食物份量估算：从像素到卡路里

Gautham Vinod Fengqing Zhu

arXiv:2602.05078v1 [cs.CV] 4 Feb 2026

摘要——**依靠图像进行膳食评估，是准确、便捷地监测个人健康状况的重要策略，也是预防和管理慢性疾病及肥胖的重要手段。然而，基于图像的膳食评估面临着一个难题：如何根据二维图像输入估算食物的三维尺寸。为克服这一关键限制，人们提出了多种策略，例如使用深度图等辅助输入、多视图输入，或采用模板匹配等基于模型的方法。深度学习也有助于弥合这一差距：它既可以利用单目图像，也可以结合图像与辅助输入，从图像输入中准确预测食物份量。本文探讨了用于准确估算份量的各种策略。**

索引词——**尺寸估算、食物、综述、基于图像的膳食评估**

一、引言

膳食评估是降低心血管问题、肥胖和糖尿病等慢性疾病风险的重要工具，而这些疾病通常与不健康的饮食模式有关 [1]。然而，传统膳食评估严重依赖自我报告方法，例如 24 小时膳食回顾 [2] 和食物频率问卷。在这些流程中，参与者需要向注册营养师报告此前一段时间的食物摄入情况。这一过程不仅给参与者带来负担，也容易产生显著误差；研究一再表明，参与者常因偏差而少报食物摄入量 [3], [4]。

基于图像的膳食评估（IBDA）旨在实现这一流程的自动化：用户只需拍摄用餐时的图像，从而减少用户付出的精力并降低报告偏差 [5]。标准的基于图像的膳食评估流程通常包括食物检测、分类和份量估计。尽管传统分类（识别餐盘上“是什么”）得益于大规模数据集而日趋成熟，但食物数据仍然特别具有挑战性，因为其类间相似性较高（例如意面酱与果冻），且类内差异较大（例如不同类型的汉堡或三明治）。更重要的是，定量（估计“有多少”）仍是长期存在的瓶颈。

份量估计的核心挑战是尺度歧义问题。当三维真实场景投影到2D图像平面时，深度信息会丢失。若没有已知参照物，靠近相机的小披萨在几何上与放置在更远处的大披萨无法区分。因此，仅凭单张图像进行份量估计在数学上属于不适定问题，因为无限种三维体积都能产生完全相同的2D投影 [6]。

为了解决这种歧义，早期方法依赖显式的几何恢复。这些方法通常需要基准标记 （例如场景中放置的一枚硬币或已知尺寸的棋盘格图案）作为物理参照来确定尺度 [7]。虽然这种方法有效，却给用户带来了额外负担，降低了系统的易用性。后续研究探索了使用专用硬件（如 RGB-D 传感器（例如 Microsoft Kinect）或智能手机激光雷达）来直接获取深度图 [8]–[10]。然而，并非所有消费级设备都配备此类传感器，而且它们往往难以捕捉区分食物密度所需的细致纹理。另一种方法是采用多视图立体视觉（MVS）技术，尝试根据餐食的多张图像重建 3D 网格 [11], ，但这需要场景保持固定，并要求用户充分配合以拍摄 360度视图。

深度学习的最新进展已将研究重点从显式几何重建转向单目推断。现代架构利用海量数据集“学习”深度先验，从而能够直接根据单张RGB图像预测体积和能量，无需物理参照物 [5], [9], [12], [13]。这些方法采用的技术涵盖单目深度估计以及隐式体积表示（如神经辐射场（NeRF）） [14],, 以补全缺失的三维结构。

本文回顾了食物份量估算面临的挑战与进展。我们首先讨论通过传感器或多视图约束恢复三维数据的几何方法，再探讨深度学习如何利用这些输入实现单目预测，并讨论准确度、用户负担与计算复杂度之间的权衡。

II. 几何尺寸估计：恢复三维数据

份量估计面临的根本挑战，是从二维投影中恢复度量三维结构。要估计食物的体积 V （或其他与尺寸相关的物理量，如重量），系统必须确定两个未知量：场景的绝对物理尺度（度量深度）和物体的形状几何。我们将几何方法分为深度感知、多视图立体和基于模型的方法，如图 1 所示。

A. 专用深度传感器

早期的 IBDA 系统利用能够捕获显式深度图的硬件， $D(u, v)$ ，以及 RGB 数据 [8], [9], [15]–[17]，绕过了单目估计的不适定性。

结构光或 RGB-D：诸如 Microsoft Kinect 或英特尔 RealSense 这样的《传感器》，会将已知的红外光图案（例如斑点或网格）投射到场景中。图案在物体表面的变形可用于计算

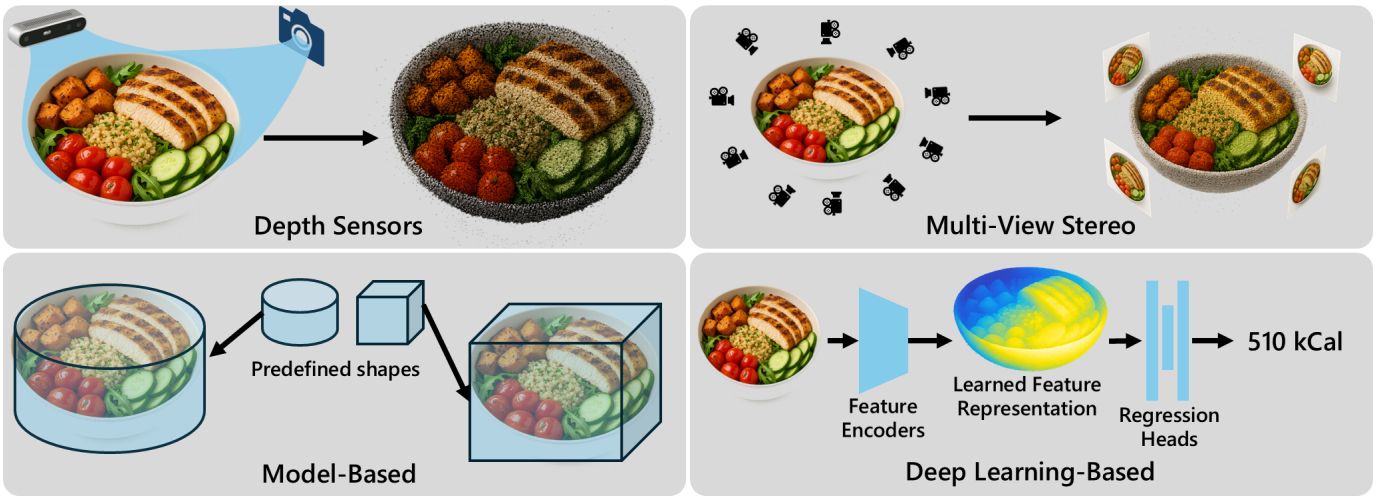

图 1：传统方法与深度学习方法。传统方法通过基于深度的方法（专用硬件）、多视图立体视觉或基于模型的方法，从二维图像中推断缺失的三维信息。深度学习则尝试通过使用辅助输入，或学习将输入图像映射到辅助输入，并结合这些方法进行份量回归，以弥合这一差距。

通过三角测量获得深度。

尽管这些系统在体积积分方面准确度很高，但存在严重的可用性限制：它们通常无法处理反光或透明表面（例如汤、金属餐具），在户外照明条件下会因红外干扰而表现不佳，而且对较小物体（如食物）的准确度较低。此外，这些系统需要笨重的外置硬件，因此不适合日常、真实场景中的饮食追踪。

消费级智能手机（如 iPhone Pro 系列）中的移动激光雷达和飞行时间（ToF）扫描仪，引发了人们对直接采集深度信息的兴趣。飞行时间传感器通过测量光脉冲的往返时间来生成深度图。然而，移动激光雷达的空间分辨率往往不足以用于膳食评估。较小的颗粒状食物（例如米粒、坚果）经常会被平滑处理，导致体积估算出现显著误差，这是由于“凸包”效应造成的：颗粒之间的空隙被计作实体质量 [18], [19]。

B. 多视图立体视觉（*MVS*）

运动恢复结构（SfM）：通过围绕餐盘拍摄视频或一系列静态图像，运动恢复结构算法（例如 COLMAP） [20]–[22] 可以同时估计相机位姿并生成场景点云。虽然这种方法对刚性物体 [23], 有效，但运动恢复结构（SfM）假定场景是静态的；拍摄期间食物的任何移动（例如蒸汽、食材移位）都会引入噪声。除非已知相机内参，否则通常无法确定尺度。

立体匹配：利用校正后的立体图像对 [24]，可以通过视差 d 计算Z 深度。视差表示某个点在左视图与右视图之间的位移，其关系由以下公式表示：

$$Z = \frac{f \cdot B}{d} \tag{1}$$

其中 f 是焦距，B 是相机之间的基线距离。

然而，对于饮食评估而言，MVS 的主要瓶颈是用户负担。要求用户进行 360° 扫描操作或从多个稳定角度进行拍摄，都十分繁琐。此外，MVS 难以解决遮挡问题：与盘子接触或被其他食物遮挡的部分无法重建，因此需要进行插值，从而引入体积误差。

C. 基于模型与模板匹配

为避免进行密集重建，基于模型的方法会将观测结果拟合到已知的形状先验，从而有效地将问题简化为参数估计任务。

几何基本体：结构规整的 Foods 可用简单的欧几里得形状近似。例如，饮料建模为圆柱体，碗建模为半球体，蛋糕建模为三棱柱。体积可通过解析方法计算（例如，$V = \pi r^2 h$ 圆柱体的体积）。这种方法计算效率高，但难以推广到不具刚性或形状不规则的食物（例如沙拉、咖喱），因为它们不符合标准几何基本体 [25], [26]。

3D模板缩放：该方法使用高保真3D网格库（例如扫描得到的香蕉或苹果模型）。系统将3D模板投影到2D图像平面上，并优化姿态参数 $\mathbf{R}$，$\mathbf{t}$ 和全局缩放因子 s，以最小化投影与分割掩码 [6]之间的对齐误差。

在这里，模板刚性是一个显著的局限。通用的披萨片模板无法适应被咬了一口或披萨饼边折叠的具体实例。理想模板与真实世界变形之间的差异会导致严重的体积误差。此外，需要已知物理锚点也限制了此类方法的适用场景。

III. 深度学习：单目转变

近年来最重要的进展，是从显式几何重建转向单目推-

断。深度神经网络（DNN）展现了从单张RGB图像中“推断”缺失三维信息的能力，即通过从大型数据集 [9]中学习先验知识来实现这一点。这一转变使份量估计成为可能，有时仅凭单目输入即可完成。

A. 单目深度预测

该方法不依赖物理传感器，而是将深度估计视为稠密回归问题。神经网络学习映射函数 $f_\theta: I \to D$，根据单张RGB图像 I [9]预测逐像素深度图 D， [27]。此处通常采用编码器-解码器架构，其中编码器提取高层语义特征（将物体识别为“食物”），而解码器逐步上采样这些特征，以回归每个像素的深度值（u, v） [5]。

一旦深度图后，利用相机内参矩阵 K将其反投影到三维空间。系统生成点云或体素网格，并通过积分体素来计算体积。

DPF-Nutrition [28]等近期方法采用跨模态融合。这些网络并非孤立处理深度与RGB特征，而是在多个尺度上将语义特征（纹理、类别）与几何特征（形状、深度）相融合，与仅深度基线相比，显著降低了体积估算误差。

B. 直接能量回归

另一种方法则完全绕过中间的三维重建。这些“端到端”架构直接学习从图像像素到营养成分的映射： $I \to \mathbb{R}^N$，其中输出表示质量、体积、千卡或其他宏量营养素 [15]。

标准骨干网络通过将最后的分类层替换为回归头进行改造。网络通过最小化预测卡路里数与真实值 [29]之间的损失函数来进行训练。

C. 隐式表示

食品计算机视觉的前沿已采用神经辐射场（NeRF）来克服离散网格的局限性 [14]。

与显式点云不同，NeRF 将食物场景表示为连续体积函数 $F_\Theta: (\mathbf{x}, \mathbf{d}) \to (\mathbf{c}\ \sigma)$，将空间坐标 $\mathbf{x}$ 和观察方向 $\mathbf{d}$映射为颜色 $\mathbf{c}$ 和密度 σ。

诸如在 NutritionVerse [18]中作为基准评测的方法，在质量上实现了显著飞跃。它们能够从极少的输入图像中合成新视角，并提取高度精确的体积网格（小样本学习）。至关重要的是，NeRF 对半透明和复杂物体（例如汤、无定形泥状食物）的处理显著优于传统多视图立体视觉，后者往往无法在无纹理或流体区域匹配特征 [30]。

然而，这些方法都面临三个根本性挑战：尺度歧义、遮挡问题和密度估计。

IV. 挑战与未来方向

尽管从主动传感器转向单目深度学习提升了易用性，但准确估计食物份量的问题仍未得到彻底解决。我们指出了三个关键瓶颈，并提出了未来研究方向。

A. 尺度与参考问题

单目估计中最持久的挑战是尺度歧义。在没有已知参考的情况下，从像素坐标（u, v）到世界坐标（X, Y, Z）的映射是不确定的。当前解决方案依赖于基准标记（例如棋盘格或信用卡） [7], ，这会破坏用户体验。深度学习策略依赖于训练数据的分布，往往无法准确预测真实场景输入图像的尺度。

目前，研究正转向无标记尺度估计，通过利用环境先验来实现这一目标。这包括使用动态参考尺度，或学习图像中常见餐饮容器（盘子、碗）或其他视觉参照物的标准尺寸 [31]。

B. 遮挡问题

视觉系统只能估算可见表面的体积。食物底部（被盘子遮挡）和内部配料（例如墨西哥卷饼的馅料或千层面的层次）仍然不可见。目前，深度学习模型依赖视觉线索来“猜测”这些隐藏的几何结构，这往往会导致对复杂、堆叠食物的体积估计偏低。

不过，我们预计，借助生成式扩散模型对食物隐藏的背向几何进行概率推断的模态补全 网络将日益增多，从而基于学到的结构先验有效地“学习”不可见的体积。个性化方法还会利用个人的饮食习惯，对食物内容作出更有依据的推测，而非仅仅依赖视觉特征 [32]。

C. 从体积到质量：密度鸿沟

膳食评估最终需要估算能量（kCal），而不只是体积（cm^3）。二者之间的关系由密度（ρ）决定，而即使在同一食品类别中，密度也可能有很大差异（例如，蓬松的羊角面包与扎实的贝果）。仅凭视觉的方法难以区分这些质地细微差别。

近年来，多模态大语言模型（*MLLMs*）通过结合视觉特征与文本上下文（例如，用户对“低碳水”或“密实”的描述），为解决这一问题提供了有前景的方案；基础模型可利用这些信息来优化密度估算 [33]。此外，将膳食数据库准确映射到分类结果（例如美国农业部的 FNDDS数据库 [34]），也能为不同食物的密度提供估算。

V. 结论

基于图像的食物份量估计领域，正经历一场根本性的转变：从主动、依赖硬件的

重建转向被动的人工智能驱动的推断。早期的几何方法采用结构光和多视图立体视觉，能够提供可靠的真实值，但由于硬件限制和用户负担过重，未能得到广泛应用。深度学习的出现使单目解决方案成为可能，虽然几何精度略有下降，但易用性大幅提升。

然而，由于固有的尺度和密度歧义，要实现真正的准确度仍然充满挑战。通过将视觉数据与大型基础模型的语义推理相结合，下一代 IBDA系统有望解决其中的一些问题，最终让自动化、精准的饮食追踪成为慢性病管理的现实。

参考文献

[1] P. Wang, M. Song, A. H. Eliassen, M. Wang, T. T. Fung, S. K. Clinton, E. B. Rimm, F. B. Hu, W. C. Willett, F. K. Tabung 等，《预防慢性病的最佳膳食模式》，《自然医学》，第29卷，第3期，第719–728页，2023年。[2] T. Baranowski，《24小时膳食回顾和膳食记录法》，《营养流行病学》，第40卷，第49–69页，2012年。[3] R.-L. Karvetti，《24小时膳食回顾的有效性》，《美国营养师协会杂志》，第85卷，第11期，第1437–1442页，1985年。[4] A. F. Subar, L. S. Freedman, J. A. Tooze, S. I. Kirkpatrick, C. J. Boushey, M. L. Neuhouser 等，《传统的自我报告膳食工具容易产生误差，因此需要采用新方法》，《营养学杂志》，第145卷，第12期，第2636–2645页，2015年。[5] Z. Shao, G. Vinod, J. He, and F. Zhu，《一种基于单目图像形状重建的端到端食物份量估算框架》，*2023*年*IEEE*国际多媒体与博览会会议（*ICME*）论文集，第942–947页，2023年7月。[6] G. Vinod, J. He, Z. Shao, and F. Zhu，《通过三维物体缩放估算食物份量》，*2024*年*IEEE/CVF*计算机视觉与模式识别会议研讨会（*CVPRW*）论文集，第3741–3749页，2024年。[7] W. Jia, Y. Yue, J. D. Fernstrom, M. Zhang, and M. Sun，《利用基准标记从数码照片估算食物份量的准确度》，载于*2014*年第*36*届*IEEE*医学与生物学工程学会国际年会。IEEE，2014年，第2028–2031页。[8] F. P. Lo, Y. Sun, J. Qiu, and B. Lo，《使用智能手机进行基于图像的食物份量估算》，《*IEEE*生物医学与健康信息学汇刊》，第23卷，第6期，第2513–2523页，2019年。[9] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Sil- berman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy，《Im2calories：迈向自动化移动视觉食物日记》，载于*IEEE*国际计算机视觉会议论文集，2015年，第1233–1241页。[10] S. Makhsous, H. M. Mohammad, J. M. Schenk, A. V. Mamishev, and A. R. Kristal，《一种用于食物三维重建和体积估算的新型移动式结构光系统》，《传感器》，第19卷，第3期，2019年。[ ] 1424 8220 19 3 564 在线。可访问：https://www.mdpi.com/ - / / / J. Dehais, M. Anthimopoulos, S. Shevchik, and S. Mougiakakou，《用于食物体积估算的双视图三维重建》，载于*IEEE*多媒体汇刊，第19卷，第5期，2016年，第1090–1099页。[12] S. Fang, Z. Shao, R. Mao, C. Fu, and F. Zhu，《单视图食物份量估算：利用生成对抗网络学习图像到能量映射》，载于*2021*年*IEEE*国际图像处理会议（*ICIP*）。IEEE，2021年，第2513–2517页。[13] G. Vinod, Z. Shao, and F. Zhu，《通过深度域自适应进行基于图像的食物能量估算》，*2022*年第五届*IEEE*国际多媒体信息处理与检索会议（*MIPR*）论文集，第262–267页，2022年。[14] B. Mildenhall, P. P. Srinivasan 等，《NeRF：将场景表示为用于视图合成的神经辐射场》，*2020*年欧洲计算机视觉会议论文集，第405–421页，2020年。[15] Q. Thames, A. Karpur, W. Norris, F. Xia, L. Panait, T. Weyand, and J. Sim，《Nutrition5k：迈向通用食物的自动营养理解》，《IEEE/CVF计算机视觉与模式识别会议论文集》，第8903–8911页，2021年。[16] V. B. Raju and E. Sazonov，《Foodcam：一种用于食物份量估算的新型结构光立体成像系统》，《传感器》，第22卷，第9期，第3300页，2022年。[17] J. Shang, G. Broster, and M. Sun，《使用手机和结构光进行膳食评估》，《2011年IEEE多媒体国际研讨会论文集》，第319–324页，2011年。[18] M. Keller, C.-e. A. Tai, Y. Chen, P. Xi, and A. Wong，《NutritionVerse-Direct：探索用于从食物图像进行多任务营养预测的深度神经网络》，arXiv预印本 arXiv:2405.07814，2024年。[19] R. Dinic, M. Domhardt, S. Ginzinger, and T. St¨utz，《Eatar tango：利用深度传感器在移动设备上估算份量》，《第19届人机交互与移动设备及服务国际会议论文集》，2017年。 [在线]。可访问：https://doi.org/10 1145 3098279 3125434 。/ 。J. L. Schonberger, E. [20] ¨ Zheng, M. Pollefeys, and J.-M. Frahm，《非结构化多视图立体视觉的逐像素视图选择》，《欧洲计算机视觉会议》（ECCV），2016年。[21] J. L. Sch¨onberger and J.-M. Frahm，《运动恢复结构再探》，《计算机视觉与模式识别会议》（CVPR），2016年。[22] M. H. Rahman, Q. Li, M. Pickering, M. Frater, D. Kerr, C. Bouchey, and E. Delp，《基于手机的膳食评估系统中的食物体积估算》，载于《2012年第八届信号图像技术与互联网系统国际会议》，2012年，第988–995页。[23] J. He, Y. Chen, G. Vinod, T. I. Mahmud, F. Zhu, E. Delp, A. Wong, P. Xi, A. AlMughrabi, U. Haroon 等，《Metafood CVPR 2024物理信息引导的三维食物重建挑战赛：方法与结果》，ArXiv预印本 ArXiv:2407.09285，2024年。[24] J. Dehais, M. Anthimopoulos, S. Shevchik, and S. Mougiakakou，《用于食物体积估算的双视图三维重建》，IEEE多媒体汇刊，第19卷，第5期，第1090–1099页，2017年。[25] C. Xu, Y. He, N. Khanna, C. J. Boushey, and E. J. Delp，《利用三维姿态进行基于模型的食物体积估算》，《2013年IEEE图像处理国际会议论文集》，第2534–2538页，2013年。[26] S. Fang, F. Zhu, C. Jiang, S. Zhang, C. J. Boushey, and E. J. Delp，《使用几何模型和深度图像估算食物份量的比较》，《2016年IEEE图像处理国际会议（ICIP）论文集》，第26–30页，2016年。[27] J. Ma, X. Zhang, G. Vinod, S. Raghavan, J. He, and F. Zhu，《Mfp3d：利用三维点云进行单目食物份量估计》，ArXiv预印本 ArXiv:2411.10492，2024年。[28] Y. Han, Q. Cheng, W. Wu, and Z. Huang，《DPF-Nutrition：通过深度预测与融合进行食物营养估计》，Foods，第12卷，第23期，第4293页，2023年。[29] A. Sanatbyek, T. Rakhimzhanova, B. Nurmanova, Z. Omarova, A. Rakhmankulova, R. Orazbayev, H. A. Varol, and M. Y. Chan，《用于食物场景识别与份量估算的多任务深度学习模型——食物份量基准（FPB）数据集》，IEEE Access，第13卷，第152 033–152 045页，2025年。[30] G. Vinod, B. Coburn, S. Raghavan, J. He, and F. Zhu，《尺寸至关重要：从单目图像重建真实尺度三维模型以估算食物份量》，arXiv，2026年。 [在线]。可访问：https://arxiv.org/abs/ 2601 20051 。A. [31] Sharma, C. Czarnecki, Y. Chen, P. Xi, L. Xu, and A. Wong，《你吃了多少?勺上食物份量估计》，《IEEE/CVF计算机视觉与模式识别会议（CVPR）研讨会论文集》，第3761–3770页，2024年6月。[32] X. Pan, J. He, A. Peng, and F. Zhu，《利用改进的马尔可夫链模拟个人食物消费模式》，《第七届多媒体辅助膳食管理国际研讨会论文集》，第61–69页，2022年。 [在线]。可访问：https://doi.org/ 10 1145 3552484 3555747 。/ 。B. Coburn, J. He, M. E. Rollo, S. S. Dhaliwal, D. A. Kerr, and F. Zhu，《评估大型多模态模型的营养分析能力：一个加入上下文元数据的新基准》，IEEE生物医学与健康信息学国际会议（BHI）论文集，[33] 2025 [ ] 2507 07048 。在线。可访问：https://arxiv.org/abs/ 。J. B. Montville, J. K. Ahuja, C. L. Martin, K. Y. He, G. Omolewa- Tomobi, L. C. Steinfeldt, J. Anand, M. E. Adler, R. P. LaComb, and A. Moshfegh，《美国农业部膳食研究食物与营养数据库（FNDDS），5.0》，Procedia Food Science，第2卷，第99–112页，2013年。[34]