘分布可以表示为：

$$q_{BB}(x_t|x_0,y) = \mathcal{N}(x_t;(1-m_t)x_0+m_t y,\delta_t I)$$

其中$m_t=\frac{t}{T}$，$\delta_t=2(m_t-m_t^2)$.

为了得到前向中间过程的每一步的分布，我们将这个过程建模为马尔科夫链（Markov Chain）：

$$q_{BB}(x_{1:T}|x_0,y) = \prod_{t=1}^{T} q_{BB}(x_t|x_{t-1},y)$$

利用马尔科夫链的性质可以得到前向中间过程的分布的表示：

$$q_{BB}(x_t|x_{t-1},y) = \mathcal{N}(x_t;\frac{1-m_t}{1-m_{t-1}}x_{t-1} + \frac{m_t-m_{t-1}}{1-m_{t-1}}y,\delta_{t|t-1}I)$$

其中$\delta_{t|t-1}=\delta_t-\delta_{t-1}\left(\frac{1-m_t}{1-m_{t-1}}\right)^2$.

之后利用马尔科夫链的无后效性，可以得到后验分布：

$$q_{BB}(x_{t-1}|x_t,x_0,y) = \mathcal{N}(x_{t-1};\frac{\delta_{t-1}(1-m_t)}{\delta_t(1-m_{t-1})}x_t + (1-m_{t-1})\frac{\delta_{t|t-1}}{\delta_t}x_0 + \left(m_{t-1}-m_t\frac{\delta_{t-1}(1-m_t)}{\delta_t(1-m_{t-1})}\right)y,\tilde{\delta}_t=\frac{\delta_{t|t-1}\delta_{t-1}}{\delta_t})$$

反向的过程也定义为马尔科夫链，但是反向中间过程的均值需要通过神经网络预测：

$$p_\theta(x_{t-1}|x_t,y) = \mathcal{N}(x_{t-1}|\mu_\theta(x_t,t),\tilde{\delta}_t I)$$

训练的过程通过最小化变分下界来对齐前向中间过程的后验分布和反向的中间过程，也即约束均值相同，经过一定的重参数化技巧可以得到最终的训练目标：

$$\mathbb{E}_{x_0,y,\epsilon}\left[\left\|m_t(y-x_0)+\sqrt{\delta_t}\epsilon-\epsilon_\theta(x_t,t)\right\|_2^2\right]$$

但是把布朗桥随机过程建模为马尔科夫链的一个弊端是反向采样的过程和前向过程的步数要完全一样，时间代价线性正比于步数T。为了提高采样效率，我们借鉴了去噪扩散隐式模型的方法，将前向过程建模为非马尔科夫链的形式，这样在保证训练目标完全不变的情况下，只需要对采样过程进行一定的改动就能实现快速采样，具体的对于采样序列$\{1,2,\dots,T\}$的任意一个子序列$\{\tau_1,\tau_2,\dots,\tau_S\}$，将后验分布定义为如下形式就可以实现跳步的采样从而大大节省时间代价：

$$q_{BB}(x_{\tau_{s-1}}|x_{\tau_s},x_0,y) = \mathcal{N}(x_{\tau_{s-1}};(1-m_{\tau_{s-1}})x_0+m_{\tau_{s-1}}y + \sqrt{\delta_{\tau_{s-1}}-\sigma_{\tau_s}^2}\frac{1}{\sqrt{\delta_{\tau_s}}}(x_{\tau_s}-(1-m_{\tau_s})x_0-m_{\tau_s}y),\sigma_{\tau_s}^2 I)$$

当$\sigma_t^2=\tilde{\delta}_t$时，这个过程与建模为马尔科夫链过程是等价的。

四、实验内容

为了验证布朗桥扩散模型在图像翻译应用上的有效性，我们在语义图像生成真实图像，边界图像生成真实图像，风格迁移，图像补全，灰度图像着色任务上进

行了实验。实验结果证明，我们提出的基于布朗桥扩散模型的图像翻译方法不仅可以生成高质量和多样的结果，而且在不同的任务上有着较好的泛化性能。部分实验结果如图 3 所示。

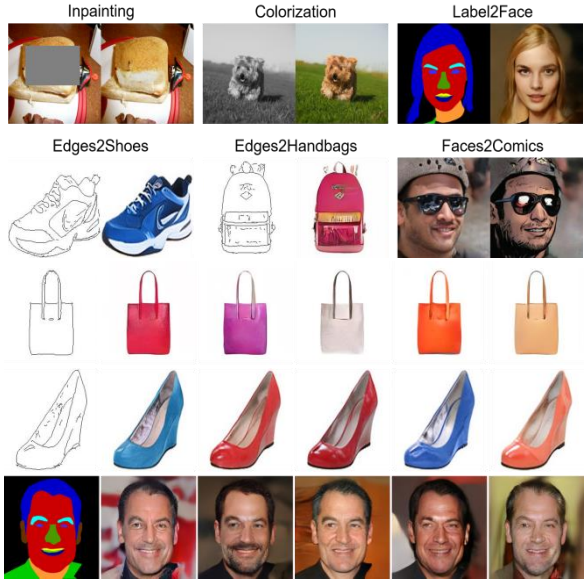


图 3　基于布朗桥扩散模型的图像翻译实验结果

五、总结和展望

本文介绍了基于布朗桥扩散模型的图像翻译方法的基本原理以及相关的实验内容。相较于传统的条件扩散模型，布朗桥扩散模型拥有更清晰的理论基础。实验的结果也证明，基于布朗桥扩散模型的图像翻译可以生成高质量和多样的结果。但是目前的布朗桥扩散模型应用于图像翻译需要成对的训练数据，而且需要较大的数据集才能有比较好的效果。因此，将布朗桥扩散模型应用于不成对的图像翻译并降低其对数据集大小的依赖是未来工作的重要方向。

该成果被国际学术会议 The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR2023) 接收。

责任编委　储珺

李波

南昌航空大学，数学与信息科学学院教授，主要研究方向为图深度学习、计算机视觉。
Email: libo@nchu.edu.cn

薛凯韬

南昌航空大学，在读硕士，CCF 学生会员。主要研究方向为深度学习、计算机视觉等。
Email: xuekt98@gmail.com

刘彬

南昌航空大学，数学与信息科学学院讲师，主要研究方向为数字几何处理、计算机视觉。
Email: nyliubin@nchu.edu.cn

Yukun Lai

Cardiff University, Professor of School of Computer Science and Informatics, research interests include computer graphics, geometric modelling and processing, computer vision, image processing.
Email: LaiY4@cardiff.ac.uk

顶会观察

ICLR 2023

上海人工智能实验室青年科学家　李弘扬

国际学习表征会议（International Conference on Learning Representations，ICLR）是深度学习的顶级会议之一，受到学术研究者们广泛认可，在谷歌发布的人工智能领域“top publication”排名中位列前十。该会议由深度学习三大巨头之二的 Yoshua Bengio 和 Yann LeCun 牵头创办。Yoshua Bengio 是蒙特利尔大学教授，深度学习三巨头之一，他领导蒙特利尔大学的人工智能实验室 MILA 进行 AI 技术的学术研究。Yann LeCun 同为深度学习三巨头之一的他现任 Facebook 人工智能研究院 FAIR 院长、纽约大学教授。ICLR 的创办旨在提供一个场所，能够让学者们交流分享表征学习领域所关心的话题，为深度学习提供一个专业化的交流平台。今年的 ICLR 于 2023 年 5 月 1-5 日在卢旺达举办第十一届会议，这是机器学习领域顶会首次在非洲国家举办，这将会极大促进非洲国家科技发展及民众对前沿科技的认知，进而带来广泛而深远的国际影响力。

一、 会议亮点

开放的评审机制：ICLR 采用了开放评审的评审制度，根据规定，所有提交的论文都会公开姓名等信息，并且接受所有同行的评价及提问，任何学者都可或匿名或实名地评价论文。而在公开评审结束后，论文作者也能够对论文进行调整和修改。ICLR 是在开放评审方面做得最公开、影响范围最大的一个会议。目前 ICLR 的历届所有论文及评审讨论的内容，都完整地保存在 OpenReview 上，它也是 ICLR 的官方投稿入口。OpenReview 是马萨诸塞大学阿默斯特学院 Andrew McCAllum 为 ICLR 2013 牵头创办的一个公开评审系统，秉承公开同行评审、公开发表、公开来源、公开讨论、公开引导、公开推荐、公开 API 及开源等八大原则，得到了 Facebook、Google、NSF 和马萨诸塞大学阿默斯特中心等机构的支持。

开源代码：随着 ICLR 影响力的进一步提升，越来越多的研究工作发布了代码或数据，开源已经成为趋势。据 GitHub 项目的不完全统计，今年已经超过 400 篇论文公开了源代码。代码开源的优势明显，免费透明，不仅可以增加研究者之间的协作机会，而且还能提升研究工作的影响力。当然，也存在一定的未知风险，这需要大家共同努力完善开源代码。

二、 录用情况

ICLR2023 收到的有效投稿和录用数量都有显著提高，大会共收到了 4922 篇有效投稿，相比上一年增加了 32.2%。最终接收了近 1573 篇论文，接收率约为 31.8%，录用率与去年基本持平。此外，今年还有一个变化是接收论文的标签会有两个，一个是论文类型（oral、spotlight、poster），另一个是 presentation 的方式。ICLR2023 会议涵盖的方向主要有强化学习、深度学习、表征学习、图神经网络等方向，并且今年 Transformer 论文接收率下降，语言模型论文增多。位于 top5% 共有 90 篇论文，内容涉及 Transformer、in-context learning、扩散模型等内容。今年 ICLR2023，共有 4 篇论文获得杰出论文奖，5 篇论文获得杰出论文奖提名。其中，来自北京大学的张博航、罗胜杰、王立威、贺笛共同获得一篇杰出论文奖，清华大学孔祥哲、中国人民

大学高瓴人工智能学院黄文炳、清华大学刘洋共同获得一篇杰出论文奖提名。据统计，清华大学朱军教授和谷歌大脑的 Dale Schuurmans 两位学者发文数量并列。

三、 主题报告

本次 ICLR2023 会议邀请了六位 Keynote 演讲者，主题报告内容涵盖了一系列不同的研究主题，从生成模型，医疗保健和健康公平中的机器学习，统计机器学习/理论，到对话式人工智能和人工智能的创造力。每位演讲者分别进行了 60 分钟的演讲，主题报告激发了研究者们深入思考深度学习的技术基础及其对社会日益增长的影响。

Entanglements, Exploring Artificial Biodiversity. 随着生成系统的快速发展，生成系统已经在艺术创作过程中得到广泛的应用。Sofia Crespo 分享了她的艺术实践和使用生成系统的旅程，特别是神经网络，作为探索推测性生命体的一种手段，以及技术如何使我们更接近自然世界。

Understanding Systematic Deviations in Data for Trustworthy AI. 微软 AI for Good Research Lab 实验室的首席研究科学家 Girmaw Abebe Tadesse 讨论了在可信人工智能研究中的数据偏差问题。随着采用机器学习（ML）模型来协助决策的趋势越来越明显，为了实现值得信赖的 ML 应用，检查模型和其相应的数据的潜在系统偏差是至关重要的。这种检查过的数据可能被用于训练、测试或由模型本身产生。对系统性偏差的理解在资源有限和/或错误敏感的领域尤其关键，例如医疗保健。在这次演讲中，报告者将反思他们最近的工作，这些工作利用系统偏差的自动识别和表征来完成医疗领域的各种任务，包括：数据质量理解；时间漂移；异质干预效果分析；以及新类检测。此外，人工智能驱动的科学发现正越来越多地使用生成模型来促进。报告者分享他们以数据为中心的多层次评估框架如何帮助量化生成式模型的能力，并以材料科学作为一个使用案例，以领域诊断和可解释的方式。除了经常用来训练 ML 模型的策划数据集的分析，类似的以数据为中心的分析也应该被考虑在传统的数据源上，比如教科书。最后，报告者害介绍最近在皮肤病学学术材料中进行的自动表示分析的合作工作。

Importance-Weighting Approach to Distribution Shift Adaptation. RIKEN 中心主任、东京大学教授 Masashi Sugiyama 介绍了分布转移自适应问题中的重要性加权方法。对于可靠的机器学习，克服分布偏移是最重要的挑战之一。在这次演讲中，报告者首先概述经典的重要性加权方法来适应分布偏移，它包括一个重要性估计步骤和一个重要性加权的训练步骤。然后，报告者介绍一种较新的方法，即同时估计重要性权重和训练一个预测器。最后，报告者讨论一个更具挑战性的连续分布转移的场景，即数据分布随时间不断变化。

AI, History and Equity. 波士顿大学副教授 Elaine Nsoesie 介绍了人工智能在社会问题中的应用。大型数据集越来越多地被用来训练人工智能模型，以解决社会问题，包括健康方面的问题。有偏见的人工智能模型的社会影响已被广泛讨论。然而，在对话中有时缺少的是历史上的政策和不公正在塑造现有数据和结果方面的作用。通过历史视角评估数据和算法可能对社会变革至关重要。

Dialogue Research in the Era of LLMs. 来自 stealth AI startup 公司的 Dilek Hakkani-Tur 对大语言模型时代的对话研究进行了介绍。最近的大型语言模型（LLMs）由于能够对任何用户的要求产生连贯的自然语言反应，使得开放域对话系统取得了重大进展。它们记忆和进行组合推理的能力使对话相关的任务得以准确执行，如语言理解和响应生成。然而，这些模型存在局限性，例如，幻觉、不希望捕捉到的偏见、对特定政策的概括性困难，以及缺乏可解释性。为了解决这些问题，自然语言处理界提出了一些方法，例如，在训练或推理过程中向语言模型注入知识，使用多步骤推理和 API/工具检索相关知识，等等。在这次演讲中，报告者对相关工作进行概述。

Learned optimizers: why they're the future, why they’ re hard, and what they can do now. 来自 Google Brain 的研究员 Jascha Sohl-Dickstein 报告了在深度学习中广泛应用的优化器所存在的问题和改进方案。深度学习的成功有赖于所学函数在许多任

务中的表现大大超过手工设计的函数。然而，研究者们目前仍然使用手工设计的优化器来训练模型，这些优化器作用于手工设计的损失函数。报告者论证了这些手工设计的组件通常与期望的行为不匹配，研究者可以期望元学习的优化器表现得更好。报告者讨论使元学习优化器难以训练的挑战和原因。其中包括：混乱和高方差的元损失景观；元训练的极端计算成本；缺乏全面的元训练数据集；设计具有正确归纳偏差的学习型优化器的挑战；解释学习型优化器的行动方法的挑战。报告者分享了其中一些挑战的解决方案。报告者展示了实验结果，在许多情况下，学习型优化器的性能优于手工设计的优化器。此外，报告者讨论了元训练学习型优化器所带来的新功能。

四、 热点论文

在今年的 ICLR2023 的获奖论文中 ，共有 4 篇论文获得杰出论文奖，5 篇论文获得杰出论文奖提名。其中，来自北京大学的张博航、罗胜杰、王立威,、贺笛共同获得一篇杰出论文奖，清华大学孔祥哲、中国人民大学高瓴人工智能学院黄文炳、清华大学刘洋共同获得一篇杰出论文奖提名。

杰出论文奖

论文 1：Universal Few-shot Learning of Dense Prediction Tasks with Visual Token Matching，来自 KAIST 和微软亚洲研究院。该论文提出了一种用于密集预测任务的少样本学习 pipeline，密集预测任务包括语义分割、深度估计、边缘检测和关键点检测等。该研究提出了一个简单的统一模型，可以处理所有密集预测任务，并包含多项关键创新。该研究将激发密集预测的进一步发展，所提方法 —— 例如视觉 token 匹配、情景（episodic）元学习 —— 可以用于相关的多任务学习问题。

论文 2：Rethinking the Expressive Power of GNNs via Graph Biconnectivity，来自北京大学，该论文基于双连通性（biconnectivity）提出一种 GNN 表达性度量新指标。具体来说，该研究提出了一种利用节点间距离的新算法，并在合成数据和真实数据中进行了演示。该研究表明：双连通性问题在理论和实践中都有着广泛的潜在应用。

论文 3：DreamFusion: Text-to-3D using 2D Diffusion，来自谷歌研究院和加州大学伯克利分校，该论文提出了一种基于文本生成 3D 模型的有效方法，而无需 3D 模型作为训练数据。该论文的关键思想是利用本生成图像的扩散模型，并通过将误差信号反向传播到 3D 模型的神经辐射场来生成 3D 模型。该方法是 SOTA 图像生成和 3D 建模的巧妙组合，在实践中效果极好，并将启发各种后续工作，包括基于文本的 3D 视频生成。

论文 4：Emergence of Maps in the Memories of Blind Navigation Agents，来自佐治亚理工学院、Meta AI 等，该论文基于认知科学和机器学习的跨学科方法，让仅具备自我运动（egomotion）（不具备其他任何感知）的导航智能体学得有效表征，并实现有效导航。该研究对表征学习具有重要意义。

杰出论文奖提名

论文 1：Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning，来自 Meta 和 MBZUAI，该论文试图从一个新的理论视角来理解知识蒸馏。作者认为对于自然的多视图结构，没有蒸馏的情况下神经网络只能训练为仅依赖于部分特征，而蒸馏可以缓解这个问题。这篇论文提供了证明这一点的简化示例，有助于人们更好地理解知识蒸馏的有效性。

论文 2：Mastering the Game of No-Press Diplomacy via Human-Regularized Reinforcement Learning and Planning，来自 Meta AI 和 MIT，该论文的主题是多回合、多阶段、多人游戏的算法开发，提出使用一种类似于自我对弈（self-play）的策略来找到游戏均衡（equilibrium）状态，并在一个受人类玩家欢迎的复杂多人棋盘游戏上测试了该算法。其中，将寻求平衡的策略与行为克隆相结合。

论文 3：On the duality between contrastive and non-contrastive self-supervised learning，来自 Meta AI 等，在自监督学习领域，各种方法似乎没有任何共同点，但在实践中却表现相似。该论文对各种自监督学习方法进行了分析探究，发现了它们的共同点。

该论文展开研究了一些流行的自监督学习方法，证明其提出的理论能用于实际方法。这篇论文对自监督学习领域具有重要意义。

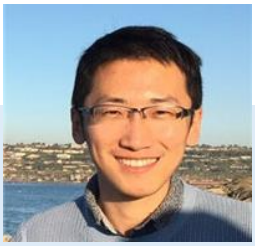
论文 4：Conditional Antibody Design as 3D Equivariant Graph Translation，来自清华大学计算机系，清华大学智能产业研究院，中国人民大学高瓴人工智能学院以及北京智源人工智能研究院，抗体设计是药物研发的一个重要问题，具有重要的应用前景。本文提出一种基于等变图神经网络的抗体设计方法 MEAN，在给定抗原、抗体重链和轻链的条件下，实现了抗体 CDRs 的设计和优化。与以往方法不同，MEAN 不但考虑了更全的「上下文信息」，而且能直接生成抗体 CDRs 的 1D 氨基酸序列及其 3D 构象，具有更高效率。在多个数据集的完整实验上，MEAN 显著优于已有方法。论文有望为后续湿实验研究提供一种高效的算法工具。

论文 5：Disentanglement with Biological Constraints: A Theory of Functional Cell Types，来自斯坦福大学、牛津大学和 UCL，该研究受生物学启发，揭示了机器学习和神经科学之间有趣的联系，并从数学上证明机器学习中的约束会导致线性网络解缠结（disentanglement）。该研究还通过实验表明，相同的约束对于非线性情况也是有效的。总的来说，这项研究从数学的角度对单个神经元和大脑结构给出了更深层的理解。

五、 总结展望

本年度 ICLR 大会中强化学习、深度学习、表征学习、图神经网络等领域依旧保持高热度。深度学习领域还面临许多挑战应该会包括：从物理世界到传感器与机器认知过程中，现有方法的能力上界是什么、人类已经发现的规律如何辅助机器认知物理世界、机器如何探知人类未曾发现的物理世界规律、深度学习模型的可解释性和可靠性等。回答好上述问题将会使得深度学习更好地迈向更高层级的智能。

责任编委 魏秀参



李弘扬

李弘扬，上海人工智能实验室青年科学家、高级工程师。2019 年香港中文大学博士毕业，在读期间荣获香港政府奖学金。以第一作者身份，在相关国际会议与期刊如 CVPR、NeurIPS、T-PAMI 等发表文章 10 余篇，累计谷歌学术引用 1600 余次，提交专利 20 余项。担任 CVPR、NeurIPS 2023 等会议领域主席。主持国家自然科学基金青年项目、上海市启明星项目。连续三年担任清华大学《计算机视觉》课程主讲人、上海交通大学行业博士生导师。2022 年获 Waymo 自动驾驶国际知名挑战赛第一名；团队提出的工作（BEVFormer、UniAD 等）在国际上取得领先地位，为多家自动驾驶公司提供了实际量产落地方案。提出的环视 3D 检测工作获 2022 年全球最有影响力的 AI 论文 Top 100。 UniAD 工作获 CVPR 2023 最佳论文提名奖。
Email: lihongyang@pjlab.org.cn

中科院自动化所何晖光研究员访谈

2023 年 6 月 5 日，《CCF-CV 专委简报》在线采访了中国科学院自动化研究所博士生导师何晖光研究员。下面是采访实录。

问题 1：何老师，您好！首先，请您分享一下您的个人学习和研究经历。

我的学习和工作经历较为简单，1990 毕业于黄冈中学，在大连海事大学本科、硕士毕业后，当了 2 年老师，99 年到中科院自动化所读博，师从戴汝为院士和田捷教授，02 年留所工作至今，期间在美国罗切斯特大学、加拿大滑铁卢大学、美国北卡教堂山分校做过博士后或访问教授。目前我是自动化所的研究员，也是国科大的岗位教授、上海科技大学的特聘教授。我的研究方向是人工智能、脑机接口和医学影像分析等。

问题 2：您在基于视觉信息编解码的深度学习类脑机制研究、基于多模态影像的医学图像数据处理等方面做出了突出贡献。能跟大家介绍一下您认为最值得骄傲的几项成果么？

我博士的工作是关于医学图像分割和三维重建的，博士毕业后延续了相关工作。我 07 年获选了北京市科技新星，认识了很多来自医院的新星，开展了广泛的医工合作，包括与天坛、协和、同仁、儿童医院等。我们将模式识别与图像处理的技术用在医学影像上，并结合临床信息加以分析，其目的是为了找到特定疾病相关的影像学参数，从而辅助医生早期发现、辅助诊断、愈后评估。刚开始由于缺少数据，因此研究是数据驱动的，和哪个医院合作，有什么样的疾病数据，就做相关的工作。比如中风、癫痫、弱视、青光眼、儿童抽动症等。尽管也发表了不少文章，后来意识到一个问题，就是每个疾病都做了一些工作，但是以我们的工科背景很难对疾病机制做进一步深入的分析，这样难以形成系统深入的工作。基于此考虑，我们的工作转向脑机接口，由于我们的工作涉及到了大脑的结构、功能，需要了解大脑不同的区域是执行什么样的功能，进一步我们想通过大脑的信号输出控制指令，就很自然地延伸到了脑机接口，其中脑机接口的编解码方法是我们主要研究的重点。

在非侵入式脑机接口方面，我们基于脑电、磁共振图像等多个模态数据研究大脑对视觉刺激的加工机制，对语义、运动意图、情绪进行解析，对图像进行重建，并且构建了在线脑-机接口系统，实现脑与机器人系统的交互。重点突破精细运动想象解码，多模态信息融合，以及针对复杂自然场景刺激的脑信息解码难题。

在视觉神经解码方面，针对大脑信号样本量少、信噪比低、图像重建模糊等难点问题，提出了贝叶斯深度多视图解码算法、多任务结构化神经解码算法、半监督对抗学习算法等一系列方法，弥补了脑活动和视觉刺激之间的信息鸿沟，克服了视觉信息重建模糊问题，为理解大脑解码过程提供了新的视角。更进一步，我们将大脑、视觉和语言知识相结合，通过多模态学习的方式，实现了从人类脑活动记录中零样本地解码视觉新类别。系列成果发表于 IEEE TPAMI,TNNLS 2018/2020,Inf.

Fusion 2021 等，获 ICME 2019 Best Paper Runner-Up Award，且被 MIT Technology Review 头条报导。

在情绪神经解码方面，研发了多源迁移学习情绪识别方法，可根据神经活动推断当前呈现刺激所含的情绪类别及个体当时的主观情绪状态，该方法有助于解决训练样本少、被试间差异大等问题。相关研究发表在 IEEE TMI 2023,IEEE TNNLS 2022,IEEE TCYB 2019 等期刊。

问题 3：能否对我们国家在上述研究方面的研究现状及国际地位进行一些分析和评价？您认为我们国家在这些方面的研究还有哪些需要突破之处呢？

脑-机接口技术已成为发展最为迅速的脑科学应用技术之一。目前大脑控制假肢和外骨骼已成功运用于残障者康复治疗，大脑意念控制飞行器等设备已成功实现原理演示，展示出巨大的应用前景。

我国脑机接口研究整体水平和欧美等发达国家还存在着一定的差距，在信号采集芯片技术方面，与国外差距较大。但是通过近几年的努力，已取得显著进展，差距正在明显缩短，在某些方面实现了与国际水平相接轨，个别方面已经走在国际前沿。

近年来，欧、美、日、韩等世界主要国家的顶尖企业和大学均加大了对脑机接口、脑机智能领域的抢先布局和资助力度，涌现出一批创新成果，覆盖了运动控制、视听觉解码、语言解码及特定脑疾病诊疗等方面。例如，Meta 公司发布了“意念打字”项目，通过非侵入式脑机接口创建一个脑机智能语言系统，实现直接根据脑信号来打字。然而，目前脑机接口领域还存在很多问题，如脑信号样本量小、多模态脑信号没有充分融合、被试个体差异性大等。

我国将“脑科学与类脑研究”列为 2030 创新重大项目，“脑机智能技术”作为“中国脑计划一体两翼”布局中的其中“一翼”。国内清华大学、浙江大学、天津大学、上海交通大学、中科院等单位均在脑机领域做出了有特色的工作。

问题 4：您先后主持包括 7 项国家自然科学基金（包括 1 项基金重点，1 项国际合作重点）、2 项 863 项目、国家重点研究计划课题等多个重要项目，请问您是如何对这些项目进行规划的？能否分享一下您在承担这些项目过程中所获得的经验、认识和建议？

由于中科院所是差额拨款，很多支出都要从课题经费中支出，申请项目就是科研人员升级打怪的必经之路。经验谈不上，有不少教训，其中有一个国际合作重点项目，我前后申请了 4 年，答辩 2 次才拿到，但正是由于坚持、不放弃，才有机会拿到。另外就是要注意平时的积累，提前做好准备，有些应急项目指南发布后所给的时间比较短，如果没有积累，很难在短时间完成申请书的准备，提前将自己最擅长的工作整理归纳好，后续就可方便地结合相应的应用场景进行扩展。另外，研究还是要聚焦，刚开始可能迫于考核或者项目的压力，为了一些经费承接一些不是自己主业的项目，如果长久下去，就会失去核心竞争力。

问题 5：您先后获得国家科技进步二等奖两项、北京市科技进步奖两项、教育部科技进步一等奖、中科院首届优秀博士论文奖、北京市科技新星、中科院“卢嘉锡青年人才奖”等奖项，请问您如何看待这些奖项，又是如何在获奖方面做到持续产出的？

我能比较幸运地获得一些奖项，首先得感谢有一个高水平的平台，以及和一群优秀的人共事。我的前两个国家级奖项是在田捷教授带领下参与完成的。田老师现已年过六旬，仍然工作在第一线，基本上是到实验室最早，走得最晚的，他忘我的工作热情和精力让我们自愧不如。

这里结合医工合作来介绍一下我的另一个工作。

我于 2007 年获选了北京市科技新星，新星给大家提供了一个很好的交流平台，有不少医生参与。由此我与很多新星开展了合作，我的研究方向之一是脑影像分

析，一方面需要从医院获取数据，另一方面研究结果也需要医生的验证。所以与医生的合作非常重要，刚开始是与影像科医生合作，后来发现光有影像科医生还不够，我们还需要与临床医生合作，这样才能获取相关的疾病方面的知识。我们和同仁医院开展了 10 余年的合作，和影像中心、眼科开展了深度的合作，一起承担自然基金的课题，合作撰写了多篇文章，

神经眼科病变常导致失明和神经功能损害，引起重大社会卫生问题。神经眼科检出手段局限，诊断困难，难以实现精准治疗，缺乏对常见神经眼科疾病全脑改变及其发生机制的认知。针对上述问题，项目组历经多年研究，构建了神经眼科疾病临床-影像大数据库，在影像学技术创新的基础上，创立神经眼科疾病的影像分类评估体系、症状导向性影像侦查策略、阐明常见神经眼科疾病脑网络功能连接与调控规律，并基于新理论、新技术、新策略逐步创建了我国神经眼科影像体系，搭建了神经眼科疾病的影像研究平台。我们组与临床医生一起致力于神经眼科的特异性影像技术创新，用多模态影像技术，首次提出青光眼为全视路、多网络神经损害及证实先天复杂性斜视仅为核下性异常。创新性地提出利用自动分割技术研发出外侧膝状体体积测量新方法。该项目获得了教育部科技进步一等奖。后续我们又在这个工作的基础上，获得国家自然基金国际合作重点项目的支持，项目在有序进展中。

问题 6：您先后指导研究生 30 余人，其中 19 人获得博士学位，且培养的学生中多人获得中科院院长优秀奖、国家奖学金博士奖、北京市优秀毕业生等奖项，请问您是如何对他们进行有效培养和指导的呢？

学生的培养是一个长期的过程，十年树木，百年树人。在我读博士的时候，戴先生和田老师经常教导我们，“做人、做事，做学问”，做研究首先要做好人，做好事。我也是秉承这个传统来教育我的学生。我们组有大组会（侧重于工作进展汇报），小组会（侧重于文献交流），我和学生每周都会有一对一的交流。组会的作用非常重要，其提供了学生展示和交流自己工作的机会，把工作清晰地表达出来有时比编程实现更重要。在组会中，我会引导学生交流讨论，这样学生在交流讨论中学到新的知识。

我也鼓励学生多外出参会交流，我也有一些国际合作伙伴，曾先后把学生送到德国、美国等组合作交流。国内类似 VALSE、MICS 等活动，我也鼓励学生参加。除了科研指导以外，也要关心学生的日常生活，特别是疫情期间，长期的封闭管理给学生的身心带来了很大的影响。我和学生的交流是平等的，亦师亦友，组里也不要求打卡。

问题 7：您在中国科学院大学作为首席教授主讲《医学图象分析》和《脑机接口技术》等课程，《医学图像分析》获得国科大校级优秀课程，2021 年获中科院朱李月华教学奖，能否分享一下您的教学经验？您又是如何将教学和科研相结合的呢？

由于国科大的研究生校区在怀柔，去上课来回需要 3 个多小时，有时还需要在学校住一晚上，谢谢国科大的支持，为老师上课提供了班车和住宿。由于老师的科研任务重，国科大允许两名老师共同承担一门课程，我也要感谢我两门课的搭档，北京邮电大学的刘勇教授和中科院半导体所的王毅军教授。

我们建设了完整的课程资料库，所准备的讲稿提前放在网上供学生提前学习，向学生提供本门课程学习过程中应当阅读的相关经典文献以及最新的学术文献。同时针对课程的交叉学科的特点，所授材料既包括人工智能的基本理论和方法，也含有介绍神经科学的基础知识。实行启发式的和讨论式的教学方法，除了教授基本概念和知识以外，更重要的是引导学生的思考。指导学生阅读文献能够归纳出文献的主要贡献点，并尝试分析作者的写作动机，还需要分析其缺点在哪，有哪些值得改进

的地方。引导学生掌握正确的学习方法和思维方法。注重学生分析问题、解决问题能力的培养。另外，我们在课堂上也安排了学生来进行汇报，老师点评的方式，使学生更容易掌握本领域的要点。

科研教学互动相长，以科研促教学。授课教师承担了多个国家课题，发表了高水平的论文。所以在向学生讲授时，可以把握本学科的发展脉络，及时介绍最新的前沿进展。同时也让学生记住一些名字，比如本领域有名的实验室，有名的期刊，代表性方法的名字。加强学生互动，培养学生团队合作，以及做报告的能力。

问题 8：在繁忙的科研之余，请问您是如何来平衡科研教学工作、社会兼职及家庭的呢？您有什么业余爱好？不谈工作，私底下您认为自己是什么样的一个人？

说实话，由于忙于工作，对家庭比较愧疚，照顾较少。我要特别谢谢我的夫人，她承担了大部分的家务和孩子的教育工作。但是我们家庭是一个民主的家庭，有事大家一起商量。

我认为我是一个真诚、坦率的人，对朋友真诚，做事情比较直接。

问题 9：您曾在美国罗切斯特大学放射系做博士后研究，在加拿大滑铁卢大学做高级研究学者，在美国北卡大学教堂山分校做访问教授，关于研究者是否需要出国交流访问，以及如何选择和把握出国交流的机会，您的看法是什么？

如果有机会，还是要出国看看，但是最后一定要回国效力，我也是向所有问我这样问题的人说，包括我对孩子的教育也是如此。另外，出国一定要有明确的规划。我第一次出国，只是觉得身边的人都出去了，也想去看看，所以有个机会就出去了，其实出国的那个组的方向和自己的方向不是很匹配，导致耽误了时间，进步也不大。我在 14 年时拿到了国家支持，这个时候我就选择了与我方向非常匹配的组，UNC 沈定刚教授组，在这个组里我也进一步学到了科研设计、组会交流等经验，更重要的是结识了一群小伙伴，可以一起开心地讨论和交流，回国后还保持联系，互相支持，他们是 MICS 的中坚力量，我也是首届 MICS 组委会轮值主席。

问题 10：如果吐露研究工作者的心声，您最想说的是什么？

希望能给科研人员减负，让破五维落在实处，使得科研人员可以体面、健康、快乐地做研究。

责任编委 余烨 赵振兵

何晖光

中科院自动化所研究员，博士生导师，自动化所学位委员会委员，中国科学院大学岗位教授，上海科技大学特聘教授，中科院朱李月华优秀教师，建国七十周年纪念章获得者。CSIG 视觉大数据和机器视觉专委会常委，CCF-CV 专委会执行委员，IEEE 高级会员，CCF/CSIG 杰出会员。先后主持包括 7 项国家自然科学基金（包括 2 项重点）、2 项 863 项目、国家重点研究计划课题等多个重要项目。先后获得国家科技进步二等奖两项（分别排名第二、第三），北京市科技进步奖两项，教育部科技进步一等奖（排名第三），获中科院首届优秀博士论文奖，北京市科技新星，中科院“卢嘉锡青年人才奖”，中科院青年创新促进会优秀会员等奖项，其研究领域为人工智能，医学影像分析，脑 - 机接口等，其研究结果在 IEEE TPAMI/TNNLS/TCYB/TMI、NeuroImage、HBM、MedIA、ICML 等核心期刊及国际主流会议上发表文章 200 余篇。自动化学报编委，中国图象图形学报编委，KJW 重大专项专家组专家。

委员好消息

✪ 2023 年 3 月 29 日，爱思唯尔发布了 2022 中国高被引学者榜单，全国共有 5216 人上榜，CCF-CV 专委会 53 位执行委员上榜：百度**王井东**，北京大学**林宙辰**、**彭宇新**，北京邮电大学**邓伟洪**，重庆邮电大学**高新波**，大连理工大学**卢湖川**、**王栋**，复旦大学**姜育刚**，国防科技大学**刘丽**，哈尔滨工业大学**徐勇**、**左旺孟**，杭州电子科技大学**俞俊**，华中科技大学**白翔**、**尤新革**，江西财经大学**方玉明**，南京大学**王利民**、**吴建鑫**，南京航空航天大学**谭晓阳**，南京理工大学**李泽超**、**潘金山**、**唐金辉**、**张姗姗**，南开大学**程明明**，清华大学**代季峰**、**郭振华**、**黄高**、**鲁继文**，厦门大学**纪荣嵘**，上海交通大学**卢策吾**、**马超**，天津理工大学**陈胜勇**，同济大学**张林**，武汉大学**荆晓远**、**夏桂松**，西安电子科技大学**邓成**、**董伟生**，西安理工大学**魏巍**，西北工业大学**程塨**、**韩军伟**、**聂飞平**、**王琦**，云南大学**陶大鹏**，中国科学技术大学**张天柱**，中国科学院计算技术研究所**韩琥**、**山世光**，中国科学院深圳先进技术研究院**乔宇**，中国科学院西安光学精密机械研究所**卢孝强**，中国科学院信息工程研究所**任文琦**，中国科学院自动化研究所**雷震**、**王亮**，中山大学**操晓春**、**林倞**、**郑伟诗**。

✪ 2023 年 4 月 10 日，2023 年陕西高等学校科学技术研究优秀成果奖拟授奖成果名单公示，CCF-CV 专委会 5 执行委员的项目入选：西北工业大学**程塨**和**韩军伟**等完成的“高分辨率光学遥感图像目标检测理论与方法研究”拟授特等奖，西北工业大学**王庆**等完成的“多视光场理论与方法”、**王鹏**参与完成的“边缘网络智能安全防护关键技术及应用”拟授一等奖，西安理工大学**蔺广逢**等完成的“视觉目标表征异质结构关系挖掘与动态演化特性研究”拟授二等奖。

✪ 2023 年 4 月 11 日，教育部公示了第二批国家级一流本科课程认定结果，CCF-CV 专委会 9 位执行委员的课程入选：南京信息工程大学**刘青山**主持的《数字图像处理》、太原理工大学**赵涓涓**主持的《面向对象程序设计基础》、南京信息工程大学**陈允杰**主讲的《计算方法》、西安电子科技大学**冯冬竹**主讲的《数字图像处理》、上海交通大学**林巍峣**主讲的《通信原理》、西安电子科技大学**苗启广**主讲的《离散数学》获评线下一流课程，华中科技大学**高常鑫**主讲的《C 语言程序设计》获评线上一流课程，上海大学**方昱春**主讲的《中国手语文化》和湖南大学**李智勇**主讲的《计算与人工智能概论》获评线上线下混合式一流课程）。

✪ 2023 年 4 月 19 日，中国计算机学会公布了 2023 上半年 CCF 高级会员评选结果，CCF-CV 专委会 10 执行委员晋升为 CCF 高级会员，他们是：中国科学院沈阳自动化研究所**丛杨**、西北工业大学**戴玉超**、百度**丁二锐**、燕山大学**顾广华**、南京信息工程大学**刘青山**、北京交通大学**阮秋琦**、美团点评**王栋**、西北工业大学**王鹏**、北京大学**王亦洲**和西安交通大学**钟德星**。

✪ 2023 年 4 月 20 日，陕西省人民政府发布了 2022 年度陕西省科学技术奖励决定，CCF-CV 专委会执行委员、西北工业大学**戴玉超**参与的项目“复杂目标多视角高光谱联合感知与智能识别的理论与方法”获自然科学一等奖。

✪ 2023 年 5 月 13 日，高校在线开放课程联盟联席会发布了慕课十年典型案例，CCF-CV 专委会执行委员、哈尔滨工程大学**刘海波**的慕课案例“突破线上课程困境提升学习实效”入选。

✪ 2023 年 5 月 14 日，2022 年度重庆市科学技术奖拟奖名单公示，CCF-CV 专委会执行委员、重庆邮电大

学**高新波**主持的“基于内容生成的视频画质提升理论与方法”拟授自然科学一等奖。

✪ 2023 年 5 月 15 日，2022 年高等教育本科和研究生国家级教学成果奖拟授奖成果公示，CCF-CV 专委会 10 位执行委员的成果拟授高等教育（本科）国家级教学成果二等奖：东南大学**耿新**，清华大学**鲁继文**，天津大学**雷建军**，复旦大学**姜育刚**，山东大学**许信顺**，中南大学**陈再良**，湖南大学**张汗灵**、**李智勇**，中山大学**郑伟诗**，西安电子科技大学**苗启广**，3 位执行委员的成果拟授高等教育（研究生）国家级教学成果一等奖：西安交通大学**薛建儒**、西北工业大学**张艳宁**和**王鹏**，5 位执行委员的成果拟授高等教育（研究生）国家级教学成果二等奖：厦门大学**王菡子**、湖南大学**李智勇**、大连理工大学**杨鑫**、重庆邮电大学**高新波**和**肖斌**。

✪ 2023 年 5 月 25 日，教育部公示了新一批享受政府特殊津贴推荐人选，CCF-CV 专委会执行委员、西安电子科技大学**邓成**入选。

责任编委 刘海波

视觉模型诊断调试工具

北京航空航天大学 贾俊龙 黄雷

基础模型因其令人印象深刻的表现和能力受到了学术界及工业界的广泛关注，人工智能正在经历范式上的转变。为了确保模型训练的可复现性、准确性和可靠性，以及满足法律和伦理要求，我们需要强大而可靠的工具来帮助我们理解和调试这些模型的行为。

本文重点介绍可用于模型诊断调试的工具，主要包括 Captum、Cockpit、Taiyi 三种即插即用的工具。

1、Captum

介绍： Captum（拉丁语中的“理解”）是由 PyTorch 社区开发的一个 Python 库，用于对构建在 PyTorch 上的模型进行可解释性分析,旨在帮助开发者深入理解模型的决策过程，并提供可视化工具来解释模型的行为。Captum 是一个开源、可扩展的库。随着模型复杂性的增加以及由此导致的透明度的缺乏，模型可解释性方法变得越来越重要。模型理解既是一个活跃的研究领域，也是使用机器学习的跨行业实际应用的重点领域。Captum 主要从 Attribution、Robustness、Concept 以及 Influential Examples 等方面来对模型进行解释分析。

Attribution Attribution 是神经网络可解释性分析的核心概念之一。Attribution 算法主要分为两类，一类是基于梯度的 Attribution 算法，另一类是基于遮挡的 Attribution 算法。Captum 集成了这两类算法，并针对模型的不同粒度进行了拆分，分为基于 Feature，基于 Neutral 以及基于 Layer。研究人员可以通过 Captum 全方位的对模型进行剖析，帮助改进模型和排除模型故障。图 1 是通过集成梯度算法对 VQA 模型进行分析的示例。

Robustness Captum 在鲁棒性 (Robustness) 方面也提供了有价值的功能。鲁棒性是指模型对于输入数据中的扰动和变化的敏感程度。Captum 提供了几种鲁棒性评估和增强方法。通过使用 Captum，研究人员可以识别模型在输入扰动或对抗攻击下的表现，并提供对抗性训练和防御策略的指导，这有助于增强模型的可信度和稳定性。

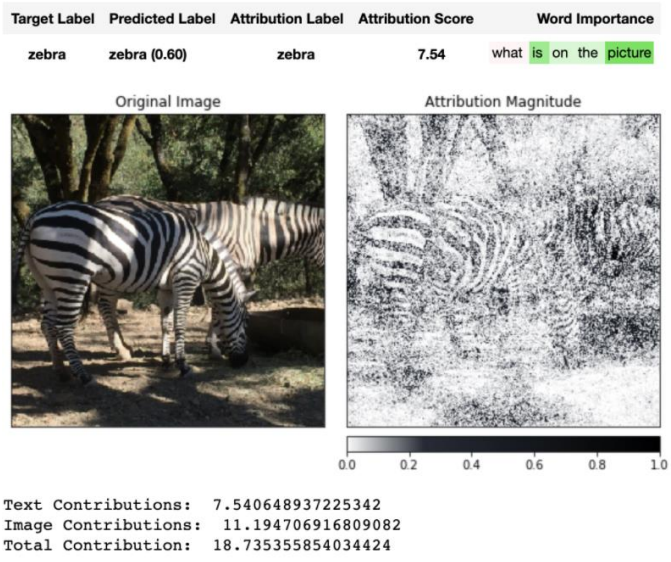


图 1 Captum Attribution 分析结果图

Concept Concept 是指关键概念特征，Captum 提供了 Concept Attribution 方法，可帮助用户理解模

型在决策时所依赖的概念。通过对模型进行概念解释，研究人员可以验证模型是否具备正确的领域知识和概念理解，图 2 是对 GoogleNet 模型的 TCAV 分数的可视化分析结果。

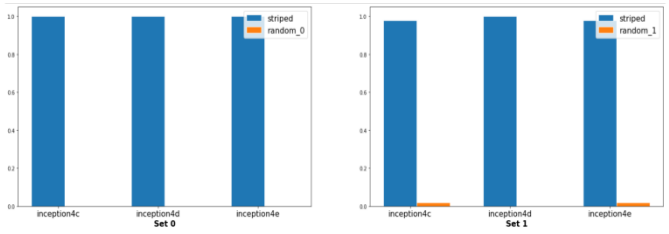


图 2 Captum 概念分析结果图

Influential Examples 这个功能指的是给定训练样例对给定测试样例的影响分数，粗略地说，表示如果将给定训练样例从训练数据集中移除，则给定测试样例的损失会高多少。Captum 可以识别测试样例中最积极影响分数的样例和最消极分数的样例，通过识别特殊样例，模型开发者对理解模型的决策有很大帮助。

用法 使用 Captum 工具进行模型解释分析的基本流程是：导入库，创建模型，选择属性方法，计算属性，可视化结果。除此之外，Captum 还提供了一套交互式可视化工具，称为 Captum Insights，它可以帮助用户通过交互式界面探索模型的属性和预测结果。可以使用 Captum.insights 模块中的类来创建一个 Captum Insights 应用程序。总而言之通过 Captum，我们可以更好地理解深度学习模型的决策过程，并获得关于输入特征的重要信息。

代码地址： https://github.com/pytorch/captum

2、Cockpit

介绍： 当研究人员训练深度学习模型时，他们会非常“Flying Blind”。常用的训练时的诊断方法（例如查看训练或者测试时的 Loss）是有限的。为了解决这个问题，作者开发了一套工具 Cockpit，中文翻译为驾驶舱，意思是研究人员可以像通过仪表盘控制飞机一样通过仔细地观察训练过程中的状态报告来控制模型的训练，如图 3 所示。模型训练中经常遇到的三类错误：第一类是实现错误，这是是低级的编程错误，可以通过强大的集成开发环境来避免或者解决这种错误；第二种是训练错误，会导致不必要的效率低下，甚至是不成功的训练。例如，它们可能源于错误的数据处理、所选的模型体系结构或错误选择的超参数，这种错误通过集成开发环境是无法解决的；第三种是预测错误，描述了一个训练过的模型对特定例子的错误预测。Cockpit 的重点是有效地识别训练错误。通过使用 Cockpit，研究人员可以科学的调整超参数、观察局部 Loss 几何的 Hession 属性以及可视化网络的内部动态等。

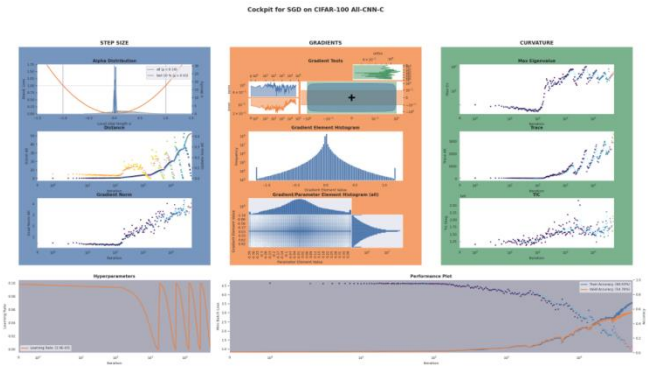


图 3 Cockpit 训练总览图

Adapting Hyperparameters 深度学习的一个大挑战是正确设置超参数，目前主要是通过参数搜索的试错来完成的。Cockpit 组合了已发表论文的方法来帮助用户更科学的设置超参数，例如通过观察 **Alpha** 属性判断当前模型是否跨过了低谷，通过观察 **Distances** 属性判断训练是不是有效的以及一些基于梯度的方法回答当前批次的训练噪声或者训练轨迹的问题。

Hessian Properties For Local Loss Geometry Hessian 属性对局部 Loss 的 Landscape 的直觉在很多方面都有帮助，例如帮助诊断训练是否被卡住，以调整 Learning Rate。深度学习的一个主要的挑战是大量的参数，仅仅将高维的权重投影到低维可能表现不是很直观，观察一些极端的特例或者平均的行为可能有助于调试。一阶信息可能得不到有效的结果，Cockpit 通过二阶的方法例如 Hessian 的特征值以及 TIC 来进行度量。

Visualizing Internal Network Dynamics 在训练期间进行监测梯度直方图以及对应的参数，整个模型的状态和动态可以在单个图中可视化出来。与跟踪参数和梯度范数相比，这提供了一个更细粒度的训练视图。通过观察梯度和参数的直方图可以更好的理解训练过程中模型发生了什么。

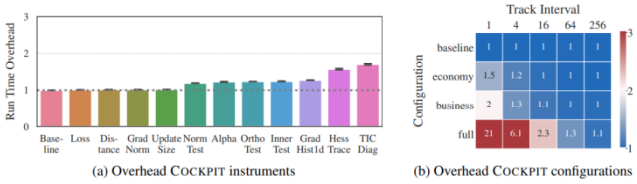


图 4 Cockpit 不同配置运行时间开销

用法 使用 Cockpit 进行优化算法分析的基本流程是：导入库，创建模型，创建模型上下文定义量，运行指标计算，分析结果。通过 Cockpit，我们可以更好地理解和调试优化算法的行为，以及监控和优化模型训练过程中的各种指标。除此之外，使用 Cockpit 还要注意观察指标和性能方面的平衡，不同配置的运行时间开销如图 4 所示。

论文地址： https://arxiv.org/pdf/2102.06604.pdf

代码地址： https://github.com/f-dangel/cockpit

3、Taiyi

介绍： Taiyi 与 Cockpit 不同，Taiyi 调试器希望能观察到模型训练过程中细粒度的变化值以及统计量，这种变化或者统计量能够帮助从业人员更好的理解网络训练过程中模型的动态变化情况，帮助从业人员减少训练过程中盲目的参数搜索以及根据细粒度的信息设计更优秀的算法。Taiyi 将需要观察的模块以及指标解耦，用户可以定制自己想要的观察结果，Taiyi 的指标分为两类，一类是单步计算的指标，另一类是多步计算的指标。

Single Step Quantity Single Quantity 指的是一次计算就能得到结果的统计量，例如条件数、梯度范数、权重范数等等，通过观察这些指标，用户可以清晰看到模块在训练过程中的变化量，判断当前训练模块是否是良态或者到当前模块信息流动是否正常，这种细粒度的变化使得用户对模型有这更精准的掌控。

Multi Step Quantity Multi Quantity 是指指标需要通过多次计算才能统计得到，例如 MeanTID，VarTID 等等。我们都知道，BN 模块在深度学习训练过程中起着至关重要的作用，但由于其本身计算方式的缺陷，这种训练和测试过程中的 Gap 使得模型非常不稳定，如图 5，当 BN 模块的 Batch 统计量和总体估计统计量差距较大时，模型的测试 Loss 也随之抖动，当我们观察到这样的现象后，我们可以通过约束其中的 Gap 使得模型更加稳定。

用法 使用 Taiyi 进行训练过程中指标监视的基本流程是：定义观察模型脚本的配置文件，根据配置文件初始化 Monitor 类，计算属性，可视化结果。Taiyi 将结果计算和可视化进行解耦，使用者可将获取得到的数据按照熟悉的可视化工具进行展示。

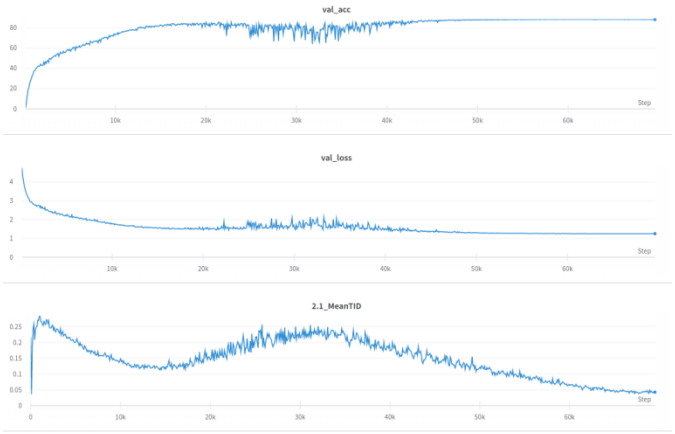


图 5 训练过程中 MeanTID 以及测试准确率结果图

论文地址： https://arxiv.org/pdf/2002.10801.pdf

代码地址： https://github.com/DLCV-BUAA/Taiyi

贾俊龙

硕士研究生，北京航空航天大学人工智能研究院，研究方向为计算机视觉。

黄 雷

北京航空航天大学人工智能研究院从事教学与科研工作，多次担任 CVPR，ICCV，NIPS 等顶级会议和期刊的审稿人。研究方向为模式识别、机器学习。

个人主页：https://huangleibuaa.github.io/

多目标跟踪数据集

南京大学 杨旖纯 崔玉涛 王利民

目标跟踪是指在视频中同时追踪多个目标的位置和轨迹，它是计算机视觉中的一个重要领域，被广泛应用于视频监控、无人驾驶、行为分析等领域。

在过去的几十年中，多目标跟踪技术取得了长足的发展。针对不同的场景，人们提出了许多多目标跟踪数据集。现有数据集侧重于跟踪人类目标，随着自动驾驶受到企业的青睐，一些专门针对自动驾驶的数据集专注于驾驶场景下的车辆和行人。其他数据集关注的对象类别更加多样化，以研究长尾分布下的对象跟踪。这些多目标跟踪数据集为评估算法的性能提供了相对客观的衡量标准，推动了该领域的标准化发展，是科学研究的重要指南。

本文重点介绍跟踪人类目标的数据集，包括 MOT17, DanceTrack, SportsMOT 等代表性的数据集。

1、MOT17 数据集

介绍： MOT Challenge 系列是目前使用最广泛的数据集，由慕尼黑工业大学发布。包括 MOT15、MOT16、MOT17、MOT20 及其他版本。MOT17 跟踪的目标是在不同的光照、天气条件下，在街道或室内的行人，具有拥挤、形变小、移速慢、外观差别大等特点。当一个行人离开画面又再次进入的时候，会被标识为一个新的目标。视频由移动或静态的摄像机拍摄，摄像机的位置可以在高处（监控视角）、中位（行人高度）或低处。由于行人密度高以及频繁遮挡，该数据集具有很高的难度。

图 1 MOT17 数据集示例

MOT17 帧率在 14FPS-30FPS 之间，总共包含 14 个视频序列（7 个用于训练，7 个用于测试），其中包括 11235 帧，1342 个轨迹和 292733 个框。数据格式为：<frame>, <id>, <bb_left>, <bb_top>, <bb_width>, <bb_height>, <conf>, -1, -1.

MOT Challenge 官方为每个视频提供了三组检测结果（边界框信息）：一组使用 Fast R-CNN，一组使用精度最高的 DPM，一组使用规模相关池 SDP。跟踪算法必须具备很好的关联能力，才能在同样的检测结果下脱颖而出；必须具有足够的通用性和鲁棒性，才能在不同的检测结果下都获得良好的性能。

MOT17 虽然遮挡频繁、人物密集，但只包含少量视频和场景，且行人的运动也非常规则线性，外观清晰独

特。即使不通过运动模型，只通过纯外观匹配的关联也取得了较好的效果。

数据集下载地址：https://motchallenge.net/

相关论文链接：MOT16: A Benchmark for Multi-Object Tracking https://arxiv.org/abs/1603.00831

2、DanceTrack 数据集

介绍：DanceTrack 于 2021 年由香港大学、卡耐基梅隆大学、字节跳动联合发布，跟踪的目标为舞者和运动员等等，包含各种各样的舞蹈场景，如街舞、流行舞、古典舞等，以及一些体育场景，如体操、中国功夫和啦啦队舞蹈。

图 2 DanceTrack 数据集示例

DanceTrack 数据集中总共收集了 100 个视频，其中 40 个视频作为训练集，25 个作为验证集，35 个作为测试集，帧率为 20FPS。该数据集包含超过 100K 的图像帧，与 MOT Challenge 数据集相比，DanceTrack 的容量要大 10 倍。数据格式为：

<frame>, <id>, <bb_left>, <bb_top>, <bb_width>, <bb_height>, 1, 1, 1.

该数据集的特点是：（1）统一的外观：视频中的人穿着非常相似甚至相同的衣服，使得他们的视觉特征很难通过 Re-ID 模型来区分；（2）复杂的运动：人们通常有大范围的运动、复杂的姿势变化和相对位置的交换，带来了大面积的遮挡和交叉，这对运动建模提出了更高的要求。

DanceTrack 关注了现实世界中复杂的情况，分析了现有数据集存在的严重偏差，提出建模复杂运动模式的能力对于构建更全面智能的跟踪器是必要的，为接下来的技术发展提供了方向。

数据集下载地址：https://dancetrack.github.io/

相关论文链接：DanceTrack: Multi-Object Tracking in Uniform Appearance and Diverse Motion https://arxiv.org/abs/2111.14690

算法评测平台：
https://codalab.lisn.upsaclay.fr/competitions/5830

3、SportsMOT 数据集

介绍：运动场景下的多目标跟踪任务是计算机视觉最底层的任务，结合其他技术，可以自动执行游戏解说、智能裁判、分数统计、球员能力评估、训练计划制定、比赛策略改进等高级任务。

尽管人们对体育分析的需求越来越大，但由于背景复杂、运动员动作迅速、镜头移动速度快等原因，目前缺乏针对各种体育场景的大型多目标跟踪数据集。

SportsMOT 数据集于 2022 年由南京大学媒体计算研究组发布，数据集中的所有视频参考 MultiSports 数据集，来自于 Youtube 上不同国家和地区的专业比赛，如奥运会，NCAA 冠军赛，以及 NBA，所有视频都没有镜头切换。

SportsMOT 分为三种运动类别：篮球、足球、排球，其中篮球场景的多目标跟踪最难，排球场景的最简单。足球比赛提供的是室外场景，篮球和排球比赛提供的是室内场景。此外，比赛场地的视角也各不相同，包括 NBA 中常见的侧视图，排球比赛中的发球区视图，足球比赛中的鸟瞰图。

SportsMOT 选择了运动场上所有的运动员（不包括观众、裁判、替补队员）作为跟踪目标。该数据集不仅规模大，而且质量高，密集注释了所有场上球员的位置边界框及独有的 ID，旨在弥补多目标跟踪领域运动场景数据集的缺失。

图 3 SportsMOT 数据集示例

SportsMOT 包含 240 个视频，其中 72 个为训练集，72 个为验证集，96 个为测试集，共有 150379 帧(相当于 MOT17 的 15 倍)和 1629490 个注释框(相当于 MOT17 的 3 倍)，帧率为 25FPS，视频总长度约为 6015s，训练集、验证集、测试集的帧数比例也与个数比例相当。数据格式与 MOT Challenge 保持一致，使用 HOTA 指标来为所有的跟踪算法排名。

表 1 现有 MOT 数据集与 SportsMOT 的数据量比较

Dataset	Videos	Frames	Length (s)	Bbox	Tracks
MOT17	14	11,235	463	292,733	1,342
MOT20	8	13,410	535	**1,652,040**	**3,456**
DanceTrack	100	105,855	5,292	-	990
SportsMOT	**240**	**150,379**	**6,015**	1,629,490	3,401

表 2 SportsMOT 三种球类的数据量比较

Category	Frames	Tracks	Track gap len.	Track len.	Bboxes per frame
Basketball	845.4	10	68.7	767.9	9.1
Volleyball	360.4	12	38.2	335.9	11.2
Football	673.9	20.5	116.1	422.1	12.8
Total	626.6	14.2	96.6	479.1	10.8

SportsMOT 数据集有以下两个特点：

1. 快速和变速运动。由于体育运动场景的特殊性，SportsMOT 具有独特的运动模式，即快速和变速运动，运动员通常快速移动，频繁地改变速度方向和快慢。因此对基于简单运动假设的跟踪器提出了重大挑战，也鼓励跟踪器以更动态和自适应的方式对物体运动进行建模。

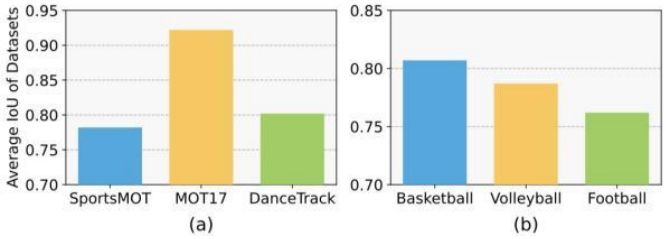


图 4 基于卡尔曼滤波器的 IoU（越低运动越多变）

2. 相似但可区分的外观。外观是跟踪器用来区分不同目标的另一种线索。在 SportsMOT 中，运动员的球衣也非常相似，然而，球衣有着不同的号码，运动员的发型、鞋子、姿势也不尽相同，从而导致相似但可区分的外观。SportsMOT 的目标相比 MOT17 是更相似的，但相比 DanceTrack 是更可区分的，这就需要外观模型挖掘出更具判别力和泛化性的表达能力。

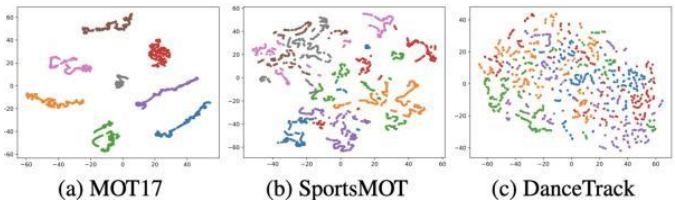


图 5 使用 t-SNE 对 ReID 特征的可视化

数据集下载地址： https://github.com/MCG-NJU/SportsMOT

相关论文链接： SportsMOT: A Large Multi-Object Tracking Dataset in Multiple Sports Scenes

https://arxiv.org/abs/2304.05170

算法评测平台：

https://codalab.lisn.upsaclay.fr/competitions/12424

责任编委 李策 王田

杨旖纯

博士研究生，南京大学人工智能学院，研究方向为计算机视觉。

崔玉涛

博士研究生，南京大学计算机科学与技术系，研究方向为计算机视觉、物体跟踪。

王利民

教授，博士生导师，南京大学教授。研究方向包括计算机视觉、视频理解和动作识别。

个人主页：http://wanglimin.github.io/

好 文 推 荐

南开大学团队关于“PoolNet+: Exploring the Potential of Pooling for Salient Object Detection”最新成果发表在 IEEE TPAMI 2023。

论文：Jiang-Jiang Liu , Qibin Hou , Zhi-Ang Liu , and Ming-Ming Cheng. PoolNet+: Exploring the Potential of Pooling for Salient Object Detection, IEEE TPAMI, 45 (1): 887-904, 2023

显著性目标检测旨在检测给定图像中最为突出的目标，常常被用于视觉追踪、内容感知的图像裁剪和编辑、图像检索、视频分割、机器人导航等计算机视觉领域。作为一项基本的视觉任务，显著性目标检测已逐渐成为计算机视觉中不可缺少的组成部分，对研究更高层次的视觉问题具有重要意义。近年来，卷积神经网络可以在多个尺度空间中同时提取高级语义和低级细节的能力极大推动了显著性目标检测的发展。

该团队通过扩展池化技术在卷积神经网络中的作用来探索池化技术在显著目标检测任务上的潜力。该团队提出了两种基于池化的模块。如图 1 所示，首先基于 U 型结构自底向上的路径构建全局引导模块（Global Guidance Module, GGM），将潜在显著目标的位置信息分层到不同的特征层次。此外，进一步设计了特征聚合模块（Feature Aggregation Module, FAM），将自顶向下路径中的粗级语义信息与细级特征无缝融合。这两个模块可以逐步细化高级语义特征，得到细节丰富的显著性映射图。实验结果表明，与现有方法相比，该方法可以更准确地定位出显著目标，并使细节更清晰，性能有了很大提高。此外，所提方法具有更高的运行速度，在 300*400 的图像上可达到 53 FPS。为了使所提方法更好应用于移动应用程序，该团队以 MobileNetV2 为骨干，重新定制基于池化的模块结构。该改进主干在手机端达到了 66 FPS 的运行速度。

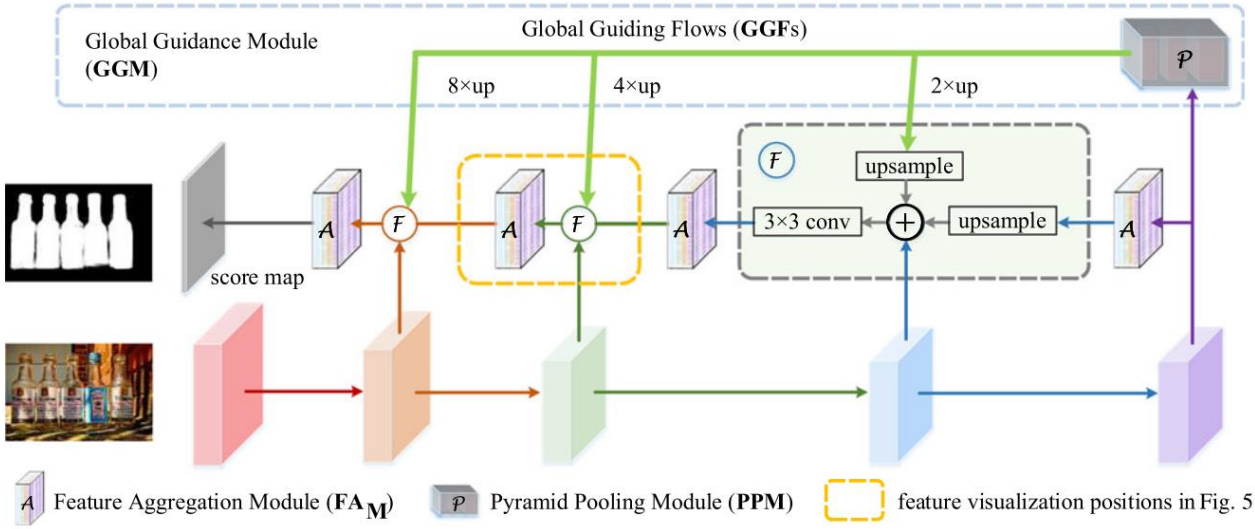


图 1 所提方法结构流程图

责任编委 樊鑫 贾同

好 文 推 荐

香港城市大学和斯坦福大学相关团队关于“Learning to Deblur using Light Field Generated and Real Defocus Images”最新成果发表在 CVPR-2022。

论文：Ruan L Y, Chen B, Li J Z, et al. Learning to Deblur using Light Field Generated and Real Defocus Images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16283-16292.

单幅散焦图像去模糊的主要目的是将输入的模糊图像恢复为清晰图像，以恢复图像更多的纹理以及细节信息，从而揭示图像中存在的潜在信息，以便用于后续的目标检测、图像分割等一些高级的计算机视觉任务中。目前处理散焦图像去模糊任务中存在的难点通常为散焦图像的模糊核未知且存在空间变化，普通双镜头相机无法获取像素级别的散焦与全聚焦图像对训练模型。

为了解决上述问题，文章提出了基于 U-net 的多尺度去模糊模型，并配备了新颖的动态残差块，以从粗到细的方式重建清晰的图像，无需进行模糊核的估计。网络总体结构如图 1 所示。首先为解决非均匀或者空间变化的图像模糊问题，文章结合了残差网络与动态局部滤波网络两种模型的优点，提出了动态残差块的概念，通过将动态残差块与解码器中相应尺度的模块进行连接来逐步重建清晰的图像；其次普通双镜头相机无法捕捉像素级别的散焦与全聚焦图像对，进而导致训练出的模型性能表现不佳。因此，文章提出了利用光场合成孔径与重聚焦技术来获得精确的散焦与全聚焦图像对。同时为了充分利用光场数据的优势以及克服光场数据不足的劣势，文章提出了一个新的训练策略即首先将光场获得的 LFDOF 数据集用于模型的训练，以获得高精确的图像对，然后再将双镜头相机捕捉得到的 DPDD 数据集用于模型的训练，以消除两域散焦模糊间的差异，进而提高模型的泛化性能。

文章在 DPDD、RealDOF、CUHK 和 PiexlDP 数据集上的实验结果表明，在基于深度学习的散焦图像去模糊方面，文章提出的方法优于现有方法，并证明了所提出的基于 U-net 的多尺度去模糊模型以及相应训练策略的有效性。

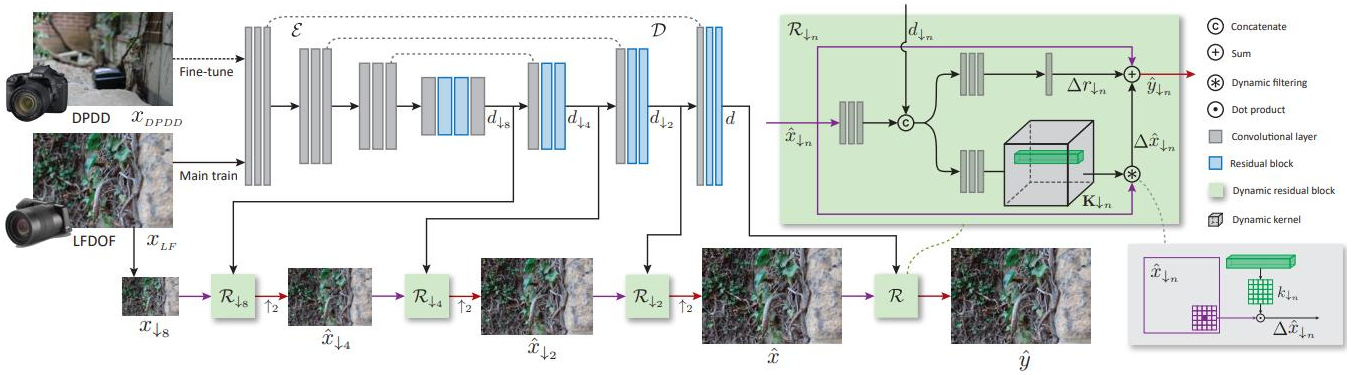


图 1 所提出的单幅散焦图像去模糊网络结构图

责任编委 贾同 樊鑫

好 文 推 荐

不列颠哥伦比亚大学和 Google 公司等相关团队关于“Light Field Neural Rendering”最新成果发表在 CVPR-2022。

论文：Suhail M, Esteves C, Sigal L, et al. Light Field Neural Rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 8269-8279.

新视图合成是计算机视觉、计算机图形学交叉领域的一个重点难题，具体指从一个场景的多张图片中创建该场景的新视图。新视图合成被广泛应用于医学影像分析、虚拟现实等领域。用于新视图合成的经典光场渲染方法可以精确地模拟复杂视图效果，例如反射、折射和半透明，但是该方法对密集视图采样依赖性强；基于几何重构的方法只需要输入稀疏视图，但不能精确模拟非朗伯效应。因此，文章提出了一个基于 Transformer 的两阶段光场神经渲染模型，结合了以上两种方法的优势并缓解了其局限性，使模型仅在稀疏视图集中学习场景几何，就能精确模拟复杂视图效果。

文章所提出模型采用两阶段 Transformer 方法将 Patch 集合映射到目标像素颜色预测中。第一阶段 Transformer 沿每条极线聚合特征，使用自注意力机制在每个参考图像上寻找目标像素的潜在对应关系；第二阶段 Transformer 沿参考视图聚合特征，负责推理视图遮挡和视线依赖效应。模型采用几何约束特征聚合的方法降低了模型对密集视图采样的依赖。与 NeRF 不同，LFNR 模型放弃了体渲染方法，使用基于图像渲染的方法精确模拟非朗伯效应。在训练过程中，模型使用两阶段 Transformer 对目标像素颜色做二次预测，将它与真实颜色的 L2 损失定义为辅助损失，进一步提高模型渲染精度。LFNR 模型架构如图 1 所示。

文章在 RFF、Shiny、Blender 数据集上进行实验，实验结果表明 LFNR 模型的渲染精度优于 NeRF、NeX 等先进模型，其 PSNR 提升幅度高达 5dB。同时，与其他基于 Transformer 的模型相比，LFNR 模型渲染速度提高了 3 倍。

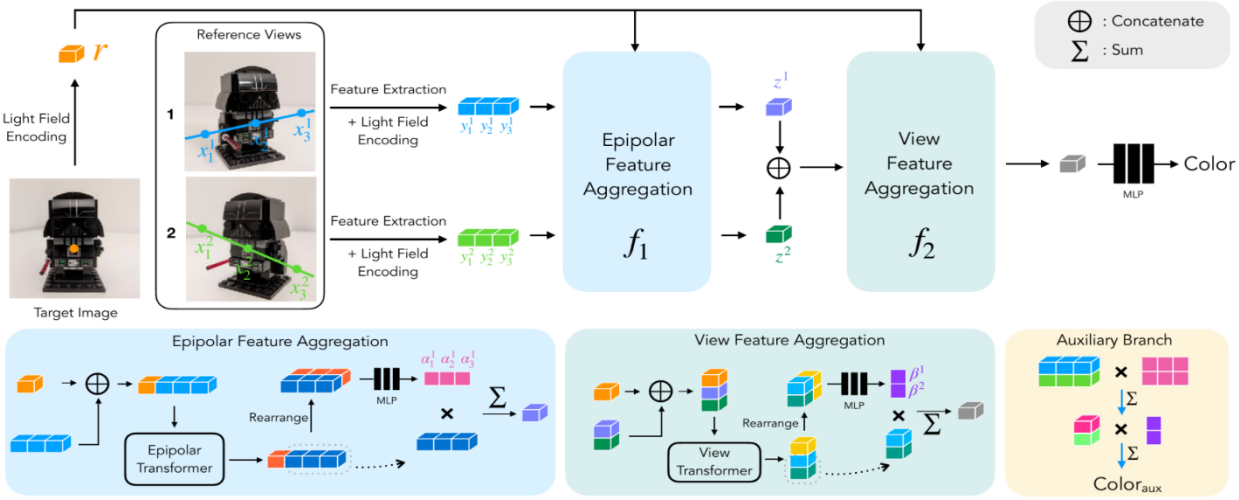


图 1 LFNR 模型架构图

责任编委 贾同 李策

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 Letters 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。

第六届 PRCV 将于 2023 年 10 月 13 日至 10 月 15 日在厦门举办，由厦门大学承办。会议旨在汇聚国内国外模式识别和计算机视觉理论与应用研究的广大科研工作者及工业界同行，共同分享我国模式识别与计算机视觉领域的最新理论和技术成果。通过此次会议，进一步加强本领域的同行与东南沿海地区的学者和企业进行学术交流和技术碰撞，从而促进模式识别与计算机视觉领域的协同合作与融合创新。

责任编委 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
AAAI 2024	2024.2.20-28	Vancouver, Canada	2023.08.16	https://aaai.org/aaai-conference/
ICDM 2023	2023.12.1-4	Shanghai, China	2023.07.02	https://www.cloud-conf.net/icdm2023/
SIGIR-AP 2023	2023.11.26-29	Beijing, China	2023.07.04	http://www.sigir-ap.org/sigir-ap-2023/

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
JBHI	New Age of Deep Learning on Drug Discovery: Research to Practice	https://www.embs.org/jbhi/wp-content/uploads/sites/18/2023/03/New-Age-of-deep-lerning-on-drug-JBHI-CFP-Improved.pdf	2023.07.05
PRL	Explainable Representation Learning for Multi-view/modal Data（ERLMD）	https://www.sciencedirect.com/journal/pattern-recognition-letters/about/call-for-papers#special-section-for-awarded-papers-from-11th-iberian-conference-on-pattern-recognition-and-image-analysis-ibpria-2023-ibpria-2023	2023.07.20
JBHI	Trustworthy Machine Learning for Health Informatics	https://www.embs.org/jbhi/wp-content/uploads/sites/18/2023/05/Trustworthy-20230401_CallforPapers_TMLH.pdf	2023.09.01
IVC	Special issue on Synthetic Data in Generalizable Video Analytics	https://www.sciencedirect.com/journal/image-and-vision-computing/about/call-for-papers	2023.09.15

潜心育人 勇攀高峰∞刘文予教授专访

自 50 年代以来，我国在计算机视觉领域展开了相关的科研工作。而今，我国已经拥有一支庞大的、在这一领域辛勤耕耘且能与世界一流水平并驾齐驱的科研队伍。在这一过程中，有一批见证了视觉领域发展、为我国计算机视觉领域的奠基做出了重大贡献的先驱者。

《视界专访》栏目希望通过对计算机视觉研究历史、进展的见证者作一个系列专访，以帮助从事计算机视觉及相关领域的科研工作者或爱好者，全方面地了解 50 年代以来信息技术、信号处理技术以及计算机视觉相关的一些历史发展及进步，也希望能帮助我们在见证这段历史的同时，展望计算机视觉领域的未来。

我是负责本次专访的主要采访人，北京邮电大学明悦。本次采访通过微信交流完成，相关问题由 CCF-CV 专委会的《视界专访》组提供。为能更好地帮助我们回顾本次采访，我们采用问答加书面回顾的形式来表述。以下是刘文予教授的简介和专访内容。


图 1 刘文予教授

明悦（采访者，后缩写为明）：您于 1986 年在清华大学获得学士学位，后在华中科技大学获得硕士和博士学位，并留校任教至今。能分享一下您的求学历程，以及期间难忘的经历故事给大家吗？是什么原因让您坚定不移地选择了学术研究这条道路？

刘文予（后缩写为刘）：我是 1986 年本科毕业于清华大学计算机系。当时清华大学计算机系有四个研究方向，计算机系统结构、计算机信息处理、智能控制（人工智能）和计算机软件。我的研究方向主要是信号处理，尤其是图像处理方向。我的本科毕业设计参与国家“六五”攻关项目，指导老师是徐光佑教授，杨士强老师当时是我们的辅导员。该项目主要是对集成电路进行逆向工程设计，将集成电路的封装打开，利用显微照相技术分析光刻胶层上刻画的几何图形结构，推断连接关系，这个工作与视觉和图像处理非常相关。本科毕业之后来到华中科技大学（当时为华中工学院）。上世纪 80 年代中期教育部属的十几所重点大学引进 DEC 公司的图像处理系统 VAX-11 和 Model75，这套系统包括图像的输入、大型阵列处理机和图像输出等组成，是当时最先进的图像处理系统，80 年代到 90 年代华中科技大学的图像处理方向的研究生都是在这套系统上进行实验的。通过世界银行贷款，华中科技大学图像实验室花费 45 万美元引进这套系统。由于之前有一些图像处理的基础，我参与了这套系统的调试、后续的开发，撰写了详细的使用

手册和二次开发的接口等指导书，之后一直在图像实验室攻读研究生和从事与图像处理和计算机视觉相关的研究工作。

获得学士学位后，我来到华中科技大学电子信息与通信学院（原无线电系），在图像实验室继续攻读硕士和博士学位，师从朱光喜、朱耀庭教授，毕业后留校工作至今。在这个过程中，我发现科学研究是一件非常有趣的事情，也很喜欢这份工作。高校的环境非常适合我，校园内各种设施完备，体育运动设施十分充足，图书馆收藏的图书丰富多样，同时也有十分详尽的资料参考，可以做自己喜欢的事。教师的工作主要是面向学生，他们年轻有朝气，给人以积极向上的感觉，让人也感觉年轻一点。相比起来，我觉得去企业工作可能会比较单调和沉闷，我更喜欢在学校里做研究的环境。另外，我个人也是非常喜欢教学工作的，所以选择留在高校当老师，从事教学和研究这样的职业规划。

明：您带领的研究团队出了多名国家级人才，能谈谈您在这方面的经验吗？对于指导青年教师有什么建议？

刘：首先，研究团队对于青年教师的成长非常重要，作为团队负责人，首要任务就是构建和谐愉快的团队氛围和确立团队的研究方向。从博士研究生阶段开始，我就着重培养学生的研究品位和研究方向的把握，挖掘有价值的研究点。一方面，支持他们参加国际顶尖的学术会议，去到世界名校访问，和世界上最好的研究学者讨论交流，开阔眼界。另一方面，帮助他们树立远大的目标，聚焦国际科学研究前沿，从而为后续留校、继续从事相关的研究工作打下基础。

图 2 刘文予教授参加国际顶尖学术会议

青年教师留校以后，我比较强调他们研究方向的持续性，长期坚持聚焦一个研究方向，不盲目追逐研究热点。比如，我们在物体形状识别的方向上已经坚持了十几年的深入研究，逐步做到世界领先的水平。上世纪 90 年代，计算机视觉这个方向并不是很热门，甚至有点偏冷门。当时正值移动通信迅猛发展的阶段，学院的研究生大都选择 3G、4G（第三代、第四代移动通讯系统）这个热点研究方向。而我们团队坚持做计算机视觉和人工智能方面的研究，直到一二十年后，随着互联网的发展，这个方向才慢慢成为科技发展的热门。所以对于青年教师来说，一定要坚持一个领域方向的长期研究和积累。在这些青年人才成长过程中，团队始终对青年教师全心地扶持，从职业方向规划、治学态度、教学方法及项目申请方法等各个方面给予他们极大的帮助。从青年教师一进实验室，通过日常和他们的交谈中深入了解他们的性格、优势以及对未来职业生涯的想法。项目申请方面，从申请书的选题、内容、文字撰写和申请技巧等多个方面给予青年老师支持。从最初的国家青年基金到国家人才计划，每本申请书都逐字逐句地进行反复修改，并多次在团队内组织预答辩。

图 3 刘文予教授指导留学生博士答辩

在科研经费和研究生指标方面，我们尽可能支持青年教师的发展，让他们把有限的精力和工作的重点都放在科学研究工作上。青年教师刚毕业尽量不要去接横向课题或与自己研究方向关联度不高的课题。科学研究工作非常需要研究生的支持，但学院往往分配给青年教师

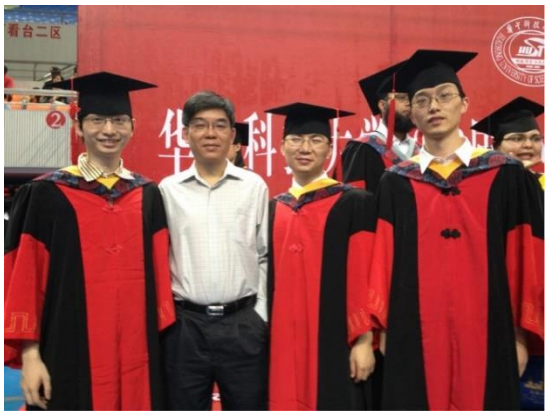


图 4 刘文予教授及其培养博士生

的研究生指标不多，尤其是博士研究生的名额，大多青年教师是没有的，这对他们的发展非常不利。我把我指导的博士研究生交给青年教师来指导，发表的论文也以青年教师作为通讯作者，以此帮助青年教师渡过前期的发展瓶颈期，积累成果，尽快度过讲师阶段。另一方面，也要给青年老师承担更多的重任，让他们尽快成为项目负责人，培养综合能力。

我们团队培养了很多国家级的人才，我也非常有幸获得华中科技大学的“伯乐奖”，这也是华中科技大学对教师的最高荣誉。

图 5 刘文予教授获得华中科技大学 2022 年教师节表彰

明：我们了解到，您在学术领域取得突破进展的同时，在教书育人方面也颇有建树，曾荣获多项教师荣誉奖。您是如何平衡科研与教学的时间精力分配的？有什么秘诀分享给青年教师吗？

刘：教学是高校教师的基本任务，我们大部分教师是科研教学并重岗，虽然承担一部分教学任务，但仍是以科学研究为主。我个人认为，不能把教学工作和科学研究当作独立的两个事情分开对待，教学工作应是我们学术研究的一个部分。我担任本科生“数据结构”和博士生“视觉计算”的课程责任教授，组织课程的建设。高校教师要对其专业领域的原理和概念非常清晰，并具有深刻的领会，才能出色地开展教学活动，而不仅仅是完成教学任务，这本身就是进行科学研究的相关工作。关于教学课件和教材，也应是自身学术研究方面的总结凝练，是逐步形成的。这样看来，教学工作本身就是学术研究的某种延伸，可以很好的平衡。

对于刚留校的青年教师，可以适当控制教学工作的投入，将主要精力放在科学研究方面，可以一年承担一门课，通过长期经验积累和时间投入来提升教学水平。随着个人的发展与成长，可以逐步增加在教学工作方面的投入，尤其到职业生涯后期退出科研一线工作后，再把主要精力放在教学工作上。我本人也是在接近退休的年龄时，开始把课程教学作为主要的工作。但对于青年老师，建议把主要精力放在科学研究上面，用少部分时间开展课堂教学活动。

图 6 刘文予教授回答学生提问

对于学生的培养，我们还会积极开展本科生创新创业活动，增强他们的学习兴趣，进行必要的科研训练。这样可以实施从本科阶段到研究生进行长期的培养方案。最主要的是培养学生从不同角度观察社会、观察问题，具有自己的独特观点。我们非常强调对每一个问题，需要讲清楚动机和解决的思路、主要的原理是什么、为什么性能会好。另一方面，我们也强调实验（编程）能力，通常刚进入实验室的学生按个人编程能力都会有一

个编程培训，大概 1 到 3 个月，尽快提升研究生的实验能力。2022 年我们指导的研究生获得全国互联网+大赛的金奖。

图 7 刘文予教授全国互联网+大赛团队

明：我注意到您是湖南人，您觉得湖南人有哪些适合做科研的品质?

刘：我父母都是湖南人，我也是在湖南出生的，上小学之后才转到湖北武汉这边来。我们一直说“惟楚有材，于斯为盛”，指的主要就是湖南、湖北这一带，湖南一直有这种重视教育的风气和传统。湖南人也具有“吃得苦，耐得烦，霸得蛮”的特点，这其实跟我们做的科学研究是非常契合的。科学研究首先要耐得住寂寞，能够抓住一个目标不放松；还要能够吃苦耐劳，这是我们科研工作者的一个基本特征。同时要有一股霸气，追求一个非常高的目标。所以湖南人在很多研究领域中都有很多突出的人才，也取得了很多突出的成果，各个领域都有很多专家来自于湖南。希望后续的湖南人里面能够发扬这种优良传统和特点，发扬湖南人吃苦耐劳的精神，取得更丰硕的成绩。

明：作为人工智能领域的专家，您如何看待目前的大模型发展状态，它对未来的计算机视觉会有何影响呢？未来计算机视觉还有哪些领域值得继续探索?

刘：当前，大模型的迅猛发展带来了人工智能内容生成（AIGC）领域涌现出大量的应用落地，例如擅长文本生成的 ChatGPT、擅长图像生成的 Stable Diffusion/Midjourney、擅长多模态对话的 GPT-4 等。大模型技术体现了人工智能的极大魅力，在精巧的算法、高质量的数据、高性能计算的加持下，之前的擅长单任务的深度学习算法已经进化成为通用人工智能。

对于计算机视觉领域而言，最近 Meta 公司发布的 Segment Anything 大模型也引起了很大的反响。在人为给定合适的提示下，Segment Anything 模型能够准确地分割几乎所有的图像。这些技术上的突飞猛进将为计算机视觉、自然语言处理等人工智能相关领域注入强大的动力，促进这些领域的算法更新和应用的落地。因此，未来的计算机视觉领域也应该拥抱大模型技术，并利用大模型在通用性强、泛化性强等方面的优势更好地处理计算机视觉问题。

图 8 刘文予教授开展学术讲座

明：在当前人工智能飞速发展的时期，很多学生和青年学者会感到有些迷茫，不知道未来计算机视觉领域还有哪些方向是值得探索的，刘老师对未来预测如何?

刘：现在许多青年学者对大模型的到来感到迷茫，因为大模型研究需要海量的标注数据和高额的计算开销，让青年学者感到举步维艰。随着人工智能技术和大规模数据、计算力的飞速发展，很多以前困难的问题，很容易解决了。我认为这是一个正常现象，应该看到大模型技术的发展带来的机遇和挑战。具体来说：（1）大模型是基础模型（Foundation Model），未来的计算机视觉研究可以基于 Segment Anything、CLIP 等基础大模型做出更好的应用效果。通过固定基础大模型的部分神经网络参数，不进行更新，实现低资源计算环境下大模型的研究与应用。（2）在现有的大模型基础上进行提示学

习（Prompt Learning）和指令微调（Instruction Fine-tuning），更好地完成更多的计算机视觉任务，也是未来的重要研究方向。（3）在多模态大模型中的计算机视觉技术中，如何将图像视频数据融入到多模态大模型也是一个十分困难、十分迫切解决的问题。（4）大模型的小型化，如何利用蒸馏、量化、轻量化架构设计等技术让大模型变得更小，给大模型提速。（5）大模型也并不是万能的，很多经典的计算机视觉技术，例如 3D 视觉、SLAM、底层视觉中的图像增强、图像去模糊等，大模型都没有办法解决。(6) 面向领域数据的个性化大模型，例如面向医学图像、工业视觉、遥感影像开发个性化的大模型，也会取得很好的推进计算机视觉的应用发展。（7）视觉的基础问题和挑战，如视觉的感知过程等。

总的来讲，只要静下心来仔细地去分析，可以做的事情还有很多，不用过分忧虑。

明：您如何看待现在从考研、读博、到青年教师成长都比较“卷”的这种状况？当代青年如何才能发挥个人优势，取得丰硕的科研成果？

刘：您说的内卷这个事情，是当前社会各界都存在的问题，并不仅仅局限于某个领域，只是在教育行业十分突出的一个现象。高校的内卷可能从本科就开始，包括考试成绩、就业、留学、考研、读博等等，究其根本，是因为好的高等教育的资源有限，在整个社会环境压力非常大的背景下，大家普遍感到比较焦虑。我认为这种现象属于一种正常的良性竞争，在面对这些所谓比较卷的竞争时，要及早地树立目标，做好长期的规划，更有利于个人的发展成长。

对于想要做好科研的青年来说，需要耐得住寂寞，需要有长期的积累，不可以盲目的内卷形成无用的内耗，必须要抓住核心问题，尽量做一些基础性的研究，或者解决国家急需解决的难题。这些问题都是需要长时间的付出，短时间内可能无法取得成果，但只要有一个好的团队，加以长时间的努力与钻研，后续的工作就希望。总的来说，不要焦虑，不盲目跟风，结合团队的方向和个人的优势，制定好自己的职业规划，避免受到外界的干扰，保持定力，就能取得丰硕的科研成果。

责任编委 明悦 张军平 贾熹滨

刘文予

华中科技大学电子信息与通信学院人工智能研究所所长、教授，中国通信学会会士，中国图象图形学学会视频通信专委会主任。获湖北省技术发明一等奖、中国图象图形学学会自然科学一等奖、“计算可视媒体”期刊（SCI Q1）2021 最佳论文奖。在国际著名期刊和会议发表论文 160 多篇，谷歌学术总引用次数超 2.3 万次。指导的博士生获全国百篇优秀博士论文提名奖，8 人获湖北省优秀博士论文奖，1 人获 2019 年 ACM SIGAI、2 人获中国图象图形学会优秀博士论文奖。

COMPUTER VISION NEWSLETTER

02 2023

总第 36 期

计算机视觉专委会简报

CCF 计算机视觉专委会