主办 CCF计算机视觉专业委员会

COMPUTER VISION NEWSLETTER

CCF 计算机视觉专委会简报

02 2026

总第 48 期

CCF 计算机视觉专委会

COMPUTER VISION NEWSLETTER

计算机视觉专委会简报

2026 年第 02 期

总第 48 期

主　办　CCF 计算机视觉专业委员会

编委会

CONTENTS
简报目录

| 专 委 动 态

04 CCF-CV 走进高校系列报告会
06 CCF-CV 2026 年执行委员增选申请开始
07 CCF-CV 专委会主任陈熙霖获第四届全国创新争先奖状

| 科 技 前 沿

08 大模型架构的下半场
14 原生三维生成先验驱动的组合式三维生成
27 ICLR 2026

| 委 员 风 采

31 武汉大学叶茫教授访谈
34 委员好消息

| 学 术 资 源

36 基于扩散模型的风格迁移方法及开源代码
39 道路交通场景风格迁移数据集
42 好文推荐

| 海 外 学 者

45 征文通知

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 151 期 南京大学

2026 年 4 月 24 日上午，由中国计算机学会计算机视觉专委会 (CCF-CV) 主办、南京大学人工智能学院承办的第 151 期 CCF-CV 走进高校系列报告会，在南京大学仙林校区人工智能学院 A101 报告厅顺利举行。本次报告会邀请了六位领域内知名专家学者作专题报告，分别是：浙江科技大学**梁荣华**教授、中国科学院自动化所**李国齐**研究员、南京理工大学**李泽超**教授、哈尔滨工业大学**左旺孟**教授、北京交通大学**魏云超**教授以及北京大学**余肇飞**研究员。与会专家围绕边缘智能、类脑智能、多模态感知、具身智能及视觉理解等前沿方向，分享了最新研究成果与未来趋势展望。本次活动由南京大学**叶翰嘉**副教授担任执行主席，报告环节由南京大学**张利军**教授、**叶翰嘉**副教授和**赵鹏**副教授依次主持。

南京大学人工智能学院院长**黎铭**教授致欢迎辞。黎铭教授首先向与会专家学者致以诚挚的欢迎与感谢，并感谢 CCF-CV 专委会对学院的信任与支持。黎铭教授简要介绍了学院的基本情况和发展历程，并预祝本次活动圆满成功。

第 152 期 苏州科技大学

2026 年 4 月 26 日，计算机视觉感知与理解前沿报告会在苏州科技大学成功举办。本次报告会聚焦计算机视觉领域的前沿问题与关键技术，围绕图像语义理解、多模态视频感知、基础模型驱动的语义分割、视觉—触觉持续学习、视觉目标分割以及第一视角多模态视频理解等方向展开深入交流。活动特邀同济大学**徐行**教授、南京邮电大学**周全**教授、中国矿业大学**姚睿**教授、西北工业大学**刘婷**副教授、浙大城市学院**魏李娜**副教授、山东大学**徐明珠**助理研究员等 6 位专家作专题报告，为与会师生带来了一场兼具理论深度、技术前沿与应用视野的学术盛会。

本次活动由苏州科技大学**徐峰磊**副教授与胡伏原教授担任执行主席。报告会由苏州科技大学徐峰磊副教授主持，他对与会专家学者表示欢迎，并对会议议程、报告安排及交流环节进行了介绍。两位执行主席围绕“计算机视觉感知与理解前沿技术与应用”主题精心组织筹备，保障了报告会各环节有序推进，为本次学术交流活动的顺利举办奠定了坚实基础。中国图象图形学报

“图图 Seminar”直播平台对报告会进行全程直播，进一步拓展了学术交流的覆盖范围，使更多相关领域师生和科研工作者能够共同参与前沿学术讨论。

第 153 期 浙江师范大学

2026 年 5 月 15 日，由中国计算机学会计算机视觉专委会（CCF-CV）主办、浙江师范大学计算机学院和 CCF 金华分部联合组织的“CCF-CV 走进高校系列报告会”在浙江师范大学成功举办。本次活动以“视觉多模态前沿探索”为主题，吸引了百余位校内外师生共同参与，共同探讨计算机视觉领域的最新发展与应用。

浙江师范大学党委书记**蒋云良**教授出席开幕式并致辞。他代表学校向莅临现场的各位专家学者致以诚挚问候与由衷感谢。他介绍了我校办学基本情况、计算机学科的发展历程与近年来建设成果。他指出，人工智能浪潮推动计算机视觉从感知向理解、生成、决策全面演进，多模态大模型进一步拓展了技术融合与智能系统发展的新路径。本次报告会聚焦视觉多模态前沿，旨在搭建学术与人才培养的桥梁，推动学科高质量发展。

中国计算机学会计算机视觉专委会主任、中国科学院计算技术研究所所长**陈熙霖**教授出席开幕式并致辞，他首先对各位拨冗出席的专家同仁表示诚挚感谢，对周密细致的会议筹备工作给予充分肯定。陈熙霖表示，此次会议是建立联系的桥梁，热忱希望现场的青年科研学子踊跃投身学术行列，汇聚力量融入这份朝气蓬勃的学术阵营，携手助力领域长远发展。

报告会开幕式由浙江师范大学计算机科学与技术学院院长、CCF 金华分部秘书长**郑忠龙**教授主持。会议邀请到了重庆邮电大学**李伟生**教授、南开大学**杨巨峰**教授、北京理工大学**付莹**教授、上海交通大学**刘伟**教授、南京理工大学**潘金山**教授、杭州电子科技大学**余宙**教授等计算机视觉领域的专家学者作特邀报告。

责任编委 毋立芳、朱安娜

中国计算机学会计算机视觉专委会（CCF-CV）2026 年执行委员增选申请开始啦！

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、计算机视觉前沿讲习班，与中国自动化学会模式识别与机器智能专委会、中国图像图形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），牵头承担第四届全国计算机名词委人工智能分委会名词起草与审定工作，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（https://space.bilibili.com/611909696）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 13 年以来，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 7 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 2026 年计算机视觉专委会的执行委员增选工作。

一、申请时间

2026 年 5 月 15 日— 2026 年 8 月 5 日。

二、申请流程

填写申请表（点击最下方阅读原文可直接下载），发送给秘书处（ccfcv@139.com），主题“2026 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（https://ccfcv.ccf.org.cn/ccfcv/zzjg/）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）。

三、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。

四、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“2026 年中国模式识别与计算机视觉大会”（http://www.prcv.cn）期间（2026 年 8 月 22 日-25 日）举行的专委工作年会上投票确定（申请者届时必须“注册参会”）。

五、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。

欢迎计算机视觉及相关领域的同仁加入！

责任编委 马伟

CCF-CV 专委会主任陈熙霖获第四届全国创新争先奖状

2026 年 5 月 30 日，第十个全国科技工作者日主场活动暨第四届全国创新争先奖表彰大会在京举行。活动现场宣读了第四届全国创新争先奖表彰决定，为获奖团队及个人举行了颁奖仪式。CCF-CV 专委会主任、中国科学院计算技术研究所**陈熙霖**研究员荣获第四届全国创新争先奖状。

全国创新争先奖由人力资源社会保障部、中国科协、科技部、国务院国资委联合设立，是国家表彰科技创新工作者的重要国家级荣誉，聚焦嘉奖在科技创新、科学普及领域做出实绩的优秀个人与集体。全国创新争先奖是国家科技奖励体系的重要组成部分和补充，是国家科技奖项与重大人才计划的有机衔接，是仅次于国家最高科技奖的科技人才大奖。

陈熙霖研究员现任中国计算机学会计算机视觉专委会（CCF-CV）主任，同时为 ACM/CCF/IEEE/IAPR Fellow，现任中国科学院计算技术研究所所长、党委书记，历任中国科学院智能信息处理重点实验室主任、中国科学院计算技术研究所副所长、中国科学院国际合作局局长。陈熙霖研究员的主要研究领域为计算机视觉、模式识别、多媒体技术以及多模式人机接口。在国内外重要刊物和会议上发表论文 400 多篇，以第一完成人获国家自然科学二等奖 1 项，作为主要完成人获国家科技进步二等奖。

责任编委 黄岩

专题综述

大模型架构的下半场

华中科技大学 王兴刚 朱良辉

太长不看：我们花了十年去扩展层内的计算能力，却忘了扩展层间的通信能力。这件事亟需被改变。

过去十年，深度学习领域取得进展的方式出奇地一致：什么都往大了整。更多参数、更多数据、更长上下文；而且确实管用：loss 在降，能力在涨，scaling law（扩展定律）精确地告诉我们还需要投入多少。

但扩展的方向不同，差异也是巨大的。序列长度的扩展需要真正的创新，也确实催生了一整套机制研究和系统工程。数据的扩展则直截了当：数据越多，loss 越低。让模型变得更宽、更深，这看起来也和数据的扩展一样简单。但宽度和深度真的在同等地发挥作用吗？

并非如此。深度在数量上增长了，但在质量上却没有。层与层之间的通信机制几乎没有变化。接下来我们将解释这一点为什么重要，这不仅关乎网络的深度本身，更关乎我们设计神经网络架构时的一个集体盲区。

一、大模型架构上半场

要看清上半场做对了什么，就看看什么被成功地扩展了，以及是怎么做到的。

先看序列长度。早期 Transformer 只能处理几百个 token。要达到 128K+，需要多个方向上的持续创新：新的注意力模式（稀疏、线性、混合）、系统工程（FlashAttention）、位置编码的进步（RoPE scaling）。研究者和工程师们共同建造了一整个生态，持续改进 token 之间的通信方式。而回报颇丰，我们不止能够处理极其长的文档，还为 OpenAI-o1 和 DeepSeek-R1 的长链推理奠定了坚实的基础。这就是当我们认真投资于“信息在序列维度上的流动方式时”，所收获的斐然成果。

参数和数据的扩展是最符合人类直觉的部分。从深度学习的最早期开始，每本教科书都在教我们同一套配

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

GPT-3, 175B Params, over 300B Tokens

DeepSeekV3, 671B Params, 14.8T Tokens

GLM-5, 744B Params, 28.5T Tokens

Qwen3, 1T Params, 36T Tokens

Kimi K2, 1T Params, 15.5T Tokens

图 1　现代 LLM 中的参数与数据规模迅速增长

方：更多数据、更宽的层、更深的网络，自然带来更好的表征。从 GPT-2 的 15 亿参数到如今的数千亿，这套配方一直管用。我们不需要新机制，只需要持续拓展这些被验证了的方向。

只不过，对网络而言，更宽和更深往往并不是一回事。宽度的扩展是自然而然的：现代 GPU 天生擅长处理更宽的矩阵乘法，注意力机制的演进越来越高效，这使得更宽的网络可以无缝接入现有架构。

而深度则是另一个故事。模型确实变深了：我们将模型加到 32 层、64 层、甚至 100 层以上。但层间通信的机制本质上还是 ResNet 在 2015 年引入的深度残差“x + F(x)”。自它诞生以来，围绕它有过不少改良（归一化位置、残差缩放、跨层连接），但没有任何改良真正取代过那个深度残差中“+”的决定性地位。

残差连接可以说是深度学习中最重要的基石。没有它，就没有 100 层的 Transformer，没有现代 LLM，没有 scaling law。但基础性方案有一个特点：它们有时会变得太过隐形，以至于没人再去质疑它到底是最优解，还是仅仅是我们探索出的第一个能用的方案。

打个比方，想象一个有特殊规则的传话游戏。在标准版本里，第 1 个人对第 2 个人耳语，第 2 个人再对第 3 个人耳语。到第 18 个人的时候，消息已经面目全非了。这就是没有残差连接的深层网络：每一层只能看到上一层的输出。

残差连接修复了这个问题：每个人在传达自己的理解的同时，也把之前积累的原始信息原封不动地往下传。第 3 个人既能听到第 2 个人的新解读，也能听到之前的所有内容。原始信号始终被保留，它成为了不断壮大的合唱中的一个声部。

但到了第 152 个人，你同时在听 152 个声音：原始信息加上 151 层叠加上去的内容，全部混在一句耳语里。理论上，前面那些人的声音依然存在，但它们已经被淹没了。如果第 152 个人需要知道第 3 个人具体说了什么，他得费力地从这首宏大的合唱声中把它挑出来。通常而言，第 152 个人是做不到这一点的。

这就是信息稀释。每一层都面临两难：倘若该层贡献新信息就可能会掩盖之前的内容，但保守不动则能保留之前层传过来的已有信息。这种状况下，很多层学会了保守不动，它们几乎不往残差流里写入任何东西。这样的深度网络在纸面上很深，实际上却很浅。我们堆了 152 层，但其中很多层却只学会了保持沉默。

这里的瓶颈不在于 152 层网络所需求的算力，而在于信息穿过这些层的通信能力。CPU 的发展在几十年前就撞过同样的墙：处理器越来越快，直到内存带宽跟不上了，逼得整个行业转向缓存和通信。组织管理也一样：一群聪明人所能发挥出的创造力，也受限于他们之间的沟通、组织方式。深度学习正在经历自己的版本：十年来不断增强每一层的能力，而层与层之间的通道始终是 2015 年那条单车道公路。

那么，有没有更好的机制？

二、大模型架构配方

在我之前已经有很多研究者注意到了深度学习的瓶颈。多年来，修补方案越来越巧妙：获评 CVPR best paper 的 DenseNet 保留了每一层的输出，但代价是平方级的开销。使用可学习加权的方案 DenseFormer、LIMe 降低了成本，但训练完成后权重就固定了，每个 token、每套上下文都用同样的权重。字节跳动的 Hyper-Connections 和 DeepSeek 的 mHC 另辟蹊径，它们把管道拓宽到 N 个通道，层间用混合矩阵连接，这相当于信息高速公路上同时多了好几条车道。但坏消息是，信息仍然在逐层流动，第 152 层没有办法直接回溯到第 3 层。

彩云公司的 MUDDFormer 让混合每层输出这件事变成动态的，它会根据每个 token 的表征来生成权重。这在根本方向上是对的：从每一层汲取多少信息本就应该取决于你正在处理的内容。但同样有个坏消息，第 152 层在决定从第 3 层汲取多少时，只依赖第 152 层本身的状态，它并不知道第 3 层实际包含了什么。它是在预测哪些层有用，而不是在查看。

以上的每一步都修复了一个真实存在的缺陷，但却鲜有哪一个方法质疑过深度残差的框架本身。

不难发现这些方法都有一个共同点。从 DenseNet

到 Hyper-Connections，每个方法都在回答同一个隐含的问题：“怎样才能更好地混合各层的输出？”更好的系数，更多的通道，自适应的权重。但自始至终都是混合，自始至终都是累加。ELMo 早就表明，不同的层编码的是截然不同的信息：浅层编码句法，深层编码语义。所有人得出的结论都是“学习更好的混合权重用来平衡句法和语义”。但还有一条被主流忽视的道路：如果不同层持有不同信息，也许每一层应该能够根据内容而非位置，从持有所需信息的那一层直接检索。

这就是范畴谬误：把层间通信当作累加（用学习到的或生成的系数来组合信号）而非检索（通过基于内容的匹配来选择信息）。在累加框架下，即使是动态方法也只从当前层的状态生成混合权重，而不去查看信息的来源层实际包含了什么。在检索框架下，Query（查询）编码的是“我需要什么”，Key（键）编码的是“我有什么”，而它们之间的运算决定了相关性。Query 和 Key 双方都应该有发言权。

合优化策略，实现几何与外观质量的协同提升。

回到传话游戏。之前所有的方法都在试图产生一个更清晰的合唱：更好的发音、更多的中继通道、自适应的音量。没有一个质疑过这个根本约束：所有声音必须累加成一个声音吗？也没有人问过：咱是否可以直接走回去，跟之前的任何一个人当面对话呢？

我觉得这种范畴谬误在架构设计中无处不在。当某个东西足够好用的时候，你不会去质疑它的概念框架，而是只会在框架内改进。经历了多年越来越巧妙的修补之后，我才看清：深度维度的残差连接需要的不是更好的系数，而是被一种根本不同的操作所替代：一种在序列维度上已经成功解决了同样问题的操作。

三、大模型架构下半场

一旦我们把层间的通信理解为检索而非累加，一个很自然的答案就是在深度维度上引入注意力机制。很多团队都独立地收敛到了这个想法：谷歌提出的 DCA、华为的 MRLA、Hessian.AI 的 Dreamer、Kimi 的 AttnRes，大家都尝试在层间应用注意力机制。这种独立趋同本身就是一个信号：方向走对了！

但找对方向和做出成品是两回事。说实话，我第一次用 Pytorch 实现运行深度注意力的时候，前向和反向传播共计耗时达到了 44,924 ms。44 秒啊！朋友们！这个时间都够我喝完一瓶 500 毫升的冰红茶了！也就是说，在深度维度上应用注意力机制的想法本身没问题，但工程现实却残酷到了极点。现代 GPU 为大规模的矩阵乘法做了大量优化，却不擅长数千个跨深度的极小规模的注意力操作。深度注意力作为一个计算量不大的算法，跑起来却可能慢得要命。

至此，之前的方法都陷入了两难：要么简化深度注意力来换速度，这种方式丢掉了完整的选择性检索这一核心价值；要么保持完整的表达能力，但运算代价变得不可接受。我们找到了一条出路：不是简化算法，而是重新组织参与计算的数据布局，从而适配 GPU 硬件。如图 2，Flash Depth Attention（https://github.com/hustvl/MoDA）让具备完整表达能力的深度检索快到可以参与实际训练。

有了高效的深度检索之后，我们注意到网络的主干流水线变成了：深度注意力→序列注意力→深度注意力→ FFN（前馈网络）。这三个连续的注意力操作作用于不同的 Key（键，缩写作 K） 和 Value（值，缩写作

```
================================================================================
Comprehensive Benchmark [fwd_bwd]  B=1 H=2 K=64 V=64 L=64 dtype=torch.float16
================================================================================
    T_kv     G        T_q |         DepthRef |           FDAv4 |   DepthRef/FDAv4
--------------------------------------------------------------------------------
     512     8       4096 |      4341.213 ms |        0.686 ms |      6328.299x
    1024     8       8192 |      9560.738 ms |        0.674 ms |     14185.071x
    4096     8      32768 |     44924.228 ms |        0.948 ms |     47388.426x

     512    16       8192 |      9199.153 ms |        0.639 ms |     14396.171x
    1024    16      16384 |     19010.542 ms |        0.869 ms |     21876.343x
```

图 2　Pytorch 实现的深度注意力（DepthRef）很慢；Flash Depth Attention（FDA）很快

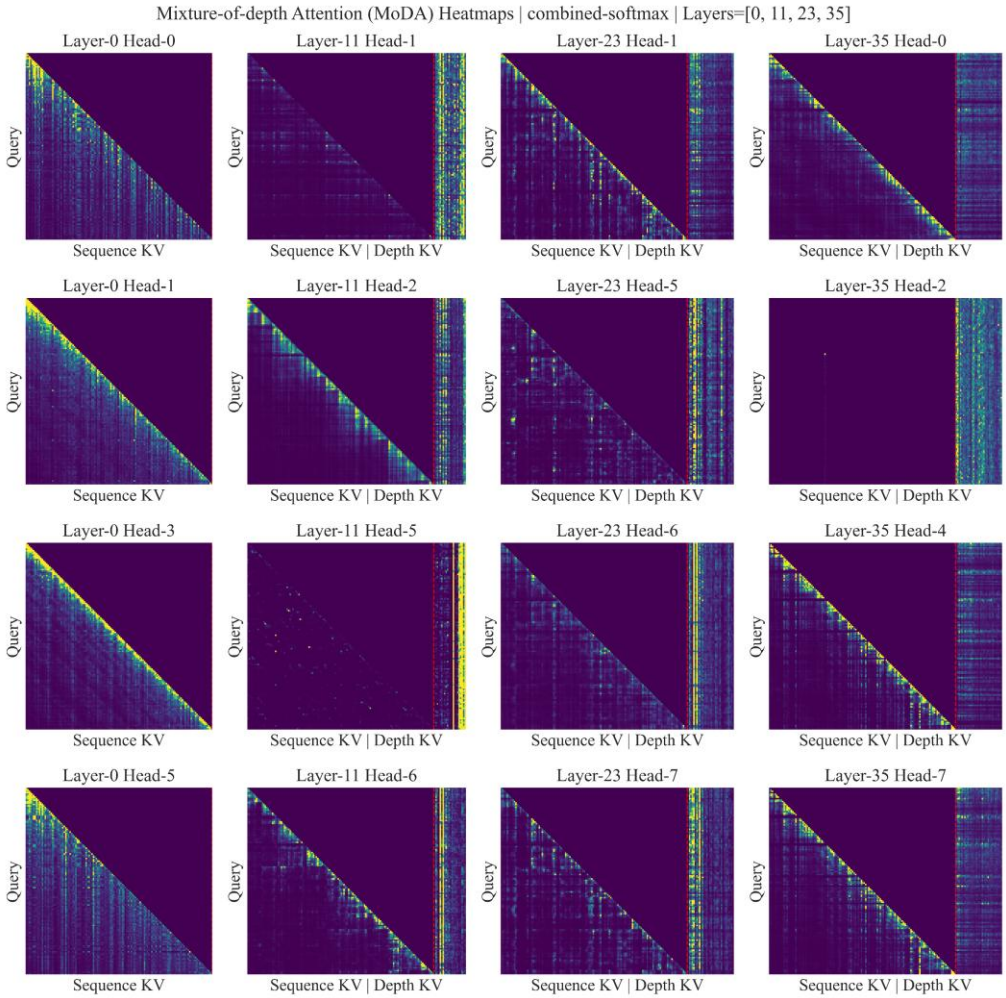


图 3　左侧区域是序列 KV，右侧区域是深度 KV。颜色越黄，注意力越强。

V），却共享着近乎相同的 Query（查询）。一个很自然的做法就是把它们融合。我们提出了混合深度注意力（Mixture-of-depths Attention, MoDA）将深度检索和序列检索合并到一个统一的 softmax 中。如图 3，每个注意力头同时关注当前层的序列 KV 对（键值对）和所有前序层的深度 KV 对（键值对）。在同一个 softmax 下，模型可以自由决定何时关注序列中的其他 token，何时跨层检索自身的历史信息。通过一次操作，我们完成了两个维度的检索。

回到传话游戏。在残差连接的版本里，第 152 个人费力地从累加的合唱中辨认第 3 个人的声音。有了深度检索，第 152 个人拍拍第 3 个人的肩膀直接问："你刚才说了什么？"没有中间人，没有累积的噪音。可视化的实验结果也印证了这个类比所预测的现象：当模型获得了通过深度 KV 从特定层进行选择性检索的能力时，它会持续且主动地使用这种能力。之前困扰模型架构研究员们的 Attention Sink（注意力沉没）现象，即模型把概率质量堆积在少数固定 token 上的行为，也随之减弱。这就是当我们尝试发展层之间而非仅仅层之内的信息流动时，所取得的有趣成果。

大模型架构的上半场是关于扩展组件的。我们扩展出更长的序列，更多的数据，更大的模型。这个阶段最关键的问题是"我们怎么把一切都做大？"。在上半场，这是正确且关键的问题，它把我们从 GPT-2 带到了 GPT-4。下半场是关于扩展通信的。新的问题是："组件之间的通信质量如何？"

深度是最明显的例子，因为现有方案（累加）和可能的方案（选择性检索）之间的差距是巨大的。而我相信这个原则是可以推广的。凡是神经网络使用静态的、与数据无关的通道来传递信息的地方，包括层与层之间、

模态与模态之间、时间步与时间步之间等等，很可能都会有一个检索机制等着替代那个累加操作。

我们花了十年掌握 token 之间如何对话，现在是时候掌握层与层之间如何对话了。而最终，我们将掌握神经网络中每个组件如何与其他任意组件对话。深度残差的“+”带我们跑过了一段极为精彩的旅程，但现在，是时候升级这座阶梯了。

欢迎来到大模型架构的下半场。

本论文内容主要来自于华中科技大学 (HUST) 电子信息与通信学院视觉实验室 (Vision Lab) 的研究工作。HUST Vision Lab 研究主要集中在计算机视觉和深度学习领域，尤其关注以下方向：多模态基础模型，视觉表征学习，目标检测、分割与跟踪，端到端自动驾驶，和新型神经网络架构。

HUST Vision Lab 突破视觉智能边界的代表性工作：CCNet (TPAMI 2020，4300+引用，1.5K Star)、Mask Scoring R-CNN (CVPR 2019，1400+引用，1.9K Star)、FairMOT（IJCV 2021，2200+引用，4.2K Star）、ByteTrack（ECCV 2022，3400+引用，6.2K Star）、EVA（CVPR 2023，1100+引用，2.7K Star）、MapTR（ICLR 2023，400+引用，1.5K Star）、Vectorized Autonomous Driving（ICCV 2023，600+引用，1.3K Star)、DiffusionDrive (CVPR 2025，200+引用，1.3K Star）、Vision Mamba（ICML 2024，3100+引用，3.8K Star）、4D Gaussian Splatting (4DGS)（CVPR 2024，1400+引用，3.5K Star）、YOLOS（NeurIPS 2021，500+引用，900+ Star)、YOLO-World (CVPR 2024，1000+引用，6.3K Star)，及 LightningDiT & VA-VAE（CVPR 2025，200+引用，1.4K Star）。

责任编委 王金甲

参考文献

[1] https://arxiv.org/abs/2603.15619.
[2] https://github.com/hustvl/MoDA.
[3] https://github.com/hustvl.

王兴刚

华中科技大学电子信息与通信学院教授，博士生导师，国家“万人计划”青年拔尖人才，现任 Image and Vision Computing 期刊（Elsevier, IF 4.2）主编，分别于 2009 年和 2014 年在华中科技大学获得了本科和博士学位，期间在美国 UCLA 和天普大学访问研究。主要研究方向为计算机视觉、深度学习、多模态基础模型、视觉表征学习、目标检测分割跟踪、自动驾驶等，发表包括 IEEE TPAMI、IJCV、CVPR、ICCV、NeurIPS 等顶级期刊会议在内的论文 100 余篇，谷歌学术引用 5 万余次，其中一作/通讯 1000+引用论文 6 篇，入选 Elsevier 中国高被引学者、斯坦福前 2%科学家。研究成果应用于地平线征程系列芯片、华为海思昇腾芯片、Deepmind 蛋白质结构预测模型 AlphaFold；获得了华为、vivo 等公司优秀合作项目奖。担任 CVPR、ICCV、NeurIPS 等顶级会议领域主席，IEEE TPAMI、Machine Vision and Application 等期刊编委。入选了中国科协青年人才托举工程，获湖北青年五四奖章、CSIG 青年科学家奖、吴文俊人工智能优秀青年奖、CVMJ 2021 年度最佳论文奖、MIR 期刊年度最高引用论文奖、湖北省自然科学二等奖等。
Email: xgwang@hust.edu.cn

朱良辉

华中科技大学电子信息与通信学院博士研究生，中共党员，师从王兴刚教授和刘文予教授，入选首届字节跳动校企联培博士。主要研究方向为基础模型架构、弱监督视觉感知和视觉语言大模型等，已发表 ICML、ICLR、CVPR、ICCV、AAAI、ACM MM、IJCV、TIP 等顶级期刊会议论文 11 篇，其中一作 8 篇，谷歌学术引用 4500 余次，开源代码获 GitHub Stars 4600+。代表作 Vision Mamba 提出首个基于 Mamba 的线性复杂度通用视觉骨干网络，入选 ICML 2024 最具影响力论文和 ArXiv 2024 计算机视觉方向最具影响力论文。主持国家自然科学基金青年学生基础研究项目和中国科协青年人才托举工程博士生专项，获博士研究生国家奖学金、华中科技大学“学术新星”、ICLR Notable Reviewer 等荣誉。

Email: lhzh@hust.edu.cn

专题综述

原生三维生成先验驱动的组合式三维生成

樊红兴 盛律
北京航空航天大学

本文主要介绍北京航空航天大学软件学院盛律团队在三维可控内容创作的最新研究成果，包括发表在CVPR 2025 的 MIDI[1]、CVPR 2026 的 LaviGen[2]、ACM SIGGRAPH 2026 Journal Track的SegviGen[3]和3DV 2026口头报告的VoxHammer[4]。本文以三维资产的可控生成为主线，重点介绍部件解析、局部编辑和场景构建三个环节，展示原生三维生成先验在组合式三维生成中的作用，并共同指向可生成、可理解、可编辑和可组合的三维场景构建。

一、三维生成中的生成式先验

与二维图像和视频生成相比，三维生成不仅要求逼真的视觉外观，还需要满足几何结构完整、跨视角一致、空间关系合理以及后续交互编辑等要求。随着应用需求的发展，三维生成的目标也正在从单体资产生成逐渐扩展到组合式场景构建和可控三维内容创作。

当前，大规模高质量三维数据仍相对稀缺，二维图像和视频生成模型因而成为三维生成的重要先验来源。以 DreamFusion[5]为代表的方法利用二维文生图扩散模型提供的得分蒸馏采样（Score Distillation Sampling，SDS）技术优化三维表示，展示了二维生成式视觉先验辅助三维生成的潜力。然而，这类方法通常需要针对每个样本进行逐场景优化，计算成本较高。相比而言，Zero-1-to-3[6]和 SyncDreamer[7]等方法通过单图到多视图图像生成，并结合三维重建模型恢复三维内容，将二维生成先验以更高效的前馈方式引入三维生成流程。这类方法显著提升了三维生成效率，但仍然面临跨视角视图不一致、细粒度控制薄弱和几何细节缺失等问题。

围绕这些问题，团队前序工作从不同角度为二维生成先验加入更强的几何约束。例如，EpiDiff[8]在多视图扩散过程中引入局部对极约束，提升多视图一致性并加速多视图注意力计算。Ouroboros3D[9]递归地将三维重建结果反馈给多视图扩散过程，使每一步去噪都能够获得显式几何指导。MV-Adapter[10]直接利用适配器结构通用学习多视图扩散能力，能够适配不同二维扩散模型和各类型风格化变体。AnimaX[11]则进一步利用多视图“视频-姿态”扩散模型，为可动三维对象生成提供满足物理运动约束的视觉先验。这些工作共同说明，二维图像和视频生成先验可以有效缓解三维/四维生成中的数据稀缺和几何约束不足问题。

尽管图像和视频生成先验显著推动了三维生成，但这类方法的核心模块通常仍依赖多视图图像、视频序列或文本化布局表示等外部表示，难以原生表达三维空间中的空间结构和物理约束等语义和几何要素。随着物体级的原生三维生成模型 TRELLIS[12]和 TRELLIS.2[13]的不断推出，利用三维生成先验直接进行语义和几何描述的能力不断提升。三维生成研究已不再满足于实现几何结构完整、视觉质量较高的单体资产生成，而是进一步关注三维内容的结构理解、局部交互编辑以及由部件到对象、由对象到场景的组合式生成与构建。

二、三维部件解析、局部编辑与场景构建

SegviGen、VoxHammer 和 LaviGen 展示了原生三维生成先验在已知三维资产的情况下如何进行部件解析、局部编辑和场景布局生成。MIDI 进一步探索了仅有图像的条件下如何进行组合式三维场景生成。

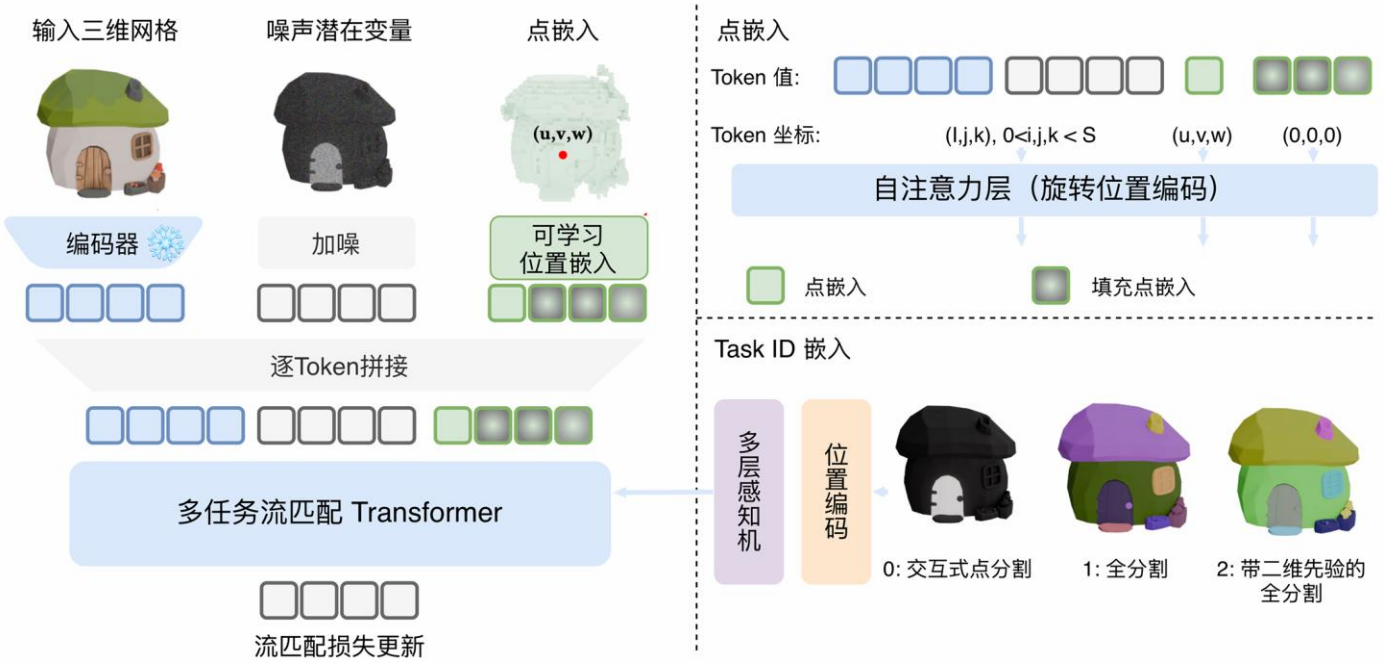


图 1：SegviGen 的统一部件解析流程。

1．部件解析

一个三维资产（比如说“椅子”）通常由多个具有语义和几何意义的部件组成（例如椅背、椅腿和扶手）。清晰的部件结构能够支撑后续丰富的选取、编辑、动画绑定和交互操作等需求，是三维资产实现结构化理解与精细化控制的基础。

传统三维部件分割通常依赖带有固定类别和部件标签的数据集，通过监督学习训练点云或网格分割模型。然而，三维部件标注成本高，不同数据集之间的部件层级、粒度和边界定义也往往不一致，导致模型泛化能力受限。另一类方法尝试利用二维基础模型，将三维对象从多个视角渲染成二维图像，再使用二维分割模型得到掩码，最后投影回三维空间并进行融合。这类方法能够借助强大的二维分割先验，将多视角二维分割结果投影并融合到三维空间中，但也容易出现跨视角不一致、边界模糊、视角覆盖不足和投影误差等问题。

相比之下，SegviGen 将部件分割转化为原生三维潜空间中的条件着色任务。模型不再直接预测离散的部件类别标签，而是在三维表示中为不同部件预测具有指示意义的颜色，再根据颜色得到部件分割结果。因为利用了预训练三维生成模型强大的结构与纹理生成先验，所以能获得边界更清晰、泛化性更强的部件分割结果。

通过这种统一的条件着色框架，SegviGen 能够同时支持交互式部件分割、全量部件分割和二维图引导的全量部件分割。

（1）交互式部件分割：用户点击目标部件后，模型将目标部件预测为前景颜色，其余区域预测为背景颜色；

（2）全量部件分割：模型为对象内部不同部件分配不同颜色；

（3）二维分割图引导的全量部件分割：模型还可以接收单视角二维分割图作为条件，将用户指定的部件划分方式迁移到完整三维对象上。

在模型结构上，SegviGen 以预训练原生三维生成模型 TRELLIS.2[13]为基础。首先，输入三维资产会被编码到结构化三维潜变量空间中，其中有效 O-Voxels 作为稀疏的三维体素单元，保存对象的几何和纹理特征，并为后续生成过程提供与几何对齐的三维支撑。随后，模型根据不同任务构造对应的部件着色目标，并将该着色目标编码到同一三维潜空间中。训练过程中，模型不是直接一步输出最终颜色，而是对部件颜色潜变量加入噪声，并通过条件流匹配（conditional flow matching）[14]学习从带噪颜色表示（noisy color latent）中恢复清晰部件颜色的去噪方向。模型主体采用多任务流匹配

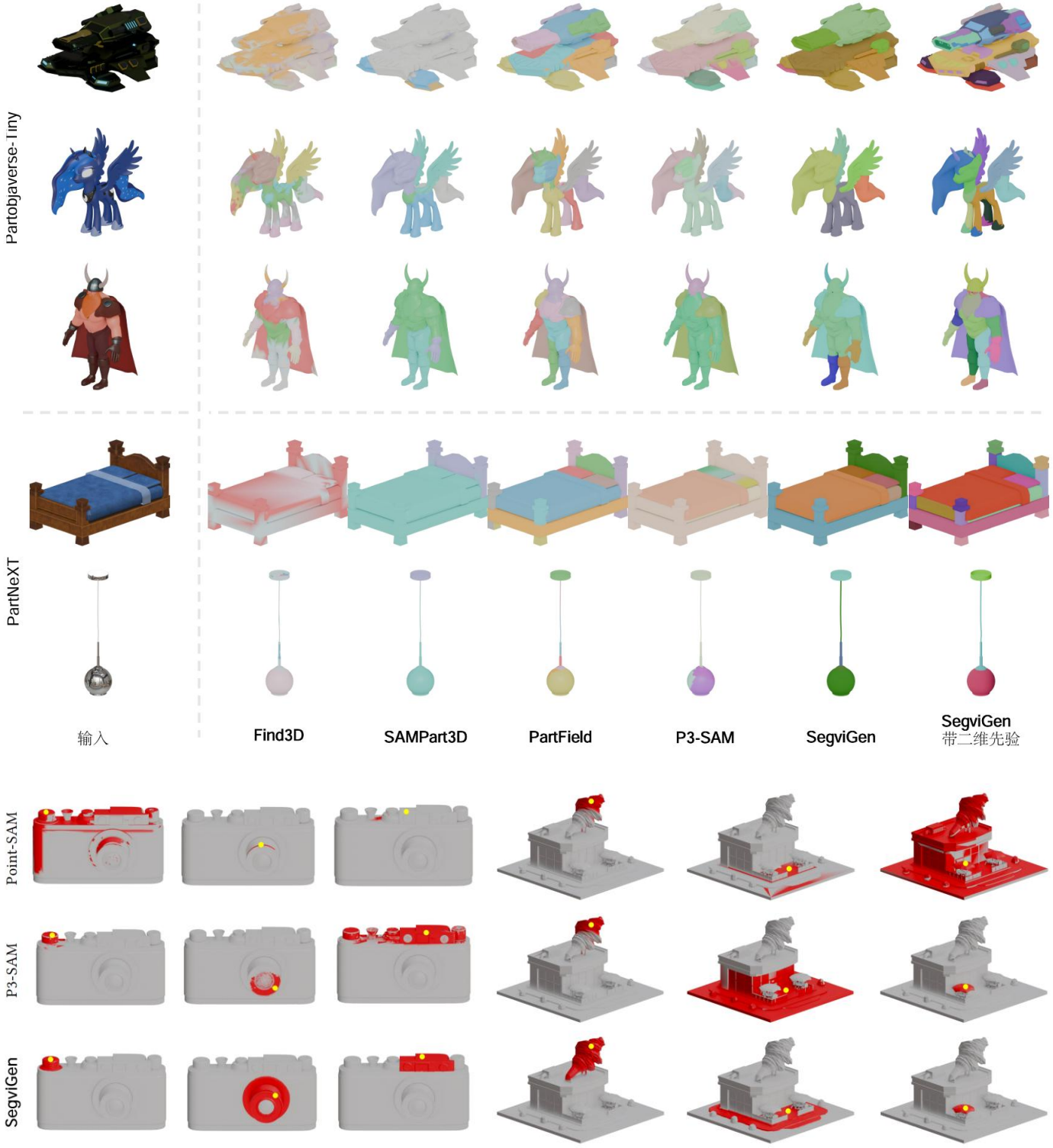


图 2：上：全量部件分割结果对比；下：交互式部件分割结果对比。

Transformer（Multi-Task Flow Transformer），同时接收输入三维资产的几何潜变量（geometry latent）、带噪的部件颜色潜变量以及任务条件，最终在生成过程中得到每个有效体素上的部件指示颜色。SegviGen 的整体流程如图 1 所示。

SegviGen 的条件设计使其能够统一处理多种部件解析任务。在交互式部件分割中，用户点击会被编码为点条件，用于指示目标部件的位置；在全量部件分割中，模型不依赖用户点击，而是根据三维资产本身生成完整部件划分；在二维图引导的全量分割中，单视角二维分割图会通过图像编码器转化为条件信息，并通过交叉注

表 1：PartObjaverse-Tiny 和 PartNeXT 上的交互式部件分割定量结果。

方法	PartObjaverse-Tiny					PartNeXT				
	IoU@1	IoU@3	IoU@5	IoU@7	IoU@10	IoU@1	IoU@3	IoU@5	IoU@7	IoU@10
Point-SAM[15]	24.87	48.99	59.67	64.33	67.99	23.90	47.50	56.71	61.23	65.04
P3-SAM[16]	33.04	50.57	53.78	54.74	55.51	35.61	51.26	52.03	52.61	53.81
SegviGen	**42.49**	**61.14**	**67.53**	**71.50**	**75.02**	**54.86**	**71.15**	**78.11**	**79.96**	**82.73**

意力注入三维生成过程。同时，模型还引入任务类型嵌入，用于区分交互式分割、全量分割和二维图引导分割三种任务，使同一个网络能够在统一框架下完成不同形式的部件解析。

表 2：全量部件分割定量结果（IoU）。其中 w. 2D Map 表示使用二维分割图引导。

Method	PartObjaverse-Tiny	PartNexT
Find3D[17]	15.62	19.04
SAMPart3D[18]	59.05	29.62
PartField[19]	51.72	41.50
P3-SAM[16]	45.36	31.94
SegviGen	**50.64**	**55.40**
SegviGen (w. 2D Map)	**62.98**	**71.53**

模型最终生成的是有效 O-Voxels 上的部件指示颜色。由于三维生成模型解码得到的网格可能与原始输入模型在拓扑结构和局部三角面划分上存在差异，SegviGen 并不直接使用解码网格作为最终分割结果，而是将预测得到的体素颜色回传到原始输入网格上。具体来说，每个顶点继承最近有效体素的颜色，每个面片再根据顶点颜色投票得到部件标签，并通过轻量级网格平滑去除边界附近的孤立噪声。这样，方法能够在保持原始三维模型结构不变的同时，得到稳定的网格级部件分割结果。

实验结果显示，SegviGen 在数据效率和分割质量上均表现突出。该方法仅使用约 0.32%的训练数据，即在交互式部件分割的单击 IoU@1 指标上取得约 40%的相对提升，并在无二维图引导的全量分割平均 IoU 上取得约 15%的提升。

如图 2（上）所示，在 PartObjaverse-Tiny 和 PartNeXT 等不同类别对象上，SegviGen 能够生成边界更清晰、结构更完整的全量部件分割的结果。与方法 SAMPart3D[18]、PartField[19]和 P3-SAM[16]等相比，SegviGen 对复杂角色、家具和细长结构具有更好的跨类别泛化能力；在引入二维分割图引导后，模型还能够进一步结合外部二维条件调整部件划分方式。图 2（下）还给出了交互式分割的结果，SegviGen 能指明更符合几何和语义准确性的部件结构，较少出现部件边界不清晰或混淆的情况。

定量结果如表 1 和表 2 所示：在交互式部件分割中，SegviGen 在 PartObjaverse-Tiny 和 PartNeXT 的 IoU@1、IoU@3、IoU@5、IoU@7 和 IoU@10 上均优于 Point-SAM 和 P3-SAM，其中在 PartNeXT 上 IoU@10 达到 82.73；在全量分割中，SegviGen 在 PartObjaverse-Tiny 和 PartNeXT 上分别达到 50.64 和 55.40，引入二维分割图引导后进一步提升至 62.98 和 71.53，分别提高 12.34 和 16.13 个百分点。更重要的是，SegviGen 的分割结果可以直接作为后续局部编辑的操作区域，使三维资产从整体模型转变为可分解、可选择、可修改的结构化对象。

2．局部编辑

局部编辑的目标是在用户指定区域内进行修改，同时保持未编辑区域的几何结构、纹理外观和整体一致性。例如，给一个三维角色添加局部装饰时，不应改变其整体姿态；替换某个物体部件时，也不应导致其他区域发生结构漂移。因此，局部编辑的关键在于平衡目标区域的可控修改与未编辑区域的一致性保持。

已有三维编辑方法大体可以分为两类。一类方法使用得分蒸馏采样等优化方式，根据文本提示逐场景优化三维表示，但通常耗时较长。另一类方法先将三维模型渲染为多视图图像，在二维图像空间中完成编辑后再重建三维模型。多视图编辑路线效率较高，但由于编辑发

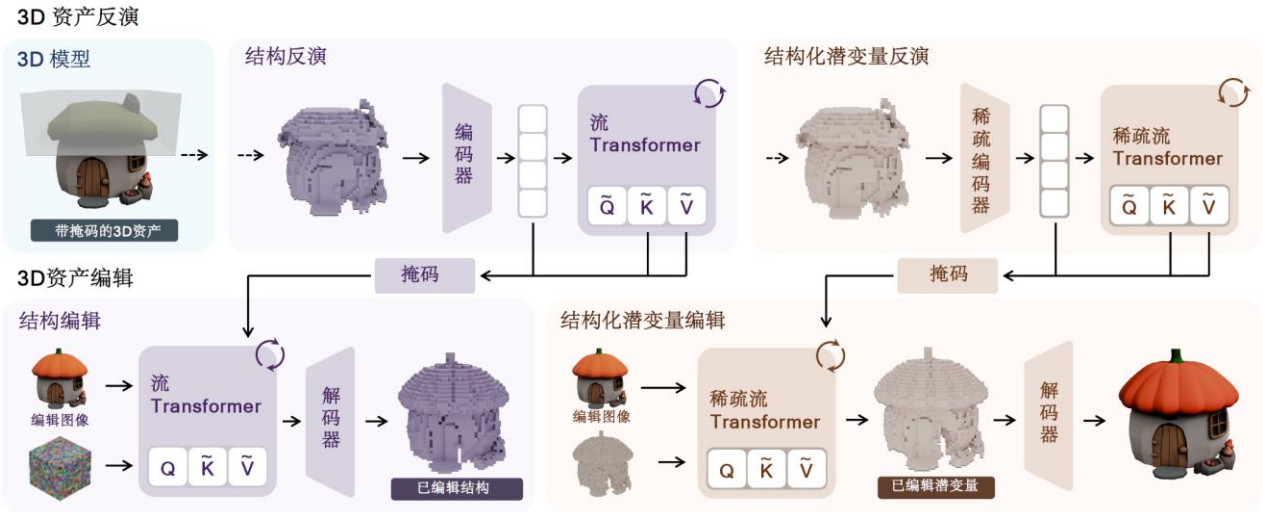


图 3：VoxHammer 的整体框架。

表 3：Edit3D-Bench 上的局部编辑定量比较结果。

方法	未编辑区域保持				整体三维质量		条件对齐	
	CD.↓	PSNR (M)↑	SSIM (M)↑	LPIPS (M)↓	FID↓	FVD↓	DINO-I↑	CLIP-T↑
Vox-E[20]	/	13.84	0.827	0.316	87.41	3000.3	0.721	0.274
MVEdit[21]	0.017	26.12	0.945	0.070	58.53	946.5	0.911	0.281
Tailor3D[22]	0.043	20.94	0.861	0.148	110.52	3812.1	0.704	0.258
Instant3DiT[23]	0.016	27.70	0.957	0.067	45.93	450.1	0.903	0.260
TRELLIS[12]	0.047	23.64	0.919	0.131	38.19	757.2	0.911	0.283
VoxHammer	**0.012**	**41.68**	**0.994**	**0.027**	**23.05**	**187.8**	**0.947**	**0.287**

生在二维空间，不同视角之间容易产生不一致，重建阶段也可能引入位置偏差和空间漂移，导致未编辑区域被破坏。

VoxHammer 针对这一问题提出了免训练的原生三维局部编辑方法。与仅依赖文本提示直接生成三维结果不同，VoxHammer 首先将输入三维资产渲染为指定视角图像，并结合用户给定的编辑区域和编辑文本，利用图像编辑模型生成二维编辑图像；随后将原始三维资产、二维编辑图像和三维编辑掩码共同作为条件，驱动预训练结构化三维生成模型 TRELLIS[12]在原生三维潜空间中完成局部编辑。如图 3 所示，VoxHammer 的核心流程包括三维资产反演（Inversion）和三维资产编辑两个阶段。第一阶段为三维反演：给定一个三维模型，VoxHammer 沿结构化三维扩散模型（structured 3D latent diffusion models）的生成过程进行反演得到对应的噪声状态，并记录每个时间步的反演潜变量（invert-ed latents）以及注意力层中的键/值（K/V）张量。第二阶段为去噪编辑：模型从反演得到的噪声状态出发，在用户指定区域内生成新的结构与外观；同时，对于未编辑区域，将当前去噪过程中的潜变量和 K/V 张量替换为反演阶段缓存的对应特征。这样，编辑区域可以根据编辑条件被重生成，未编辑区域则受到原始特征约束，从而减少几何漂移和纹理变化。

该方法的核心在于利用预训练三维生成模型已有的生成能力，并在去噪编辑过程中通过三维掩码区分编辑区域与保留区域，使模型既能在目标区域内完成指定修改，又能维持未编辑区域的几何和纹理一致性。三维反演、未编辑区域的潜变量替换以及注意力键值张量替换，是实现这一约束的关键技术手段。由于编辑过程直接发生在原生三维潜空间中，该方法能够减少多视图图像编辑中常见的视角不一致和重建偏差问题，并在局部修改效果与整体结构一致性之间取得更好的平衡。

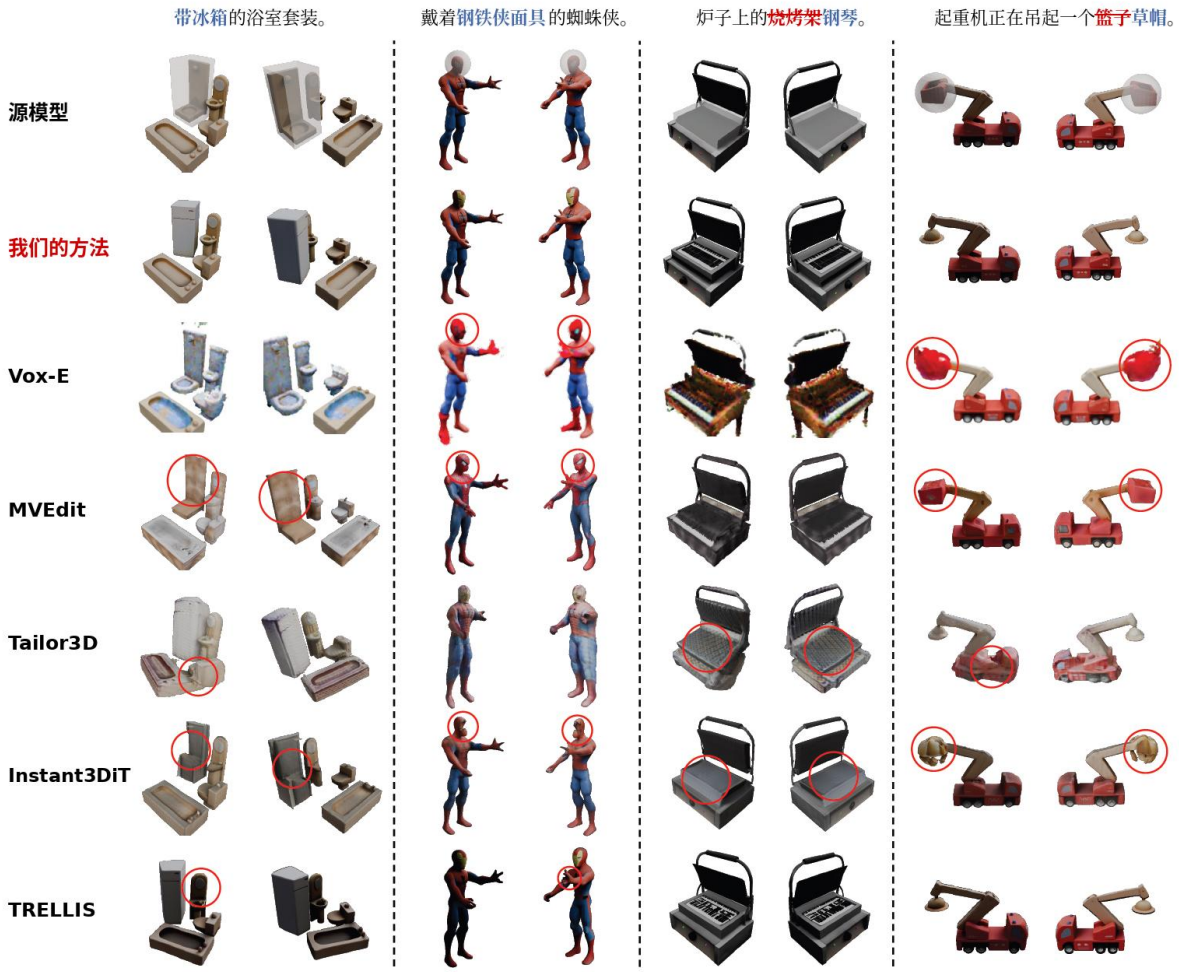


图 4：VoxHammer 与不同三维编辑方法在 Edit3D-Bench 上的局部编辑效果对比。

图 5：SegviGen 辅助 VoxHammer 进行部件级局部编辑。

为了系统评估局部编辑效果，VoxHammer 构建了 Edit3D-Bench。该基准包含人工标注的三维编辑区域，

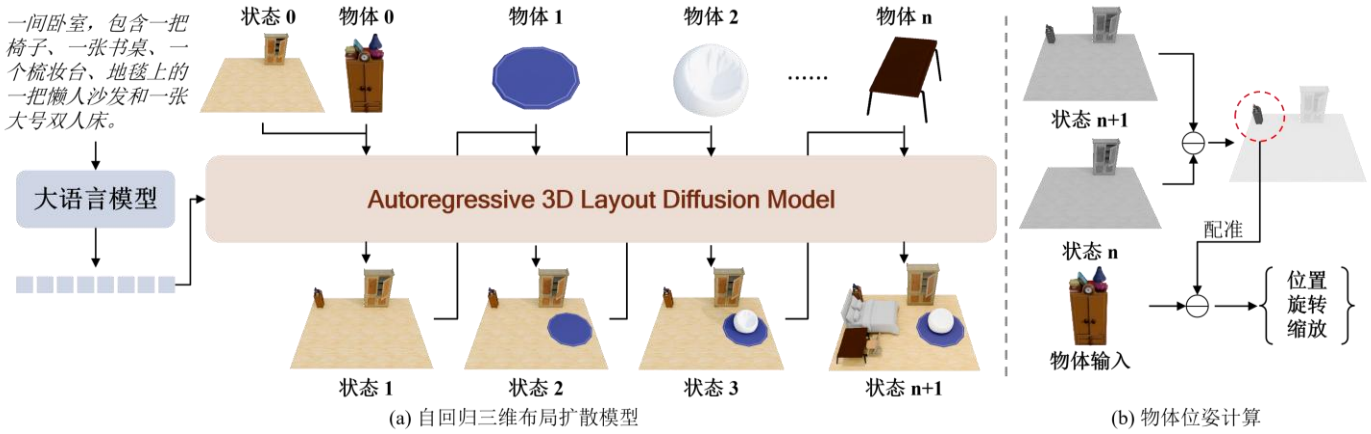


图 6：LaviGen 的自回归三维布局生成流程。

并从未编辑区域保持、整体三维质量以及条件对齐（编辑结果与输入条件的一致性）三个方面进行评估。实验结果表明，VoxHammer 在未编辑区域保持和整体编辑质量上整体优于 Vox-E[20]、MVEdit[21]、Tailor3D[22]、Instant3DiT[23]和 TRELLIS[12]等基线方法。图 4 给出了不同方法在 Edit3D-Bench 上的定性对比。可以看到，多视图编辑或重建式方法容易在未编辑区域产生结构变形、位置漂移或纹理破坏，而 VoxHammer 能够在完成目标区域修改的同时，更稳定地保持原始三维结构和外观一致性。如表 3 所示，VoxHammer 在未编辑区域保持上取得最低 CD（0.012）、最高 PSNR（41.68）和 SSIM（0.994），LPIPS 也降至 0.027；同时，其 FID/FVD 分别降低到 23.05 和 187.8，DINO-I 与 CLIP-T 达到 0.947 和 0.287，说明局部修改并未牺牲整体质量和条件对齐。

此外，VoxHammer 所需的三维编辑掩码也可由 SegviGen 提供：SegviGen 先将三维对象划分为语义部件，用户选定目标部件后将其作为编辑区域，从而形成“部件解析—区域选择—局部编辑”的连续流程，如图 5 所示。

3．布局生成

三维场景的布局生成任务不再关注物体内部的解析和编辑，而是考虑在给定一组物体和语言指令后，在三维空间中如何生成各物体合理的位置、朝向和尺度，重点强调这些实例之间的语义关系和物理约束。

早期布局生成方法通常利用三维场景数据训练模型直接预测物体位置和朝向，但受限于场景数据规模和真实空间关系建模能力，模型往往难以泛化到复杂布局。随着大语言模型的发展，一些方法（例如 LayoutGPT[24]）开始将布局表示为 JSON 或结构化文本，由语言模型根据指令生成物体坐标和旋转角。这类方法能够利用语言模型的语义知识理解指令中的空间关系，并生成看似合理的物体位置、朝向等布局参数，但由于文本坐标并不直接包含三维几何和物理约束，生成结果容易出现碰撞、悬浮、遮挡和越界等问题。另一些方法（例如 LayoutVLM[27]）通过渲染视图和视觉反馈优化布局，能够在一定程度上改善视觉合理性，但通常需要多轮渲染与优化，计算成本较高，对三维接触、支撑和边界约束的建模也不够直接。

LaviGen 的核心思路是将布局生成视为原生三维空间中的条件生成问题。如图 6 所示，给定当前场景、待放置物体和布局指令，模型以自回归方式逐步生成更新后的场景状态。每一步中，模型根据已有场景状态、待放置物体的三维几何信息和语言指令，生成包含新增物体的下一场景状态；随后通过比较更新前后的场景差异定位新增区域，并对输入物体进行拟合，从而估计其位置、旋转和尺度参数。由此，布局生成不再只是输出一组抽象坐标，而是在三维空间中直接建模物体与场景之间的空间关系。

为了区分当前场景和待放置物体，LaviGen 引入身份感知嵌入，将场景状态、目标物体和带噪潜变量放入统一三维潜在表示中建模。由于布局生成是长序列自回

表 4：LaviGen 主要定量比较与消融结果。

方法	物理合理性		空间一致性		PSA↑	运行时间(s)↓
	CF↑	IB↑	Pos.↑	Rot.↑		
LayoutGPT[24]	83.8	24.2	**80.8**	**78.0**	16.6	**21.3**
Holodeck[25]	77.8	8.1	62.8	55.6	5.6	58.2
I-Design[26]	76.8	34.3	68.3	62.8	18.0	179.2
LayoutVLM[27]	81.8	94.9	77.5	73.2	58.8	75.5
模型基线	75.6	64.8	45.1	44.7	16.7	145.7
+ id-aware emb.	89.1	96.8	68.8	66.5	71.4	144.1
+ L_holistic.	79.5	81.9	61.4	58.7	59.7	24.5
+ L_step. (完整模型)	**97.3**	**98.6**	76.9	77.1	**78.8**	24.3

归过程，下一步生成需要依赖模型前一步生成的结果，容易产生误差累积。为缓解这一问题，LaviGen 进一步提出双重引导的自展开蒸馏策略（dual-guidance self-rollout distillation）。该策略不再只依赖真实历史状态训练模型，而是让模型在训练阶段基于自身生成的中间场景继续展开，并分别引入场景级整体引导和步骤级物体引导：前者约束最终场景的整体合理性，后者在每一步提供与当前待放置物体的校正信号，从而提升长序列布局生成的稳定性、空间准确性和物理合理性。

实验结果表明，LaviGen 在三维布局生成中具有更好的物理合理性和空间一致性。如图 7 所示，LayoutGPT[24]虽能生成语义相关的布局，但容易出现碰撞、越界和悬浮；LayoutVLM[27]通过渲染视图优化缓解了部分越界问题，但仍存在碰撞和悬浮现象。相比之下，LaviGen 直接在原生三维空间中建模物体关系，在拥挤场景中也能生成更稳定的空间布局。 定量结果如表 4 所示，LaviGen 完整模型在物理合理性 CF 和 IB 指标上达到 97.3 和 98.6，PSA 指标达到 78.8，显著高于 LayoutGPT[24]、Holodeck[25]和 LayoutVLM[27]等基线。消融结果表明，身份感知嵌入与双重蒸馏约束能够逐步提升布局质量，并在保持语义对齐的同时增强物理可行性。

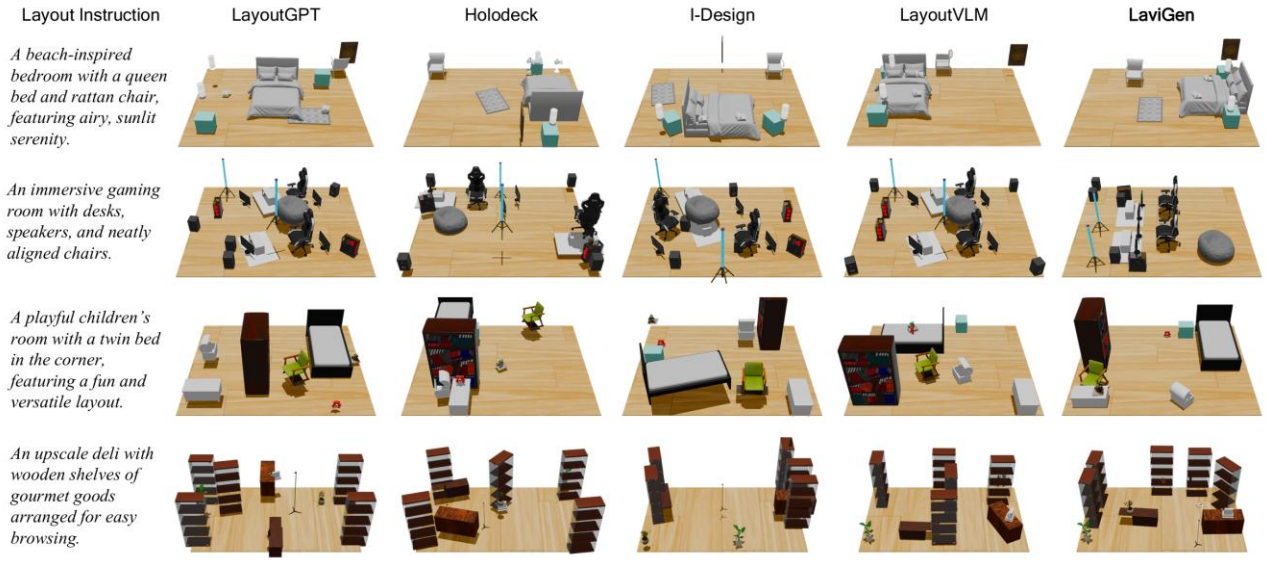


图 7 文本驱动三维布局生成的定性比较。

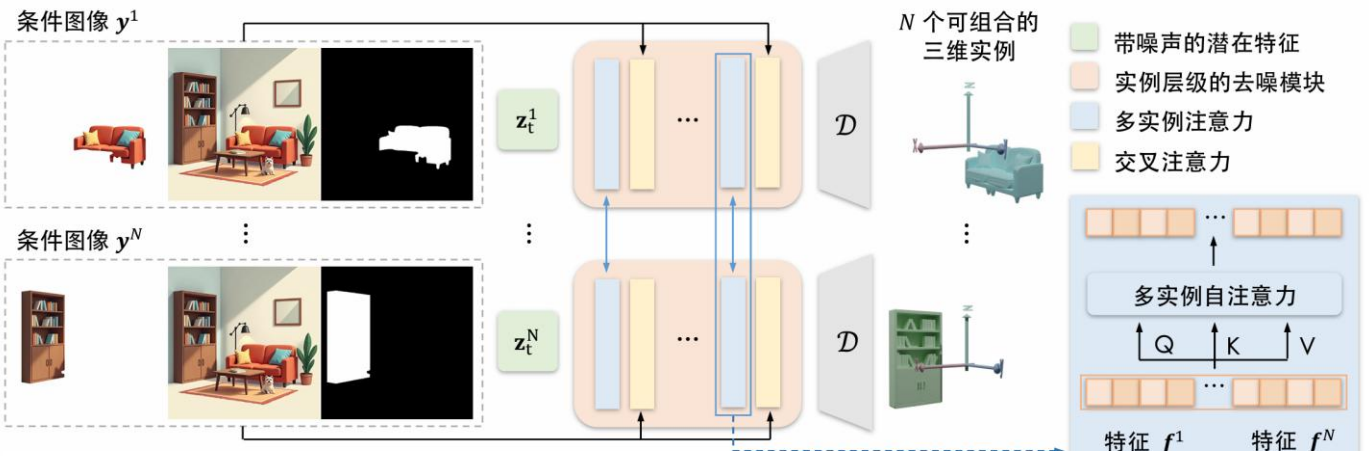


图 8：MIDI 的多实例注意力机制示意图。

4．场景生成

除了基于已知三维资产情况下的布局生成，团队还探索了在单图像条件下的组合式三维场景构建。

如图 8 所示，MIDI 为每个物体实例构造条件图像，将局部物体图像、实例掩码与全局场景图像共同编码，并把多个实例对应的带噪三维潜变量同时输入共享权重的扩散 Transformer（DiT）[28]去噪模块，在同一扩散过程中并行生成多个三维物体。为避免各实例孤立地生成，MIDI 在去噪网络中引入多实例注意力机制（Multi-Instance Attention），使每个实例的特征能够与场景中其他实例的特征交互，从而在生成阶段直接建模物体之间的相对位置、空间依赖与整体一致性；同时，交叉注意力用于融合局部物体细节和全局场景上下文。最终，各实例的三维潜变量经解码器还原为可组合的三维物体，并形成具有空间关联的三维场景。进一步地，如图 9 所示，相比“图像分割—物体补全—逐实例生成—布局优化”的传统多阶段流程，MIDI 将实例形状生成和空间关系建模统一到同一生成过程中，减少中间误差累积，天然满足形状生成与布局生成的协同优化。

基于上述多实例扩散框架，MIDI 能够从不同类型的单张图像中生成组合式三维场景。图 10 展示了其在

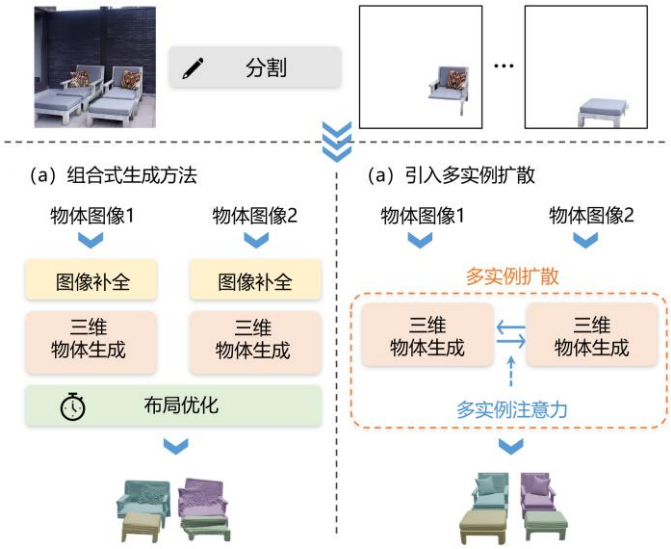


图 9：传统组合式生成与 MIDI 多实例扩散流程对比。

合成数据、真实场景图像以及风格化图像上的生成结果：模型不仅能够恢复多个物体实例的三维形状，还能保持较合理的相对位置和场景布局。表 5 进一步给出了合成数据集上的定量结果：在 3D-Front 和 BlendSwap 上，MIDI 均取得最低的场景级和物体级 Chamfer Distance（CD-S/CD-O）以及最高的 F-Score 和 IoU-B 结果，说明其同时提升了实例形状质量和场景空间关系。运行时间约 40 秒，也明显短于 Gen3DSR[34]和 REPARO[35]等分钟级方法。

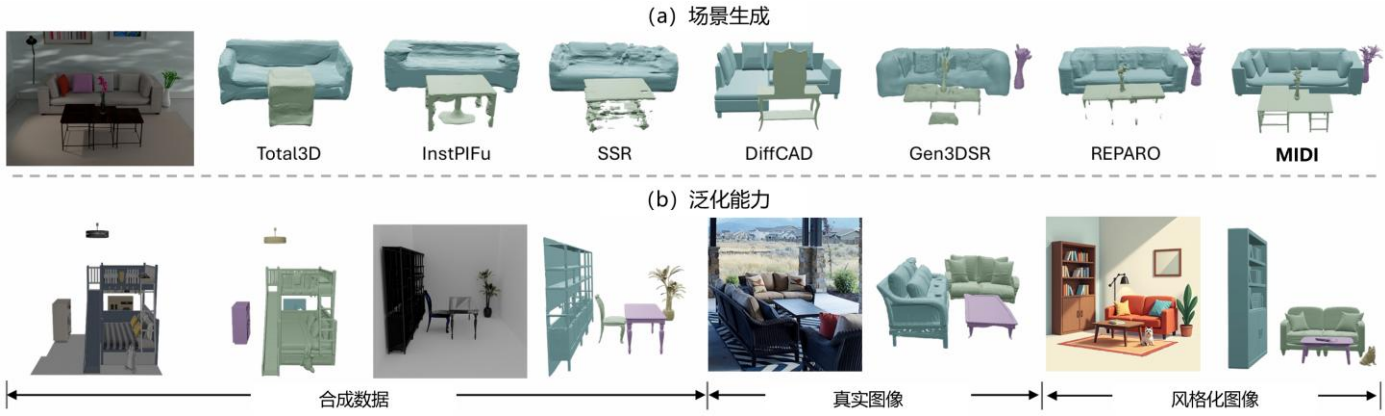


图 10：MIDI 的单图像三维场景生成与泛化效果。

表 5：合成数据集上的场景级与物体级定量比较结果。

方法	3D-Front 数据集					BlendSwap 数据集					时间↓
	CD-S↓	F-Score-S↑	CD-O↓	F-Score-O↑	IoU-B↑	CD-S↓	F-Score-S↑	CD-O↓	F-Score-O↑	IoU-B↑	
PanoRecon[29]	0.150	40.65	0.211	35.05	0.240	0.427	19.11	0.713	13.06	0.119	32s
Total3D[30]	0.270	32.90	0.179	36.38	0.238	0.258	37.93	0.168	38.14	0.328	39s
InstPIFu[31]	0.138	39.99	0.165	38.11	0.299	0.129	50.28	0.167	38.42	0.340	32s
SSR[32]	0.140	39.76	0.170	37.79	0.311	0.132	48.72	0.173	38.11	0.336	32s
DiffCAD[33]	0.117	43.58	0.190	37.45	0.392	0.110	52.83	0.169	38.98	0.457	64s
Gen3DSR[34]	0.123	40.07	0.157	38.11	0.363	0.107	60.17	0.148	40.76	0.449	9min
REPARO[35]	0.129	41.68	0.160	40.85	0.339	0.115	62.39	0.151	42.84	0.410	4min
Ours	**0.080**	**50.19**	**0.103**	**53.58**	**0.518**	**0.077**	**78.21**	**0.090**	**62.94**	**0.663**	40s

三、面向空间智能的三维世界构建

当三维生成模型从单体资产生成走向结构化、可编辑和组合式场景化生成时，生成结果不再只是视觉资产，也可以成为空间智能系统的数据来源、仿真环境和评测对象。SegviGen[3]和 VoxHammer[4]使三维资产具备可解析、可选择和可编辑的局部结构，LaviGen[2]和 MIDI[1]则进一步面向布局生成和组合式场景生成，建模多物体之间的空间关系。由此得到的三维内容不仅包含几何和外观信息，还保留部件结构、对象关系、空间布局和可编辑区域，可为遮挡关系理解、可达性分析、操作路径规划和多物体交互建模提供结构化三维数据支撑。进一步看，面向机器人、具身智能和多模态空间理解等应用，三维生成还需要从静态资产生成走向更大规模、更开放的动态组合式三维世界构建，在场景扩展过程中保持全局结构一致性、功能语义连贯性和物理可行性，并进一步引入材质、质量、摩擦、受力、可动关节和动力学行为等物理属性。未来，原生三维生成先验可提供结构分布，语言模型可提供任务理解和规划能力，视觉模型可提供感知反馈，物理模拟器可提供约束验证，使三维生成系统从一次性前馈模型逐步发展为可迭代搭建、持续编辑和交互控制的闭环式空间智能工具。

责任编委 张杰

参考文献

[1] Huang Z, Guo Y C, An X Q, et al. MIDI: Multi-Instance Diffusion for Single Image to 3D Scene Generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2025: 23646-23657.

[2] Feng H R, Niu Y F, Huang Z H, et al. Repurposing 3D Generative Model for Autoregressive Layout Generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2026: 3231-3243.

[3] Li L, Feng H R, Huang Z H, et al. SegviGen: Repurposing 3D Generative Model for Part Segmentation[EB/OL]. arXiv:2603.16869, 2026.

[4] Li L, Huang Z H, Feng H R, et al. VoxHammer: Training-Free Precise and Coherent 3D Editing in Native 3D Space[C]//International Conference on 3D Vision (3DV). 2026.

[5] Poole B, Jain A, Barron J T, et al. DreamFusion: Text-to-3D using 2D Diffusion[C]//International Conference on Learning Representations. 2023.

[6] Liu R, Wu R, Van Hoorick B, et al. Zero-1-to-3: Zero-shot One Image to 3D Object[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 9298-9309.

[7] Liu Y, Lin C, Zeng Z, et al. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image[C]//International Conference on Learning Representations. 2024.

[8] Huang Z, Wen H, Dong J, et al. EpiDiff: Enhancing Multi-View Synthesis via Localized Epipolar-Constrained Diffusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 9784-9794.

[9] Wen H, Huang Z, Wang Y, et al. Ouroboros3D: Image-to-3D Generation via 3D-aware Recursive Diffusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2025: 21631-21641.

[10] Huang Z, Guo Y C, Wang H, et al. MV-Adapter: Multi-View Consistent Image Generation Made Easy[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2025: 16377-16387.

[11] Huang Z, Feng H, Sun Y, et al. AnimaX: Animating the Inanimate in 3D with Joint Video-Pose Diffusion Models[EB/OL]. arXiv:2506.19851, 2025.

[12] Xiang J, Lv Z, Xu S, et al. Structured 3D Latents for Scalable and Versatile 3D Generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2025: 21469-21480.

[13] Xiang J, Chen X, Xu S, et al. Native and Compact Structured Latents for 3D Generation[EB/OL]. arXiv:2512.14692, 2025.

[14] Lipman Y, Chen R T Q, Ben-Hamu H, et al. Flow Matching for Generative Modeling[C]//International Conference on Learning Representations. 2023.

[15] Zhou Y, Gu J, Chiang T Y, et al. Point-SAM: Promptable 3D Segmentation Model for Point Clouds[C]//International Conference on Learning Representations. 2025.

[16] Ma C, Li Y, Yan X, et al. P3-SAM: Native 3D Part Segmentation[EB/OL]. arXiv:2509.06784, 2025.

[17] Ma Z, Yue Y, Gkioxari G. Find Any Part in 3D[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2025: 7818-7827.

[18] Yang Y, Huang Y, Guo Y C, et al. SAMPart3D: Segment Any Part in 3D Objects[EB/OL]. arXiv:2411.07184, 2024.

[19] Liu M, Uy M A, Xiang D, et al. PartField: Learning 3D Feature Fields for Part Segmentation and Beyond[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2025: 9704-9715.

[20] Sella E, Fiebelman G, Hedman P, et al. Vox-E: Text-Guided Voxel Editing of 3D Objects[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023: 430-440.

[21] Chen H, Shi R, Liu Y, et al. Generic 3D Diffusion Adapter Using Controlled Multi-View Editing[EB/OL]. arXiv:2403.12032, 2024.

[22] Qi Z, Yang Y, Zhang M, et al. Tailor3D: Customized 3D Assets Editing and Generation with Dual-Side Images[EB/OL]. arXiv:2407.06191, 2024.

[23] Barda A, Gadelha M, Kim V G, et al. Instant3dit: Multiview Inpainting for Fast Editing of 3D Objects[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2025: 16273-16282.

[24] Feng W, Zhu W, Fu T J, et al. LayoutGPT: Compositional Visual Planning and Generation with Large Language Models[C]//Advances in Neural Information Processing Systems. 2023: 18225-18250.

[25] Yang Y, Sun F Y, Weihs L, et al. Holodeck: Language Guided Generation of 3D Embodied AI Environments[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 16227-16237.

[26] Çelen A, Han G, Schindler K, et al. I-Design: Personalized LLM Interior Designer[EB/OL]. arXiv:2404.02838, 2024.

[27] Sun F Y, Liu W, Gu S, et al. LayoutVLM: Differentiable Optimization of 3D Layout via Vision-Language Models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2025: 29469-29478.

[28] Peebles W, Xie S. Scalable Diffusion Models with Transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023: 4195-4205.

[29] Dahnert M, Hou J, Nießner M, et al. Panoptic 3D Scene Reconstruction From a Single RGB Image[C]//Advances in Neural Information Processing Systems. 2021: 8282-8293.

[30] Nie Y, Han X, Guo S, et al. Total3DUnderstanding: Joint Layout, Object Pose and Mesh Reconstruction for Indoor Scenes From a Single Image[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 55-64.

[31] Liu H, Zheng Y, Chen G, et al. Towards High-Fidelity Single-View Holistic Reconstruction of Indoor Scenes[C]//European Conference on Computer Vision. 2022: 429-446.

[32] Chen Y, Ni J, Jiang N, et al. Single-View 3D Scene Reconstruction with High-Fidelity Shape and Texture[C]//International Conference on 3D Vision (3DV). 2024: 1456-1467.

[33] Gao D, Rozenberszki D, Leutenegger S, et al. DiffCAD: Weakly-Supervised Probabilistic CAD Model Retrieval and Alignment from an RGB Image[J]. ACM Transactions on Graphics, 2024, 43(4): 106:1-106:15.

[34] Ardelean A, Özer M, Egger B. Gen3DSR: Generalizable 3D Scene Reconstruction Via Divide and Conquer From a Single View[C]//International Conference on 3D Vision (3DV). 2025: 616-626.

[35] Han H, Yang R, Liao H, et al. REPARO: Compositional 3D Assets Generation with Differentiable 3D Layout Alignment[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2025: 25367-25377.

樊红兴

北京航空航天大学计算机学院 2021 级博士研究生，导师为盛律教授，主要研究方向为视觉生成，多模态大语言模型。在 IEEE TIP/CVPR/AAAI/ACM MM 等重要国际期刊和会议发表论文 10 余篇，Google Scholar 显示被引用数 300 余次，担任多个顶会顶刊审稿人。

Email: fanhongxing@buaa.edu.cn

盛 律

北京航空航天大学教授、博士生导师、国家级青年人才、智源学者和小米青年学者，入选斯坦福全球前 2%顶尖科学家排行榜单。主要研究方向为三维视觉、多模态大模型和具身智能。在 IEEE TPAMI/ACM TOG/IJCV 以及 CVPR/ICCV/ICLR/ECCV 等重要国际期刊和会议发表论文 80 余篇，Google Scholar 显示被引用数 9500 余次。现任 ACM Computing Surveys 编委，CVPR/ICLR/ECCV/ACM Multimedia/AAAI 等领域主席，以及多个领域顶会顶刊审稿人和程序委员。CCF 高级会员，任 CCF 和 CSIG 多个专委会执行委员，VALSE 执行领域主席。

Email: lsheng@buaa.edu.cn

顶会观察

ICLR 2026

南京大学助理教授 傅朝友

国际表征学习会议（International Conference on Learning Representations, ICLR）作为全球人工智能领域的顶级盛会，备受学术界和产业界的关注。 第 14 届 ICLR 大会于 2026 年 4 月底在巴西里约热内卢成功举办。本届会议在经历了激增的投稿量、严苛的 LLM 审查机制以及开放评审安全突发事件考验后，依然保持了学术前瞻性与严谨性。本届会议不仅组织了涵盖核心机器学习与交叉学科的特邀报告，还开展了 40 场聚焦智能体、AI for Science、基础模型对齐等热门议题的专题研讨会，并开展 35 场 Oral 报告，系统性地展现了该领域的最新研究进展。

一、 会议亮点

评审机制革新：规范 LLM 使用与“幻觉文献”零容忍。面对 LLM 在论文写作与审稿中的全面渗透，本届会议出台了严格的管控与检测政策。组委会在审稿端部署了双重 LLM 内容检测器，以辅助 AC 甄别低质量的 AI 自动生成审稿意见，同时在投稿端重点排查大模型生成的幻觉参考文献。通过自动化提取比对与多级人工核查（每篇涉事论文至少由 3 名人类复核）相结合，所有被证实编造虚假文献的论文均被作 Desk Reject 处理，切实维护了学术严谨性。

突发危机响应：采取分数重置与人员重分配机制保障评审公正。在本届会议的 Rebuttal 阶段，遭遇了 OpenReview API 恶意抓取事件，导致部分作者、审稿人与 AC 身份泄露，并引发了学术骚扰与串通打分威胁。面对该违规事件，组委会紧急采取了冻结讨论、将评分重置回 Rebuttal 前状态、并为受波及论文重新分配 AC 等应对措施。此外，官方还适当延长了 Meta-Review 周期，最终有效保障了整体录用标准。

学术生态拓展： “开放科学”与 Workshop 赛道规模显著增长。本届大会注重开放科学与跨学科交流，Workshop 赛道提案数量创下新高，共收到 151 份提案（较去年增长约 24%），最终精选出 40 场。议题呈现出明显的跨学科前沿导向，诸如智能体系统、AI for Science（涵盖偏微分方程、基因组学、材料科学）、基础模型 Test-Time Updates 等成为主要探讨方向，进一步拓宽了表征学习的研究边界。

二、 录用情况

投稿与录用：大会共收到了 **19,525** 篇有效投稿，相较于 2025 年的 11,672 篇出现了爆发式增长。经过多轮严格评审，最终接收 **5,355** 篇论文，录用率约为 **27.4%**，整体延续了 ICLR 近年来将录用率稳定在 30% 上下的趋势。在录用的论文中，有 223 篇被选为 Oral，占比仅为 1.13%。其余 5,120 篇录用论文以 Poster 形式展示。此外，受本届严格的合规性与 LLM 使用审查影响，Desk Reject（779 篇）与 Withdraw（5,042 篇）的数量也创下新高。

评审团队：为了应对海量且激增的投稿，大会组织了由 18,054 名审稿人构成的评审团，共计提供了逾 7.6 万份评审意见。在遭遇突发事件的压力下，大会紧急优化了 AC 的接管与校准流程，通过延长 Meta-Review 周期，在维持一贯录用标准的前提下，确保了评审质量。

区域与机构表现： 本届 ICLR 的论文区域分布呈

现出显著的结构性变化。据第三方开源项目 TradingView 提取的论文机构署名数据显示，在接收的论文中，中国大陆机构参与的论文占比为 43.7%，超过了美国的 31.9%。高校方面，清华大学以 332 篇录用论文位列全球第一；上海交大（240 篇）、浙江大学（232 篇）紧随其后。企业及新型机构方面，阿里巴巴和上海 AI Lab 以 135 篇录用位居国内企业前列、华为（111 篇）、字节跳动（107 篇）等机构亦有显著产出。不过值得注意的是，在代表更高学术评价的 Oral 论文中，美国机构占比约为 40%（中国约 30%），表明其在顶尖研究上仍保持一定的引领优势。

三、 热门研究方向

根据主题演讲、Workshop 议题及录用论文分布，ICLR 2026 呈现出以下核心技术趋势：

智能体系统与自主可靠性全面深化：纯文本大模型的单点研究趋于收敛，智能体在复杂环境中的应用成为重点探讨方向。本届会议设置了多个相关的专题 Workshop，重点涵盖大模型记忆机制（Memory）、终身学习智能体（Lifelong Agents）以及真实世界可靠自主性（Agentic AI in the Wild）等议题，探讨模型在动态环境下的规划与执行能力。

AI for Science 领域的跨学科融合：AI 与自然科学的结合正向底层机制探索演进。本届大会的特邀报告与多个 Workshop 集中探讨了这一交叉领域。具体方向包括偏微分方程求解（AI&PDE）、加速材料设计（AI4MAT）、生成式基因组学（Gen^2）等。研究重点涉及如何将物理先验知识、空间对称性及热力学原理整合至表征学习模型中，以辅助科学问题的计算与发现。

开放模型开发与推理时计算（Test-Time Compute）：针对当前前沿基础模型训练细节不够透明的现状，大会特邀报告探讨了借助 Marin 等开源平台，实现从数据配比、优化器设计到 Scaling Laws 的公开实验与开发。此外，在推理架构层面，相关 Workshop 集中探讨了超越传统 CoT 的“Latent & Implicit Thinking”以及“Test-Time Updates”等技术，旨在研究模型在推理阶段动态分配计算资源的有效机制。

四、 热点论文

本届 ICLR 官方评审委员会最终在浩如烟海的录用论文中，评选出 **2 篇杰出论文**（Outstanding Papers）和 **1 篇荣誉提名**（Honorable Mention）：

(Outstanding Paper) Transformers are Inherently Succinct[1]：随着 LLM 的广泛应用，理解 Transformer 架构的理论表达能力成为核心议题。该论文从形式语言理论的角度，提出“简洁性（Succinctness）”作为衡量 Transformer 表达能力的新标准。研究在理论上证明了：相较于有限状态自动机（Finite Automata）、线性时序逻辑（LTL）以及传统 RNN（含近期热门的 SSMs），固定精度的 Transformer 能够以“指数级”甚至“双指数级”的简洁性来编码复杂的概念模式。这项纯理论工作不仅为“Transformer 架构为何如此强大”提供了深刻的数学解答，同时也更加严谨证明了对 Transformer 内部属性进行形式化验证是一个困难（EXPSPACE-Complete）的计算问题。

(Outstanding Paper) LLMs Get Lost In Multi-Turn Conversation[2]：这篇研究揭示了当今 LLM 在真实交互场景中的关键局限。现有的 LLM 评测多集中于“单轮、指令明确”的理想环境，而现实中用户与模型的交互往往是多轮且指令不明。该研究团队设计了一种可扩展的多轮对话评估方法，对 15 款主流大模型进行了逾 20 万次模拟测试。结果表明：当面临多轮、信息不全的交互时，无论是开源小模型还是顶尖闭源模型（如 GPT-4o, Gemini 2.5 Pro），其整体性能均出现显著下滑（平均降幅达 39%），且输出可靠性大幅降低。论文指出，这种“在对话中迷失”的根源在于模型倾向于在早期做出错误假设，并过度依赖先前的错误尝试而无法完成自我纠错。这项工作为未来评测基准的演进提供了新的方向。

(Honorable Mention) The Polar Express: Optimal Matrix Sign Methods and their Application to the Muon Algorithm[3]：随着深度学习规模的扩大，高效的底层优化器至关重要。这篇研究聚焦近期在训练大模型时具备竞争力的 Muon 优化

器，提出了一种名为“Polar Express”的全新矩阵极分解算法。该方法利用近似理论设计最佳多项式，适配 GPU 的高吞吐量需求并规避了低精度计算陷阱。实验证明，将其集成至 Muon 优化器后，能显著且稳定地降低 GPT-2 等语言模型的验证损失。

五、 时间检验奖

时间检验奖（Test of Time Award）是 ICLR 的最高荣誉之一，旨在回溯并表彰 10 年前（即 2016 年）发表，且经长期实践证明对人工智能领域产生深远、基石性影响的经典工作。本年度获得该殊荣的是两篇在生成式人工智能与深度强化学习领域具有里程碑意义的奠基性工作：

Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks[4]。在 GAN 被提出的初期，其训练很不稳定且难以生成高分辨率图像。在这篇论文中，研究团队首次成功将深度卷积神经网络与 GAN 结合，提出了一系列网络架构设计规范（如使用步长卷积代替池化、引入 Batch Normalization、取消全连接层等），有效缓解了早期 GAN 的训练崩溃难题。DCGAN 不仅率先实现了对复杂自然图像的稳定合成，其在隐空间中展示的“向量算术”特性也揭示了模型深层的语义组合能力，为现代计算机视觉中生成模型（及后续演进架构）的底层设计确立了范式。

Continuous Control with Deep Reinforcement Learning[5]。该工作大幅拓展了深度强化学习在物理控制领域的应用边界。在此之前，经典的 DQN 算法主要处理离散动作空间，面对物理机器人等“连续且高维”的动作空间时存在显著局限。为了解决“维度灾难”与动作离散化带来的精度折损，该论文创新性地将确定性策略梯度（DPG）与 Actor-Critic 架构、深度神经网络及经验回放机制相融合，提出了 DDPG 算法。该算法率先实现了端到端（直接从高维像素输入到连续动作输出）的运动控制，极大推动了深度强化学习在复杂机器人与连续控制任务中的广泛应用。

六、 总结展望

从 ICLR 2026 创纪录的近两万篇投稿，以及仅 1.13%的 Oral 录用率可以看出，全球深度学习领域的研究热度仍在持续攀升，而顶级学术成果的筛选标准依然保持着极高门槛。在这一背景下，中国大陆研究机构的论文录用占比首次在顶级人工智能会议中超过美国（达到 43.7%），不仅反映出中国人工智能研究生态的快速成熟，也体现了其在基础研究领域长期积累后的集中释放。

纵观本届会议的研究图景，一个值得关注的变化是：学术界的关注重点正逐步从单纯追求模型规模扩张（Scaling Laws），转向对模型底层机制、系统可靠性以及高效优化方法的深入探索。研究者开始更加关注大模型在真实复杂环境中的行为表现与能力边界。一方面，越来越多工作系统性揭示并尝试解决大模型在复杂多轮交互、长程推理以及信息不完备场景下出现的脆弱性与性能退化问题；另一方面，研究社区也在从更加严谨的数学视角重新审视神经网络的表达能力与理论边界，并探索在算力约束条件下，通过最佳多项式近似等数学工具实现训练与推理效率提升的新路径。

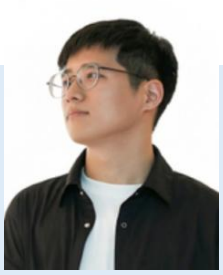
这一趋势表明，人工智能研究正在逐渐告别以静态基准测试为核心的能力评估范式，转而更加关注模型在开放环境中的实际表现，以及在有限算力、高可靠性和可持续部署条件下的综合能力。如何构建真正稳定、可信且具备长期泛化能力的智能系统，正成为学界与产业界共同关注的核心议题。

展望未来，随着多模态原生架构、推理时计算以及 AI for Science 等方向的持续融合，人工智能领域有望迎来新的技术突破。如何通过系统架构创新赋予模型动态规划、反思与自我纠错能力，如何将物理规律、领域知识与先验约束有效融入神经网络的学习过程，从而提升模型的推理效率、可靠性与泛化能力，都有望成为推动下一代人工智能向更高水平通用智能演进的重要研究命题。

责任编委 **魏秀参**

参考文献

[1] Pascal Bergsträßer, et al. Transformers are Inherently Succinct. ICLR 2026.
[2] Philippe Laban, et al. LLMs Get Lost In Multi-Turn Conversation. ICLR 2026.
[3] Noah Amsel, et al. The Polar Express: Optimal Matrix Sign Methods and their Application to the Muon Algorithm. ICLR 2026.
[4] Alec Radford, et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. ICLR 2016.
[5] Timothy P. Lillicrap, et al. Continuous Control with Deep Reinforcement Learning. ICLR 2016.



傅朝友

南京大学智能科学与技术学院研究员、助理教授、博导，入选中国科协“青年人才托举工程”。2022 年博士毕业于中科院自动化所模式识别实验室。主要研究方向为多模态内容分析，谷歌学术引用 9800 余次，开源项目累计获 2 万余次 GitHub Stars。担任 Pattern Recognition/IEEE T-BIOM 期刊编委、ICLR/ICML 会议领域主席。

Email: cyfu@nju.edu.cn

武汉大学叶茫教授访谈

2026 年 4 月 23 日，《CCF-CV 专委简报》在线采访了武汉大学博士生导师叶茫教授。下面是采访实录。

叶老师，您好！首先，请您分享一下您的个人学习和研究经历。

大家好！我本科与硕士阶段均在武汉大学度过，打下了扎实的计算机与人工智能基础。之后我赴香港浸会大学攻读博士学位，系统开展计算机视觉与多媒体方向研究；在博士就读期间，我有幸在美国哥伦比亚大学完成了半年的访问学习。博士毕业后加入阿联酋起源人工智能研究院担任研究科学家。2020 年底，我回到母校武汉大学计算机学院任教，现任计算机学院教授、智能科学系主任、泰康生命医学中心 PI。主要从事可信多模态计算、联邦学习、医学 AI 方向研究。围绕“多模态可信协同机制”这一科学问题，从多模态深度语义认知、跨场景表征泛化、多中心可信协同三个方面展开研究，旨在实现“认知有深度、表征可泛化、跨域可协同”的异构多模态数据的可信协同计算，破解数据协同分析的核心难题，为社会治理和公共安全提供支撑。

您曾在香港浸会大学攻读博士，也曾任阿联酋起源人工智能研究院研究科学家、美国哥伦比亚大学访问学者，具有非常丰富的学术经历，您能分享下这些经历对您的成长有何影响吗？

这些跨地区、跨机构的经历，对我的学术格局、研究范式和做事风格都产生了非常深远的影响。最大的感受是“视野被打开了”，不同地方关注的问题不一样。在香港浸会大学读博期间，我接受了严格、系统的学术训练，培养了严谨的逻辑思维、扎实的理论推导能力和规范的论文写作习惯，为我在计算机视觉与多模态领域的研究打下了不可替代的根基。在美国哥伦比亚大学访学期间，我近距离接触全球顶尖团队的前沿探索，在自监督学习等方向拓宽了国际视野，学会从更高维度审视领域发展趋势，做高质量的研究。在阿联酋起源人工智能研究院工作时，我面对的是真实场景、大规模数据、高并发工程需求，必须把算法做稳、做快、能用，这段经历让我深刻理解学术研究不能悬空，必须面向真实问题、解决真实痛点。

您承担了多项国家和省部级重点科研项目，请问您是如何领导团队高效推进这些项目的具体实施的？

承担国家和省部级重点项目，既要保证高度，又要保证进度，更要保证成果质量。我在团队管理和项目推进上主要坚持三点：

第一是目标清晰化、任务模块化。把大项目拆解为可执行、可验收、可追溯的子任务，明确每个阶段的里程碑，让每个人都知道“做什么、做到什么程度、何时交付”，避免内耗。

第二是分工到人、权责匹配。根据我们团队内部各个学生的特长分配任务，擅长理论的做算法创新，擅长

工程的做系统实现，擅长实验的做验证分析，人尽其才、协同高效。

第三是高频同步，定期复盘。每周会开短会同步进度，遇到问题及时调整。但同时不会过度干预大家的计划和想法，让大家在自己的思路上深入做下去。

您的研究领域涉及多模态计算、联邦学习、医学人工智能等多个不同领域，可否分享下您对于研究方向选择的一些看法和建议？

我的研究以计算机视觉为入口，以多模态为桥梁，以可信、可用、安全为目标，解决真实场景里的智能问题，当然很多时候也是来源于学生自己的兴趣，对自己的研究有兴趣才能把工作做好。关于方向选择，我有几点体会想和青年学者分享：第一，不要只追热点，要追价值。热点会过去，但真问题、硬需求长期存在。第二，立足主线，适度交叉。先把一个核心领域做深做透，再自然向外延伸。第三，把个人兴趣与国家需求结合起来。人工智能发展很快，在不同的领域需求迫切，把研究扎进实际需求，更容易做出高影响力、高转化价值的成果。第四，保持耐心，长期主义。真正有分量的成果，都需要长时间积累。沉下心打磨一两项看家本领，树立标签，才能在领域里站稳脚跟。

您加入武汉大学后不久就在人才培养上成果斐然，指导学生先后获得国家自然科学基金本科生和博士生项目，请问您对招生和学生培养有什么独特的理念和方法？

在招生和学生培养上我始终坚持重内在、重成长、重实效的理念，最看重学生两点核心品质：一是好奇心，二是自驱力。专业能力与技术基础可以在科研过程中逐步培养，但如果缺少主动探索的热情、自主钻研的意愿，很难在长期科研道路上走得稳、走得远。所以我招生时不唯分数、不唯出身，更愿意选择真正热爱科研、踏实肯干、愿意主动思考的学生。

在具体培养方法上，我坚持“放手但不放任”的原则。不会一开始就给学生划定过于细致的路径，而是鼓励他们大胆尝试、主动探索，独立去思考科学问题、设计研究方案、推进课题实验等，在实践中真正锻炼科研能力。主要做好方向把关、思路引导和细节打磨，当他们遇到瓶颈时及时纠偏、提供支持，确保研究始终沿着有价值、有创新的方向推进。

我始终把解决真实问题、塑造完整科研人格作为培养目标。我希望从实验室走出去的学生，不只是能发表高质量论文，更能成长为有独立见解、有责任担当、有创造能力、能解决实际问题的优秀科研人才，把学术理想落到实处，真正为领域发展、为社会需求创造价值。

在当前论文数量激增的背景下，您是如何从中高效筛选出真正有价值的论文进行阅读？您是如何指导学生进行科技论文写作的？

现在 AI 领域论文数量爆炸，高效筛选、精准阅读非常关键。我自己筛选论文一般遵循三个标准：第一，看源头。优先阅读 CVPR 等顶会顶刊大组和大团队的文章，这些论文经过严格评审，创新度和完整性更有保障。第二，看创新。重点关注提出新问题、新框架、新范式的工作，而不是在已有方法上小修小补、刷点精度的增量工作。第三，看价值。看它是否能启发理论思考，是否能推动工程落地，是否能打开新的研究分支。

在指导学生论文写作上，我强调三点：一是逻辑至上。从问题动机、方法设计、实验验证，到结论分析，必须环环相扣、自洽可信。二是表达清晰。把复杂问题讲简单、把创新点讲明白，不堆砌术语、不故弄玄虚。三是实验扎实。充分对比、充分消融、充分可视化，用扎实数据支撑结论，让审稿人信服、让同行可复现。

大模型时代，学术界的研究受算力等多项资源限制，请问您如何看待当前学术界和工业界的关系？未来 3-5 年，您和团队计划围绕哪些研究领域重点突破？

大模型时代算力、数据、工程化门槛大幅提高，很多人觉得学术界被工业界“拉开差距”，但我认为这恰恰是学术界与工业界协同共赢的黄金时代。工业界的优势是算力充足、场景丰富、工程能力强，适合做大规模系统优化、产品落地和生态构建。学术界的优势是更自由的探索空间、更基础的理论创新、更交叉的学科视野，适合提出新思想、新范式、新框架，解决工业界来不及深入思考的底层问题。两者不是竞争关系，而是互补关系：学术界做 “从 0 到 1” 的突破，工业界做 “从 1 到 N” 的落地。

未来 3-5 年，我们团队将紧紧围绕新一代智能社会治理这一国家重大需求，在已有的多模态可信协同计算基础上，拟突破物理空间感知的局限，将计算对象由单一的个体行为延伸至复杂的社会群体认知， 数据由常规的多模态数据（视频、音频、文本）扩展至更复杂高维的异构媒体数据（如个体行为数据、政策新闻数据等）， 以实现复杂社会系统的模拟。同时，基于前期目标表征泛化和跨域协同技术，完成由被动静态分析向主动动态调控的突破，助力实现智能社会治理新手段。形成一套可模拟、可演进、可调控的智能社会治理技术体系，让 AI 从 “感知物理世界” 走向 “认知社会系统”，把科学研究和国家民生需求相结合，为社会治理和公共安全提供支撑，并在社会治理、智能家居、智慧医疗等领域实现规模化应用。

刚刚步入学术圈的一线青年教师，面临着教学和科研的多重压力，对于他们，您有什么吐露心声的建议？

学术这条路本质上是一场长跑，而不是短暂的冲刺，所以首先要稳住心态，不必过度焦虑。不用被身边人的速度带乱自己的节奏，每个人都有积累期、探索期，慢慢来、走扎实，反而会走得更稳更远。更重要的是尽早找准自己的研究主线，在属于自己的方向上长期深耕、持续沉淀，慢慢做出特色与口碑，远比四处跟风、频繁切换赛道要更有价值。

教学和科研从来不是对立的两面，完全可以相互促进。把科研最前沿的思考带进课堂，让教学更有深度；再通过授课梳理知识体系，反过来让科研思路更清晰，两者相辅相成、彼此成就。同时也一定要主动融入 CCF-CV 这样的学术共同体，多参会交流、多合作探讨、多向前辈请教，在开放的学术氛围里找方向、找机会、找伙伴。

最后也想提醒大家，无论压力多大、任务多繁重，都不要忘记最初做科研的那份初心，不要忘记身体健康的重要性。祝愿每一位青年学者都能在自己的领域里沉下心、稳步走，不负时光，终有所成。

责任编委 余烨 张青

叶茫

叶茫，武汉大学计算机学院教授、智能科学系主任、国家高层次青年人才，科睿唯安全球高被引科学家。长期从事可信多模态计算、联邦学习、医学人工智能等领域研究，以第一/通讯作者发表 CCF-A 类论文 100 余篇，谷歌学术引用 1.8 万余次。担任 CCF-A 类 IEEE TIP、IEEE TIFS 等期刊编委，CVPR、ICLR、ICML 等顶级会议领域主席等学术职务。主持国家重点研发计划、海外优青、国自科-香港联合基金、湖北省重点研发计划等科研项目。牵头获中国图像图形学会科技进步一等奖（序 1）、吴文俊人工智能自然科学二等奖（序 1）。指导学生获首届国自科博士科研基金、国自科本科生科研基金、中国电子学会优秀硕士论文、大学生“挑战杯”主体赛全国一等奖/湖北省特等奖，育人成果获 2025 年度湖北省本科教学成果特等奖。

委员好消息

- 2026 年 3 月 6 日，四川省科学技术厅关于 2025 年度四川省科学技术奖拟受理通用项目和人选的公示，CCF-CV 专委会 2 位执行委员的项目入选。电子科技大学**李永杰**等人的项目《类脑视觉计算模式、关键技术与应用》拟受理自然科学奖；同济大学**张林**等人的项目《面向智慧农业的多源感知分析与智 能决策关键技术及应用》拟受理科学技术进步奖人工智能组。四川省科学技术奖是由四川省人民政府设立的科技类奖项，由省科学技术行政部门组织实施，授予在科学发现、技术发明和科技进步等领域作出贡献的个人或组织。

- 2026 年 3 月 24 日，2026 CAAI-腾讯犀牛鸟研究计划入选名单公示，CCF-CV 专委会 3 位执行委员的项目入选。西湖大学**凌海滨**的《细粒度几何和光度增强的视觉语言模型研究》、中山大学**李冠彬**的《GUI Agent 多轮自主决策与在线强化学习训练框架》、北京大学**刘洋**的《3D 场景生成理解统一大模型技术研究》入选。CAAI-腾讯犀牛鸟研究计划由中国人工智能学会（CAAI）和腾讯公司联合发起，旨在构建强强联手的产学合作生态，共同攻克 AI 关键核心技术，并布局前沿科研方向。

- 2026 年 4 月 2 日，2025 年度辽宁省科学技术奖受理项目公示，CCF-CV 专委会 8 位执行委员的项目入选。大连海事大学**付先平**等人的项目《海洋自主作业智能机器人关键技术及应用》、大连理工大学**徐易**等人的项目《航空推进系统极端工况能量自适应协同控制关键技术与工程应用》、大连理工大学**杨鑫**等人的项目《复杂场景的感知、建模与推演关键技术及应用》获辽宁省科学技术进步奖一等奖；大连理工大学**卢湖川、王一帆、贾旭、王栋**等人的项目《结构化视觉感知与跟踪理论及方法》、西安交通大学**孟德宇**等人的项目《面向复杂环境感知增强的低秩建模与机器学习理论方法》获得辽宁省自然科学奖一等奖。辽宁省科学技术奖是辽宁省政府批准并报请省长签署的省级科技奖项，由辽宁省科学技术奖励委员会审定，旨在表彰为辽宁省科技进步和经济社会发展作出突出贡献的科技工作者及组织。

- 2026 年 4 月 10 日，中国电子学会 2025 年度会士评定结果公布。CCF-CV 专委会执行委、中山大学**操晓春**入选中国电子学会会士（Fellow of the Chinese Institute of Electronics），以肯定其牵头组织学会多场专题论坛，参与并承办青年年会，任《电子学报》领域编委；担任学会理事、青工委副主任等职，主持学术研讨与沙龙，助力学会学术交流及行业影响力提升等方面的贡献。该称号是会员的最高学术称号，授予在电子信息科技领域科学研究、促进产业发展、人才培养等方面作出杰出贡献，以及对学会事业发展作出重大贡献的个人会员。

- 2026 年 4 月 20 日，中国计算机学会关于报送第四届全国创新争先奖推荐对象的公示，CCF-CV 专委会主任中国科学院计算技术研究所**陈熙霖**被推荐拟授予第四届全国创新争先奖奖状。

- 2026 年 4 月 22 日，2024-2025 年度湖南省科学技术奖建议授奖通用项目公示，CCF-CV 专委会执行委员湖南大学**佃仁伟**入选。湖南省科学技术奖由省人民政府设立，主要奖励在科学技术进步活动中做出突出贡献的组织和个人。

- 2026 年 4 月 22 日，2026 上半年 CCF 新晋高级会员名单公布，CCF-CV 专委会 4 位执行委员入选。

西北工业大学**闫庆森**、西安电子科技大学杭州研究院**廖良**、北京科技大学**徐婧林**、中国科学院自动化研究所**朱翔显**入选。CCF 高级会员是 CCF 会员级别中较高级别的会员荣誉，授予在计算机及相关领域做出较高成就或对 CCF 服务方面有显著贡献的 CCF 专业会员。

- 2026 年 4 月 24 日，中国电子学会 2026 年第一季度高级会员评定结果发布，CCF-CV 专委会执行委员哈尔滨工业大学（深圳）**文杰**入选。中国电子学会高级会员，是为彰显在电子信息领域取得显著成绩，并对中国电子学会事业发展作出一定贡献的会员而设置的会员等级。
- 2026 年 4 月 27 日，中国图象图形学学会 2026 年上半年新晋高级会员名单公布，CCF-CV 专委会清华大学**徐静**、南京理工大学**舒祥波**、北京邮电大学**梁孔明**、天津大学**冀中**、燕山大学**顾广华** 5 位执行委员入选。中国图象图形学学会高级会员，是为了表彰在图像图形科学技术领域取得一定成就，对本学会发展有显著贡献的会员。
- 2026 年 4 月 29 日，2025 年 CSIG 科普先进工作者名单公布，CCF-CV 专委会上海交通大学**沈为**、河北大学**刘帅奇** 2 位执行委员入选。CSIG 科普先进工作者是中国图象图形学学会科普与教育工作委员会为弘扬科学精神，普及科学知识，进一步加强图像图形应用知识等宣传推广，充分调动广大研究工作者参加科普工作的积极性组织开展的一个重要评选工作。
- 2026 年 5 月 30 日，第四届全国创新争先奖评选表彰结果揭晓，CCF-CV 专委会主任、中国科学院计算技术研究所**陈熙霖**被授予第四届全国创新争先奖奖状。全国创新争先奖由中国科协、科技部、人力资源社会保障部、国务院国资委共同主办，表彰在基础研究和前沿探索、重大装备和工程攻关、成果转化和创新创业、社会服务等方面作出突出贡献的集体和个人。全国创新争先奖已成为国家科技奖励体系的重要组成部分和补充，是国家科技奖项与重大人才计划的有机衔接，是仅次于国家最高科技奖的科技人才大奖。
- 2026 年 6 月 7 日，广东省科学技术厅公示 2025 年度广东省科学技术奖拟奖，CCF-CV 专委会 3 位执行委员入选。中山大学**操晓春**等人的项目《超大规模跨模态人工智能应用平台关键技术与产业化应用》拟获科学进步特等奖；西安交通大学**孟德宇**等人的项目《疏重构理论、算法及应用研究》拟获自然科学奖二等奖；中山大学**李冠彬**拟获青年科技创新奖。广东省科学技术奖由广东省科学技术厅主办评选，主要授予为促进科技进步和经济社会发展做出突出贡献的个人或组织，是省级科技成成果奖励荣誉。
- 2026 年 6 月 12 日，北京市科学技术委员会、、中关村科技园区管理委员会公示 2026 年北京市高层次创新创业人才支持计划科技新星计划拟入选人员名单，CCF-CV 专委会执行委员中国科学院心理研究所**李婧婷**拟入选。
- 2026 年 6 月 12 日，北京市科学技术委员会、、中关村科技园区管理委员会公示 2026 年北京市高层次创新创业人才支持计划“首都高端领军人才聚集培养工程”拟入选人员名单，CCF-CV 专委会执行委员北京大学**施柏鑫**拟入选科技创新领军人才。
- 2026 年 6 月 22 日，辽宁省科学技术厅公示先进技术类计划拟立项项目，CCF-CV 专委会执行委员东北大学**张云洲**的项目《基于卫星地图的 WRJ 视觉地面目标实时定位》拟立项为应用基础研究类项目。

责任编委 徐天阳

基于扩散模型的异常检测方法及开源代码

东北大学 贾同 任天翔 王昊

无监督异常检测（UAD）是工业产品质量检测、智能制造等领域的核心技术，其仅依赖正常样本训练的特性，完美解决了工业场景中异常样本稀缺、缺陷类型多变且标注成本高昂的痛点。在无监督异常检测方法中，基于重建的方法是其中重要的组成部分，其原理在于重建模型仅能重建出正常样本，当输入测试图像时，图像被重建为正常样本，因此输入图像与重建图像产生重建误差进而检测异常。

近年来，扩散模型凭借强大的分布建模和图像重建能力，逐渐取代传统自编码器、GAN 等方法，成为异常检测领域的主流技术路线。然而，现有扩散模型异常检测方法仍存在诸多局限性：固定去噪步骤无法平衡异常重建效果与正常区域细节保留；全局加噪机制导致正常区域信息退化，难以实现异常区域的选择性修正；大尺度异常（如组件缺失、结构变形）和多类统一检测场景下性能显著下降。本文主要介绍具有里程碑意义的基于扩散模型的异常检测方法的研究成果，其重建与检测效果示例 1 所示。

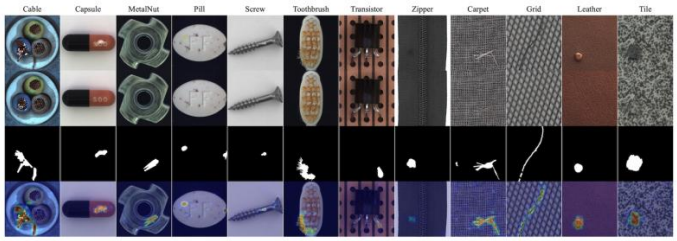


图 1 基于扩散模型的异常检测方法效果示例

1、Global and Local Adaptive Diffusion Models for Unsupervised Anomaly Detection（GLAD）

GLAD（Global and Local Adaptive Diffusion）利用全局 - 局部自适应扩散机制用于工业无监督异常检测，核心模块主要包括自适应去噪步骤（ADP）模块、异常导向训练范式（ATP）模块和空间自适应特征融合（SAFF）模块。其中，自适应去噪步骤模块通过逐步骤比较重建样本与加噪输入的特征差异，为不同严重程度的异常动态匹配最优去噪步数；异常导向训练范式模块通过引入合成异常样本并泛化扩散损失函数，使模型突破标准高斯分布限制学习异常区域的偏差模式；空间自适应特征融合模块通过生成像素级异常概率掩码，加权融合原始图像与生成图像，整个网络结构如图 2 所示。

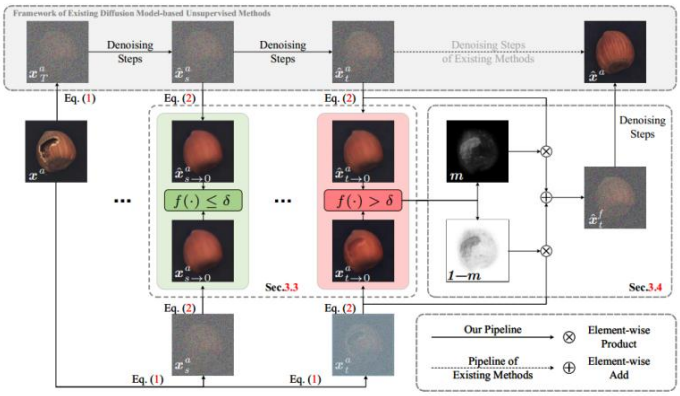


图 2 GLAD 结构图

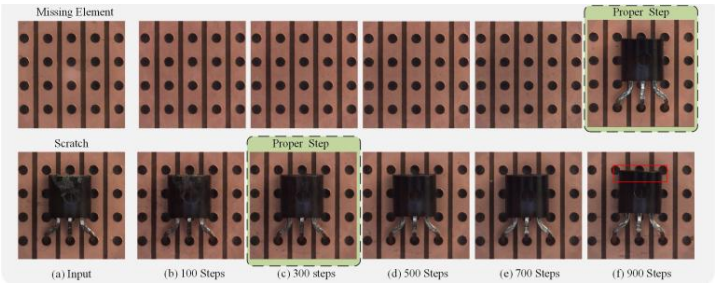


图 3 去噪步骤对重建的影响

实验结果以及图 3 均表明对于基于扩散模型的异常

检测方法，针对缺陷的大小来动态的选择去噪步骤对于整个检测流程的性能有至关重要的影响。

更多有关 GLAD 的详细内容可参考发布该方法的论文“GLAD: Towards Better Reconstruction with Global and Local Adaptive Diffusion Models for Unsupervised Anomaly Detection”。

论文链接：
https://arxiv.org/abs/2406.07487
代码链接：
https://github.com/hyao1/GLAD

2、Deviation correction diffusion（DeCo-Diff）

DeCo-Diff（Deviation Correction Diffusion）利用偏差修正扩散范式用于多类无监督工业异常检测，核心模块主要包括随机掩码前向扩散模块、偏差方向（DoD）预测模块和潜在空间编码解码模块。其中，随机掩码前向扩散模块通过训练时仅对随机选择的掩码区域加噪、保留其余正常区域不变，使模型学习利用周围正常上下文信息修正异常；偏差方向预测模块通过将扩散模型的预测目标从噪声改为偏离正态分布的方向，实现无需前置加噪、直接从原始图像潜在表示出发修正异常区域；潜在空间编码解码模块通过预训练 VAE 将图像映射到低维潜在空间，将异常建模为潜在空间中的噪声，缓解恒等映射问题并大幅降低计算量；DeCo-Diff 融合了像素 - 潜在双空间几何平均异常评分机制，从而提升多类场景下的异常检测与定位精度，该机制兼顾细微颜色异常的检测和大尺度结构异常的定位，整个网络结构如图 4 所示。

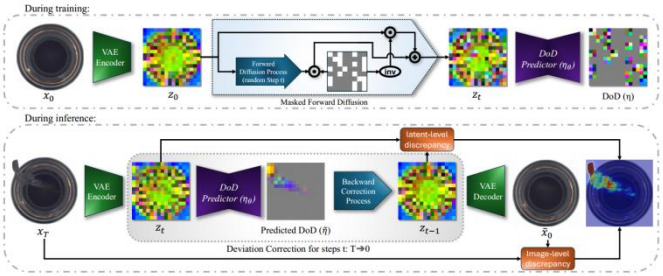


图 4 DeCo-Diff 结构图

实验结果以及图 5 均表明对于基于扩散模型的异常检测方法，在去噪过程中不对推理图像进行加噪处理而是仅仅进行对应的放缩处理，能够有效的解决多类无监督异常检测方法中的假阳性以及背景退化等问题。

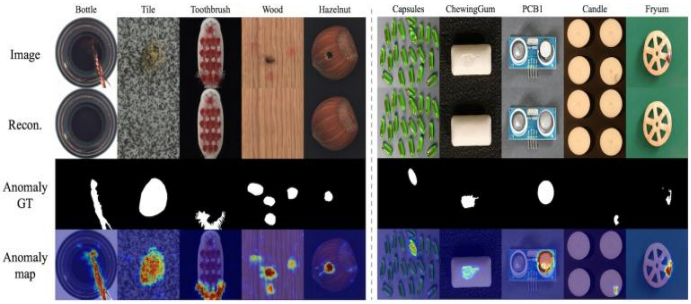


图 5 DeCo-Diff 重建结果可视化图

更多有关 DeCo-Diff 的详细内容可参考发布该方法的论文“Correcting Deviations from Normality: A Reformulated Diffusion Model for Multi-Class Unsupervised Anomaly Detection”。

论文链接：
https://arxiv.org/abs/2503.19357
代码链接：
https://github.com/farzad-bz/DeCo-Diff

3、Inversion-based Reconstruction-Free Anomaly Detection with Diffusion Models（InvAD）

InvAD（Inversion-based Reconstruction-Free Anomaly Detection）利用基于反演的无重建扩散机制用于工业和医学无监督异常检测，核心模块主要包括 DDIM 潜空间反演模块、特征空间扩散建模（FDM）模块和多步快速反演模块。其中，DDIM 潜空间反演模块通过追踪概率流常微分方程（PF-ODE）确定性轨迹，直接从输入图像推断最终扩散步的潜变量，彻底绕开传统扩散方法的显式重建过程；特征空间扩散建模模块通过预训练网络将图像映射到高维语义特征空间进行扩散建模，利用预训练特征对低阶变化的不变性提升检测精度并进一步降低计算开销；多步快速反演模块采用欧拉法近似求解 PF-ODE，仅使用 3 步反演对测试图像进行自适应加噪处理，在不损失检测性能的前提下实现了数量级的推理加速；InvAD 利用了混合异常评分机制，从而解决高维空间的反向评分问题并提升异常检测与定位的鲁棒性，该机制通过结合潜变量的负对数似然和空间欧几里得范数的绝对差异，生成细粒度的异常热力图，原理图如图 6 所示。

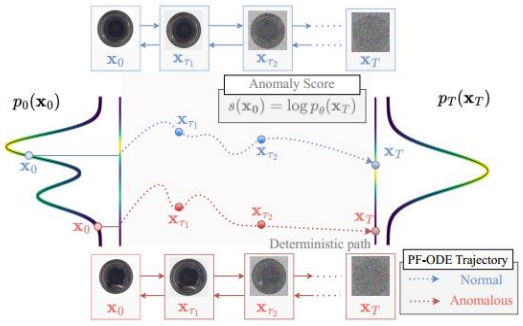


图 6 InvAD 原理图

InvAD 的方法与其他基于扩散模型的异常检测方法均不同，它丢弃了繁琐且复杂的去噪过程。利用反演过程来得到输入图像的带噪版本，并对带噪图像进行分析得到异常热力图，完美解决了扩散模型在重建过程中多步骤去噪的繁琐过程以及推理速度慢的问题。

更多有关 InvAD 的详细内容可参考发布该方法的论文”InvAD:Inversion-based Reconstruction-Free Anomaly Detection with Diffusion Models”。

论文链接：

https://arxiv.org/abs/2504.05662

代码链接：

https: //github.com/SkyShunsuke/InversionAD

责任编委 李策 樊鑫

贾 同

东北大学教授、博士生导师，机器人科学与工程学院、未来技术学院党委书记（双肩挑）。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

任天翔

博士研究生，东北大学信息科学与工程学院，研究方向为工业检测。
电子邮箱：rentx@mails.neu.edu.cn

王 昊

东北大学信息科学与工程学院特聘副研究员，硕士生导师，研究方向为物体检测、开放场景理解。
电子邮箱：wanghao@ise.neu.edu.cn

道路交通场景风格迁移数据集

西安交通大学 韩翼 李垚辰

在自动驾驶与智能交通系统快速发展的背景下，道路交通场景理解已成为计算机视觉领域的重要研究方向。传统的道路场景感知任务主要集中于目标检测、语义分割、车道线检测和可行驶区域识别等判别式任务。然而，真实交通环境具有高度复杂性：同一条道路在白天、夜晚、雨天、雾天、雪天、阴天等条件下会呈现出显著不同的视觉外观，这使得模型在跨天气、跨时间和跨域场景中的泛化能力面临严峻挑战。

道路交通场景风格迁移与图像翻译任务正是在这一背景下受到关注。该类任务旨在保持道路结构、车辆、行人、交通标志等关键语义内容不变的前提下，将图像从一种视觉域转换到另一种视觉域，例如从正常天气转换为雾天、从白天转换为夜晚、从合成场景转换为真实场景等。与普通艺术风格迁移不同，道路交通场景风格迁移更强调结构一致性、语义可靠性和真实感，因为生成图像往往会进一步服务于自动驾驶感知模型的训练、增强和鲁棒性评估。道路交通场景风格迁移任务的发展离不开高质量道路场景数据集的支撑。本文将重点介绍四个具有代表性的道路交通场景风格迁移相关数据集，分别是：ACDC（2021）、SYNTHIA（2016）、BDD100K（2020）和 VIPER 数据集（2017）。

1、ACDC 数据集

介绍：CDC（The Adverse Conditions Dataset with Correspondences）是道路交通恶劣天气场景理解领域的重要数据集，由 Sakaridis 等人于 2021 年 ICCV 上提出。该数据集的核心目标是推动自动驾驶视觉系统在恶劣视觉条件下的鲁棒感知能力，尤其关注雾天、夜晚、雨天和雪天等真实道路场景。ACDC 共包含 4006 张恶劣条件图像，并且这些图像在 fog、nighttime、rain 和 snow 四种条件之间均匀分布；每张恶劣条件图像还配有像素级语义标注、对应的正常条件图像，以及用于区分清晰区域和语义不确定区域的二值掩码。

图 1 ACDC 数据集图片示例

ACDC 对道路交通场景风格迁移研究具有重要价值。该数据集涵盖雾天、雨天、雪天和夜晚等多种恶劣天气与光照条件，能够自然构建 Normal→Foggy、Normal→Rainy、Normal→Snowy 和 Normal→Night 等道路场景外观转换任务。同时，ACDC 为每张恶劣条件图像提供了对应的正常条件图像，使研究者能够更方便地分析模型在风格迁移过程中是否保持道路结构、物体布局和语义一致性。此外，ACDC 所包含的恶劣场景具有较强现实意义，例如雾天会降低远处目标可见性，雨天会引入反光与模糊，雪天会改变道路和背景颜色分布，夜晚则带来低照度和强光源干扰。因此，该数据集常被用于验证道路交通场景风格迁移模型能否生成真实、自然且语义可靠的恶劣天气图像。

数据集地址：https://acdc.vision.ee.ethz.ch/

2、SYNTHIA 数据集

介绍：SYNTHIA（SYNTHetic collection of Imagery

and Annotations）是一个面向自动驾驶场景理解的合成道路场景数据集，由 Ros 等人于 2016 年在 CVPR 上提出。该数据集通过虚拟城市环境自动生成具有真实感的道路交通图像，并提供自动生成的像素级语义标注。SYNTHIA 的提出主要是为了解决真实道路场景中像素级标注成本高、采集难度大和场景多样性不足等问题。

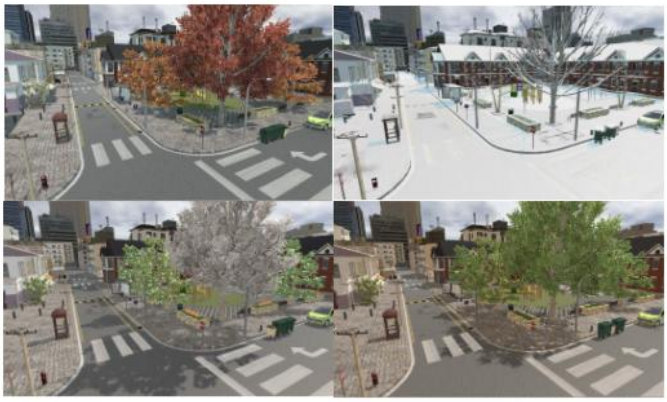


图 2 SYNTHIA 数据集图片示例

SYNTHIA 的核心特点是“可控合成”。与真实采集数据集不同，SYNTHIA 可以通过虚拟环境控制场景中的光照、天气、季节、视角和动态物体分布，从而生成不同视觉条件下的道路交通图像。根据原论文介绍，SYNTHIA 包含超过 213,400 张合成图像，既包括随机快照，也包括虚拟城市中的视频序列；图像模拟了不同季节、天气和光照条件，并提供像素级语义标注与深度信息。这种可控性使其成为道路交通场景风格迁移研究中重要的合成数据来源。

对于道路交通场景风格迁移而言，SYNTHIA 的价值主要体现在合成到真实、白天到夜晚以及跨天气外观转换等任务中。由于其图像具有清晰的结构、完整的语义标注和可控的外观变化，SYNTHIA 非常适合作为源域数据，用于研究从合成场景向真实道路场景的图像翻译。例如，研究者可以将 SYNTHIA Day→Night 作为跨时间风格迁移任务，也可以将 SYNTHIA 与 Cityscapes、BDD100K 等真实数据集结合，构建 Synthetic→Real 的域迁移任务。相比真实道路图像采集受限于天气、地点和安全条件，合成数据能够以较低成本生成大量不同视觉风格的图像，有助于系统性分析模型在不同光照、季节和天气变化下的表现，并进一步研究如何通过风格迁移缩小合成数据与真实数据之间的域差异。

数据集地址：https://synthia-dataset.net/

3、BDD100K 数据集

介绍：BDD100K 是由 Berkeley DeepDrive 团队构建的大规模真实道路交通数据集，也是自动驾驶视觉研究中最常用的数据集之一。该数据集于 2020 年发布，目标是提供一个大规模、多样化、具有真实道路环境变化的数据资源。BDD100K 包含 100,000 段驾驶视频，每段视频约 40 秒，分辨率为 720p，帧率为 30fps，并覆盖美国多个地区的真实驾驶场景。

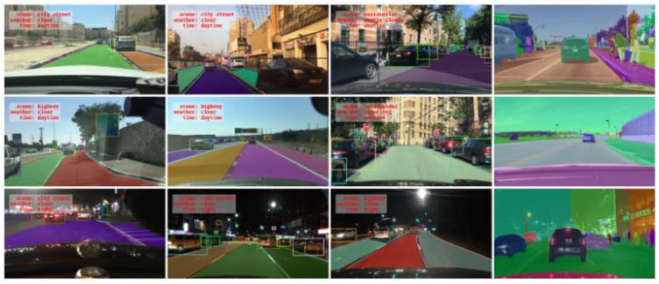


图 3 BDD100K 数据集图片示例

BDD100K 的核心优势在于规模大、场景多样且标注类型丰富。该数据集覆盖不同天气、时间和道路环境，其官方标注中包含 rainy、snowy、clear、overcast、partly cloudy、foggy 等天气属性，以及 daytime、night、dawn/dusk 等时间属性。因此，BDD100K 可以较方便地组织为 Clear→Overcast、Daytime→Night、Daytime→Dawn/Dusk 或 Clear→Foggy 等道路交通场景风格迁移任务。此外，BDD100K 还提供目标边界框、车道线、可行驶区域、语义分割、实例分割和多目标跟踪等多种自动驾驶相关标注，为下游感知任务评估提供了支持。

对于道路交通场景风格迁移研究而言，BDD100K 的重要性主要体现在真实场景多样性和应用价值上。相比 ACDC 更侧重恶劣天气，BDD100K 覆盖了更广泛的日常驾驶场景；相比 SYNTHIA 的合成数据，BDD100K 更接近真实自动驾驶系统所面对的视觉输入。因此，该数据集常被用于验证风格迁移方法在复杂真实道路环境中的适用性。同时，由于其同时具备风格相关标签和感知任务标注，研究者不仅可以评估生成图像的视觉真实感，还可以进一步分析其是否有助于提升目标检测、

语义分割或车道线检测等下游任务的鲁棒性。

数据集地址：
https://bair.berkeley.edu/blog/2018/05/30/bdd/

4、VIPER 数据集

介绍：VIPER 是 Richter 等人于 2017 年在 ICCV 提出的一个大规模合成视觉感知数据集。该数据集基于开放世界游戏环境构建，可生成道路场景视频并自动提供多种精确标注，为自动驾驶和视觉感知研究提供高质量、低成本的合成数据。

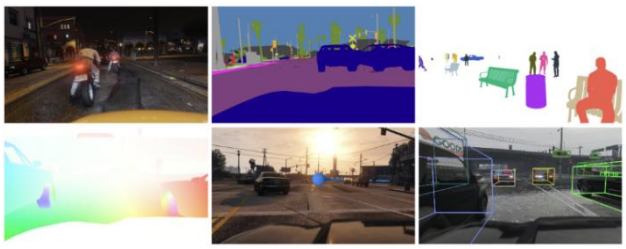


图 4 VIPER 数据集图片示例

VIPER 的核心特点是规模大、标注丰富且具有视频连续性。该数据集包含超过 250,000 帧高分辨率视频图像，数据采集过程覆盖驾驶、骑行和步行等多种运动方式，总行驶距离约 184 公里。每一帧图像均提供多种视觉任务的真值标注，包括语义实例分割、目标检测与跟踪、光流、视觉里程计以及三维场景布局等。因此，该数据集不仅可以用于静态道路场景理解，也适合用于研究视频风格迁移、时序一致性建模和跨域感知鲁棒性评估。

对于道路交通场景风格迁移与图像翻译研究，VIPER 兼具跨域和视频级应用价值。其合成道路场景可与 Cityscapes、BDD100K 等真实数据集结合开展 Synthetic→Real 图像翻译研究，同时依托连续视频帧及光流、目标跟踪等时序标注评估风格迁移的时序一致性。此外，数据集涵盖昼夜、雨雪等多种环境条件，可支持不同天气和时间场景之间的外观转换任务。凭借大规模精确标注、良好的视频连续性和丰富的多任务信息，VIPER 被广泛用于验证道路交通场景风格迁移模型的跨域泛化能力与实际应用效果。

数据集地：
https://playing-for-benchmarks.org/

责任编委 樊鑫 贾同

韩 翼

博士研究生，西安交通大学软件学院，研究方向为计算机视觉，风格迁移，图像生成，图像增强等。

李垚辰

教授，博士生导师，西安交通大学软件学院副院长。担任 CCF 智慧交通分会执行委员、CCF 多媒体专委会执行委员等。研究方向为计算机视觉与模式识别、人工智能与机器学习。
个人主页：https://gr.xjtu.edu.cn/yaochenli/

好 文 推 荐

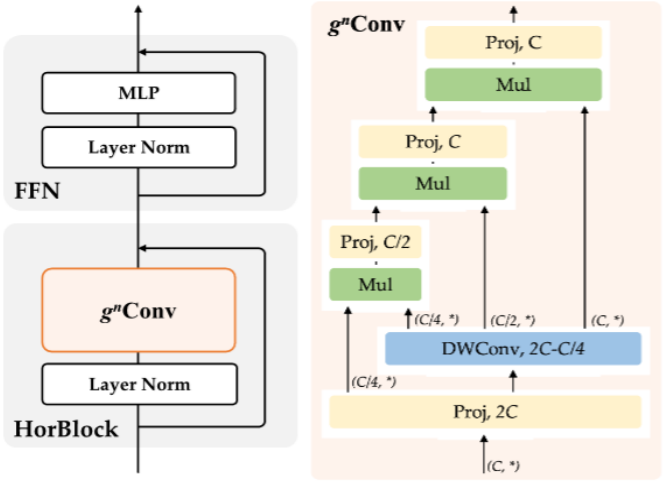
清华大学的最新成果“Efficient High-Order Spatial Interactions for Visual Perception”发表在 IEEE TPAMI 2026。

论文：Zuyan Liu, Yongming Rao, Wenliang Zhao, Jie Zhou, Jiwen Lu. Efficient High-Order Spatial Interactions for Visual Perception, IEEE TPAMI, 33-46 (2026)

自 AlexNet 问世以来，卷积神经网络（CNN）极大推动了深度学习与计算机视觉的发展。CNN 拥有诸多天然适配各类视觉任务的优良特性：平移不变性可为主流视觉任务提供有效的归纳偏置，同时保证模型在不同输入分辨率下仍具备迁移能力；高度优化的算子实现使其在 GPU 与边缘设备上均可高效运行；此外，各类架构的迭代进化进一步拓宽了 CNN 在视觉任务中的应用场景。然而，基于 Transformer 的模型架构出现后，CNN 的主导地位受到极大挑战。

但现有研究尚未从高阶空间交互视角，剖析点积自注意力在视觉任务中的建模优势。深度网络中，非线性变换会让空间两点之间产生复杂的高阶关联；自注意力与其他动态网络的成功证明：通过网络结构显式引入高阶空间交互，能够有效提升视觉模型建模能力。

本文研究发现 Transformer 取得成功的核心本质：依靠自注意力实现输入自适应、长距离、显式高阶空间交互的全新空间建模范式。如图 1 所示，该三大核心特性均可依托卷积框架高效实现，由此提出递归门控卷积（g^nCONV），同时结合门控卷积与递归结构实现高阶空间交互。不同于单纯复刻自注意力的设计，所提框架具备三项独特优势：高效性：基于卷积实现，规避自注意力二次复杂度；空间交互过程中通道维度渐进扩张的设计，能够在计算量可控的前提下实现高阶交互；可拓展性：将自注意力仅能实现的二阶交互拓展至任意阶，进一步提升建模能力；且不限制卷积空间核类型，可兼容各类大核、全局空间混合策略；平移不变性：完整继承标准卷积的平移不变性归纳偏置，适配绝大多数视觉任务，同时规避局部注意力带来的不对称建模缺陷。实验证明，g^nCONV 可作为视觉建模全新基础算子，有效融合视觉 Transformer 与卷积神经网络两者的优势。



PyTorch-style code for g^nConv

```
class gnconv(nn.Module):
    def __init__(self, dim, order, dwtype):
        super().__init__()
        self.order = order
        self.dims = [dim // 2 ** i for i in range(order)].reverse()
        self.proj_in = nn.Conv2d(dim, 2*dim, 1)
        self.dwconv = dwconv(sum(self.dims), type=dwtype)
        self.projs = nn.ModuleList(
            [nn.Conv2d(self.dims[i], self.dims[i+1], 1)
            for i in range(order-1)])
        self.proj_out = nn.Conv2d(dim, dim, 1)

    def forward(self, x):
        x = self.proj_in(x)
        y, x = torch.split(x, (self.dims[0], sum(self.dims)), dim=1)
        x = self.dwconv(x)
        x_list = torch.split(x, self.dims, dim=1)
        x = y * x_list[0]
        for i in range(self.order -1):
            x = self.projs[i](x) * x_list[i+1]
        return self.proj_out(x)
```

图 1 g^nCONV 结构流程图

责任编委 王田 李策

好 文 推 荐

东南大学的最新成果“Foundation Model for Skeleton-Based Human Action Understanding”发表在 IEEE TPAMI 2026。

论文：Hongsong Wang, Wanjiang Weng, Junbo Wang, Fang Zhao, Guo-Sen Xie, Xin Geng, Liang Wang. Foundation Model for Skeleton-Based Human Action Understanding, IEEE TPAMI, 47-61 (2026)

基于骨架的人体行为理解在机器人、人机交互、沉浸式虚拟环境、辅助助残、等领域拥有巨大应用价值。近十年来，基于深度学习的有监督骨架行为识别技术飞速发展，代表性算法包括分层循环神经网络、时空长短期记忆网络、多模态表征方法、时空图卷积 ST-GCN、双流图卷积、ShiftGCN、MS-G3D 网络、轻量化图卷积等。上述方法虽在训练集上能够取得较高精度，但泛化至全新场景、未见过行为类别时效果往往大幅下滑。此外，有监督方案依赖海量标注数据，数据采集与标注成本高昂；在数据集规模有限、样本多样性不足的场景下，模型极易发生过拟合。

为解决上述痛点，自监督骨架表征学习成为近年研究热点。现有自督方法可分为两大主流范式：掩码序列建模、对比学习。 但两类方法均存在明显缺陷：需要额外解码器、记忆库等组件，掩码策略、采样逻辑设计复杂；同时现有方案大多仅学习粗粒度行为表征，忽略了稠密预测任务至关重要的细粒度特征。

针对该现状，本文提出一套带自监督稠密表征学习能力的骨架基础模型--统一骨架稠密表征学习框架 USDRL（见图 1）。不同于对比学习、掩码序列建模两类传统自监督范式，本文采用特征解相关实现自监督行为表征学习。为学习高质量稠密表征，本文设计多粒度特征解相关模块 MG-FD，从时序、空间、实例多粒度维度对特征做解相关处理，兼顾单样本内部特征一致性与不同样本间特征区分度。同时本文搭建基于 Transformer 的骨干网络--稠密时空编码器 DSTE，该模块是挖掘多粒度特征、生成稠密表征的核心组件，能够显著提升模型稠密预测性能。稠密时空编码器包含卷积注意力（CA）与稠密移位注意力（DSA）两个子模块：卷积注意力负责捕获局部特征关联，稠密移位注意力挖掘序列中隐藏的长距离依赖。最后本文提出全新多视角一致性训练框架 MPCT，融合显式视角信息与多类型骨架模态数据，在保证计算效率的前提下优化表征学习效果。

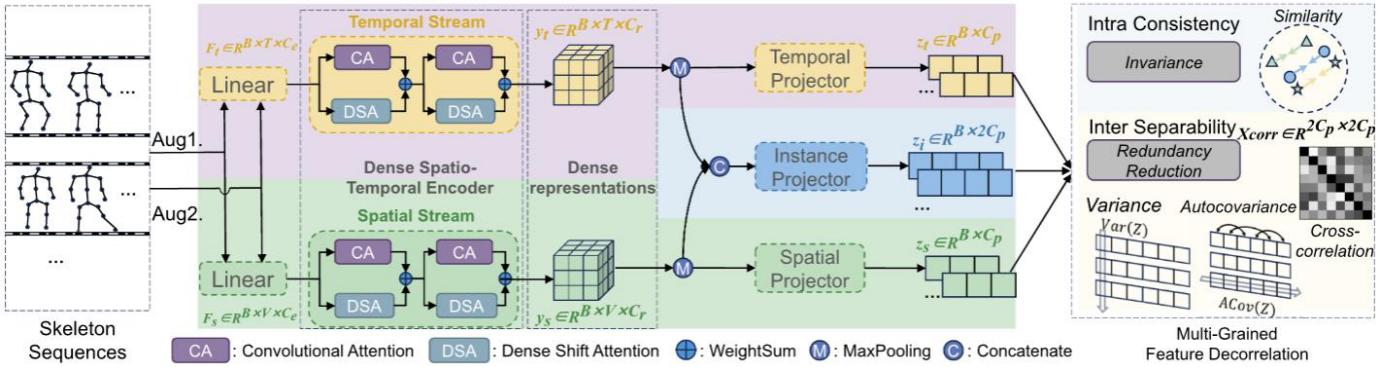


图 1 USDRL 的整体架构

责任编委 李策 樊鑫

好 文 推 荐

华中科技大学、中国地质大学（武汉）、湖北工业大学、国防科技大学和中国人民解放军总医院的“Game Theory Inspired Cross-view Interaction Alignment for Partially View-aligned Clustering”最新成果发表在 IEEE TPAMI 2026。

论文：Shanghui Deng, Xiao Zheng, Chang Tang, Lianbo Guo, Xinwang Liu, Kunlun He. Game Theory Inspired Cross-view Interaction Alignment for Partially View-aligned Clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026.

本文解决了多视图聚类中的部分视图对齐问题。在现实场景中，多视图数据常因传感器故障、时间错位或存储问题导致跨视图样本对应关系部分或完全缺失，这严重制约了传统多视图聚类方法的应用。部分视图对齐聚类旨在应对这一挑战，即在仅有少量甚至没有预对齐样本的情况下，仍能实现有效的跨视图样本匹配与聚类。然而，现有方法过度依赖预定义的对齐样本作为锚点，在完全视图错位等极端场景下性能急剧下降，缺乏动态适应能力。为从根本上突破这一预对齐依赖瓶颈，本文该方法的核心创新在于将合作博弈论引入部分视图对齐聚类问题。

具体而言，每个视图中的样本都被视为博弈中的玩家，所有样本构成一个完整的博弈联盟。在此框架下，Banzhaf 指数被用来精准衡量任意两个跨视图样本之间的协作价值——通过比较将二者作为联合体参与博弈与作为独立个体参与博弈时系统总收益的差异，差异越大则表明二者的语义关联越紧密、协作越重要。这种设计使得模型完全不依赖预定义的锚点样本，仅从数据自身的协作收益分布中就能动态发现跨视图样本间的真实对应关系。同时，本文设计了双重损失约束机制：Banzhaf 增益损失动态捕获跨视图样本对的边际贡献以增强样本相关性，对比损失则抑制弱相关样本的干扰，二者协同形成可自主学习的样本匹配框架。为进一步降低 Banzhaf 指数计算偏差，BIN 引入了跨视图融合重建表示策略，将单视图表示与跨视图表示自适应融合，在具有高语义相关性和低偏置的表示空间中确保 Banzhaf 指数的精确计算。在多个基准多视图数据集上的大量实验表明，BIN 在从 10%到 100%的各种视图错位率下均取得最优的聚类性能，显著优于现有方法，有力验证了其有效性和实用性。

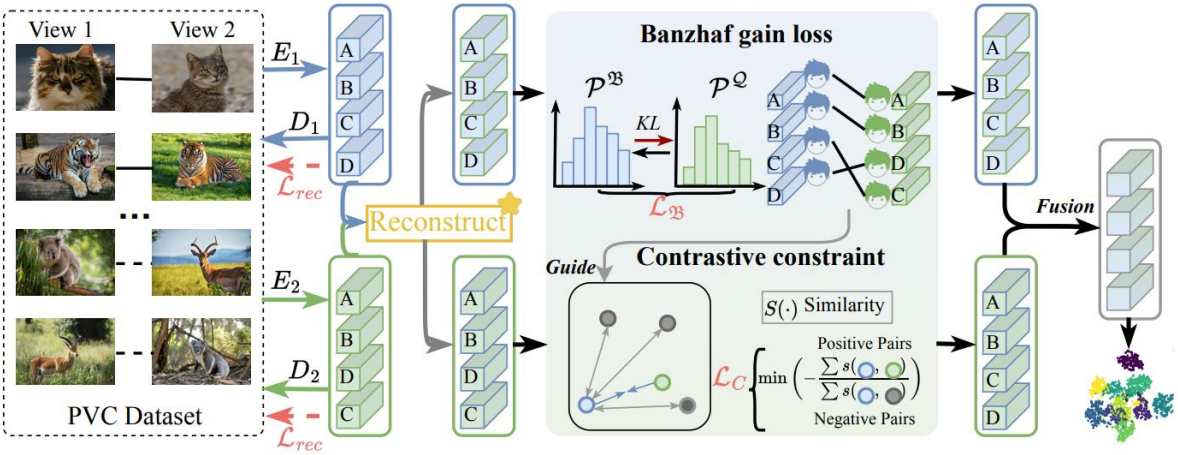


图 1 BIN 方法结构图

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 Pattern Recognition，IEEE Transactions on Multimedia 和 Computers & Graphics。

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)，由中国图象图形学学会（CSIG）、中国人工智能学会（CAAI）、中国计算机学会（CCF）和中国自动化学会（CAA）联合主办。旨在汇聚全球学者展示前沿研究成果，推动技术创新与应用发展，已进入 CCF 分区（CCF-C）。会议通常包括主旨报告、专题论坛及产业交流环节，覆盖学术研究、智能驾驶、智能制造等热点方向。

第九届 PRCV 将于 2026 年由哈尔滨工程大学承办。本届会议将秉持团结模式识别与计算机视觉领域科技工作者的宗旨，进一步推动开放合作，广泛吸引学术界和工业界的人才，提升会议的国际化水平，力求打造一个高品质的学术交流平台。大会的举办将为学术界与工业界提供更多产学研合作机会，推动模式识别与计算机视觉领域的协同创新和可持续发展。

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
ACCV 2026	2026.12.14-18	Osaka, Japan	2026-07-06	https://accv2026.org/
AAAI 2027	2026.11.23-26	Turin, Italy	2026-07-28	https://aaai.org/conference/aaai/aaai-27/
DAI 2026	26.11.29-12.02	HK, China	2026-08-04	https://www.adai.ai/dai/2026/

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
PR	New Trend of Multimodal Intelligence: Perception, Computing, Understanding, and Generation via Large AI models	https://www.sciencedirect.com/special-issue/331523/new-trend-of-multimodal-intelligence-perception-computing-understanding-and-generation-via-large-ai-models	2026.08.25
PR	Multimodal Agent AI for Video-Centric Understanding, Generation, and Embodied Intelligence	https://www.sciencedirect.com/special-issue/332957/multimodal-agent-ai-for-video-centric-understanding-generation-and-embodied-intelligence	2026.07.31
IEEE TMM	IEEE TMM Special Section on Multimodal Video Compression and Reconstruction: Theory, Algorithms, and Applications	https://signalprocessingsociety.org/events/ieee-tmm-special-section-multimodal-video-compression-and-reconstruction-theory-algorithms	2026.07.30
CG	The International Conference on Computer-Aided Design and Computer Graphics 2026	https://www.sciencedirect.com/special-issue/333475/the-international-conference-on-computer-aided-design-and-computer-graphics-2026	2026.09.10

COMPUTER VISION NEWSLETTER

02 2026

总第 48 期

计算机视觉专委会简报

CCF 计算机视觉专委会