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 2023

总第 38 期

CCF 计算机视觉专委会

COMPUTER VISION NEWSLETTER

计算机视觉专委会 简报

2023 年第 04 期

总第 38 期

CCF 计算机视觉专委会

主　办　CCF 计算机视觉专业委员会

编委会

CONTENTS
简报目录

CCF 计算机视觉专委会
CCFCV.CCF.ORG.CN
CCFCVN@GMail.com

CCF-CV 走进高校系列报告会

第 129 期 沈阳航空航天大学

2023 年 9 月 21 日，由中国计算机学会计算机视觉专委会主办、沈阳航空航天大学人工智能学院承办的 CCF-CV 走进高校系列报告会第 129 期活动在沈阳航空航天大学图书馆国际报告厅成功举办。本次活动邀请了北京大学**彭宇新**教授、东北大学**王骄**教授、中国科学院自动化研究所**张俊格**研究员做特邀报告。沈阳航空航天大学人工智能学院院长**叶长龙**教授、副院长**王琳霖**教授担任本次活动的执行主席。

活动首先由沈阳航空航天大学人工智能学院院长叶长龙教授致欢迎辞。叶教授表示随着人工智能技术的快速发展，多模态感知、博弈对抗、智能决策等前沿技术受到广泛关注，成为国内外专家学者和产业界研究的焦点，本次报告会为推动人工智能技术的应用，深入探讨和交流相关技术发展前沿具有重要的意义。随后，王骄教授、张俊格研究员、彭宇新教授分别做主题报告。最后，沈阳航空航天大学人工智能学院院长叶长龙教授进行了活动总结，对三位专家的精彩报告表示感谢，同时对师生们的热情参与给予充分肯定，最后再次感谢 CCF-CV 专委会和学校对本次活动的大力支持，并祝贺本次活动取得了圆满成功！

第 130 期 山东省计算中心（国家超级计算济南中心）

2023 年 10 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、山东省计算中心（国家超级计算济南中心）承办的走进高校系列报告会第 130 期在国家超级计算济南中心成功举办。本次活动邀请了北京航空航天大学**史振威**教授、浙江大学**许威威**教授、上海大学**沈礼权**研究员、大连理工大学**刘日升**教授、天津大学**张长青**副教授做特邀报告，由山东省计算中心（国家超级计算济南中心）**韩晓晖**研究员和崔慧博士担任本次活动的执行主席。在本次报告会上，专家们围绕“图像处理与计算机视觉前沿技术及应用”主题进行了精彩的报告。

活动首先由山东省计算中心（国家超级计算济南中心）的赵大伟副主任发表了欢迎词。随后，史振威教授、许威威教授、沈礼权研究员、刘日升教授、张长青副教授分别做主题报告。参加本次活动的老师和同学认真聆听了报告，并与报告嘉宾进行了交流互动。最后，由执行主席韩晓晖研究员进行总结，他对与会人员的积极参与和贡献表示感谢，祝贺本次报告会的成功举行。

CCF-CV 走进高校系列报告会

第 131 期 苏州科技大学

2023 年 11 月 3 日，由中国计算机学会计算机视觉专委会（CCF-CV）主办，苏州科技大学电子与信息工程学院承办的走进高校系列报告会第 131 期隆重举行。本次活动邀请了南京邮电大学**刘青山**教授、清华大学自动化系**鲁继文**副主任、北京大学**林宙辰**教授、江南大学**吴小俊**教授、浙江大学**李玺**教授进行主题汇报，执行主席由苏州科技大学**胡伏原**教授，**陈静**博士，**沈忠伟**博士和**程涵婧**博士担任。在本次报告会上，专家们围绕“计算机视觉前沿技术及应用”主题进行了精彩的报告。

活动首先由刘青山教授、鲁继文教授、林宙辰教授、吴小俊教授、李玺教授分别做主题报告。最后，由执行主席胡伏原教授进行活动总结。他对与会人员的积极参与和贡献表示感谢，祝贺本次报告会的成功举行。他强调学术研究的不断前行，鼓励研究生们继续深入探讨相关问题，相互学习，共同推动相关领域的发展。最后，胡伏原教授祝愿大家在未来的学术研究中取得更加丰硕的成果。

第 132 期 中国海洋大学

2023 年 11 月 4 日下午，由中国计算机学会（CCF）主办，CCF 计算机视觉专委会（CCF-CV）和 CCF 青岛分部联合承办的走进高校系列报告会第 132 期在中国海洋大学西海岸校区成功举办。本次活动邀请了南京理工大学**肖亮**教授、中国海洋大学**何波**教授、中国科学院海洋研究所**高乐**副研究员做特邀报告，由 CCF-CV 执行委员中国海洋大学**王胜科**副教授、**仲国强**教授担任本次活动的执行主席。在本次报告会上，专家们围绕“海洋大数据智能分析”主题进行了精彩的报告。

活动首先由中国海洋大学计算机科学与技术学院曲海鹏副院长致欢迎词。曲海鹏副院长对中国海洋大学和计算机科学与技术学院进行了简介，并预祝本次报告会成功举办。随后，肖亮教授、何波教授、高乐副研究员分别做主题报告。最后，CCF-CV 执行委员、CCF 青岛秘书长王胜科副教授对本次报告会进行了总结，他对与会人员的积极参与和贡献表示感谢，祝贺本次报告会的成功举行。

CCF-CV 走进高校系列报告会

第 133 期 武汉大学

2023 年 11 月 5 日，由中国计算机学会计算机视觉专委会 (CCF-CV) 主办，武汉大学计算机学院承办的走进高校系列报告会第 133 期隆重举行。本次活动邀请了大连理工大学**卢湖川**教授、山西大学**钱宇华**教授、同济大学**何良华**教授、国防科技大学**刘新旺**教授、同济大学**史淼晶**教授、小米公司**张帆**研究员做特邀报告，由武汉大学计算机学院**杜博**教授和**武宇**教授担任本次活动的执行主席。在本次报告会上，专家们围绕“视觉基础模型和多模态感知前沿技术”主题进行了精彩的报告。

活动首先由武汉大学计算机学院院长杜博教授发表了欢迎词。随后卢湖川教授、钱宇华教授、何良华教授、刘新旺教授、史淼晶教授、张帆研究员分别做主题报告。参加本次活动的老师和同学认真聆听了报告，并与报告嘉宾进行了交流互动。最后，杜博教授进行活动总结。他对与会人员的积极参与和贡献表示感谢，祝贺本次报告会的成功举行。

第 134 期 南开大学

2023 年 11 月 19 日，由中国计算机学会计算机视觉专委会 (CCF-CV) 主办、南开大学计算机学院承办的走进高校系列报告会第 134 期活动在南开大学津南校区成功举办。本次活动邀请了北京大学**查红彬**教授、中国科学院自动化研究所**王亮**研究员、上海科技大学**虞晶怡**教授、北京交通大学**魏云超**教授做特邀报告。南开大学**杨巨峰**教授、**刘夏雷**副教授主持了本次会议。

活动首先由北京大学查红彬教授、中国科学院自动化研究所王亮研究员、上海科技大学虞晶怡教授、北京交通大学魏云超教授分别做主题报告。会议现场师生听众踊跃提问，和专家进行了深入的交流与讨论。南开大学程明明教授对本次活动进行了简短的总结。程老师再次感谢了各位专家的精彩报告，对与会人员的积极参与和贡献表示感谢，祝贺本次报告会的成功举行。

第 135 期 深圳北理莫斯科大学

2023 年 12 月 16 日下午，由中国计算机学会计算机视觉专委会(CCF-CV)主办，深圳北理莫斯科大学(以下简称“深北莫大学”)、广东省智能感知与计算重点实验室联合承办的第 135 期 CCF-CV 走进高校系列报告会——“智能感知与计算”，在深北莫大学图书馆国际中心举行。本期报告会邀请了四川大学**彭玺**教授、大连理工大学**王栋**教授、中山大学**任文琦**副教授、天津大学**王旗龙**副教授四位专家学者做特邀报告。北京理工大学计算机学院**武玉伟**长聘副教授担任本次报告会的执行主席。

活动首先由深北莫大学工程系主任、广东省智能感知与计算重点实验室主任贾云得教授致欢迎词。随后，四川大学彭玺教授、大连理工大学王栋教授、中山大学任文琦副教授、天津大学王旗龙副教授围绕智能感知与计算的前沿进展，分别做主题报告。来自深北莫大学、北京理工大学、南方科技大学、香港中文大学(深圳)和中山大学等高校的师生们聆听了 4 位专家的精彩报告。最后，报告会在热烈的掌声中圆满结束。

责任编委 朱安娜

第14期 视觉质量评价前沿进展与未来趋势

CCF-CV专委会十周年纪念活动成功举办

CCF-CV专委会十周年纪念活动2023年11月4日在南京大学苏州校区国际学术交流中心成功举办。该活动旨在回顾CCF-CV专委会伴随人工智能和计算机视觉领域的发展，自2013年成立至今十年的发展历程，分享专委会取得的重要成果，并探讨未来的发展方向。

活动开场由CCF-CV专委会副主任、南京邮电大学副校长刘青山教授主持。南京大学党委书记、中国科学院谭铁牛院士和CCF秘书长、中国科学院计算技术研究所唐卫清研究员致辞。中国人工智能学会副理事长、中国自动化学会模式识别与机器智能专委会主任、中国科学院自动化研究所刘成林研究员和中国计算机学会人工智能与模式识别专委会主任、北京交通大学于剑教授作为兄弟学会/专委会代表进行致辞。

谭铁牛院士作为 CCF-CV 专委会创始主任，回顾了专委会创立初心和十年期间取得的重大进展，强调了南京大学在计算机视觉领域的研究实力和学科优势，号召与会者和专委会委员为国家人工智能和计算机视觉事业的发展做出贡献。

唐卫清秘书长肯定了 CCF-CV 专委会在过去十年取得的重要成果，表扬了 CCF-CV 专委会的工作表现，并鼓励专委会继续带头为学会做贡献，在未来取得更大进步。

刘成林副理事长和于剑主任表示学会/专委间通过 PRCV 大会等方式紧密合作，共同见证了 CCF-CV 专委会和计算机视觉技术的飞速发展，希望未来继续携手并进，更好地推动计算机视觉、人工智能和模式识别等领域的发展，提升国家在这些研究领域的国际影响力。

随后，CCF-CV 专委会主任、北京大学查红彬教授进行了题为《专委会十周年回顾、新十年启幕》的报告，从组织建设平稳推进、特色活动有序开展、学术影响不断扩大、宣传平台稳固建设、国际声誉日益显著等方面，对专委会 2013 年以来取得的进展进行了总结和梳理，并对未来十年进行了展望。他表示下个十年专委会在学会的指导下、在各位专委会委员的共同努力下，一定会取得更大的进步与发展。

报告结束后，谭铁牛院士、查红彬教授、陈熙霖研究员三届专委会主任一同上台，开启了《计算机视觉十讲》发布仪式。《计算机视觉十讲》是首批入选的中国计算机学会计算机科学前沿系列丛书之一，由 CCF-CV 专委会牵头撰写，内容涵盖了计算机视觉领域的热点主题。

2022 年度 CCF-CV 专委工作会议顺利举办

圆桌论坛由 CCF-CV 专委会副主任、中国科学院自动化研究所王亮研究员主持，论坛的主题是：计算机视觉的过去、现在与未来。嘉宾包括：谭铁牛院士、查红彬教授、陈熙霖研究员、于海斌研究员、权龙教授、赖剑煌教授。在讨论中，嘉宾们从不同角度探讨了计算机视觉的发展历程、现状以及未来的趋势，并且指出了计算机视觉技术的挑战和机遇。总体而言，计算机视觉是一个富有活力和创新性的领域，需要我们不断地学习和探索。随着人工智能技术的不断发展，计算机视觉将会取得更多的突破和成就，为我们的生产和生活带来更多美好的变化。

CCF-CV 专委会副主任、上海科技大学虞晶怡教授主持了第二场学术论坛。论坛邀请了北京航空航天大学王蕴红教授，重庆邮电大学校长高新波教授，北京大学黄铁军教授，赢彻科技 CTO 杨睿刚博士，中国科学院计算技术研究所山世光研究员。报告题目分别为面向视觉生成的表征学习，人脸隐私保护与伪造检测，脉冲视觉：计算机视觉的一场革命，从几何规则到深度学习：3D 视觉的转变和机遇，从像素到语义的探索和实践--以“读脸”为例。

2022 年度 CCF-CV 专委工作会议顺利举办

第三场论坛由 CCF-CV 专委会候任秘书长，中国科学院计算技术研究所王瑞平研究员主持。论坛邀请了清华大学黄高副教授，西北工业大学戴玉超教授，微软亚洲研究院首席研究员胡瀚博士，上海交通大学卢策吾教授。报告题目分别为动态计算与高效视觉识别，动态场景三维重建：优化、学习与生成，面向视觉和语言的统一，具身智能“感知-想象-执行”研究。

接下来，举行了 CCF-CV 颁奖活动，由 CCF-CV 专委会秘书长、北京邮电大学马占宇教授主持。南京大学校长助理、苏州校区党工委常务副书记、管委会常务副主任索文斌致辞，祝贺 CCF-CV 专委会十周纪念活动顺利开展，欢迎各位嘉宾光临南京大学苏州校区，介绍了校区的宏图规划，并期待 CCF-CV 专委会和南京大学苏州校区未来有更加辉煌的发展。

颁奖仪式由 CCF-CV 专委会副主任、CCF-CV 提名与奖励工作组组长、南京邮电大学副校长刘青山教授主持。上海交通大学自动化系刘允才教授获得 2023 年度 CCF-CV 终身学术贡献学者，由 CCF-CV 专委会主任查红彬教授颁奖。刘允才教授发表了获奖感言，感谢 CCF-CV 专委会对自己在计算机视觉领域的工作成果给予肯定和荣誉，期待年轻学者在专委平台上具有更好的发展，做出更多的杰出贡献。

本年度杰出成就学者授予了 CCF-CV 专委会顾问委员会委员、西北工业大学副校长张艳宁教授，由 CCF-CV 专委会候任主任陈熙霖研究员颁奖。张艳宁教授感谢了专委平台、三届主任、专委会委员以及其团队的支持，表示将继续为国家战略需求和平台发展做出贡献。

2023 年度 CCF-CV 持久影响力工作授予了发表于 IEEE TPAMI 2013 的一项工作：Robust Recovery of Subspace Structures by Low-rank Representation，作者为 Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Jun Sun, Yong Yu, Yi Ma。第一作者刘光灿教授接受了重庆邮电大学校长高新波教授的颁奖。

2023 年度 CCF-CV 中科视拓 Seeta 学术新锐学者授予了清华大学刘世隆、上海交通大学孙健华、南开大学郑兆晖三位同学，由 CCF-CV 常委会委员、中山大学赖剑煌教授颁奖。

本年度的 CCF-CV 中科视拓 Seeta 服务贡献奖颁发给了五位专委会执行委员：大连理工大学樊鑫，东北大学贾同，厦门大学孙晓帅，西安电子科技大学王楠楠，北京航空航天大学于茜，武汉理工大学朱安娜。CCF-CV 常委会委员、中国科学院计算技术研究所山世光研究员为获奖委员颁奖。

专委会主任查红彬教授感谢中科视拓（北京）科技有限公司为 CCF-CV 奖项提供赞助。

最后，CCF-CV 专委会主任查红彬教授对 CCF-CV 专委会十周年活动进行了总结。查主任表示，十周年纪念活动是对专委会过去十年发展的回顾，同时也是对未来十年工作的展望和启航，期待在新的十年里 CCF-CV 专委会平台在专委会主任带领下有大发展、新贡献。

本次 CCF-CV 专委会十周年纪念活动受到了新华日报（交点汇）、苏州日报（引力播）、苏州电视台等多家媒体的报道。

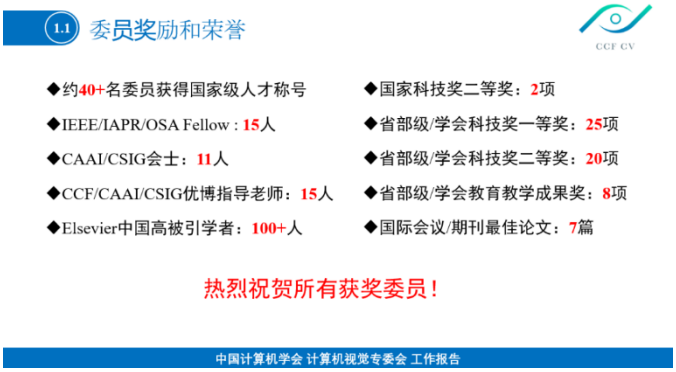
责任编委 毋立芳

2023 年度 CCF-CV 专委工作会议暨换届选举顺利召开

计算机视觉专委会（CCF-CV）年度工作会议暨换届选举于 2023 年 10 月 14 日下午在厦门国际会议中心顺利举办。来自全国高校、科研院所、企业的现任执行委员共计 289 位参加了会议。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。查主任代表 CCF 计算机视觉专委会欢迎并感谢百忙之中现场参加会议的各位委员和 CCF 学会代表、华南理工大学副校长许勇教授，肯定了本届专委会过去四年中通过线上和线下的方式在学术交流及社会服务等多方面取得的丰硕成果，鼓励委员们利用好专委会这个平台聚焦学术前沿、做出高质量的研究工作，希望专委会大家庭加强互动交流合作，进一步提升各项专委活动品牌质量，为推进国家计算机视觉学术研究继续发挥积极的引领作用。

接下来，专委会秘书长、北京邮电大学马占宇教授向各位执行委员做了本届专委会工作报告。报告简要介绍了专委会的组织结构、党的工作小组、顾问委员会、国际顾问委员会，通报了专委会组织的学术活动、企业交流与教育、宣传与联络平台，总结了过去四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，指出了工作中存在的问题和改进方案，最后介绍了专委会未来工作计划。



根据会议日程，本届会议进行了专委会的换届选举工作，由专委会指导委员会委员、深圳北理莫斯科大学贾云得教授主持了选举过程。贾教授通报了本次换届选举提名工作组成员、选举工作组成员，宣读了 CCF 学会制定的专委会主任、副主任、秘书长和常务委员选举规则，以及本次换届选举流程。

随后，中国计算机学会（CCF）代表、华南理工大学副校长许勇教授代表学会致辞。许教授介绍了 CCF 学会的专委会换届选举要求和兄弟专委会选举案例，以及 CCF 代表对专委换届选举工作的监督职责，并预祝专委会选举工作顺利成功。

按照 CCF 学会选举规则，本年度共有 2 位委员申请竞选主任、4 位委员申请竞选副主任、2 位委员申请秘书长，20 位委员申请常务委员。按照流程，候选人先后上台进行了竞选陈述，并回答了现场委员的质询。经过现场参会的现任执行委员投票表决和 CCF 学会代表确认，选出了新一届专委会主任、副主任、秘书长，以及常务委员。

2022 年度 CCF-CV 专委工作会议顺利举办

专委会主任查红彬教授对本次会议进行了总结发言。查主任首先祝贺新当选的新一届专委班子，指出专委会工作将主要围绕两件事情：一是如何促进专委委员之间、学术界与工业界之间更深层次的互动合作，二是如何加强国际同行间的交流合作。查主任期待专委会平台得到更大提升，促进国内计算机视觉研究领域更深层、更高水平的学术交流。

查主任再次感谢现场参加会议的执行委员，并祝愿大家万事如意、身体健康。最后，专委会 2023 年度工作会议暨换届选举圆满结束，期待明年在乌鲁木齐再聚！

责任编委 毋立芳

CCF-CV 常务委员会 2023 年度第二次工作会议顺利召开

2023 年 11 月 3 日于苏州召开中国计算机学会计算机视觉专委会 (CCF-CV) 常务委员会 2023 年度第二次工作会议，本次工作会议由专委会主任、北京大学查红彬教授主持，专委会顾问委员会委员、候任主任、中科院计算研究所陈熙霖研究员和常委会委员参会，秘书处成员列席。

首先，专委会党小组成员、常委会委员、中科院计算研究所山世光研究员组织了党小组学习。

接下来，查红彬主任和陈熙霖候任主任带领常委会委员，就专委会执行委员提出、由秘书处收集的热点议题进行了讨论，包括如何加强专委会委员之间、学术界与工业界、计算机视觉研究不同地域之间的合作交流，如何提升专委会开展的各项特色活动，以及有哪些途径扩大专委会国际化等内容，形成了具体可行的指导性建议。

随后，查红彬主任、陈熙霖候任主任和常委会委员就学术研究价值和技术产业转化等话题进行了自由讨论。

最后，查红彬主任作了总结发言。会议在紧张而有序的热烈讨论氛围中结束。

责任编委 黄岩

专题综述

基于任意粒度语言描述的图像着色方法

北京邮电大学 常征 张沛瑄 李思
北京大学 翁书晨 施柏鑫

本文是北京邮电大学与北京大学团队合作研究的成果，发表在NeurIPS 2023并获得Spotlight的工作L-CAD[1]（Language-based Colorization with Any-level Descriptions）。论文研究的问题是基于语言描述的图像着色。该任务要求模型能够在用户友好的自然语言描述指导下为灰度图像添加结构合理且主观视觉效果满意的颜色。先前的方法假设用户为图像中的大多数物体提供全面的颜色描述，而忽略了只描述主要物体、甚至完全缺少描述的情况。论文提出了一个统一的模型，用于解决基于任意粒度语言描述的图像着色问题。该方法利用预训练的跨模态生成模型，凭借其强大的语言理解能力和丰富的颜色先验知识来处理描述粒度的不确定性。该方法进一步设计了语义对齐模块，以使着色结果保持和输入灰度图一致的局部空间结构并防止鬼影效果。通过提出的新型采样策略，所提出的模型能够在多样且复杂的场景中实现实例感知的着色效果。广泛的实验结果显示了论文提出方法的优势，包括有效处理任意粒度的描述，以及在基于文本条件和自动着色方面超越其他模型的表现。

一、研究背景

图像着色是一项具有挑战性的任务，其目的在于将灰度图像转换为合理且视觉上令人愉悦的彩色图像。基于语言的着色方法[2, 3, 4]利用自然语言描述作为指导，以产生更可控的彩色图像。这些方法能够满足用户的特定需求，使他们能够提供更具体和细腻的颜色偏好。由于自动着色[5,6,7]在为常见对象（例如，花的颜色）确定颜色时经常遇到歧义，基于语言的着色在生成高质量和可定制的彩色图像方面显示出了令人满意的结果，且具有对用户友好的交互方式。尽管基于语言描述的着色方法通过特征融合[4, 8]、解耦颜色-对象空间[2, 9]和聚合相似区块[3]改进了语言描述与彩色化结果之间的一致性，但它们隐式地假设用户为图像中的大多数物体提供全面的颜色描述。这种假设通常导致次优的性能，尤其是对于没有相应颜色描述的对象。此外，本文观察到用户通常只为他们感兴趣的物体分配颜色。

图 1　基于任意粒度语言描述的图像着色

为了用具有不同粒度的语言描述给图像着色，本文提出了一个统一的模型，它能够自适应地理解任何粒度的描述，并按如下方式着色：(1) 对于包括所有物体的“完全”粒度描述，模型会根据用户的要求精确着色 (图 1 第一行，对四杯鸡尾酒的详尽描述)；(2) 对于只关注感兴趣物体的“部分”粒度描述，模型会根据图像语义来着色未提及的物体（图 1 第二行，仅对罐子和花朵的选择性描述)；(3) 对于缺乏有意义的颜色信息的极简粒度描述，它会切换到自动上色模型（图 1 第三行，遗漏了对披萨和餐厅的描述)。

为了实现上述目标，论文提出了 L-CAD，用于完成

基于任意粒度语言描述的图像着色。鉴于任意粒度描述中提及的物体存在固有的歧义，该方法利用了预训练的跨模态生成模型（即，Stable Diffusion[10]），通过其强大的语言理解能力以及丰富颜色先验知识，来帮助完成基于任意粒度描述的着色。然而，由于生成模型并不是专门为上色而设计的，它在与输入灰度图的空间对齐方面面临挑战。为了解决这个问题，该方法设计了一个灰度图指导的隐变量解码模块，它使着色结果在像素空间中与灰度图像对齐，保留了灰度图中的局部空间结构。此外，该方法通过通道扩展的卷积运算，使隐空间中的特征图与语言描述对齐，以防止鬼影的出现。另外，为了处理具有不同粒度和复杂场境下的描述，论文提出了一种实例感知的采样策略，它在隐空间中粗略估计物体轮廓并将颜色特征分配给它们对应的区域。由于这些改进，该方法可以不受描述中提及物体数量的限制，能够有效处理任意粒度的描述。

二、L-CAD方法介绍

2.1 前言：扩散模型

扩散模型[11, 12]作为一种生成模型，在图像生成方面取得了显著的成就。在前向过程中，随机采样一个高斯噪声对原图进行加噪：

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon_t$$

在反向过程中，它训练一个神经网络ϵ_θ来预测噪声。通过加入额外的条件，扩散模型能够生成与给定条件一致的结果，以完成条件生成任务。条件扩散模型的训练过程中，仍然使用均方误差作为损失函数：

$$\mathcal{L}_{dm} = \mathbb{E}_{t,x_0,\epsilon\sim\mathcal{N}(0,1)}[\| \epsilon_t - \epsilon_\theta(x_t, t, y) \|^2]$$

为了缓解高分辨率图像生成时的资源消耗，Stable Diffusion[10] 引入了感知压缩模型，将像素空间的图像压缩到隐空间，这使得扩散过程可以在尺度更小的隐空间中进行。具体来说，它采用了一个压缩编码器$\mathcal{E}$来将给定的图像x映射到隐空间$z = \mathcal{E}(x)$，并使用一个压缩解码器$\mathcal{D}$来重建图像$\tilde{x} = \mathcal{D}(z)$。这样，噪声预测网络的训练目标变成了学习隐变量$z$的分布，而不是图像$x$。

该方法的目标是利用 Stable Diffusion 的强大语言理解能力和丰富的颜色先验知识，实现基于任何级别描述的语言驱动的着色。

2.2 灰度图指导的隐变量解码

尽管 Stable Diffusion 在文本生成图像任务中展现出卓越性能，但它缺乏在着色任务中保留输入灰度图像局部空间结构的能力。为了解决这一问题，该方法提出在像素空间中加入额外的灰度图编码器。如图 2 所示，彩色图像被压缩编码器映射成隐变量，而灰度编码器从灰度图像中提取多尺度特征，保留局部结构语义。之后，该方法将这些特征使用跳跃链接的方式直接加到压缩解码器的相应尺度中，引导从隐空间到像素空间的解码过程。灰度编码器采用与压缩编码器相同的架构，其压缩编码器和解码器的权重固定，以保留来自预训练模型的先验知识。

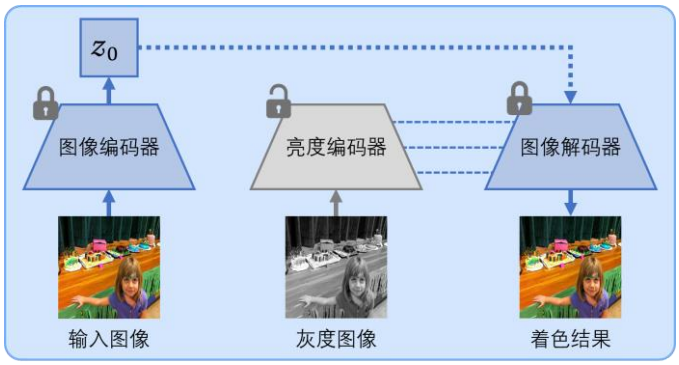


图 2　灰度图引导的隐变量解码

因为观察到在具有规则形状和锐利结构的区域中的错误像素显著损害了视觉感知，该方法估计了一个局部误差图$M^{art}_{h,w}$，并将其用于指示在解码结果中特定空间位置(h,w)遇到错误像素的概率：

$$M^{\text{art}}_{h,w} = \sum_{p\in\Omega_{(h,w)}}\left(\frac{\delta_p-\mu_p}{N_{win}}\right)^2,\ \mu_p = \sum_{p\in\Omega_{(h,w)}}\left(\frac{\delta_p-\mu_p}{N^2_{\text{win}}}\right)^2$$

其中，p表示以(h,w)为中心的局部窗口Ω的位置索引，N_{win}是局部窗口的大小。之后，该方法将局部误差图作为一个权重应用于图像重建损失中：

$$\mathcal{L}_{rec} = \| M^{\text{art}} \odot (x - \tilde{x}) \|_1$$

该方法在像素空间训练模型的总损失如下：

$$\mathcal{L}_{pix} = \mathcal{L}_{\text{rec}} + \alpha\mathcal{L}_{\text{dis}} + \beta\mathcal{L}_{\text{per}}$$

其中$\mathcal{L}_{dis}$为判别器损失，$\mathcal{L}_{per}$为感知损失。

2.3 隐空间语义对齐

在 Stable Diffusion[10]中，语言描述通过交叉注意力机制被注入隐空间的特征图中。然而在图像着色任务中，颜色特征被分配到预期区域之外会使生成的结果中产生鬼影。因此，需要在隐空间内保持语言描述和灰度图像之间的语义对齐。

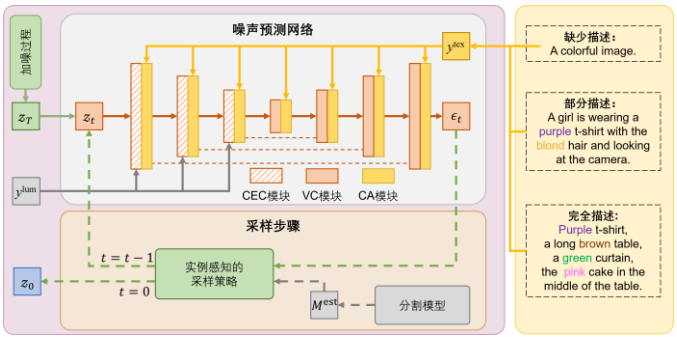


图 3　隐空间语义对齐

如图 3 所示，该方法用通道扩展卷积（CEC）模块替换了噪声预测网络中的普通卷积(VC)模块。这些 CEC 模块位于下采样网络内，它们接收来自像素空间的灰度编码器的亮度特征y^{lum}作为额外的引导。通过利用扩展的通道，CEC 模块可以有效地捕获隐空间中灰度的局部结构语义。VC 模块和 CEC 模块的区别如图 4 所示：

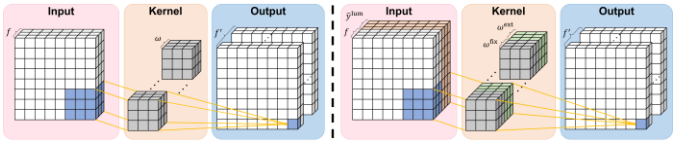


图 4　VC 模块和 CEC 模块

从数学角度来看，CEC 模块等同于使用卷积操作来提取特征，并将它们的输出加到下采样网络中。为了保留预训练生成模型的强大语言理解能力和丰富的颜色先验知识，该方法在训练过程中保持原始通道的参数w^{fix}不变。此外，扩展通道的权重w^{ext}被初始化为零，以确保在训练之前的模型保持与预训练的生成模型的功能等效性。该方法在下采样和上采样模块之间使用跳跃连接，以引导上采样过程中的亮度特征，训练损失为：

$$\mathcal{L}_{lat} = \mathbb{E}_{t,z_0,\epsilon\sim\mathcal{N}(0,1)}[\| \epsilon_t - \varepsilon_\theta(z_t, t, y^{\mathrm{tex}}, y^{\mathrm{lum}}) \|^2]$$

2.4 实例感知的采样策略

为确保语言描述中的颜色能准确地体现在图像里指定的物体上，本文提出了一种实例感知的采样策略。该策略采用了一个外部的分割模型（针对文本描述中提到的物体或区域进行分割，例如 SAM[13]）来估计描述中提到的物体的轮廓M^{est}。考虑到在第l个 CA 模块中的注意力图M_l^{att}控制每个颜色词的着色区域，该方法使用 Sigmoid 函数对注意力图归一化，并通过迭代优化来使它与缩放后的物体的轮廓$\widehat{M}^{\mathrm{est}}$对齐。之后，该方法应用 DDIM[12]完成去噪过程，如算法 1 所示。

算法 1：实例感知的采样策略

输入：粗略估计的物体轮廓 M^{est}
输出：着色过的隐变量 z_0
For $t = T \cdots 1$ do:
　　$\_, M_*^{\mathrm{att}} = \epsilon_\theta(z_t, t, y^{\mathrm{lum}}, y^{\mathrm{tex}})$
　　For $l = 1 \cdots L$ do:
　　　　$\widehat{M}_l^{\mathrm{est}} \leftarrow \mathrm{Downsampling}(M^{\mathrm{est}}, l)$
　　　　$\mathcal{M} \leftarrow \mathrm{Sigmoid}(M_l^{\mathrm{att}})$
　　　　$\widehat{M}_l^{\mathrm{att}} \leftarrow M_l^{\mathrm{att}} - \lambda\nabla_{\mathcal{M}}\mathcal{L}_{\mathrm{BCE}}(\mathcal{M}, \widehat{M}_l^{\mathrm{est}})$
　　End
　　$\hat{\epsilon}_t, \_ = \epsilon_\theta(z_t, t, y^{\mathrm{lum}}, y^{\mathrm{tex}})\{M_*^{\mathrm{att}} \leftarrow \widehat{M}_*^{\mathrm{att}}\}$
　　$z_{t-1} = \mathrm{DDIM}(z_t, \hat{\epsilon}_t, t)$
End

三、实验结果

该方法在基于语言的图像着色数据集上进行实验：（1）扩展的COCO-Stuff数据集[9]，该数据集是在COCO-Stuff数据集[14]的基础上构建的，包括59K训练图像和2.4K评估图像；（2）多实例数据集[3]，它提供了在单一图像内有多个不同实例的样本，包括65K训练图像和7K评估图像。对于这两个数据集，每个图像都有相应的语言描述。

3.1 与基于语言的着色模型对比

L-CAD 与基于语言的图像着色方法进行了比较，例如 ML2018 [4]、L-CoDe [9]、L-CoDer [2]以及 L-CoIns [6]。先前基于语言的图像着色方法假设用户提供全面的颜色描述，因而在处理部分描述的样本时性能降低。相比之下，该方法利用了 Stable Diffusion[10]的先验知识以及新提出的实例感知采样策略，即使在描述程度不的情况下也能展现出生动的着色结果，如图 5 所示。

图 5　提出方法与基于语言的着色模型对比

3.2 与自动着色模型对比

L-CAD 还与全自动的图像着色方法进行了比较，例如 CIC[7]，InstColor[5]，ChromaGAN[15]，BigColor[16]，DISCO[17]以及 CT² [6]。如果用户有特殊要求，自动着色方法无法按照用户的需求改变图像中物体的颜色，只能按照数据集中物体的颜色分布着色，如图 6 所示。

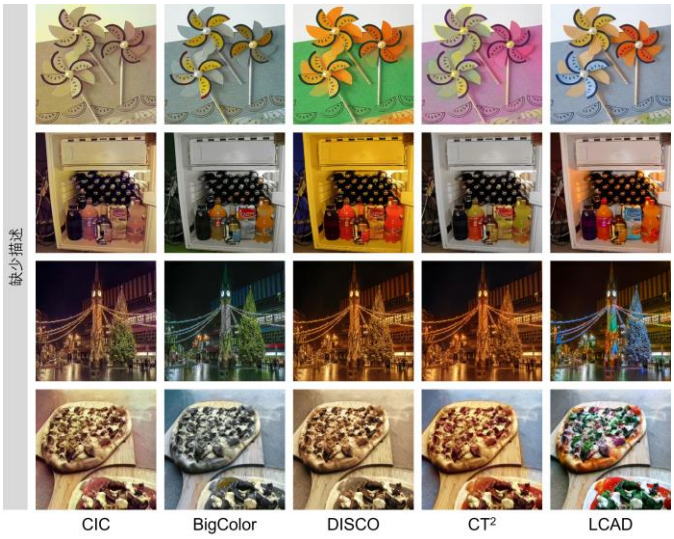


图 6　提出方法与自动着色模型对比

3.3 与基于扩散模型的图像编辑方法对比

一些图像编辑方法例如 ControlNet[18]、Pix2PixZero[19]和 SDEdit[20]也能够基于预训练的扩散模型，使用描述来编辑图像。然而，这些方法并不是专门为着色任务设计的，这导致它们在保持局部空间结构、利用颜色先验以及学习物体与颜色词之间的对应关系方面存在挑战。这些限制使得它们难以完成任意粒度描述指导的实例感知着色。图 7 展示了它们的不适用性。

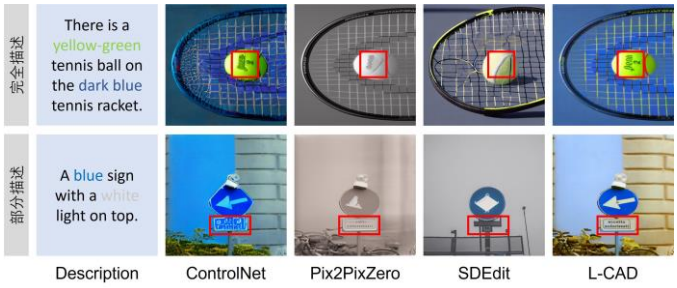


图 7　提出方法与基于扩散模型的图像编辑方法对比

3.4 消融实验

为了研究该方法提出的模块和采样策略的影响，论文创建了三个基准模型进行消融实验。我们在图 8 中展示了消融实验的定性结果。

灰度引导的图像压缩（LIC）。该实验禁用了像素空间中的亮度编码器，导致模型缺少多尺度灰度特征进行引导，进一步使得模型无法正确保持局部空间结构。

语义对齐的隐变量（SLR）。在移除了指导隐空间语义对齐的灰度特征，并将扩展通道的卷积块替换为普通的卷积模块以后，着色图像中出现了明显的鬼影。

实例感知采样策略（ISS）。用标准的 DDIM[12]替换了实例感知采样策略以后，模型正确地根据文本描述的为物体分配颜色的性能显著降低了。

图 8　消融实验的定性结果展示

四、总结

该论文提出了一种基于任意粒度语言描述的图像着色模型。利用预训练模型的先验知识，其设计了新颖

的模块，以保持局部空间结构并防止出现鬼影，并进一步提出了实例感知采样策略，在复杂场景下实现实例感知的图像着色。与基于语言的着色方法进行比较的结果展示了该模型可以处理任何粒度的语言描述，这些粒度包括完全描述和部分描述，以及缺少描述。定性和定量结果都证明了该模型的优越性能。

五、前期相关工作

在L-CAD发表之前，北京邮电大学与北京大学的团队围绕语言引导的图像着色算法已经发表一系列工作，包括：L-CoDe[9]（Language-based Colorization using Color-object Decoupled Conditions）提出了颜色与对象解耦条件的着色网络，解决了颜色与对象耦合与不匹配的问题；L-CoDer[3]（Language-based colorization with Color-object Decoupling transformer）首次将transformer引入了到基于语言的图像着色任务中，统一了颜色描述与图像实例的特征表示；L-CoIns[4]（Language-based Colorization with Instance awareness）进一步设计了自动聚合的分组机制，强化了着色模型里实例感知的能力。

5.1 L-CoDe

基于语言描述的图像着色方法普遍面临颜色与物体的耦合和不匹配的问题：颜色与物体的耦合会导致难以将香蕉着色为红色，因为模型未曾见过红色的香蕉；而颜色与物体不匹配的问题会导致未被描述的物体会被错误地着色为其他物体的颜色。这些物体导致现有的模型难以准确地将句子中描述颜色的形容词映射为图像中指定物体的颜色。

为解决以上问题，L-CoDe 提出了颜色与对象解耦条件的着色网络。为了解决颜色与对象耦合的问题，其引入了颜色物体对应矩阵的预测器和新颖的注意力转移模块，这确保了颜色描述的特征被注入到对应物体的图像区域上；同时，为了解决颜色与物体不匹配问题，其采用了一个软门控的注入模块来有效地过滤掉不对应的颜色引导。进一步，它还提出了一个包含标注的“颜色与物体对”新数据集，以提供监督信号解决耦合问题。

5.2 L-CoDer

尽管L-CoDe在基于语言的着色任务中取得了一定的进展，但其性能受到三个关键问题的限制。（1）图像和语言属于不同的模态，因而提取的特征之间存在较大差距，这增加了理解语言描述的难度。（2）图像特征表示的语义在网络中是由局部到全局逐层演化的，单一的语言特征难以与多尺度的图像特征匹配，这降低了颜色表示的准确性。（3）基于卷积神经网络设计的着色模型往往是基于局部感知的，因此在着色局部亮度变化强烈的区域时容易出现伪影。

L-CoDer 首次将 transformer 引入了到基于语言的图像着色任务中，同时保持语言条件中颜色物体的解耦，以解决上述问题。L-CoDer 使用的 transformer 结构将图像与语言的特征统一表示为 tokens，并进一步支持语言描述中的颜色条件根据图像特征的变化的自适应调整。此外，由于 transformer 具备全局感受野，L-CoDer 对局部强烈的亮度变化具有鲁棒性。

5.3 L-CoIns

尽管 L-CoDe 和 L-CoDer 引入额外的标注来防止颜色-对象耦合和不匹配问题，但它们仍然难以处理包括多个不同实例的场景。例如，一张多个人物的合照，且每个人的衣着被语言描述指定为不同的颜色。

CoIns引入了多个可学习的分组向量，并提出了一个自动聚合具有相似颜色图像块的分组机制，促使分组向量能够自适应地表示图像中的多个实例，并最终能够在没有任何外部先验引导的情况下实现了实例感知的图像着色。此外，L-CoIns还提出了的亮度增强和颜色对比损失，打破了亮度和颜色词之间的统计相关性，驱动模型合成与语言描述更加一致的颜色。该工作进一步收集了一个多实例数据集，为同一图像中的多个实例提供了详细的语言描述。

责任编委 王金甲

参考文献

[1] Z. Chang, S. Weng, P. Zhang, Y. Li, S. Li, and B. Shi. L-CAD: Language-based Colorization with Any-level Descriptions using Diffusion Priors. In NeuIPS 2023.

[2] Z. Chang, S. Weng, Y. Li, S. Li, and B. Shi. L-CoDer: Language-based colorization with color-object decoupling transformer. In ECCV, 2022.

[3] Z. Chang, S. Weng, P. Zhang, Y. Li, S. Li, and B. Shi. L-CoIns: Language-based colorization with instance awareness. In CVPR, 2023.

[4] V. Manjunatha, M. Iyyer, J. Boyd-Graber, and L. Davis. Learning to color from language. In NAACL, 2018.

[5] J.-W. Su, H.-K. Chu, and J.-B. Huang. Instance-aware image colorization. In CVPR, 2020.

[6] S. Weng, J. Sun, Y. Li, S. Li, and B. Shi. CT2 : Colorization transformer via color tokens. In ECCV, 2022.

[7] J. Chen, Y. Shen, J. Gao, J. Liu, and X. Liu. Language-based image editing with recurrent attentive models. In CVPR, 2018.

[8] Zhigang Li, Gu Wang, Xiangyang Ji. CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation. In ICCV, 2019.

[9] S. Weng, H. Wu, Z. C. Chang, J. Tang, S. Li, and B. Shi. L-CoDe: Language-based colorization using color-object decoupled conditions. In AAAI, 2022.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022.

[11] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In NeurIPS, 2020.

[12] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In ICLR, 2021.

[13] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. arXiv preprint arXiv:2304.02643, 2023.

[14] H. Caesar, J. Uijlings, and V. Ferrari. COCO-Stuff: Thing and stuff classes in context. In CVPR, 2018.

[15] P. Vitoria, L. Raad, and C. Ballester. ChromaGAN: Adversarial picture colorization with semantic class distribution. In WACV, 2020.

[16] G. Kim, K. Kang, S. Kim, H. Lee, S. Kim, J. Kim, S.-H. Baek, and S. Cho. BigColor: Colorization using a generative color prior for natural images. In ECCV, 2022.

[17] M. Xia, W. Hu, T.-T. Wong, and J. Wang. Disentangled image colorization via global anchors. TOG, 2022

[18] L. Zhang and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[19] G. Parmar, K. K. Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu. Zero-shot image-to-image translation. arXiv preprint arXiv:2302.03027, 2023.

[20] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In ICLR, 2021.

常征

北京邮电大学人工智能学院 2021 级硕士研究生，导师为李思副教授，主要研究方向为图像合成。

Email:zhengchang98@bupt.edu.cn

张沛瑄

北京邮电大学人工智能学院 2023 级硕士研究生，导师为李思副教授，主要研究方向为图像合成。

Email: pxzhang@bupt.edu.cn

李思

北京邮电大学人工智能学院副教授，博士生导师。北京邮电大学博士，美国布兰迪斯大学博士后，曾在新加坡国立大学、日本国立信息学研究所从事科研工作。研究方向为多模态智能信息处理，在 TPAMI、CVPR、ICCV、NeurIPS 以及 ACL 等期刊和会议上发表论文多篇。主持和参与国家自然科学基金、科技创新 2030—“新一代人工智能”重大项目课题、北京市自然科学基金等多个项目。

Email:lisi@bupt.edu.cn

翁书晨

北京大学计算机学院 2019 级博士生，导师为施柏鑫研究员，主要研究方向为跨模态图像编辑。

Email:shuchenweng@pku.edu.cn

施柏鑫

北京大学计算机学院多媒体信息处理全国重点实验室、视频与视觉技术国家工程研究中心研究员、博士生导师（“博雅青年学者”）；北京智源人工智能研究院青年科学家。2013 年博士毕业于日本东京大学，曾先后在麻省理工学院媒体实验室、新加坡科技设计大学、南洋理工大学、日本国立产业技术综合研究所从事研究工作。研究方向为计算摄像学与计算机视觉，发表论文 180 余篇（包括 TPAMI 论文 23 篇，计算机视觉三大顶级会议论文 69 篇）。论文获评国际计算摄像会议（ICCP）2015 年 Best Paper - Runner Up、入选 IJCV 专刊 Best Papers from ICCV 2015，2021 年获得日本大川研究助成奖。主持科技创新 2030—“新一代人工智能”重大项目、国家自然科学基金重点、国家级青年人才等多个项目。担任国际顶级期刊 TPAMI、IJCV 编委，顶级会议 CVPR、ICCV 领域主席。IEEE、CCF、CSIG 高级会员，APSIPA 杰出讲者。

Email:shiboxin@pku.edu.cn

热点追踪

动态场景新视角合成

西北工业大学 郭相 戴玉超

一、引言

新视角合成（NVS）是计算机视觉和图形学中一个长期且具有挑战性的问题，在虚拟现实、增强现实、数据增强、图像编辑等领域有很多应用。最近，可微神经渲染技术[1][2][3]特别是神经辐射场（NeRF）[1]的引入在短时间内极大地推动了这一领域的快速发展，并引起了广泛关注。NeRF[1]通过多层感知器（MLP）表示三维世界，将输入的三维坐标和视角方向映射到对应的不透明度和颜色，从而通过渲染生成逼真的图像。

最初的 NeRF 只能对静态场景建模，一系列工作将基于 NeRF 的框架从静态场景扩展到了动态场景[4][5][6][7][8][9][10][11]。其中一个很有前景的思路是使用规范空间表示法[8][9][12]。这种表示法将一个时刻设置为规范时刻，并用神经辐射场对规范时刻的静态场景进行建模。为了渲染其他时刻的图像，需要使用一个变形场来估计三维点从当前时刻移动到规范时刻的后向流。虽然基于后向流的规范表示法很容易实现，但后向流场是非光滑的。如图 1(b)所示，对于时间轴上的一个固定三维位置 p，会有不同类型的物体点覆盖位置 p，这就需要不连续的后向流将它们映射回规范空间（图 1(d)）。因此，常用的平滑运动模型（如 MLP）无法很好地拟合后向流。此外，由于运动模型的失效，规范空间的静态场景模型也会发生扭曲变形。

为了解决后向流的问题，提出使用前向流（Forward Flow)作为变形模型。通过使用前向变形流，将整个规范空间的辐射场从规范时刻翘曲到其他时刻，并在相应的时刻进行渲染。这样，对于时间轴上的同一个位置，变形模型估计的前向流将是平滑和连续的（图 1(c)和(e)）。需要注意的是，SNARF[13]使用了基于皮肤模型的前向翘曲，但它是为动态人体建模而设计的，不能用于一般场景。我们的目标是实现一般场景的动态建模，这意味着我们必须对整个空间进行翘曲。

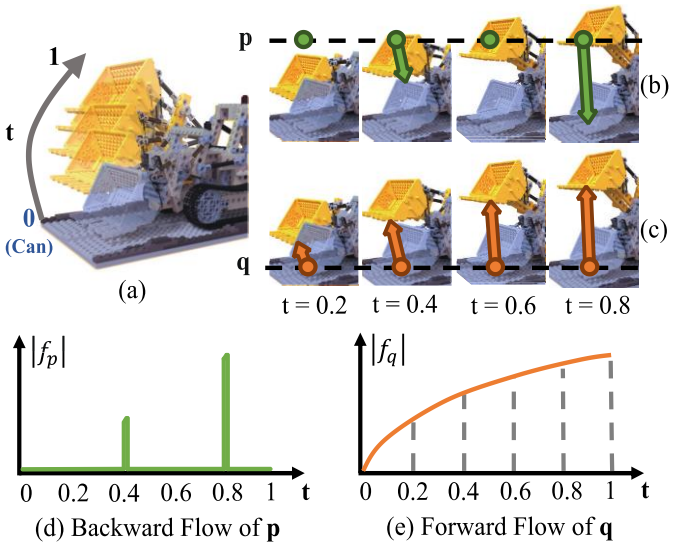


图 1 后向流 vs 前向流。本图展示了后向流和前向流变化的示例。(a) 动态场景示例。(b) 随着铲斗升起，不同类型的点覆盖了绿色的位置 p，这就需要非常不同的后向流来将这个点映射回规范空间。(d) 位置 p 后向流的模长随时间的变化不平滑。(c) 位置 q 的前向流将一个特定物体点从规范空间映射到其他时间，它是平滑和连续的。(e) 显示了位置 p 的前向流的模长变化是平滑的。

然而，将前向翘曲引入基于规范空间的动态 NeRF 方法仍有三个主要问题亟待解决。首先，现有方法中的传统辐射场无法进行显式的翘曲，这是因为辐射场是由 MLP 参数化的连续函数表示的。为了解决这个问题，我

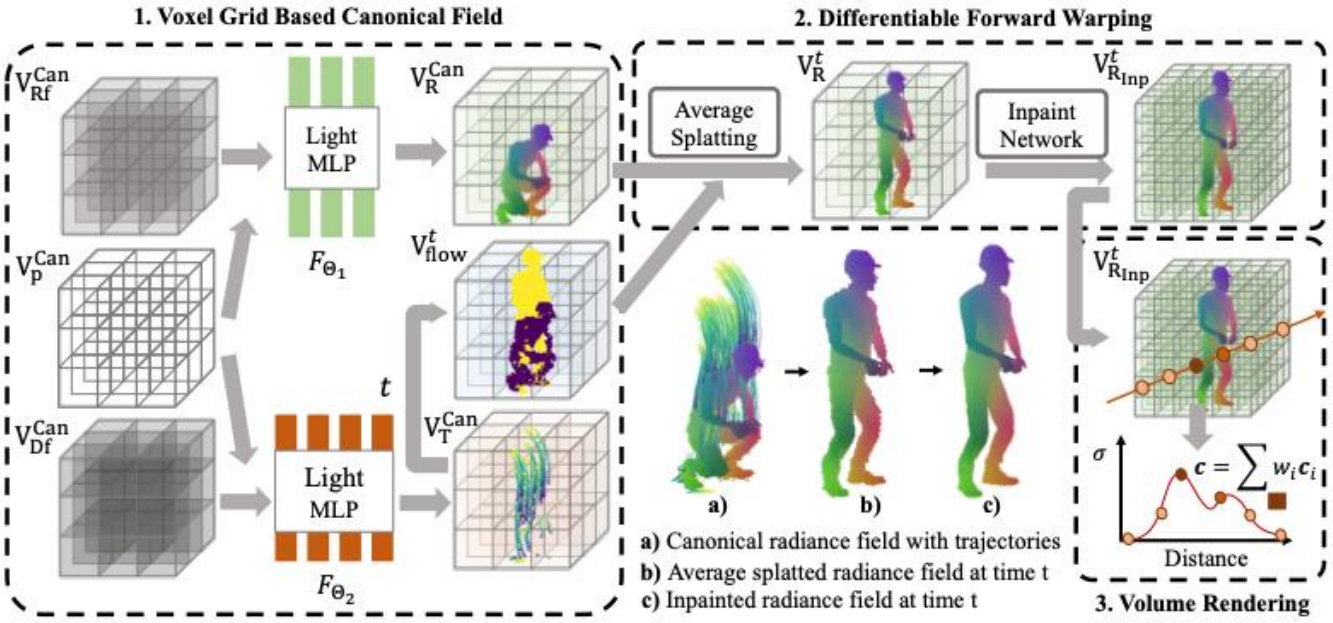


图 2　方法框架。a）我们用基于体素网格的辐射场来表示规范空间静态场景的不透明度和颜色，基于体素网格的轨迹场来表示变形；b）我们提出使用平均拼接的前向流对规范空间的静态辐射场进行前向翘曲；c）使用填充网络对翘曲后的辐射场进行填充。 具体来说： 1. 基于体素网格的规范空间中包含两个模型，包括了一个基于体素网格的静态辐射场和一个基于体素网格的轨迹场。静态辐射场V_R^{Can}是由一个 MLP 估计的，它将辐射特征 V_{Rf}^{Can} 和相应的三维坐标 V_p^{Can}作为输入，估计辐射场三维点的颜色和不透明度。轨迹场 V_T^{Can} 由另一个 MLP 估计，它将形变特征V_{Df}^{Can}和坐标V_p^{Can}作为输入，估计每个点的轨迹，然后就可以得到从规范时刻到时间 t 的前向流V_{flow}^t。 2. 可微分的前向翘曲操作首先将规范空间的辐射场V_R^{Can}根据V_{flow}^t翘曲，得到时间 t 的辐射场 V_R^t。 然后通过填充网络得到填充后的辐射场$V_{R_{Inp}}^t$；3. 基于辐射场$V_{R_{Inp}}^t$，使用体积渲染（Volume Rendering）来渲染 t 时刻的图像。

们提出使用体素网格来表示规范空间辐射场，因为它是离散的并且有限的。基于体素的方法[14][15][16]已经证明了这种表示方法的有效性。另外两个问题是前向翘曲操作的固有特性带来的多对一和一对多映射问题。针对这两个问题，我们提出了一种由平均拼接操作和填充网络组成的可微分前向翘曲方法，分别解决了多对一和一对多的问题。

二、动态场景新视角合成相关工作

将 NeRF 从静态场景扩展到具有非刚性可变形物体的动态场景是一个热门的研究领域。一种可行的方法是构建 4D 时空表示。例如，Yoon 等人[17]结合了单视角深度和多视角立体深度来渲染具有三维变形的新视角图像。Gao 等人[5]使用时间不变模型（静态）和时间变化模型（动态）来表示场景，并通过场景流估计对动态模型进行正则化。NeRFlow[4]从一组 RGB 图像中学习动态场景的 4D 时空表示。Xian 等人[11]建立了一个 4D 时空辐射场，将时空位置映射到点的颜色和不透明度。同样，NSFF[6]将动态场景建模为关于外观、几何结构和三维场景运动的连续函数。DCT-NeRF[10]使用离散余弦变换(DCT)捕捉动态运动，即学习空间中每个点随时间变化的平滑稳定轨迹。

另一方面，D-NeRF[8]、Nerfies[7]、HyperNeRF[18]和 NR-NeRF[9]使用静态规范辐射场捕捉场景的几何和外观，然后在每个时刻学习规范空间的变形/位移场。具体来说，要渲染不同时刻的图像，就需要使用变形场来估算后向场景流，将三维点从当前时刻移动回规范辐射场。然而，对于时间轴上的同一 3D 位置，后向流场并不能保证平滑和连续。因此，规范辐射场通常会出现扭曲，类似于移动物体的平均形状。本文将重点解决后向流的不平滑问题。

除了这两个主要方向之外，目前还有一种趋势是加速基于体素网格表示的动态 NeRF 的训练。TiNeuVox[19]

使用轻量化的 MLP 对变形进行建模，并使用多分辨率为辐射网络获取特征，从而估算不透明度和颜色。V4D[20]使用三维特征体素对四维辐射场进行建模，并将额外的时间维度串联起来，同时提出了用于像素级的查找表。然而 V4D 主要侧重于提高图像质量，与 TiNeuVox 相比，其训练速度并不明显。DeVRF[21]也是以体素网格表示法为基础，它提出使用多视角数据来克服单目设置带来的奇异性问题。与其他使用单目设置的方法相比，多视角数据简化了运动和几何的学习。

三、动态场景新视角合成方法

我们使用一个基于体素网格的静态辐射场和一个基于体素网格的轨迹场来对规范空间中的场景进行建模。为了合成动态图像，我们提出将规范空间中的辐射场向前翘曲到相应的时刻，并根据翘曲得到的辐射场使用体渲染技术渲染图像。图 2 显示了提出方法的框架。该方法有三个主要组成部分：基于体素网格的规范空间模型、可微的前向翘曲方法和体渲染方法。

3.1 基于体素网格的规范空间模型

基于体素网格的规范空间中包含两个模型，包括了一个基于体素网格的静态辐射场和一个基于体素网格的轨迹场。静态辐射场包含了一个可学习的辐射特征V_{Rf}^{Can}和一个轻量化 MLP 网络F_{θ_1}，静态辐射场定义如下：

$$V_R^{Can} = F_{\theta_1}(V_{Rf}^{Can}, V_p^{Can}) \tag{1}$$

其中V_p^{Can}是体素网格在世界坐标中的坐标。

我们提出使用离散余弦变换（DCT）[10]来表示三维点的运动轨迹，以确保运动的平滑性。与静态辐射场类似，我们也使用了可学习的变形特征V_{Df}^{Can}和轻量化 MLP 网络F_{θ_2}，来估计轨迹场，其定义为：

$$V_T^{Can} = F_{\theta_2}(V_{Df}^{Can}, V_p^{Can}) \tag{2}$$

其中V_T^{Can}包含了每个体素的 DCT 轨迹系数。在给定时刻 t 的情况下，我们可以通过以下公式得到这些体素从规范空间到时刻 t 的前向流：

$$V_{flow}^t = f_{DCT^{-1}}(V_T^{Can}, t) - f_{DCT^{-1}}(V_T^{Can}, Can) \tag{3}$$

其中$f_{DCT^{-1}}$是 DCT 变换的逆变换，详细公式可以参考 DCT-NeRF[10]。

3.2 可微的前向翘曲方法

为了根据计算得到的前向流，将规范空间的静态辐射场从规范空间的时刻前向翘曲到对应的时刻，我们提出了一个可微的前向翘曲方法。该方法包含了两步：平均拼接和填充网络。对于平均拼接，受 Softmax-Splatting[22]的启发，我们提出将源网格中可能存在的多个值使用平均拼接融合到对应的目标网格。具体来说，我们提出一种简单而有效的方法：用三线性核计算这些值的"平均值"。形式上，假设我们需要通过流$f_{S\to T}$将源网格V^S翘曲到目标网格V^T，而 $\boldsymbol{p}$、$\boldsymbol{q}$ 是体素网格的索引。我们将$V^T = F_{warp}(V^S, f_{S\to T})$定义如下：

$$V^T[\boldsymbol{p}] = \frac{\sum_{\forall \boldsymbol{q}\in V^S} b[\boldsymbol{u}]\cdot V^S[\boldsymbol{q}]}{\sum_{\forall \boldsymbol{q}\in V^S} b[\boldsymbol{u}]} \tag{4}$$

$$b[\boldsymbol{u}] = \prod \max(0, 1 - |\boldsymbol{u}_i|), i \in \{x, y, z\} \tag{5}$$

$$\boldsymbol{u} = (\boldsymbol{q} + f_{S\to T}[\boldsymbol{q}]) - \boldsymbol{p} \tag{6}$$

其中x, y, z代表了体素网格的三个坐标轴。

通过以上的翘曲公式，我们可以将静态辐射场V_R^{Can}翘曲时刻 t：

$$V_R^t = F_{warp}(V_R^{Can}, V_{flow}^t) \tag{7}$$

由于一对多的问题存在，通过平均拼接后的辐射场存在空洞。为了解决这个问题，我们提出一个填充网络F_{θ_3}来填充V_R^t可能存在的空洞：

$$V_{R_{Inp}}^t = F_{\theta_3}(V_R^t) \tag{8}$$

填充网络是一个基于三维卷积的 UNet 网络结构，它可以通过学习，利用领域信息，填补存在的空洞。

3.3 体渲染方法

在得到时间 t 的辐射场$V_{R_{Inp}}^t$后，就可以使用体渲染技术[23]渲染图像射线的像素颜色。给定一条射线 $r(w) = \boldsymbol{o} + w\boldsymbol{d}$ 从像机中心 $\boldsymbol{o}$ 出发，以 $\boldsymbol{d}$ 为视角方向穿过图像平面上给定的像素，我们通过体渲染方法渲染出对应像素的颜色$C_{Inp}(r) = F_{render}(V_{R_{Inp}}^t, r)$。为此，我们获取射线$r$和体素网格相交的所有三维点$p$。然后，应用三线插值法获得每个三维点的密度σ和颜色$\mathbf{c}$，

$$(\sigma, \mathbf{c}) = F_{inter}(V_{R_{Inp}}^t, p) \tag{9}$$

最后，像素的颜色可以通过如下公式渲染得到：

$$C(r)=\sum_{k=1}^{K}T(w_k)\alpha(\sigma(w_k)\delta_k)\boldsymbol{c}(w_k) \quad (10)$$

$$T(w_k)=\exp(-\sum_{j=1}^{k-1}\sigma(w_j)\delta_j) \quad (11)$$

$$\alpha(\sigma(w_k)\delta_k)=1-\exp(-\sigma(w_k)\delta_k) \quad (12)$$

其中δ_k是射线上，相邻的两个采样点之间的距离。

三、动态场景新视角合成实验结果

在 D-NeRF 数据集上进行了相关测试，量化结果如表 1 。可以看到提出的方法在性能上与其他方法比较，有明显并且一致性的优势。

表 1　在 D-NeRF 数据集上的测试结果

方法	PSNR↑	SSIM↓	LPIPS↓
T-NeRF[8]	29.51	0.95	0.08
TiNeuVox-S[19]	30.75	0.96	0.07
TiNeuVox-B[19]	32.67	0.97	0.04
D-NeRF[8]	30.50	0.95	0.07
NDVG[12]	30.54	0.96	0.05
Ours	32.68	0.97	0.04

图 3 提供了一些可视化展示。可以渲染出准确而具有细节的图像，例如顶部场景中的头盔和手臂，也可以生成更清晰的边界，例如底部场景中的手和脚。

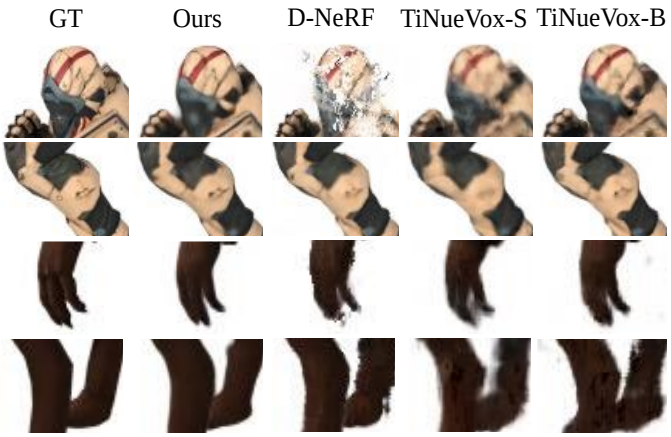


图 3　可视化结果

图 4 比较了提出方法与 D-NeRF[8]重建的规范空间场景模型。提出的方法可以恢复规范空间中的正确几何结构。例如，D-NeRF[8]所生成的球和机械臂位于整个轨迹的"平均"位置，而提出的方法则位于正确的位置。这表明提出的运动模型可以估计出更加精度的运动，从而减少了规范空间中场景模型的误差。

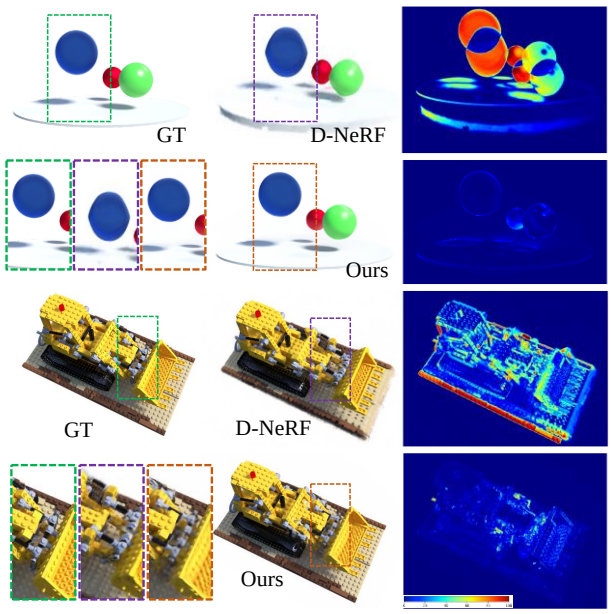


图 4　规范空间场景模型重建比较

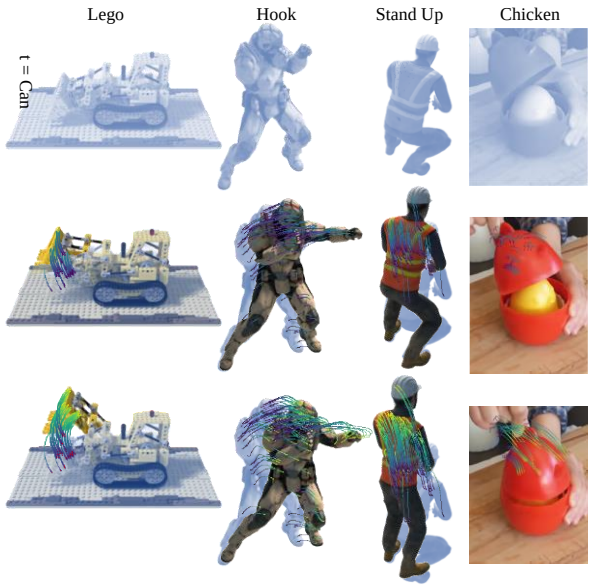


图 5　轨迹可视化

图 5 展示了由轨迹场学习到的轨迹。可以看到提出的方法可以学习到合理的运动轨迹。利用这一特点，在未来的工作中可以考虑引入几何约束、运动模型和先验知识等，来帮助模型提高轨迹的估计质量。

四、总结与展望

本文介绍了一种基于规范空间表示法的前向翘曲方法，用于动态场景的新视角合成。提出的方法在规范空间中对静态场景进行建模，并将整个场向前翘曲到其他时刻，以进行动态场景的渲染。为了解决多对一和一对多映射的问题，提出了一种由平均拼接和填充网络组

成的可微前向翘曲方案。提出的前向翘曲方法在公开数据集上实现了最优的性能。

方法的局限性和未来发展方向：我们目前实现的方法消耗显存相对较大，尤其是在真实场景中。此外，训练速度也相对较慢 (每个场景一天)。由于我们的方法采用了前向翘曲技术，该方法可以获得平滑的轨迹场。因此在未来的工作中，我们还可以引入额外的约束条件和运动模型，来帮助模型学习更加精确的轨迹。

责任编委 储珺

参考文献

[1] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In Proceedings of the European Conference on Computer Vision (ECCV), 2020.

[2] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[3] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2020.

[4] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021.

[5] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021.

[6] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[7] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. Proceedings of the IEEE International Conference on Com- puter Vision (ICCV), 2021.

[8] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[9] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollho ̈fer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021.

[10] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. Neural trajectory fields for dynamic novel view synthesis. arXiv preprint arXiv:2105.05994, 2021.

[11] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[12] Xiang Guo, Guanying Chen, Yuchao Dai, Xiaoqing Ye, Jiadai Sun, Xiao Tan, and Errui Ding. Neural deformable voxel grid for fast optimization of dynamic view synthesis. In Proceedings of the Asian Conference on Computer Vision (ACCV), 2022.

[13] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021.

[14] Thomas Mu¨ller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (TOG), 2022.

[15] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct Voxel Grid Optimization: Super-fast convergence for radiance fields reconstruction. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[16] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021.

[17] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[18] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. HyperNeRF: A higher- dimensional representation for topologically varying neural radiance fields. ACM Transactions on Graphics (TOG), 2021.

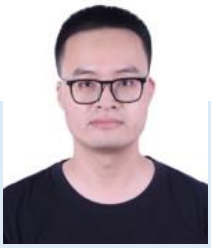
[19] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. ACM SIGGRAPH Asia, 2022.

[20] Wanshui Gan, Hongbin Xu, Yi Huang, Shifeng Chen, and Naoto Yokoya. V4d: Voxel for 4d novel view synthesis. arXiv preprint arXiv:2205.14332, 2022.

[21] Jia-Wei Liu, Yan-Pei Cao, Weijia Mao, Wenqiao Zhang, David Junhao Zhang, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Devrf: Fast deformable voxel radiance fields for dynamic scenes. arXiv preprint arXiv:2205.15723, 2022.

[22] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn- thesis. In Proceedings of the European Conference on Computer Vision (ECCV), 2020.



郭相

西北工业大学电子信息学院博士生，研究方向：新视角合成，视觉定位。
Email: guoxiang@mail.nwpu.edu.cn

戴玉超

西北工业大学电子信息学院教授，研究方向：机器视觉与人工智能
Email: daiyuchao@nwpu.edu.cn

热点追踪

RGBD1K：一个用于 RGB-D 目标跟踪的大规模数据集和基准

江南大学 朱学峰 徐天阳 吴小俊

一、摘要

RGB-D 目标跟踪近年来引起了广泛的关注，这主要得益于视觉和深度通道之间的信息合作，使其在性能上取得了令人瞩目的成果。然而，由于有限数量的带标注的 RGB-D 跟踪数据，大多数最先进的 RGB-D 目标跟踪器只是高性能 RGB 跟踪器的简单扩展，在离线训练阶段未充分挖掘深度通道的潜力。为缓解训练数据集不足的问题，我们发布了一个名为 RGBD1K 的新 RGB-D 数据集。RGBD1K 包含 1,050 个序列，总共约 250 万 RGB-D 图像对。为了展示在更大的 RGB-D 数据集上训练模型的好处，我们开发了一种基于 Transformer 的 RGB-D 跟踪算法，可作为未来使用新数据集 RGBD1K 进行视觉目标跟踪研究的基线方法。通过进行的大量实验表明，RGBD1K 数据集作为训练集可以显著提升 RGB-D 目标跟踪的性能，这为未来有效的跟踪器设计开辟了更多可能性。有关数据集和代码的详细信息在项目主页上提供：https://github.com/xuefeng-zhu5/SPT。

二、引言

视觉目标跟踪（Visual Object Tracking, VOT）旨在根据给定的目标初始状态，在视频的后续每一帧中预测给定目标对象的位置和尺度。目标跟踪在计算机视觉和模式识别领域扮演着重要角色，视觉目标跟踪技术的发展已经持续了几十年。特别是近年来，随着大规模标注数据集 (例如 GOT10K[1]、TrackingNet[2]和 LaSOT[3] 等）的发布，深度学习进一步加速了高性能视觉目标跟踪算法的发展。使用数百万帧标记的图像进行离线训练，跟踪网络能够学习强大的特征表示，相比传统的在线学习方法，取得了显著的性能提升[4]。

最近，随着低成本 RGB-D 传感器的广泛普及，视觉目标跟踪的研究已经从单模态 RGB 数据扩展到了多模态 RGB-D 视频数据。RGB-D 图像由三通道 RGB 图像和单通道距离深度（Depth）图组成。与传统的 RGB 跟踪相比，RGB-D 数据的附加深度图像提供了额外的空间信息，有助于在复杂场景中实现稳定的目标跟踪[5,6]。然而，现有的大多数 RGB-D 跟踪方法是建立在高性能的 RGB 跟踪器基础上，仅在在线跟踪阶段采用深度信息以支持部分遮挡物体的推理和重新检测消失的目标[7,8]。与 RGB 目标跟踪算法相比，多模态 RGB-D 跟踪的发展速度不如人意。其中主要原因是 RGB-D 跟踪的训练数据不足。公开可用的带有标注的 RGB-D 视频无法支撑 RGB-D 跟踪网络的离线训练。具体而言，现有的 RGB 跟踪数据集包含数千个视频序列，其中有数百万帧带标注图像，但现有的 RGB-D 跟踪标注数据数量要少很多，远远不足以推动 RGB-D 目标跟踪算法的快速发展。

为了进一步激发对 RGB-D 跟踪及其应用的研究，我们采集了一个名为 RGBD1K 的新 RGB-D 数据集。RGBD1K 数据集共包含 1050 个序列，其中，有 1,000 个视频用于训练，50 个视频用于测试。考虑到训练视频的标注成本以及长时视频的视频片段也能包含目标代表性的视觉和深度外观变化，足以支持跟踪模型的学习，因此，我们只对每个视频的前 600 帧进行标注。这样，RGBD1K 包含 60 万个标注图像帧可用于端到端基于深度网络的 RGB-D 跟踪方法的监督学习。对于测试集，所有图像帧都被标注，总共包含约 11.8 万帧。此外，我

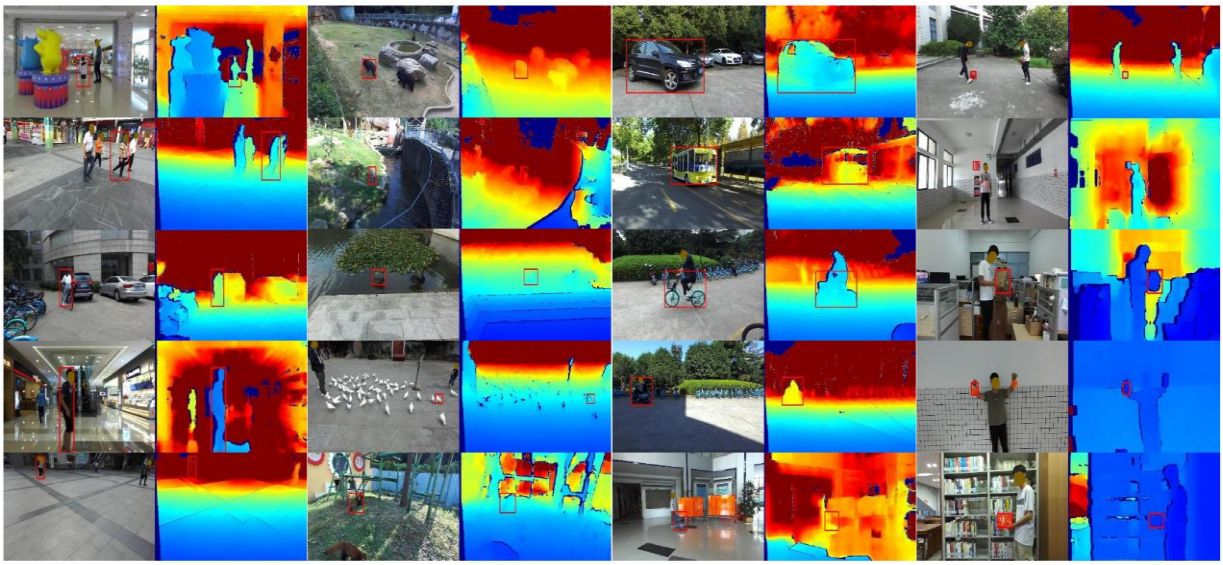


图 1　所提出多模态数据集 RGBD1K 的部分场景示例

们对测试集的每一帧都使用 15 个不同场景挑战属性进行标注。数据的挑战属性标注有助于对跟踪器性能与缺陷进行分析。表 1 对比了现有的 RGB-D 数据集，包括 PTB[9]、STC[10]、CDTB[11]、DepthTrack[12]以及所提出的 RGBD1K 数据集。从表 1 中可以看出，提出的 RGBD1K 具有最多的序列数、RGB-D 图像对数、标注量以及序列的平均长度。另外，为了展示新数据集对 RGB-D 跟踪性能的影响，我们提出了一种基于 Transformer 的跟踪算法，具体而言，我们将基于 RGB 的目标跟踪网络 STARK[13]扩展为 RGB-D 版本，并设计了一个用于融合两种模态特征的模块。该方法使用 RGBD1K 数据集的 1000 个训练序列进行离线训练。在 RGBD1K、DepthTrack 和 CDTB 数据集上进行的大量验证评估和相应的结果表明了 RGBD1K 数据集的有效性以及所提出的 RGB-D 跟踪方法的性能优势。

表 1　RGBD1K 和现有 RGB-D 跟踪数据集的对比

数据集	视频数	帧数	平均帧数	标注数	挑战属性
PTB	100	21,542	215	21,542	5
STC	36	9,009	250	9,009	12
CDTB	80	101,956	1,274	101,956	13
DepthTrack	200	294,591	1,473	294,591	15
RGBD1K	1,050	2,503,400	2,384	717,900	15

三、RGBD1K 数据集

3.1 视频序列

RGBD1K 包含 1,000 个训练序列和 50 个测试序列。总体而言，训练集共包含 2,385,500 个 RGB-D 图像对，测试集包含 117,900 图像对。RGBD1K 数据集的所有 1,050 个序列都是使用立体摄像头 ZED 在室内或室外采集的。ZED 相机提供了时间同步和像素对齐的 RGB 和 Depth 图像对。每个视频的帧率为每秒 25 帧。其中，RGB 图像以 24 位（每通道 8 位）的 JPEG 格式存储，深度图以 16 位 PNG 格式存储。

RGBD1K 涵盖了大量的对象类别，包括涉及人类、动物、交通工具和日常用品的 100 多种不同类型的目标物体，图 1 展示了不同目标类别的一些示例。我们选择了数十种不同的场景来拍摄这些序列，例如办公楼、购物中心、动物园、体育场等。此外，一些视频是从第一人称视角和俯视视角捕获的，用以模拟移动机器人、无人机和监控摄像机的视角。

3.2 数据标注

对于每个视频，我们使用目标边界框对图像进行标注。众所周知，数据标注对于科学研究至关重要但非常耗时。考虑到视频序列的片段可以包含足够的目标视觉和深度外观变化，同时为了减少标注的时间成本，对于训练集，我们只标注了每个视频中一个片段的图像帧。具体来说，对于训练集的每个序列，我们只标注了前 600 帧图像。尽管平均每个视频只标注了其长度的 1/4，但我们认为标注的片段中的外观变化足以让模型学习到目标和场景的时空变化[14]。此外，未标记的部分与标记的部分紧密相关，这样部分标记的数据可以直接用于监

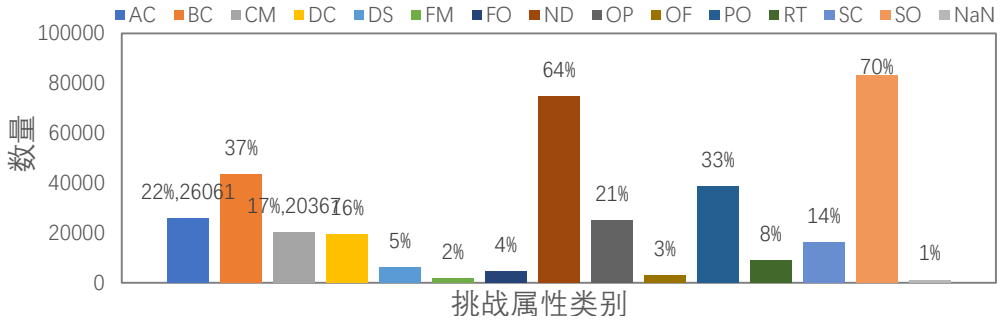


图 2　测试集各种挑战属性的数据分布

督学习同时也存着用于半监督学习的可能。对于测试集，每个序列的所有图像都进行了标注。

为了进一步分析跟踪算法的性能，我们使用了CDTB[11]和 DepthTrack[12]提出的 15 种场景挑战属性为测试集中的每一帧进行了标注。这些挑战属性包括纵横比变化（Aspect-ratio Change，AC），背景杂乱（Background Clutter，BC），摄像机运动（Camera Motion，CM），深度变化（Depth Change，DC），暗场景（Dark Scene，DS），快速运动（Fast Motion，FM），完全遮挡（Full Occlusion，FO），非刚性形变（Non-rigid Deformation，ND），平面外旋转（Out-of-plane Rotation，OP），超出视野（Out of Frame，OF），部分遮挡（Partial Occlusion，PO），反光目标（Reflective Target，RT），尺度变化（Size Change，SC），相似对象（Similar Objects，SO）和未分配（Unassigned, NaN）。其中 AC、DC、FM、SC 和 NaN 这些属性是根据 RGB-D 图像和目标边界框标注计算得到，其余 10 种挑战属性是人工手动标注的。这些场景的挑战属性有助于分析跟踪器在特定挑战下的优劣势。

此外，RGBD1K 测试集中每种挑战属性类别的帧数分布如图 2 所示。从图中可以看出，RGBD1K 数据集只有 1%的图像被标记为没有任何挑战属性，这表明 RGBD1K 测试集对于目标跟踪来说极具挑战性。在所有序列中，大约 64%的图像中的目标属于非刚性可形变物体。通常，可形变物体意味着极端外观变化出现的概率较高，这对于稳定跟踪来说更加困难。此外，70%的图像被标记为相似对象的挑战属性。背景中相似对象的干扰是实现鲁棒单目标跟踪的一个值得研究的重要问题。此外，背景嘈杂和局部遮挡也是 RGBD1K 测试集中重要的挑战因素。尽管某些属性包含的帧数较少，例如 FO 和 OF 只占 4%和 3%，但它们仍然对实际应用非常有价值。具有 FO 或 OF 属性的图像意味着目标在当前图像中是不可见的。尽管总体上只有 7%的图像中的目标是不可见的，但这意味着测试集中的每个视频平均有大约 165 帧的目标消失。目标频繁的长时间消失和重新出现使跟踪问题变得复杂，需要 RGB-D 跟踪器具备较强的感知能力。

3.3 性能评价指标

虽然对于 RGBD1K 的使用没有明确的限制，但在测试集上评估跟踪算法性能时，我们提倡使用长时跟踪评估协议[15]。这个长时跟踪评估协议被主要用于 VOT 竞赛的长时跟踪赛道和多模态 RGB-D 跟踪赛道。在 RGBD1K 数据集中有一定比例的图像中目标是不可见的，即目标可能在一个视频中多次消失和重新出现。因此，在 RGBD1K 上评估跟踪算法时，算法定位目标以及预测目标消失并再次捕获重新出现的目标的能力对于稳健的跟踪系统至关重要。因此，长时 VOT 评估协议非常适用于在我们的数据集上评估跟踪算法。其中主要的评价指标有跟踪精度（Precision，Pr）和召回率（Recall, Re）。具体而言，精度定义为在检测到目标的图像帧上，计算预测目标框和实际目标框的平均重叠比率。召回率表示在目标可见的图像帧上，计算预测目标边界框与标签边界框的平均重叠比率。最后最主要的性能指标是通过计算结合了跟踪精度和召回率的跟踪 F-分数（F-score）。此外，可以使用 VOT 竞赛的评测工具包[]非常方便地在 RGBD1K 上评估跟踪器。具体的三个评价指标计算公式如下：

$$\mathrm{Pr}(\tau_\theta) = \frac{1}{N_p} \sum_{t\in\{t:A_t(\tau_\theta)\neq\emptyset\}} \Omega\left(\mathrm{A}_{\mathrm{t}(\tau_\theta)}, \mathrm{G}_\mathrm{t}\right),$$

$$\mathrm{Re}(\tau_\theta) = \frac{1}{N_g} \sum_{t\in\{t:G_t\neq\emptyset\}} \Omega\left(\mathrm{A}_{\mathrm{t}(\tau_\theta)}, \mathrm{G}_\mathrm{t}\right),$$

$$F(\tau_\theta) = \frac{2\mathrm{Pr}(\tau_\theta)\mathrm{Re}(\tau_\theta)}{\mathrm{Pr}(\tau_\theta) + \mathrm{Re}(\tau_\theta)}.$$

其中，$\mathrm{G_t}$表示实际边界框，$A_{t(\tau_\theta)}$表示在第 t 帧的预测边界框。$\Omega\left(A_{t(\tau_\theta)}, G_t\right)$表示实际边界框和跟踪预测之间的交并比 (Intersection-over-Union, IoU) 。τ_θ是一个置信度阈值。评估协议要求跟踪器一并报告预测边界框和置信度分数。如果在第 t 帧的预测置信度分数θ_t低于τ_θ，则$A_t(\tau_\theta) = \emptyset$。$N_p$是跟踪算法给出预测的帧数，即$A_t(\tau_\theta) \neq \emptyset$的帧数，$\mathrm{N_g}$是目标在视野里可见的帧数，即$G_t \neq \emptyset$的帧数。

四、基线 RGB-D 跟踪器

为了展示 RGBD1K 数据集的重要性，并激发新的 RGB-D 跟踪算法设计，我们提出了一种名为 SPT 的新的 RGB-D 跟踪基线方法。SPT 是从最近的基于 Transformer 的跟踪器 STARK[13]发展而来的。STARK 是一种高性能的基于 RGB 数据的目标跟踪器。SPT 是通过将 STARK-S（STARK 没有使用时序结构的版本）扩展为具有专用特征融合模块的 RGB-D 版本而构造的。SPT 的架构在图 3 中给出。首先，将两个模态的搜索区域和初始模板输入到骨干网络中以分别提取深度 CNN 特征。这里使用的骨干网络是 ResNet-50[16]网络。每个模态的搜索区域和模板的特征都是$\mathrm{H}\times\mathrm{W}\times\mathrm{C}$和$h\times\mathrm{w}\times\mathrm{C}$大小的张量。然后，我们将每个模态的特征进行展平并级联，然后通过一个 6 层编码器层堆叠成的 Transformer 编码器来融合每个模态的模板-搜索区域特征信息。最后，两个模态特定编码器的输出通过设计的特征融合模块进行融合。

关于所提出的特征融合模块，首先，深度模态的 Transformer 编码器的输出和 RGB 模态的编码器的输出在通道维度上进行级联。然后采用一维卷积层来减少级联特征的通道数，从2C通道减少到C通道。最后，我们引入一个 2 层编码层组成的 Transformer 编码器，以进一步融合和增强两个模态的特征。每个编码器层包含多头自注意模块和前馈网络[13]。

网络框架的其余部分包括目标 query、Transformer 解码器和目标边界框预测头都保持和 STARK 算法一致。其中 Transformer 解码器由 6 层解码层堆叠而成，通过将可学习的目标 query 和融合特征作为输入生成输出，以学习目标通用的鉴别信息。每个

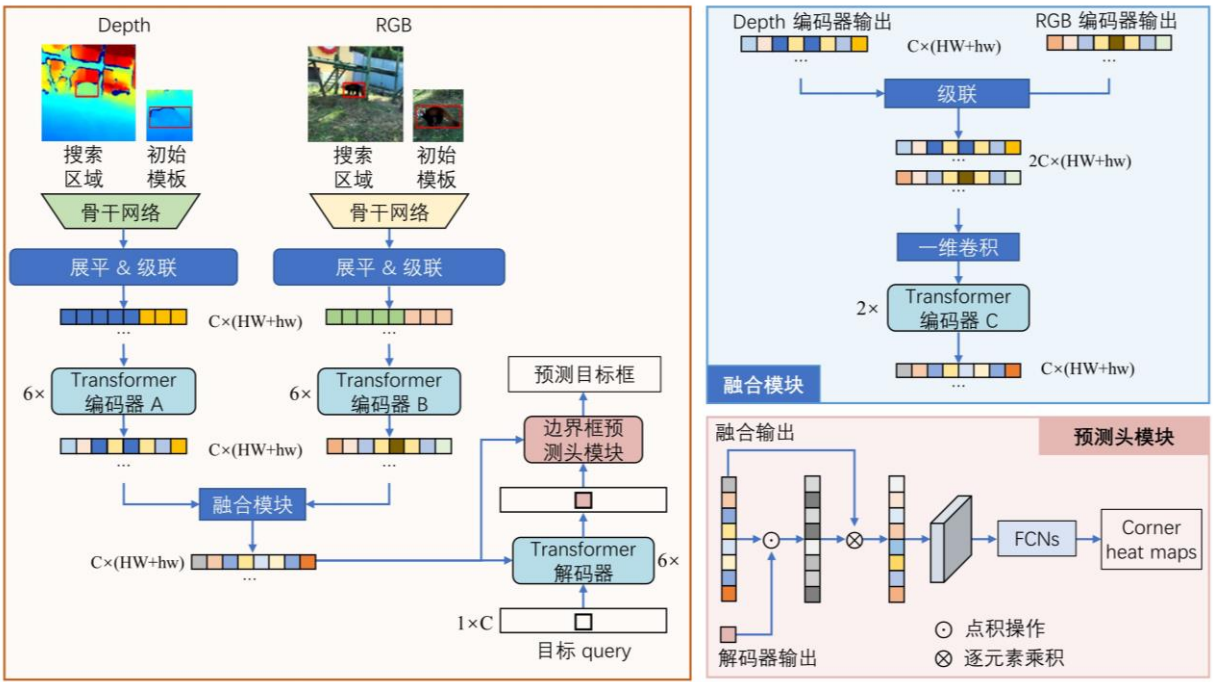


图 3　所提出的 RGB-D 目标跟踪基线算法的网络框架图

表 2　在 RGBD1K、DepthTrack 和 CDTB 上的消融对比实验结果

数据集	RGBD1K			DepthTrack			CDTB		
评价指标	Pr	Re	F-score	Pr	Re	F-score	Pr	Re	F-score
STARK-S	0.480	0.510	0.495	0.490	0.511	0.500	0.630	0.701	0.664
STARK-S-FT	0.509	0.537	0.522	0.497	0.517	0.507	0.638	0.706	0.670
SPT	0.545	0.578	0.561	0.527	0.549	0.538	0.654	0.726	0.688

解码器层由自注意模块、编码器-解码器注意模块和前馈网络组成。然后，Transformer 解码器的输出和融合特征一起输入到边界框预测头，以预测目标框坐标。

在边界框预测模块中，首先计算解码器输出与融合特征的相似度，并使用相似度来增强融合特征。然后，增强后的特征张量被调整形状并通过全卷积网络传递，生成左上角热图和右下角热图。通过左上角和右下角的点，目标边界框就可以被确定。最后，和 STARK 算法一样，SPT 网络训练损失函数是L_1损失和IoU损失组合[13]。

五、实验与分析

我们在 RGBD1K、DepthTrack 和 CDTB 数据集上进行了实验。在本节中，我们描述了所提出算法的实现细节，包括参数设置和实验平台。然后，本节呈现了消融研究的结果，以展示我们数据集以及提出的特征融合模块的有效性。最后，我们提供了比较实验的结果和相应的分析。

5.1 实验环境与设置

我们提出的 SPT 跟踪器是在有一块 Intel i9-CPU 和一块 NVIDIA GeForce RTX 3090 GPU 的计算机平台上进行训练和评估的。训练和测试参数设置与 STARK 算法相同，除了学习率和训练周期的设置。SPT 训练的学习率设置为10^{-5}，训练周期数为 250。对于 SPT 的骨干网络、Transformer 编码器 A、B、Transformer 解码器和边界框预测头网络，我们在训练时使用官方发布的 STARK-S 模型相应组件的权重进行初始化。SPT 的训练数据集为 RGBD1K 的训练集。

5.2 消融研究

为了展示提出的 RGBD1K 数据集对提升 RGB-D 跟踪性能的有效性，首先我们构建了三个跟踪器，包括 STARK-S[13]、STARK-S-FT 和我们的 SPT。STARK-S 是 ResNet-50 作为骨干网络（没有时间分支）的 STARK 跟踪器，是 SPT 的基线方法。对此，我们使用官方发布的 STARK-S 训练模型。STARK-S-FT 是在 RGBD1K 上仅使用训练集的所有 RGB 图像对 STARK-S 进行微调训练后的跟踪器。SPT 使用 RGBD1K 的 RGB-D 图像进行训练。

在 RGBD1K 测试集上的结果如表 2 所示。在使用 RGBD1K 训练集的 RGB 图像进行微调后，STARK-S-FT 在精度、召回率和 F-score 方面的性能从 0.480、0.510 和 0.495 提高到 0.509、0.537 和 0.522。同样使用 RGBD1K 训练集的 RGB-D 图像进行训练后，SPT 在精度、召回率和 F-score 方面进一步提高到 0.545、0.578 和 0.561。这个提高可以使我们得出结论，RGBD1K 的带标注的 RGB 图像和深度图像都有助于提高 RGB-D 跟踪性能。为了进一步确认 RGBD1K 的有效性，我们在另外两个数据集 DepthTrack 和 CDTB 上进行相同的实验，以探索 STARK-S、STARK-S-FT 和 SPT 之间的性能。值得注意的是，这些跟踪器仅使用 RGBD1K 进行训练，没有使用 DepthTrack 或 CDTB 的序列来微调跟踪网络或相应的超参数。在 DepthTrack 和 CDTB 上的结果也在表 2 中展示。

从表中可见，在 RGBD1K 数据集上训练后，STARK-S-FT 和 SPT 在 DepthTrack 和 CDTB 数据集上取得显著的性能提升。在使用 RGBD1K 的 RGB-D 数据训练，SPT 将 STARK-S 在两个数据集上的精度、召回率和 F-score 分别从 0.490、0.511 和 0.500 提高到 0.527、0.549 和 0.538，从 0.630、0.701 和 0.664 提高到 0.654、0.726 和 0.688。在 F-score 度量方面，SPT 分别提高了 STARK-S 在 DepthTrack 和 CDTB 上的性能 7.6%和 3.6%。显然，表 2 结果证明了提出的 RGBD1K 数据集在训练端到端的 RGB-D 跟踪器算法方面有着广泛优势。

表 3　在 RGBD1K、DepthTrack 和 CDTB 上的对比实验结果

跟踪算法	RGBD1K			DepthTrack			CDTB		
	Pr	Re	F-score	Pr	Re	F-score	Pr	Re	F-score
ATCAIS	0.511	0.451	0.479	0.500	0.455	0.476	0.709	0.696	0.702
DDiMP	0.557	0.534	0.545	0.503	0.469	0.485	0.703	0.689	0.696
TALGD	0.511	0.451	0.479	0.494	0.424	0.456	**0.728**	0.717	**0.722**
DAL	**0.562**	0.407	0.472	0.512	0.369	0.429	0.647	0.571	0.607
DeT	0.438	0.419	0.428	**0.560**	0.506	0.532	0.674	0.642	0.657
SPT	0.545	**0.578**	**0.561**	0.527	**0.549**	**0.538**	0.654	**0.726**	0.688

5.3 对比研究

我们将提出的 SPT 与最近的一些 RGB-D 跟踪算法进行了比较，这些跟踪器包括 VOT-RGBD 竞赛[18,19,4]上的算法 ATCAIS、DDiMP、TALGD 和 Siam_LTD，以及 DAL[17]和 DeT[12]。表 3 中呈现了这些算法在 RGBD1K, DepthTrack 和 CDTB 三个数据集上的跟踪结果。一般来说，F-score 是 VOT 协议中最重要的性能度量，跟踪器按照 F-score 值进行排名。从表中可以看出，在 RGBD1K 测试集上，SPT 实现了最佳的 F-score，而 RGB-D 跟踪器 DDiMP 是第二优秀的跟踪器。与 DDiMP 跟踪器相比，我们的 SPT 跟踪器在 F-score 方面取得了 2.9%的提高。此外，与 ATCAIS、TALGD、DAL 和 DeT 等 RGB-D 跟踪器相比，提出的 SPT 跟踪器也有显著优势。SPT 的跟踪性能提升表明使用提出的 RGBD1K 进行训练模型有助于实现更稳健的 RGB-D 目标跟踪。

为了更好地反映我们的 RGBD1K 对于跟踪性能提升的普遍优势，我们同样将 SPT 跟踪器与最先进的跟踪器在 DepthTrack 和 CDTB 数据集上进行了比较。值得注意的是，我们的 SPT 跟踪器是使用 RGBD1K 训练然后直接在 DepthTrack 和 CDTB 数据集上测试的，没有微调任何参数。在表 3 中可以看出，SPT 在 DepthTrack 数据集上实现了最佳的 F-score（0.538）和召回率 Recall（0.549）。在 CDTB 数据集上，SPT 在 Recall 方面显著优于其他最先进的跟踪器。尽管在 Precision 和 F-score 方面，SPT 相对于 DDiMP、ATCAIS 和 TALGD 较差，但这主要是因为这些算法使用了多个跟踪器的组合，且在 CDTB 数据集上过拟合。因此，这些算法的效率较低，而且在另外两个数据集 RGBD1K 和 DepthTrack 上性能大打折扣。以上结果证实，使用大量标注的 RGB-D 数据离线训练的 SPT 跟踪器可以实现较为鲁棒精准的 RGB-D 目标跟踪。另一方面，这些结果也证实了所提出的 RGBD1K 数据集对于 RGB-D 目标跟踪发展的重要性。

六、总结

在这项工作中，我们提出了一个用于 RGB-D 目标跟踪的大规模数据集，以及一个基于端到端深度网络的基线跟踪器。这项工作的动机是现有带标注 RGB-D 视频的稀缺阻碍了 RGB-D 跟踪的发展。所提出的 RGBD1K 数据集大大提升了现有 RGB-D 目标跟踪的标注数据量。为了展示 RGBD1K 数据集的实用性，我们设计了一种新的 RGB-D 跟踪基线方法，并使用 RGBD1K 训练集的所有 RGB-D 数据进行离线训练。使用 RGBD1K、DepthTrack 和 CDTB 数据集测试得到的广泛实验结果展示了在 RGBD1K 上进行训练算法的好处以及其推动未来 RGB-D 跟踪发展的潜力。

责任编委　崔海楠

参考文献

[1] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild [J]. IEEE TPAMI, 2019, 43(5): 1562-1577.

[2] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, Bernard Ghanem; "Trackingnet: A large-scale dataset and benchmark for object tracking in the wild." ECCV. 2018, pp. 300-317.

[3] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, Haibin Ling; "Lasot: A high-quality benchmark for large-scale single object tracking." CVPR, 2019, pp. 5374-5383.

[4] Matej Kristan, et al. "The ninth visual object tracking vot2021 challenge results." ICCV. 2021.

[5] Meshgi, Kourosh, et al. "An occlusion-aware particle filter tracker to handle complex and persistent occlusions." CVIU, 150 (2016): 81-94.

[6] Timur Bagautdinov, Francois Fleuret, and Pascal Fua. "Probability occupancy maps for occluded depth images." CVPR. 2015, pp. 2829-2837.

[7] Sion Hannuna, et al. "DS-KCF: a real-time tracker for RGB-D data." Journal of Real-Time Image Processing 16 (2019): 1439-1458.

[8] Ugur Kart, Joni-Kristian Kamarainen, and Jiri Matas. "How to make an rgbd tracker?." ECCVW. 2018.

[9] Shuran Song, and Jianxiong Xiao. "Tracking revisited using RGBD camera: Unified benchmark and baselines." ICCV. 2013, pp. 233-240.

[10] Jingjing Xiao, Rustam Stolkin, Yuqing Gao, Aleš Leonardis. "Robust fusion of color and depth data for RGB-D target tracking using adaptive range-invariant depth models and spatio-temporal consistency constraints." IEEE TCYB 48.8 (2017): 2485-2499.

[11] Alan Lukezic, Ugur Kart, Jani Kapyla, Ahmed Durmush, Joni-Kristian Kamarainen, Jiri Matas, Matej Kristan. "Cdtb: A color and depth visual object tracking dataset and benchmark." ICCV. 2019, pp. 10013-10022.

[12] Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Aleš Leonardis, Joni-Kristian Kämäräinen. "Depthtrack: Unveiling the power of rgbd tracking."ICCV. 2021, pp. 10725-10733.

[13] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, Huchuan Lu. "Learning spatio-temporal transformer for visual tracking." ICCV, 2021, pp. 10448-10457.

[14] Jack Valmadre, Luca Bertinetto, Joao F. Henriques, Ran Tao, Andrea Vedaldi, Arnold W.M. Smeulders, Philip H.S. Torr, Efstratios Gavves. "Long-term tracking in the wild: A benchmark." ECCV. 2018, pp. 670-685.

[15] Alan Lukežič, Luka Čehovin Zajc, Tomáš Vojíř, Jiří Matas, Matej Kristan. "Now you see me: evaluating performance in long-term visual tracking." arXiv preprint arXiv:1804.07056 (2018).

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. "Deep residual learning for image recognition." Proceedings of the CVPR. 2016, pp. 770-778.

[17] Yanlin Qian, Song Yan, Alan Lukežič, Matej Kristan, Joni-Kristian Kämäräinen, Jiří Matas. "DAL: A deep depth-aware long-term tracker." ICPR., 2021, pp. 7825-7832.

[18] Matej Kristan, et al. “The eighth visual object tracking VOT2020 challenge results.” ECCVW 2020: 547-601.

[19] Matej Kristan, et al. "The seventh visual object tracking VOT2019 challenge results." ICCV, 2019.

朱学峰

江南大学博士生，导师为江南大学吴小俊教授，主要研究方向为计算机视觉、目标跟踪。
Email：xuefeng.zhu@stu.jiangnan.edu.cn

徐天阳

博士，江南大学副教授，主要研究方向为模式识别、计算机视觉和人工智能。
Email：tianyang.xu@jiangnan.edu.cn

吴小俊

博士，江南大学二级教授、至善教授、研究生院院长，IAPR Fellow，AAIA Fellow，主要研究方向为模式识别与人工智能。
Email：wu_xiaojun@jiangnan.edu.cn

顶会观察

ICCV 2023

西澳大学 武子杰
湖南大学 王耀南

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

一、大会概况

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

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

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

二、论文录用情况

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

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

三、主题报告

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

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

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

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

四、最佳论文

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

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

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

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

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

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

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

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

五、大会奖项

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

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

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

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

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

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

六、精彩报告选介

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

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

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

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

七、总结展望

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

责任编委 **魏秀参**

参考文献

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

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

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

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

武子杰

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

王耀南

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

河北大学刘帅奇教授访谈

2023 年 12 月 4 日，《CCF-CV 专委简报》在线采访了河北大学博士生导师刘帅奇教授。下面是采访实录。

刘老师，您好！首先，请您分享一下您的个人学习和研究经历。

我于 2009 年从山东科技大学信息科学与工程学院信息与计算科学专业获得学士学位，同年 9 月进入北京交通大学计算机与信息技术学院进行硕博连读，于 2014 年 1 月获得工学博士学位。2014 年进入河北大学电子信息工程学院工作，期间 2016 年 8 月到 2017 年 1 月曾赴加拿大渥太华大学做访问学者，2020 年 10 月赴中国科学院自动化所做博士后。我的研究方向是计算视觉与多维信号处理，主要包括遥感图像处理和医学影像处理。

您在计算机视觉与多维信号处理等领域内已有所建树，能否介绍一下您在这些领域中最突出的几项研究成果？针对这些领域的研究者，您有什么建议？

我从 2010 年进入到计算机视觉与多维信号处理的研究领域，主要针对遥感图像的智能解析进行研究，重点对机场异物检测、图像背景减除、数据特征自动化提取技术进行了深入的研究。

最早的时候是研究 synthetic aperture radar（SAR）图像去噪，这也是我们研究最深入的一个方向，研究的内容包括多尺度几何变换、稀疏表示和深度学习，构造了一系列的新型多尺度几何变换和深度学习模型，相应的研究成果发表于 IEEE TGARS、IEEE JSTARS 等期刊，入选了 ESI 高被引论文 1 篇，而且也将这些技术应用到了实际工程中。我们构造一种基于距离-时间维的移不变混合变换以抑制机场雷达图像的噪声，该算法可以有效地去除机场雷达图像噪声，显著地改善图像的视觉效果，具有很强的实时性，已经应用于中航工业某所研制的机场异物检测雷达中。

我们和科研院所联合攻关，取得了一系列多源图像融合和快速背景减除的关键技术成果，具有重要的理论意义和工程价值，成功解决了光纤自动熔接、图像融合、绝缘子憎水性检测、视频动态监控等关键技术难题，该项目成果在八家单位获得了应用，为相关公司累计获得利润一千一百多万元。

与五一九大队联合攻关，优化了道路塌陷隐患雷达检测工作流程，研制了道路病害快速检测设备，基于道路地下病害雷达图谱特征库开发了三维地质雷达图谱快速识别处理软件，有效提高了工作效率、节约了施工成本；开发了道路塌陷监测预警平台，实现了检测成果的信息化管理，提升了道路塌陷隐患综合风险评估能力，助力公司累计签订合同金额达一千三百多万元，特别是在 2023 年 7 月底涿州发生特大洪涝灾害后，助力五一九大队援涿道路检测小组快速完成了涿州市区 70 余公里市政道路的应急检测工作，发现病害隐患点 125 处，为消除道路病害隐患提供了数据支撑，在保障救灾通道畅通以及人民群众出行安全方面发挥了重要作用。

进入河北大学工作之后，由于河北大学是综合性的大学，有附属医院，因此倡导进行医工交叉，我们也在医学影像处理方面做了一些工作，尤其是在脑电情绪识别领域做了一系列的工作，构建多种基于时空特征的脑电情绪识别算法，有助于解决训练样本少、被试间差异大等问题。相关研究发表在 IEEE JBHI 2022, Knowledge-Based Systems 2023 等期刊。

计算机视觉领域是一个非常大的领域，或许在一开始大家进入这个领域的时候都是觉得这个领域很火才选择这个领域。希望大家进入这个领域以后能够选择其中的一个细分领域一直深耕，这样才会产出一些好的成果。另外，计算机视觉领域的研究往往不仅需要编程技术好，还需要有深厚的数学知识，因此刚入门的学者们一定要打好数学的基础。最后，要想把论文写在大地上就需要给研究的技术找一个实用的场景，这个需要跟企业做好对接，了解企业的真实需求。

作为青年学者，您获得了河北省燕赵英才、河北省高等学校科学技术研究项目青年拔尖人才等荣誉，您认为这些荣誉对您的学术研究有什么样的影响呢？

我能比较幸运地获得一些荣誉，首先得感谢有一个高水平的平台，以及和一群优秀的人共事。这些荣誉是大家对我所做工作的认可和肯定，增加了个人在学术界的声誉和知名度，这有助于获得更多的学术机会。这些荣誉更像是一种激励，可以激励我和团队成员在研究中保持高度的自我要求和创新精神，不断追求卓越。这些荣誉也是一种展示和交流的机会，能够让学者们与同行交流学术观点，促进学术成果的共享与学科交叉的发展，拓宽研究领域的视野。

近年您发表了多篇高水平学术论文，其中有多篇高被引论文，能跟大家分享一下您是如何做到持续产出高水平论文的呢？您在学术影响力方面做了哪些努力呢？

我认为要持续产出高水平论文，首先要做到在相应的研究领域有很深厚的基础，其次，要有一个优秀的团队，最后，要有一个高水平的平台。很幸运，这三个要素我们这个团队都具备，剩下的就是要找准方向，寻找创新点，做实验，最后进行投稿。在这其中有一些好的习惯可以分享给大家。

第一，工程技术领域的研究往往都是需要很深厚的数学基础的，而数学类的课与其他课程不太一样，没有老师教授，仅凭教材很难学通学透。因此，建议在学生时代或者青年学者一定要多学数学类的课程。我在大学和研究生期间学了《数学分析》、《高等代数》、《概率论》、《数理统计》、《平面几何》、《离散数学》、《运筹学》、《偏微分方程数值解》、《常微分方程》、《矩阵论》、《最优化理论》、《随机过程》、《数值代数》、《数值逼近》和《复变函数论》等数学课程，我觉得这些课程为我后续的研究打下了良好的基础，极大地助力了我的研究工作。

第二，选择参考文献进行阅读时一定要选高质量的期刊论文，这样的论文往往水平很高，逻辑性很强，有助于建立完善的领域知识体系。另外，读论文时要做笔记，最好是专门建立一个 EXCEL 表格，用于记录文献阅读笔记，比如文章的框架、研究思路等，这对论文阅读质量有了更高要求。

第三，在做学术研究的同时不要忽略工程实际问题。学术研究尤其是应用研究如果不和工程相结合就是无源之水。大多数的应用基础研究领域的主要目标是解决工程问题，而实际问题的解决往往难度更大，会有更多的限制条件，如果能够克服这些问题，极有可能开辟出一个新的细分领域。

我非常认同左旺孟老师说的“无论是从学术界还是产业界的角度来看，底层视觉和图像生成都属于我们国家与国外差距相对较小的领域。”我所做的研究大部分属于底层视觉领域，要想做出好的成果，需要踏踏实实，耐下心来深耕一个细分领域。另外，建议年轻学者要提高公开发布论文的质量，尽可能保证论文的贡献和完成

质量，积极参与学术交流和合作，注重社会服务和学术服务。

您是 1986 年出生的，非常年轻，现在已经是博导、教授，有丰富的科研经历且取得了很多科研成果，能否跟大家分享一下您的快速成长历程，以及您的成长感悟？

我感觉我的经历非常普通，从一所普通高校本科毕业，到了一所 211 高校读博，在读博的时候因为论文和项目两手抓，完成度比较高，因此毕业还算顺利。然后博士毕业到了一所普通的学校工作，按部就班评了副教授、教授和博导。现在看起来，很多年轻的人都非常优秀，做出了很好的成果。我唯一的优点应该是老实肯干，做事情有恒心，所以在一个工科不是优势学科的综合性大学一直工作到现在。我觉得不管做什么事情，首先是格局要大，要能够规划未来 5-10 年内事情。其次是要坚持不懈地去做一件事，肯吃亏，不管是做研究还是做人都是如此。最后是组建一个优秀的团队，这样才会有一帮志趣相投的人一起做喜欢的事情，避免单打独斗、势单力薄，也更符合有组织的科研要求，能够集中力量去解决一些大的难题。另外，一定要找到自己的研究方向和企业需求的结合点，这样才能实现成果转化，将论文写在祖国的大地上。

可否请您谈一下在第三代人工智能时代，计算机视觉将如何发展？面临哪些挑战？哪些研究方向会特别有价值呢？

清华大学人工智能研究院院长、中国科学院院士张钹教授在「纪念《中国科学》创刊 70 周年专刊」上发表署名文章，首次全面阐述第三代人工智能的理念，此后第三代人工智能的概念被学术界广泛使用。作为 CCF YOCSEF 保定的学术委员，我在 2021 年作为执行主席举办了“谁将引领第三代人工智能领域的科技创新浪潮：大学还是企业？”的观点论坛，对第三代人工智能的发展做了一些探讨。我个人认为第三代人工智能刚刚起步，处于探索阶段，计算视觉的发展还任重而道远。虽然现在各国学者在各类计算视觉任务中获得了巨大成功，但是这些视觉模型的可解释性不足，难以形成可信的数学描述，这是未来研究中面临的挑战之一。另外，计算机视觉领域的研究将会从单一单任务模型向通用视觉模型发展，这就对视觉模型的鲁棒性和可靠性提出了更高的要求。作为教师，我认为人工智能基础的普及教育也非常值得投入精力去研究，具有挑战性。另外一个值得研究的领域就是隐私保护和真伪辨别，这对个人的信息安全非常重要。

您获批了多项科研项目，能跟大家分享一下您成功申请的经验和体会吗？

申请项目是科研人员必须要做的事情。想要申请成功就要注意平时积累，提前做好准备。可以在平时的研究中将自己的工作整理归纳好，后续可方便地结合相应的应用场景进行扩展。另外，申请材料的内容结构和格式规范也是非常重要的，我们会听取不同领域专家的意见，进行多次会议讨论，仔细修改完善申请材料。

您担任了行政职务和多个期刊编委，这必然占据您大量时间，能否跟大家分享一下您是如何协调本职工作和兼职工作的？能分享一下您的经验吗？

在这个世界上，每个人都是多重角色的扮演者，往往是身兼数职，学会平衡和分配时间是一件很重要的事情。我一般会把要做的事情分成轻重缓急的目录，然后对照条目，一条一条去执行，优先处理最重要和最着急的事情，然后再利用闲暇的时间处理轻松的工作和不那么着急的工作。比如我会每周参与每个科研小组的会议讨论，定时与学生进行沟通，讨论科研中的问题和进展，利用零碎的时间去处理零碎的任务。当然一个好的团队也非常重要，有很多琐碎细致的工作都是我在团队的帮助下完成的。最后，人的精力是有限的，希望大家能够将精力集中到最有意义的事情上，忽略一些不那么重要

的事情，实现人生价值的最大化。

您出版了多部专著与教材，请问花费较多时间完成这些专著与教材对科研与教学发展有哪些益处呢？

我的这些专著和教材撰写实际上是完全基于我在教学和科研中的困惑而发起的。在教学的过程中，我发现以前的教材很难图文并茂、以示例的形式将知识展示给学生，这对学生学习相关知识来说非常痛苦。因此我首先跟学生去深入地探讨，到底什么样的教材能够让他们容易接受新知识，在此基础上我们编写出版了几本教材，例如《MATLAB 程序设计基础与应用》，事实上这本教材确实也达到了相应目标，在二十几所高校被选为教材，发行量达到一万七千余册。

在科研的过程中我也存在很多困惑，比如我一开始做遥感图像去噪，由于不是雷达工程专业出身，所以很多雷达相关的知识都不懂，因此费了很大的力气才算入门，并且在一段时间内只能去复现其他人的算法，找不到源代码，导致算法对比时非常头大。于是我一直在想，怎样去帮助像我一样的人快速入门我的研究领域，并且提供自己力所能及的帮助，因此我们出版了两本专著，其中一本专著已经发行了三千余册，也算是对该领域的研究者有所帮助。

当然这些专著和教材的编写过程也是对自己所教授和所掌握知识的再提炼过程，可以帮助自己更好地去掌握所有需要教授和研究的内容，有助于持续提升自己的教学和科研水平。

您曾获河北大学首届学生最喜爱的教师称号，这一荣誉是您甘为人梯的结果，可否分享您在教书和育人方面的心得？

人生的许多寻找，不仅仅在于千山万水，还在于咫尺之内、眉眼之间。因为热爱，教育不只教书，更是育人，我希望用我的人格去感染，用我的才能去培养具有家国情怀的社会主义现代化接班人。实际上，我挺喜欢和学生待在一起。在学院里面做过辅导员、做过系主任，是通信工程省级教学团队负责人，很喜欢和学生打交道。想要学生信服，就需要恪己治学，修身前行，努力在平凡的工作中服务学校，服务学生，做好学生成长的引路人。

作为高校教师，立德树人是最根本的任务。因此从参加工作以来，我一直都在为本科生和研究生上课。在课堂上除讲授教学知识外，我喜欢跟大家分享这些知识背后的故事，这样的好处一方面在于能够激发学生们的兴趣，另一方面可以通过这些故事潜移默化地影响学生价值观和人生观。另外，教书育人，育人就需要我们在上课的过程中去观察每个学生的状态，然后鼓励那些对课程感兴趣的学生积极主动地去学习和表达自己的观点，鼓励不同的声音。最后，要坚持走科研教学相长之路，科研促教，科研促学，把科研成果转化为教学所用，以更好地促进教学。

您领导着非常优秀的团队，请问您是如何管理和运作您的团队的？您是如何管理研究生的？您对他们的要求是什么？

我所在的团队是河北大学计算视觉与多维信号处理创新团队，由 1 名教授、9 名教师和 50 余名硕博研究生组成。我们团队主要从事模式识别、计算机视觉和无线通信等领域的应用研究与技术创新。团队目前正处于转型期，将从原有的多个研究方向转到专注于其中的三四个优势的研究方向。目前团队的管理正在走向精细化，有些老师会专注于科研一些，有些老师会专注于教学一些，通过教学科研搭配，让每一位团队成员都能发挥出自己的优势，贡献自己的力量。

对于研究生的管理，我们在制定研究方向时比较宽松，学生在兴趣点的基础上结合老师的指导进行选题。在管理方面，学科制定了一系列的规范要求，在此基础上，我们团队还要求每周有一次例会，主要是进行文献讲解；每个研究方向需要一周进行一次小组会，主要是对项目难点和论文中的难点进行讨论。我对研究生的要

求是，不管是做研究还是搞技术，必须侧重一方面，做精一方面。

如果吐露研究工作者的心声，您最想说的是什么？

希望大家都能专注于做自己擅长的事情，做一些有价值的事情，而不是一直在申请基金或者项目的路上。希望大家能够更多地去关注双非类院校年轻教师，给这一广大的人群更多的机会。希望能切实减轻科研人员的负担，让年轻的学者们能够活得有尊严，不要让项目、论文、职称等帽子压垮他们的脊梁。

责任编委 赵振兵 余烨

刘帅奇

工学博士，河北大学教授、博士生导师，电子信息工程学院副院长，河北省机器视觉技术创新中心主任，河北省燕赵英才，河北省高等学校科学技术研究项目青年拔尖人才，南通市江海英才，河北大学坤舆青年学者，保定市新时代好青年，获宝钢优秀教师奖，入选 2023 年全球前 2%顶尖科学家榜单。研究方向为计算视觉与多维信号处理，主要涉及图像去噪、图像融合、边缘检测和医学影像处理等研究领域。2014 年毕业于北京交通大学计算机与信息技术学院，同年进入河北大学电子信息工程学院工作，曾获河北大学首届学生最喜爱的教师称号。2016 年 8 月至 2017 年 1 月赴加拿大渥太华大学做访问学者。近年来，先后主持参与纵向项目 15 项，包括主持国家基金 2 项、省基金 3 项。先后主持参与横向项目 14 项，合同金额达到了 260 万元。近年来在包括 IEEE TGARS、IEEE J-STARS、IEEE J-BHI、IEEE TBIOM、ACM TOMM、IEEE TNSRE、IEEE-ACM TCBB 等国内外权威期刊上发表论文 60 余篇，包括 50 余篇 SCI 检索期刊论文，入选 ESI 高被引论文 2 篇，入选中国知网高被引论文 2 篇。获授权中国发明专利 5 项，发明专利受理 2 项。现为 International Journal of Imaging System and Technology 期刊编委、兵器装备工程学报编委。

委员好消息

✪ 2023 年 5 月 8 日，亚洲青年科学家基金项目宣布了该基金项目的首届研究员名单，共评选出 12 位获得者，CCF-CV 专委会执行委员、香港科技大学**陈浩**入选。

✪ 2023 年 9 月 26 日，CCF 公布了 2023 下半年新晋高级会员、杰出会员名单，CCF-CV 专委会 15 位执行委员晋升杰出会员：南开大学**程明明**、北京航空航天大学**黄迪**、中国科学院计算技术研究所**蒋树强**、北京航空航天大学**李甲**、南京邮电大学**刘青山**、北京航空航天大学**陆峰**、西安交通大学**孟德宇**、中国科学院深圳先进技术研究院**乔宇**、北京交通大学**阮秋琦**、中国科学院自动化研究所**申抒含**、中国科学院自动化研究所**兴军亮**、上海海事大学**周日贵**、北京航空航天大学**史振威**、华北电力大学**翟永杰**、北京师范大学**邬霞**，11 位执行委员晋升高级会员：复旦大学**姜育刚**、中国科学院计算技术研究所**山世光**、中国科学院自动化研究所**王伟**、华南理工大学**吴永贤**、北京理工大学**杨健**、华中科技大学**尤新革**、北京理工大学**付莹**、湖南大学**李智勇**、上海大学**曾丹**、天津大学**周圆**、清华大学**刘烨斌**。

✪ 2023 年 9 月 28 日，在 2023 国际生物特征识别大会上，CCF-CV 专委会执行委员、中国科学院自动化研究所多模态人工智能系统全国重点实验室**朱翔昱**副研究员因其在人脸生物特征识别领域对人脸识别、三维人脸重建等方面的突出贡献，获得国际模式识别学会颁发的“国际生物特征识别青年学者奖”。

✪ 2023 年 10 月 7 日，中国图象图形学学会 2023 年度会士增选名单揭晓，CCF-CV 专委会副主任委员、中国科学院自动化研究所**王亮**当选。

✪ 2023 年 10 月 14 日 PRCV2023 最佳论文评审结果发布，CCF-CV 专委会 4 位执行委员执导的 3 篇论文获奖：西安电子科技大学**苗启广**指导的 *Auto-Learning-GCN: An Ingenious Framework for Skelenton-based Action Recognition* 获最佳论文，南京理工大学**张珊珊**指导的 *OKGR: Occluded Keypoint Generation and Refinement for 3D Object Detection* 获最佳学生论文奖，西北工业大学**孙瑾秋**和**张艳宁**指导的 *An Internal-external Constrained Distillation Framework for Continual Semantic Segmentation* 获最佳论文提名。

✪ 2023 年 10 月 29 日，CCF 第十三届会员代表大会选举产生新一届理事会、监事会，CCF-CV 专委会 17 位执行委员当选：中国科学院计算技术研究所**蒋树强**当选监事，北京师范大学**黄华**、北京航空航天大学**王蕴红**、西北工业大学**张艳宁**当选常务理事，江西财经大学**方玉明**、东南大学**耿新**、中国科学院自动化研究所**赫然**、中国科学院大学**黄庆明**、新疆大学**库尔班•吾布力**、西安电子科技大学**苗启广**、中国科学院心理研究所**王甦菁**、航天宏图信息技术股份有限公司**王涛**、大连海事大学**王新年**、哈尔滨工业大学**邬向前**、北京工业大学**毋立芳**、南京理工大学**肖亮**、英特尔中国研究院**张益民**当选理事。

✪ 2023 年 11 月 3 日，2023 年度河北省科学技术奖总评审结果公布，CCF-CV 专委会位执行委员、燕山大学**吴培良**参与完成的“石油产液剖面三相流多参数系列组合测井仪及应用研究”获科技进步二等奖。

✪ 2023 年 11 月 4 日，CCF-CV 专委会举行十周年纪念活动，会上举行了颁奖仪式。CCF-CV 顾问委员会委员、西北工业大学**张艳宁**获杰出成就学者称号。CCF-CV 专委会常务委员、北京大学**林宙辰**等 2013 年发表于 IEEE TPAMI 的完成的论文 Robust Recovery of Subspace Structures by Low-Rank Representation

被授予持久影响力工作。CCF-CV 专委会 6 位执行委员获服务贡献学者称号：北京航空航天大学**于茜**、西安电子科技大学**王楠楠**、武汉理工大学**朱安娜**、厦门大学**孙晓帅**、大连理工大学**樊鑫**、东北大学**贾同**。

✪ 2023 年 11 月 6 日，2022 年度北京市科学技术奖公布，CCF-CV 专委会 5 位执行委员获奖：中科院自动化所**张兆翔**、**谭铁牛**等完成的“面向动态复杂环境的多源融合智能感知技术及产业化”、香港科技大学**陈浩**参与完成的“环骨盆严重创伤智能化微创救治体系和临床应用”获科技进步一等奖，北京交通大学**白慧慧**等完成的“空天地协同的视觉智能关键技术及应用”、北京工业大学**胡永利**参与完成的“超大城市交通视觉大数据高效表达与运行决策关键技术与应用”获科技进步二等奖。

✪ 2023 年 11 月 23 日，2024 IEEE Fellow 名单出炉，CCF-CV 专委会 6 位执行委员入选：西安电子科技大学（重庆邮电大学）**高新波**因对混合增强智能和图像质量评估的贡献、复旦大学**姜育刚**因对大规模视频分析和开源数据集的贡献、中国科学院自动化研究所**雷震**因对人脸分析和目标检测的贡献、中山大学**林倞**因对多媒体内容分析的贡献、清华大学**鲁继文**因对视觉内容分析和识别的贡献、大连理工大学**卢湖川**因对视觉对象跟踪和显着对象检测的贡献、台湾交通大学**郑文皇**因对对智能多媒体计算和应用的贡献入选。

✪ 2023 年 11 月 24 日，2022-2023 年度高校计算机专业优秀教师奖励计划名单公布，CCF-CV 专委会 5 位执行委员入选：北京大学**刘家瑛**、同济大学**张林**、西南交通大学**龚勋**、华南理工大学**许勇**、哈尔滨工程大学**刘海波**。

✪ 2023 年 12 月 1 日，2023 年度中国图象图形学学会博士学位论文激励计划遴选结果公布，8 位执行委员指导的论文入选：复旦大学**姜育刚**指导的《视觉与语言结合的视频理解方法研究》、中国科学院大学**黄庆明**指导的《标签缺失条件下的机器学习算法研究》、浙江大学**李玺**指导的《动态开放环境下自动驾驶结构化场景感知与学习研究》、北京邮电大学**马占宇**指导的《基于概率模型的深度神经网络研究》、西北工业大学**夏勇**指导的《面向有限标注的医学影像分割及分类方法研究》、江西财经大学**方玉明**指导的《立体视觉信号质量评价方法研究》入选 2023 年度中国图象图形学学会博士学位论文激励计划，中科院计算所**山世光**指导的《基于对抗学习的人脸属性编辑研究》和北京航空航天大学**李甲**指导的《基于局部关系学习的物体细粒度解析与识别》获得提名。

✪ 2023 年 12 月 1 日，2023 年度中国图象图形学学会科学技术奖评选结果公布，CCF-CV 专委会 20 位执行委员获奖：西北工业大学**聂飞平**等完成“平动态特征的不变量学习理论与方法”、中国科学院大学**黄庆明**和中国科学院信息工程研究所**任文琦**等完成“异构媒体的协同计算与泛化推理”、北京航空航天大学**刘偲**和浙江大学**杨易**等完成的“ 知识引导的视觉内容理解”获自然科学一等奖，大连理工大学**王立君**、**卢湖川**等完成的“鲁棒目标检测与跟踪理论及方法”、北京科技大学**马惠敏**等完成的“面向视觉任务的鲁棒特征表示与学习”、北京交通大学**丛润民**和天津大学**雷建军**等完成的“面向显著属性挖掘的跨模态交互与远距离感知理论和方法”、大连理工大学**刘日升**、**樊鑫**和北京大学**林宙辰**等完成的“时复杂环境感知的多域协同视觉计算理论与方法”获自然科学二等奖，北京邮电大学**马占宇**等完成的“智能多媒体内容安全中间件关键技术及应用”获技术发明一等奖，山东大学**张敬林**等完成的“多模态数据理解技术及其在典型灾害天气识别与预报中的应用”、北京交通大学**韦世奎**等完成的“面向交通场景的多传感器融合智能感知系统”获科技进步奖二等奖，西北工业大学**程塬**、上海交通大学**马超**、中科院计算技术研究所**闵巍庆**获青年科学家奖，北京理工大学**付莹**获石青云女科学家奖。

✪ 2023 年 12 月 5 日，第九届中国科协青年人才托举工程人选名单公示，共 752 人 (不包含特殊科技领域人选)， CCF-CV 专委会 3 位执行委员、北京航空航天大学**于茜**、华中科技大学**刘禹良**和南开大学**刘夏雷**入选。

✪ 2023 年 12 月 20 日，2023CCF 会士名单公布，CCF-CV 专委会副主任、中科院自动化所**王亮**和 CCF-CV 顾问委员会委员、西北工业大学**张艳宁**当选。

责任编委 刘海波

光学三维测量开源代码

东北大学 贾同 郭俏梅

随着科学技术和工业的发展，光学三维测量技术在自动化生产、质量控制、机器人视觉、反求工程、CAD/CAM 以及生物医学工程等方面的应用日益重要，逐步成为人们的研究重点。其难点在于在低信噪比（SNR）的环境下精度低鲁棒性差，解决这些问题将促进其在现实中的潜在应用。本文主要从被动式三维测量方面来介绍该领域的研究成果。

1、Depth Estimation by Combining Binocular Stereo and Monocular Structured-Light

介绍：针对弱纹理目标这一问题，该文提出了一种新型的立体成像系统，可以整合单目结构光和双目立体视觉的优势，如图 1 所示。它由两个相机(RGB 相机和红外相机)和一个红外散斑投影仪组成。RGB 摄像机用于深度估计和纹理采集。红外相机和散斑投影仪可以组成一个单目结构光子系统，而两个相机可以组成一个双目立体子系统。由单目结构光子系统生成的深度图可以为立体匹配网络提供外部引导，显著提高匹配精度。

深度估计流程如图 2 所示。首先对目标当前红外图像和散斑图像进行匹配，生成视差图 d_m。利用单目结构光子系统的标定参数，获得深度图 Z_m，投影到 RGB 相机坐标系统中。Z_m' 表示与 RGB 图像对齐的深度图，d_m' 表示对应的视差图。然后将 RGB 图像、IR 图像和视差图 d_m' 送入立体匹配网络（PSMNet 和 RAFT）估计最终的视差图。

该文收集了一个室内场景的真实测试数据集。定量结果表明，该系统的 Bad 2.0 误差为经典被动立体系统的 28.2%。

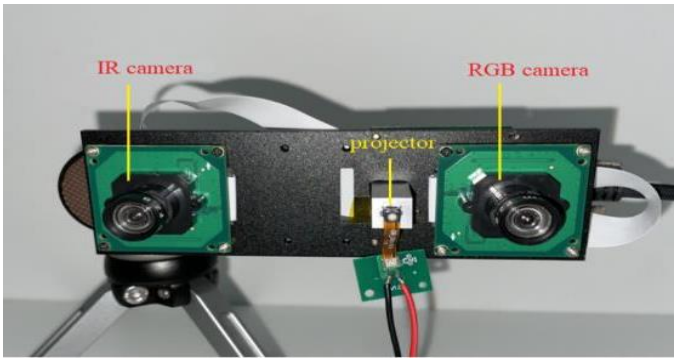


图 1 立体成像系统

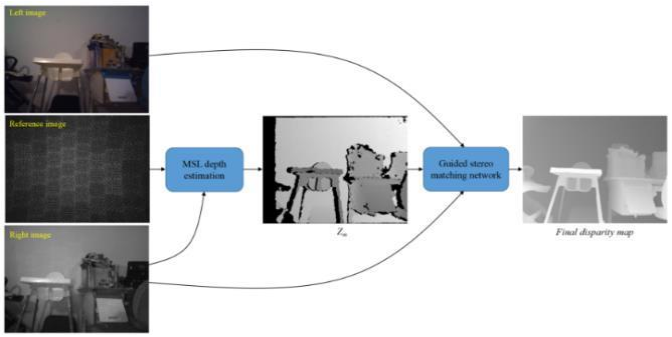


图 2 深度估计流程

论文地址：

https://arxiv.org/pdf/2203.10493v1.pdf

代码地址：

https://github.com/yuhuaxu/monostereofusion

2、AdaBins: Depth Estimation Using Adaptive Bins

介绍：该文提出了一种基于 Transformer 的结构块来执行场景信息的全局处理，它将深度范围划分为多个单元，每个单元的中心值自适应估计每幅图像，图像最终深度值估计为单元中心的线性组合。该方法解决了从单个 RGB 输入图像估计高质量密集深度图的问题。

架构由两个主要组成部分组成：(1) 基于预先训练的 EfficientNet B5 编码器和标准特征上采样解码器构建的编码器-解码器块；(2) 提出的自适应单元宽度估计块 AdaBins。第一个部分主要是基于 Alhashim 和 Wonka 的简单深度回归网络，并进行了一些修改。第二个部分是该文的关键贡献，即 Adabins 模块。Adabins 模块的输入是大小为($H\times W\times C_d$)的解码特征，输出是大小为 (H×W×1) 的张量。由于当前 gpu 硬件的内存限制，文章用 h=H/2 和 w=W/2 以便在批量较大的情况下更好地学习。通过简单的双线性上采样到 H×W×1 计算得到最终的深度图。总体结构如图 3 所示。

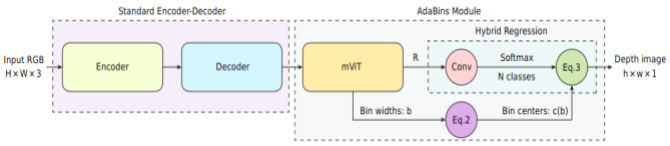


图 3 深度估计体系结构

该文对于两种最流行的数据集 NYU 和 KITTI 进行了实验，结果显示，在所有指标上与最先进的技术相比有了决定性的改进，并通过消融实验验证了该模型的有效性。

论文地址：https://arxiv.org/pdf/2011.14141.pdf

代码地址：https://github.com/shariqfarooq123/AdaBins

3、Attention Concatenation Volume for Accurate and Efficient Stereo Matching

介绍：立体匹配是许多视觉和机器人应用的基本组成部分。代价体在立体匹配中至关重要，该文提出了一种新的代价体构建方法，该方法利用相关线索生成注意权重，以抑制连接体中的冗余信息，增强匹配相关信息。此外，还提出了一种多层级自适应块匹配策略来得到具有区分性的匹配特征，该方法可以提高弱（无）纹理区域的特征表达能力。

所构建的代价体 Attention Concatenation Vloume (ACV) 可以作为子模块接入到其它模型中，使得模型轻量化的同时提升精度。ACV 的构建过程包括三个步骤:初始级联体构建、注意权重生成和注意过滤。利用生成的注意权重对初始级联量进行过滤，可以抑制冗余信息，增强匹配相关信息，得到注意级联量。

在 ACV 的基础上，该文设计了一个精确高效的端到端立体匹配网络，命名为 ACVNet。模型架构如图 4 所示。该网络由一元特征提取、注意力级联量构建、成本聚合和视差预测四个模块组成。

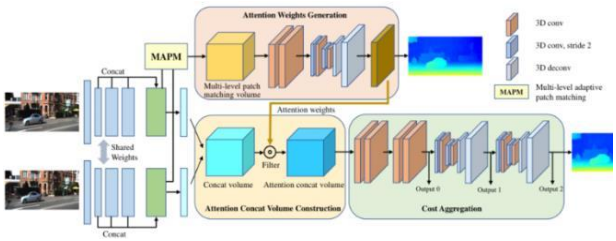


图 4 ACVNet 模型架

该文还构建了一个实时版本的 ACVNet，命名为 ACVNet-fast。ACVNet-fast 采用与 ACVNet 相同的特征提取方法，但减少了层次和视差预测模块。图 6 展示了 ACVNet-fast 的架构，ACVNet-fast 与 ACVNet 的主要区别在于 ACV 的构建和聚合。

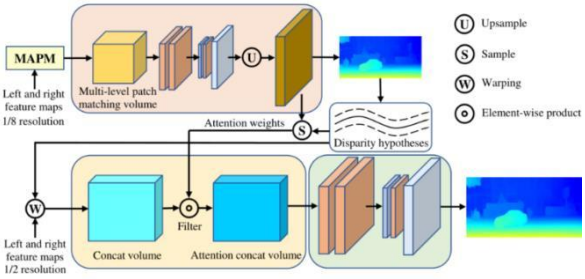


图 5 ACVNet- fast 架构

ACVNet 方法在 KITTI 2012&2015、Scene Flow 和 ETH3D 四个公共基准上都表现出了最先进的性能，也是唯一一个在四个数据集上同时排名前 5 的方法，这说明该方法对各种场景具有良好的泛化能力。

论文地址： https://arxiv.org/pdf/2203.02146.pdf

代码地址： https://github.com/gangweiX/ACVNet

责任编委 王田 李策

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、图像处理和深度学习等领域。电子邮箱：jiatong@ise.neu.edu.cn

郭俏梅

博士研究生，东北大学信息科学与工程学院，研究方向为三维感知、计算机视觉和深度学习。

电子邮箱：2210300@stu.neu.edu.cn

AI 生成图像检测数据集

华为 朱铭健 陈汉亭 胡海林 王云鹤

在这个 AIGC 技术快速发展的时代，人人都可以利用 AI 算法生成高质量的文本，图像，音频内容。其中，由 Midjourney，Stable Diffusion 等图像生成方法制作的图像已经能够以假乱真，让人眼难以分辨了。这不禁唤起了人们的隐忧：大量虚假图片将会在互联网上广泛传播。进而引发多种社会安全问题。例如，虚假新闻会扰乱社会秩序，混淆视听。恶意的人脸图片造假则会引发金融欺诈，造成信任危机。近期一五角大楼起火的 AI 生成虚假图片骗过了几家主要新闻机构，并导致美国股市大幅下跌。因此，对这些 AI 生成的图像进行有效监管是非常有必要的，AI 生成图像检测目标在于对真实的图片和 AI 生成的图片进行有效的二分类。为了有效训练和评估 AI 生成图像检测模型，人们提出了多种多样的 AI 生成图像检测数据集。

表 1 现有 AI 生成图像检测数据集比较

Dataset	Image Content	Generator Category		Public Availability	Real Images	Fake Images
		GAN	Diffusion			
UADFV [30]	Face	✓	✗	✗	241	252
FakeSpotter [25]	Face	✓	✗	✗	6,000	5,000
DFFD [3]	Face	✓	✗	✓	58,703	240,336
APFDD [6]	Face	✓	✗	✗	5,000	5,000
ForgeryNet [9]	Face	✓	✗	✓	1,438,201	1,457,861
DeepArt [27]	Art	✗	✓	✓	64,479	73,411
CNNSpot [26]	General	✓	✗	✓	362,000	362,000
IEEE VIP Cup [24]	General	✓	✓	✗	7,000	7,000
DE-FAKE [22]	General	✗	✓	✗	20,000	60,000
CiFAKE [1]	General	✗	✓	✓	60,000	60,000
GenImage	**General**	✓	✓	✓	**1,331,167**	**1,350,000**

本文重点介绍 AI 生成图像检测的数据集，包括 GenImage，ForenSynths，CiFAKE。上图从图像内容，生成器类别，可获得性和真假图片数量这几个角度对比了现有的数据集。AIGC 检测的难点在于跨生成器的图片检测和退化图像的检测。为了解决这个问题，数据集含有更多图片，更多种类的生成器以及更丰富的图片内容就很重要了。

1、ForenSynths 数据集

介绍：这个数据集由 UC Berkeley 等单位发布于 CVPR2020，是目前比较常用的数据集之一。这个数据集包含了 ProGAN，StyleGAN，BigGAN，CycleGAN，StarGAN，GauGAN，CRN，IMLE，SITD，SAN 和 FaceForensics++生成的图像，约 36 万张。真实图片则来自于 LSUN，ImageNet，Style/object transfer，CelebA，COCO，GTA，Raw camera，Standard SR benchmark 和 Videos of faces，约 36 万张。

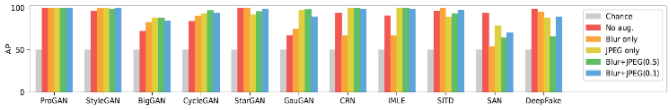


图 1 采用数据增强可以提升性能

这个数据集提出通过在训练阶段对数据进行模糊，压缩等预处理，能够提高检测性能。这个方法在实验上得到了验证，见图 1。这个数据集的缺点在于不包括近两年备受关注的扩散模型生成的图片，而这种模型生成图片的质量很高，很可能被用于生成假新闻图片。

相关论文链接：CNN-Generated Images Are Surprisingly Easy to Spot...for Now
https://arxiv.org/abs/1912.11035

数据集下载地址：

https://github.com/PeterWang512/CNNDetection/tree/master

2、CiFAKE 数据集

介绍： CiFAKE 数据集由 Nottingham UK 等单位于 2023 年发布，其基于 CiFAR 数据集的 60000 张真实图片来构建。通过使用 Stable Diffusion V1.4 来生成 60000 张假图。在数据内容方面，CiFAKE 数据集包含 Airplane，Automobile，Bird，Cat，Deer，Dog，Frog，Horse，Ship 和 Truck。由于 CIFAR 的图片数量和类别数量都不多，而且 CiFAKE 只采用了一种生成器来生成图像，所以这个数据集在图片数量，生成器数量以及图片内容丰富程度上都不如 ForenSynths。

相关论文链接： CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images

https://arxiv.org/pdf/2303.14126.pdf

数据集下载地址：

https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images

3、GenImage 数据集

介绍：GenImage 是一个百万级别大规模 AI 生成通用图像检测数据集，由华为在 NeurIPS 2023 Track on Datasets and Benchmarks 发布。这个数据集的重要意义在于为 AI 生成图像检测任务提供了最新的经过精心调控的大规模数据集，而且在生成器选择，数据集规模和图片内容上都具有显著优势。采用这个数据集将对虚假新闻识别，人脸造假等问题有所帮助。

这个数据集中真实的图片采用了 ImageNet 数据集。虚假的图片采用 ImageNet 的标签进行生成。GenImage 是 Midjourney, Stable Diffusion V1.4, Stable Diffusion V1.5, ADM, GLIDE, Wukong, VQDM 和 BigGAN。下图展示了 GenImage 的一些生成样本。

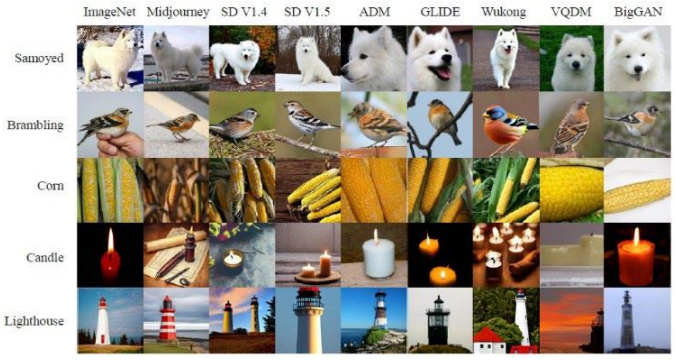


图 2 GenImage 样本示例

这个数据集具有以下优势：1. 大量的数据：超过百万对图片对；2. 丰富的图片内容：利用 ImageNet 进行构建，具有丰富的标签；3. 先进的生成器：覆盖 Midjourney, Stable Diffusion 等 Diffusion 生成器。

表 2 跨生成器检测能力评估

Method	Testing Subset								Avg
	Midjourney	SD V1.4	SD V1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Acc.(%)
ResNet-50 [8]	54.9	**99.9**	99.7	**53.5**	61.9	98.2	56.6	52.0	72.1
DeiT-S [23]	55.6	**99.9**	99.8	49.8	58.1	98.9	56.9	53.5	71.6
Swin-T [13]	**62.1**	**99.9**	99.8	49.8	**67.6**	99.1	**62.3**	**57.6**	**74.8**
CNNSpot [26]	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
Spec [31]	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8
F3Net [19]	50.1	**99.9**	**99.9**	49.9	50.0	**99.9**	49.9	49.9	68.7
GramNet [14]	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9

表 3 退化图像检测能力评估

Method	Testing Subset						Avg
	LR (112)	LR (64)	JPEG (q=65)	JPEG (q=30)	Blur (σ=3)	Blur (σ=5)	Acc(%)
ResNet-50 [8]	96.2	57.4	51.9	51.2	**97.9**	69.4	70.6
DeiT-S [23]	97.1	54.0	55.6	50.5	94.4	67.2	69.8
Swin-T [13]	97.4	54.6	52.5	50.9	94.5	52.5	67.0
CNNSpot [26]	50.0	50.0	**97.3**	**97.3**	97.4	77.9	78.3
Spec [31]	50.0	49.9	50.8	50.4	49.9	49.9	50.1
F3Net [19]	50.0	50.0	89.0	74.4	57.9	51.7	62.1
GramNet [14]	**98.8**	**94.9**	68.8	53.4	95.9	**81.6**	**82.2**

这个数据集的难度在于跨生成器检测问题以及图像传播过程中的图像退化问题，这也是现实场景下往往会遇到的问题。由于目前的生成器多种多样，未来也会不断出现新的生成器。训练集往往无法包含现实场景中需检测图片所对应的生成器。因此检测器需要具有跨生成器检测能力，即能够检测出不在训练集中的生成器所生成的虚假图片。GenImage 提出了一个 benchmark 去评估现有检测器的跨生成器检测能力。将来新提出的检测器也可以在此基础上进行有效评估，见表 2。模型在 Stable Diffusion V1.4 上面训练，然后在多个检测

器上测试验证。图像在互联网传播过程中往往会遇到图像退化问题，例如模糊，噪声，JPEG 压缩。GenImage 在 Stable Diffusion V1.4 上进行了图像退化操作，并对现有检测器进行了有效评估，见表 3。

这个数据集不仅提供了海量且丰富的数据，还提供了非常全面的评测数据和指标，为 AIGC 检测打造了坚实的基础。未来这个数据集的应有价值在于识别各类 AI 生成的图片以解决虚假新闻识别问题，此外，AI 生成检测图片任务还可以作为版权保护流程中的一环。目前 GenImage 涉及的是拍摄类图片，未来可以改进的方向是加入更多风格，比如漫画，油画等。此外，更多提示词生成的图片也可以加入。

相关论文链接： GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image https://arxiv.org/abs/2306.08571

数据集下载地址： https://github.com/GenImage-Dataset/GenImage

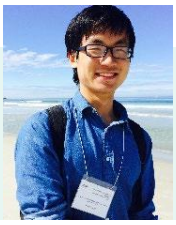
责任编委 李策 樊鑫

朱铭健

华为诺亚方舟实验室研究员，浙江大学博士，研究方向包括计算机视觉，高效神经网络部署等

陈汉亭

华为诺亚方舟实验室高级研究员，北京大学计算机专业博士，研究方向包括计算机视觉，自然语言处理，高效神经网络部署等



胡海林

华为诺亚方舟实验室高级研究员，清华大学博士，研究方向包括自然语言处理，序列分析，自监督学习等

王云鹤

华为诺亚方舟实验室算法应用部部长，北京大学计算机专业博士，研究方向包括计算机视觉，自然语言处理，高效神经网络部署等

好 文 推 荐

哈尔滨工业大学的“CLIP2Point: Transfer CLIP to Point Cloud Classification with Image-Depth Pre-Training“最新成果发表在 ICCV 2023。

论文：Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson W.H. Lau, Wanli Ouyang, Wangmeng Zuo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22157-22167.

得益于大规模的预训练，视觉-语言模型可为视觉和语言数据提供统一的知识表征，因而可以广泛应用于各种计算机视觉和跨模态下游任务。然而，许多实际应用中还会涉及更多的模态如 3D 点云、语音等，这些模态因为缺乏大规模的文本对训练数据，难以构建与自然语言的统一表达。为了解决这一问题，本工作以 3D 点云为例，提出了以视觉为中介，建立 3D-视觉-语言统一知识表征的预训练方法。本工作通过 3D 和视觉间的对比学习预训练，以对齐 3D-视觉表征的方式关联视觉-语言的表征，进而实现 3D-视觉-语言的统一知识表征。此外，在下游任务的适配上，提出了一种轻量化的双路适配模块，将预训练知识有监督地应用在点云分类中，取得了目前最好的分类性能。本工作的研究方法可以进一步运用在点云的分割、检测等下游任务中，对融合其他模态的知识表达也有一定的借鉴和启发意义。

本工作提出了一个基于对比学习的视觉-点云自监督预训练方法，将预训练分为模态间和模态外两部分，模态间对比学习能够拉近 2D 视觉的彩色图像和点云投影的深度图存在的域差异，而模态内对比学习可以增强深度编码器对距离信息编码的鲁棒性。具体而言，对于给定的 3D 物体和相对应的点云数据，本方法分别渲染出彩色图像和深度图像的跨模态匹配对，其中深度图像通过随机的距离采样进一步得到模态内的匹配对。彩色图像通过预训练的视觉-语言模型（CLIP）提取特征，而深度图像通过可学习的深度编码器提取特征，并通过两组匹配关系的对比学习，逐步和 CLIP 对齐。这其中，模态间训练的提出动机比较自然，是为了将彩色图像和深度图的特征域拉近；而模态内的训练，是考虑到点云在三维空间中的分布是离散且不规则的。即使从同一距离的不同视角来投影深度图，得到的距离分布也可能是完全不一样的，通过随机采样投影距离的方式能够让深度编码器更加稳定地编码出相应深度特征。

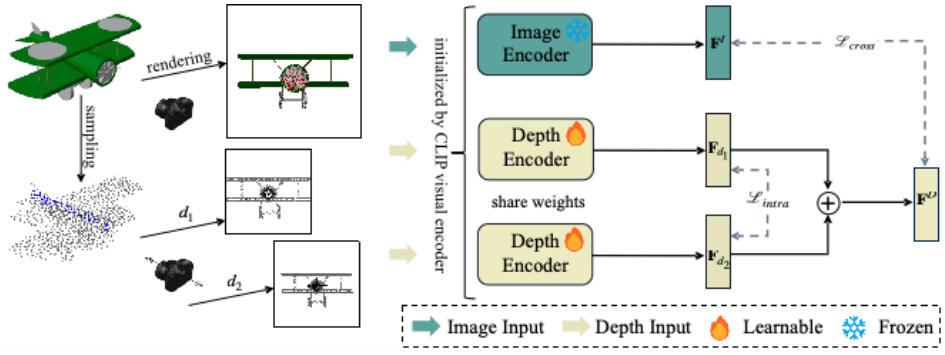


图 1 视觉-点云对比学习自监督预训练示意图

好 文 推 荐

西安电子科技大学的最新成果“TIB: Detecting Unknown Objects via Two-Stream Information Bottleneck”发表在 IEEE TPAMI 2023。

论文：Aming Wu, Cheng Deng. TIB: Detecting Unknown Objects via Two-Stream Information Bottleneck, IEEE TPAMI, 2023.

基于深度神经网络的目标检测方法已经实现了出色的检测效果。通常，大多数现有方法往往遵循一个“封闭”的假设，即训练集和测试集的类别空间是一致的。然而，现实场景是开放的，充满了各种各样的未知物体。极大限制了基于“闭集”假设的方法的应用。为了加速检测网络在实际场景中的安全部署，最近提出了分布外目标检测任务（Out-of-Distribution Object Detection, OOD-OD），旨在让基于分布内数据训练的目标检测器能够在没有给定任何分布外数据的情况下准确检测出相应的分布外物体。由于缺乏用于训练的分布外数据，导致基于分布内数据训练的检测器无法学到一条辨别分布内和分布外物体的边界，出现较多误检和漏检的情形。一种可行的解决方案是基于分布内特征来合成大量的用于训练的分布外特征，从而提升模型的能力。因此，本文提出了一种被称为双流信息瓶颈（TIB）的新方法，旨在从信息论的角度来解决这个任务。

在假设分布外数据往往是和当前任务无关的基础上，给定输入图像，本文首先使用一个由多个卷积层组成的骨干网络来提取相应的特征图，进而定义了一个标准信息瓶颈网络来处理当前特征图。由于标准信息瓶颈具有弱化任务无关信息的优势，因此，经过标准信息瓶颈网络输出的特征图将包含丰富的物体相关内容和更少的背景信息，有助于提升物体的定位能力。接下来，为了缓解缺失分布外数据的影响，该方法定义了一个反向信息瓶颈网络来处理骨干网络提取的特征图。本文发现通过简单地反转标准信息瓶颈的优化目标可以促进输出的结果包含更多的任务无关信息，从而获得期望的分布外特征。同时，本文还发现由检测模型的提取的候选物体特征往往包含较多的背景信息，影响了检测器定位和分类物体的准确性。为此，本文提出了一个混合信息瓶颈模块来动态地净化候选特征中物体相关的内容，去除物体无关信息的干扰，提升了候选物体特征的判别性。在实验中，本文在三个不同的检测任务上对该方法进行了评估。多个数据集的实验结果以及大量的可视化分析表明该方法能够有效提升检测分布外物体的能力。

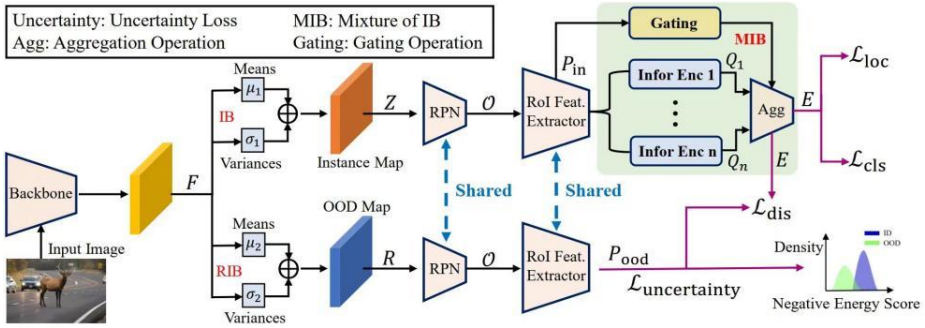


图 1 TIB 分布外物体检测流程

责任编委 贾同 李策

好 文 推 荐

加拿大多伦多大学的最新成果“Shape-Based Measures Improve Scene Categorization”发表在IEEE TPAMI 2023。

论文：Morteza Rezanejad, John Wilder, Dirk B. Walther, Allan D. Jepson, Sven Dickinson, Kaleem Siddiqi. Shape-Based Measures Improve Scene Categorization, IEEE TPAMI, 2023.

场景分类旨在通过理解整个图像将场景图像分类为预定义的场景类别之一，是计算机视觉中一个长期存在、基础且具有挑战性的问题。场景类别通常由图像上下文中一组对象定义。例如，海滩场景通常包含遮阳伞、沙滩椅和穿着泳衣的人，这些都位于水体旁边。街景可能包括带有汽车、自行车和行人的道路，以及沿路两侧的建筑物。

当前基于卷积神经网络的计算机视觉系统在各种任务中实现了优越的性能。尽管基于卷积神经网络的算法在近年来改变了计算机视觉和机器学习的格局，但这些模型仍然缺乏人类视觉系统具有的一些关键能力。目前的深度神经网络模型通常对数据需求极大，并且未必能够代表所有结构性视觉线索。具体而言，大量实验数据表明，在大型数据集上训练的深度神经网络模型更偏好颜色和纹理信息。相比之下，人类更易从图像以及边界轮廓中识别图像中的对象和场景。

人类中层视觉通过 Gestalt 规则，将简单的初级特征重新组合并组织成更复杂的特征。尽管这些感知组合规则在大量文献中有定性描述，但目前尚未对其在场景分类应用中进行实际实现。本文研究发现，尽管这些轮廓组合度量可以直接从图像中的轮廓计算得出，但当前的卷积神经网络模型没有提取或利用这些组合线索。由此，本文提出一套基于轮廓线索的复杂场景分类算法。如图 1 所示，本文使用中轴变换（MAT, Medial Axis Transform），根据组合规则在图像局部对轮廓进行评分。本文通过两种方式展示轮廓线索在场景分类中的益处：(i) 当强调感知组合信息时，人类观察者和卷积神经网络模型都能准确地对场景进行分类。(ii) 使用度量对轮廓进行加权相比于不加权轮廓的方式显著提高了卷积神经网络模型的性能。大量实验说明本文所提方法可大幅提升卷积神经网络的场景分类性能。

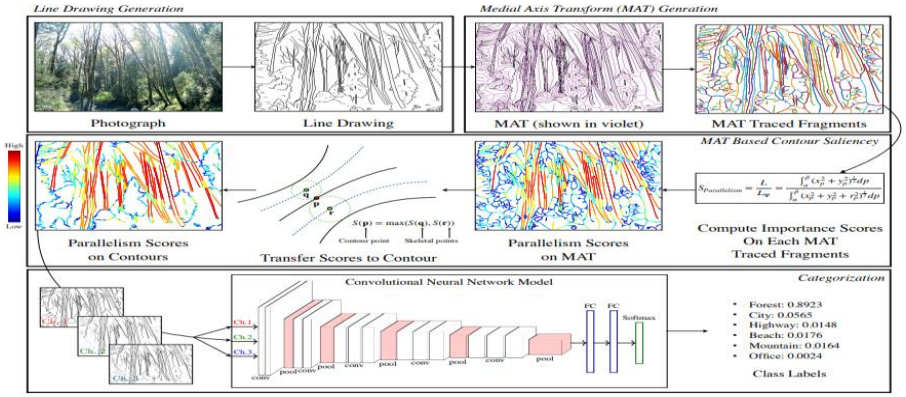


图 1 所提方法示例说明

责任编委 贾同 王田

征文通知

1 会议征文

计算机视觉领域相关国内外会议的征文通知如表 1 所示。同时，可继续关注每个会议举办的 workshop 或 special session。

2 期刊征文

计算机视觉领域近期相关期刊专刊的征文通知如表 2 所示，包括 IEEE Journal of Biomedical and Health Informatics, Pattern Recognition Letters, Image and Vision Computing 和 IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing。

3 会议简介

中国模式识别与计算机视觉学术会议 PRCV (Chinese Conference on Pattern Recognition and Computer Vision)，由中国计算机学会（CCF）、中国自动化学会（CAA）、中国图象图形学学会（CSIG）和中国人工智能学会（CAAI）联合主办，定位国内顶级的模式识别和计算机视觉领域学术盛会。

第七届 PRCV 将于 2024 年在乌鲁木齐举办，由新疆大学承办。会议旨在汇聚国内国外模式识别和计算机视觉理论与应用研究的广大科研工作者及工业界同行，共同分享我国模式识别与计算机视觉领域的最新理论和技术成果。通过此次会议，进一步加强本领域的同行与东南沿海地区的学者和企业进行学术交流和技术碰撞，从而促进模式识别与计算机视觉领域的协同合作与融合创新。

责任编辑：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
CHIL 2024	2024.06.27-28	New York, American	2024.02.06	https://www.chilconference.org/index.html
ICML 2024	2024.07.21-27	Vienna, Austria	2024.02.02	https://icml.cc/Conferences/2024
ECCV 2024	2024.09.29-10.04	Milan, Italy	2024.03.07	https://eccv2024.ecva.net/
MICCAI 2024	2024.10.06-10	Marrakech, Morocco	2024.03.08	https://conferences.miccai.org/2024/en

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
PRL	Multiscale Pattern Recognition for Multifractal Data (MSPR)	https://www.sciencedirect.com/journal/pattern-recognition-letters/about/call-for-papers	2024.01.20
IEEE J-STARS	Advances in Satellite Image Quality Enhancement with Deep Learning Techniques	https://www.grss-ieee.org/wp-content/uploads/2023/04/cfp_Advances-in-Satellite-Image-Quality-Enhancement-with-Deep-Learning-Techniques.pdf	2024.01.31
IVC	AI on Digital Health: Computer vision applications in medical imaging	https://www.sciencedirect.com/journal/image-and-vision-computing/about/call-for-papers#ai-on-digital-health-computer-vision-applications-in-medical-imaging	2024.02.01
IEEE JBHI	Foundation Models in Medical Imaging	https://www.embs.org/jbhi/wp-content/uploads/sites/18/2023/09/JBHI_Foundation-Models_SI.pdf	2024.03.15

心底无私视界宽∞刘允才教授专访

近 50 年来，我国在计算机视觉领域展开了相关的科研工作。而今，我国已经拥有了一支庞大的、在这一领域辛勤耕耘且能与世界一流水平并驾齐驱的科研队伍。在这一过程中，有一批见证了视觉领域发展、为我国计算机视觉领域的奠基做出了重大贡献的先驱者。

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

我是负责本次专访的主要采访人，复旦大学张军平。本次采访通过微信交流完成，相关问题由 CCF-CV 专委会《视界专访》和刘老师的学生严骏驰等提供。为能更好地帮助我们回顾本次采访，我们采用了问答加书面回顾的形式来表述。以下是刘允才教授的简介和专访内容。

张军平（采访者，后缩写为张）：刘老师您于上世纪 80 年代去美留学深造，是什么样的契机让您拜在 Thomas S. Huang 黄煦涛院士门下读博？当时计算机视觉发展还属于早期阶段，您是先选择了这个专业方向再联系的黄教授，还是先联系好了黄教授再确定的方向？

刘允才（后缩写为刘）：我是国内 78 届的研究生，研究方向是图像处理。那时自认为是相当前沿的研究领域。1984 中国科学院研究生院和清华大学计算机系等联合邀请黄煦涛教授来北京讲学，我惊喜地获知国外还有一个计算机视觉的学术领域，已经在国际上取得了相当的成就。当即我就决定将这个研究方向作为自己的追求目标。我立刻写信与教授联系，向他汇报了我的硕士研究课题和成果情况，提出了追随他学习的愿望。很快就得到了回复。于是 1984 年 12 月中旬我以访问学者的身份加入黄教授的实验室，1985 年秋季入学正式成为黄教授的博士研究生。这里还有一个小插曲。据说黄教授 1985 年的博士招生人选本已经确定为陈长汶先生，由于我的出现，长汶不得不延迟一年入校，变成我的师弟。

张：能否分享一些您在 UIUC 读博士学位期间的一些印象深刻的事。听闻黄教授在山东大学访问期间公开给大家讲了您做梦解决了一个研究难题，这是真的吗？

刘：我记忆最深刻的就是圣诞节。我在 UIUC 学习了七年，赶上了八个圣诞节。这八个圣诞节都是在黄教授家中度过的，有时是单独邀请我们一家，有时是以 Party 的形式邀请所有的学生。记得第一次过圣诞节，是我刚到美国十多天，对一切还感到新奇。那天大雪纷飞，黄教授亲自开车和他的小儿子一起去我的住所接我。在黄教授家，我第一次看到了挂满彩球和装饰物的圣诞树，第一次品尝到了美味的烤火鸡。顺便说一下，每次圣诞节，黄教授一定会在餐前将刚烤好的整只火鸡端到餐厅展示给大家，然后才拿回厨房分切。那天我们都非常愉快，黄太太高兴地给我讲述圣诞树上每个挂物的来由，哪个是女儿制作的，哪个是某年某朋友赠送的。晚餐后黄教授和黄太太还兴致勃勃地开车载我到香槟城去看城市的雪景和灯光，然后送我回家。出国前我只从电影

和小说里看到过圣诞节的情景，是在黄教授的家中，我才真正地体会到了圣诞节的温馨、友爱与美好。

图一：（上）1987 年在黄教授家过圣诞节，女儿刚到美国还很怕生；（中）黄教授端出刚刚烤好的火鸡；（下）1991 年在美国的最后一个圣诞节，与黄教授夫妇合影

黄教授经常邀请著名学者来实验室进行学术交流，对我印象最深刻的是邀请德克萨斯大学的 J. K. Aggarwal 教授。他有一项研究与我的博士研究课题相同——基于图像线对应元的三维运动分析，因而印象很深刻。Aggarwal 是印度裔美籍教授，非常友善，与我的导师黄教授是同一辈人。由于是原创性课题，大家都希望能最先发表自己的研究成果。我们和 Aggarwal 教授在一起讨论分析课题，阐述各自的研究思路，分手后又都争分夺秒地做实验写论文，与时间赛跑。结果，我们和 Aggarwal 教授的研究成果同时在会议 ICPR 1986 上发表，以不同的方法解决了同一个计算机视觉领域的难题。不过，Aggarwal 的论文结论有些偏差，颇为之惋惜。我回国后，还特邀 Aggarwal 教授来我们交大的实验室交流访问。

关于做梦解决研究难题，那是黄教授对我工作状态的幽默描述。在发表了图像线元的三维运动分析的非线性算法后，我们开始寻求线性算法。以前没有人搞过，因而不知如何下手，只能不断地试错。这项工作实际上是一系列的数学推导，找出求解运动参数的线性公式。我对这项研究太感兴趣了，或许有点走火入魔，日以继夜地公式推导。白天推导走不通，晚上接着推导；晚上的推导行不通，躺在床上继续思考。有时半夜突然醒来，觉得有了思路，立刻从床上爬起来按照蒙眬的思路推导，往往还是不能成功。不记得这样不分昼夜地推演了多少时间，最终我导出了运动分析的线性算法。

张：您有诸多原创性成果，您最得意的成果有哪些？

刘：在 UIUC 读书时做的线元运动分析、角元运动分析、常运动分析和我的第一个博士生所做的连接刚体运动分析都是原创性研究，其他的研究成果应该属于创新或者改进。回国后很多研究课题属于应用基础性的研究，很难出原创性的成果。

我在上海交通大学工作中做的最得意、投入最多、也是最遗憾的项目是一个应用性的研究课题：全膝关节置换机器人外科手术。该项目与著名的上海膝关节手术专家合作，历经十二年、八代博士研究生的刻苦研究与实验，制造成了符合手术室消毒要求的临床手术机器人样机。在这十二年的研究中，进行了 400 多例骨骼模型手术实验和数十例人尸骨骼准临床手术实验，膝关节的

切割精度高于医生的手工操作。在此过程中，升级更新了三代手术机器人样机，也获得了国家科技发明奖和上海市科技进步奖。但在积极准备申请活体临床手术实验的时候，我到了退休年龄，与我合作的骨科专家不久也相继退休。结果，十多年的辛苦研究成果只能搁置。如果这架手术机器人在我退休前能够能完成数例甚至一例活体临床手术，该项研究成果或许就有了应用推广的意义。也许，我对课题的进度规划不周，或者提出的这个课题超前了一些，造成了我这段学术生涯的遗憾。

张：作为计算机视觉泰斗级人物黄煦涛院士的学生，能否谈谈您从他那学到了哪些治学理念，您又是如何考虑您自己的研究生培养呢？

刘：黄教授一生培养出数以百计的计算机视觉领域中的栋梁人才，通过学术交流而受益者不计其数。在黄先生的实验室学习过程中，我体会到作为培养人才之地，首先要有一个清正严谨的学风，其次要有一个自由开放的学术环境。在培养学生方面，我最强调的也是学风。学风正才能出人才，学风正才能出成果。学生们在一个风清气正、积极向上的环境中工作学习，才能把自己的研究当作兴趣去追求，才能充分发挥自己才能和潜力。为此，我在每个学期的第一例会上都要讲学风。其次就是自由开放的学术环境，要与学生在学术上做到真正平等，鼓励学生自由思考。我曾多次对我的学生讲，在课堂上和办公室里我是老师；在实验室、会议室就没有老师学生之分，人人平等，欢迎提出不同的意见，允许争论。再则，要时刻留意学生的具体困难。硕士、博士研究生都是成年人了，偶尔会出现身体上、经济上或者感情上的问题。这时就需要老师及早发现，及时给与帮助和疏导。作为导师，不仅要注重学生的学业进步，也要关心他们身心健康的发展。教书育人永远是一名教师的宗旨。

张：您博士毕业在 UIUC 贝克曼研究院工作后，去了住友电气公司工作了约十年时间，是什么原因促使您和夫人于 2000 年回国加入了上海交通大学，回归了学术界？

刘：我去日本住友电气公司工作是导师黄教授推荐的，公司领导和员工对我都十分友好，工作顺心应手，我也很喜欢这个工作。但工作久了发现，在公司工作没有那么“自由”。研究课题必须与公司的方向和发展利益一致；由于公司不同于学术界，有些研究成果也不一定能得到及时发表。因此，我想改变一下工作环境，给自己一个更加自由的工作空间，回归学术界当然是最好的选择。另一方面，随着年龄的增长，希望能把自己的毕生所学传授给年轻一代，在他们身上结出更加丰硕的果实，这也是我希望到学校工作的动力。当然，回国是我的第一选择，我从小学到研究生毕业都是国家培养的，应该给国家做一些事情。

图二： 2009 年上海交大视觉实验室成员合影

张：您在工业届和学术界都取得了很大的成就，请问这两者在工作上的主要区别在哪里？解决问题的思路有什么共通的地方？

刘：对于基础理论的研究，在工业届和学术界的研究思路及研究方法的差异不是很大，可以说是共通的。但是在研究项目的选题和内容的侧重方面，在学术界会有较大的自由度，可以根据研究者的兴趣申请课题。但是在企业界，首先要考虑你的研究课题是否会对公司的发展能有帮助，多长时间公司可以从中获得效益。另外，在学术界研究成果大多以学术论文的形式展现，而企业界多以技术报告的形式提交。研究成果能否公开发表，一般需要较多方面的考虑。还有，在撰写学术论文时，作者往往会把自己最好的实验结果放进论文，而技术报告则要求将成功的实验结果和失败的结果都要写进报告，并对不满意的结果进行分析。从这一角度看，工业届的研究相对学术界要求似乎更高了一些。

张：您在计算机视觉和智能交通领域都有卓越的建树，

您如何看待马斯克在研究自动驾驶只希望依赖视觉摄像头进行检测导航?

刘: 马斯克研究的自动驾驶已经做到了相当成熟的程度。国内不少企业也在研究自动驾驶，并且取得了非常好的成绩，如百度、华为、文行知远等。深圳市已经出现了无人驾驶出租车试验。实际上，人们很早就追求汽车的自动驾驶，早在上个世纪的七十年代末，日本的筑波工程实验室就开发出第一辆基于摄像头检测导航的自动驾驶汽车；八十年代，德国就完成了能在高速公路上以每小时 70 英里的速度行驶的、基于计算机视觉技术的汽车自动驾驶实验。目标商业化的自动驾驶的研究是近十几年的事情。视觉摄像头是当下汽车自动驾驶的主要传感器。这当然不是唯一的，一个实际的应用系统应该尽可能采用它所能够使用的传感器，使系统的控制变得简单可靠，譬如，除了视频摄像头外，激光传感器和超声波传感器也是自动驾驶中最常用的传感器。设计者也一直希望能够将声音传感信息有效地融合到自动驾驶系统。

图三：1994 年在新加坡组织 IVHS 东南亚区议会时在会议入口处留影，IVHS(智能车辆与高速公路系统）是智能交通系统（ITS）的前身

目前，自动驾驶的汽车一般仅限应用在转运货场内、停车场内以及交通专线上。在公共交通环境中的普遍应用还需要等待相关法律法规的出台。如，交警是否要拦截无人驾驶的汽车？自动驾驶的汽车发生了交通事故应该由谁赔偿，保险公司、车主还是汽车制造商，等等。自动驾驶也是智能交通系统研究的范畴。智能交通系统的远期目标是人-车-路之间高度协调。在道路、车辆上都安装了各种传感器和网路装置，道路与道路之间、道路与车辆之间、车辆与车辆之间都有无线网络连接，传送各种协调控制数据，协调路网上的车辆以最佳的状态运行。那时，人们将会乘坐更安全、更快速、更舒适的无人驾驶的汽车在道路上行驶。

张: ChatGPT 展现出了非常强大的人机交互功能。图像与视频的人机交互性能似乎不及文字语音，挑战性在哪里?

刘: 人工智能技术正在突飞猛进地发展，GPT-4 Turbo 已经发布，展现了非常强大的功能。ChatGPT 的人机交互优势当前主要侧重于文字语音方面，对图像视频的交互仍然有着巨大的发展空间。图像与视频的人机交互主要涉及到两个技术领域，计算机视觉和计算机图形学。我这里只谈计算机视觉，传感器是摄像头，它处于人机交互的输入端，感知外界的图像信息，是计算机的眼睛。计算机视觉已经有了近半个世纪的发展历程，但是相对于语言和文字，仍然是一个相当稚嫩的领域，许多研究成果目前尚不能很好地应用于实际。语言与文字乃是人类数千年乃至上万年文明发展的结晶，每个字或者单词都有相对固定的含义。在语言中，特别是西文 (法、德、俄、英等)，句子中的单词之间又具有严格的语法约束。这些人类文明发展的成果对计算机理解文字世界起到了巨大的帮助。然而对于视觉来说，人机交互所输入的图像或视频对于计算机基本上是一大宗原始数据。尽管深度学习方法对视频图像的识别与理解具有很大的推动，但理解图像中的人物或者目标之间的语义关联对于计算机仍然是一个相对困难的任务。如果有一天，计算机成为真正的“电脑”，它将摄像头“看到”的图像或视频能够准确、快速、高效地转换成文字描述，视觉人机交互的时代或许即将到来。

张: 能否给现在从事计算机视觉研究的青年工作者一个寄语?

刘：年轻的计算机视觉工作者朋友们，很高兴看到你们选择了计算机视觉这个充满创新和挑战的领域。在过去几十年中计算机视觉的研究取得了巨大的进步，但仍然是一个充满无限可能性的领域，等待着你们新的发现和贡献。在这里，我想给你们谈一下我的研究心得。作为一名青年研究者，要始终保持你对这个领域的好奇心和热情。计算机视觉是一个快速发展的领域，不断涌现出新的技术和方法。只有持续学习和保持兴趣，你才能跟上最新研究趋势的发展，为这个领域的进步做出贡献。其次，研究中不要畏惧困难的挑战。研究中不可避免地会遇到困难或困惑，千万不要放弃，坚持下去。这些困难与困惑或许是你的创新机遇，尝试新的思路和方法，你可能做出突破。此外，与国内外的技术同行学术交流非常重要。每个人对问题都有不同的观察角度和思考方法，通过交流，你可以受到启示，发现新的问题，创造出新的方法。当前计算机视觉正在蓬勃发展，已经在医疗、工业、交通、民生等领域有了较为广泛的应用，但仍然具有极其巨大的探索和发展空间。年轻的计算机视觉工作者朋友们，你们是这个领域的未来和希望。愿你们勇敢地面对挑战，坚持不懈地追求，为国家和人类的进步创造出令人瞩目的成就。祝福你们！

责任编委 张军平 贾熹滨 明悦

刘允才

1990 年于伊利诺伊香槟分校 ECE 系获得博士学位，教育部长江学者特聘教授，曾任上海交通大学特聘教授、博士生导师。长期从事计算机视觉理论研究。主要研究方向为计算机视觉、智能交通系统、机器人外科手术。培养博士、硕士研究生近百人，其学生获全国优秀博士论文提名奖、上海市优秀博士论文奖。实验室多名毕业生目前在大学和科研院所担任教授、研究员和副教授。刘教授曾荣获 2012 年度国家技术发明二等奖，2002 年度教育部提名国家科学技术奖自然科学奖；上海市白玉兰奖，上海市 2003、2004、2005、2009、2011 年度科技进步奖；在包括 IEEE Trans. PAMI、IEEE Trans. Image Processing、CVPR、ICCV 等国际权威学术期刊与会议上发表论文 400 余篇、获国家发明专利授权 80 余项。刘教授获 2023 年度 CCF-CV 终身学术贡献学者荣誉。

COMPUTER VISION NEWSLETTER

04 2023

总第 38 期

计算机视觉专委会简报

CCF 计算机视觉专委会