顶会观察

CVPR 2025

清华大学 叶栩冰 唐彦嵩

国际计算机视觉与模式识别会议（CVF/IEEE Conference on Computer Vision and Pattern Recognition，CVPR）是计算机视觉和模式识别领域最重要的会议之一。CVPR 于 1983 年在美国华盛顿特区举办，每年举办一次，一般在美国举办。CVPR 2025 于 2025 年 6 月 11 日至 15 日在美国田纳西州纳什维尔的音乐城中心（Music City Center）举办。

一、 会议概况

CVPR 2025 共收到 13,008 篇有效投稿论文，经过严格评审后录用 2,878 篇，录用率 22.1%，两项指标均创历史新高。投稿作者总数超过 42,000 人，本届会议注册参会人数突破 10,000 人，其中线下参会约 9,000 人，来自全球 70 多个国家和地区。美国本土注册人数最多，中国大陆紧随其后，韩国、德国、加拿大和日本分列三至六位。

会议前两天（6 月 10–11 日）为 Tutorial 与 Workshop 时段，共组织了超过 120 场 Workshop 与 20 余场 Tutorial。主会议论文展示继续采用“口头报告+海报”双轨制：口头报告限时 8 分钟，仅少数论文入选；所有录用论文均须以海报形式展示，每轮海报展约 400–500 篇论文，持续 90 分钟。

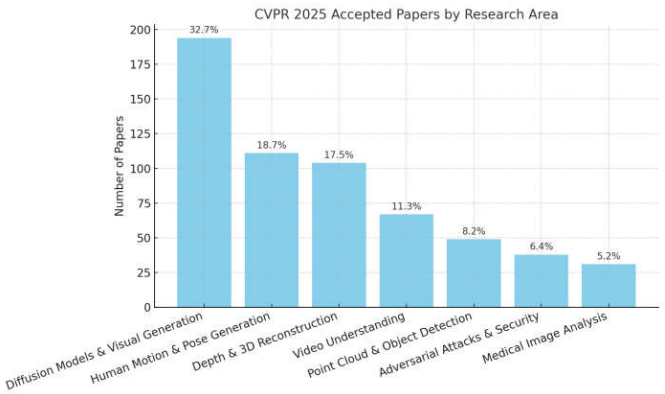

图 1　所有中稿文章的领域统计

图 2　纳什维尔会场大幅海报

CVPR 官方公布了各个细分领域的论文接收情况，如图 1 所示。可以看到，图像与视频生成领域今年度的论文接收数量最多。

根据会方统计，今年大会共收到 4 万多名作者提交的 13008 份论文。相比去年（11532），今年的投稿数量增长了 13%，最终有 2872 篇论文被接收，整体接收率约为 22.1%。在接收论文中，Oral 的数量是 96（3.3%），Highlights 的数量是 387（13.7%）。今年共有 14 篇论文入围最佳论文评选，最终 5 篇论文摘得奖项，包括 1 篇最佳论文、4 篇最佳论文荣誉提名。

此外，大会还颁发了 1 篇最佳学生论文、1 篇最佳学生论文荣誉提名。

计算机视觉技术的火热给大会审稿带来了空前的压力。本届投稿作者数量、论文评审者和领域主席（AC）数量均创下新高。今年前来现场参会的学者也超过

9000 人，他们来自 70 余个国家和地区。

二、 参会感受

飞机落地纳什维尔国际机场，海关大屏就滚动播放“Welcome Nashville”，行李转盘旁竖起了巨型霓虹吉他——会徽与田纳西音乐之城主题第一次同框。接驳大巴一路驶过百老汇大街，沿街酒吧里白天就响起乡村摇滚，司机干脆把车内音响也调到同一频道，途中已有人跟着节奏打拍子，旅途疲劳瞬间减半。

主会场 Hall A 被布置成双层同心圆：内圆 360°环形屏幕滚动播放论文短视频，外圆 380 块海报板呈放射状排布。环形顶部悬挂 36 只音箱，每到整点播放一段 AI 生成的乡村旋律——提醒换场。作者们在“舞台”中央站成一圈，观众像歌迷一样高举手机扫码、递贴纸索要签名。由于场地回音大，讨论基本靠吼，两天下来不少人嗓子沙哑，于是展台免费润喉糖成了最抢手周边。

三、 大会获奖论文

会议共选出了 5 篇论文摘得奖项，包括 1 篇最佳论文、4 篇最佳论文荣誉提名。此外，大会还颁发了 1 篇最佳学生论文、1 篇最佳学生论文荣誉提名。其中的最佳论文、1 篇最佳论文荣誉提名和 1 篇最佳学生论文如下。

最佳论文 1：VGGT：Visual Geometry Grounded Transformer[1]：在计算机视觉领域，3D 场景理解一直是极具挑战性的任务。传统方法，如通过 SfM（结构光运动恢复）、BA（光束平差法）和 MVS（多视图立体视觉）等多阶段流程实现 3D 重构，不仅步骤繁琐，计算成本高，而且耗时较长。而今年 CVPR 最佳论文中的 VGGT（Visual Geometry Grounded Transformer），作为一项开创性成果，为这一领域带来了全新的解决方案。VGGT 旨在实现 “全栈一次性 3D 场景理解”，构建了一个拥有 12 亿参数、24 层交替注意力的纯前馈 Transformer 框架。在这个创新框架下，原本复杂的 3D 重构任务得以大幅简化，仅需一次前向传播，就能在不到 1 秒的时间内，将 1 张到数百张无约束图像，同时解析成相机内外参、稠密深度、视点统一坐标系的点云，以及可跨帧跟踪的 3D 点轨迹，极大地提升了处理效率。

图 3　纳什维尔会场的 CVPR 标识

图 4　最佳论文 VGGT 的颁奖现场

在模型构建方面，VGGT 首先利用已冻结的 DINO-V2，将每张输入图像切分成 14×14 的 patch token，以此作为基础视觉单元。为了有效区分参考帧，模型特别引入了可学习的 “相机 token” 与四个注册 token 。其中，“相机 token” 专门负责学习相机相关参数，而注册 token 则用于捕捉全局场景特性。

在主干网络设计上，VGGT 采用了“逐帧自注意力 + 全局自注意力” 交替堆叠 24 次的独特架构。这种设计可谓独具匠心，一方面，逐帧自注意力能够处理每一帧图像内的 patch tokens，确保局部信息的一致性和准确性；另一方面，全局自注意力实现了不同帧间 tokens 的交互，有效整合多视角信息，从而全面理解场景。并且，这种交替设计巧妙地避免了传统交叉注意力可能导致的显存爆炸问题，同时保证了模型对任意帧排列的等变性，极大提升了模型的实用性和稳定性。

为了让模型具备强大的泛化能力，研究团队在 17 个公开的 3D 数据集上，使用 1500 万张图像对 VGGT 进行端到端训练。这些数据集涵盖了丰富多样的场景，包括室内场景（如 ScanNet、Replica）、室外场景（如 MegaDepth、Mapillary）、合成场景（如 Kubric、Objverse），以及手持设备、无人机、车载等多源采集场景 。在训练过程中，采用多任务损失函数来优化模型，它由相机 Huber 损失、深度和点云的不确定性感知 L1 及梯度损失，以及点跟踪 L1 损失共同组成，同时辅以大规模的颜色、模糊、灰度等数据增强策略，进一步提升模型的鲁棒性。研究人员还发现，VGGT 的预训练特征具有出色的通用性，可作为强大的通用 3D 先验。VGGT 为 3D 场景理解开辟了新路径，尽管存在一定局限，但无疑为该领域的后续研究提供了极具价值的参考和方向。

最佳论文荣誉提名 1： MegaSaM: Accurate, Fast, and Robust Structure and Motion from Casual Dynamic Videos [2]：在计算机视觉领域，从动态场景的单目视频里，精准、高效且稳健地估算相机参数与深度图，始终是极具挑战性的研究热点。传统的运动恢复结构（SfM）以及单目 SLAM 技术，大多依赖于静态场景，并且要求输入视频具备大量视差。一旦这些条件无法满足，例如在无约束摄像机运动、未知视野范围，或者存在动态场景干扰时，就极易导致估算结果出现偏差 。近年来，神经网络方法虽尝试打破这些局限，然而在面对复杂动态视频时，却暴露出计算量过大，或者可靠性欠佳的问题。在此背景下，研究团队创新性地提出了 MegaSaM 这一视觉 SLAM 框架，致力于攻克野外动态场景下单目视频的相机跟踪与深度估计难题 。

MegaSaM 框架的核心亮点，在于对现有深度视觉 SLAM 架构进行了深度扩展与优化 。一方面，它巧妙借鉴了 DROID-SLAM 等系统中可微分捆集调整（BA）层的优势。通过迭代更新场景几何与相机姿态变量，同时借助相机和光流监督，从海量数据中学习中间预测结果，为在复杂动态场景中实现精准的相机姿态估计筑牢根基 。另一方面，MegaSaM 开创性地将单目深度先验和运动概率图融入可微分 SLAM 范式。这种融合策略，极大地增强了模型对动态场景的适应能力，使其能够更好地应对复杂多变的实际情况 。

MegaSaM 深入剖析了视频中结构和相机参数的可观测性，并据此引入了不确定性感知的全局 BA 方案。当相机参数受输入视频的约束较弱时，该方案能够显著提升系统的稳健性，同时还达成了在测试时无需对网络进行微调，就能高效获取一致视频深度的目标 。

MegaSaM 在合成数据集与真实世界数据集上开展了大量实验。结果显示，MegaSaM 在相机姿态与深度估计的精度方面，远超先前以及同期的方法。并且，在运行时间上，MegaSaM 也表现出色，要么比其他方法更快，要么与之相当 。这充分验证了 MegaSaM 在处理无约束相机路径、复杂动态场景以及低视差视频等具有挑战性的场景时的有效性，为动态场景下的单目视觉定位与建图提供了创新的解决方案，有望推动相关领域迈向新的发展阶段 。下的单目视觉定位与建图提供了革新性的解决方案。

最佳学生论文 1：Neural Inverse Rendering from Propagating Light [3]：此论文首次提出了基于物理的多视角动态光传播神经逆渲染系统。其核心创新点在于对神经辐射缓存技术进行了时间分辨维度的拓展 。神经辐射缓存作为一种加速逆向渲染的技术，通过存储从任意方向抵达任意点的无限反射辐射来实现加速效果。研究团队创新性地将时间因素融入其中，构建出能够精确计算直接和间接光传输效应的模型 。在实际应用场景中，当面对闪光激光雷达系统捕获的测量结果时，该模型优势尽显，即便是在强间接光存在的复杂场景下，也能够实现当前最先进水平的三维重建 。

从具体实现过程来看，在模型搭建初期，系统借助先进的算法，对多视角视频中的光线传播数据进行细致分析与处理，构建起初始的光线传播模型。随后，基于拓展后的神经辐射缓存技术，模型能够持续跟踪光线在不同时间、空间维度下的传播路径与反射情况。在处理间接光时，模型通过对多次反射光线的精准捕捉与分析，有效克服了传统方法中易出现的光线信息丢失或错误计算的问题 。多视图时间分辨重新照明这一创新功能，更是允许用户在不同时间维度下，对捕获场景进行重新

照明模拟，进一步挖掘场景中的光线细节与潜在信息。

四、 总结展望

CVPR 2025 再次刷新规模与质量纪录，投稿量、录用论文、参会人数均创新高。研究主题呈现“三维化、多模态、大模型、可解释”四大趋势：NeRF 与 3D GS 继续深化，扩散模型走向高效与可控，多模态大模型参数突破百亿，同时可解释性与评测方法成为焦点。数据、算力与标注成本持续攀升，催生“开放权重+开放数据”的新研究范式；工业界与学术界合作更加紧密，现场招聘与技术 Demo 成为会议标配。如何在数据墙与资源墙日益逼近的背景下保持创新，将是 CV 社区在 2026 及以后必须回答的问题。

责任编委 张 青

参考文献

[1] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, David Novotny; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 5294-5306

[2] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, Noah Snavely; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 10486-10496

[3] Anagh Malik, Benjamin Attal, Andrew Xie, Matthew O'Toole, David B. Lindell; Neural Inverse Rendering from Propagating Light. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 10534-10544

叶栩冰

清华大学深圳国际研究生院硕士生，研究方向多模态大模型。CVPR 2025 录用论文为 VoCo-LLaMA: Towards Vision Compression with Large Language Models 和 ATP-LLaVA: Adaptive Token Pruning for Large Vision Language Models，两篇论文对视觉语言大模型进行视觉 token 剪枝和压缩，取得了显著效果。

Email: yxb_tongji@163.com

唐彦嵩

清华大学深圳国际研究生院副教授、博士生导师、科研处副处长。分别在清华大学自动化系获得工学学士和博士学位，并于英国牛津大学从事博士后工作。主要从事具身智能、计算机视觉、模式识别等领域的相关工作，以第一/通讯作者发表 TPAMI 等 IEEE 汇刊和 CVPR 等 CCF-A 类会议论文 40 余篇，主持广东省杰青、国家重点研发计划课题、中国科协青年托举工程等项目，获 2024 年公安部科学技术奖一等奖、2024 年广东省科学技术奖（科技进步）二等奖和国际顶会竞赛冠军 3 项，担任 CVPR、FG 等国际会议领域主席、国际期刊 JVCI 编委以及中国人工智能学会模式识别专业委员会（CAAI-PR）常务委员兼副秘书长等学术职务。

Email：tang.yansong@sz.tsinghua.edu.cn