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

04 2021

总第 30 期


CCF 计算机视觉专委会

COMPUTER VISION NEWSLETTER

计算机视觉专委会简报

2021 年第 04 期

总第 30 期

**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 105 期 北京工商大学

2021 年 9 月 27 日下午，中国计算机学会计算机视觉专委会 (CCF-CV) 走进高校系列报告在北京工商大学阜成路校区综合楼一层报告厅顺利举行。本次报告会邀请到的讲者和专家有北京大学查红彬教授、北京交通大学赵耀教授、厦门大学纪荣嵘教授、重庆大学张磊教授、中科院自动化研究所王亮研究员。北京工商大学副校长徐丹丹教授出席报告会并致辞，报告会采取线下+线上结合形式，由活动执行主席北京工商大学计算机学院执行院长李海生教授主持，计算机学院全体师生和相关领域部分科研人员参加了本次报告会。

报告会进行过程中，与会师生积极向专家学者提问，现场展开了积极热烈的讨论交流。同学们表示通过这次报告会收获了新知，了解了新的研究角度、研究方法和研究内容，希望能够在今后有更多的机会与学术前沿的教授进行这样的学习和交流，报告会取得圆满成功。

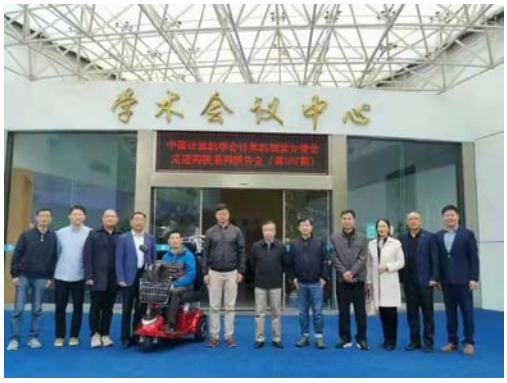
## 第 106 期 山东师范大学


2021 年 9 月 28 日，由中国计算机学会计算机视觉专委会 (CCF-CV) 主办、山东师范大学承办的第 106 期走进高校系列报告会在山东师范大学成功举办。本次活动邀请了上海交通大学林巍峣教授、华中科技大学白翔教授以及江南大学吴小俊教授三位专家学者莅临现场做特邀报告。山东师范大学信息科学与工程学院张化祥教授、郑元杰教授和朱磊教授担任本次会议的执行主席。希望借此机会能够和各位专家及同行加强学术交流与合作，扩展师生科研视野，促进学院进步发展。

报告结束后，张维军书记对三位报告专家以及 CCF-CV 专委会表示衷心的感谢，并欢迎各位专家学者继续来山东师范大学进行学术指导和交流。最后，张书记对此次学术报告会进行总结发言，对信工学院师生提出新的期望，本次活动取得了圆满成功。

# CCF-CV 走进高校系列报告会

## 第 107 期 江苏海洋大学




2021 年 10 月 16 日，由中国计算机学会计算机视觉专委会（CCF-CV）主办、江苏海洋大学计算机工程学院承办的第 107 期走进高校系列报告会在江苏海洋大学苍梧校区学术会议中心成功举行。本次活动邀请了中国海洋大学董军宇教授、深圳大学赖志辉教授、同济大学赵才荣教授、中国科学院心理研究所王甦菁副研究员四位专家莅临现场做特邀报告。会议由江苏海洋大学计算机工程学院副院长仲兆满教授担任执行主席。专家们围绕“计算机视觉前沿技术及应用”主题做精彩报告。

此次 CCF-CV 走进高校系列报告会专家们的讲解深入浅出，从实际出发分享自己学术上的经验，报告内容十分精彩，让所有参加会人员享受了一场学术盛宴。国家自然科学基金项目申请指导会专家们的指导切中要害，为切实提升青年教师的国家自然科学基金项目申请能力提供了帮助。希望能够通过 CCF-CV 走进高校系列报告会，促进计算机视觉学科的专业内容更好地开展交流、发展战略研究，促进国内学者间的了解与合作，推动国内计算机视觉学科发展，提升我国计算机视觉研究在国际领域的影响力。

## 第 108 期 北京科技大学


2021 年 10 月 24 日下午，中国计算机学会计算机视觉专委会（CCF-CV）走进高校系列报告会第 108 期活动“计算机视觉前沿技术及应用”通过线上直播的形式在北京科技大学成功举行。本期报告会由北京科技大学自动化学院“机器视觉与智能感知”梯队承办，邀请了清华大学季向阳教授，西北工业大学韩军伟教授，中科院自动化所王亮研究员三位专家做特邀报告，由北京科技大学自动化学院“机器视觉与智能感知”梯队樊彬教授和刘红敏教授担任本次报告会的执行主席。在本次报告会上，专家们围绕“计算机视觉前沿技术及应用”做了精彩报告，并在问答环节就计算机视觉领域的多个前沿学术问题、热点应用问题进行了深入的探讨。

报告会最后，主持人樊彬教授对报告会进行总结发言。他首先对进行报告的各位嘉宾以及 CCF-CV 专委会表示衷心的感谢，并希望以本次报告会为契机，欢迎全国计算机视觉领域的同行们来北京科技大学交流指导，拉近在校学生和专家学者之间的距离，为计算机视觉的发展持续地注入新鲜活力。

# CCF-CV 走进高校系列报告会

## 第 109 期 北京工业大学

2021 年 11 月 27 日下午，由中国计算机学会计算机视觉专委会（CCF-CV）主办、北京工业大学承办的 CCF-CV 走进高校系列报告会第 109 期活动在线上成功举办。本次活动邀请了中国科学院自动化所研究院胡卫明研究员、浙江大学潘纲教授、重庆邮电大学李伟生教授、复旦大学姜育刚教授以及北京师范大学邬霞教授五位专家学者做特邀报告。北京工业大学信息学部毋立芳教授、施云惠教授和简萌副教授担任本次活动的执行主席。北京工业大学信息学部副主任杨震教授致欢迎辞，他首先代表信息学部对五位报告嘉宾表示热烈的欢迎和衷心的感谢，也感谢 CCF-CV 专委会对北工大的信任，希望这类高水平、高创新、高收获的会议能够多在北京工业大学举办，相信这次报告会能够促进我校师生和各位专家同行的学术交流与合作，拓展师生的科研视野，促进信息学部的进步发展，并预祝本次报告会取得圆满成功！

报告结束后，毋立芳教授进行总结，首先感谢五位报告专家的精彩报告，其次感谢线上听众的热情参与和高质量的提问，最后再次感谢北京工业大学信息学部和 CCF-CV 专委会对活动的大力支持！并欢迎各位专家学者经常来北京工业大学进行学术指导和交流！本次活动取得了圆满成功！

## 第 110 期 北京化工大学

2021 年 11 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，北京化工大学承办的 CCF-CV 走进高校系列报告会第 110 期活动，通过线上直播的方式成功举行。本期活动邀请到中国科学院自动化所徐常胜研究员、西安交通大学薛建儒教授、复旦大学张军平教授共 3 位专家学者做报告。北京化工大学信息学院的王坤峰教授和汪凌峰教授担任本次报告会的执行主席。报告会开始由北京化工大学信息学院王友清院长致欢迎词。王院长首先对三位特邀专家表示热烈的欢迎。然后简要介绍了北京化工大学信息学院在学科建设和国家一流本科专业建设方面的成绩，以及在计算机视觉方向的发展状况，希望未来进一步加强学院与计算机视觉专家们的交流合作，并祝愿本次报告会成功举办。

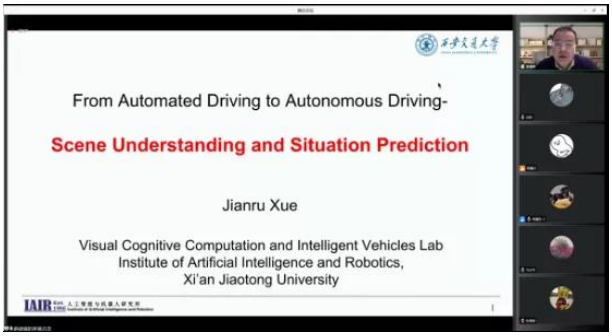
报告会最后，执行主席王坤峰教授进行总结发言，对三位特邀专家走进北京化工大学表示衷心感谢，并欢迎校外专家和听众在条件允许时亲自走进北化校园进行交流。活动全程通过哔哩哔哩平台进行直播，在线听众人数超过 1700 人，听众对专家们的报告积极提问。活动受到广泛好评，取得圆满成功。

责任编委 毋立芳

第 11 期 智能系统环境感知的前沿进展与未来趋势

# CCF-CV 视界无限系列研讨会


2021 年 12 月 12 日，由中国计算机学会计算机视觉专委会主办的第 11 期 CCF-CV “视界无限”系列活动“智能系统环境感知的前沿进展与未来趋势”研讨会在线上成功举办，天津大学朱鹏飞副教授、张长青副教授担任执行主席。研讨会邀请了天津大学胡清华教授、专委会秘书长北京邮电大学马占宇教授致辞，西安交通大学薛建儒教授、西北工业大学韩军伟教授、武汉大学夏桂松教授、大连理工大学王栋教授、一飞智控 (天津) 科技有限公司研发总监吴冲博士做主题报告。天津大学胡清华教授等及以上五位讲者参与了深度研讨。计算机视觉专委会 B 站公众号对本次会议进行了全程直播，直播人气峰值达到 3100+。




薛建儒教授的报告题目是“交通场景的理解与预测”。报告分为三个方面，首先，介绍了自动驾驶现阶段的发展情况以及在场景理解和情景预测等基础的问题；其次，是在视觉导航里视觉理解、场景理解所面临主要的难点问题；最后针对用于驾驶预测的交通情景的预测展开介绍。最后，薛建儒教授指出，自动驾驶正在不断落地应用。然而，实现真正的自主驾驶依然面临着诸多极具挑战性问题。本报告主要探讨自主驾驶的动态场景理解与预测问题，并报告课题组近年来所取得的一些研究进展。


韩军伟教授的报告题目是“AI 赋能高分遥感目标检测与识别：挑战、对策及展望”。韩军伟教授指出，高分辨率遥感图像目标检测与识别是空天地海一体化观测系统的一项关键技术，广泛应用在侦察、监视、预警、搜救等军民领域。与自然图像相比，高分辨率遥感图像具有目标方向多变、目标类型及数量繁杂、特定领域样本稀缺 (如导弹阵地)、成像视角单一等特点，此外，不同平台、不同光照、天气条件、大气参数等都会对遥感图像获取产生影响。这些综合因素使得高分辨率遥感图像目标检测与识别，与自然图像理解相比，面临着更大的挑战和更多的难点问题。本报告首先总结分析高分辨率遥感图像目标检测与识别面临的挑战，重点汇报韩

教授团队在旋转不变目标检测、有向目标检测、弱监督目标检测、小样本目标检测以及目标型号识别等方向取得的研究进展和典型应用，最后展望了未来的研究工作。


夏桂松教授的报告题目是“图像几何结构矢量化感知及应用”。夏桂松教授指出几何结构是诸多低、中层视觉计算任务的重要特征，在视觉场景结构理解和环境感知等方面有重要应用。本报告分享夏教授团队在自然图像几何结构（如 wireframe、line segment、junction 等）建模和计算问题上的新进展，包括图像 junction、line segment 的非监督计算模型和有监督深度感知方法、自然图像一维线结构与二维区域之间的对偶计算框架、wireframe 的深度计算模型等，及其在室内图像三维重建、高分遥感图像解译等方面的应用。


王栋教授的报告题目是“高性能视觉跟踪算法探索”。视觉跟踪算法近年已取得突破性进展，但复杂现实环境中的各种场景变化和平台限制对跟踪算法提出了以“强鲁棒性、高精准度、易迁移性、低计算量”为导向的高性能要求。王栋教授团队近年来从针对鲁棒外观模型更新控制、模板和搜索区域特征融合、通用尺度估计模块设计及嵌入式平台部署需求这四方面开展研究，分别提出了元更新器模型、Transformer 跟踪模型、Alpha-Refine 尺度估计模块和 NAS 跟踪网络搜索模型，显著提升跟踪算法精度的同时降低了参数量和计算量。


吴冲博士的报告题目是“环境感知技术在行业无人机中的应用”。吴冲博士指出，随着无人机行业应用的快速发展，环境感知技术对推动无人机应用的深入发展发挥着越来越重要的作用，报告中以不同行业无人机对环境感知技术的需求为牵引，从无人机对地安全探测、无人机紧急降落安全区域评估、多无人机协同定位、无人机拒止环境定位等多个方面介绍环境感知技术在无人机中的应用场景和应用方法。

在 Panel 环节，与会嘉宾就“智能系统的环境感知中复杂环境和开放环境应该怎么定义？”、“自主智能系统对于小目标或突发情况造成的感知困难怎么解决？”、“如何把环境感知与决策控制等相关学科做交叉研究？”、“智能感知系统中的大规模预训练模型”等问题展开热烈讨论，各位专家就上述问题分享各自的观点。最后，第 11 期“视界无限”研讨会在中午 12 点 20 分圆满结束。

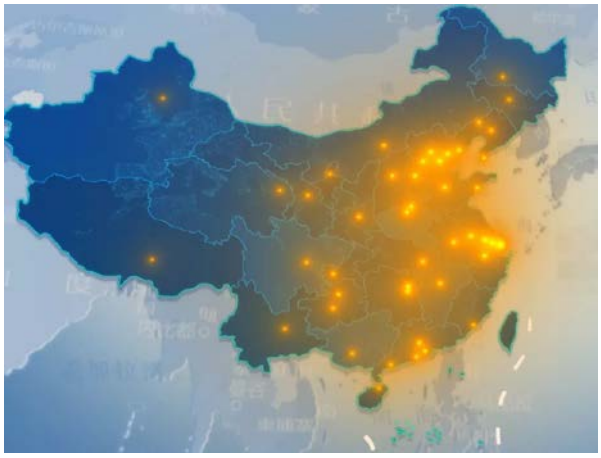
责任编委 杨巨峰

# CCF-CV 专委宣传视频和走进高校活动 100 期纪念视频发布

中国计算机学会计算机视觉专委会（CCF-CV)专委宣传视频和 CCF-CV 走进高校 100 期活动总结视频已经在专委 B 站账号发布，直接拷贝下面视频链接到浏览器即可观看，欢迎关注!

CCF-CV 专委宣传视频链接：

https://www.bilibili.com/video/BV1A64y187Hj



CCF-CV 走进高校 100 期活动纪念视频链接：

https://www.bilibili.com/video/BV1nq4y1X7DV

## CCF-CV 走进高校活动背景

自 2015 年 11 月起，CCF 计算机视觉专委会(CCF-CV)在全国范围内率先开展走进高校系列报告会、走进企业系列交流会等特色活动，在学术界、工业界产生了热烈反响，受众遍及祖国大江南北。CCF 计算机视觉专委会欢迎各兄弟学会、专委会借鉴，共同推动我国相关领域的学术繁荣和产业发展!。

## CCF-CV 走进高校活动申请

如您想了解活动申请相关信息，请看活动申请链接。如您有意申请 CCF-CV 活动，请与专委会秘书处联系。联系方式：

毋立芳：lfwu@bjut.edu.cn

杨巨峰：yangjufeng@nankai.edu.cn

责任编委 黄岩

# CCF-CV 专委会 2021 年执行委员增选申请

自 2013 年 10 月成立以来，中国计算机学会（CCF）计算机视觉专业委员会（ccfcv.ccf.org.cn）发展迅速，举办了很多有影响力的活动，如计算机视觉前沿进展研讨会（RACV）、CCF-CV 走进高校系列报告会、CCF-CV 走进企业系列交流会、CCF-CV 视界无限系列研讨会，与中国自动化学会模式识别与机器智能专委会、中国图象图形学学会视觉大数据专委会、中国人工智能学会模式识别专委会共同举办中国模式识别与计算机视觉大会（PRCV），定期出版专委简报，建设专委中英文网站，专委微信公众号文章平均阅读上千次，专委活动视频在专委 Bilibili 账号发布。搭建了全方位、高水平、大规模的计算机视觉领域交流平台。专委会成立八年以来，已经发展执行委员 354 人，在 CCF 专委评估中获得“特色活动奖”、“综合进步奖”、“优秀专委奖”、“年度特别奖”等 6 个奖项。为了保持专委会的活力、促进国内外视觉领域人员的交流和合作，专委会现开放 2021 年计算机视觉专委会的执行委员增选工作。

## 申请时间：

2021 年 6 月 15 日— 2021 年 10 月 15 日

## 申请流程：

填写申请表，发送给秘书处（ccfcv@139.com），主题“2021 新执行委员申请-姓名-单位”。（注：推荐人必须是现任专委执行委员，名单可以从专委网站查询。电子版申请表中需填写推荐人姓名和意见，执行委员增选成功后可以补签签名）。

## 申请资格：

任职国内外学术界或企业界副教授或等同级别以上的人员，拥有计算机视觉相关领域的高水平研究成果，是 CCF 计算机视觉专委委员，且积极参加计算机学会计算机视觉专委会的各项活动。特别优秀的讲师、企业人士亦可考虑。

## 申请需知：

现任专委执行委员每人可推荐最多 3 名候选人。本次申请结果将在 “2021 年中国模式识别与计算机视觉大会” （http://www.prcv.cn）期间举行的专委工作年会上（时间：2021 年 10 月 29 日 15:40-17:00，地点：北京国际会议中心第二会议厅（B+C））投票确定（申请者届时必须“注册参会”）。

## 特别说明：

按照 CCF 的新规定，CCF 专业会员通过 CCF 会员系统关注相关专委后即加入专委并成为其委员，其后每年可以更改一次关注的专委。委员在专委中无选举权和被选举权，但具有对专委的评价权。原来的专委委员自动升级为专委执行委员，享有选举权、被选举权以及对专委的评价权。

责任编委 毋立芳

# RACV2021 计算机视觉前沿进展研讨会圆满召开


2021 年 10 月 16 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）在湖北恩施圆满召开。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。


本次会议开幕式由华中科技大学白翔教授主持，湖北民族大学副校长、祝建波教授和专委会主任、北京大学查红彬教授进行开幕式致辞。根据常委委员前期的讨论票选，本次会议设置了 3 项研讨主题。每项主题首先由特邀嘉宾们进行主题发言，之后所有与会人员进行自由讨论。

上午首先进行了主题一“大规模多模态预训练模型：现状与趋势”的研讨。该主题由专委会副主任、中科院自动化所王亮研究员、中科院计算所山世光研究员、中科院计算所杨双副研究员 3 位委员负责组织，邀请中国人民大学卢志武教授、中科院自动化所刘静研究员、华为常建龙博士、阿里巴巴周畅博士 4 位嘉宾进行主题发言。近几年多模态预训练模型吸引大量学者的注意，但是伴随而来的高计算复杂度和大规模训练数据需求却令人望而却步，我们应该如何应对这种研究趋势呢？几位嘉宾围绕数据收集与清洗、模型知识学习与记忆、工业界应用等方面的未来发展趋势进行精彩的观点分享。

下午首先进行了主题二“视觉感知算法怎么适应开放环境”的研讨。该主题由南开大学程明明教授和杨巨峰教授、华中科技大学王兴刚教授 3 位委员负责组织，邀请了中科院自动化所刘成林研究员、南开大学程明明教授、上海交通大学林巍峣教授、清华大学黄高助理教授 4 位嘉宾进行主题发言。近年来，很多视觉算法被用到了实际需求中，大大便捷了人们的日常生活。但是在更多开放环境下，视觉算法的精度和鲁棒性下降明显，如何缓解这个问题并推进算法实用化发展？嘉宾们围绕开放环境设定、模型学习与评测、先验知识使用等议题展开了深入探讨。

下午还进行了主题三“视觉 transformer 从主干 encoder 到任务 decoder：现状与趋势”的研讨。该主题由微软亚洲研究院王井东博士、大连理工大学卢湖川教授、北京邮电大学马占宇教授、北京大学刘洋助理教授 4 位委员负责组织，邀请了复旦大学邱锡鹏教授、旷视研究院张祥雨博士、微软亚洲研究院胡瀚博士、华中科技大学王兴刚教授 4 位嘉宾进行主题发言。视觉 transformer 的兴起为视觉领域的发展带来了蓬勃生机，其发展现状、核心挑战、关键应用和未来趋势有哪些？嘉宾们围绕网络结构调试、transformer 与 CNN 比较、生物机制联系等议题展开了深入探讨。

最后，研讨会闭幕式由专委会副主任、南京信息工程大学刘青山教授主持。受疫情影响，本次研讨会延期召开可谓一波三折，但是本次研讨会在短短一天内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。本次研讨会由华中科技大学白翔教授主要负责组织，湖北民族大学相关老师和专委秘书处成员协助会务组织。

责任编委 杨巨峰

# 2021 年度 CCF-CV 专委工作会议顺利举办



中国计算机学会计算机视觉专委会（CCF-CV）年度工作会议于 2021 年 12 月 19 日在珠海海泉湾维景国际大酒店顺利举办。

大会以线上和线下结合的方式举办，来自全国高校、科研院所、企业的现任委员和新申请委员共计 320 多位参加了工作会议。会议由专委会秘书长、北京邮电大学马占宇教授主持。

会议首先由专委会主任、北京大学查红彬教授致辞。查主任感谢从祖国各地来到现场参会的新老委员，感谢因疫情影响而在线参加会议的委员，他积极肯定了专委会在过去一年中取得的丰硕成果，鼓励委员们聚焦学术前沿、做出高质量研究工作，希望专委会大家庭一起精诚合作，各项活动进一步提升品牌质量，打造精品学术交流活动，为推进计算机视觉学术研究与产业发展继续发挥积极的引领作用。

随后，中国计算机学会（CCF）常务理事、浙江大学卜佳俊教授和 CCF 专委工委、中科院计算所蒋树强研究员代表学会致辞，对专委工作会议的召开表示祝贺，对专委开展的各项活动给予了积极评价，特别肯定了专委建立的一系列学术交流品牌活动。

接下来，专委会秘书长马占宇教授向与会委员做了年度工作报告。报告简要介绍了专委的组织结构、专委顾问委员会和国际顾问委员会，通报了 4 月份专委年度常委会议和 2 月和 7 月秘书处工作会议所作出的若干新举措，总结了过去一年专委工作的重要成就，以及专委委员获得的各项奖励和荣誉，全面回顾了专委过去一年的学术交流活动，指出了工作中存在的问题和改进方案，最后介绍了下一年度专委工作计划。

根据日程，工作会议还进行了新增执行委员的选举，由专委副主任、中科院自动化研究所王亮研究员主持。本年度共收到 70 位候选执行委员申请，由现任常务委员通过投票表决。按照 CCF 学会规定，新当选委员数不超过现任执行委员人数的 10%，因此最终共有 35 位候选人当选。


接下来，进入 CCF-CV 颁奖环节。颁奖仪式由专委会副主任、提名与奖励工作组组长、南京信息工程大学刘青山教授主持。刘青山教授详细介绍了本年度设立的奖项评选范围与评选规则，包括：终身学术贡献奖、杰出成就奖、服务贡献奖、学术新锐奖和持久影响力论文奖。本年度的终身学术贡献奖获奖者为上海交通大学施鹏飞教授，杰出成就奖获奖者为旷视首席科学家、旷视研究院院长孙剑博士，持久影响力论文奖获奖者为南开大学程明明教授作为第一作者发表于 CVPR 2011 的论文 Global Contrast based Salient Region Detection，服务贡献奖获奖者为中科院自动化所黄岩副研究员、哈尔滨工程大学刘海波教授、中山大学任传贤副教授、燕山大学王金甲教授、北京科技大学殷绪成教授、华北电力大学赵振兵副教授，学术新锐奖获奖者为大连理工大学代克楠、清华大学王语霖、上海交通大学汪润中。五个奖项的获奖人既有早年投入我国计算机视觉奠基性研究的老一代科学家，又有新一代崭露头角的学术新星，也有为专委发展尽心尽力、无私服务的中生代科研工作



者，充分展示了计算机视觉专委会大家庭的繁荣兴盛。

委员建言献策环节由专委副主任、上海科技大学虞晶怡教授主持。虞晶怡教授介绍了 CVPR 2021 采用基于 AI 的领域选择机制、将由投稿者担任审稿工作，欢迎更多专委委员担任 CVPR、ICCV 等国际顶级会议的领域主席和审稿专家，提升中国计算机视觉研究者在国际学术界的影响力。中山大学赖剑煌教授和林宙辰教授建议更大地放开新执行委员当选名额，吸收优秀年轻学者壮大计算机视觉队伍。华北电力大学赵振兵副教授提出希望更多关注计算机视觉在电力产业方面的应用研究。厦门大学杨晨晖教授提出增加科普教育，让计算机视觉让中小学生也能亲近和理解。针对北京邮电大学明月副教授关于计算机视觉教材编撰方面的建议，虞晶怡教授指出专委为此成立了教育工作小组，已开展计算机视觉教材编辑工作，不久将交付出版。


专委主任查红彬教授祝贺施鹏飞教授、孙剑博士等获奖者，欢迎更多有建树的年轻学者加入专委会，再次感谢在线和现场参加本次专委工作会议的专委委员，并祝愿大家新年快乐、成果丰硕。最后，专委会 2021 年度工作会议在热烈的掌声中圆满结束，期待明年再聚！

责任编委 黄岩

专题综述

# Transformer 的视频背景音乐生成

北京航空航天大学 狄尚哲 姜泽仁 王肇凯 朱乐岩 何泽欣 刘偲

随着新媒体技术与相关产业的发展，人们对短视频的编辑和发布变得越来越便捷。通常，为了让视频更吸引人，视频制作者会为视频搭配背景音乐。然而，对于不具备音乐制作、视频剪辑等技能的人而言，这是一个困难工作。除此之外，这一过程还面临着版权等许多问题。

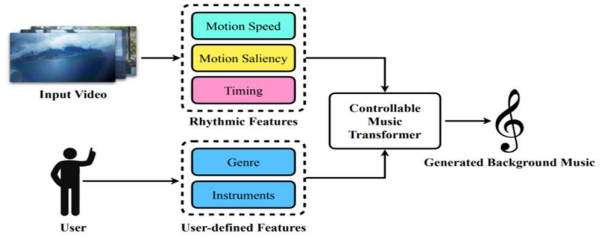



图 1 方法整体示意图

因此根据给定视频生成背景音乐成为一个十分重要的任务。然而，目前尚未有针对这一任务的有效研究。虽然在音乐生成领域已经有了一些研究工作[19][2][7]，但这些工作均未考虑视频信息。为了解决这个问题，我们提出一种新的音乐表示形式，并基于这种表示形式设计了音乐生成模型，实现了视频背景音乐生成的功能。

虽然先前尚未有与一般视频的背景音乐生成相关的工作，但已经有与音乐表示学习、音乐生成以及从静音演奏视频复原音乐的相关研究。大多数音乐生成研究工作是以类似 MIDI 的事件序列[7][12]作为输入。REMI[8]提出了一种表示音乐的结构，这种结构清晰地标注了小节、节拍、和弦、音高等信息。这种新的表示形式有助于维护音高局部变化的灵活性，提供了一种可以人为控制的节奏与和声结构。 Compound words[6] 将 REMI 的标记转换为一系列的复合词，大大缩短了序列的长度。

先前关于视频音乐生成的任务主要针对于乐器演奏的视频，例如小提琴、钢琴和吉他[4][15][16]。由于大部分生成结果，例如乐器类型、节奏、音调等都可以从人的手部动作判断，因此生成的音乐也基本是固定的，无法适应一般的视频背景音乐生成任务。

## 一、音乐与视频的关系

视听关系在心理学与认知科学等领域已经有了几个世纪的研究，特别是音乐与视频，它们在许多方面存在着联系。例如，一个人会希望在看浪漫电影时听到抒情的音乐，或者在观看战斗场景时听到激昂的音乐。

为了更好地使生成的背景音乐匹配视频，我们分析并建立了若干音乐-视频关系。首先建立音视频之间的时间与节奏对应关系；基于此，我们进一步建立视频光流强度与音符密度的联系；最后我们将视频的运动显著性与音乐的音符强度进行对应。以上三种关系将用于指导如何从视频中提取信息监督音乐的生成。

### 1.1. 视频帧与音乐节拍

理想情况下，生成的背景音乐随着视频的开始与结束应当平滑地出现和减弱。我们考虑将需要生成的音乐分成固定数量的片段(音乐节拍)，并给每一个片段设置位置相关的独特编码，使得模型能够学习到相关的开始和结束等位置信息。通过这种设计，我们能方便地对生成的音乐长度进行控制，使得与相应的视频匹配。

### 1.2. 动作速度和 Simu-note 密度

我们发现视频的运动速度与音乐的 Simu-note 密度间存在正相关的关系，即快速的画面应当对应激烈的背景音乐，而缓慢的画面应当对应舒缓的背景音乐。如图 2 所示，音乐的 Simu-note 密度定义为一个小节内

Simu-note 的数量。

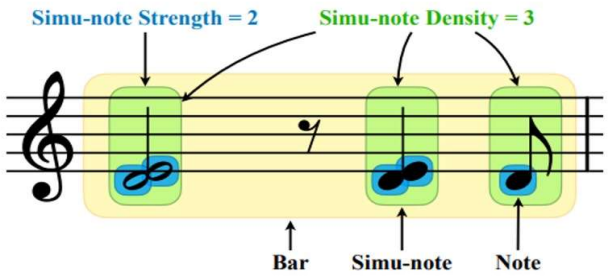



图 2 Simu-note 强度与密度示意图

1.3. 动作显著性和 Simu-note 强度

视觉节拍可以类比为音乐节拍在视觉上的表示形式，其代表视觉动作在时间上的分布。当画面中出现比较明显的转场时，视觉节拍强度应当是一个比较大的值；相反，当画面没有比较大的变化时，视觉节拍强度应当是一个比较小的值。我们对 Simu-note 的强度与视觉节拍强度建立起了一个正相关的对应关系，通过这种方式，视频中的转场点在生成出的音乐中将表现为强拍。如图2所示，Simu-note 强度定义为单个 Simu-note 发生时刻音符的个数。

## 二、可控音乐生成模型

在上述建立的视频-音乐节奏关系的基础上，我们提出了一种基于 Transformer 的方法以生成视频背景音乐，称为 Controllable Music Transformer (CMT)。整体架构如图 3 所示。我们从视频和 MIDI 文件中分别提取节奏特征(参见前一部分)。在训练时，我们只使用了音乐数据集及其中的节奏特征。在生成时，我们将节奏特征替换为视频中提取出的节奏特征，用来对音乐的生成过程进行控制。

2.1. 音乐表示形式

我们为可控的多轨音乐生成模型设计了一种结构化的表示形式。启发于 PopMAG[13] 与 CWT[6]，我们将一组相互关联的属性合并为同一个 token，来缩短序列的长度。如图 3 所示，每个 token 共有 7 种属性：类别、乐器、时值、音高、强度、密度、拍数。上述属性被分为两组：其一是节奏相关的属性(图 3 中标识为 R)，包括强度、密度和拍数；另一组是音符本身的属性(图 3 中标识为 N)，包括乐器、时值和音高。我们使用类别属性(图 3 中 R/N 的一行)来区分这两组属性。

我们将节奏相关 token 中的乐器、时值和音高三个属性的值设为 None，将音符相关 token 中的强度、密度和拍数三个属性设为 None。图 3 中的空位即对应于该位置的值为 None 的情况。每个节奏相关 token 包含强度属性，表示后续音符 token 的数量。此外，密度属性在每个小节内单调递减，表示该小节内剩余 simu-note 的数量。每个 token 的不同属性分别进行 embedding 并连接到一起，作为 token 的 embedding。此外，我们提取流派和乐器种类作为每段音乐的初始 token，对其使用独立的 embedding 层。

2.2. 对生成过程的控制

在训练结束后，CMT 已经理解了强度和密度两个属性的含义。由此，我们在生成过程中只需将这两个属性替换为我们需要的值，从而生成出与指定视频更和谐匹配的音乐。

2.2.1. 密度属性替换

为了使生成的音乐各处音符密度与视频的光流强度相符，我们将每个小节的密度属性替换为从视频中提取出的光流强度信息，按照特定比例位数进行替换。由于 CMT 已经理解了小节 token 上密度属性的含义，模型将在这个小节中自动生成出对应数量的音符，从而对音符的密度进行控制。

2.2.2. 强度属性替换

类似地，我们利用视频的视觉节拍信息来控制每个 simu note 的强度。如果 CMT 在一个视觉节拍处生成出了一个 simu note，这个 simu note 的强度会被替换为这个视觉节拍的强度(根据比例位数)。然后 CMT 会在这一拍中预测出指定数量的音符，从而控制音符强度。

2.2.3. 调节控制程度

我们使用了一个超参数 C 来调节对模型生成音乐过程的控制程度。在这个过程中，我们需要权衡两个因素。其一是视频与旋律的匹配程度，其二是音乐的质量。也就是说，在生成的过程中，加入的限制条件越多，生成的音乐听起来也越不自然。为了解决这个问题，我们设计了一个超参数 C，用来表示对生成过程的控制程度。C 的值越大，生成过程中加入的限制条件也就越多。当 C 为 0 时，我们生成音乐的过程是完全不受控制的；当

C 为 1 时，我们将得到与视频节奏完全匹配的音乐。 C 的值可以由用户根据需求来指定。

2.2.4. 节拍时间编码

为了利用视频的时间(长度)特征，我们在训练和生成过程的 embedding 中加入了节拍时间编码。也就是说，它指导 CMT 在合适的时间开始和结束生成过程。节拍时间编码的 embedding 表示了当前节拍在整个视频中的位置比例。我们将该比例等分为 M 个区间，并使用一个可学习的 embedding 层来将其映射到和其他 token 相同的维度，并将其一同作为 CMT 的输入。

2.2.5. 流派和乐器种类

我们的方法中包含六种音乐流派(乡村，舞蹈，电子，金属，流行和摇滚)以及五种乐器(鼓，钢琴，吉他，贝斯和弦乐)，将它们各作为 CMT 模型的初始 token。用户可以改变不同的初始 token 以选择不同的流派和乐器，从而使得音乐与视频的情感相互对应。

2.3. 序列建模

音乐 token 序列(如 2.1 小节所述)被输入进 Transformer[17]模型以建模元素之间的依赖关系。我们使用 Linear Transformer[9] 作为模型主干结构，考虑到其轻量级的架构和注意力机制的线性复杂度等优势。

Multi-head 输出模块(依照[6]的设计)，按照两阶段的方式预测每个 token 的 7 种属性。第一阶段中，模型将 transformer 的输出进行线性投影，预测出类别属性；第二阶段中，使用类别属性通过六个前馈 head 同时预测剩余的 6 种属性。

在生成阶段，上述提到的控制策略被结合在一起。我们使用了随机温度控制采样策略[5]以提升生成序列的多样性。

## 三、实验评估

我们针对提出的音乐生成模型进行了一系列消融实验。在实验中，我们兼用了客观性评价指标与主观性评价指标。 我们使用 Lakh Pianoroll Dataset (LPD) 来训练 CMT 模型。LPD 是从 Lakh MIDI Dataset (LMD) 提取出的 174154 首多轨音乐的集合。我们使用的是 LPD-5-Cleansed 版本的 LPD 数据集，这个版本是由 LMD 经过一系列数据清洗，并将所有音轨合并为鼓、钢琴、吉他、贝斯和弦乐五个音轨后得到的。

3.1. 客观评价性实验

我们使用了 3 个统计性的评价指标来客观地评价生成出的音乐。 Pitch Histogram Entropy，评估音乐的音调质量； Grooving Pattern Similarity，衡量音乐的节奏，和 Structureness Indicator，衡量音乐的结构重复性。如表 1 所示，我们使用消融实验来评价提出的可控性属性在音乐生成过程中发挥的作用。这里列出的所有指标越接近数据集 (Data) 中的指标代表效果越好。在进行客观指标评估时我们不对音乐施加控制，只增加 3 种可控属性。从结果上看，增加了一些额外的属性后，对音乐生成的节奏和结构性帮助较大。

表 1 客观指标实验结果

| Model | Data | Baseline | Ours |
|---|---|---|---|
| Pitch Histogram Entorpy | 4.452 | 3.634 | 3.617 |
| Grooving Pattern Similarity | 0.968 | 0.677 | 0.810 |
| Structure Indicator | 0.488 | 0.219 | 0.241 |

3.2. 主观评价性实验

对于主观性的用户评价方法，我们设计了一个调查问卷，并邀请了 36 人来对我们提出的可控性指标进行评价。我们主要从两个方面设计主观性评价指标来评价生成出的音乐，首先是音乐本身的音乐性，这里我们主要考虑以下几点：(1) 丰富性：生成出音乐的多样性和趣味性；(2) 正确性：作曲技巧与演奏错误；(3) 结构性：音乐是否具有某种风格或音乐模式。另一个方面是生成的音乐和视频的匹配程度，主要考虑以下几点：(1) 节奏性：生成音乐与视频运动的匹配程度。举例来说，一个激烈体育运动 vlog 会与一首快节奏的音乐相匹配，而一个平缓的旅行 vlog 会与一首轻柔的慢节奏音乐相匹配。(2) 同步性：音乐的重音或边界是否与视频的视觉节拍相匹配。举例来说，对于一个比较有节奏感的视频，例如舞蹈视频，音乐的重音应当落在主要的舞步上。(3) 视频结构性：音乐的起点和终点是否与视频的起点和终点相匹配。类似地，音乐与视频都有序章、插曲和终章，音乐与视频的相应部分应当相互匹配。问卷

需要大约十分钟来完成。结果如表 2 所示，我们生成的音乐虽然在音乐性上不如匹配的人工创作的音乐，但是在音视频匹配程度上和最后的排名上，我们提出的算法均领先。

表 2 主观指标实验结果

| Model | Baseline | Matched | Ours |
|---|---|---|---|
| Melodiousness↑ | 3.4 | 4.0 | 3.8 |
| Compatibility↑ | 3.4 | 3.7 | 3.9 |
| Overall Rank↓ | 2.3 | 1.9 | 1.8 |

## 四、总结

本文中我们针对未被探索过的视频背景音乐生成任务，提出了关联信息可控的生成模型 CMT，使用基于复合词的音乐表示形式，并加入了“组合音符”的密度与强度两个属性来与视频中的运动强度和视觉节拍相关联。我们的模型无需配对的视频-音乐训练数据，在训练时只使用音乐，而在生成时直接利用视频提取出的关联信息。经过实验，我们的模型生成出了能够与视频的节拍、感情风格相匹配的音乐，生成音乐的质量也比较可观。该项研究对于短视频制作、直播与电商等场景具有实际应用价值。

责任编委 储珺

## 参考文献

[1] Chris Donahue, Huanru Henry Mao, Yiting Ethan Li, Garrison W Cottrell, and Julian McAuley. 2019. LakhNES: Improving multi-instrumental music generation with cross-domain pre-training. arXiv preprint arXiv:1907.04868 (2019).

[2] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. 2018. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32.

[3] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. 2018. GANSynth: Adversarial Neural Audio Synthesis. In International Conference on Learning Representations.

[4] Chuang Gan, Deng Huang, Peihao Chen, and Joshua B Tenenbaum. [n.d.]. Foley music: Learning to generate music from videos. ([n. d.]).

[5] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751 (2019).

[6] Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang. 2021. Compound Word Transformer: Learning to Compose Full-Song Music over Dynamic Directed Hypergraphs. arXiv preprint arXiv:2101.02402 (2021).

[7] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music Transformer: Generating Music with Long-Term Structure. In International Conference on Learning Representations.

[8] Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop Music Transformer: Beat-based modeling and generation of expressive Pop piano compositions. In Proceedings of the 28th ACM International Conference on Multimedia. 1180–1188.

[9] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. 2020. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In Proceedings of the International Conference on Machine Learning (ICML).

[10] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. 2016. SampleRNN: An unconditional end-to-end neural audio generation model. arXiv preprint arXiv:1612.07837 (2016).

[11] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016).

[12] Christine Payne. 2019. MuseNet. OpenAI Blog 3 (2019).

[13] Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. Popmag: Pop music accompaniment generation. In Proceedings of the 28th ACM International Conference on Multimedia. 1198–1206.

[14] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. 2018. A hierarchical latent vector model for learning long-term structure in music. In International Conference on Machine Learning. PMLR, 4364–4373.

[15] Kun Su, Xiulong Liu, and Eli Shlizerman. 2020. Audeo: Audio generation for a silent performance video. arXiv preprint arXiv:2006.14348 (2020).

[16] Kun Su, Xiulong Liu, and Eli Shlizerman. 2020. Multi-Instrumentalist Net: Unsupervised Generation of Music from Body Movements. arXiv preprint arXiv:2012.03478 (2020).

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems. 6000–6010.

[18] Ziyu Wang, Yiyi Zhang, Yixiao Zhang, Junyan Jiang, Ruihan Yang, Junbo Zhao, and Gus Xia. 2020. Pianotree vae: Structured representation learning for polyphonic music. arXiv preprint arXiv:2008.07118 (2020).

[19] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. 2017. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. arXiv preprint arXiv:1703.10847 (2017).

[20] Andrea Valenti, Antonio Carta, and Davide Bacciu. 2020. Learning Style-Aware Symbolic Music Representations by Adversarial Autoencoders. arXiv preprint arXiv:2001.05494 (2020).

刘偲

北航副教授，博导。研究方向是跨模态多媒体智能分析(跨模态包含自然语言，计算机视觉以及语音等)以及经典计算机视觉任务(目标检测、跟踪和分割)。
个人主页：http://colalab.org/
Email: liusi@buaa.edu.cn

专题综述

# 动物姿态估计研究与展望

悉尼大学 张敬

## 一、引言

近年来，对动物的保护和动物行为的分析受到越来越多的关注。准确的动物姿态估计是动物行为分析的重要基础。动物姿态估计旨在对动物运动的关键点进行准确的描述和估计，从而为后续对动物行为的分析创造条件。此外，准确的动物姿态估计在动画制作，仿真设计，动物保护等应用场景中发挥重要作用。

现有的姿态估计方法主要聚焦于人体姿态估计，即识别检测人体上具有语义信息的关键点[8][9][10][12][15][16]，如图 1 所示。自 2014 年以来，人体姿态估计算法经历了快速的发展并在一系列应用中展现出极佳的识别准确度和应用前景。

图 1 人体姿态估计示意图

相较于人体姿态估计，动物姿态估计任务存在更多挑战。其中，最主要的挑战是如何应对自然界中动物种类的多样性。自然界中存在大量不同种类的动物。由于物种、生存环境等不同，不同种类的动物在皮毛，行为，姿态等方面存在巨大差异。先进的人体姿态估计算法是否可以应对这样巨大的差异性是未知且需要探索的。

目前用于动物姿态估计的数据集大多聚焦于特定的动物种类，如老虎，马，猫，狗等，而忽略了自然界中动物种类的多样性。使用这些数据集进行训练和测试只能对算法在某一特定动物种类上的表现进行评估。考虑到动物种类的多样性，这样的数据集是远远不够的。这也让动物姿态估计领域的一个重要问题悬而未决，即现有姿态估计算法是否可以很好的泛化到自然界中各种各样的动物上?

为了探索这个问题的答案，迫切需要构建一个包含多种动物和其姿态标注的大规模数据集。此外，这样一个的数据集也有利于推动动物姿态估计算法的研究从聚焦于单一动物种类的动物姿态估计到一个普适的具有良好泛化性的通用动物姿态估计的发展。

## 二、姿态估计算法

现有人体姿态估计算法方面的研究可以粗略划分为两类，即自下向上的方法和自上向下的方法。前者指直接根据输入图片回归出人体的关键点信息，并对其进行分组，以得到不同个体的人体关键点检测结果；自上向下的方法指根据人体在图片中所在的位置，将人从图片中分割出来，并针对各个人体进行独立的人体姿态估计。虽然自下向上的方法具有较快的推理速度，尤其是在图像中具有较多个体的情况下，自上向下的方法往往能得到更好的人体姿态估计效果。这些方法往往在通用的人体姿态估计数据集上进行训练和效果评估，包括 MS COCO[6]和 MPII[1]。其中，MS COCO 包含大概 250000 多个人体实例以用于训练和测试，MPII 数据集中包含超过 40000 张有标注的人体实例。为了评估人

体姿态估计算法在更复杂场景中的表现能力，一些包含复杂场景的人体姿态估计数据集也被提出，比方说包含拥挤人群场景的数据集 CrowdPose[4]和包含遮挡场景的数据集 OCHuman[13]。这些数据集极大的帮助了人体姿态估计算法的快速发展，并帮助这些人体姿态估计算法应对各种挑战的场景，例如不同光照条件的变化，人体姿态的变化，人体尺度的变化，以及人群拥挤和遮挡等场景。这些丰富的数据集提供的测试基准为各个先进的人体姿态估计算法在各种应用场景中发挥出良好的效果提供了重要参考。

动物姿态估计和人体姿态估计算法本质上是相似的，即都是将包含动物或者人体的图片作为输入，通过网络预测对应的关键点信息。两者的区别更多的在于由于目标动物或人体姿态、纹理、生物结构等的不同而导致的关键点定义不同。然而，相较于人体姿态估计数据集的多样性和丰富程度，目前只有少数几个关于动物姿态估计的数据集提供给动物姿态估计识别算法进行训练和测试。然而，由于深度学习算法是数据驱动的，这样有限的数据集会影响和限制目前的动物姿态估计模型的性能和泛化性。此外，由于动物数据整理和标注的难度，这些数据集往往只关注特定的动物种类，例如，马，斑马，蚊子，老虎等。这样有限的动物种类使这些数据集中只包含有限的动物姿态信息、纹理信息和动物栖息地的背景信息等。Animal Pose Dataset[2] 尝试包含更多种类的动物，以帮助模型学到更具备泛化性的特征。然而，Animal Pose 数据集也仅仅包含 5 类动物，

这还是难以让网络学到足够具有泛化性的特征表示。此外，尽管不同种类的动物有不同的外观、行为模式和骨骼分布情况，他们往往会遵循一定的生物学规律，例如生物进化过程中自然产生的科，目，种等分类方式。属于同一属的动物往往会比属于不同属的动物有更为接近的生活习性，行为模式和姿态分布等。具体来说，相较于牛和黑猩猩，牛和马有更为相近的关键点分布。这是因为牛和马同属于偶蹄目，而黑猩猩则并不属于偶蹄目。利用不同动物生物学上的相似特性可以帮助在有限动物种类上训练好的动物姿态估计网络更好的泛化到未知的动物种类上。因此，一个大规模的，按照生物学规律进行整理的动物姿态估计数据集是有必要的。

表 1 动物关键点定义

| Keypoint | Definition | Keypoint | Definition |
|---|---|---|---|
| 1 | Left Eye | 10 | Right Elbow |
| 2 | Right Eye | 11 | Right Front Paw |
| 3 | Nose | 12 | Left Hip |
| 4 | Neck | 13 | Left Knee |
| 5 | Root of Tail | 14 | Left Back Paw |
| 6 | Left Shoulder | 15 | Right Hip |
| 7 | Left Elbow | 16 | Right Knee |
| 8 | Left Front Paw | 17 | Right Back Paw |
| 9 | Right Shoulder | | |

## 三、 动物姿态估计数据集

### 3.1 数据集收集

本文构建了一个大规模动物姿态估计数据集 AP-10K[11]。为了得到高质量动物数据，AP-10K 以 9 个公开发布的用于动物分类的数据集为基础，经过仔细清洗、鉴别、再组织和标记，构建了一个包含 59658 张图片的动物数据集。在这个数据集中，不同动物按照科和物种的生物学概念进行了准确划分，物种之间的生物学关系得到了清晰的体现。在此基础上，经过仔细分析和挑选，本着“每个物种选取 200 张作为基础，稀有物种充分标记”的原则，我们对其中 54 类动物进行标记，最终得到了 10015 张包含姿态信息的图片。表 1 展示了 17 个关键点的定义，图 2 展示了一幅黑猩猩图片其对应的标记。

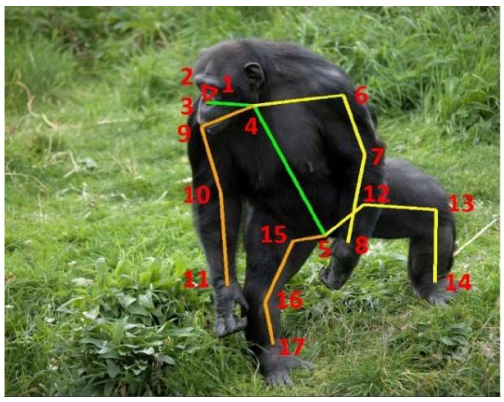



图 2 黑猩猩图片及其对应标记点

表 2 不同动物数据集对比

| Dataset | Species | Family | Labeled image | Unlabeled image | Keypoint | Instance |
|---|---|---|---|---|---|---|
| Animal-Pose Dataset [2] | 5 | N/A | 4,666 | 0 | 20 | 6,117 |
| Horses-10 [7] | 1 | N/A | 8,110 | 0 | 22 | 8,110 |
| ATRW [5] | 1 | N/A | 8,076 | 0 | 15 | 9,496 |
| AP-10K | 54 | 23 | 10,015 | 50k | 17 | 13,028 |

表 3 全监督学习结果比较

| | HRNet-w32 [9] | HRNet-w48 [9] | ResNet50 [3] | ResNet101 [3] | Hourglass [8] |
|---|---|---|---|---|---|
| w/o pretraining | $0.703_{\pm 0.002}$ | $0.713_{\pm 0.002}$ | $0.646_{\pm 0.001}$ | $0.667_{\pm 0.002}$ | $0.686_{\pm 0.006}$ |
| w/ pretraining | $0.738_{\pm 0.006}$ | $0.744_{\pm 0.004}$ | $0.699_{\pm 0.004}$ | $0.698_{\pm 0.002}$ | $0.729_{\pm 0.001}$ |

### 3.2 数据集整理

为了利用好生物进化规律以帮助使用有限动物种类进行训练的动物姿态估计模型更好的泛化到自然界中各种动物上，我们在构建 AP-10K 的过程中对收集到的数据按照分类阶元进行了重新整理和标注。为了简化分类层级，我们没有按照科-属-种关系对动物进行整理，而是按照科-种关系对动物进行整理，即对属和种不作进一步区分。此外，按照生物学进化规律对动物进行划分可以对网络泛化能力进行更为公平的评估，并为提升网络在特定动物类别关键点估计能力提供依据和指导。

### 3.3 数据集标注

为了获得高质量的标记效果，我们招募了 13 名经过训练的志愿者对数据进行标注。此外，我们提供了详尽的文档对于标记者可能遇到的标记状况进行了详细的解说，其中包括对于多个体、遮挡情况等情形的处理情况等。这些举措保证了多个体、遮挡等有难度的少见样本的准确标记效果。为了更进一步保证标注信息的质量，我们采取了自动化和人工两种校验手段。其中自动化校验是指根据预设规则对于标记好的坐标信息进行自动化检查，去除一些低质量标记和错误标记。例如标记点落在检测框外侧，同一个实例出现重复的标记名称等。人工校验是指组织者和标记者进行了三轮检查，这确保了高质量的标注信息。三轮检查过程如下：首先，标记者在分配的标记工作完成后，将标记结果提交组织者进行检查，组织者将检查出的错误信息反馈给标记者，这是一轮检查；标记者根据反馈的勘误表对标记进行修改，并将二次修改结果反馈给组织者，这是二轮检查；最后组织者拿到二次标记结果，对于标记进行最后的检查，如果发现错误就进行本地修改，这是三轮检查。三轮检查的过程如同 TCP 协议的三次握手一般，加强了标注可靠性。

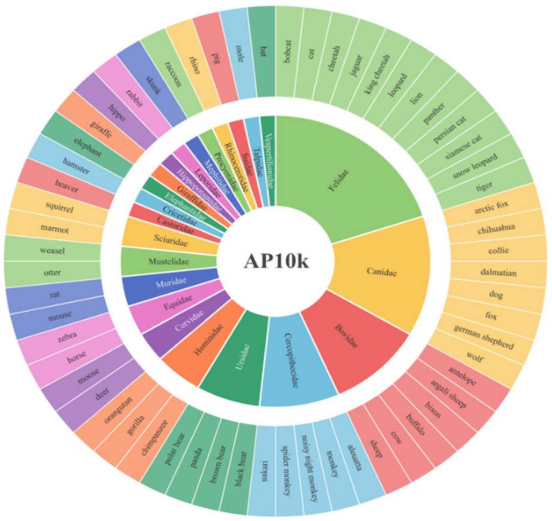



图 3 AP-10K 动物种类分布情况

### 3.4 数据集统计指标

AP-10K 数据集中包含 10015 张完全标记的动物图片及其对应标注，共 13028 个不同的动物个体，涵盖了 23 科，54 种不同的动物种类。这使得 AP-10K 有更为复杂的动物姿态分布和纹理信息。这样的特点让使用 AP-10K 进行训练的模型展现出更好的泛化性能。下图展示了 AP-10K 数据集的特点(表 2)和动物种类分布(图 3)。由图中可以看出，AP-10K 数据集不论是在动物种类还是在标记数量均具有显著优势。值得一提的，AP-

10K 数据集的标记图片具有长尾分布的特点，比如对于猫科(Felidae)来说，一共有 10 个标记物种，1913 张标记图片，而河狸科只包含 1 个物种，178 张标记图片。这些特性对于小样本学习、零样本学习或者元学习等研究方向是很有意义的。此外，AP-10K 数据集中额外包含 50K 张含有类别标注但是缺少关键点标注的动物图片。这些图片和对应的生物学标注可以为研究跨物种动物姿态估计的自监督[15]和半监督学习等课题提供条件。

表 4 人体姿态估计模型迁移学习效果比较

| epoch | AP | $AP_{.5}$ | $AP_{.75}$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|
| 20 | $0.606_{\pm 0.004}$ | $0.906_{\pm 0.005}$ | $0.635_{\pm 0.006}$ | $0.501_{\pm 0.037}$ | $0.610_{\pm 0.003}$ |
| 30 | $0.642_{\pm 0.002}$ | $0.921_{\pm 0.010}$ | $0.680_{\pm 0.002}$ | $0.521_{\pm 0.044}$ | $0.645_{\pm 0.002}$ |
| 40 | $0.667_{\pm 0.003}$ | $0.934_{\pm 0.004}$ | $0.714_{\pm 0.007}$ | $0.547_{\pm 0.059}$ | $0.671_{\pm 0.003}$ |
| 210 | $0.753_{\pm 0.005}$ | $0.962_{\pm 0.002}$ | $0.827_{\pm 0.003}$ | $0.616_{\pm 0.031}$ | $0.756_{\pm 0.004}$ |

表 5 牛科科内泛化实验

| Train / Test | Bov./Ant | Bov./A.S. | Bov./Bis | Bov./Buf | Bov./Cow | Bov./She. | Average |
|---|---|---|---|---|---|---|---|
| Antelope | $0.607_{\pm 0.010}$ | $0.742_{\pm 0.013}$ | $0.775_{\pm 0.004}$ | $0.842_{\pm 0.005}$ | $0.729_{\pm 0.002}$ | $0.853_{\pm 0.002}$ | $0.788_{\pm 0.051}$ |
| A.S. | $0.836_{\pm 0.016}$ | $0.655_{\pm 0.015}$ | $0.805_{\pm 0.022}$ | $0.840_{\pm 0.007}$ | $0.725_{\pm 0.027}$ | $0.697_{\pm 0.008}$ | $0.781_{\pm 0.059}$ |
| Bison | $0.731_{\pm 0.017}$ | $0.646_{\pm 0.009}$ | $0.530_{\pm 0.006}$ | $0.605_{\pm 0.006}$ | $0.616_{\pm 0.009}$ | $0.693_{\pm 0.014}$ | $0.658_{\pm 0.047}$ |
| Buffalo | $0.783_{\pm 0.010}$ | $0.748_{\pm 0.031}$ | $0.726_{\pm 0.017}$ | $0.658_{\pm 0.004}$ | $0.794_{\pm 0.022}$ | $0.750_{\pm 0.008}$ | $0.760_{\pm 0.025}$ |
| Cow | $0.597_{\pm 0.011}$ | $0.691_{\pm 0.004}$ | $0.740_{\pm 0.007}$ | $0.732_{\pm 0.009}$ | $0.586_{\pm 0.006}$ | $0.683_{\pm 0.002}$ | $0.689_{\pm 0.051}$ |
| Sheep | $0.707_{\pm 0.012}$ | $0.607_{\pm 0.006}$ | $0.681_{\pm 0.004}$ | $0.676_{\pm 0.007}$ | $0.645_{\pm 0.005}$ | $0.520_{\pm 0.001}$ | $0.663_{\pm 0.034}$ |

表 6 狗科科内泛化实验

| Train / Test | Can./Dog | Can./Fox | Can./Wolf | Average |
|---|---|---|---|---|
| Dog | $0.224_{\pm 0.011}$ | $0.699_{\pm 0.009}$ | $0.699_{\pm 0.003}$ | $0.699_{\pm 0.000}$ |
| Fox | $0.614_{\pm 0.013}$ | $0.627_{\pm 0.005}$ | $0.732_{\pm 0.013}$ | $0.673_{\pm 0.059}$ |
| Wolf | $0.663_{\pm 0.024}$ | $0.694_{\pm 0.013}$ | $0.633_{\pm 0.006}$ | $0.679_{\pm 0.016}$ |

## 四、应用

### 4.1 全监督学习

AP-10K 评估了五种主流的人体姿态估计模型在动物姿态估计任务上的表现，它们分别是 HRNet-w32[9]，HRNet-w48[9]，SimpleBaseline[10] (ResNet50[3]骨干网络)[3]，SimpleBaseline[10](ResNet101[3]骨干网络)和 Hourglass[8]，然后又对比了使用流行的 ImageNet 预训练模型和随机初始化网络进行训练的效果。实验结果(表 3)表明：使用 ImageNet[14] 预训练比随机初始化的效果要更好，ImageNet 预训练能够提升上述 5 种模型的性能。随着网络规模的增大，HRNet[9] 和 SimpleBaseline[10] 的训练指标也逐渐提升，这展现了大网络具有更好的表征能力这一特性。值得注意的是，增加训练代数可以提升随机初始化模型的最终性能，降低其与采用 ImageNet 预训练权重初始化模型的性能差距，详细结果可以参考 AP-10k 论文[11]。

### 4.2 人体姿态估计模型的迁移学习

因为人和四足动物的相似性，评估人体姿态估计模型到动物姿态估计模型的泛化能力是一个很有必要的事情。AP-10K 使用 HRNet-w32 模型，加载基于 COCO 的人体姿态估计任务预训练模型的权重，然后在 AP-10K 数据集上进行微调并测试。实验结果(表 4)表明当训练 epoch 较少时，人体姿态估计算法迁移到动物姿态估计的结果不够好，这是因为动物和人在外形

表 7 猫科科内泛化实验

| Train / Test | Fel./Bob. | Fel./Cat | Fel./Che. | Fel./Jag. | Fel./K.C. | Fel./Leo. | Fel./Lio. | Fel./Pan. | Fel./S.L. | Fel./Tig. | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bob. | 0.631 ±0.005 | 0.714 ±0.016 | 0.664 ±0.004 | 0.674 ±0.013 | 0.673 ±0.013 | 0.663 ±0.006 | 0.691 ±0.016 | 0.623 ±0.004 | 0.669 ±0.005 | 0.713 ±0.008 | 0.676 ±0.026 |
| Cat | 0.638 ±0.002 | 0.332 ±0.004 | 0.625 ±0.018 | 0.552 ±0.010 | 0.629 ±0.007 | 0.641 ±0.009 | 0.601 ±0.004 | 0.609 ±0.010 | 0.582 ±0.014 | 0.608 ±0.007 | 0.609 ±0.027 |
| Che. | 0.715 ±0.002 | 0.716 ±0.012 | 0.660 ±0.003 | 0.762 ±0.013 | 0.731 ±0.014 | 0.747 ±0.010 | 0.734 ±0.021 | 0.790 ±0.008 | 0.713 ±0.008 | 0.662 ±0.008 | 0.730 ±0.034 |
| Jag. | 0.757 ±0.005 | 0.770 ±0.017 | 0.754 ±0.006 | 0.704 ±0.008 | 0.750 ±0.004 | 0.759 ±0.012 | 0.798 ±0.013 | 0.724 ±0.008 | 0.756 ±0.011 | 0.734 ±0.005 | 0.756 ±0.020 |
| K.C. | 0.961 ±0.008 | 0.804 ±0.035 | 0.692 ±0.042 | 0.771 ±0.028 | 0.779 ±0.010 | 0.958 ±0.008 | 0.713 ±0.017 | 0.924 ±0.026 | 0.864 ±0.033 | 0.838 ±0.016 | 0.836 ±0.094 |
| Leo. | 0.730 ±0.005 | 0.697 ±0.007 | 0.766 ±0.014 | 0.741 ±0.006 | 0.682 ±0.005 | 0.686 ±0.009 | 0.700 ±0.012 | 0.705 ±0.012 | 0.775 ±0.010 | 0.744 ±0.004 | 0.727 ±0.031 |
| Lio. | 0.623 ±0.016 | 0.582 ±0.023 | 0.639 ±0.012 | 0.694 ±0.010 | 0.688 ±0.002 | 0.690 ±0.018 | 0.528 ±0.002 | 0.638 ±0.007 | 0.630 ±0.011 | 0.625 ±0.024 | 0.645 ±0.036 |
| Pan. | 0.705 ±0.020 | 0.722 ±0.011 | 0.718 ±0.020 | 0.720 ±0.023 | 0.727 ±0.013 | 0.785 ±0.014 | 0.763 ±0.026 | 0.511 ±0.014 | 0.719 ±0.004 | 0.684 ±0.018 | 0.727 ±0.028 |
| S.L. | 0.792 ±0.011 | 0.776 ±0.008 | 0.810 ±0.018 | 0.779 ±0.019 | 0.790 ±0.024 | 0.818 ±0.004 | 0.821 ±0.009 | 0.760 ±0.015 | 0.724 ±0.010 | 0.855 ±0.012 | 0.800 ±0.027 |
| Tig. | 0.754 ±0.008 | 0.741 ±0.018 | 0.751 ±0.012 | 0.715 ±0.015 | 0.768 ±0.021 | 0.753 ±0.015 | 0.797 ±0.005 | 0.848 ±0.023 | 0.744 ±0.011 | 0.675 ±0.007 | 0.763 ±0.036 |

表 8 科间泛化实验

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| train | Bov. | 0.782±0.002 | Bov. | 0.782±0.002 | Bov. | 0.782±0.002 | Cerc. | 0.695±0.007 |
| | Ant. | 0.856±0.001 | Ant. | 0.856±0.001 | Ant. | 0.856±0.001 | Alo. | 0.697±0.020 |
| | A.S. | 0.887±0.006 | A.S. | 0.887±0.006 | A.S. | 0.887±0.006 | Mon. | 0.725±0.013 |
| | Bis. | 0.643±0.005 | Bis. | 0.643±0.005 | Bis. | 0.643±0.005 | N.N.M. | 0.750±0.027 |
| | Buf. | 0.815±0.004 | Buf. | 0.815±0.004 | Buf. | 0.815±0.004 | S.M. | 0.581±0.008 |
| | Cow. | 0.737±0.004 | Cow. | 0.737±0.004 | Cow. | 0.737±0.004 | Uak. | 0.720±0.009 |
| | She. | 0.754±0.005 | She. | 0.754±0.005 | She. | 0.754±0.002 | | |
| test | Cer. | 0.641±0.007 | Equ. | 0.468±0.019 | Hom. | 0.015±0.001 | Hom. | 0.446±0.007 |
| | Der. | 0.724±0.004 | Hor. | 0.618±0.005 | Chi. | 0.005±0.000 | Chi. | 0.446±0.011 |
| | Moo. | 0.558±0.010 | Zeb. | 0.319±0.035 | Gor. | 0.026±0.003 | Gor. | 0.445±0.011 |

和纹理上有较大的差异性。随着训练时间的增加，微调的效果也逐渐增加，并显著优于采用 ImageNet 预训练模型进行训练的结果。该结果表明，人体姿态估计和动物姿态估计任务之间域间隔(Domain Gap)相比姿态估计任务和图像分类任务之间域间隔更小。

### 4.3 动物姿态估计模型在科内和科间的泛化性能

为了验证动物姿态估计模型在同一科内和相似动物科之间的泛化性能，我们选择了 AP-10K 中三个数量最多的科(牛，狗和猫)进行实验。在每科中，一个物种被用作测试集而剩下的物种构成训练集。科内实验结果(表 5-7)表明，在三个不同科中，测试物种的分数虽然不如在第一部分中使用大量物种进行训练的效果好，但是也能达到一个不错的结果。这是因为同科物种在生物学关系和外形上具有高度相似性。实验结果中狗(Dog)的分数偏低，这是因为相比狐狸(Fox)和狼(Wolf)，狗(Dog)包含了更多的图片，将其排除之后训练集图片数量较少。其次，狗(Dog)中包含了许多人工培育的宠物类型，它们的外形差异较大，类似现象也存在于猫(Cat)中。

在科间实验中，牛科被用作为训练集，鹿科(Cervidae)、马科(Equidae)和人科(Hominidae)被分别用作测试集。科间实验结果(表 8)表明，使用牛科作为训练集的模型在鹿科和马科的泛化结果很好，但是在人科上泛化效果较差。因为牛科和鹿科、马科的生物学关系相近，外形差异也较小。而人科物种和牛科生物学关系

表 9　科间迁移学习和少样本学习效果

| Species | Setting | Performance | Species | Setting | Performance |
|---|---|---|---|---|---|
| Deer | Generalization | $0.723_{\pm 0.036}$ | Moose | Generalization | $0.587_{\pm 0.025}$ |
| | Few-Shot | $0.742_{\pm 0.034}$ | | Few-Shot | $0.648_{\pm 0.025}$ |
| | Transfer | $0.751_{\pm 0.024}$ | | Transfer | $0.726_{\pm 0.011}$ |
| Horse | Generalization | $0.592_{\pm 0.047}$ | Zebra | Generalization | $0.324_{\pm 0.021}$ |
| | Few-Shot | $0.635_{\pm 0.034}$ | | Few-Shot | $0.480_{\pm 0.029}$ |
| | Transfer | $0.718_{\pm 0.023}$ | | Transfer | $0.708_{\pm 0.024}$ |
| Chimpanzee | Generalization | $0.009_{\pm 0.006}$ | Gorilla | Generalization | $0.017_{\pm 0.006}$ |
| | Few-Shot | $0.022_{\pm 0.010}$ | | Few-Shot | $0.144_{\pm 0.121}$ |
| | Transfer | $0.550_{\pm 0.032}$ | | Transfer | $0.662_{\pm 0.039}$ |

表 10 跨数据集泛化效果比较

| | Direct Test(mAP) | Finetune&Test(mAP) | Train&Test(mAP) |
|---|---|---|---|
| Animal-Pose Dataset[2] → AP-10K | 0.424 | 0.722 | 0.727 |
| AP-10K → Animal-Pose Dataset[2] | 0.913 | 0.935 | 0.932 |

较远，外形和生存环境也差异较大，所以泛化效果不好。作为对照，表格最后一列使用了猴科(Cercopithecidae)作为训练集来测试人科的物种，性能得到大幅提升，这再次证明了 AP-10K 在构建过程中采用生物学进化规律的必要性：生物学关系和外形相似的物种，彼此之间域差异也越小，更利于姿态估计模型的泛化。

### 4.4 科间的迁移学习和少样本学习

在科间泛化实验的基础上，AP-10K 进一步探究了少样本学习和迁移学习带来的性能提升。与科间迁移实验相同，牛科图片被作为训练集，然后鹿科(Deer 和 Moose)、马科(Horse 和 Zebra)和人科(Chimpanzee 和 Gorilla)图片被用于微调和测试，其中少样本学习对每个物种抽样 20 张进行微调，而迁移学习采用该物种全部训练集图片进行微调。实验结果(表 9)表明少样本学习和迁移学习效果均相对于直接泛化测试有了不同程度的提升。即便是对人科这样和训练集差距较大的测试集，采用更多的图片进行迁移也能得到性能的提升。

### 4.5 跨数据集泛化能力比较

如表 10 所示，我们使用 Animal Pose Dataset (包含 5 类动物)和 AP-10K 数据集分别训练姿态估计模型并对比了它们的双向泛化效果。结果表明，采用包含更多物种的 AP-10K 数据集进行(预)训练的模型的泛化性能优于使用少量动物数据进行训练的模型。

## 五、展望

AP-10K 是第一个大规模的哺乳动物姿态数据集。它的物种数量、姿态多样性，以及按照生物学关系组织上的优势可以极大的促进相关领域的研究，例如动物保护和动物行为研究等。我们基于 AP-10K 训练了 5 种经典的姿态估计模型并测试了它们的在不同物种上的表现能力，初步探究了动物和人体姿态估计之间的联系以及不同物种之间的泛化效果。总得来说，AP-10K 数据集为动物姿态估计领域提供新的可能性和发展发向。

**致谢：**本文由博士生徐宇飞(悉尼大学)、喻航(西安电子科技大学)撰写初稿，指导老师张敬(悉尼大学)进行修改。本文对应发表在 NeurIPS2021 的学术论文作者还包括赵伟教授(西安电子科技大学)、管子玉教授(西安电子科技大学)、陶大程教授(京东探索研究院)。

论文链接：
https://openreview.net/forum?id=rH8yliN6C83

数据集和代码链接：
https://github.com/AlexTheBad/AP-10K

责任编委　崔海楠

## 参考文献

[1] Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. In CVPR.

[2] Cao, J., Tang, H., Fang, H. S., Shen, X., Lu, C., & Tai, Y. W. (2019). Cross-domain adaptation for animal pose estimation. In CVPR.

[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In CVPR.

[4] Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H. S., & Lu, C. (2019). Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In CVPR

[5] Li, S., Li, J., Tang, H., Qian, R., & Lin, W. (2020). ATRW: A Benchmark for Amur Tiger Re-identification in the Wild. In Proceedings of the 28th ACM International Conference on Multimedia.

[6] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In ECCV.

[7] Mathis, A., Biasi, T., Schneider, S., Yuksekgonul, M., Rogers, B., Bethge, M., & Mathis, M. W. (2021). Pretraining boosts out-of-domain robustness for pose estimation. IEEE/CVF Winter Conference on Applications of Computer Vision.

[8] Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In ECCV.

[9] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[10] Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In ECCV.

[11] Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D. (2021). AP-10K: A Benchmark for Animal Pose Estimation in the Wild. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

[12] Zhang, J., Chen, Z., & Tao, D. (2021). Towards high performance human keypoint detection. International Journal of Computer Vision, 129(9), 2639-2662.

[13] Zhang, S. H., Li, R., Dong, X., Rosin, P., Cai, Z., Xi, H., ... & Hu, S. M. (2019). Pose2seg: Detection free human instance segmentation. In CVPR.

[14] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In CVPR.

[15] Xu, Y., Zhang, Q., Zhang, J., & Tao, D. (2021). RegionCL: Can Simple Region Swapping Contribute to Contrastive Learning? arXiv preprint arXiv:2111.12309.

[16] Xu, Y., Zhang, Q., Zhang, J., & Tao, D. (2021). ViTAE: Vision Transformer Advanced by Exploring Intrinsic Inductive Bias. In Thirty-fifth Conference on Neural Information Processing Systems.

**张敬**

悉尼大学工程学院计算机系，博士后研究员。主要研究方向为计算机视觉和深度学习，已经在计算机视觉及人工智能相关领域的国内外著名学术期刊和会议发表论文 40 余篇，担任多个国际学术期刊和会议的审稿人，以及 IJCAI、AAAI 的 Senior Program Committee Member。
Email: jing.zhang1@sydney.edu.au

顶会观察

# ICCV 2021

微软亚洲研究院视觉计算组研究员 元玉慧

国际计算机视觉大会（International Conference on Computer Vision）是计算机视觉领域的顶级学术会议，与 CVPR 和 ECCV 并称为计算机视觉领域三大顶会。ICCV 属于中国计算机学会推荐国际学术会议中的人工智能领域 A 类会议。今年大会的主席成员包括：来自 Facebook 的 Tamara Berg 和 Tal Hassner、来自麦吉尔大学的 James Clark、来自大阪大学的 Yasuyuki Matsushita、来自宾夕法尼亚大学的 Camillo Jose Taylor、来自蒙特利尔工程学院的 Christopher Pal、来自东京大学的 Yoichi Sato、来自布里斯托大学的 Dima Damen。今年大会最大的研究热点非 Transformer 莫属，在被接收论文中，题目中带有 Transformer 的就有 59 篇，并且最佳论文 Swin Transformer 也是关于 Transformer 的。从被收录论文第一作者的单位来看，43%来自中国，23%来自美国。可以看出，中国在计算机视觉领域的国际舞台上有着举足轻重的地位。

ICCV 2021 于美东部夏令时间的 2021 年 10 月 11 日至 17 日举行。由于新冠疫情的原因，原本安排在加拿大第二大城市蒙特利尔举办的会议转移到了线上平台举办。在会议的线上平台上，参会者可以提前收藏关注自己想要参加的活动或者报告，并且可以直接通过平台提供的 Zoom 会议链接进入直播房间参与讨论。此外，线上平台也支持查询其他参会者的信息、被接收论文的题目和补充材料、获奖论文信息等。大会的第一天和最后两天是以 workshop 和 tutorial 为主，线上平台也提供了 workshop 和 tutorial 的主界面入口。主会是在 10 月 12 日到 15 日之间四天举办的。在主会举办期间，参会者可以自由地与被接收的各篇论文的作者进行面对面的视频交流。下面本文分别从会议概况、论文录用情况、获奖论文的研究工作介绍和精彩的专家观点分享这四个方面进行详细地介绍。

## 一、 会议概况

James Clark 代表 ICCV 2021 的主席团成员(general chairs)致辞欢迎所有参会者并介绍了会议的安排：82 场 workshops、12 场 tutorials、20 多场企业展览、37 场 mentorship 会议、7 场 affinity 小组会议等。大会最大的三家赞助商包括：谷歌研究院、索尼公司、摩根斯坦利公司(据说这是摩根斯坦利第一次赞助 ICCV 会议)。根据介绍，今年有 4000 多名线上参会者注册了 ICCV 2021。大会还安排了几个特殊环节：例如，两位设计家和艺术家给了一个关于技术发展与社会之间关系的 Keynote Lecture，多位知名计算机视觉领域专家参与了一个关于“计算机视觉中的深度学习方法和传统方法”的论坛等。

## 二、 论文录用情况

大会的程序主席们(program chairs)对 ICCV 2021 论文的投稿和收录情况作了详细介绍： ICCV 2021 收到了 6152 篇有效投稿，相比于 ICCV 2019 增加了 1800 篇。其中有 11%的投稿选择了撤稿，因此最终有 5486 篇投稿被审稿人评审。大会论文的接受率是 26%，即收录了 1621 篇论文，其中 210 篇论文被选做 oral，另外 1412 篇论文选做 poster。今年会议组织方邀请了 233 位专家作为领域主席(area chair)，平均每位领域主席需要负责 27 篇投稿。组织方邀请了 4216 位

从业者作为审稿人参与论文评审，其中有 2746 位经验丰富的审稿人且每人被分配 7 篇论文、1462 位学生审稿人且每人被分配 4 篇论文、622 位紧急审稿人。每篇论文会收到最少三个评审意见且最终由两位领域主席一起讨论决定是否予以接收。

在被接收论文中，数量最多的研究领域包括：transfer/low-shot/unsupervised learning, image and video synthesis, recognition and classification, detection and localization in 2D and 3D 等。这四个研究领域都有超过 75 篇被录用的论文，其中关于 transfer/low-shot/unsupervised learning 的论文录用数目接近 125 篇。数量最少的研究领域包括 visual reason and logical representation, biometrics, faces 等。大会也按照不同的研究领域分别统计了接收率，其中接收率最高的研究领域包括 gestures and body pose 和 vision for robotics and autonomous vehicles。整体上不同领域的论文接收率区别并不大。

## 三、 获奖论文选介

大会程序主席 Dima Damen 宣布了 ICCV 2021 的颁奖信息，首先宣布了获得最佳审稿人的名单，从经验丰富的审稿人和学生审稿人中分别选取了前 5%共 210 位作为最佳审稿人。然后宣布了今年的马尔奖(Marr prize)的评委成员并宣布了获奖论文信息：有四篇论文荣获最佳论文提名奖、有一篇论文荣获最佳学生论文奖、有一篇论文荣获最佳论文奖。

最佳论文提名奖：

1 Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. 这篇论文的第一作者 Jonathan T. Barro 来自谷歌研究院，曾因工作 NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis 荣获过 ECCV 2020 最佳论文提名奖。 Mip-NeRF 是为了解决原始 NeRF 在处理不同分辨率的图片时会碰到渲染效果模糊的问题。即使与在不同分辨率的图片上训练的 NeRF 模型相比，Mip-NeRF 的渲染效果也要好很多。不同于 NeRF 采用一个离散的(可阻挡且可发光的)粒子构成连续的体积场(continuous volumetric field)来表示要渲染的场景，Mip-NeRF 提出在一个连续尺度的空间中构建体积场来表示要渲染的场景并采用圆锥的截面而非射线的形式来编码采样空间中的点。实验表明 Mip-NeRF 可以显著提高 NeRF 渲染效果的精细程度且比 NeRF 快 7%，另外， Mip-NeRF 在速度提升 22 倍的情况下取得了与暴力上采样的 NeRF 可比的渲染效果。

2 OpenGAN: Open-Set Recognition via Open Data Generation. 这篇论文的第一作者 Shu Kong 来自卡耐基梅隆大学。为了解决如何判断来自开放集合(即真实世界中)的图片是否包含封闭训练集中出现的类别的问题，OpenGAN 提出结合两种传统做法：学习一个二类的判别器来判断图片是来自开放集合还是封闭训练集合，在封闭训练集合上训练一个对抗生成网络模型(GAN)并利用该模型的判别器来判断图片属于开放集合的可能性。OpenGAN 还发现选择合适的判别器网络结构、对抗地生成一些开放集合的训练数据和在封闭集合的分类器网络抽取的特征上训练判别器对提高最后的实验结果都很重要。

3 Viewing Graph Solvability via Cycle Consistency. 这篇论文的第一作者 Federica Arrigoni 来自特伦托大学。这篇论文是关于运动恢复结构(structure-from-motion)的工作，提出一种新的算法来判断一个视角图(viewing graph)是否可解，即给定的视角图能否推断出一组唯一的投影相机参数。论文中分析了已有的根据多张图片进行唯一三维重建的理论条件并为三维重建理论提供了新的基石。

4 Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction. (CO3D) 这篇论文的第一作者 Jeremy Reizenstein 来自 Facebook AI Research。类似于图像识别领域中用于图像检测和分割的 COCO 数据集，CO3D 是目前最大的面向真实世界的三维识别与重建数据集。 CO3D 包含有 150 万个样本，其中每一个样本都包含多个视角拍摄的图像以及图片中物体的三维点云。另外，论文中也提出了 NerFormer 网络结构来利用 Transformer 结构实现三维重建，即根据

几张从不同视角拍摄的图片来重构物体的三维结构。

最佳学生论文奖：

Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. 这篇论文的第一作者 Philipp Lindenberger 来自苏黎世联邦理工大学。论文针对运动恢复结构(structure-from-motion)问题中的两个关键步骤进行了改进：首先，调整初始二维关键点位置先验到任意一个几何估计条件下，然后，对三维点云和相机坐标做后处理。这样的改进操作对检测噪声以及外观变化都非常鲁棒。实验验证了该改进方案在各种关键点检测任务中可以提高相机姿态估计和场景几何估计的效果。

最佳论文奖：

Swin Transformer: Hierarchical Vision Transformer using Shifted Window. 这篇论文的共同第一作者 Ze Liu、 Yutong Lin、 Yue Cao、 Han Hu 来自微软亚洲研究院、中国科技大学、西安交通大学。 Swin Transformer 是一种通用网络结构，其核心思想是：利用一个滑动窗机制来实现局部区域之间的信息传播、只在局部窗口区域内应用自注意力从而降低计算复杂度、采用分层次的网络结构设计来学习多尺度的特征表示等。实验结果验证了 Swin Transformer 在图像分类、物体检测和语义分割任务上都取得了比卷积神经网络显著更好的结果，从而验证了 Transformer 在计算机视觉领域的巨大潜力。

大会也宣布了 PAMI-TC 奖的获得者：来自加州伯克利大学的 Ruzena Bajcsy 获得了 2021 Rosenfeld Lifetime Achievement Award，来自加州理工学院的 Pietro Perona 和来自 INRIA 的 Cordelia Schmid 荣获了 2021 PAMI Distinguished Researcher Award，The KITTI Vision Benchmark Suite 团队和 The Detectron object detection and segmentation software 团队荣获了 2021 Everingham prize。最后也对三篇发表于十年前 ICCV 2011 上的工作颁发了 2021 Helmholtz Prize。

1 ORB: An efficient alternative to SIFT or SURF, ICCV 2011. 这篇论文的第一作者 Ethan Rublee 来自机器人实验室 Willow Garage。 ORB 在 2011 年论文发表的时候是一种高效的视觉任务描述子(descriptor)且比 SIFT 的速度快 2 个数量级。实验结果验证了 ORB 在当时移动端(运行在手机上)的物体检测和区块跟踪等任务上都表现很好。

2 HMDB: A large video database for human motion recognition, ICCV-2011. 这篇论文的第一作者 Hildegard Kuehne 来自德国的卡尔斯鲁厄理工学院(KIT)且目前在德国的法兰克福大学工作。这篇论文的贡献是构建了在当时最大的用于动作识别的视频数据集 HMDB, HMDB 数据集包括接近 7000 段视频片段，定义了 51 个不同的语义类别标签。

3 DTAM: Dense tracking and mapping in real-time, ICCV 2021. 这篇论文的第一作者 Richard Newcombe 来自帝国理工大学，目前 Facebook Reality Labs 工作。DTAM 在当时是一个用于实时相机跟踪和三维重建的算法。DTAM 不依赖于当时基于复杂特征提取算法的特征匹配方法，而是直接利用 RGB 信息对图像原始像素做匹配。

## 四、 精彩报告选介

大会内容精彩纷呈，由于篇幅受限，这里仅仅选取其中最有代表性的几位计算机视觉专家在“A discussion about deep learning vs classical methods and their roles in computer vision”论坛上的精彩分享为例作详细地介绍。

来自华盛顿大学的 Richard Szeliski 分享了他编写的最新版的教材 Computer Vision: Algorithms and Applications(第二版)中关于会议主题的讨论。另外，Richard Szeliski 也分享了深度学习方法已经被华盛顿大学、密歇根大学、麻省理工学院的计算机视觉课程所囊括。然后， Richard Szeliski 选取了多篇关于应用深度学习方法解决传统计算机视觉任务的工作，比如三维重建、深度估计、相机参数估计、光流估计、图像合成等。最后， Richard Szeliski 给出了他对这个问题的建议：深度学习仅仅是很多有效工具中的一种方法但

是不是唯一方法、深度学习并不适合依赖几何/光学/物理约束的问题和建模不确定性的问题、深度学习方法在训练数据不够的情况下很难具有较好的泛化能力、基于深度学习的特征提取速度在有些任务上比传统方法快很多。

来自加州伯克利大学的 Jitendra Malik 从 1973 年的诺贝尔生理学奖获得者 Nikolaas Tinbergen 提出的关于研究动物行为的四个经典问题(注解：(1) adaptive function and (2) phylogenetic history; and the proximate explanations, in particular the (3) underlying physiological mechanisms and (4) ontogenetic developmental history https://en.wikipedia.org/wiki/Tinbergen's_four_questions )出发，用人类如何估计深度作为例子来解释计算机视觉任务。 Jitendra Malik 认为未来应该研究如何设计支持小数据规模而非大数据规模、利用自然信号而非人工标注作为监督信号的算法。最后， Jitendra Malik 也鼓励计算机视觉领域的研究人员去学习人类思想史并分享了自己如何汲取计算机视觉领域的思想史来做出更好研究工作的经验。

来自伊利诺伊大学厄巴纳-香槟分校的 Svetlana Lazebnik 从回顾 AlexNet 在 ECCV 2012 ImageNet 挑战上取得了非常好的分类结果开始，提到了当时她和周围的研究员最开始对深度学习方法的推广(因为 GPUs 不普及)保持怀疑态度。随着更方便的深度学习框架 Caffe 的出现，在 2013 年、2014 年、2015 年发表深度学习方面的顶会论文非常容易。例如，当时只需要训练一个 AlexNet 或者把 AlexNet 抽取的特征用于各种各样的计算机视觉任务就可以发一篇顶会论文。现在由于很多 low-hanging fruits 快都被摘完了，所以深度学习研究也进入了一个瓶颈期。最近随着 GPT 和 Transformer 的成功应用，Svetlana Lazebnik 也表示她对深度学习发展没有之前那么悲观。

来自加州伯克利大学的 Alexei Efros 回顾了自己早在 2012 年就访问纽约大学研究深度学习的计算机视觉实验室。Alexei Efros 也坦言自己不喜欢追随热点去研究网络结构，而是喜欢思考计算机视觉问题的本质。Alexei Efros 也回顾了早在 2015 年自己就带领团队开始专注于自监督学习并发表了多篇顶级论文。

## 五、 总结与展望

通过深度参与今年的 ICCV 大会，我们不仅关注到 Transformer 快速席卷各种计算机视觉领域各个重要任务并有替代卷积神经网络的趋势，也关注到最前沿的专家们分享了关于计算机视觉领域的研究工作目前面临的重要挑战与局限性，同时还关注到专家们对深度学习方法和传统方法之间关系的热烈讨论，以及工业界专家们分享如何将计算机视觉领域的前沿技术落地到自动驾驶系统中等。笔者相信未来计算机视觉领域的研究热点应该会包括：如何推动计算机视觉领域的大模型预训练、如何构建视觉大模型预训练所需要的大规模高质量的数据集、如何提高算法模型在开放世界中的泛化能力等。笔者认为回答好这些问题将大大推动计算机视觉向更通用的智能迈进。

责任编委 **魏秀参**

元玉慧

微软亚洲研究院视觉计算组研究员。主要研究方向为语义分割和物体检测。
Email: yuhui.yuan@microsoft.com

顶会观察

# ACM Multimedia 2021

电子科技大学 徐行

ACM International Conference on Multimedia (简称 ACM Multimedia) 是世界多媒体领域最重要的顶级会议，也是中国计算机学会推荐的该领域唯一的 A 类国际学术会议。自 1993 年首次召开以来， ACM Multimedia 已成功举办了二十八届。2021 年第 29 届 ACM Multimedia 会议已于 2021 年 10 月 21 日至 25 日在中国成都举办，此次会议是第二次在中国、首次在大西南地区举办。

## 一、 论文录用情况

本次会议吸引了来自中国、美国、德国、澳大利亚、瑞典、法国、日本等 19 个国家和地区约 1100 余名人员注册参会，其中 ACM/SIGMM Member 有 396 名，Non-ACM/SIGMM Member 有 377 名，Student-ACM 有 109 名，Student-Non ACM 有 124 名；共收到了来自近 40 个国家共 2544 篇投稿论文，有效投稿 1942 篇，共有 219 名 Area Chair 以及 1233 名审稿人，最终录用了 542 篇。

本次 ACM Multimedia 内容覆盖了 Engaging Users with Multimedia、Experience、Multimedia Systems 和 Understanding Multimedia Content 四个主题。在被录用的论文中，占比最多的是 Deep Learning for Multimedia，占比 36%；其次是 Vision and Language，占比 14%；排名第三的是 Emerging Multimedia Applications，占比 8%；另外 Multimedia Search and Recommendation 和 Media Interpretation 均占比为 7%； Multimodal Fusion and Embedding 占比为 6%； Multimodal Analysis and Description 占比为 5%；Emotional and Social Signals in Multimedia 和 Multimedia HCI and Quality of Experience 均只占 3%，而 Multimedia Art， Entertainment and Culture 占比只有 2%。

## 二、 会议日程概要

根据国内外疫情形势，本次会议采用独特的线上-线下混合会议模式。来自国内的参会者大多通过线下参会的方式来到会议现场参会，而在海外的参会者或者国内中高风险地区的参会者则通过线上参会方式参与会议。本次会议也针对性地提供了不同的线上参会软件和直播平台，方便线上参会者参与会议并进行在线交流。为期 5 天的大会学术报告共进行了 6 场 Keynote、13 场 Workshop、8 场 Tutorial、2 场 Panel 和 37 场 Session，充分展现了多媒体领域全球最新研究成果和前沿动态，以及对该领域未来发展方向的创新引领。会议的第 1 和第 5 个会议日安排了 Workshop 和 Tutorial。第 2 到 4 个会议日则安排了主会期间的各个主旨报告，口头发表 session，特定主题 session 以及海报展示环节。各组织者，演讲者以及论文发表者通过现场参会或在线报告的方式介绍论文内容，参会者通过线下参会或加入线上会议的方式听报告或交流。本次会议每天中场休息和中午休息期间为现场参会者提供了茶歇和冷餐食，方便参会者的现场交流。从现场来看，交流氛围十分热烈，大家也十分珍惜疫情之下难得的线下交流机会。

会议最引人瞩目的四项多媒体领域大奖在大会晚

宴 Banquet 的热烈氛围中揭晓，包括最佳演示奖“ViDA-MAN: Visual Dialog with Digital Humans”、最佳开源奖“X-modaler: A Versatile and High-performance Codebase for Cross-modal Analytics”、最佳学生论文奖“aBio: Active Bi-Olfactory Display Using Subwoofers for Virtual Reality”，及最佳论文奖“Video Background Music Generation with Controllable Music Transformer”。

## 三、 会议主旨演讲

今年大会主旨演讲邀请到了多媒体领域世界级的专家学者作为重量级嘉宾，分别带来了涵盖视频编码技术、多模态模型与学习、AI 教育等主题的 6 场演讲，从多个角度见证了多媒体技术的巨大潜力。

中国工程院院士，北京大学教授，鹏城实验室主任高文院士带来了“面向机器的视频编码技术”报告，剖析了传统视频编码技术在智能机器时代的缺陷，提出了社会发展对新一代针对智能机器的视频编码技术的迫切需要，并介绍了鹏城实验室在此方向上的最新实践与成果。密歇根大学安娜堡分校的 H. V. Jagadish 教授带来了“语义媒体转换”主题的演讲，分析了智能机器时代不同媒介的多媒体信息相互转换越发灵活便利的最新发展趋势，介绍了通过结构化的数据表来高保真地转换多媒体信息的最新技术。

法国 INRIA 和 Google 的研究员 Cordelia Schmid 带来了针对多模态视频的大规模训练的最新研究发展，介绍了 VideoBert，ViViT 等最新基于 Transformer 架构的多模态神经网络模型，并展示了跨模态监督训练范式的强大潜力。蚂蚁集团副总裁周靖人博士介绍了阿里巴巴和蚂蚁集团研发的最新大规模多模态预训练模型 M6 及其在大量下游学习任务上的卓越表现，以及一种最新的基于预训练对抗生成网络的图像编辑技术。

北卡罗莱纳州立大学的 James Lester 教授分享了 AI 技术对社会带来的机遇与挑战，介绍了计算机视觉、自然语言处理、机器学习等 AI 技术给人类教育事业带来的变革与发展。腾讯 AI 实验室、腾讯 Robotics X 实验的主任张正友博士提出了“虚实集成世界”的新概念，介绍了虚拟现实、增强现实等技术带来的物理-虚拟世界融合的未来发展方向，以及其中涉及的人工智能、多媒体关键技术。

## 四、会议亮点 Sessions

Best Paper Session: 本次会议的 Best Paper Session 安排在主会第二天上午开场进行，共有 5 篇最佳论文提名的论文，内容涉及 Co-Speech 手势生成研究[1]、视频背景音乐生成[2]、场景文字识别[3]、多模态语音增强[4]、虚拟现实[5]，分别来自马里兰大学[1]、北京航空航天大学[2]、中国科学院信息工程研究所[3]、赫尔辛基大学[4]、台湾大学[5]。报告者通过线上或者现场的方式对论文内容进行介绍。最终，北京航空航天大学“Video Background Music Generation with Controllable Music Transformer”胜出获得了最佳论文奖。在该篇论文中，作者创造性地提出了视频背景音乐生成问题，能够依据视频播放内容自动生成合适的背景音乐。此外，该团队成员还在现场针对视频内容生成的音乐进行了实时演奏，体现了丰富的多媒体交互形式和视听觉体验。

Brave New Ideas Session: 此 Session 一直是作为 ACM Multimedia 会议的代表性的会议议程，探索未来多媒体领域最具研究和应用潜力的主题。本次 session 共包含 5 篇论文，来自中国科学技术大学[6]，纽约州立大学布法罗分校[7]，上海交通大学[8]，清华大学[9]和香港中文大学[10]。内容涉及图片质量评估、视频运动公式学习、用于抑郁症检测的音频编码、多媒体计算的新范式以及基于元宇宙(Metaverse)构建的虚拟大学原型。其中来自香港中文大学的 Haihan Duan 通过在元宇宙构建一个大学原型为例，介绍了元宇宙的概念和雏形，让大家对未来的元宇宙有了一个初步的认识。

Industrial Track Session: ACM Multimedia 会议自去年开始引入此 session 的形式，旨在介绍工业界所研究的前沿热点多媒体问题及产业应用情况，同时加强多媒体领域工业和学术领域研究者的交流。此次 session 共包含两个特邀报告和四篇口头发表论文。特邀报告中，来自阿里巴巴的 Xiansheng Hua 通过在线

的方式介绍了阿里巴巴公司内部中所使用的视觉搜索技术及典型应用案例；来自 OPPO 的 Yandong Guo 以风趣成熟的演讲风格给大家介绍了在 Metaverse 时代，针对智能移动设备的多媒体技术，重点对基于知识图谱的大规模视觉理解进行介绍，让大家对如何针对大规模多媒体数据构建知识图谱有了一个清晰的认识。四篇口头发表论文分别来自 OPPO、微软亚洲研究院、阿里巴巴、香港科技大学、京东 AI Lab 以及腾讯。论文内容涵盖多媒体算法如何应用在具体的工业场景中，例如，如何大规模视频数据自监督，对于语音识别的评分机制，通过知识蒸馏来提高行人检索精度和对遮挡文本的重建方法等。

## 五、总结与展望

在 ACM Multimedia 会议的线上-线下混合举办期间，还有多个丰富多彩的 sessions，例如 Workshops、Tutorials、Art & Culture、Best Demo、Open Source Competition、 Posters、艺术展、赞助企业展、午餐会等。笔者的理解仅代表个人观点，如有偏差还请各位读者指出。2022 年的 ACM Multimedia 会议将于 10 月 10 日至 14 日在葡萄牙里斯本举行，期待疫情显著缓解，全世界各地多媒体领域研究者能通过线下参会的方式相聚，共同探讨多媒体领域未来的发展。

推荐委员 金鑫　责任编委 王金甲

## 参考文献

[1] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, Dinesh Manocha.Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. 2021. In ACM Multimedia (ACM MM), Chengdu, China, 20-24 October, 2021, 2027—2036.

[2] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, Shuicheng Yan. Video Background Music Generation with Controllable Music Transformer. ACM MM 2021, 2037—2045.

[3] Zhi Qiao, Yu Zhou, Jin Wei, Wei Wang, Yuan Zhang, Ning Jiang, Hongbin Wang, Weiping Wang. PIMNet: a parallel, iterative and mimicking network for scene text recognition. ACM MM 2021, 2046—2055.

[4] Abhishek Kumar, Tristan Braud, Lik Hang Lee, Pan Hui.Theophany: Multimodal speech augmentation in instantaneous privacy channels. ACM MM 2021, 2056—2064.

[5] You-Yang Hu, Yao-Fu Jan, Kuan-Wei Tseng, You-Shin Tsai, Hung-Ming Sung, Jin-Yao Lin, Yi-Ping Hung. aBio: Active Bi-Olfactory Display Using Subwoofers for Virtual Reality. ACM MM 2021, 2065—2073.

[6] Zhu Y, Ma H, Peng J, et al. Recycling Discriminator: Towards Opinion-Unaware Image Quality Assessment Using Wasserstein GAN. ACM MM 2021.

[7] Song L, Liu S, Liu C, et al. Learning Kinematic Formulas from Multiple View Videos. ACM MM 2021.

[8] Zhang P, Wu M, Dinkel H, et al. Depa: Self-supervised audio embedding for depression detection. ACM MM 2021.

[9] Kang Z, Li J, Zhu L, et al. Retinomorphic Sensing: A Novel Paradigm for Future Multimedia Computing. ACM MM 2021.

[10] Duan H, Li J, Fan S, et al. Metaverse for social good: A university campus prototype. ACM MM 2021.

**徐行**

电子科技大学副教授。主要研究方向为多模态信息感知与计算、跨媒体智能分析，已在多媒体、计算机视觉及人工智能相关领域的国内外学术期刊和会议累计发表论文 100 余篇，获得包含 2017 年国际多媒体大会 ACM MM (CCF-A)最佳论文奖，2017 年国际多媒体展览会 ICME (CCF-B)的最佳会议论文铂金奖等国际会议奖项 6 项。

Email: xing.xu@uestc.edu.cn

# 北京科技大学马惠敏教授访谈

2021 年 10 月 6 日，《CCF-CV 专委简报》线上采访了北京科技大学博士生导师马惠敏教授。下面是采访实录。

马老师，您好！首先，请您分享一下您的个人学习和研究经历。

我是在北京理工大学上的研究生，专业是模式识别与智能系统。刚进入这个领域的时候还没有像现在这样火热，我记得当时学习的课程有知识工程、模式识别、神经网络以及信号处理等，这些课程到现在都非常有用。这段学习经历确实让我在整个计算机视觉和人工智能方面打下了非常坚实的基础。

我认为人如果能够在自己所学的专业上继续做同样的工作，是一件非常幸福的事情。于是在博士毕业之后，我选择在清华大学电子工程系工作，成立了三维目标认知与仿真实验室，研究如何让无人系统学习人的思维模式，实现复杂环境下的目标识别、检测和跟踪。

学生时代是一个无忧无虑的阶段，主要任务就是学习知识，但是真正到了工作研究的时候就需要不断地去挑战、去创新。2001 年博士毕业进入清华大学电子工程系成为一名承担教学科研工作的老师，自强不息、厚德载物的校训一直激励着我，其中有两个重要的转折点。

第一个转折点是 2007 年，也就是我工作的第 6 年。在那时候我发现认知心理学其实是可以和计算机视觉进行结合的。我可能是咱们国家最早开始把认知心理学引入到计算机视觉之中的一个学者。当时清华的心理系刚刚成立，我就跟心理系的傅世敏教授实验室联合组会。这段经历对于我的计算机视觉与认知心理学交叉学科研究是非常重要的，我发现心理学领域传统的量表问答方法其实是可以用计算机视觉技术来客观实现的，而计算机视觉领域的很多技术其实是可以用认知心理学理论来指导的。于是，我确定了三维图像认知这个新的研究方向，我的团队从 2007 年一直到现在都在做这个交叉研究，包括建立了全新的基于图像认知的心理测评方法和系统、在自动驾驶领域引入人的认知机理，也取得了一系列原始性创新成果。

第二个转折点是在 2019 年，北京科技大学成立了人工智能研究院，我调到北京科技大学工作，担任北京科技大学计算机与通信工程学院物联网与电子工程系主任、人工智能研究院副院长。人工智能相比于计算机视觉是一个更宽泛的领域，我组建了一个优秀的团队，希望能够在视觉心智理论和通用人工智能系统方面做一些前沿工作。

您主要从事三维图像认知与多模态学习交叉学科研究，将计算机视觉与认知心理学结合，取得了三维图像认知理论的原创性成果，建立了复杂环境仿真、视觉感知、认知学习、智能决策系统新理论和新技术，处于国际领先水平。能否分享一下您对这个领域的研究现状和未来发展情况的认识？

对于三维图像认知和多模态学习而言，它涉及到

“强人工智能”这样一个未来的方向。因为我们现在的人工智能相对而言还是在特定的条件下、不涉及人的意识和思维的智能。十年前我做计算机视觉和认知心理学交叉方向时，几乎没人看好这个方向，但是十年后，我和我的团队把它做成了最前沿、最热门的方向。我认为这个方向下一步就是引入心智理论，能够让机器学习具备人的思维模式和共情等心智能力，这一定是未来发展的一个方向。现在视觉领域的很多著名的科学家都在关注认知科学，希望通过研究人是怎么做的、怎么思考的，来让机器接近人的智能。

目前比较热门的有可解释的人工智能、小样本学习等。如果机器能真正地像人一样举一反三，将省去大量数据标注的过程。我们在去年获批了一个国家自然科学基金中电科联合基金重点项目，基于我们特有的复杂场景仿真技术、视觉心智理论、脑与认知启发的目标认知算法深入开展不确定环境下小样本认知学习及目标识别理论和方法研究。

在二十年前，我就开展了复杂环境仿真研究，因为我觉得我们能看到的这个世界的样本是很有限的，有很多特殊的样本是很难看到的，因而不可能形成数据用来学习，所以我们用仿真来生成特殊样本。仿真在国外非常重视，但国内之前的重视度不够。仿真可以为机器学习算法提供特殊的实例，甚至一些现实中不存在的实例，可以提升小样本学习模型的能力。同时研究人的视觉心智理论，把人的认知模型加进计算机视觉感知算法中，实现类人水平的机器学习，例如我们在自动驾驶视觉感知领域提出的基于 Thinking in 3D 思路的 Mono3D、MV3D、3DOP 等代表性方法。

现在小样本学习在深度学习之后已经成为一个研究热点，但是说实话，我觉得理论和方法还很不成熟，与目前已有的大数据深度学习结果相比，性能差得非常远，但这是人工智能技术未来发展的方向。

**您首个在国际上建立了图像认知心理测评智能系统，并获得了第六届吴文俊人工智能科学技术创新一等奖，能跟大家介绍一下这个系统以及建立这个系统的初衷么？**

这个初衷就是希望能够用人工智能技术早期发现高焦虑、抑郁等心理问题，帮助更多的人及时得到专业的干预和救治。心理是主观的东西，一直以来主要通过做量表、医生的晤谈方式来判断心理状态，全世界还没有一种办法能够像测量血压一样来测量心理。医生通过面对面的交互判断人的心理状态，当我看到这个现象时，我就思考为什么我们不能用科学的方法来测量心理。实际上利用认知心理学中的注意偏向理论，建立心理的视觉量化评估模型，对于计算机视觉和心理学而言，这都是一个突破性的新理论。所以当时我就把心理学的量表问题转变成了图像形式。图像领域有几个典型的任务，比如图像处理、图像识别和看图示意等，但是我们做了一个相反的任务，将文字转化为图像。就是根据一句话生成了一幅图，这在十几年前也是开创性的工作。例如：“我害怕从高处往下看”，这是一道量表的题目。我们把“从高处往下看“变成了一幅图像或者一个视频，要做得非常逼真，人害怕、不害怕是他的心理活动，从高处往下看就相当于一个刺激物，我们通过捕捉人看的时候产生的眼动、瞳孔变化等生理行为来判断他的心理活动。这样就变成了类似测血压一样客观地测量心理，于是我的团队建立了国际上首个心理图像库、首个图像认知心理测评系统。

**您的研究工作多次获奖，包括吴文俊人工智能科学技术创新一等奖、日内瓦国际发明展览会银奖、教育部技术发明奖二等奖、中国图象图形学学会技术发明一等奖等，这些获奖的工作中，哪一些或者哪几项是您最值得骄傲的？可否跟大家分享一下？**

通过仿真来获得高价值的样本，这个是教育部技术发明二等奖的内容。通过仿真获得了高价值或者复杂场

景的图像之后，我们去观察人是怎么理解这个过程，研究人的心理和心智，然后建立人的视觉认知模型，这就是吴文俊人工智能科技创新一等奖的主要内容。而同时将该模型用于解决心理状态量化评估的问题，为心理学界提供了一个新的人工智能研究手段，这也是日内瓦国际发明展览会奖的内容。同时这个视觉认知模型服务于无人系统目标认知技术，这个是中国图象图形学会的技术发明一等奖的内容。

所有获奖的成果其实是一体的，只不过在每一个点上，我们都能够做到前沿，所以我比较自豪的不是获得了几个奖，而是我们建立的这套体系，从复杂环境仿真到感知、认知、学习、决策这样的一个独特的、由心智理论指导的完整的体系，这个是我非常自豪的。过程很坎坷，刚开始很多人不能理解，我坚持了 15 年，从未放弃过，现在很多人开始关注这个研究方向了。

您已获批及申请专利十余项，其中两项专利完成了科研成果转化，能否介绍一下您这两项专利的内容及其成果转化情况？

这两项专利就是前面提到的像测血压一样测心理的心理图像库和基于图像认知心理状态测评方法和系统。我很感谢清华电子系专门负责成果转化的老师们，2016 年清华率先在国内开始做成果转化，我的这两项专利是清华大学转化的第一批专利成果之一。清华专利作价入股成立了北京清视野科技有限公司，把实验室级别的一些实验设备做成了专用的设备，已经在大学新生普测、民航飞行员筛查、员工体检、全国人大机关、国家民政部、企事业单位、医院使用，实现了便捷、客观、准确的心理状态评估。很开心能够用自己的科研成果服务人民健康，我虽然不是心理专业的，但我花了很多精力去学习一个新的领域，我觉得这一切是值得的。

在“未来简说”中，您做了一个报告，“让机器看懂抑郁症患者的悲伤”，您能跟大家分享一下这个报告的主要观点及内容么？

这就是我之前提到的初衷，能像测血压一样简单量化地测我们的心理状态，能够让每个人很方便地对自己的心理状态有一个正确的评估，在评估的基础上，来享受定制的干预，希望能够帮助更多焦虑、抑郁的人，让全世界的人都能获得人工智能的福祉。我们这个设备，自从研发出来以后，有很多家庭和我们联系，希望能帮到他们的孩子。我当时做这件事情，把它用在医学上，也是因为我带的学生里有抑郁症患者，他们真的很不容易，我希望能够用科学的方法帮到他们，所以希望能够“用视觉的方法，让机器看懂抑郁症患者的悲伤”。抑郁人群的心理状态，包括引发他们抑郁状态的定向因素，我们都能够测量，能够真正地走到他们的内心深处，给他们做科学的干预，这就是当时的一个主要观点和内容。也是前面提到的吴文俊人工智能奖和两个转化的专利的内容。

您作为负责人承担了国家重点研发计划子课题、国家自然科学基金、专项重点基金、国际国内企业合作等 30 余项科研项目，请问您承担这些项目后的主要感悟是什么？在承担国家级项目和企业合作项目时，您觉得最大的区别是什么？您能否为刚走向科研岗位、即将承担这样一些项目的年轻人提出一些建议？

在纵向项目方面，我觉得关键的一点是“千万别犯懒”，即在做科研过程中一定要让我们的思维活跃。例如这次重点研发计划中我承担的子课题，把心智理论引入到小样本学习中、引入到自动驾驶视觉感知与决策中，就是一个全新的方法。另外，我觉得应该保持一种敢为人先的状态和精神。具体来说，国家级项目的目的是解决国家的重大需求、卡脖子问题，所以一定要知道问题的关键在哪儿，要打准它而且有办法解决它，如果达不到这个水平就很难拿到国家级项目。

总而言之，一定要让自己思维活跃，不要停，清楚地知道关键问题在什么地方，而且还要有自己的钥匙来解决这个问题，这个钥匙就是你自己能够创立的、引领的方向，我觉得这是国家级项目里最根本的东西。不仅

要有理论支持，还要有理论创新和积累。

横向合作项目实际上是在某个小点上，企业希望能够解决的某一个实际问题。企业会提供明确的细则和具体某个点上的需求。很多时候企业并不关注你的那些创新思想，它只关心能不能达到自己的要求，不过现在企业也开始越来越关心核心技术了。企业合作需要比较强的工程能力，当然国家级项目也需要强的工程能力，不然你很难做出好的算法和好的结果。其次，企业合作项目的指标和需求要讨论得非常清楚，不然就容易出现指标完成不了的情况，一定要科学地制定指标。企业肯定希望你完成 100%，但如果目前科技水平大家都只能达到 90%的时候，要坚持自己的想法。

您做了一个具有颠覆性的原创性工作，做原创性的工作是需要勇气和底气的，您能给要做原创性工作的学者一些建议么？

首先，要知道如何才能有原创性的工作。我认为一定要有很多在不同学术领域的好朋友，这样在讨论中能碰撞出很多很有意思的事情，然后自己去深入地思考。现在基金委提倡原创，真正原创的东西会得到一些支持，但是在我们那个年代原创性工作是非常难的，就像我们老系主任说的：十年前没有人看好你这个方向，十年后你把它做成了最热的方向。其实原创性工作还需要在自己的领域有深厚的积累，如果在自己领域都不能做到比较高的水平，原创很难保证扎实。所以，当时大家听不懂我在做什么时，我就告诉他们：我有一只左手，一只右手，左手就是以前积累下来的“仿真+识别”，右手是“图像认知心理学“，右手拿不到钱，我也要做这件事情，我拿我自己的左手养我的右手，用我自己的已经有的科研来养我的新方向，因为它在未来会成为我的左手，到那时我还会有一只新的右手要养”。也就是说我们需要具备能够养活原创性工作的能力。在头几年，我申请什么都申请不到，包括国家自然科学基金，因为视觉领域的人根本看不懂你在说什么，心理领域的人说你这个东西根本不是我们心理领域的。发论文也很难。但我认准了这个是一个好方向，我就会去做，虽然经历了很多坎坷。

能否介绍一下您的研究团队，以及您是如何管理您的团队的？

我在清华大学的团队很交叉，周鹏老师是从国外引进回来的视听觉认知科学的，他的很多观点对我的启发很大，我们还有一支优秀的博士生和硕士生队伍。我在北京科技大学的团队有 7 位老师，还有 20 多名有扎实计算机基础的研究生。关于团队的管理，我们团队会分成多个小团队，每一个人和每个过程我都清楚，我自己也不可能面面俱到，年轻老师、博士和硕士和我形成了梯队。

另外我的管理理念是以身作则。习主席 4 月 19 日在清华说过要做大先生，为学、为事、为人的示范。“为学”就是要保持自己的学术领先，学术上的每个点我都会弄透，每个学生都能和我进行学术上的探讨，我也会提供很多有建设性的意见。“为事”就是对待工作和研究的态度，学者要能站在科研的最前沿，要严谨地对待科研，我和我的学生说过，我承担的每一件事情都要把它做到最好，我的学生都非常地认真和严谨，这也是一种风格的传承。“为人”方面，我本身也有很多社会工作，这也是我自己一直以来坚守的，希望在我力所能及的情况下，能够更多地服务社会，利用自己的知识和力量来让这个世界更美好，我的学生们和我团队的老师们也都承担了很多相应的社会工作。

论文是科研成果的主要表现方式之一，您是如何规划和看待论文发表的？您又是如何激励团队青年教师及研究生们发表高水平论文的？

对这个问题我还真有自己的观点。我并不是一个高产作家。刚开始大家不怎么发论文的时候，我就在扎扎实实地做研究，每年发表几篇代表性的论文，后来大家开始批量发论文的时候，我仍然是每年几篇论文，我从来没有追求过论文的数量。我认为高水平论文实际上取

决于科研的水平，是需要积淀的。

我会隔一段时间与团队每个老师和学生进行交流，例如我会让他们画关于最近研究的思路图。如果你直接跟同学和老师们说，“你要发表高水平论文”，说了等于白说，因为高水平论文需要有高水平才可以发出。在高校里讨论特别重要，与其说我们激励学生不如说是“助力学生”。我发现青年老师和学生们有时候不能站在一个系统级上看待问题，很难有特别好的思路，换句话说视野不够高。所以我们需要讨论，在讨论过程中间把高水平思想注入进去，这样才能助力学生，发出高水平的论文。

我们注意到从 2002 年至今，您一直承担研究生课程“数字图像技术及应用”的教学工作，从 2002 年到现在，随着深度学习的发展，可以说数字图像技术发生了非常大的革新和进展，那请问您对这门课程的规划都发生了哪些变化？您是如何将最新技术融入教学过程的呢?

“数字图像技术及应用”这门课非常特殊，我从 2002 年在清华开始上这门课，而国内其他高校的课程叫“数字图像处理”。开这门课的时候我就一直在思考一个问题，就是如何跨越数字图像理论和应用之间的鸿沟。因为我发现很多人学了很多种技术，但是不知道怎么去应用，因为他们没有形成系统级的概念，这制约了他们的发挥，于是产生了“学了一堆算法，但到用的时候不知道用哪个”的现象。我觉得这是一个很大的问题，所以当时我开这门课的时候就是立足于去建立一个数字图像技术的系统级的概念，然后再去解决实际的应用问题。

我自己也做过很多横向和纵向的课题，所以我做了一个梳理，从图像处理到图像识别、到应用系统三个层级，从基础理论上把视觉和认知加进来（当然那个时候还没有加认知，主要是加入视觉特性，这几年已经引入了认知理论）。这个课程是清华大学的校级公开课，特别受欢迎。我主要通过问题和技术来引导，探讨我们面临的这个世界上关于数字图像关键的技术有哪些，然后让同学们直接面对问题，然后从问题来讲理论和算法。通过这种方式教学，学生的能力有很大提升。

当时这门课是给研究生开设的校级公开课，这个课难度系数很高。首先，这个课程的应用面特别广，老师必须对图像图形领域熟悉。第二，这个课涉及的内容日新月异，所以授课的内容需要不停地修改。

最早的时候只是加入了视觉的特性，通过了解视觉的高层特性和低层特性来解决图像识别的特性到底是什么；后来把认知也引了进来。当然深度学习之后，技术本身就发生变化，技术演变包括之前的生物特征识别、在线检测、视频监控这些经典应用系统，我们还加进了自动驾驶等内容，这也是我们自己的研究方向。总而言之，就是通过技术引导解决问题的思路。这门课的内容到目前为止每年都还在更新中。

您从清华到北科大，对研究工作有什么影响，个人又有什么收获呢？

从清华到北科大我确实下了一个很大的决心。做科研，尤其是做方向，不像一篇论文一篇论文地产出，它是一个体系，就像前面提到的从仿真到感知、认知、学习和决策。因此个人力量做这个工作是有困难的，尽管说我们在理论上做到了创新，但是真正想在一个领域里做好、做深，没有团队是不行的。我来到北科大以后，无论是我的教师团队还是我的研究生团队，都和我清华的团队形成了一个很好的协同。而且北科大这边对我们方向非常地重视和支持，从学校校长、书记到学院院长、书记等都给我们全力的支持，所以我是非常感谢的。让我的平台更大了，很多大的成果是需要大舞台的。

作为北京科技大学计算机与通信工程学院物联网与电子工程系主任、人工智能研究院副院长，您觉得您的使命是什么？

我很有使命感，我觉得我最大的愿望就是我们的科研成果能造福全世界，这是我一直的想法。所以我在北京科技大学组建团队，最希望的事情就是能够让我们这些志同道合的各个年龄段的老师们团结在一起，把我们整个的从智能理论到关键技术、到代表性应用做实。希望我们中国自主产权的最前沿的智能系统能让全世界人来用。简而言之，我的期待就是用我们的知识推动社会的进步。

**您现任中国图象图形学学会副理事长兼秘书长，工作繁重，您是如何协调学会和研究工作的？学会工作和研究工作关系是怎样的？**

这个对时间的要求很高。我基本上把时间分成三段：早上 5:00-7:00，这个时间段是我自己的时间，这是一个非常好的思考的时间，万籁俱寂，我会去做我的学术型思考和一些阅读的工作，保持自己学术上不要掉队；7:30-17:30 是我科研的时间段，早上有些思路要落地，或者要和学生讨论，或者写一些 proposal、建议书，或者要论证一些东西以及推导一些东西，或者开工作上的会议，都会在这个阶段来完成。基本上我是不吃午饭的，在这个时间段，外面是刮风还是下雨，我统统都不知道。以前中午时间会有一些午餐会，或者学会的事情，不过这两年会休息半个小时。为什么到 17:30 呢，因为我还是一个母亲，我对孩子的教育很重视，17:30 我会去接我儿子，把晚饭的时间段留给家人。20:00 以后，我会处理一些临时的讨论、学会或其他的事情，有的也会约晚上 23: 00 的会议。晚上一般 12 点会去休息，这样的节奏我已经习惯了。

**作为女性科研工作者，请问您是如何协调家庭和工作的？**

在家庭方面，我很重视孩子的教育。我认为孩子的成长是不可逆的，所以我会亲自带孩子，就像我以身作则带学生一样。下班后，我会陪他玩、给他讲故事，再累我也会在他睡觉之前给他读书。我的很多课都是连三的，每次上完课都很累，话都不想说，但是我回到家，晚上还是会陪他读书。在周末的时候，如果全家人出去游玩，我都会陪同到达游玩地点，让他们感受到我的参与和陪伴，尽管我可能大概率会自己呆在车上工作。

工作的时间认真工作，全身心投入。能够陪伴孩子的时间，我还是会尽可能地去陪伴孩子。我尽量做到能保持平衡。

**作为北京市“三八红旗奖章”获得者，您能跟大家分享一下您的获奖情况及获奖感言么？**

当时因为我是学科的带头人，在这个领域里有些代表性的成果，同时学会的工作确实做得比较投入，也比较有成效，所以获得了北京市的“三八红旗奖章”。我想跟大家分享的是努力做最好的自己，实际上是能让自己很快乐的，同时如果能够帮助别人，那我觉得会更幸福。我觉得这是个社会责任。

**如果吐露研究工作者的心声，您最想说的是什么？**

自强不息，厚德载物，努力作一名“大先生”。

责任编委 赵振兵 余烨

# 马惠敏

马惠敏教授，博士生导师，北京市“三八红旗奖章”获得者。2001 年博士毕业后在清华大学电子工程系承担教学科研工作，担任三维图像认知与仿真实验室负责人，2019 年担任北京科技大学计算机与通信工程学院物联网与电子工程系主任、人工智能研究院副院长，现任中国图象图形学学会副理事长兼秘书长、CCF-CV 专委执行委员。从事三维图像认知与多模态学习交叉学科研究，将计算机视觉与认知心理学结合，取得了三维图像认知理论的原创性成果，建立了复杂环境仿真、视觉感知、认知学习、智能决策系统新理论和新技术，处于国际领先水平。首次在国际上建立了图像认知心理测评智能系统，2016 年获得吴文俊人工智能科技创新一等奖（排名第一），2017 年获得日内瓦国际发明展览会银奖，教育部鉴定为“原始性创新，达到国际领先水平”；提出的基于 GPU 的高效能复杂环境仿真方法及应用，2017 年获得教育部技术发明奖二等奖（排名第一）；提出的复杂环境中三维目标认知方法，2015-2017 年连续在国际最大的自动驾驶数据集（KITTI）评测中获得第一名，2018 年在驾驶员状态预测国际数据集（Brain4Cars）上获得最好的成绩，属于国际领先水平，2020 年获得中国图象图形学学会技术发明一等奖（排名第一）。作为通讯作者在 TPAMI、TIP、TITS、PR、CVPR、NIPS、ICCV、ICIP 等发表论文 100 余篇，单篇他引超过 1000 次。作为负责人承担了国家重点研发计划子课题、国家自然科学基金、专项重点基金、国际国内企业合作等 30 余项科研项目，获批及申请专利十余项，两项专利完成了科研成果转化，实现了从基础理论、核心技术到应用的突破性进展。

# 委员好消息

✪ 2021 年 3 月下旬，陕西省科学技术厅公布了 2021 年陕西省杰出青年科学基金资助项目获得者名单，全省共有 50 名青年学者入选，CCF-CV 专委会委员、西北工业大学程塨研究员获得陕西省杰出青年科学基金资助。

✪ 2021 年 4 月 8 日，清华大学 AMiner 发布了人工智能全球最具影响力学者 AI 2000 榜单，CCF-CV 专委会委员、西北工业大学聂飞平教授、南京大学李武军教授入选经典人工智能领域最具影响力学者榜单，CCF-CV 专委会委员另有 13 人次入选最具影响力学者提名榜单：西北工业大学聂飞平教授入选数据挖掘领域最具影响力学者提名榜单，南开大学程明明教授、中国科学院深圳先进技术研究院乔宇研究员入选计算机视觉领域最具影响力学者提名榜单，山东大学聂礼强教授入选信息检索与推荐领域最具影响力学者提名榜单，上海交通大学卢策吾研究员入选计算机图形学领域最具影响力学者提名榜单，厦门大学纪荣嵘教授、山东大学聂礼强教授、复旦大学姜育刚教授、中国科学院计算技术研究所张勇东研究员、中山大学林倞教授、中国科学院大学黄庆明教授、西安交通大学孟德宇教授入选多媒体领域最具影响力学者提名榜单，北京航空航天大学刘偲副教授入选人工智能全球女性榜单。

✪ 2021 年 5 月 10 日，Guide2Research 网站公布了 2021 年世界顶尖 1000 名计算机科学家排名，53 位中国学者上榜，CCF-CV 专委会 4 位委员上榜：中国科学院自动化研究所谭铁牛院士、西北工业大学聂飞平教授、中国科学院计算技术研究所陈熙霖研究员、中国科学院计算技术研究所山世光研究员。

✪ 2021 年 5 月，CCF-CV 专委会委员、西北工业大学韩军伟教授和程塨教授等合著的论文 A Unified Metric Learning-Based Framework for Co-saliency Detection 获得多媒体领域顶级国际期刊《IEEE Transactions on Circuits and Systems for Video Technology》最佳论文奖。

✪ 2021 年 7 月 17 日，在布鲁塞尔举行的 IGARSS (地球科学与遥感大会) 上，CCF-CV 专委会委员、西北工业大学韩军伟教授和程塨教授等合著的论文 Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images 获得 2021 年度 IEEE 地球科学与遥感学会最有影响力论文奖。

✪ 2021 年 10 月 22 日，CCF-CV 专委会委员、北京航空航天大学刘偲副教授等指导完成的论文 Video Background Music Generation with Controllable Music Transformer 荣获 ACM Multimedia 2021 最佳论文奖。

✪ 2021 年 10 月 28 日，陕西省教育厅公示了 2021 年陕西省优秀博士学位论文 100 篇，由 CCF-CV 专委会委员、西北工业大学张艳宁教授指导的《基于稀疏优化的图像复原方法研究》以及由 CCF-CV 专委会委员、西安电子科技大学邓成教授指导的《面向多媒体最近邻检索的深度紧致编码学习》入选。

✪ 2021 年 10 月 28 日，四川省科技厅公示了 2021 年度四川省科学技术奖拟奖项目，由 CCF-CV 专委会委员、中国科学院空天信息创新研究院孙显等合作完成的“高分辨率复杂 SAR 图像场景建模理论与解译方法”拟授自然科学二等奖。

✪ 2021 年 11 月 3 日，2020 年度国家科学技术奖励名单公布，CCF-CV 专委会常务委员、东南大学耿

新教授等参与的“面向多义性对象的新型机器学习理论与方法”获国家自然科学二等奖，CCF-CV 专委会委员、北京百度网讯科技有限公司总监丁二锐等参与的“知识增强的跨模态语义理解关键技术及应用”获国家技术发明二等奖。

✪ 2021 年 11 月 11 日，山西省科技厅公示了 2021 年度山西省科学技术奖评审委员会评审结果，共有 208 个项目（企业）通过评审委员会评审，其中一等奖 19 项（含科技合作奖 2 项）、二等奖 91 项（含科技合作奖 2 项）、三等奖 91 项，企业技术创新奖 7 项，CCF-CV 专委会委员、太原理工大学赵涓涓教授参与完成的基于 CT 影像的肺癌长时程演变规律及智能诊断模型研究拟授二等奖。

✪ 2021 年 11 月 16 日，科睿唯安发布了 2021 年度全球“高被引科学家”名单，来自全球 70 多个国家和地区的 6602 人次入选高被引科学家名单，CCF-CV 专委会 6 位委员入选：杭州电子科技大学俞俊教授、西北工业大学程塨研究员、韩军伟教授、聂飞平教授、王琦教授、中国科学院西安光学精密机械研究所卢孝强研究员。

✪ 2021 年 11 月 22 日，陕西省科技厅公布了 2021 年陕西省创新人才推进计划入选人员名单，CCF-CV 专委会委员、西北工业大学孙瑾秋教授入选中青年科技创新领军人才。

✪ 2021 年 11 月 24 日，IEEE Fellow of Class 2022 公布，CCF-CV 专委会委员、西北工业大学韩军伟教授（因在视觉显著性检测与图像理解领域的贡献）、中科院计算所山世光研究员（因在视觉信号处理与识别领域的贡献）和微软亚洲研究院王井东研究员（因在视觉内容理解与检索领域的贡献）当选。

✪ 2021 年 12 月 2 日，江苏省科学技术厅公示了 2021 年度江苏省科学技术奖综合评审结果， CCF-CV 专委会 6 位委员完成的 2 个项目进入公示名单：南京信息工程大学刘青山教授、袁晓彤教授、中科院自动化所程健研究员等共同完成的高维视觉大数据的紧致化表示理论与方法、哈尔滨工业大学左旺孟教授、徐勇教授和南京邮电大学高广谓副研究员等共同完成的图像复原与稳健识别的理论与方法拟授一等奖。

✪ 2021 年 12 月 3 日，中国人工智能学会公示了 2021 年度中国人工智能学会会士增选结果， CCF-CV 专委会副主任、中科院自动化所王亮研究员和 CCV-CV 专委会委员、重庆邮电大学校长高新波教授入选。

✪ 2021 年 12 月 4 日，中国图象图形学学会发布了 2021 年度中国图象图形学学会石青云女科学家奖评选结果公告，CCF-CV 专委会委员、南京理工大学张姗姗教授入选。

✪ 2021 年 12 月 4 日，中国图象图形学学会发布了当选 2021 年度中国图象图形学学会会士名单，CCF-CV 专委会 5 位委员当选：北京大学林宙辰教授、彭宇新教授、北京科技大学马惠敏教授、西安交通大学薛建儒教授、西北工业大学张艳宁教授。

✪ 2021 年 12 月 4 日，中国图象图形学学会发布了 2021 年度优秀博士学位论文奖评选结果公告，CCF-CV 专委会委员指导的 5 篇博士学位论文入选：中科院计算所山世光研究员指导的《开放场景中面部表情分析方法研究》、华南理工大学金连文教授指导的《基于深度学习的自然场景文本检测及端到端识别的研究》、北京科技大学马惠敏教授指导的《基于同物性学习的复杂场景图像分割研究》、中山大学郑伟诗教授指导的《面向开放环境的行人重识别》获优秀博士学位论文奖，西北工业大学张艳宁教授指导的《高动态范围图像重建方法研究》获优秀博士学位论文提名奖。

✪ 2021 年 12 月 6 日，中国人工智能学会公示了 2021 年度吴文俊人工智能优秀博士学位论文评选获奖和提名论文名单，CCF-CV 专委会 3 位委员指导的博士学位论文获奖，分别是：中山大学林倞教授指导的《视觉表征的高效学习：深度神经网络的持续演进》、西安电子科技大学邓成教授指导的《多模态数据的图表示学习》、北京大学林宙辰教授指导的《深度卷积神经网络结构的设计与搜索研究》。

✪ 2021 年 12 月 9 日，中国图象图形学学会发布了 2021 年度中国图象图形学学会自然科学奖、技术发明奖、科技进步奖评选结果公告，CCF-CV 专委会 13 位委员完成的 8 个项目获奖：华中科技大学白翔教授等完成的复杂场景文字检测与识别、中科院自动化所赫然研究员、谭铁牛院士等完成的异质图像表示和生成理论与方法获自然科学一等奖；中科院自动化所张俊格副研究员等完成的基于结构化认知学习的图像语义理解理论和方法、浙江大学李玺教授、微软公司王井东研究员（现百度）等完成的基于语义关联建模和结构知识表达的智能学习理论与方法获自然科学二等奖；中科院自动化所董晶副研究员、王伟助理研究员、谭铁牛院士完成的视觉伪造可信鉴别关键技术及应用获技术发明二等奖；北京理工大学刘越教授等完成的多通道协同的长时沉浸虚拟现实关键技术及应用获科技进步一等奖；华南理工大学金连文教授等完成的复杂场景文档图像识别与理解关键技术及应用、银河水滴科技（北京）有限公司创始人兼 CEO 黄永祯博士、中科院自动化所谭铁牛院士、王亮研究员、深圳职业技术学院杨金锋教授等完成的面向公共安全的步态识别技术与应用获科技进步二等奖。

✪ 12 月 12 日，中国人民政治协商会议北京市海淀区第十一届委员会第一次会议进行大会选举，CCF-CV 专委会副主任、中科院自动化所王亮研究员当选为政协北京市海淀区第十一届委员会常务委员。

✪ 2021 年 12 月 15 日，ACM 公布了 2021 年度 ACM 杰出科学家（Distinguished Member）名单。CCF-CV 专委会副主任、上海科技大学虞晶怡教授入选。虞晶怡教师是上海科技大学副教务长、信息学院执行院长。在加入上海科技大学前，任职美国特拉华大学计算机与信息科学系正教授。虞晶怡教授长期从事计算机视觉、计算成像、计算机图形学、生物信息学等领域的研究工作，已发表 120 多篇学术论文，其中超 70 篇发表于国际会议 CVPR/ICCV/ECCV 和期刊 TPAMI。目前已获得美国发明专利 20 余项，并于 2009 和 2010 年分别获得美国国家科学基金的杰出青年奖和美国空军研究院的杰出青年奖。此外还是 IEEE TPAMI、IEEE TIP 和 Elsevier CVIU 的编委，担任 ICPR 2020，IEEE CVPR 2021，IEEE WACV 2021，和 ICCV 2025 的大会程序主席。因为他在计算机视觉和计算成像上的贡献，当选 IEEE Fellow。

✪ 2021 年 12 月 15 日，中国电子学会公示了 2021 中国电子学会科学技术奖拟授奖项目，CCF-CV 专委会 6 位委员参加完成的 5 个项目上榜：中国科学院自动化研究所雷震研究员等完成的多模态高鲁棒细微情感分析关键技术与系统、西安电子科技大学董伟生教授等完成的时空谱编码耦合与深度网络解耦超限成像技术拟授技术发明一等奖，浙江大学李玺教授和赵洲副教授等完成的超大规模高性能图神经网络计算平台及其应用拟授科技进步一等奖，中国科学院自动化研究所朱翔昱副研究员参加的基于多模态身份识别的智能金融终端及跨域云服务平台拟授科技进步二等奖。

✪ 2021 年 12 月 15 日，科技部公示了全国科技系统抗击新冠肺炎疫情先进集体和先进个人拟表彰对象名单，CCF-CV 专委会常务委员、华中科技大学白翔教授入选先进个人。

✪ 2021 年 12 月 19 日，CCF-CV 专委会在珠海召开全体工作会议，会上举办了颁奖仪式。CCF-CV 专委会常务委员、南开大学程明明教授获持久影响力论文奖，CCF-CV 专委会 6 位委员、北京科技大学殷绪成教授、哈尔滨工程大学刘海波教授、华北电力大学赵振兵副教授、燕山大学王金甲教授、中山大学任传贤副教授、中科院自动化所黄岩副研究员获中科视拓 Seeta 服务贡献奖。

✪ 2021 年 12 月 20 日，PRCV 2021 举办最佳论文颁奖仪式，CCF-CV 专委会委员、北京工业大学简萌副教授和毋立芳教授指导的论文获最佳论文提名奖。

责任编委 刘海波

# 基于活动轮廓模型的图像分割算法开源代码

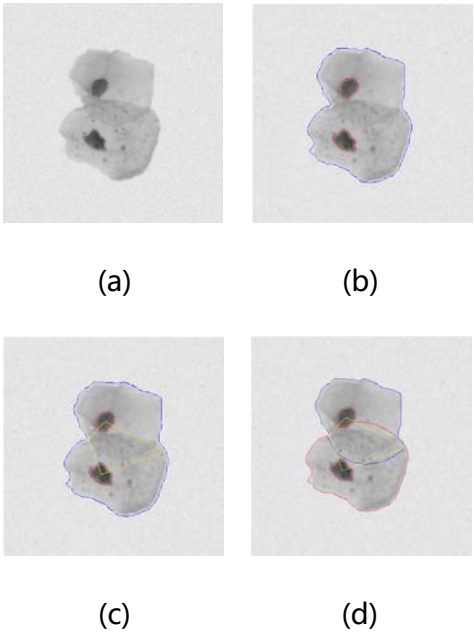
西安交通大学 郑尧月 田智强

活动轮廓模型是一种广泛应用的图像分割方法，其基本思想是使用连续曲线表达目标轮廓。通过定义一个关于轮廓曲线的能量泛函，求解能量函数对应的欧拉方程，使分割过程转变为求解能量泛函最小值的过程。活动轮廓模型的典型代表是几何活动轮廓模型，也被称为水平集方法。水平集方法最初由 Osher 和 Sethian 提出，用于计算物理中对界面的捕捉。水平集方法隐式地将轮廓曲线表示为高维函数的零水平集，通过演化一个高维函数来跟踪目标的边缘曲线，并将分割结果表示为该函数的零水平面曲线。

在计算机视觉领域，尤其是图像分割任务中，水平集方法已获得了广泛关注并成功地应用于医学图像分割任务中。传统的基于水平集的分割方法大多将其作为后处理工具以提高分割结果的精度，近年来，有相关工作提出将深度神经网络与传统水平集方法结合起来，进行端到端的训练。本文将从 Multi-LS 开始，重点介绍基于水平集的分割方法，以及关于深度神经网络与水平集方法相结合的研究成果。

## 1、Multi-LS

**工作：**在显微镜图像中，对重叠细胞的自动检测和分割是最具体挑战性的问题。在宫颈癌筛查中，需要将细胞在显微镜下进行进一步检查，而细胞间的重叠、细胞质对比度差、血液和炎症细胞的存在都影响了检测精确度。该论文提出一种改进的水平集方法，利用多个水平集函数的联合优化处理重叠的宫颈细胞并分割出细胞质和细胞核。图 1 为重叠的宫颈细胞分割示意图，分别为：(a)重叠的宫颈细胞，(b)检测到细胞团和细胞核，(c)细胞团和细胞核轮廓点关联，(d)重叠细胞边缘外扩。



(a) (b) (c) (d)

图 1 重叠的宫颈细胞分割示意图

**论文：**https://ieeexplore.ieee.org/document/7005499

**代码：**https://github.com/luzhi/cellsegmentation_TIP2015

## 2、DCLSM

**工作：**传统水平集方法对初始轮廓位置敏感，若初始轮廓位置设定与分割目标距离过远，可能导致曲线轮廓无法收敛到目标位置。该论文提出深度卷积水平集方法(DCLSM)，主要思想是将通过迁移学习训练的卷积神经

网络作为水平集方法的先验，以获得比传统水平集方法更高的精确度。图 2 为 DCLSM 方法的整体框架图，其中 DCP 表示深度卷积先验(Deep Convolutional Prior)，LSM 表示水平集方法（Level Set Method)。

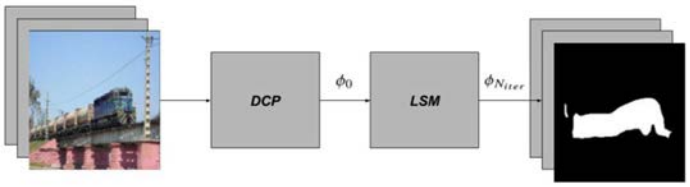



图 2 DCLSM 方法框架

**论文：**https://www.semanticscholar.org/paper/Deep-Convolutional-Level-Set-Method-for-Image-Kristiadi-Pranowo/5b9251bc024a52762537c9c798cf48a40fa16872

**代码：**https://github.com/wiseodd/cnn-levelset

## 3、SPLS

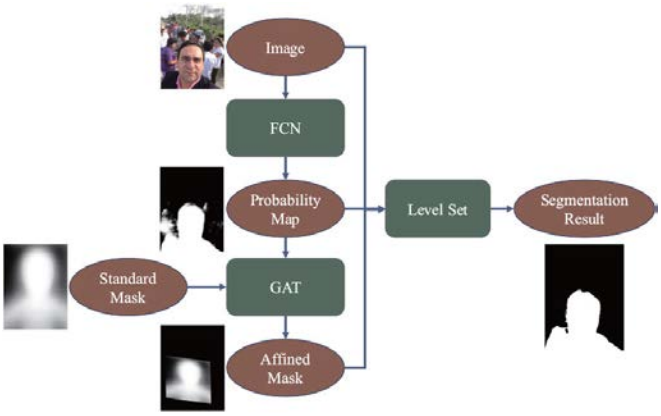



图 3 基于 FCNs 的水平集方法网络框架

**工作：**目前，深度卷积网络在图像分割任务中获得了出色的性能，但是这些方法存在分割结果有噪声、边缘粗糙和无形状先验等缺陷。该论文提出一种基于全卷网络(FCNs)学习先验的水平集分割方法。首先通过 FCNs 从训练数据中学习图像的高级语义特征，并将 FCNs 的输出表示为概率映射。接着通过全局仿射变换，得到先验形状的最优仿射变换。与传统水平集方法相比，该论文将原始图像、概率图和先验信息融入分割网络，利用高级语义信息对复杂场景图像进行分割，解决了 FCNs 的不足。图 3 为所提出方法的整体框架。

**论文：**https://ietresearch.onlinelibrary.wiley.com/doi/epdf/10.1049/iet-ipr.2018.6622

**代码：**https://github.com/zsh965866221/LevelSet-ShapePrior-DeepLearning

## 4、DELSE

**工作：**基于卷积神经网络的分割方法目前已表现出了优秀的性能，但其在处理目标受到遮挡、形状变化大等复杂情况仍存在缺陷。该论文提出交互式深度水平集分割算法，将分割过程定义为曲线演化，通过设计能量函数保证曲线收敛到分割目标边界。论文以端到端的方式将卷积神经网络与水平集演化相结合。模型采用多分支结构预测水平集演化参数，并对预测的初始轮廓进行演化以分割目标。此外，该论文通过结合极端点实现交互式操作。所提出方法能够处理形状和拓扑变化复杂的目标。图 4 为 DELSE 方法示意图，显示了分割框架及卷积神经网络预测的水平集演化参数。

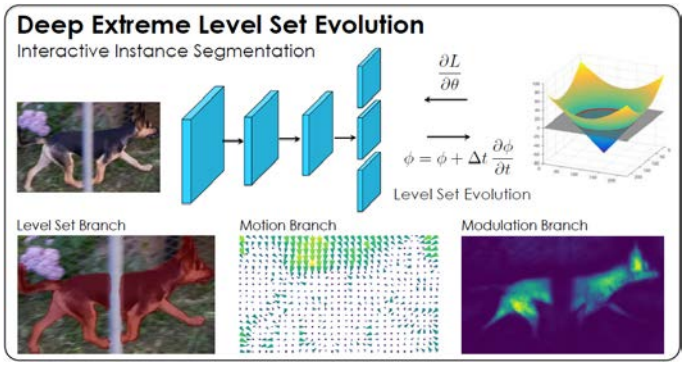



图 4 DELSE 方法框架

**论文：**http://www.cs.toronto.edu/~zianwang/DELSE/zian19delse.pdf

**代码：**https://github.com/fidler-lab/delse/

## 6、DLS

**工作：**图像分割是医学图像处理的重要步骤，在临床分

析和应用中得到了广泛研究和发展。目前基于深度学习的图像分割方法性能受限于独立预测每个像素的类别。该论文基于活动轮廓模型的思想，提出将感兴趣区域内部和外部的面积以及学习过程中目标轮廓的长度集成到学习模型中。论文提出基于活动轮廓思想的损失函数用以监督网络训练，使得在网络训练过程中，同时监督目标内部面积和边缘轮廓信息。该方法在医学图像分割任务上达到了更高的精度。图 5 为所提出方法的示意图。

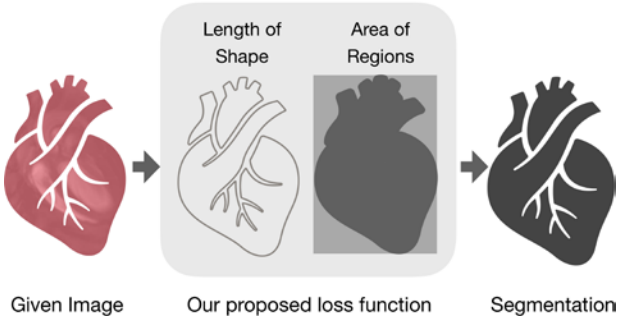



图 5 活动轮廓模型

**论文：**https://openaccess.thecvf.com/content_CVPR_2019/html/ CVPR_2019_paper.html

**代码：**https://github.com/lc82111/Active-Contour-Loss-pytorch

**工作：**青光眼是致盲的主要原因，而测量视杯视盘比是青光眼筛查的主要方法之一。由于视盘和视杯间边缘模糊、生理病变导致的区域变形和不同眼底数据集间图像分布差异大，目前已有的分割方法不能达到分割精度的要求。该论文提出一种两阶段的深度水平集方法实现视杯和视盘的分割，使用多尺度卷积神经网络预测水平集初始轮廓和演化参数，并进一步对初始轮廓进行 T 步水平集演化。网络能够进行端到端训练，结合目标先验知识提升对视杯和视盘的分割精度。

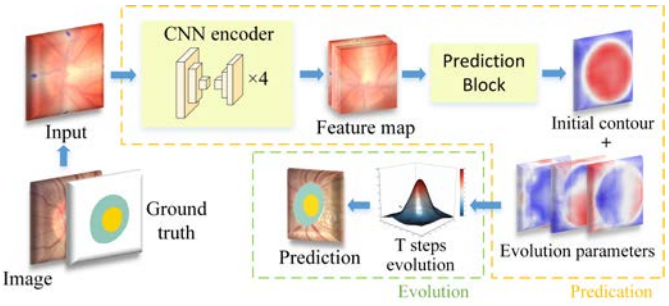



图 5 基于深度水平集的视杯和视盘分割框架

**论文：**https://www.osapublishing.org/boe/fulltext.cfm?uri=boe-12-11-6969&id=460679

**代码：**https://github.com/yaoyz96/deep-level-set

责任编委 李策 贾同

郑尧月

西安交通大学电信学部博士研究生，研究方向为图像分割、计算机视觉。

田智强

西安交通大学电信学部副教授，主要研究方向为视频目标分割、场景理解，视频摘要生成，医学图像处理，计算机视觉等。

# 医学影像数据集

东北大学 贾同 贾娜娜

在过去的几十年里，计算机断层扫描(CT)、磁共振成超声波像(MRI)、正电子发射断层扫描术(PET)、乳房 X 光检测、超声波、X 射线等医学成像技术，已被广泛用于疾病的早期发现、诊断和治疗。在临床上，医学图像的解释大多是由放射科医生和内科医生等人类专家进行的。然而，考虑到病理上的巨大差异和人类专家的潜在疲劳，研究人员和医生已经开始从计算机辅助干预中受益。深度学习可以发现或学习数据固有规律或模式的有效特征，其在医学图像分析的各种任务中起着至关重要的作用。

随着医学成像技术和计算机技术的不断发展和进步，医学图像分析已成为医学研究、临床疾病诊断和治疗中一个不可或缺的工具和技术手段。深度学习已经迅速发展成为医学图像分析的研究热点，它能够从医学图像大数据中自动识别隐含的疾病诊断特征，在医学图像分类、检测、分割、配准、检索、图像生成和增强等各个领域中都起着重要的作用。如何充分利用人工只能深度学习方法分析处理医学图像大数据，为临床医学中各个重大疾病的筛查、诊断、治疗评估提供科学方法，是当前医学图像分析领域急需解决的重大科学问题和前沿医学影像关键技术。本文重点介绍几个在医学影像分析领域常用、基础的数据集，具体包括 LIDC-IDRI、COVID-19 Radiography Database 、 Chest XR COVID-19、LiTS、ADNI 和 LOLA 11。

## 1、LIDC-IDRI 数据集

**介绍:** LIDC-IDRI (The Lung Database Consortium)，该数据集由胸部医学图像文件(.dcm)(如 CT、X 光片)和对应的诊断结果病变标注（.xml）组成。数据是由美国国家癌症研究所（National Cancer Institute）发起收集的，目的是为了研究高危人群早期癌症检测。

该数据中，共收录了 1018 个研究实例。对于每个实例中的图像，都由 4 位经验丰富的放射科医师进行两阶段的诊断标注。在第一阶段，每位医师分别独立诊断并标注病患位置，标注三种类别：1）>=3mm 的结节；2）<3mm 的结节；3）>=3mm 的非结节。在随后的第二阶段中，各位医师都分别独立的复审其他三位医师的标注，并给出自己的最终诊断结果，通过两阶段的标注来实现尽可能完整的所有标注结果。

图像文件位 Dicom 格式，是医疗图像的标准格式，其中除了图像像素外，还有一些辅助的元数据，如图像类型、图像时间等信息。一张 CT 图像的大小为 512×512，如图 1 所示左图为 CT 图像，右图为诊断标注信息，该标注信息以 XML 格式提供。

图 1 LIDC-IDRI 数据集

**数据集地址**

https://pan.baidu.com/s/15PiJ7BSy4JvxFQ9S2cdeRA **提取码**：u366

### 2、COVID-19 Radiography Database 数据集

**介绍：**COVID-19 Radiography Database 是由卡塔尔大学的研究员卡塔尔、多哈以及来自巴基斯坦和马来西亚的合作者和医生，为 COVID-19 阳性病例、正常以及病毒性肺炎建立的胸部 X 光数据库。

COVID-19 Radiography Database 数据集包含 3616 个 COVID-19 阳性病例，10192 个正常例子，6012 个肺部不透明 (非肺部感染) 和 1345 个病毒性肺炎图像，图像格式为.png 格式，图像大小为 256×256，同时每种类型的 CT 图像都有对应的标注。该数据集用于对新冠肺炎、其他病毒性肺炎和正常人三个类别的 CT 检查图像进行分类。图 3 展示了 COVID-19 Radiography Database 数据集中部分对象。如图 2(a)、(b)、(c)、(d)分别为新冠肺炎、肺浑浊、正常人以及病毒性肺炎的 CT 图像。

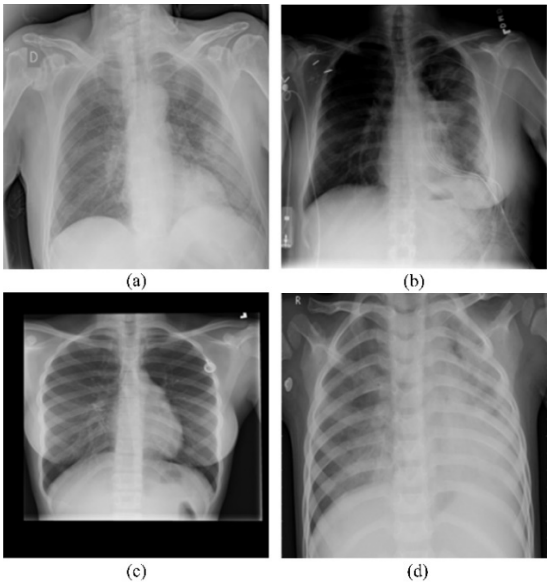



图 2 COVID-19 Radiography Database 数据集

**数据集地址**

https://www.kaggle.com/tawsifurrahman/covid19-radiology-database

### 3、LiTS

**介绍：**LiTS（Liver Tumor Segmentation）数据集来自 ISBI 2017 和 MICCAI 2017 联合举办的 LiTS 肝脏肿瘤分割挑战赛。

LiTS 数据集分为训练数据集和测试数据集，训练数据集中有 130 例，测试数据集中有 70 例，数据格式为标准 Dicom 格式，如图 3 所示为部分数据集示意图，左图为原始图像，右图为标签图像，其中，红色表示肝脏，绿色表示肝脏上的病变。

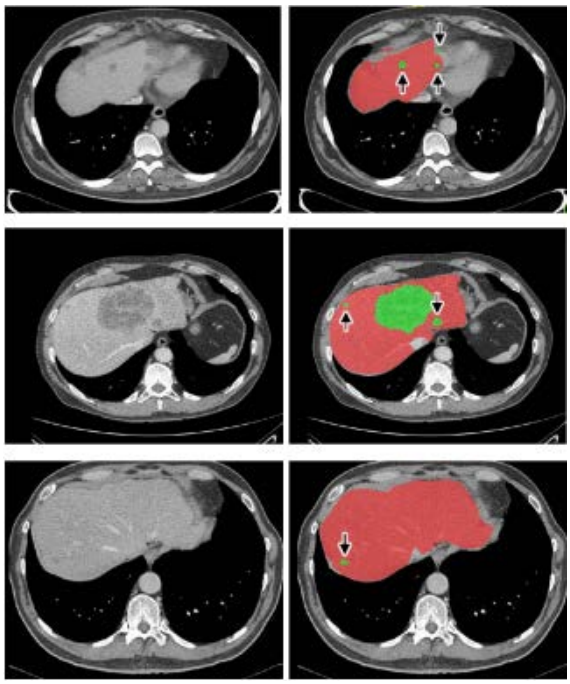


图 4 LiTS 数据集

**数据集地址**

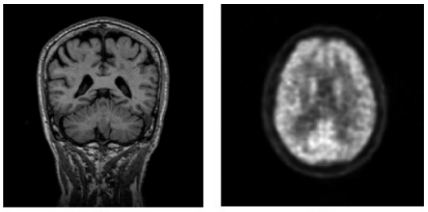
https://www.kaggle.com/andrewmvd/lits-png

### 4、ADNI

**介绍：**ADNI（Alzheimer's Disease Neuroimaging Initiative）数据集是目前研究阿尔茨海默症的权威数据中心，在 2004 年由美国国家卫生研究所和国家老年问题研究所共同资助创建而成，致力于收集阿尔茨海默症病人数据，跟踪病人的发病过程，发掘发病过程的变化与起因，以揭示阿尔茨海默症的发病原理，寻找对应的

治愈方案。

ADNI 数据集包括从轻度认知障碍（MCI）到阿尔茨海默症（AD）共 1721 个病例。ADNI 数据目前分为四个阶段，ADNI-GO、ADNI-1、ADNI-2 和 ADNI-3，其中 ADNI-GO 与 ADNI-1 为基线数据，ADNI-2 与 ADNI-3 主要为后续跟踪数据和新加入的模态数据。该数据集是一个多模态数据集，数据集包括以下几部分：Clinical Data（临床数据）；MR Image Data（磁共振成像）；PET Image Data（正电子发射计算机断层扫描）；Gennetic Data（遗传数据）;Biospecimen Data（生物样本数据）。ADNI 的关键目标是提供将遗传学与影像学和临床数据相结合的机会，以帮助调查疾病的机制，如图 4 所示(a)、(b)、(c)分别为 MRI 图像、PET 图像以及生物样本数据的部分示意图。



(a) (b)

| | | |
|---|---|---|
| All Participants: | Urine | all visits |
| | Plasma from Blood | all visits |
| | Serum from Blood | all visits |
| | CSF | baseline and month 12 |
| | Blood for NCRAD | baseline |
| | Blood for APOE | screening |
| | Cell Immortalization | baseline |

(c)

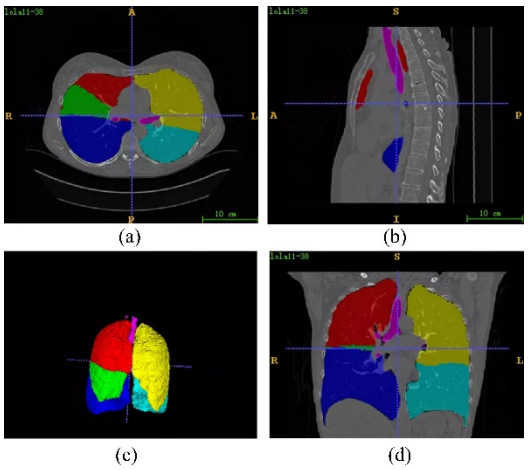
图 4 ADNI 数据集

**数据集地址**

adni.loni.usc.edu

## 5、LOLA 11

**介绍：** LOLA 11（Lobe and Lung Analysis 2011）数据集的主要目的是评估用于胸部 CT 的肺叶分割方法的性能，提供了具有不同异常情况的胸部 CT 扫描数据集，同时建立了肺叶分割参考标准。LoLa 11 数据集共有 55 套 CT 数据，数据格式为 (.mha) 格式。如图 5(a)、(b)、(c)、(d)分别为横断面、矢状面、三维图、冠状面图像。



(a) (b) (c) (d)

图 5 LOLA 11 数据集

**数据集地址**

https://doi.org/10.5281/zenodo.4708800

责任编委 沈沛意

贾同

东北大学信息科学与工程学院教授、博士生导师，智能感知与机器人研究所所长。研究方向为计算机视觉、模式识别与机器学习等。电子邮箱：jiatong@ise.neu.cn

贾娜娜

博士研究生，东北大学信息科学与工程学院，研究方向为医学影像处理。
电子邮箱：2010284@stu.neu.edu.cn

# 好 文 推 荐

厦门大学团队的“Perturbed Self-Distillation: Weakly Supervised Large-Scale Point Cloud Semantic Segmentation”成果发表在 ICCV-2021 上。

论文：Yachao Zhang, Yanyun Qu, Yuan Xie, Zonghao Li, Shanshan Zheng, Cuihua Li. Perturbed Self-Distillation: Weakly Supervised Large-Scale Point Cloud Semantic Segmentation, ICCV, 2021: 15520-15528.

大规模点云语义分割任务是当前计算机视觉领域非常活跃的研究方向，因其在自动驾驶、人机交互、虚拟现实、机器人等环境感知领域中广泛应用而备受关注。随着深度学习的发展，近年来点云分割取得重大进展，然而，大多数点云分割方法都建立在全监督学习的基础上，这种学习模式需要对点云进行手工标记。由于点云的尺度较大，因此完全标记样本导致费时、耗力、成本高。例如，标记 ScanNet 数据集中的一个场景平均需要 22.3 分钟。

为了减小注释代价，弱监督学习点云分割方法（标注一小部分点）引起学者们的注意。已有方法要么需要额外的数据集进行预训练，要么因标签稀少导致缺乏上下文学习机制，或者无法直接扩展到大规模点云上。针对上述问题，本文提出了一中扰动自蒸馏方法（其网络框架图如图 1 所示）通过解决以下两个问题来挖掘数据集自身知识以提升点特征的判别性。

1）如何为未标记的点设计辅助任务进行监督学习，从而在点云的任意两点之间建立一个良好的拓扑结构？

2）除了仅在点级别进行监督之外，如何利用上下文正则化来对标记点之间的关系进行建模？

首先，本文构建了扰动分支并保持扰动分支与原始分支之间预测分布的一致性来引入扰动自蒸馏。一致性约束为所有点提供了额外的监督信息，使引入的图卷积网络（GCN）能够在所有点之间建立起图拓扑。这种可学习的图结构，也为两分支蒸馏引入了一种新的支交互方式，实现了标记点和未标记点之间的有效信息流动。

其次，为了细化图拓扑，本文还提出了上下文感知模块，该模块对有标记点的语义相关性进行了编码，以监督特征相关性学习。由于标记点像锚点一样分布在图拓扑中，一定程度上保证锚点之间的关系准确，那么将会对未标记点之间的相关性关系学习产生积极的影响。

在三个大规模点云数据集上进行了分割性能评估，消融实验分析了所提模型的关键部件的作用。实验结果表明本文所提算法取得了更好的弱监督点云语义分割性能，表现出良好的方法可扩展性。

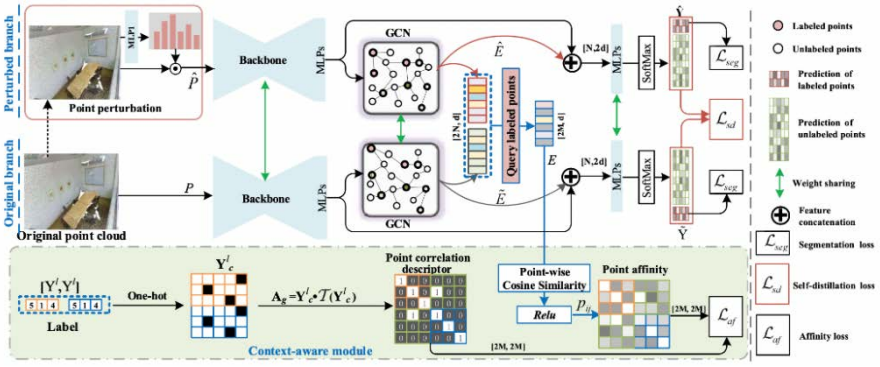



图 1 PSD 算法流程图

# 好文推荐

长安大学团队在“多目标跟踪”最新成果发表在ECCV2020。

论文：ShiJie Sun, Naveed Akhtar, XiangYu Song, HuanSheng Song, Ajmal Mian, Mubarak Shah, Simultaneous Detection and Tracking With Motion Modelling for Multiple Object Tracking, ECCV，2020:626-643.

多目标跟踪（MOT）是计算机视觉中一个长期存在的问题。当代基于深度学习的 MOT 广泛采用检测-跟踪的范式，这种范式将多目标跟踪问题自然地划分为两个子问题：检测和目标关联。在标准的 MOT 评估协议中，假设物体检测是已知的，且在评估序列上提供了公共检测，MOT 算法被用于通过解决数据关联问题来输出物体轨迹。虽然这类方式被广泛采纳，但是也会有一些负面后果，如：数据关联任务中采用的深度学习模型过度依赖于某个检测器，检测器也可能成为跟踪器的平静，此外由于牺牲了端到端的训练，无法充分利用深度学习的能力解决 MOT 问题。

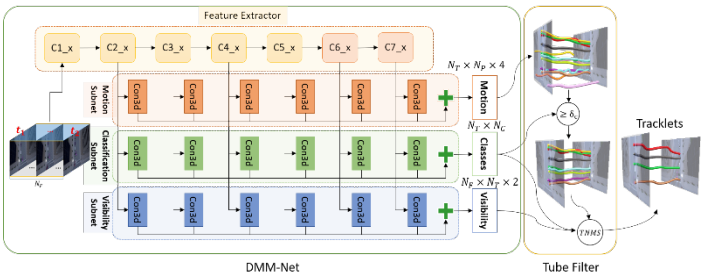



图 1 深度运动建模网络结构

为此，长安大学宋焕生团队利用了深度学习的表示能力。提出了一个端到端的检测跟踪一体化网络，深度运动建模网络（DMM-Net），如图 1 所示。它包含四个部分：主干网络、运动估计子网络、分类估计子网络和可见度估计子网络，具有估计多目标的运动参数、分类及可见度功能，从而实现端到端方式一体化实现检测和数据关联。

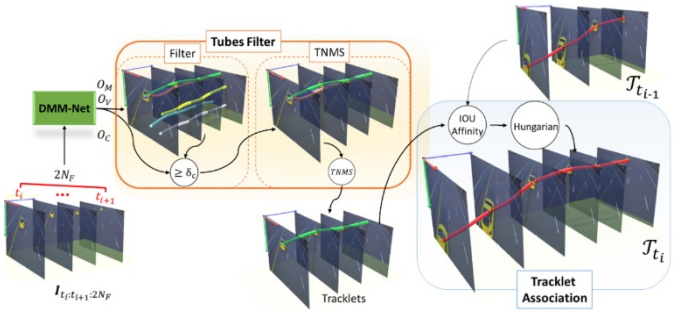



图 2 DMM-Net 部署图

如图 2 为 DMM-Net 的部署图，在图中，$2N_F$幅视频帧，输入网络中，网络输出目标短轨迹、可见度及目标分类，依次利用分类置信度及短轨迹非最大值抑制方法对目标短轨迹过滤，进而，利用匈牙利分配算法对前一时段的短轨迹关联，更新轨迹集合。

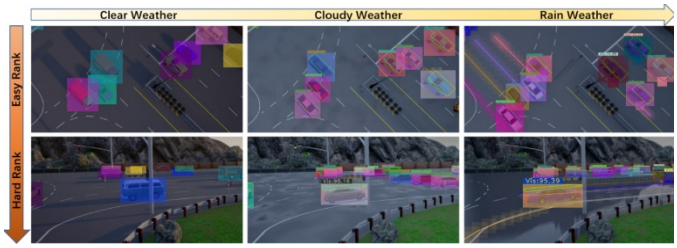



图 3 Omni-MOT 数据集示例图

此外，该团队还发布了基于 CARLA 仿真器的大规模检测跟踪数据集 Omni-MOT，该数据集包含 5 个虚拟城市、3 类天气条件、3 类相机视角、3 类交通量下的视频，数据集包含 14M 幅视频帧，其规模是 MOT17 的 **1000** 多倍，UA-DETRAC 的 **100** 多倍，Waymo 的 **50** 多倍。

DMM-Net 对多个帧的物体特征进行建模，并同时推断出物体的类别、可见度和它们的运动参数。在流行的 UA-DETRAC 挑战中，DMM-Net 达到了较好的性能 PR-MOTA 得分 12.80 @ 120+ fps，此外，实验表明该网络可在 Omni-MOT 数据集上能够取得较好结果。

责任编委 樊鑫 贾同

# 好 文 推 荐

台湾交通大学“PARALLEL RESIDUAL BI-FUSION FEATURE PYRAMID NETWORK FOR ACCURATE SINGLE-SHOT OBJECT DETECTION”最新成果发表在IEEE TIP 2021。

论文：Ping-Yang Chen, Ming-Ching Chang, Jun-Wei Hsieh, Yong-Sheng Chen. DParallel Residual Bi-Fusion Feature Pyramid Network for Accurate Single-Shot Object Detection, IEEE TIP, 30: 9099-9111, 2021

为了使单阶段目标检测方法实现快速准确的检测，本文提出了一种并行残差双融合特征金字塔网络（PRB-FPN）。近年来，特征金字塔（Feature Pyramid, FP）在视觉检测中得到了广泛应用。然而，由于池化移位（pooling shifiting）的原因，FPN 自上而下的路径无法保持精确的定位。特别地，当 FP 被应用于深层主干网络时，它增强目标特征的优势将被削弱。此外，它不能同时对大、小物体实现精确检测。为了解决上述问题，本文提出了一种新的并行 FP 结构，该结构具有双向(自顶向下和自底向上)融合方向并对相关部分进行了重新设计。因此，所提结构可以保留高质量特征从而实现精确定位的目标。接下来，将简单介绍本文所提的并行 FP 结构（图 1 是所提模块的结构图）：

(1)所提结构包括自底向下和自上而下的融合模块，能够同时检测大、小目标，从而具有较高的精度。

(2) 连 结 - 重 建 (concatenation and re-organization，CORE)模块提供了自底向上的特征融合路径，可产生双向融合的特征金字塔。该模块设计有助于从底层特征映射图中恢复丢失的信息。

(3)进一步纯化 CORE 特征以保留更丰富的上下文信息。纯化操作可以使 CORE 模块无论经过哪条特征融合路径都可以只需要几次迭代就可以完成。

(4)在 CORE 模块中添加了残余设计，形成了一个新的 Re-CORE 模块，可以使该模块更容易训练，同时，更方便其移植。

PRB-FPN 在 UAVDT17 和 MS COCO 数据集上实现了最先进的性能。

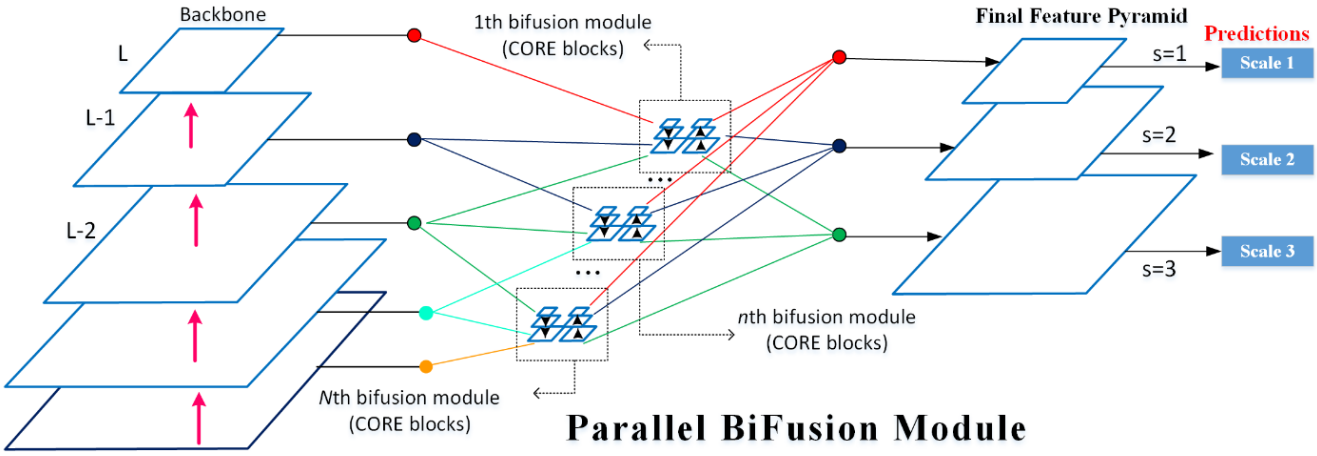
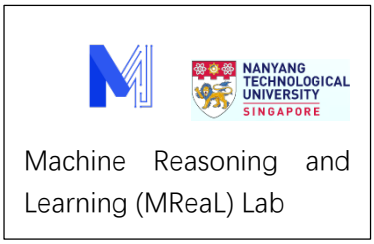



图 1 PRB-FPN 结构图

责任编委 贾同 樊鑫

# 新加坡南洋理工大学张含望教授团队




新加坡南洋理工大学 MReaL 实验室现由张含望教授领导，聚焦于面向计算机视觉的深度学习与因果推理融合重要问题，包括视觉-文本各项任务 (如视觉问答，对话，指称等)，中高层视觉各项任务 (分割，检测等)，以及机器学习基础算法 (小/零/增量学习，半/自监督学习，预训练等)。

## 团队发展历史

MReaL（读作: me real； 译作：吾真斋）实验室由张含望教授于 2018 年创立。她的标志有“左右”两部分，代表我们的左右脑：左边负责逻辑思维，右边负责艺术感知。在标志中，右边是清晰的电路图，代表目前深度学习对于感知这样的“快处理”已经日臻成熟；左边是一整块颜色，代表目前对于机器的逻辑推理这样的“慢处理”还处于黑箱萌芽状态。MReaL 的终极目标，就是若干年以后，让她标志的左边也像右边一样。

MReaL 没有独立的场地，而是寄居在南洋理工计算机学院地下一层，和其他众多老师的团队共享的实验室。所以，MReaL 最大的资产就是“人”——一群聪明、年轻、活泼、逐梦的年轻人，聚在一起，心无旁骛地痴迷“非主流”的研究，从而创造新的“主流”。

MReaL 的座右铭，也是唯一的原则，就是“绝不跟风”。她的目标，就是把每一个成员，培养成该领域未来五到十年的潮流引领者。在 2018 年，创始人团队（导师张含望与博士生杨旭、汤凯华、胡心亭，以及博士交流生陈隆、刘大庆、史佳欣）就果断放弃当时在视觉语言领域颇为流行的注意力堆叠机制，主推由符号和结构等离散归纳偏执的“神经-符号”推理框架。果不其然，注意力堆叠机制在 2019 年以后被资本以及资本带动的大规模预训练模型吞噬。MReaL 也因此幸免于难。至此，团队已迅速扩充至 20 位成员。MReaL 本可以“集团军”作战的方式复制以前的成功模式，寻求“确定”的产量，但团队却轻装上阵，再次起航，孤注一掷地走向“因果关系“应用的前沿研究。短短两年时间内，MReaL 已经在计算机视觉和机器学习的大部分领域，建立了视觉因果关系的新范式。在 2021 年，团队认为因果关系本质毕竟还只是统计归纳，如果要探究因果关系中的“因果关系”，就必须更深一步，抛弃统计理论皆为“事后诸葛亮“的弊端，主动对数据的动态演化进行建模，创立了基于群论的不变性特征学习理论。

MReaL 的品质，也是唯一的品牌印记，就是“极致打磨“。标上 MReaL 印记的论文，每一篇都会着重讲一个一气呵成的“好故事”，以理服人，直击问题的本质，大胆的公布假设的缺陷，从理论上指出为什么好，为什么差，以及可能的改进方向。MReaL 对论文的细节逻辑有着近乎疯狂的偏执，论文中没有一个词，甚至一个标点符号是多余的，她不希望浪费每一位读者的宝贵时间，

让他们都能获得启发和反思。

MReaL 的底气，也是唯一的源动力，就是不受资金支持方项目的裹挟。她的主要经费来源于两部分，1）新加坡政府，2）新加坡企业联合研究院——都是非盈利纯学术机构。这给了她无限畅想的自由度，每一位成员都不需要去充当廉价劳动力，参与跟科学研究无关的工程项目。这使得她的研究方向越来越宽，越来越深，越来越大胆，越来越冒险，越来越经典。

## 研究方向 1：基于结构化归纳偏置的视觉-语言推理

物体不是独立的，妥善地利用物体间的相互依存关系即为结构化的归纳偏置（inductive bias）。大量的生物和认知研究已经证明：人类的大脑非常善于利用这种归纳偏置去合理地在时间尺度上推测事物的发展，或者在空间尺度上由点及面地以小窥大。比如，我们看到杯子倾斜就会提前预感到茶水翻洒的风险而及时扶好，我们听到背后的引擎声就会意识到有车辆的出现。这些都构成了我们人类推理能力的一部分。而我们 MReaL 实验室的一个重要的研究分支，便是探索如何更好地将这样结构化的归纳偏置运用于视觉-语言推理中来。

基于上述研究思路，我们目前构建了如下的推理框架：首先，我们将结构化的场景图作为图片的表达形式，构建一套鲁棒的场景图提取算法来作为图片的预处理。之后，基于结构化的表达，我们就可以在下游视觉-语言任务中结合具体的场景和应用去学习有效的归纳偏置来构建和提升模型的推理能力。值得欣慰的是，目前在提供完美场景图的受控环境下，我们组的算法可以做到100%的推理准确性，进一步证明这套研究思路的潜力。

详细来说，在第一阶段的结构化场景图提取算法上，我们提出了视觉上下文动态树（VCTree[1]）结构来更好地编码视觉上下文信息用于场景图的生成。基于物体间的相关性矩阵，我们构建了物体相关度的最大生成树并进行了二叉树转换，即 VCTree。上述 VCTree 用树状长短期记忆网络（Tree LSTM）编码后，便可用来预测更鲁棒的视觉场景图。而 VCTree 的建模过程也可以通过混合训练的方式，即同时利用有监督学习和强化学习，来进一步优化。该算法的优点主要体现在两个方面：1）由于该结构是动态学习的，这避免了传统的静态结构对训练数据中物体布局的过拟合。2）由于二叉树增加了 VCTree 的深度，提取上下文信息时无关的边缘物体信息可以更好地被排除在重要的相关物体之外。VCTree 算法不仅提升了传统的场景图生成效果和鲁棒性，更大大提升所生成场景图中细粒度视觉关系的准确性。该算法也入选了 CVPR 2019 的最佳论文入围列表（Best Paper Finalists）。

在具体的下游任务上，利用结构化的图片表达以及对结构化归纳偏置的学习能力，我们在多个视觉-语言的匹配和推理任务中，包括看图说话和看图问答等任务，都取得了卓越的效果。具体来说，在看图说话（Image Captioning）领域，如图 1 所示，由于语言层面的表达

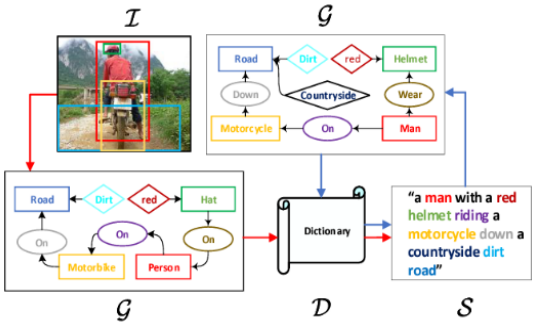



图 1：基于场景图自动编码的看图说话

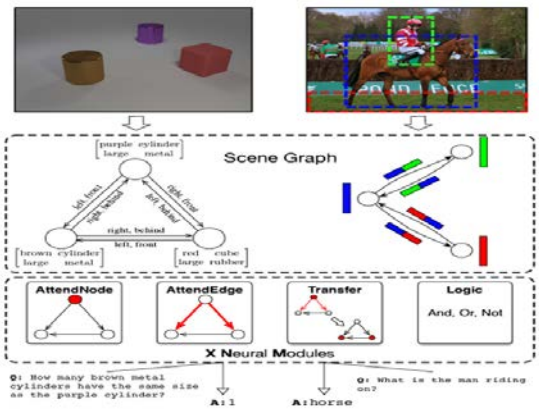



图 2：基于场景图地可解释显性视觉推理

结构往往更符合人类的描述习惯，我们建立了一套基于人类语言结构的归纳偏置“字典”[2]，然后通过对场景图的自动编码去查询这个“字典”，从而获取最恰当的对应描述文本。而在看图问答（Visual Question Answering）的推理领域，我们更进一步地利用归纳偏置将视觉推理抽象为四种模块化网络结构，即 1）节点注意力模块，2) 连接注意力模块，3) 注意力转移模块，和 4）逻辑推理模块。如图 2 所示，该算法巧妙地归纳了人类视觉推理的核心要素，并将这些推理要素转化为可以在场景图上进行计算的神经网络推理模块。利用模块化网络的可解释性和泛化性，我们的算法[3]不仅提供了一套显性的可解释的视觉推理框架，更在性能上冠绝同类算法：在给定真实场景图的可控环境下（CLEVR 数据集），我们的算法在推理能力上做到 100%的准确性。

综上，我们意识到只要巧妙地利用结构化归纳偏置构建合理的推理模块，现有的模块化神经网络可以近乎完美地建模了视觉推理中的各种要素。这证明了视觉-语言推理任务的瓶颈并非高层次的逻辑推理能力，而恰恰是底层的识别能力的鲁棒性和可解释性。因此，我们 MReaL 实验室近年也逐渐将科研重点导向了后续的两个研究方向。

## 研究方向 2：视觉因果推理

虽然大部分情况下，归纳偏置可以在视觉-语言推理中对问题本身和解决问题的过程进行有效的抽象来辅助神经网络进行推理。但归纳偏置并不一定总是好的，尤其是在涉及底层识别和认知时，对其过分依赖可能导致对训练数据分布的过度拟合从而丧失一定的泛化能力。为了解决该问题，进一步提升计算机视觉算法的鲁棒性和泛化性，我们 MReaL 实验室由此在因果推理领域寻到了新的思路：让我们的算法在分布变换（Distribution Shift）的苛刻条件下依然保持可靠性。

如果将人类的认知能力分为从易到难的不同层级[4]，那么我们可以将其大致划分出如下三层：1）关联，2）干预，和 3）反事实。而传统的机器学习算法不论多复杂的结构都仅仅只是做到将数据和标注进行关联，即第一层级。这一层级是对分布变换最为脆弱且最容易产生偏见的。在此基础上，为了去除算法的偏见，我们不仅提出了数个视觉任务中的干预算法和反事实算法，更进一步尝试结合干预和反事实能力，达到最终对任意分布鲁棒的视觉算法。

所谓干预算法，即是保证模型的预测结果不会受到特定分布变换的影响，比如为了使无人驾驶系统不会受到天气影响，一个基于干预算法的视觉系统应当需要保证同一场景下在所有因素不变时，单纯改变场景的天气不会对模型的预测产生任何影响。在这个思路下，MReaL 提出的基于干预的弱监督语义分割算法[5]，就可以消除上下文环境信息引入的偏见，从而更好地划分存在歧义的语义边界。同理，在少样本学习中，不足的数据量往往也会导致模型学习虚假的关联信息，从而产生偏见。因此利用基于后门调整公式[4]，MReaL 组提出的干预算法[6]显著地提升了少样本训练环境下，模型的鲁棒性和综合性能。

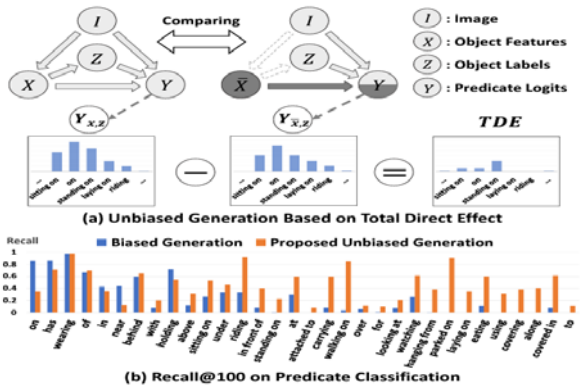



图 3：无偏见的场景图生成算法

所谓反事实算法，则是为了进一步确保模型面对未知分布变换时举一反三的能力，即人类可以合理猜想未发生事件的影响，从而判断当前行为和预测的合理性。MReaL 实验室通过引入直接/间接因果效应的概念，在计算机视觉领域的多个方向实现了以反事实推理去除偏见的算法。以长尾分布为例，目前不管是对具体任务的还是对通用的长尾问题，MReaL 都有一套以总直接

因果效应（Total Direct Effect/TDE）为核心的完善解决方案[7,8]。以图 3 中的场景图生成算法[7]为例，传统的视觉关系往往存在严重的长尾偏见，使得模型过度偏好缺乏信息量的模糊描述，而通过引入反事实推理，我们将系统性的长尾问题抽象为一种盲猜偏见，所以我们构建了一个反事实的盲猜模型，保留除物体详细特征外的所有其他因素不变，从而通过事实和反事实推理的差值来获取不受其他因素影响的纯视觉偏好，实现细粒度高信息量的场景图生成。同理，在通用的长尾问题中类似的 De-confound-TDE 算法[8]也首次实现了不依赖训练分布矫正的长尾鲁棒算法。此外在多模态的视觉-语言推理领域，MReaL 实验室提出的反事实因果推理[9]也可以显著地抵消单一模态引入的偏见，例如视觉问答系统中纯文本的问题-答案匹配偏见。

虽然单纯的干预算法和反事实算法在各自领域都能起到显著的去除偏见，进行鲁棒推理的作用。对于一个任意新问题，怎样从上述两种思路中做抉择依然是个需要详细分析和权衡的复杂过程。因此如何让模型同时结合上述两个算法的优点，对任意给定问题自动适配所需的因果推理方式便是一个新的重要目标了。如图表 4 所示，单纯的干预或反事实算法只能在同分布下的表现和异分布下的表现之间进行权衡，而无法保证任意分布下的最优。而我们 MReaL 近期提出的自省蒸馏算法（Introspective Distillation）[10]，则可以同时在两种不同分布下都取得理想的效果，即实现任意分布下的鲁棒性。这也为我们的视觉因果推理研究方向补完了最后一块拼图。

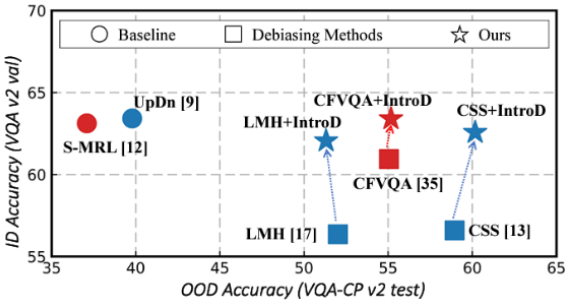



图 4：容易干预与反事实后在任意分布下同时保持最优

一言以蔽之，因果分析追求的是找到具有决定性的“不变”因素，因为分布鲁棒性是合理运用结构化归纳偏置的前提。但由于机器学习是一门纯粹由数据驱动的学科，在没有任何先验信息的情况下，如何从数据中自动区分“变”和“不变”的信息却又是一个新的难点。

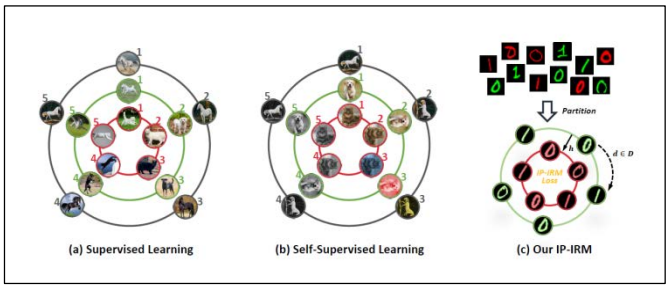



图 5：群论中轨道定理解释了“类”和“属性”的形成 [12]

## 研究方向 3：基于群论的特征不变性学习

因果关系追求的是亘古不变的原理。要想从繁复的自然观察中获取不变性，那么首先要拆分 (disentangle) 出来什么是“变”和“不变”。如果拆分不成功，因果推理就注定是空中楼阁。但是，机器如何才能拆分和理解什么是“变化“呢？不巧，目前的深度学习框架是基于统计的线性代数建模，该框架很难“主动”建模数据中的“变化”。例如，开普勒行星轨道理论就是基于统计的线性代数，它只能完美地解释已经观察到的行星运动轨迹；牛顿力学则是主动研究“变化”，能让行星运动起来的原因是重力，它不但可以推导出已经存在的行星轨道，还可以预测不存在的人造卫星可能的轨道。统计的局限在于“事后诸葛亮”，线性代数的局限在于只能描述一个具体的变化，但不能描述一类抽象的变化。

MReaL 用抽象代数的群论来建模变化，拆分“变”与“不变”。具体来讲，一个群代表图片上面的一组变化，群作用在图片上就像是重力作用于行星，让原本静态的一张图片动起来，形成一个变化后图片的轨道。例如控制动物形态变化的群，作用在一张马的图片上，就得到了一个包含形形色色的马的轨道（图 5 （a）外环）。

有趣的是，群论的轨道完美解释了全监督（SL）和自监督学习（SSL）中模型拆分出的“变”和“不变”。SL 中数据被标注成多个类别，如图 5（a），同类的图片就是一个轨道，我们用群$\mathcal{D}$ =“*类共享属性*”代表每个轨道内的变化（例如站立→奔跑），用群$\mathcal{C}$ =“*类相关属性*”代表轨道间的变化（例如马→狗）。而 SSL（图 5 （b））基于数据增强（如裁剪、噪声），每张图和其增强后的图是一个轨道，因此轨道内$\mathcal{D}$ =*数据增强*，轨道间$\mathcal{C}$ =*图片变换*（例如 1→2）。SL 和 SSL 都采用“轨道间不同，轨道内相同”的训练目标，通过“$\mathcal{C}$动我就动”的“变”来区分不同的类别，通过“$\mathcal{D}$动我不动”的“不变”实现分类的鲁棒性。

然而 SSL 拆分出对数据增强的不变性显然不够好，例如当物体背景和形状的群没有被拆分，那么我们用形状去分类的同时，就失去了对与分类无关的背景的不变性。因此我们提出基于自监督学习的 IP-IRM 算法进一步拆分群的变换[12]。支撑 IP-IRM 的核心定理是“*群$\mathcal{D}$被拆分⇒模型具有对$\mathcal{D}$的不变性*”。如图 5 (c)，IP-IRM 首先将训练数据集分为两个轨道，使得当前模型无法对轨道间的变换$h$达到“不变”，那么根据定理的逆否命题，$h$一定尚未被拆分；然后更新模型达到对$h$的不变性，定理则保证$h$被拆分出来。通过重复这两个步骤，IP-IRM 最终将所有的群拆分出来，使得模型能够在下游任务中以不变应万变。

群论的“变”与“不变”还可以完美解释为什么大多数机器学习任务的因果图是“三角形”的，即图 6 (d)中$(X, X_0)$，群$\mathcal{D}$和$Yes/No$之间的三角结构。其中，“三角形”的两条斜边蕴含着群论的“不变”，底边蕴含着群论的“变”：以图像分类任务为例，图片$X$是由某个模板$X_0$通过群$\mathcal{C}$和$\mathcal{D}$提供的变换$c$和$d$得到的（即$X = cdX_0$，例如棕色的马是由一个空白模板$X_0$通过$c$ =“变为马”和$d$ =“变为棕色”得到），这对应因果图中$\mathcal{C}$和$\mathcal{D}$指向$(X, X_0)$的部分；

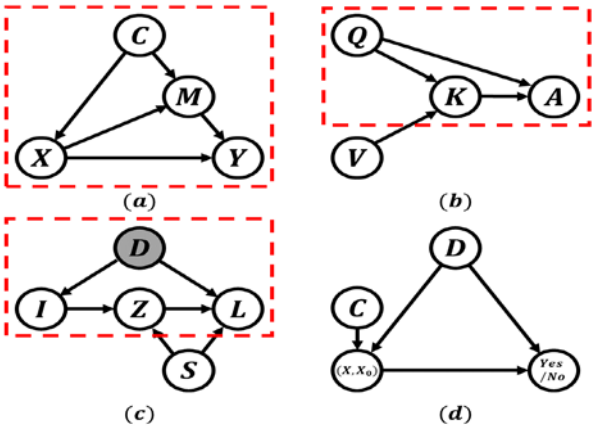



图 6：(a)，(b) 和 (c) 分别是[5]，[9]和[11]中的因果图，红色虚线框内的“三角形”可以归纳为图(d)：群论对因果图“三角形”的解释

分类任务的最终目标是判断$X$属于什么类（判断结果为$Yes/No$节点，代表$X$是否等于$cX_0$），即$X$属于哪一个$c$所对应的轨道，这个过程对应“三角形”的底边；为了实现这个判断，我们需要从$X$中“恢复”出$cX_0$，即去除群$\mathcal{D}$的影响，所以群$\mathcal{D}$需要提供逆变换$d^{-1}$，这对应“三角形”的右斜边。在“三角形”因果图中，根据群论中的“变”与“不变”，我们得到了两条从图片$X$到分类结果$Yes/No$的路径：群论中的“变”要求分类结果要跟随$\mathcal{C}$的变化（即$\mathcal{C}$产生的是因果效应），对应“三角形“中$\mathcal{C}$经过底边到达$Yes/No$的因果路径；而群论中的“不变”要求分类结果不跟随$\mathcal{D}$的变化（即$\mathcal{D}$产生的是虚假关联），对应“三角形”中$(X, X_0)$经过两条斜边到达$Yes/No$的后门路径。

群论的“变”与“不变”指导我们判断时要跟随三角形底边而“变”，对斜边“不变”，这恰巧对应了我们前面因果理论一节中提到的反事实算法所追求的目标：寻求输入输出间的因果（底边）效应，去除虚假关联（斜边）的影响。这就是群论对因果理论的深层解释，进一步证明了群论是机器学习更基础而有力的工具。

**责任编委** **张汗灵**

## 参考文献

[1] Tang, K., Zhang, H., Wu, B., Luo, W., & Liu, W. (2019). Learning to compose dynamic tree structures for visual contexts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6619-6628).

[2] Yang, X., Tang, K., Zhang, H., & Cai, J. (2019). Auto-encoding scene graphs for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10685-10694).

[3] Shi, J., Zhang, H., & Li, J. (2019). Explainable and explicit visual reasoning over scene graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8376-8384).

[4] Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. Basic books.

[5] Zhang, D., Zhang, H., Tang, J., Hua, X. S., & Sun, Q. (2020). Causal Intervention for Weakly-Supervised Semantic Segmentation. Advances in Neural Information Processing Systems, 33.

[6] Yue, Z., Zhang, H., Sun, Q., & Hua, X. S. (2020). Interventional Few-Shot Learning. Advances in Neural Information Processing Systems, 33.

[7] Tang, K., Niu, Y., Huang, J., Shi, J., & Zhang, H. (2020). Unbiased scene graph generation from biased training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3716-3725).

[8] Tang, K., Huang, J., & Zhang, H. (2020). Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect. Advances in Neural Information Processing Systems, 33.

[9] Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X. S., & Wen, J. R. (2021). Counterfactual vqa: A cause-effect look at language bias. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12700-12710).

[10] Niu, Y., & Zhang, H. (2021, May). Introspective Distillation for Robust Question Answering. In Thirty-Fifth Conference on Neural Information Processing Systems.

[11] Yang, X, Zhang, H, Cai J. (2021). Deconfounded image captioning: A causal retrospect. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[12] Wang, Tan, Yue, Z, Huang, J., Sun, Q. & Zhang, H. (2021, May). Self-Supervised Learning Disentangled Group Representation as Feature. In Thirty-Fifth Conference on Neural Information Processing Systems.

## 张含望

2009 年在浙江大学获得计算机科学与工程学士学位（竺可桢荣誉学位），2014 年在新加坡国立大学计算机科学院取得博士学位（最佳博士论文荣誉）。随后，张博士留校从事博士后研究工作直到 2016 年底，随后又在美国哥伦比亚大学从事博士后研究。张博士于 2018 年加入新加坡南洋理工大学，被聘为“南洋”助理教授，工作至今。张博士的研究方向主要为计算机视觉，机器学习，多模态分析，以及因果推理在以上领域的应用。张博士和团队多次获得多媒体以及推荐领域的最佳论文或提名奖，以及学术竞赛冠亚军。张博士凭借其团队在因果推理应用的开创性成果，被评为 2021 年新加坡总统奖-青年科学家奖和 2021 年 IEEE AI’s 10 to Watch。

# 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 Computers & Electrical Engineering（CEE），Machine Learning（ML）和 Pattern Recognition Letters（PRL)。

## 3 会议简介

国际多媒体博览会（IEEE International Conference on Multimedia and Expo）是 IEEE 一年一度的学术性会议，会议的主要内容是多媒体信号处理技术。ICME 是多媒体研究领域的旗舰类国际学术会议之一。

2022 年 ICME 在中国台北举行，本届会议将汇聚全世界从事多媒体理论与应用研究的广大科研工作者及工业界同仁，共同分享多媒体研究领域的最新理论和技术成果，为大家提供精彩的学术盛宴。

**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 |
|---|---|---|---|---|
| **FAccT 2022** | 2022.06.21-24 | Seoul, South Korea | 2022.01.15 | https://facctconference.org/2022/ |
| **ICML 2022** | 2022.07.17-23 | Baltimore, USA | 2022.01.28 | https://icml.cc/Conferences/2022 |
| **SIGIR 2022** | 2022.07.11-15 | Madrid, Spain | 2022.01.29 | https://sigir.org/sigir2022/ |
| **ECCV 2022** | 2022.10.24-28 | Tel-Aviv, Israel | 2022.03.08 | https://eccv2022.ecva.net/ |
| **IJCAI-ECAI 2022** | 2022.07.23-29 | Vienna, Austria. | 2022.01.15 | https://ijcai-22.org/ |
| **SIGGRAPH 2022** | 2022.08.08-11 | Vancouver, Canada | 2022.01.27 | https://s2022.siggraph.org/ |

**表 2 计算机视觉领域相关国内外期刊专刊**

| 期刊名称 | 专刊题目 | 投稿网址 | 截稿日期 |
|---|---|---|---|
| **CEE** | Artificial Intelligence for Smart Image Recognition and Analysis based on 3D Deep Learning and Fuzzy Logic | https://research.com/special-issue/artificial-intelligence-for-smart-image-recognition-and-analysis-based-on-3d-deep-learning-and-fuzzy-logic | 2022.01.01 |
| **PRL** | Self-Learning Systems and Pattern Recognition and Exploitation (SeLSPRE) | https://www.journals.elsevier.com/pattern-recognition-letters/call-for-papers/self-learning-systems-and-pattern-recognition-and-exploitati | 2022.01.31 |
| **ML** | Safe and Fair Machine Learning | https://www.springer.com/journal/10994/updates/18786592 | 2022.02.15 |
| **PRL** | Deep Learning for Acoustic Sensor Array Processing (DL-ASAP) | https://www.journals.elsevier.com/pattern-recognition-letters/call-for-papers/deep-learning-for-acoustic-sensor-array-processing-dl-asap | 2022.03.20 |

# 心底无私视界宽∞马颂德研究员专访

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、在该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。

2021 年 7 月 29 日，《CCF-CV 专委简报》委托中国科学院自动化所娄文利老师，专访了马颂德研究员。在本次访谈中，我们预先准备了一些与本专栏内容相关的问题，并根据马颂德研究员的专访记录，在不改变实质性内容的前提下，进行了整理。

图 1 中科院自动化所马颂德研究员

**马颂德研究员简历：**

1946 年 7 月生。1968 年毕业于清华大学自动控制系，1983 年在法国巴黎第六大学获博士学位，1986 年获法国国家博士学位，专业为计算机视觉与图象处理。1983 年至 1986 年在法国国立信息与自动化研究所任研究员。1986 年 7 月回国后，历任中国科学院自动化所研究员、国家模式识别重点实验室主任、副所长（1992-1997）、所长（1997-2000）。 2000 年 4 月至 2006 年 10 月任科学技术部副部长。

现任中科院自动化所研究员、中科院自动化所学术委员会顾问、欧美同学会副会长（2021 年换届结束）和欧美同学会留法分会会长（2021 年换届结束）。

马颂德研究员从事计算机视觉、图象处理、模式识别等方面研究，在国内外学术刊物与学术会议上发表 200 余篇学术论文，专著一本。在留学期间曾获欧洲计算机图形学年会（EUROGRAPHIC'85）最佳技术奖和最佳论文奖。回国后曾获国家自然科学二等奖。

1996 年与法国国立信息与自动化研究院、法国国家科研中心等共同创建中法信息、自动化与应用数学联合实验室（LIAMA）。实验室于 1997 年 7 月正式建于中国科学院自动化研究所内，马颂德研究员任第一届中方主任。由于马颂德研究员回国后长期努力促进中法科技交流和合作，2000 年法国政府授予马颂德研究员法国国家荣誉勋章（L'ordre National Du Merite）。

2007 年 11 月 18 日至 22 日在日本东京召开第八届

亚洲计算机视觉国际会议。这是亚洲国家主持的计算机视觉领域的最重要的两年一次的国际会议。从本届会议开始，亚洲计算机视觉会议在每次会议上评选三篇论文为最佳论文，并颁以最优论文奖。该奖分别以中日韩三国对计算机视觉的研究长期做出重要贡献的资深研究人员的名字命名，他们是日本的 Saburo Tsuji、中国的马颂德、韩国的 Sang Uk Lee。马颂德研究员应邀前往颁奖。

2017 年 10 月 12 日，在天津举行的中国计算机视觉大会上，马颂德研究员获中国计算机学会颁发的“中国计算机视觉终身学术成就奖”。

**问**：您是从何时开始从事科学研究工作的?

**马老师**：可以说是 1980 年吧。我虽然是 1963 年入学清华大学，但 1966 年开始的“文化大革命”中断了我们的学业，并于 1968 年“毕业”离校，在北京燕山石化做了 10 年工人。1977 年恢复高考，1978 年恢复研究生招生，我从工厂考上了中科院自动化所研究生，并于 1979 年公派法国巴黎第六大学计算机系做博士研究生。1980 年我在博士研究生期间，就开始了图像处理、计算机视觉的研究。

**问**：您最觉得自豪的一项科研成果 (以计算机视觉为例) 是什么?

**马老师**：我从事计算机视觉、计算机图形学几十年，说起研究成果，说“自豪”恐怕不确切，可以说有些值得回忆和欣慰的事。一是，我在博士研究生期间，提出具有创新思想的根据纹理图像特征（两阶自相关参数）合成自然纹理图像的方法。有启发意义的是，我是刚刚接触该领域研究的新人，说明年轻人具有更强的创新意识。另外，这个方法连接了计算机视觉和计算机图形学。论文发表在 EUROGRAPHIC'88 上，并获最佳论文奖和最佳技术奖，相关论文也发表于《Computer Vision, Graphics and Image Processing》，受到两个学科的共同关注。

当然，最值得欣慰的是，我在 1986 年回国后，在中科院自动化所，参与创建了模式识别国家重点实验室，系统性地推动了计算机视觉、语音识别和自然语言理解的研究。这些研究还派生出多个专门的应用研究方向，如遥感图像分析和识别、生物特征识别（Biometrics），由脑 FMRI 图像分析进入的脑功能连接分区研究。

图 2 马颂德研究员（右）参加河南欧美同学会（河南留学人员联谊会）成立大会

**问**：您觉得学生应该如何培养比较好，能分享下您的经验吗?

**马老师**：研究生（包括硕士生和博士生）是在本科扎实基础课程学习的基础上，学习和实践科研工作，导师主要在选择研究方向上给予指导（对硕士生可能要更具体和明确些）。实际上，只要研究方向在学科前沿，青年人创新性强，或大或小都能做出有创新性的工作，学生也从中体会出独立从事科研工作的乐趣和经验。目前的问题是，有些地方招了很多研究生，但导师并没有具有国际前沿的研究方向，常常让学生做一些应用性的课题，这就失去了研究生教育培养研究能力的作用。研究生培养的另一个常见的问题是，不敢进入一个较新的方向，导师和学生都怕没把握在较短的时间内取得成果，于是只好在一些“有把握”的方向上，对已经做过的工作，做一些改进。虽然也有贡献，也可以发表论文，但对培养真正的创新型思维和方法贡献不大，相信这些都是研究生教育中的世界性普遍问题。

**问**：您对现在人工智能普遍关注深度学习有何看法？

**马老师**：近年来的深度学习，带来了人工智能的快速发展和创新时期。也许可以说，由几十年前的神经网络及反向传播（back-propagation）的“复兴”而来的深度学习，很大程度上得益于大数据和计算能力的不断提高，神经网络规模变得越来越大，能力也越来越强，在全球范围内从计算机视觉、语言识别，到机器翻译、自然语言处理，几乎全面代替了传统的人工智能技术。最近，强化学习网络和对抗网络的发展，也获得了重大的应用进展。

应该说，80 年代后期的计算机视觉研究中，不少人已经注意到神经网络的应用前景。我本人也发表过两篇文章，介绍用神经网络实现“图匹配” 图像压缩和纹理映射。为了提高精度，我曾试图将 3 层神经网络增加到 4 层、5 层，发现计算量根本无法承受而放弃。在我和张正友合写的、于 1997 年出版的《计算机视觉》一书的最后一章，即“计算机视觉系统体系结构讨论和展望”中，我写过这样一段：“非线性，自适应。自学习与人工神经网络”，“神经网络”是一种巨大的互连网络。每个神经元的结构比较简单，处理速度也不快，但却具有比神经元的个数还高出三个以上数量级的连接。这种体系结构目前已被称为体系结构理论中的连接主义，它独特的体系结构，对知识表达和信息处理过程均提出了新的思路。人工神经网络允许在理论不完善的情况下，构成一种具有自学习、自适应的体系结构，在与外界信息的交互作用中，形成一种非线性映射或非线性动力系统，以正确反映输入和输出的关系而不必预先知道这种关系的精确数学模型。当然，由于条件限制，目前的人工神经网络也只是真正的神经网络的一种“过分简化，但已确实具有一种自适应自学习的机制。”

**问**：您认为在科研创新中，大团体模式好，还是单兵作战模式好？在团队模式中，团队负责人如何协调管理个人发展及考核压力与团队工作发展要求？您对如何鼓励科研创新能力的机制有什么建议吗？

图 3 马颂德研究员（左）在欧美同学会留法分会举办的中法建交 50 周年纪念暨新春联欢会上致辞

**马老师**：在前沿科学研究中，不存在一个人领导下的所谓“大团体模式”。“大团体模式”是大型工程项目，当然，所谓“单兵作战”是另一个极端，恐怕只存在于理论物理、数学等领域。对于我们从事人工智能研究的，几个人，最多几十个人在一起研究，是国际上通用的组织形式。现在为了“有影响”，为了“便于交流”，把上游的基础研究、中游的应用基础研究和工程开发硬捏在一起，搞成大项目、大团体，不利于科研创新，也增加了许多组织成本。

**问**：您知道国内在计算机视觉领域有哪些做得好的资深前辈，他们各自有哪些特色研究呢？

**马老师**：我是 1986 年从法国回国，那时中科院自动化所已经有人开始了模式识别的研究，并在此基础上正在筹建国家模式识别重点实验室。当时有戴汝为院士从事汉字识别工作，胡启恒院士开展邮政编码识别，黄泰翼老师开始了汉语语音识别工作，洪继光老师开始了显微图像分析识别集成电路芯片的工作，这些工作虽然当时也仅仅是开始，但后来都有很重要的进展。

其他院校有影响的工作，除了我看到的你们已经采访的复旦大学吴立德老师，北方交大袁保宗老师等，应该特别提到的是以下几个老师：

✪ 北京大学石青云老师（于 2002 年逝世）是在我国最早从事图像处理、模式识别的前辈，她在北大组

建了认知国家重点实验室。实验室在图像数据库，指纹识别等方面都取得过重大进展和应用。

✪ 清华大学自动化系边肇琪老师（于 2021 年去世）长期从事模式识别研究和教育，在指纹识别和系统做出开创性研究和应用。

✪ 上海交通大学李介谷老师，主要在研究显微图像分析。

✪ 西安交通大学宣国荣老师，也是郑南宁院士的硕士生导师，研究方向是计算机控制与模式识别。他于 1986 年创立了我国第一个人工智能领域的专职科研机构“人工智能与机器人研究所”。

这些老师都是在 70 年代后期就开始了我国模式识别方面的研究。

**问**：马尔曾提出过一套计算机视觉的理论，强调局部到整体；我国陈霖院士有不同的观点，认为大范围优先。您觉得人在认知上，他们俩的观点哪个更合理呢？能谈谈您的观点吗？

**马老师**：马尔的视觉计算理论早在 1977 年提出，该理论认为，视觉是一个多级的、自下而上的信息处理过程，有一定的心理学和神经科学方面的根据，也符合西方科学的还原论思维方式。

陈霖老师的大范围拓扑特征优先，打破了“自下而上”的传统思维方式，也有一些认知心理实验的根据。现在的深度学习神经网络，最初因为 AlexNet 在 ImageNet 的识别比赛中取得了很好的效果，AlexNet 这种多层神经网络虽然不能完全准确地说清各层的数学物理意义，但大致还是可以理解为自下而上的信息处理的学习网络，和马尔所说的多级信息处理顺序是一致的。现在，不仅是识别，计算机视觉的各级任务，从图像处理、特征组织、图像分割，甚至摄像机定标、运动参数估计、三维重建、图像检索等许多定性和定量的各级工作，都可以设计专门架构的神经网络通过机器学习来实现。但这些架构的基本思路还是自下而上的，由多层神经网络通过学习来实现马尔当初描述的计算过程。当然对马尔的计算理论的理解不能绝对化，不是所有的视觉“任务”都是要完成自下而上的“全过程”，比如人脸识别、指纹识别，就不需要 2.5 维、3 维的“重建”，从特征可以直接跳到“识别”。

**问**：您觉得要做科研创新，最难的地方在哪里？

**马老师**：恐怕这本身就是一个很难回答的问题。科研中的创新，可以有大有小，有不同层次。如果说我们的研究生，尤其是博士生阶段的研究生，在一个局部问题上，能提出一种新的思路、新的方法，或者对已有方法的重大改进，也都可以说是创新。所以，只要从事的是前沿领域的研究，研究生都是可以做到的，也是我们对研究生阶段工作的要求。

对于更高层次的创新，引进全新的方法、全新的研究方向、新的网络结构、新的跨学科理论、新的学科方向，甚至有新的科学发现，都是我们研究人员梦寐以求的，这方面中国的研究人员虽然也在逐步提高，但还很不够。这里有多方面的原因，包括研究环境、研究基础、甚至人文思想方面的原因。

**问**：年轻工作者或者学生在做科研过程和个人发展中遇到困难挑战该如何应对？有什么建议？如何培养对一个研究方向长期持久的热爱，坚持不懈地做下去？能给从事计算机视觉的年轻工作者一些寄语吗？

**马老师**：任何人，做任何事都不可能一帆风顺，碰到困难或发展的低谷是“常态”，只要我们老老实实，坚持不懈的做，就不辜负自己的人生。我们从事的“人工智能”研究，从上世纪 40 年代后期计算机出现后，经历了多次起伏，高潮和低谷，几代人的努力，现在又面临一个大发展的最好时期，无论是研究和应用都处于爆发时期。我们的青年研究人员更应该顺应潮流，满腔热情地投身这一洪流中。

责任编委 张军平 明悦 贾熹滨

COMPUTER VISION NEWSLETTER

04 2021

总第 30 期

# 计算机视觉专委会简报

CCF 计算机视觉专委会