计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(8):1−13 [doi: 10.15888/j.cnki.csa.009907] [CSTR: 32024.14.csa.009907]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

单目度量深度估计综述①

张久伶[1], 朱美玲[2], 徐馨润[2], 吴玉荣[2], 杨建文[2], 张秋鸿[2], 王　伟[1], 王　佳[1], 姜慧龙[1,3]

[1](中车科技创新 (北京) 有限公司 十二研究所, 北京 100083)
[2](中国科学院 软件研究所, 北京 100190)
[3](中车大连机车车辆有限公司, 大连 116045)
通信作者: 张久伶, E-mail: zhangjiuling19@mails.ucas.edu.cn

摘　要: 单目深度估计 (monocular depth estimation, MDE) 是计算机视觉中的核心任务之一, 在空间理解、三维重建、自动驾驶等领域扮演着重要角色. 基于深度学习的单目深度估计方法能够从单张图像中预测物体的相对深度, 但由于缺乏度量尺度信息, 常面临尺度不一致的问题, 限制了其在视觉 SLAM、三维重建和新视角合成等下游任务中的应用效果. 为此, 单目度量深度估计 (monocular metric depth estimation, MMDE) 应运而生, 通过对场景尺度的精确推断, 解决了深度预测中的一致性难题, 不仅显著提升了在时序任务中的深度估计稳定性, 还简化了下游任务的适配, 进一步拓展了实际应用场景. 本文系统回顾了深度估计技术的发展历程, 从传统几何方法到深度学习方法的转向, 全面梳理了该领域的技术演进及其关键突破. 在此基础上, 重点讨论了尺度不可知 (scale-agnostic) 方法在零样本 (zero-shot) 泛化中的贡献, 分析其如何为 MMDE 的进一步发展奠定基础. 本文还深入探讨了零样本 MMDE 的最新研究进展, 聚焦当前的核心挑战, 包括模型的泛化能力、边缘细节丢失等问题. 针对这些问题, 研究社区通过无标数据扩充、图像分块、模型结构优化和生成式方法等创新途径, 取得了一定进展. 本文详细剖析了这些方向的最新成果及其解决思路, 揭示了当前研究的前沿路线与技术局限. 最后, 总结了零样本 MMDE 领域内最新研究成果之间的内在联系, 梳理了尚待解决的关键问题, 并展望了未来研究方向. 通过对领域现状与发展趋势的全面分析, 旨在为研究者提供清晰的技术脉络和前沿洞察, 助力研究者更快掌握 MMDE 的研究现状, 为推动更广泛的应用和技术创新提供启示.

关键词: 单目深度估计; 度量深度估计; 深度学习; 计算机视觉

引用格式: 张久伶,朱美玲,徐馨润,吴玉荣,杨建文,张秋鸿,王伟,王佳,姜慧龙.单目度量深度估计综述.计算机系统应用,2025,34(8):1–13. http://www.c-s-a.org.cn/1003-3254/9907.html

Survey on Monocular Metric Depth Estimation

ZHANG Jiu-Ling[1], ZHU Mei-Ling[2], XU Xin-Run[2], WU Yu-Rong[2], YANG Jian-Wen[2], ZHANG Qiu-Hong[2], WANG Wei[1], WANG Jia[1], JIANG Hui-Long[1,3]

[1](12th Research Institute, CRRC Technology Innovation (Beijing) Co. Ltd., Beijing 100083, China)
[2](Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)
[3](CRRC Dalian Co. Ltd., Dalian 116045, China)

Abstract: Monocular depth estimation (MDE) is a core task in computer vision, essential for spatial understanding, 3D reconstruction, autonomous driving, and other applications. Deep learning-based MDE methods can predict relative depth from a single image, but they often lack metric information, leading to scale inconsistency and limiting their utility in

① 收稿时间: 2024-12-12; 修改时间: 2025-01-02, 2025-01-26; 采用时间: 2025-02-18; csa 在线出版时间: 2025-06-20
CNKI 网络首发时间: 2025-06-23

downstream tasks such as visual SLAM, 3D reconstruction, and novel view synthesis. To address these limitations, monocular metric depth estimation (MMDE) has been introduced. By enabling scene-scale inference, MMDE addresses depth consistency issues, enhances temporal stability in sequential tasks, and simplifies the integration of downstream applications, significantly expanding its practical use cases. This study offers a comprehensive review of the development of depth estimation technologies, from traditional geometric approaches to modern deep learning methods, highlighting key milestones and breakthroughs in the field. Special emphasis is placed on scale-agnostic methods and their role in enabling zero-shot generalization, which has been foundational for the progress of MMDE. The study also examines the latest advancements in zero-shot MMDE, focusing on critical challenges such as improving model generalization and preserving fine edge details. To address these issues, innovative solutions have been explored, including advanced data augmentation techniques, refined model architectures, and the integration of generative approaches, leading to significant advancements. This review analyzes these solutions and their contributions in depth. The study concludes by synthesizing the connections between recent achievements in zero-shot MMDE, identifying unresolved challenges, and exploring potential future directions for research. By providing an in-depth analysis of the current state of the field and emerging trends, this study aims to serve as a valuable resource for researchers, offering a clear roadmap for understanding and advancing MMDE technology across a wide range of applications.

Key words: monocular depth estimation (MDE); metric depth estimation; deep learning; computer vision

深度估计是从图像中恢复场景深度信息的关键技术, 对许多下游任务具有深远的影响. 精确的深度信息不仅在传统应用中至关重要, 如三维重建[1–3]、导航[4]、自动驾驶[5]和视频理解[6]等, 而且在新兴领域中, 如人工智能生成内容 (AI-generated content, AIGC) 中的图像[7,8]、视频[9]和三维场景[10–12]等, 也展现了巨大潜力, 彰显了深度估计在传统与现代应用中的双重价值. 早期的深度估计方法主要依赖于视差成像与立体摄影技术, 或通过双目摄像头实现精确的深度估计. 然而, 随着计算机视觉和人工智能领域的快速发展, 尤其是深度学习的崛起, 单目深度估计 (monocular depth estimation, MDE) 的概念应运而生. 与传统的双目或多目深度估计方法不同, MDE 仅依赖单张图像即可预测场景的深度, 显著降低了硬件成本和系统复杂性, 扩大了应用场景的灵活性. 近年来, MDE 领域的研究得到了广泛关注. 2025 年, CVPR 已确认将再次举办单目深度估计挑战 (the 4th monocular depth estimation challenge, MDEC), 这是继 2023 年和 2024 年后 CVPR 连续第 3 年举办该挑战. 作为计算机视觉领域最顶级的国际会议, CVPR 对 MDE 的重视进一步凸显了其在该领域中的重要性.

在 MDE 的基础上, 单目度量深度估计 (metric monocular depth estimation, MMDE) 因下游应用需求的强烈推动, 近年来取得了显著的进展. 工业界如 Intel[13]、Apple[14]、DeepMind[15]、TikTok[16,17]等公司纷纷发表了相关的研究成果. 与传统 MDE 方法仅感知无尺度的相对深度不同, MMDE 能够生成包含绝对尺度信息的场景度量深度图, 这使得 MMDE 在实际应用中的优势更加明显. 度量信息的感知不仅拓宽了方法的应用范围, 还对模型的泛化能力和预测精度提出了更高的要求, 尤其是在处理复杂场景时, 模型对图像细节的处理和尺度推断的准确性将直接影响下游任务的效果. 近年来, 得益于海量数据、强大计算资源以及模型设计的不断升级, MMDE 在零样本泛化能力、精度提升和细节保留方面取得了重要突破.

尽管在 MDE 领域已有一些综述性成果, 但这些文献[18–21]大多发布时间较早 (2020 年及之前), 或聚焦于特定的领域[22–25], 且主要讨论的是相对深度估计而非 MMDE[26–28]. 针对 MMDE 的综述性文献尚存在空白, 随着该领域的快速发展, 迫切需要对 MMDE 的已有成果和问题展开全面的综述. 此外, 我们注意到在 2024 年多个国际顶级会议 (如 CVPR 2024、ECCV 2024、NeurIPS 2024) 中, MDE 研究的两个主要趋势是: 零样本 MMDE 研究和生成式模型在 MDE 中的应用. 基于

这些背景, 本文综述了最新的研究成果, 系统地梳理了 MMDE 的核心挑战及其解决方案, 分析了生成式方法与度量深度估计相结合的挑战, 展望了 MMDE 未来的发展方向, 填补了该领域综述性文献的空白.

1 深度估计

深度估计是计算机视觉领域的核心任务之一, 旨在从二维图像中推断三维场景的深度信息, 为视觉数据赋予第三维度. 通过预测每个像素到相机的距离, 生成深度图 (depth map), 这项技术能够准确刻画场景的几何结构和空间关系, 为视觉感知和环境交互奠定了基础.

具体而言, 深度估计任务从输入的二维图像 $I := \mathbb{R}^{H\times W\times 3}$ 中预测对应的深度图 $D := \mathbb{R}^{H\times W}$, 其中每个深度值 $d_{i,j} \in D$ 表示图像像素 $i_{i,j} \in I$ 与相机间的物理距离. 然而, 这一过程本质上是一个欠定问题. 由于二维图像是三维世界的投影, 深度维度的信息在投影过程中被压缩或丢失. 因此, 在单目深度估计的场景下, 缺乏视角差异或其他辅助信息往往会导致问题解的不确定性, 增加了技术实现的复杂性.

深度估计不仅在技术挑战上具有研究价值, 更在广泛的应用领域中展现了重要意义[29]. 通过准确的深度信息, 计算机可以识别场景中物体的相对位置和距离, 为场景理解和分析提供深度支持. 在自动驾驶和机器人技术中, 深度估计显著提升了障碍物检测、路径规划和环境感知的能力, 增强了系统的可靠性和安全性. 在增强现实 (AR) 和虚拟现实 (VR) 领域, 深度信息支持精确的场景建模与交互, 大幅提升了用户的沉浸体验. 此外, 深度估计在图像后处理和计算摄影中也发挥着关键作用, 为多焦点成像、3D 视频生成和背景虚化等应用提供了技术支撑. 由此可知, 深度估计技术不仅是计算机视觉领域的重要研究课题, 还通过赋予计算机视觉系统理解和交互三维世界的能力, 为学术研究和实际应用带来了巨大的价值与潜力.

1.1 传统方法

深度学习兴起之前, 深度估计主要依赖于几何学原理或专用硬件传感器, 这些传统方法以明确的物理和数学基础为特点, 通常通过解析图像的空间关系或借助额外硬件来推断深度信息. 尽管它们在某些受控场景下表现出色, 但其在实际应用中存在诸多局限性.

● 基于硬件传感器的方法. 硬件传感器是早期深度估计的重要手段之一. Microsoft Kinect v1 是这一领域的经典代表, 其利用结构光技术, 通过向场景投射特定光图案, 分析其在物体表面上的变形来计算深度信息. 另一种常见方法是飞行时间 (time of flight, ToF) 传感器, 它通过测量光脉冲从发射到返回的时间差来推算距离. 这类方法在受控环境中能够提供较高精度, 但其硬件成本较高, 对环境光、表面反射等条件敏感, 尤其在动态或非结构化场景中适用性较差. 此外, 这类设备的复杂性限制了其在便携设备和广泛场景中的推广.

● 基于立体视觉的方法. 立体视觉 (stereo vision) 通过模拟人眼的视差原理来推算深度信息, 是另一种常见的传统深度估计方法. 它依赖两台相机从不同角度拍摄同一场景, 匹配两幅图像中的同名像素, 通过计算视差得出场景深度. 这种方法对相机的标定和图像匹配精度要求极高, 且在纹理稀疏区域、低光环境或动态场景下表现较差. 此外, 硬件成本和复杂的设置过程限制了其在许多实际应用中的普及.

● 基于几何学的多帧图像方法. 几何学方法如结构运动 (structure from motion, SfM) 和同时定位与建图 (simultaneous localization and mapping, SLAM), 通过分析多帧图像中的视差信息, 逐步重建场景的三维模型. 这些方法通常分为间接和直接两类. 间接方法依赖特征点的检测与匹配, 通过最小化重投影误差优化相机位姿和三维点云; 直接方法通过建模图像生成过程, 基于光度误差进行优化, 能够捕捉更多细节信息, 如边缘和强度变化[30]. 然而这些方法通常对场景光照条件和纹理特性敏感, 难以应对复杂动态环境或无纹理区域.

虽然上述传统方法推动了深度估计技术的早期发展, 但它们普遍依赖附加的硬件设备[31]、特定的环境条件或计算资源. 这些限制显著阻碍了技术在复杂动态场景和低成本移动设备中的应用. 深度学习技术的兴起为克服这些问题提供了新的思路. 通过学习图像中的高维特征, 深度学习方法能够以更灵活和高效的方式预测深度信息, 大幅提升了传统深度估计的鲁棒性和适用性, 为低成本、轻量化设备以及复杂环境下的应用奠定了基础.

1.2 深度学习方法

随着深度学习技术的迅猛发展, 单目深度估计领域经历了颠覆性变革. 由传统几何方法向基于神经网络的学习方法转变, 深度学习不仅为深度估计提供了全新的技术路径, 还显著扩展了其应用场景. 相比于依赖几何投影和多视角成像的传统方法, 这些基于深度

学习的技术摆脱了对特定硬件 (如立体相机或激光雷达) 及严格校准条件的依赖, 通过神经网络直接从单张图像中推断深度信息, 大幅降低了实现复杂性和成本. 这一突破为低成本、高灵活性的深度估计解决方案开辟了新路径, 尤其适用于移动设备上的增强现实 (AR)、无人机导航等场景.

深度学习的核心优势在于其从大规模数据中提取场景先验知识的能力, 从根本上弥补了深度估计中的欠定问题. 在传统几何方法中, 单张图像缺乏深度维度的信息, 仅凭二维像素分布难以直接推断三维场景结构. 深度学习通过神经网络捕捉图像中的全局和局部特征, 包括纹理、形状和语义信息, 能够间接学习场景的三维属性. 通过这种方法, 网络不仅能够识别显著物体及其相对位置关系, 还可以基于场景上下文对模糊区域进行合理推测. 例如, 远处的天空和地平线通常对应较大的深度值, 而地面的纹理变化则可以提供深度梯度分布的线索. 这种能力使得深度学习模型在没有几何信息支持的情况下, 依然能够生成高质量的深度预测.

深度学习方法的多层次特征表示能力是单目深度估计取得显著进展的关键. 以卷积神经网络 (CNN) 为核心的架构能够提取图像中多尺度的特征, 从低层纹理信息到高层语义信息均被有效利用. 通过融合这些不同尺度的特征, 深度学习模型可以同时考虑图像的局部细节和全局结构, 从而生成更为精确的深度图. 例如, 在处理建筑场景时, CNN 能够捕捉墙面纹理的局部变化, 同时识别整体几何布局. 这种基于特征层次的分析相比于传统方法依赖像素级几何计算, 极大提升了深度估计的精度与鲁棒性.

此外, 深度学习方法在复杂场景中的适应能力显著增强. 通过对纹理稀疏区域和非规则场景的特征建模, 神经网络能够有效解决传统方法在这些场景下的表现瓶颈. 例如, 在自动驾驶领域, 深度学习模型可以准确识别道路、行人、车辆等多种对象并估计其深度分布, 从而为路径规划和障碍物检测提供关键支持. 同时, 在机器人导航中, 基于深度学习的单目深度估计方法以更低的硬件要求实现了对动态环境的高效感知, 为更广泛的应用场景奠定了基础.

2 单目深度估计

单目深度估计是深度学习在深度估计领域的一个典型应用方向, 其目标是在不依赖多视角图像或复杂硬件的情况下, 仅通过单张 RGB 图像直接推断场景的深度信息. 与传统多帧深度估计方法相比, MDE 通过深度学习模型提取图像特征进行推断, 显著简化了实现难度并降低了成本. 早期研究主要基于监督学习框架, 利用带深度标注的数据集训练神经网络模型.

2014 年, Eigen 等[32]提出了一个多尺度卷积神经网络框架, 通过联合预测全局和局部深度信息, 显著提高了单目深度估计的精度. 这一框架将图像分解为粗略深度图和局部细节图, 并结合多层次特征信息, 为后续研究奠定了基础. 2015 年, Eigen 等[33]进一步扩展了这一框架, 提出了一个多任务学习模型, 同时预测深度、表面法线和语义标签. 通过共享特征提取部分, 该模型不仅提升了深度预测的精度, 还减轻了单任务模型的过拟合问题, 成为多任务联合学习领域的重要进展.

随着深度学习特别是卷积神经网络 (CNN) 的发展, MDE 的网络结构逐渐演变为编码器-解码器框架. 编码器用于捕获图像的全局上下文信息, 而解码器则通过逐步上采样生成高分辨率的深度图. 结合多尺度特征融合的方法进一步增强了网络对不同深度层次的捕捉能力. 例如, 研究者设计了跨尺度损失函数, 通过约束局部和全局深度一致性来优化模型性能, 从而显著提升了深度估计的效果.

针对 MDE 的欠定性问题, 一些研究引入了几何先验作为额外约束, 例如透视线索、物体大小等. 这些方法将几何知识与深度学习的特征学习能力结合, 改善了深度估计的合理性, 同时增强了模型在特定场景中的泛化能力. 通过这种方式, 模型能够更准确地理解场景的三维结构, 特别是在纹理稀疏或复杂环境中展现出更高的鲁棒性.

尽管早期尝试主要集中于特定领域 (如室内或室外场景) 进行深度值回归, 这些方法在单一数据集上取得了成功, 但其泛化能力有限. 当模型应用于新的场景时, 往往因缺乏跨领域适应性而产生显著偏差. 为了解决这一问题, 研究者开始探索更加鲁棒的网络结构和优化方法. 例如, 通过构建更具普适性的特征提取模块, 以及在模型训练中引入对领域不变特征的约束, 逐步提高模型的泛化性能, 为单目深度估计在多样化应用场景中的部署奠定了基础.

3 零样本单目深度估计

零样本 (zero-shot) 单目深度估计的研究源于对模

型泛化能力的迫切需求. 在早期的深度估计研究中, 模型通常直接回归度量深度值. 这种方法在训练数据与测试场景高度一致时表现良好, 但由于深度度量强烈依赖于特定场景的尺度和相机参数, 模型的迁移能力受到严重限制. 这一局限性使得模型在跨场景应用中表现不佳, 难以满足实际需求. 为克服这一挑战, 研究者提出了一种新的方向, 即通过问题简化提升模型的零样本能力.

相对深度估计 (relative depth estimation, RDE) 帮助零样本领域实现重要突破. 与度量估计不同, RDE 不再直接预测像素的真实度量深度, 而是专注于像素之间的相对深度关系. 通过消除对绝对尺度信息的依赖, RDE 关注像素间深度的排序, 显著增强了模型的泛化性. 这种方法通过引入尺度无关 (scale-agnostic) 和尺度-位移无关 (scale-and-shift-invariant) 的损失函数, 使得模型能够在异构数据集上进行训练并适应多样化的场景. RDE 的这种特性使其在零样本设置下对新领域展现出较高的适应能力.

MiDAS 是零样本深度估计领域中里程碑式的成果, 首次提出了具有普适性的零样本深度估计概念[34]. MiDAS 通过在多个异构数据集上联合训练, 并结合尺度-位移无关损失, 大幅提升了模型的跨域预测能力. 随着版本的迭代, MiDAS 从早期基于卷积的架构演变为采用 vision Transformer 的现代网络结构[35]. 在捕捉全局信息和多尺度特征融合方面, vision Transformer 提供了显著优势, 使得 MiDAS 在多样化场景中的适应能力进一步增强. 虽然 MiDAS 的预测结果仅是相对深度, 但其创新性架构和训练策略为后续研究奠定了重要技术基础.

尽管 RDE 通过问题简化提升了模型的迁移性, 但其本质是一种妥协策略. 这种方法放弃了绝对尺度信息, 将深度估计从复杂的度量问题转化为相对简单的排序任务. 这种简化虽然增强了模型的跨域适应性, 但也带来了明显的局限性. 例如, 缺乏绝对深度信息使 RDE 难以适用于需要精准三维数据的任务, 如 SLAM、增强现实 (AR) 和自动驾驶. 此外, RDE 的结果由于缺乏统一的尺度标准, 在连续帧场景中的应用容易出现抖动, 影响预测结果的稳定性.

综上所述, 零样本深度估计通过创新性思路开辟了新的研究方向, 为解决深度估计中的核心问题提供了宝贵经验. 当前研究正在积极探索如何结合相对深度估计和绝对深度估计的优势, 以实现既具跨域泛化能力, 又能够提供尺度信息的单目深度估计方法, 从而满足更广泛的实际需求. 这一方向的进展将为 SLAM、无人驾驶及其他需要高精度和高鲁棒性的场景带来全新的解决方案.

4 单目度量深度估计

度量深度估计研究在下游应用需求的强烈驱动下重新受到关注, 成为深度学习领域的重要方向. 三维重建、新视角生成、SLAM 系统等实际场景对高精度几何信息的需求与日俱增, 而传统深度估计方法由于仅能提供相对深度值, 难以满足动态场景中帧间一致性和几何稳定性的要求. 此外, 模型结构革新 (如 vision Transformer 的引入)、模型规模扩大 (million 到 billion), 以及标注深度数据集的爆发式增长 (百万量级), 共同推动了研究社区重拾这一方向. 与早期深度学习方法局限于领域内的度量深度估计不同, 此次“复兴”更加注重于零样本单目度量深度估计 (zero-shot monocular metric depth estimation) 的能力, 即无需依赖训练阶段的特定相机内参或深度标注数据, 也能在未知场景中生成一致的度量深度值.

度量深度估计能够输出具有物理量纲的绝对深度值, 显著提升了跨场景应用的适应性. 模型不仅能够在室内外等多领域表现出色, 还在动态场景中提供帧间一致的深度信息, 极大满足了实际需求. 然而, 这一领域的核心挑战在于如何在未知场景中实现泛化, 同时确保几何稳定性和空间一致性.

早期的度量深度估计方法多假设相机内参已知. 比如, Metric3D 通过将图像和深度图映射到规范空间, 并依据焦距重新调整深度值, 解决了不同相机设置下的尺度和偏移问题[36]; ZeroDepth 则通过变分框架学习相机特定嵌入. 这些方法在一定程度上提升了性能, 但对精确相机内参的依赖限制了泛化能力[37]. 研究逐步向未知相机参数的情况下推进. 例如, 部分方法通过单独的网络预测相机嵌入, 或直接在球形空间中进行深度预测, 绕过对传统内参的需求[38].

随着研究的深入, 度量深度估计的方法从单一全局深度分布学习逐步转向细粒度的自适应深度区间划分策略. AdaBins 通过引入自适应机制, 根据图像内容动态调整深度区间的位置和分布, 显著改善了深度差异较大场景的表现[39]; LocalBins 进一步细化了这一策

略, 通过将图像划分为多个局部区域, 并在局部区域内学习深度分布, 从而提升复杂场景中的深度估计精度[40]. 但这种局部自适应策略的计算复杂度较高, 推理时间也随之增加. 基于 Transformer 架构的 BinsFormer 则在端到端训练中结合全局与局部信息优化了自适应深度区间划分, 不仅提升了模型对全局上下文信息的捕捉能力, 也显著提高了估计精度[41]. 此外, New CRFs 通过结合神经网络与条件随机场 (CRFs), 在深度估计中引入全局优化机制, 通过全连接的 CRFs 确保像素间的关联性, 从而大幅提升深度预测的全局一致性与对不确定性的处理能力[42].

ZoeDepth 的提出标志着零样本度量深度估计迈入新阶段[13]. 该模型结合 MiDAS 与自适应深度区间方法, 通过引入轻量化的度量区间模块与动态调整机制, 实现了更加精确的度量深度估计, 同时能够自动分类输入图像并路由到合适的网络头, 在室内外场景中均表现优异. ZoeDepth 基于多数据集联合训练, 结合少量微调, 显著提升了跨领域泛化性能. 其多样化的训练数据与统一架构设计, 使其在多个室内外数据集上的零样本测试中均取得了出色表现, 为度量深度估计的进一步研究奠定了坚实基础.

5 主要挑战与改进方法

尽管零样本单目度量深度估计已取得显著进展, 但仍存在诸多亟待解决的问题, 其中最核心的挑战是模型的泛化能力仍有较大提升空间. 在未知场景中, 如何确保深度估计的精度和稳定性, 仍是当前研究的关键难题[43]. 此外, 单次推理模型面临诸多实际问题, 例如边缘平滑导致的几何细节模糊、场景细节丢失、高分辨率输入的适配性不足等. 这些问题直接影响模型在实际应用中的表现和鲁棒性. 为应对这些挑战, 研究社区提出了多种改进方向, 并在这些方向上涌现了一系列具有代表性的最新成果. 本节将对相关研究进展进行总结和讨论.

5.1 泛化性问题与改进方法

为提升零样本度量深度估计 (MMDE) 的泛化能力, 研究主要集中在数据扩充与模型方法改进两个方向. 数据扩充旨在通过更丰富的训练数据和优化的学习策略, 提高模型应对复杂场景的能力; 模型方法改进则聚焦于设计更鲁棒的架构与算法, 从根本上增强模型的跨域适应性和预测精度.

在数据扩充方面, Depth anything 提出采用半监督自学习策略的大规模数据扩充方法, 通过生成约 6 200 万张自标注数据, 显著提升了模型的泛化能力[16]. 该方法结合优化的训练策略, 推动模型在更复杂多样的场景中主动学习额外的视觉知识. 同时, 通过辅助监督机制, Depth anything 充分利用预训练编码器中丰富的语义先验, 显著降低了泛化误差. 这种方法在室内和室外场景中的零样本深度估计能力均表现优异, 展示了大规模自标注数据在构建强大单目深度估计基础模型中的潜力[44–47].

在模型方法改进方面, UniDepth 提出了一种直接预测度量 3D 点云的创新深度估计方案, 无需依赖额外的相机参数或元数据[48]. 其核心是通过自提示相机模块 (self-promptable camera module) 生成稠密的相机表示, 并采用伪球面输出表示解耦相机和深度特征, 从而增强模型对相机属性的抗噪性. 借助几何不变性损失, UniDepth 显著提升了深度特征的鲁棒性和泛化性能. 此外, 通过相机自举 (bootstrapping) 和显式标定相机内参的双重机制, UniDepth 实现了更稳定、精准的深度估计. 其解耦相机参数与深度特征的学习策略, 为提高 MMDE 模型的泛化能力提供了强有力的技术支持.

5.2 细节丢失与边缘平滑问题

细节丢失和边缘平滑是深度学习模型在预测稠密信息时的普遍问题, 这种现象在深度估计、图像分割和目标检测等领域都广泛存在. 它使得深度回归过程中容易忽视图像中的细节, 尤其是物体边缘和精细结构 (如头发、毛发等), 从而导致生成的深度图缺乏足够的精确度, 难以准确反映图像中的真实几何特性. 这种不足在复杂场景中尤为显著, 例如物体的遮挡边界或结构复杂的区域, 限制了模型在高精度应用中的实际表现. 此外, 如何在算力可承受范围内有效处理高分辨率输入也是一个共性挑战. 深度估计模型往往在高分辨率场景下难以同时兼顾全局一致性和局部细节表达, 生成的深度图可能缺乏精细度, 进一步加剧了遮挡轮廓和复杂边界区域的模糊化问题. 为应对这些挑战, 研究者提出了多种解决方案. 例如, SharpNet 通过引入法线和遮挡轮廓约束提升边缘锐利度, 但这类方法需要额外的监督信号, 增加了训练的复杂性. 而另一种方法 BoostingDepth 则通过将低分辨率网络独立应用于图像块实现细节增强, 但由于图像块缺乏全局上下文, 该方法需要复杂的多步骤融合管道, 带来了额外的计算成本

和实现复杂性[49].

面对这些问题, 研究社区围绕细节丢失提出了 3 种主要的改进方向, 力图在保持高效性的同时提升深度估计的精细化水平.

5.2.1 Patch 方法

为了解决细节丢失和边缘平滑的问题, 研究者提出了一系列基于图像分块 (patch-based) 的方法, 这些方法通过局部深度估计与全局上下文的融合, 有效提升了深度图分辨率与细节表现, 同时展示了在处理复杂场景中的独特优势.

PatchFusion 是对 BoostingDepth 思路的一种全新拓展, 其核心在于利用内容自适应的多分辨率合并技术来提升单目深度估计的分辨率[50]. 具体来说, PatchFusion 将图像划分为多个块 (patch), 独立进行深度估计, 再通过全局到局部 (global-to-local, G2L) 模块结合全局上下文信息进行一致性融合. 同时, PatchFusion 提出了“一致性感知训练与推理策略 (consistency-aware training and inference, CAT & CAI)”, 从几何和色彩一致性角度优化了拼接区域. 然而, 该方法需要多个步骤完成推理, 包括下采样、深度估计、拼接对齐等, 推理时间较长. 此外, 由于图像块的局部性, PatchFusion 在某些场景中可能会误将局部内容当作实际深度, 导致全局场景理解上的偏差. PatchRefiner 在 PatchFusion 的基础上进一步优化, 将高分辨率深度估计重新定义为一种细化过程[51]. PatchRefiner 通过提出“细节与尺度分离损失 (detail and scale disentangling, DSD)”, 在增强边缘锐利度的同时, 确保了深度尺度的准确性. 该方法特别针对高分辨率场景的挑战, 引入基于合成数据的伪标签策略, 构建了真实数据与合成数据间的知识迁移桥梁. 相比 PatchFusion, PatchRefiner 的模块化设计简化了多步骤管道, 显著提升了推理效率, 并且在深度图的细节捕捉和全局一致性上表现更为出色.

Depth pro 则将重点放在实际应用中的推理效率以及细节保留能力上[14]. 通过集成高效的多尺度视觉 Transformer (ViT) 和真实与合成数据的联合训练协议, Depth pro 在快速推理的同时保持了细节保留的能力. 其创新的切片方法利用少量重叠的图像块, 既保证了局部细节的完整性, 又有效缓解了 PatchFusion 和 PatchRefiner 方法中上下文丢失的问题. 此外, Depth pro 支持直接从单张图像推断绝对深度, 免去了对相机内参的依赖, 大幅降低了方法的复杂性. 然而, Depth pro 的设计更关注近距离物体的深度预测性能, 对于场景整体一致性和远距离深度估计的效果略显不足.

5.2.2 合成数据集微调方法

现实世界的深度监督数据集通常存在一些质量问题, 例如标签不准确、高光或透明区域的深度标签缺失、深度值与场景不匹配, 以及物体边界定义失败. 这些问题源于数据采集流程及标注方法的局限性, 是导致模型细节丢失的重要原因. 相比之下, 合成数据集通过视觉渲染引擎直接生成 3D 几何信息, 可以提供像素级、精确的深度监督标签. 这些标签不仅细致全面, 还覆盖了现实数据难以应对的挑战性条件, 如高反射和透明区域[51].

Depth anything V2 利用了合成数据的这些优势, 提出了一种以合成数据为核心的深度估计模型优化策略[17]. 这一方法突破了对现实数据的依赖, 通过高质量的合成数据支持更细粒度和多样化的场景训练, 避免了现实数据采集中常见的伦理和隐私问题, 并且能够快速扩展数据容量. 然而, 合成数据与真实数据在颜色分布和场景布局上存在显著差异, 这种分布偏移可能限制模型的泛化能力, 特别是在合成数据中未覆盖的场景条件. 为解决这一问题, Depth anything V2 设计了一种伪标注策略, 通过训练教师模型为真实数据生成伪标签, 以缩小合成数据和真实数据之间的分布差异. 同时, 该模型引入了梯度匹配损失函数, 以增强深度图的锐度并缓解细节丢失现象. 这种损失函数特别适用于合成数据的训练, 能够忽略训练中损失最大的区域, 从而避免模型过拟合到少量难以对齐的样本. 尽管如此, 合成数据的局限性依然存在. 由于图形引擎生成的场景覆盖范围有限, 模型在测试未见过的真实场景时可能表现不佳.

5.2.3 生成式方法

生成扩散模型近年来在解决单目深度估计中的边缘平滑和细节丢失问题方面取得了显著的进展. 这类模型通过模拟图像的退化过程, 在生成过程中逐步恢复深度信息, 展现出强大的细节还原能力[52–55]. 例如, Marigold 首次将扩散模型的潜力引入深度估计领域, 其生成的深度图在结构一致性和细节丰富度上远超过传统判别模型[55]. 尤其是在处理高反射或透明区域时, Marigold 的表现尤为出色. 然而, 由于多场景复杂布局中的歧义性问题, Marigold 的生成深度图可能会出现意外的布局错误, 限制了其在更复杂场景中的应用.

针对这些局限, GeoWizard 对 Marigold 的方法进行了重要改进[56]. GeoWizard 引入了一种新的 Decoupler 结构, 将不同场景的分布分离进行学习, 从根本上减少了混合分布带来的模糊和歧义. 此外, 通过嵌入一维场景分类向量 (如室内、室外和物体场景), GeoWizard 显著增强了模型在多场景中的泛化能力. 这种改进使得 GeoWizard 在复杂场景中的深度估计表现更加稳健, 特别是在前后景区分和室外场景的几何布局预测方面. 例如, 它能够有效避免前景元素被错误压缩的问题, 并更准确地还原场景中的三维几何结构. GeoWizard 还结合估计的法线图与 BiNI 算法生成的伪度量深度, 实现了表面重建, 从而生成更加合理的 3D 几何形状.

与此同时, DeepMind 提出的 DMD (diffusion for metric depth) 进一步推动了扩散模型在度量深度估计中的应用[15]. DMD 的核心创新在于采用对数尺度深度参数化策略, 从而解决了室内外场景中深度表示能力分配不均的问题. 通过引入视场 (FOV) 条件, DMD 还解决了因摄像机内参差异引起的深度尺度歧义. 在训练阶段, DMD 通过裁剪和噪声填充来模拟不同视场的训练数据, 并将垂直 FOV 作为条件输入, 以提高模型对多样化场景的适应性和预测精度. 此外, DMD 采用 ε-参数化以显著加速推理, 仅需一步去噪即可完成深度估计, 提升了推理效率.

5.3 方法对比与问题

在深度估计任务中, 单次推理方法因其快速高效的特点依然是研究社区的主流方向. 这类方法通过一次模型推理即可完成深度预测, 特别适合对实时性要求较高的场景, 如交互式视图合成和实时导航. 然而, 由于缺乏对高频细节的精确捕捉能力, 这些方法在处理毛发、建筑纹理或肢体等复杂结构时表现不足, 生成的深度图细节往往不够丰富. 同时, 这些方法对大规模高质量标注数据的依赖较强, 而现实中的数据标注质量参差不齐, 使得模型在泛化能力上受到限制.

与单次推理相比, 基于分块 (patch) 的推理方法通过将输入图像划分为多个小块 (patch), 逐块完成深度估计并融合结果, 从而在细节表现上取得了显著改进. 通过增加 patch 的数量, 模型可以进一步提升分辨率和预测的精确度, 同时具备潜在的并行化推理能力. 然而, 这类方法存在效率问题: 推理时间随着 patch 数量的增加呈线性增长, 且性能提升逐渐趋于饱和. 例如, Patch-Fusion 虽然在 patch 权重融合上有显著优化, 但其较慢的推理速度难以满足实时需求. 此外, 多次推理的特性使其在高分辨率场景中应用面临瓶颈.

生成式扩散模型为解决深度估计中的细节丢失问题提供了全新思路. 这类模型通过模拟图像退化过程, 逐步生成细腻的深度图, 在捕捉图像内在结构和复杂场景几何关系上表现卓越. 例如, Marigold 在复杂场景中能准确理解墙壁、家具和装饰物之间的几何关系, 生成的深度图细节丰富且自然. 然而, 扩散模型的多步推理特性带来了效率问题: 每次推理都会引入一定随机性, 从而增加生成结果的不确定性. 此外, 这类方法在度量深度估计 (MMDE) 方面的研究较少, 目前大多停留在相对深度估计 (RDE), 限制了其面对下游任务的实用价值.

此外, 生成式方法在数据依赖性上具有显著优势. 与判别式方法需要大量标注数据不同, 生成式扩散模型在无标注或少标注场景下也能取得良好的性能. 例如, DMD 通过引入视场条件和对数尺度深度参数化, 显著提高了零样本场景下的深度估计能力. 这种创新方法能够高效处理多样化场景, 并在无标注条件下实现卓越的深度预测. 然而, 多次推理和去噪步骤的复杂性使得生成式扩散模型难以满足实时性的要求, 仍需进一步优化. 图 1 以定性方式, 直观对比了室外/室内、街道/建筑、大/小场景、城市/自然、明/暗光照等场景下模型的深度图输出结果, 用不同颜色标记出了场景类型和方法类别. 其中, 生成式方法的输出为相对深度, 其余方法输出均为绝对深度值. 可以看到 Patch 方法和生成式方法有效缓解了单次推理方法的细节丢失和边缘平滑问题.

图 2 对比了单次推理模型、Patch 分块模型和生成式模型的典型方法在推理时间和内存开销上的对比, 测试基于 400 帧的 1080p 视频在 RTX 3090 上进行推理实验, 注意横轴时间开销以 s 为单位, 呈指数增长.

综合来看, 单次推理方法以高效和实时性为核心优势, 但在细节建模上仍有不足; 基于分块的推理方法在分辨率和细节表现上优于单次推理, 但推理效率随着 patch 数量增加显著下降; 生成式扩散模型则凭借卓越的细节还原能力和复杂场景几何理解能力成为深度估计的有力工具, 但其多步推理的效率瓶颈限制了实时应用的可能性.

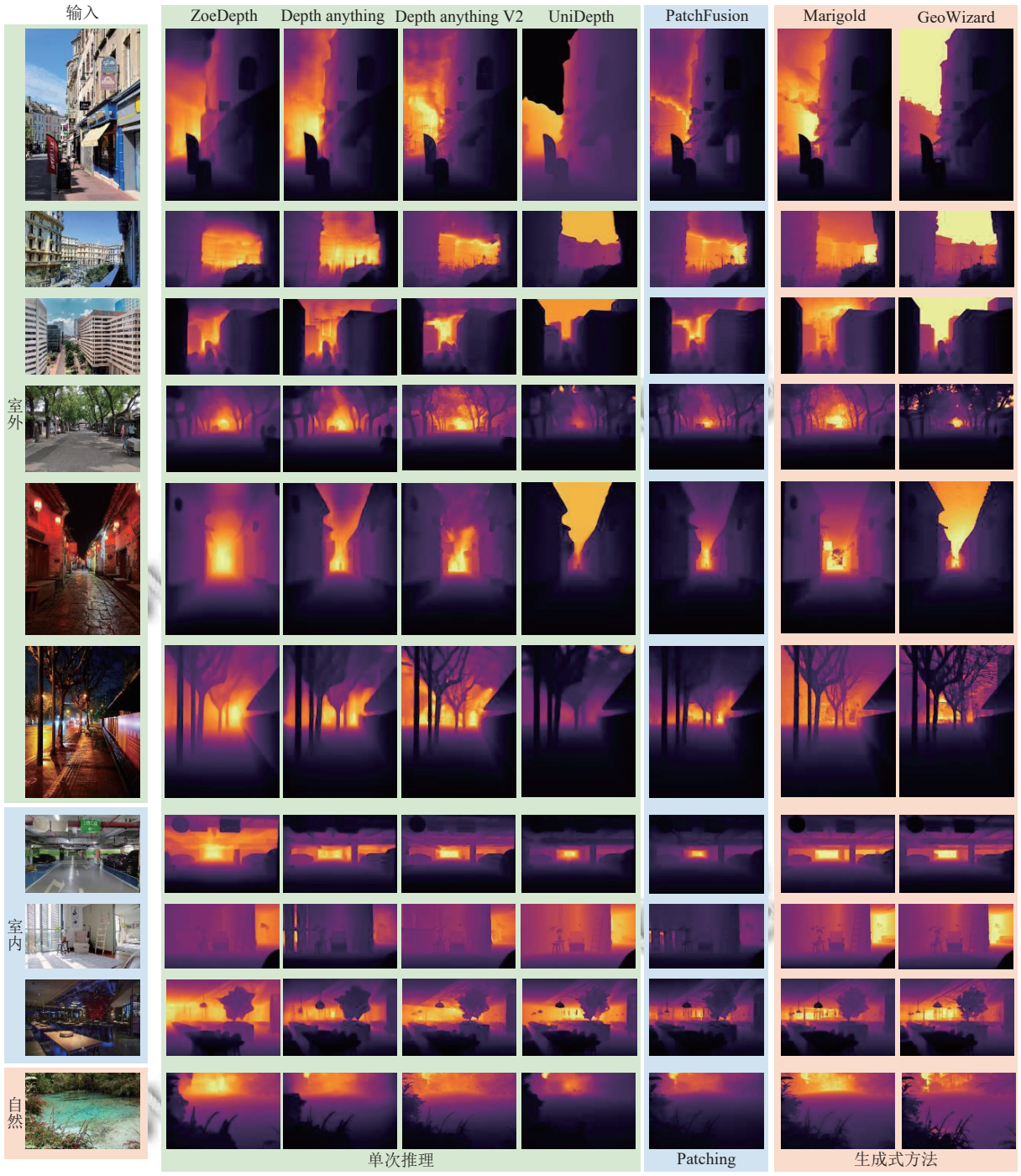

图 1 不同类型方法输出的深度图对比

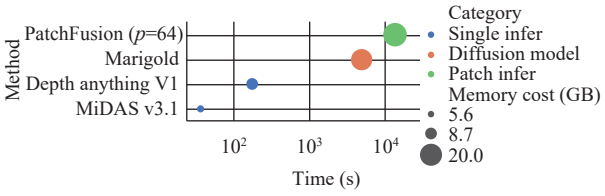

图 2 不同类别的模型推理过程中的时间和内存开销

6 总结与展望

近年来, MMDE 的研究取得了显著进展, 从传统的单一任务优化逐步迈向结合生成模型和多场景泛化能力的创新方向. 表 1 展示了 MMDE 最新研究成果的方法特征总结, 按照时间顺序排列. 其中生成式模型方法大多输出深度的相对值, 生成模型输出度量深度的

唯一实践 DMD[15]未开源. 生成式方法在 MMDE 上的潜力依然有待挖掘.

表 2 定量对比了现有方法在零样本数据和非零样本数据上的深度感知性能. 可以看出在某些数据集的测试中, ZeroDepth 因超出存储空间而无法完成, Metric-3D 需要依赖相机参数, Depth anything 的性能则不完全符合严格意义上的零样本要求. 此外, 从表中可以看出, 不同模型在不同领域的性能仍然存在显著差异, 这表明 MMDE (单目深度估计) 模型的泛化性还有较大的提升空间. 表 2 所示前 6 个数据集使用δ_1作为评价指标 (↑表示值越大越好), 用于评估模型的零样本性能, 测试成绩引用自 Depth pro[14]; 最后 2 个数据集使用 AbsRel 作为评价指标 (↓表示值越小越好) 评价模型的非零样本性能. 表 2 中的性能具有一定参考价值, 但当前 MDE 领域尚缺乏广泛认可的统一标准来对训练数据、模型规模及推理开销等因素进行对齐, 从而更公平地比较模型性能. 从表 2 可以看出, 模型改进和数据资源的优化相辅相成, 推动了深度估计技术在多样化场景中的适配能力, 为三维场景重建和空间理解带来了全新的可能性. 然而, 现有方法仍然面临诸多挑战, 如高频信息的损失、复杂场景的几何一致性不足以及在实际应用中性能与效率的平衡问题. 通过更先进的损失函数设计、数据增强策略以及生成模型的引入, 研究正逐步靠近三维场景几何关系精准还原的目标.

表 1 单目度量深度估计最新研究成果的方法特征

方法	方法来源	模型类别	推理次数	训练数据	模型输出	是否开源
ZoeDepth[13]	arXiv	判别	单次	真实	度量	是
Depth anything[16]	CVPR2024	判别	单次	真实	度量	是
PatchFusion[50]	CVPR2024	判别	多次	真实	度量	是
UniDepth[48]	CVPR2024	判别	单次	真实	度量	是
Marigold[55]	CVPR2024	生成	多次	合成	相对	是
DMD[15]	arXiv	生成	多次	真实	度量	否
Depth anything V2[17]	NeurIPS2024	判别	单次	真实+合成	度量	是
GeoWizard[56]	ECCV2024	生成	多次	真实+合成	相对	是
PatchRefiner[51]	ECCV2024	判别	多次	真实+合成	度量	是
Depth pro[14]	arXiv	判别	多次	真实+合成	度量	是

表 2 现有方法在零样本数据和非零样本数据上的深度感知性能分析

方法	δ_1 ↑						AbsRel ↓	
	Booster	ETH3D	Middlebury	NuScenes	Sintel	Sun-RGBD	NYU v2	KITTI
	室内	室外	室外	室外	室外	室内	室内	室外
Depth anything[16]	52.3	9.3	39.3	35.4	6.9	85.0	4.3	7.6
Depth anything V2[17]	**59.5**	36.3	37.2	17.7	5.9	72.4	4.4	7.4
Metric3D[36]	4.7	34.2	13.6	64.4	17.3	16.9	8.3	5.8
Metric3D v2[57]	39.4	87.7	29.9	82.6	38.3	75.6	4.5	3.9
PatchFusion[50]	22.6	51.8	49.9	20.4	14.0	53.6	—	—
UniDepth[48]	27.6	25.3	31.9	**83.6**	16.5	**95.8**	5.78	4.2
ZeroDepth[37]	—	—	46.5	64.3	12.9	—	8.4	10.5
ZoeDepth[13]	21.6	34.2	53.8	28.1	7.8	85.7	7.7	5.7
Depth pro[14]	46.6	41.5	**60.5**	49.1	**40.0**	89.0	—	—

损失函数的改进是度量深度估计的核心技术之一. 传统损失函数多集中于全局深度一致性和局部平滑性, 但在保留高频信息方面表现不足, 导致深度图在处理纹理复杂的区域时细节丢失明显. 为了解决这一问题, 研究人员开始引入高频保留约束, 如边缘感知损失和基于梯度的结构约束, 使得模型能够更有效地捕捉细微的深度变化. 这些改进不仅增强了模型对复杂纹理区域的适配能力, 也为高精度深度估计在实际场景中的应用奠定了基础. 此外, 结合生成式方法的特性, 将生成过程中逐步还原的细节作为监督信号, 进一步丰富了深度估计的细节表现力, 推动了方法的性能提升.

数据资源的丰富性和质量直接决定了深度估计模型的泛化能力和应用范围. 合成数据和现实场景数据的结合成为解决数据不足问题的重要途径. 高质量合成场景深度数据的生成技术不断进步, 通过模拟光学特性和多样化的场景布局, 生成的数据在复杂度和细

节表现上接近真实世界. 同时, 针对高质量现实场景深度数据的收集方法也有了显著改进, 诸如采用激光雷达和多视图融合等技术获取精确的深度信息. 此外, 合成数据与现实数据的混合使用策略通过增强数据分布的多样性, 有效提升了模型在多领域的适配能力. 例如, 通过设计数据增强方法和领域对齐策略, 模型能够更高效地利用合成数据的多样性和现实数据的真实性, 在不同场景中取得一致的性能表现. 这些数据改进方法不仅增强了深度估计模型的训练效果, 也为未来更复杂任务的研究提供了坚实的数据支持.

扩散生成模型的引入为度量深度估计领域带来了革命性的变化. 这类方法通过模拟图像退化和逐步生成的过程, 使得深度估计能够在细节表现和全局一致性之间取得平衡. 例如, Marigold 和 GeoWizard 在复杂场景的几何结构理解上表现出色, 特别是在处理高反射区域和透明物体时, 生成的深度图相比传统方法更加自然且细腻. 此外, DMD 等创新方法通过对数尺度深度参数化和视场条件输入, 不仅解决了不同摄像机参数下的深度尺度歧义问题, 还显著提升了模型在多样化场景中的泛化能力. 这些生成式方法的实践证明, 其在细节捕捉和复杂场景几何理解方面具备巨大潜力. 尽管当前扩散生成模型在度量深度估计中的研究尚处于初期, 但其多步推理和去噪策略的优化正不断推进, 使其在效率和性能之间找到更优的平衡点, 为未来研究提供了全新思路.

从单一领域的度量深度估计逐步迈向广泛场景的零样本深度估计, 是该领域发展的重要趋势. 通过结合大规模无标注数据和多领域迁移能力, 模型正逐步具备在未见场景中实现高效深度预测的能力. 例如, Zoe-Depth 和 UniDepth 等模型通过网络架构设计的创新和训练策略的改进, 在复杂场景中展现出卓越的性能. 这些方法不仅扩展了模型的应用范围, 也显著提升了其对多样化场景的适应能力. 特别是在动态场景和高分辨率场景中, 零样本单目深度估计方法表现出色, 展现了实现通用三维感知的潜力.

未来的研究方向将重点集中在方法效率、模型泛化性和数据资源优化上. 效率方面, 多步推理方法需要进一步简化推理过程, 以满足实时应用的需求. 结合单次推理的高效性与生成式方法的细节捕捉能力, 可能成为提升方法效率的突破口. 泛化性方面, 通过更有效的迁移学习和领域对齐技术, 使模型能够适应更广泛的场景分布. 同时, 研究几何一致性约束的优化方法, 使得深度估计结果在多视图和三维重建任务中表现更加稳健. 数据资源方面, 进一步提升合成数据的真实性和多样性, 同时加强现实数据的采集效率, 为模型提供更全面的训练基础. 此外, 探索合成数据与现实数据的动态平衡策略, 将有助于更好地利用不同数据的优势, 提升模型性能.

综合来看, 度量深度估计领域正朝着更加通用、精细和高效的方向迈进. 通过损失函数改进、高质量数据资源利用以及生成式方法的持续创新, 研究人员正在努力构建能够真实理解和还原三维场景几何关系的深度估计系统. 随着更多通用性和几何一致性增强方法的引入, 零样本单目深度估计有望成为推动计算机视觉与三维场景感知技术进一步发展的重要基石.

参考文献

1 Mildenhall B, Srinivasan PP, Tancik M, *et al*. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 2022, 65(1): 99–106. [doi: 10.1145/3503250]

2 Kerbl B, Kopanas G, Leimkühler T, *et al*. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG), 2023, 42(4): 139.

3 Ye CJ, Nie YY, Chang JH, *et al*. GauStudio: A modular framework for 3D Gaussian splatting and beyond. arXiv: 2403.19632, 2024.

4 Szeliski R. Computer vision: Algorithms and applications. 2nd ed. Cham: Springer, 2022.

5 Zheng JH, Lin CH, Sun JH, *et al*. Physical 3D adversarial attacks against monocular depth estimation in autonomous driving. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 24452–24461.

6 Leduc A, Cioppa A, Giancola S, *et al*. SoccerNet-Depth: A scalable dataset for monocular depth estimation in sports videos. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 3280–3282.

7 Zhang LM, Rao AY, Agrawala M. Adding conditional control to text-to-image diffusion models. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 3813–3824.

8 Khan N, Xiao L, Lanman D. Tiled multiplane images for practical 3D photography. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision.

Paris: IEEE, 2023. 10420–10430.
9 Liew JH, Yan HS, Zhang JF, *et al*. MagicEdit: High-fidelity and temporally coherent video editing. arXiv:2308.14749, 2023.
10 Xu DJ, Jiang YF, Wang PH, *et al*. NeuralLift-360: Lifting an in-the-wild 2D photo to a 3D object with 360° views. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 4479–4489.
11 Shahbazi M, Claessens L, Niemeyer M, *et al*. InseRF: Text-driven generative object insertion in neural 3D scenes. arXiv: 2401.05335, 2024.
12 Shriram J, Trevithick A, Liu LJ, *et al*. RealmDreamer: Text-driven 3D scene generation with inpainting and depth diffusion. arXiv:2404.07199, 2024.
13 Bhat SF, Birkl R, Wofk D, *et al*. ZoeDepth: Zero-shot transfer by combining relative and metric depth. arXiv:2302. 12288, 2023.
14 Bochkovskii A, Delaunoy A, Germain H, *et al*. Depth pro: Sharp monocular metric depth in less than a second. arXiv: 2410.02073, 2024.
15 Saxena S, Hur J, Herrmann C, *et al*. Zero-shot metric depth with a field-of-view conditioned diffusion model. arXiv: 2312.13252, 2023.
16 Yang LH, Kang BY, Huang ZL, *et al*. Depth anything: Unleashing the power of large-scale unlabeled data. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 10371–10381.
17 Yang LH, Kang BY, Huang ZL, *et al*. Depth anything V2. Proceedings of the 38th Conference on Neural Information Processing Systems. Vancouver: NeurIPS, 2024. 1–30.
18 Ruan XG, Yan WJ, Huang J, *et al*. Monocular depth estimation based on deep learning: A survey. Proceedings of the 2020 Chinese Automation Congress (CAC). Shanghai: IEEE, 2020. 2436–2440.
19 Zhao CQ, Sun QY, Zhang CZ, *et al*. Monocular depth estimation based on deep learning: An overview. Science China Technological Sciences, 2020, 63(9): 1612–1627. [doi: 10.1007/s11431-020-1582-8]
20 Khan F, Salahuddin S, Javidnia H. Deep learning-based monocular depth estimation methods—A state-of-the-art review. Sensors, 2020, 20(8): 2272. [doi: 10.3390/s20082272]
21 Bhoi A. Monocular depth estimation: A survey. arXiv: 1901.09402, 2019.
22 Dong XS, Garratt MA, Anavatti SG, *et al*. Towards real-time monocular depth estimation for robotics: A survey. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10): 16940–16961. [doi: 10.1109/TITS.2022.3160741]
23 Vyas P, Saxena C, Badapanda A, *et al*. Outdoor monocular depth estimation: A research review. arXiv:2205.01399, 2022.
24 Tosi F, Ramirez PZ, Poggi M. Diffusion models for monocular depth estimation: Overcoming challenging conditions. Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2024. 236–257.
25 Lahiri S, Ren J, Lin XK. Deep learning-based stereopsis and monocular depth estimation techniques: A review. Vehicles, 2024, 6(1): 305–351. [doi: 10.3390/vehicles6010013]
26 Masoumian A, Rashwan HA, Cristiano J, *et al*. Monocular depth estimation using deep learning: A review. Sensors, 2022, 22(14): 5353. [doi: 10.3390/s22145353]
27 Arampatzakis V, Pavlidis G, Mitianoudis N, *et al*. Monocular depth estimation: A thorough review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(4): 2396–2414. [doi: 10.1109/TPAMI.2023.3330944]
28 Rajapaksha U, Sohel F, Laga H, *et al*. Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey. ACM Computing Surveys, 2024, 56(12): 315.
29 Jampani V, Chang HW, Sargent K, *et al*. SLIDE: Single image 3D photography with soft layering and depth-aware inpainting. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 12498–12507.
30 Wofk D, Ranftl R, Müller M, *et al*. Monocular visual-inertial depth estimation. Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). London: IEEE, 2023. 6095–6101.
31 Singh AD, Ba YH, Sarker A, *et al*. Depth estimation from camera image and mmWave radar point cloud. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 9275–9285.
32 Eigen D, Puhrsch C, Fergus P. Depth map prediction from a single image using a multi-scale deep network. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2366–2374.
33 Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 2650–2658.
34 Birkl R, Wofk D, Müller M. MiDaS v3.1—A model zoo for robust monocular relative depth estimation. arXiv:2307. 14460, 2023.
35 Han K, Wang YH, Chen HT, *et al*. A survey on vision

Transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 87–110. [doi: 10.1109/TPAMI.2022.3152247]

36 Yin W, Zhang C, Chen H, *et al*. Metric3D: Towards zero-shot metric 3D prediction from a single image. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 9009–9019.

37 Guizilini V, Vasiljevic I, Chen D, *et al*. Towards zero-shot scale-aware monocular depth estimation. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 9199–9209.

38 Spencer J, Russell C, Hadfield S, *et al*. Kick back & relax++: Scaling beyond ground-truth depth with SlowTV & CribsTV. arXiv:2403.01569, 2024.

39 Bhat SF, Alhashim I, Wonka P. AdaBins: Depth estimation using adaptive bins. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 4008–4017.

40 Bhat SF, Alhashim I, Wonka P. LocalBins: Improving depth estimation by learning local distributions. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 480–496.

41 Li ZY, Wang XY, Liu XM, *et al*. BinsFormer: Revisiting adaptive bins for monocular depth estimation. IEEE Transactions on Image Processing, 2024, 33: 3964–3976. [doi: 10.1109/TIP.2024.3416065]

42 Yuan WH, Gu XD, Dai ZZ, *et al*. New CRFs: Neural window fully-connected CRFs for monocular depth estimation. arXiv:2203.01502, 2022.

43 Spencer J, Tosi F, Poggi M, *et al*. The third monocular depth estimation challenge. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2024. 1–14.

44 Wang YH, Liang YJ, Xu H, *et al*. SQLdepth: Generalizable self-supervised fine-structured monocular depth estimation. Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2024. 5713–5721.

45 Shao SW, Pei ZC, Chen WH, *et al*. MonoDiffusion: Self-supervised monocular depth estimation using diffusion model. arXiv:2311.07198v1, 2023.

46 Haji-Esmaeili MM, Montazer G. Large-scale monocular depth estimation in the wild. Engineering Applications of Artificial Intelligence, 2024, 127: 107189. [doi: 10.1016/j.engappai.2023.107189]

47 Marsal R, Chabot F, Loesch A, *et al*. MonoProb: Self-supervised monocular depth estimation with interpretable uncertainty. Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2024. 3625–3634.

48 Piccinelli L, Yang YH, Sakaridis C, *et al*. UniDepth: Universal monocular metric depth estimation. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 10106–10116.

49 Miangoleh SMH, Dille S, Mai L, *et al*. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 9680–9689.

50 Li ZY, Bhat SF, Wonka P. PatchFusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 10016–10025.

51 Li ZY, Bhat SF, Wonka P. PatchRefiner: Leveraging synthetic data for real-domain high-resolution monocular metric depth estimation. Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2024. 250–267.

52 Duan YQ, Guo XD, Zhu Z. DiffusionDepth: Diffusion denoising approach for monocular depth estimation. Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2024. 432–449.

53 Zavadski D, Kalšan D, Rother C. PrimeDepth: Efficient monocular depth estimation with a stable diffusion preimage. Proceedings of the 17th Asian Conference on Computer Vision. Hanoi: Springer, 2024. 21–40.

54 Patni S, Agarwal A, Arora C. ECoDepth: Effective conditioning of diffusion models for monocular depth estimation. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 28285–28295.

55 Ke BX, Obukhov A, Huang SY, *et al*. Repurposing diffusion-based image generators for monocular depth estimation. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 9492–9502.

56 Fu X, Yin W, Hu M, *et al*. GeoWizard: Unleashing the diffusion priors for 3D geometry estimation from a single image. Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2024. 241–258.

57 Hu M, Yin W, Zhang C, *et al*. Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 10579–10596. [doi: 10.1109/TPAMI.2024.3444912]

(校对责编: 王欣欣)