用于学习网格物理增强型可扩展代理模型的科学计算算法

Brian R. Bartoldson[†]、**Yeping Hu**[†*]、**AmarSaini**[†]、[*] **JoseCadena**[†]、[*] **Yucheng Fu**[‡]、**Jie Bao**[‡]、**Zhijie Xu**[‡]、**Brenda Ng**[†] 和 **Phan Nguyen**[†]
[†]劳伦斯利弗莫尔国家实验室 [‡]太平洋西北国家实验室

arXiv:2304.00338v1 [cs.LG] 1 Apr 2023

摘要

数据驱动建模方法能够快速构建代理模型，用于研究大规模物理问题。其中，基于网格数据运行的图神经网络（GNN）因具有促进物理保真度的归纳偏置而备受青睐，但硬件限制一直阻碍其应用于大型计算域。我们证明，可以在三维网格上训练一类图神经网络代理模型。我们通过域分解方法扩展 MeshGraphNets（MGN）——一种用于基于网格的物理建模的图神经网络子类——从而实现训练；在特定条件下，该训练在数学上等价于在整个计算域上进行训练。借助这一方法，我们能够在包含数百万 个节点的网格上训练 MGN，以生成计算流体力学（CFD）模拟。此外，我们还展示了如何通过更高阶数值积分增强 MGN，从而降低其误差和训练时间。我们使用一个配套数据集验证了这些方法，该数据集包含基于 310 万节点网格的三维 CO_2捕集计算流体力学模拟。本研究为将 MGN 扩展至实际应用提供了一条切实可行的途径。

1 引言

理解物理系统和工程过程通常需要对其底层模型进行大量数值模拟。然而，生成这些模拟通常计算成本高昂，这可能会限制它们在大规模问题中的适用性。在这种情况下，可以使用代理模型来模拟原始模型，并高效地生成计算模拟。

代理模型方法可以揭示效率与保真度之间的权衡（Willard 等人，2020年；Wang，2021年）。例如，基于网格的图神经网络在物理保真度和对未见域的泛化能力方面表现出色，但其加速效果（11–290 倍）（Pfaff 等人，2020年）不如泛化能力较弱的卷积神经网络（125–716 倍）（Kim 等人，2019 年）。此外，基于网格的图神经网络受硬件限制和内存需求所限，无法应用于大型图。据我们所知，这类方法尚未应用于节点数超过 20K 的图（Fortunato 等人，2022年）。尽管存在这种权衡，许多代理模型应用都需要同时 具备对未见域的泛化能力，并能高效处理大型数据集。

因此，我们希望借助科学计算算法，在不损害对准确性至关重要的归纳偏置的前提下，提高图神经网络的适用性。我们聚焦于 MeshGraphNets（MGN）（Pfaff 等人，2020年），通过域分解和高阶数值积分来改进 MGN 的学习——这些方法属于我们称为 **科学计算算法：学习增强型可扩展代理模型（SCALES2）** 的框架。我们的贡献如下：

1. 我们在包含数百万个节点的计算流体力学网格上训练图神经网络代理模型，其规模约为以往研究中最大网格的 100 倍（Pfaff 等人，2020年；Fortunato 等人，2022年）。2. 我们提出了分块训练，这是一种用于在大型网格上进行训练的区域分解方法，并给出了该方法与直接在大型网格上训练达到数学等价的条件。3. 我们证明，采用能够增强代理模型忠实度的高阶数值积分方案，可以大幅提升 MGN 的准确性和效率。4. 我们发布了 C3FD，这是一个大规模数据集，包含在三维网格上进行的碳捕集计算流体力学模拟，网格含有 310 万个节点，旨在促进基于图的代理模型的基准测试。

[*]贡献相同。通讯作者：bartoldson@llnl.gov、nguyen97@llnl.gov。

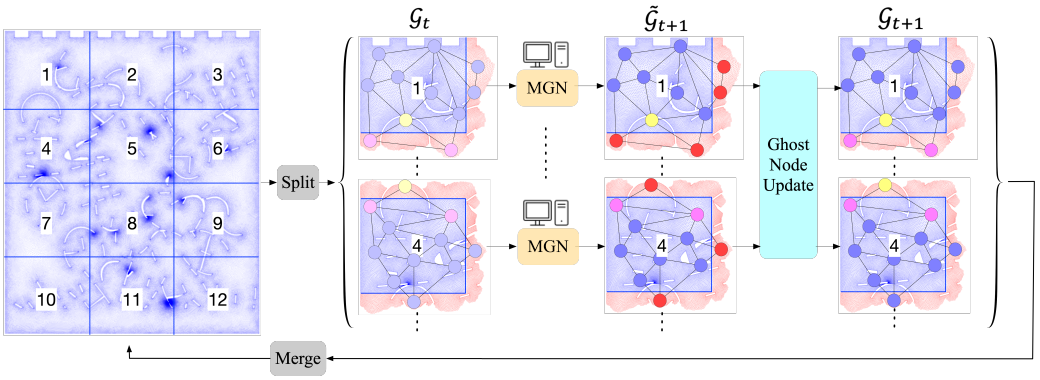


图 1：**通过域分解实现 MGN 扩展**。我们将图 $\mathcal{G}_t$ 划分为 P 子区域（正方形），在子区域中加入虚节点以获得重叠补丁（例如，补丁 1 和补丁 4 彼此包含对方的虚节点，分别以粉色和黄色节点表示），然后在这些补丁上运行 MGN，$\mathcal{G}_t^{(p)}$ 从而获得 $\tilde{\mathcal{G}}_{t+1}^{(p)}$。在推理过程中，由于仅与部分邻居进行“不完整”的消息传递，节点可能具有不正确的状态（红色）；此时通过跨补丁通信更新这些节点，以获得 $\mathcal{G}_{t+1}^{(p)}$。在训练过程中，由于我们只预测下一个时间步，因此不需要更新虚节点。

2 方法

2.1 MeshGraphNets

MGN 使用消息传递（MP）图神经网络从基于网格的模拟中学习。在每个消息传递步骤中，MGN 通过聚合关联节点的特征来更新每条边的特征，然后通过聚合关联边的特征来更新每个节点的特征。经过 k 次迭代后，每个节点和边都会受到其 k 跳邻域的影响。随后，利用每个节点的最终特征，根据以下更新规则预测该节点的下一个状态：$y_{t+h} = y_t + \text{MGN}(y_t)$，其中 y_t 是系统在时间 t 的数值估计状态，h 是时间步长。

2.2 用于扩展的域分解

我们开发了一种域分解框架，使基于网格的图神经网络代理模型能够在大型域上进行训练（见图 1）。当在用于推理或学习的硬件上将代理模型应用于某个域会产生内存溢出错误时，域分解便十分有用。广义而言，它将定义在某个域上的模拟问题划分为一组联合等价（或相似）的问题，每个问题都定义在其自身的子域上（Tang 等人，2021）。

对于我们的数据集，我们将域的每个维度划分为相等的部分，从而创建大小均匀的子域。受重叠域分解方法（Tang 等人，2021）的启发，我们为每个子域增加其 k 跳邻域，并将其称为宽度为 k 的幽灵区。我们将子域及其幽灵区的并集称为补丁。由于每个子域节点都可以访问其补丁中的 k 跳邻域，因此，具有 m 消息传递步数的 MGN（其中 $m \leq k$）将更新每个子域节点，就像整个域都可用一样。因此，在满足以下三个条件的情况下，在补丁上进行训练等价于在完整域上进行训练：(1) $m \leq k$；(2) 我们仅计算子域节点的损失（不包括虚节点）；(3) 我们在多个批次上累积梯度，直到访问完域中的所有补丁后再更新权重。实际上，我们违反了第三个条件：在看到域中随机子集的补丁后便更新权重，我们将由此产生的过程称为“补丁训练”。

2.3 用于增强学习的高阶积分

MGN 的更新规则基于前向欧拉（FE）方法。该方法通过忽略以下泰勒级数中的二次项和高阶项，引入一阶全局误差和二阶局部截断误差：$y(t+h) = y(t) + hf(y(t)) + \mathcal{O}(h^2)$，其中 $y(t)$ 是系统在时间 t 的真实状态。因此，MGN 的预测误差为 $|y(t+h) - y_{t+h}| = \left|[hf(y(t)) - \text{MGN}(y(t))] + \mathcal{O}(h^2)\right|$，

它可以分解为（缩放后的）时间导数的预测误差 $[hf(y) - \text{MGN}(y)]$, , 以及二阶局部误差 $\mathcal{O}(h^2)$。[1]

虽然 MGN 在学习估计时间导数时可能会学会抵消 $\mathcal{O}(h^2)$ 项，但通过改变 MGN 将 y_t 更新为 y_{t+h} 的方式，可以降低局部误差，从而提高正确预测时间导数所带来的收益。具体而言，我们将前向欧拉法替换为 Heun 二阶和三阶方法（Heun 等人，1900年；J.C.Butcher，1996年），它们的局部误差更小（分别为 $\mathcal{O}(h^3)$ 和 $\mathcal{O}(h^4)$）。在附录 A 中，我们推测这可以加快并增强学习过程。

3 数据集

CylinderFlowCylinderFlow 包含通过数值求解器，利用纳维–斯托克斯方程在欧拉网格上生成的绕圆柱体无黏、不可压缩流二维模拟（Pfaff 等人，2020年）。我们训练代理模型，以预测网格每个节点处的流体动量。

C3FD 我们的碳捕集计算流体力学（C3FD）数据集包含 STAR-CCM+ 模拟，其中 CO_2- 捕集溶剂流经塔器中的填料结构。该数据集包含 50 个二维模拟和 50 个三维模拟，每个模拟均跨越 500 个时间步。二维和三维模拟网格分别包含 150K 个节点和 3.1M 个节点。每个节点包含其物理位置、类型、液体体积分数（与捕集效率相关的量）、单位体积动量和接触角等信息。我们训练代理模型，以预测每个节点处的液体体积分数和动量。附录 B 提供了更多数据集详情和可视化结果。C3FD 数据集可访问：https://doi.org/10.25584/1963908。

4 实验

作为合理性检查，我们首先在完整的二维 C3FD 域上复现 Pfaff 等人（2020年）的实验。在图 3 和图 5 中，我们展示了如何将多重图网络训练至较低误差，并生成与真实值相似的 500 时间步滚动预测。此外，我们还展示了在二维 C3FD 域上训练的多重图网络能够泛化到不同的二维域（即三维域中的新切片）。

这些关于二维 C3FD 的初步实验需要约 50 GB 的加速器内存才能应用原始 MGN。将 MGN 应用于更大的数据集（如三维C3FD数据集）并不可行，因为这需要约 1 TB 的加速器内存，而常见硬件通常无法提供如此容量。因此，为了提高基于网格的图神经网络代理模型的效率，我们测试了用于扩展其规模并增强其学习能力的算法。最终，我们的分块训练成功扩展了 MGN 训练，使其能够从三维 C3FD 数据中学习。我们还展示了通过采用高阶积分可以增强 MGN 训练。

4.1 分块训练是扩展代理模型的关键

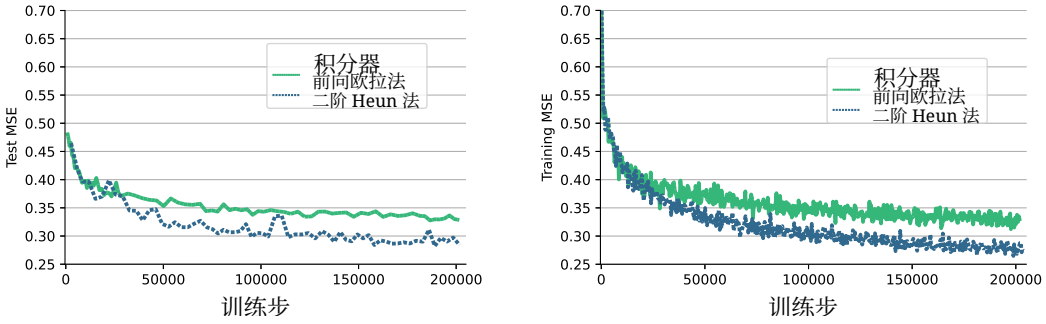
在第 2.2 节中，我们概述了分块训练。然而，像 MGN 这样的网格代理模型在超大域上的表现仍不明确。当域非常大时，在采用域分解的完整域上进行训练可能会造成浪费，因为必须遍历域中的每个补丁后才能更新权重。我们探究了这样一种可能性：采用分块训练或许同样有效，因为这种方法只需遍历域中少量样本补丁便可更新权重。

不同于先在较小域上训练、再推广到大型测试域（图 6），我们发现，与直接在大型域上训练相比，分块训练不会导致性能下降。这一结论适用于 CylinderFlow（图 8）和二维 C3FD（图 7）。基于此，我们首次使用分块训练，在包含数百万个节点的三维网格（三维 C3FD）上高效训练多重图网络（图 2）。为了优化该模型以实现实际部署，仍需开展进一步工作（例如超参数调优）。我们分别在附录 C 和 D 中提供了更多实验细节与结果。

4.2 高阶积分增强学习

基于第 2.3 节中的讨论，我们希望探究：FE 的 $\mathcal{O}(h^2)$ 截断误差有多显著？如果误差显著，MGN 在学习纠正这些误差方面有多有效？

[1]实际上，这一分析并不精确，因为 MGN 误差使用的计算流体力学数据仅近似于 $y(t+h)$。



Fi图 2：**在三维 C3FD 上进行分块训练**。实验使用的数据每个时间步包含 310 万个节点。

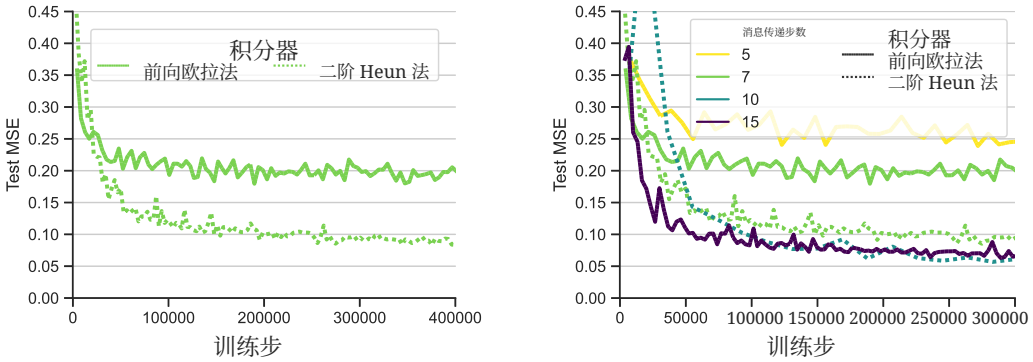


图 3：**高阶积分增强学习效果**。实验使用二维 C3FD 数据，每个时间步包含 15 万个节点。（左）采用 Heun 二阶方法（H2）且具有 7 个消息传递步数（MP）的 MGN，相较于前向欧拉法（FE），在测试数据（如图所示）和训练数据（见图 9）上的误差都能减半。（右）尽管增加消息传递步数有其优势，但具有 7 个消息传递步数的 H2 与具有 15 个消息传递步数的 FE 表现相当。具有 10 个消息传递步数的 H2 优于具有 15 个消息传递步数的 FE。图 10 显示，高阶积分的优势在不同训练运行中保持稳定。

虽然很明显，高阶积分方法应当能够减少截断误差，但如果 MGN 可以轻松纠正这些误差，那么这些高阶积分器可能无法带来益处。

使用二维 C3FD 数据，我们采用三种不同的积分器训练多重图网络：前向欧拉法（FE）、Heun 二阶方法（H2）和三阶 Heun 法（H3）。比较 FE 和 H2 可见，使用 H2 时测试误差大致可减半（图 3左）。增加消息传递步数通常会提升代理模型性能，因此我们还比较了这些积分器在不同消息传递步数下的表现（图 3 右）。使用 7 个消息传递步数的 H2 代理模型，其误差大致与使用 15 个步数的 FE 代理模型相同。此外，具有 10 个消息传递步数的 H2，其误差甚至低于具有 15 个消息传递步数的 FE。这表明，高阶积分是提升 MGN 性能的重要手段，而且还可用更少的消息传递步数达到相同的精度。附录 D 中提供了更多讨论和结果，附录 E 中回顾了相关工作。

5 讨论

我们推出了科学计算算法（SCALES2），以提升并扩展网格物理领域图神经网络代理模型的训练。我们的方法使 MGN 能够在包含数百万节点的三维网格上进行训练。我们还证明，在二维实验中，高阶积分可将误差减半。为推动代理模型模拟在扩展规模和提升性能方面的进一步研究，我们发布了 C3FD 基准数据集，其中包含基于二维和三维网格的 CO_2 捕集计算流体动力学模拟（分别包含 15 万和 310 万个节点）。由于本研究侧重于下一步预测，SCALES2 对大规模区域中的长序列滚动预测能带来多大整体提升，仍有待验证；我们在附录 F 中讨论了本研究的其他局限性和扩展方向。

致谢

本研究由劳伦斯利弗莫尔国家实验室代表美国能源部在合同 DE-AC52-07NA27344 项下开展。研究使用了国家能源研究科学计算中心（National Energy Research Scientific Computing Center）的资源。该中心是美国能源部科学办公室用户设施，由美国能源部科学办公室根据合同编号 DE-AC02-05CH11231 提供支持；本研究使用了 NERSC 项目编号 BES-ERCAP0024437 下的资源。本研究作为面向产业影响的碳捕集模拟（CCSI2）项目的一部分开展。

LLNL 发布编号（论文）：LLNL-CONF-845156。
PNNL 发布编号（数据集）：PNNL-SA-182276。

参考文献

Brian Bartoldson、Rui Wang、Yucheng Fu、David Widemann、Sam Nguyen、Jie Bao、Zhijie Xu 和 Brenda Ng。用于碳捕集设计优化的潜在空间模拟。*AAAI*人工智能大会论文集，36(11)：12447–12453，2022年6月。

Wei-Lin Chiang、Xuanqing Liu、Si Si、Yang Li、Samy Bengio 和 Cho-Jui Hsieh。Cluster-GCN：一种用于训练深层大型图卷积网络的高效算法。载于第*25*届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第257–266页，2019年。

Meire Fortunato、Tobias Pfaff、Peter Wirnsberger、Alexander Pritzel 和 Peter Battaglia。多尺度网格图网络。*arXiv* 预印本 *arXiv:2210.00612*，2022年。

Yucheng Fu、Jie Bao、Rajesh Singh、Chao Wang 和 Zhijie Xu。利用局部计算流体力学数据研究散堆填料塔中的逆流流动特征和液体滞留量。《化学工程科学》，221:115693，2020年。

Will Hamilton、Zhitao Ying 和 Jure Leskovec。大型图上的归纳式表示学习。神经信息处理系统进展，30，2017年。

Karl Heun 等人。关于近似积分独立变量微分方程的新方法。《数学物理杂志》，45:23–38，1900年。

Steeven Janny、Aurélien Bénéteau、Madiha Nadri、Julie Digne、Nicolas Thome 和 Christian Wolf。EAGLE：使用网格变换器大规模学习湍流流体动力学。发表于国际学习表征会议，2023年。

J.C.Butcher。《应用数值数学》，第20卷。爱思唯尔有限公司，1996年。

Zhihao Jia、Sina Lin、Mingyu Gao、Matei Zaharia 和 Alex Aiken。利用 ROC 提高图神经网络的准确性、可扩展性和性能。载于 I. Dhillon、D. Papailiopoulos 和 V. Sze（编），《机器学习与系统论文集》，第2卷，第187–198页，2020年。

Byungsoo Kim、Vinicius C. Azevedo、Nils Thuerey、Theodore Kim、Markus Gross 和 Barbara Solenthaler。Deep Fluids：用于参数化流体模拟的生成网络。《计算机图形论坛》，38(2)：59–70，2019年。

Marten Lienen 和 Stephan Günnemann。通过稀疏观测利用有限元网络学习物理系统的动力学。*arXiv* 预印本 *arXiv:2203.08852*，2022年。

Tobias Pfaff、Meire Fortunato、Alvaro Sanchez-Gonzalez 和 Peter Battaglia。使用图网络学习基于网格的模拟。载于国际学习表征会议，2020年。

Anthony Ralston。具有最小误差界的龙格-库塔方法。计算数学，16(80):431–437，1962年。

HS Tang、RD Haynes 和 G Houzeaux。《流体流动模拟中的区域分解方法综述：概念、算法与应用》。计算工程方法档案，28：841–873，2021年。

Pauli Virtanen、Ralf Gommers、Travis E. Oliphant、Matt Haberland、Tyler Reddy、David Cournapeau、Evgeni Burovski、Pearu Peterson、Warren Weckesser、Jonathan Bright、Stéfan J. van der Walt、Matthew Brett、Joshua Wilson、K. Jarrod Millman、Nikolay Mayorov、Andrew R. J. Nelson、Eric Jones、Robert Kern、Eric Larson、C J Carey、 ̇Ilhan Polat、Yu Feng、Eric W. Moore、Jake VanderPlas、Denis Laxalde、Josef Perktold、Robert Cimrman、Ian Henriksen、E. A. Quintero、Charles R. Harris、Anne M. Archibald、Antônio H. Ribeiro、Fabian Pedregosa、Paul van Mulbregt 和 SciPy 1.0 Contributors。《SciPy 1.0：Python 科学计算的基础算法》。*NatureMethods*，17：261–272，2020年。数字对象标识符：10.1038/s41592-019-0686-2。

Rui Wang. 动力系统的物理引导深度学习：综述。*arXiv* 预印本 *arXiv:2107.01272*，2021年。

Jared Willard、Xiaowei Jia、Shaoming Xu、Michael S. Steinbach 和 Vipin Kumar。将基于物理的建模与机器学习相结合：综述。*CoRR*，abs/2003.04919，2020年。

Ao Zhou、Jianlei Yang、Yeqi Gao、Tong Qiao、Yingjie Qi、Xiaoyi Wang、Yunli Chen、Pengcheng Dai、Weisheng Zhao 和 Chunming Hu。行业简报论文：优化边缘计算平台上图神经网络的内存效率。发表于 *2021 IEEE 第27届实时与嵌入式技术及应用研讨会 (RTAS)*，第445–448页。IEEE，2021年。

高阶积分分析

足够高阶的积分器实际上可以消除局部截断误差，这表明，基于此类积分器进行更新的 MGN，仅通过学习计算时间导数 $f(y)$，就能达到极高的精度。因此，我们探索将其作为提升代理模型性能的一种途径。在第 4.2 节和附录 D.4 中，我们训练了使用前向欧拉法（FE，一阶方法）以及 Heun 二阶（H2）和三阶（H3）方法的代理模型。
H2 方法如下：$y_{t+h} = y_t + \frac{h}{2}(f(y_t) + f(y_t + hf(y_t)))$。与 FE（公式 1）相比，我们发现，H2 和 FE 一样，要求 MGN 学习 $f(y)$，但 H2 的截断误差更低。当只能以符合物理规律的方式学习一阶导数信息时，高阶积分器或许能通过减小截断误差相对于一阶导数误差的尺度，使 MGN 更符合物理规律。

$$\begin{aligned} y(t+h) - y(t) &= \underbrace{hf(y(t))}_{\text{FE update}} + \underbrace{\mathcal{O}(h^2)}_{\text{truncation error}} \\ &= \underbrace{\frac{h}{2}\Big(f\big(y(t)\big) + f\big(y(t) + hf(y(t))\big)\Big)}_{\text{H2 update}} + \underbrace{\mathcal{O}(h^3)}_{\text{truncation error}} . \end{aligned} \tag{1}$$

使用基于 H2 的状态更新，MGN 的误差可以再次（参见第 2.3 节）分解为两项：一项是 MGN 预测（缩放）时间导数的能力所带来的误差，

$$\left[\frac{h}{2}\Big(f\big(y(t)\big) + f\big(y(t) + hf(y(t))\big)\Big) - \frac{1}{2}\Big(\text{MGN}(y(t)) + \text{MGN}\big(y(t) + \text{MGN}(y(t))\big)\Big)\right], \tag{2}$$

另一项是 $\mathcal{O}(h^3)$ 项，即三阶局部截断误差。请注意，尽管 MGN 仍然只需学习预测缩放时间导数，但由于截断误差项现在更小，基于 H2 的更新会使 MGN$(y(t))$ 模仿 $hf(y)$ 的益处增大。

因此，高阶积分可能通过明确消除无法通过基于前向欧拉的训练去除的 $\mathcal{O}(h^2)$ 截断误差，提供更准确的学习（最终误差更低）；因为原始 MGN 并不能保证学会抵消不断变化的 $\mathcal{O}(h^2)$ 误差。这些校正能否学会，取决于数据和代理模型的特性（例如，如果二阶导数可以学习，那么就有可能实现此类校正）。我们还注意到，高阶积分可能通过将学习重点放在 $f(y)$ 上（即提供更高的信噪比），实现更快的学习。

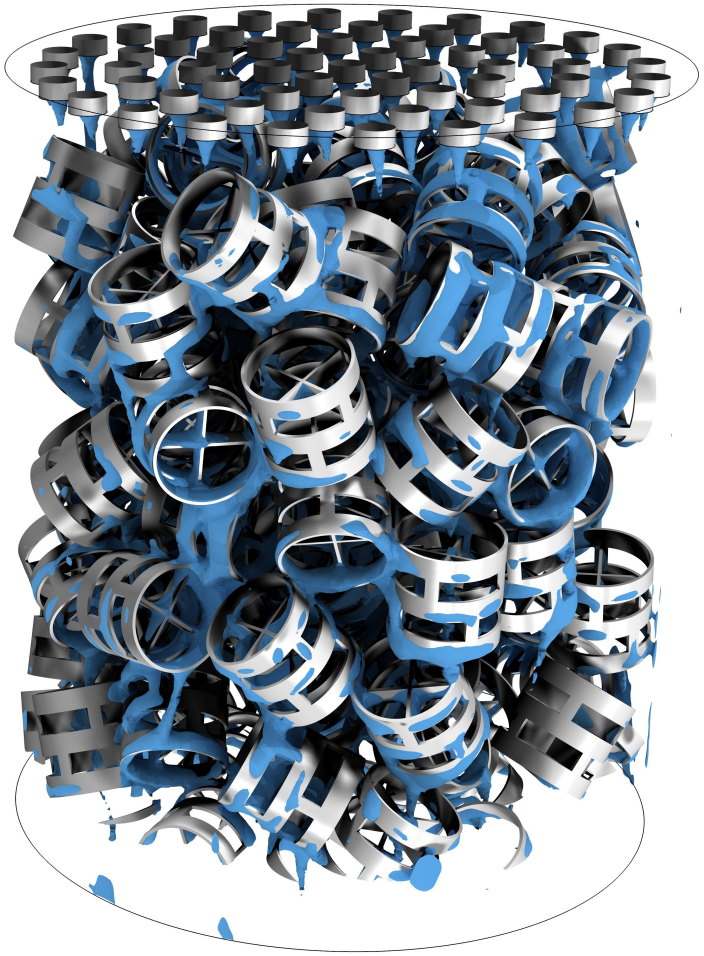


图 4：三维 C3FD 域的可视化。该域是一个宽 100 毫米、高 200 毫米、填充着鲍尔环的圆柱体。气体从下方进入并从上方流出，而液体则相反。

B 数据集详情

B.1 C3FD

图 4 展示了三维 C3FD 域，它是一个台架规模填料塔中 CO_2-捕集系统的代表性截面。二维域（示意图见图 1）有 150K 个多边形网格点，而三维域有 3.1M 个点（确切数量有所不同，因为每次模拟使用三个不同网格中的一个）。共有 50 次三维模拟和 50 次二维模拟。不同模拟对应于从域顶部入口流入的液体具有不同的速度；其他变量（如域的几何结构）在这 50 次模拟中保持不变。我们划分了训练集与验证集，以便通过验证测试外推和插值。验证模拟涵盖了入口速度最慢的四次、最快的四次以及处于中间水平的四次模拟。

C3FD遵循Fu等人（2020年）所述的设置：通过求解连续性方程和动量方程，并采用流体体积法模拟多相流体分离，生成了溶剂与气体逆流的欧拉模拟。使用STAR-CCM+ 进行求解

节点类型	二维数据边界索引	三维数据边界索引	接触角
流体	1.79769313486232e+308	1.79769313486232e+308	N/A
液体_入口	34	47	N/A
出口	31	46	N/A
气体_入口	30	45	N/A
入口_to_出口 _wall	29	44	180
side_wall	28	43	33.5
填料	27	42	33.5

表 1：C3FD 数据中边界索引到节点类型和接触角的映射。

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \mathbf{u} = 0, \tag{3}$$

$$\frac{\partial (\rho \mathbf{u})}{\partial t} + \nabla \cdot (\rho \mathbf{u}\mathbf{u}) = -\nabla p + \mu \nabla^2 \mathbf{u} + \rho \mathbf{g} + \mathbf{F}_\sigma, \tag{4}$$

其中$\mathbf{u}$ 表示速度， ρ 表示密度， p 表示压力，$\mathbf{g}$ 表示重力， μ 表示黏度，$\mathbf{F}_\sigma$ 表示气液界面处的表面张力。密度和黏度通过液体（α）和气相（1-α）的体积分数平均值计算。以下输运方程描述了 α的演化：

$$\frac{\partial \alpha}{\partial t} + \nabla \cdot (\mathbf{u}\alpha) = 0. \tag{5}$$

二维模拟域是三维域的一个横截面，但它是在独立的基础上求解的。对于二维情况，求解器使用的时间步长为 $h = 0.001s$，并且每十个时间步输出一次数据。对于三维情况，求解器使用 $h = 0.0002s$，并且每五十个时间步保存一次数据。因此，C3FD 中提供的二维和三维数据是使用 $h = 0.01s$ 保存的。

为了创建边，我们使用了 SciPy 对德劳内三角剖分的实现（Virtanen 等，2020年）。值得注意的是，我们在此过程后发现了一些碎片状单元；可以进一步进行网格细化或过滤来解决这一问题，例如采用 Lienen & Günnemann（2022年）所描述的过程。

计算域中每个节点均提供以下量：位置、速度、压力、液体体积分数和边界索引。对于速度，每个空间维度均提供一个值。我们没有将动量向量显式包含在数据集中，而是按以下方式计算：**动量**$_i = [\alpha\rho_L + (1-\alpha)\rho_G]\mathbf{u}_i$，其中 $\rho_L = 1010$ 和 $\rho_G = 1.18415$ 分别为液体和气体的密度。液体体积分数与捕集效率相关，因此从 CO_2 捕集的角度来看，它是该数据集中最重要的预测量。

数据中未明确提供节点类型和接触角信息，但可根据边界索引信息推导，如 Table1 所示。

C 实验详情

C.1 C3FD 实验详情

对于 C3FD，代理模型的输入包括：动量、液体体积分数、接触角和节点类型。我们还采用了与 Pfaff 等人（ 2020 ）相同的边特征。代理模型需要预测动量和液体体积分数的变化，而其他变量在每个节点处均保持不变。我们发现，将压力排除在模型输入和输出之外时，预测性能更好。

对于 MGN 输入噪声，我们为水平动量（二维中的 x 方向；三维中的 x、 y 方向）设置了 0.02，为垂直动量（二维中的 y 方向；三维中的 z 方向）设置了 0.03，并为液体体积分数设置了 0.05。这些数值并未经过充分优化，初步实验表明，使用较大的数值可能会带来性能提升。

Pfaff 等人（2020）发现，15 个消息传递步数能够取得良好的性能。在本研究中，我们也使用了 15 个消息传递步数，但三维实验（3 个消息传递步数）和高阶积分器实验除外。

（图例中有标示）。与 Pfaff 等人（2020）一样，我们使用 128 个潜在维度对二维数据集的节点和边进行编码。对于三维数据集，我们使用了 150 个潜在维度。

与 Pfaff 等人（2020）相同，我们使用 Adam 优化器和指数衰减的学习率训练模型，在 4,000,000 步内，学习率从 1×10^{-3} 衰减至 1×10^{-7}。通常，当性能趋于平稳时，我们会在第 400,000 步或更早停止训练，如图所示。

使用分块训练时，我们从不同模拟和时间步中随机抽取补丁。对于二维训练，我们创建了 12 个补丁（沿 100mm 宽度均匀划分为 3 个分区，沿 200mm 高度划分为 4 个分区）。对于三维训练，我们创建了 2,160 个补丁（沿每个水平轴均匀划分为 12 个分区，沿垂直轴划分为 15 个分区），其中 1,980 个非空。每个二维和三维补丁平均包含约 1×10^4 个节点。未来的工作或可采用比均匀空间划分更复杂的分解方案，以实现更好的三维负载均衡。

在采用分块训练的三维实验中，我们使用了 32 张 16 GB V100 GPU，每张 GPU 的批量大小为 1 个补丁。在采用分块训练的二维实验中，我们使用了 12 张 16 GB V100 GPU，每张 GPU 的批量大小为 1 个补丁。由于二维 C3FD 的计算域足够小，可以在某些 GPU 上完整运行，因此我们还使用了一张 80 GB A100 GPU，以批量大小为 1 个全域样本进行对比。需要指出的是，在二维 C3FD 中，MGN 的权重更新在分块训练和非分块训练时使用的节点数量大致相同。

验证误差基于随机选取的 1,000 个全域样本计算；但在二维和三维分块训练实验中，误差是基于 1,000 个补丁计算的。

C.2 CYLINDERFLOW 实验详情

我们使用 8 张 16 GB V100 GPU，在 CylinderFlow 数据集上进行了实验（Pfaff 等人，2020年）。每张 GPU 的批量大小设为 8（有效批量大小为 64）。对于分块训练，我们还将网格划分为四个大小相等的子区域。遵循 Pfaff 等人（2020年）的方法，我们使用 15 个消息传递步数、128 个潜在维度来编码节点和边，并将节点类型和动量作为输入，同时加入 0。02 的输入噪声。我们还使用了相同的边特征。我们预测动量，并采用与 C3FD 数据相同的学习率调度方法，在损失趋于平稳时停止训练。

D 支持性结果

D.1 滚动预测

图 5 展示了使用 MGN 进行滚动预测的若干帧，该模型采用了 15 个消息传递步数和有限元更新（见第 4.2 节和图 3）。在二维 C3FD 上完成训练后，我们在具有中间速度的测试模拟中评估了该模型（测试速度插值）。此外，我们还在一个新的二维域上对其进行了评估，该域是通过对三维 C3FD 域取新的横截面创建的。

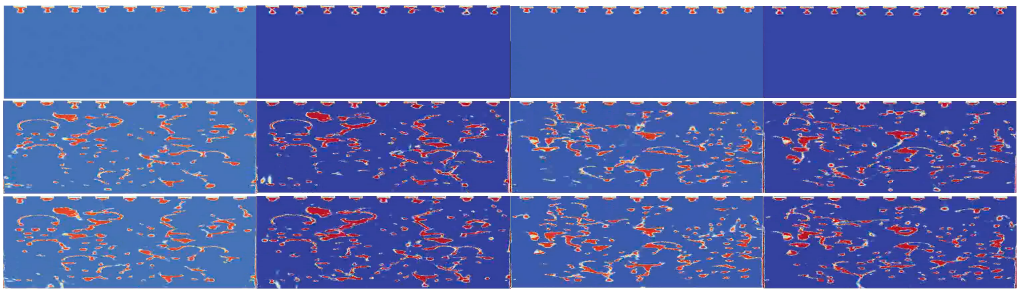


图 5：**MGN 在节点数超过 100K 的域上，即使泛化到新域，在较长时间跨度内仍具有稳定且视觉上准确的性能**。顶部、中部和底部各行分别对应时间步 30、250 和 500。从左到右，各列分别表示：MGN 在二维 C3FD 域上的滚动预测、该域的真实值计算流体动力学模拟、MGN 在新的二维域上的滚动预测（该列的填料配置有所不同），以及该新域的真实值计算流体动力学模拟。图中展示了液体体积分数。

D.2 在较大域上训练优于泛化到该域

通过域分解在大域上训练的一种替代方法，是直接在较小的域上训练。例如，Pfaff 等人（2020年）表明，多重图网络能够从包含约 2×10^3 个节点的训练网格泛化到包含约 2×10^4 个节点的测试网格。然而，目前尚不清楚依赖这种泛化能力是否足够，还是我们需要在尺度和形状与测试模拟相近的域上进行训练。关于这一问题的初步结果表明，改变域的形状会对性能造成虽小但值得注意的影响（Pfaff 等人，2020年）。

为说明分块训练的必要性，图 6 展示了这样一个结果：在二维 C3FD 域左侧三分之一部分上训练，其测试性能不如在完整域上训练。即使加速器无法容纳整个域，分块训练也能通过使在完整域上训练成为可能来解决这一问题。

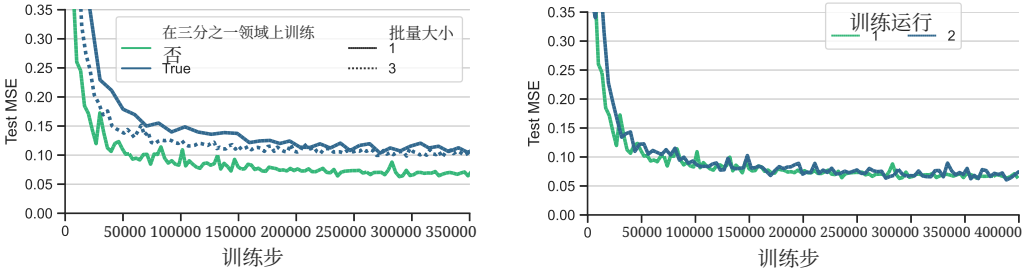


图 6：**从较小网格进行泛化不如在完整域上训练**。仅在我们的 150K 节点二维 C3FD 域左侧三分之一部分上训练，然后在测试时进行泛化，会导致更差的测试性能（左图）；无论是否调整批量大小以适应在子域上训练，结果都是如此。基线模型的训练在不同运行之间方差较低（右图）——两个基线模型达到的最小测试均方根误差的标准差为 0.003。相对于这一标准差，从较小网格进行泛化与基线方法之间的最小测试均方根误差差距（左图）达到 0.036，十分显著。

D.3 分块训练是进行大域训练的有效方法

我们证明，分块训练通过从不同模拟和时间步中随机抽取补丁来更新权重，其产生的误差与通过使用整个域更新权重的常规训练相当。这对 C3FD（图 7）和 CylinderFlow（图 8）均成立。

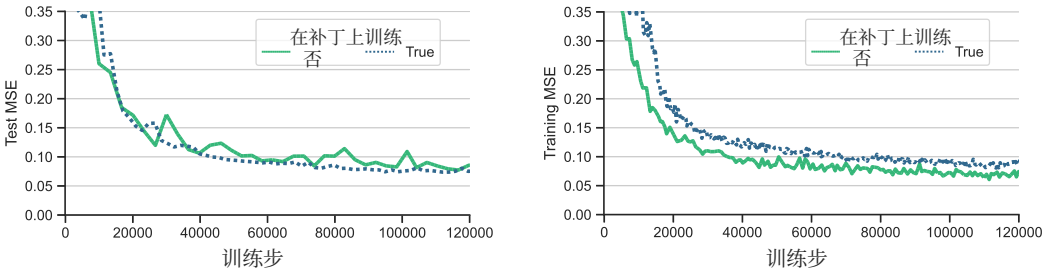


图 7：**分块训练可取代在完整域上的常规训练**。对通过域分解生成的补丁进行训练，能够获得相近的测试（左）和训练（右）性能，尽管分块训练使用的是从不同模拟和时间步中随机抽取的补丁。这表明，域分解有助于在通常会导致内存溢出错误的域上进行训练。实验基于二维 C3FD 模拟进行，每个时间步有 150K 个节点。

D.4 高阶积分的补充结果

在图 3 中，我们观察到 Heun 二阶方法有助于降低测试误差；图 10 显示，这种优势在不同训练运行中都很稳定。在图 9 中，我们还展示了这种优势同样延伸到了训练过程。值得注意的是，采用 Heun 三阶方法时，训练误差下降得甚至更快（见图 11）。

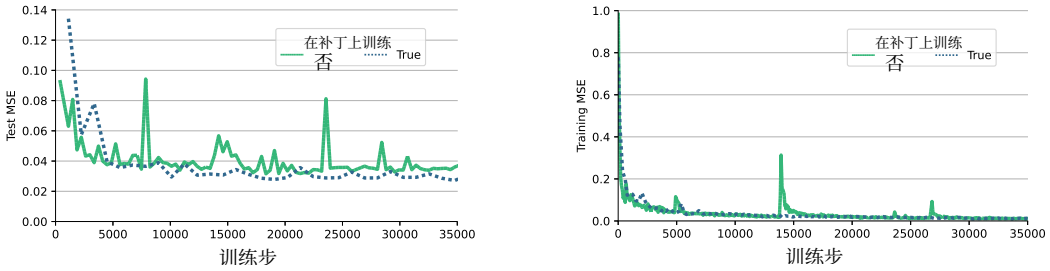


图 8：**分块训练对 CylinderFlow 有效**。在 CylinderFlow 模拟上进行分块训练，可降低测试集（左）和训练集（右）上的损失。

这与以下假设一致：像 MGN 这样的基于网格的图神经网络不仅难以消除 $\mathcal{O}(h^2)$，也难以消除 $\mathcal{O}(h^3)$ 局部截断误差。然而，我们还观察到，三阶积分器并不能带来最佳的测试误差，这表明高阶积分器需要额外的正则化才能发挥更好的作用。换言之，无法消除或无法学习的局部截断误差可能会在训练过程中提供正则化，而高阶积分器将其消除后，可能需要更多正则化才能获得良好的测试性能。

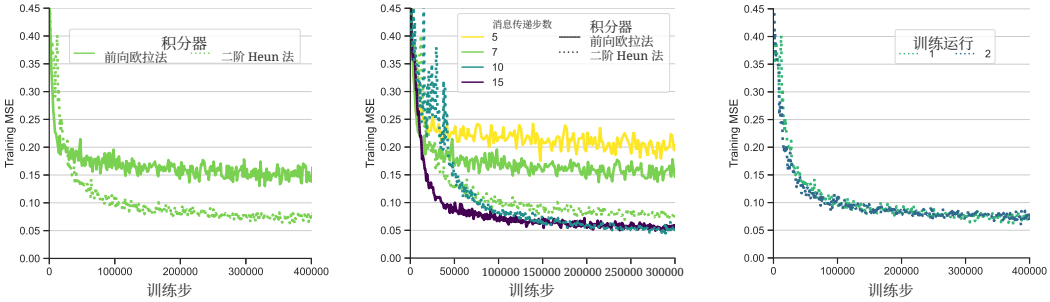


图 9：**高阶积分有助于利用训练数据进行学习**。（左）在采用 7 个消息传递步数的 MGN 上使用 Heun 二阶（H2）方法，与基线前向欧拉（FE）方法相比，训练误差减半。（中）尽管更多消息传递步数的优势显而易见（且已有文献记录），H2 方法配合 7 个消息传递步数的效果仍可媲美 FE 方法配合 15 个消息传递步数。（右）在两次训练运行中，高阶积分的优势均保持稳定。实验在二维 C3FD 上进行。

E 相关工作

此前已有研究探讨了用于类似大规模计算流体动力学模拟的神经网络代理模型（Bartoldson 等人，2022年），但据我们所知，此类场景中尚无研究使用能够跨领域泛化且忠实反映物理规律的图神经网络代理模型。具体而言，Pfaff 等人（2020年）证明，MeshGraphNets（多重图网络，MGN）能够从网格中学习可泛化的物理场更新规则；多尺度 MGN 在许多任务上达到了最先进的性能（Fortunato 等人，2022年），但其内存效率低下，因而无法应用于大型、与实际应用相关的图。

更广泛而言，如何将图神经网络扩展到大型图是一个重要课题，人们已通过多种方式加以应对，包括并行化、使用加速器外存储器（Jia 等，2020），以及更高效的学习算法。以往的算法研究通过采样（Hamilton 等，2017；Chiang 等，2019），或以其他方式修改低效图神经网络所计算的函数（Zhou 等，2021），提高了大型图上的运行效率——当平均节点度较高时，这些方法可能至关重要。相比之下，我们关注的是对物理保真度有严格要求的领域，因此近似方法的吸引力较低。[2] 因此，我们利用物理网格中适中的节点度这一特点，

[2]如果根据物理定律，节点 u 的状态会影响节点 v 的状态，那么代理模型似乎不应基于可能排除 u 的采样算法来学习 v 的更新规则。尽管采样方法与其他效率方法兼容，但为了避免损害我们所考虑的代理模型的归纳偏置，我们的研究并未采用基于采样的方法。

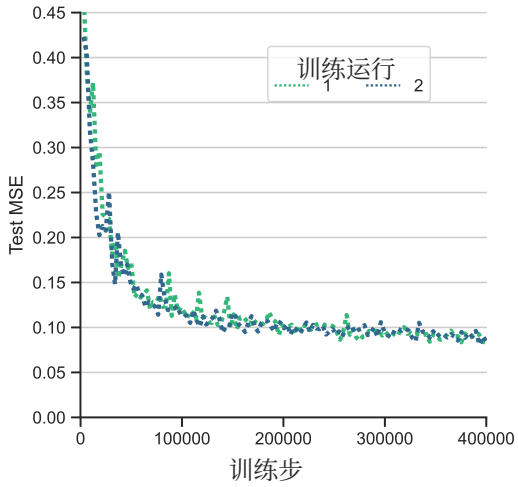


图 10： **高阶积分的优势在不同训练运行中表现稳定。**

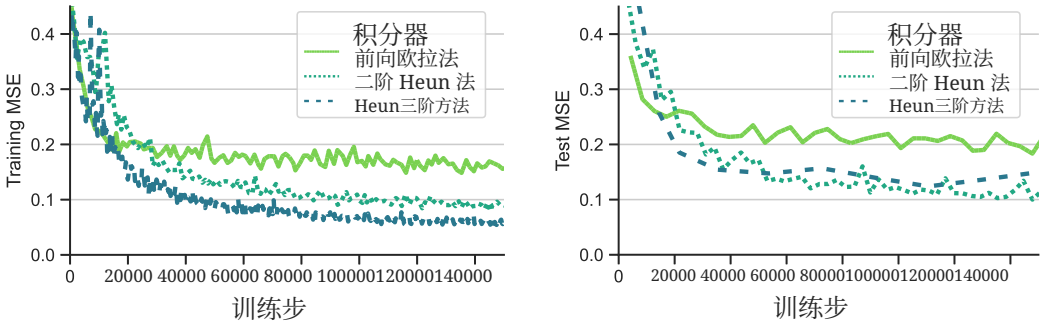


图 11： **三阶积分提升了训练性能，但未提升测试性能。**所有 MGN 模型均使用 7MP 步。

提出一种图神经网络物理代理模型扩展方法，该方法能够实现与在通常会导致内存溢出错误的网格上进行训练在数学上等价的训练。

在我们的工作之前，Janny等人（2023）推出了 EAGLE，这是一个用于物理代理模型学习的基准数据集，包含 1,000 多个二维模拟。EAGLE 中的网格平均每个时间步有 3,388 个节点，而我们的三维 C3FD 数据集在每个时间步使用包含 310 万个节点的网格（大 ∼ 1000 倍）。其他差异包括：我们的数据集使用与计算流体动力学求解器相同的网格，而 EAGLE 对计算流体动力学网格进行了大幅下采样；我们采用静态网格，而非动态网格；我们的计算流体动力学模拟涉及 CO_2 捕集塔器，这与理解捕集性能如何随模拟参数变化有关；我们的几何结构还包含多种以塔填料形式呈现的复杂障碍物。此外，Janny等人（2023）指出，可以通过使用注意力机制和图池化来减少消息传递迭代次数。相比之下，我们证明了可以通过使用高阶积分器来减少消息传递迭代次数。

基于网格的图代理模型此前已被用于生成时间导数估计，以应用于高阶积分方案（Lienen & Günnemann，2022年）。然而，这项先前的工作使用了嵌入微分方程求解框架（有限元方法）中的图神经网络，该框架赋予图神经网络所学习的规则以结构，而我们关注的更通用的网格代理图神经网络并不具备这种结构。此外，这项先前的工作将图神经网络插入运行多个时间步的微分方程求解器中，由此产生了高阶积分这一副产物。相比之下，我们通过分析高阶积分对学习任务以及任何预测时间导数的网络的归纳偏置所产生的影响，促使人们考虑高阶积分；同时，我们通过将高阶积分与基线一阶方法进行比较来检验相关假设。

F 局限性与未来工作

我们的工作仅初步探索了 SCALES2 改进网格图神经网络代理模型机器学习的各种机会。例如，或许可以在不产生其内存和计算开销的情况下获得高阶积分器的优势——预测–评估–校正（PEC）框架可能有助于在推理时间实现这一点。[3] 此外，采用其他积分器可能会获得更好的性能，例如 Ralston 方法能够最小化局部截断误差（Ralston，1962年）。与此相关的是，我们的结果表明，训练细节（例如正则化强度）可能需要进行调整，以匹配积分器的阶数。

需要特别指出的是，在消息传递步数固定的情况下，高阶积分器会增加加速器上的计算量，并且（如果加速器上的计算构成瓶颈）延长训练步的耗时。不过，它也可以减少达到目标准确率所需的训练步数。事实上，尽管我们没有对高阶积分带来的加速效果进行全面刻画，但在某些参数设置下，采用高阶积分确实能够提升训练速度。具体而言，我们的三维模型使用 H2 训练时，只需极少的训练步数就能达到目标误差水平，显著少于使用 FE 的模型（见图 2），而每个训练步的耗时相同（≈ 1.05 秒）：例如，使用 H2 的模型在约 12 小时（$\approx 40K$ 步）内即可达到 0.33 的测试均方误差，而使用 FE 的模型则约需 54 小时（$\approx 185K$ 步）。

我们工作的另一个局限是，为了使用较大的有效批量大小（32 个补丁）快速在三维 C3FD数据集上进行训练，我们使用了 32 个加速器。未来的工作可以考虑使用更小的批量大小（或许相应地采用更小的学习率），以降低加速器数量对于在超大计算域上训练类似 MGN 的模型的重要性，例如 C3FD 中的计算域。

此外，更复杂的域分解方法可能会带来收益。我们的方法会构造补丁，每个补丁都包含一个带幽灵区边界的子域，以便进行在数学上等同于直接在整个域上训练的训练。在补丁上训练比仅在子域上训练需要更多计算。此外，在我们的三维网格上，平均节点度高于二维网格（边由德劳内三角剖分生成）。这使子域的 k 跳邻域增长得更快；使用 2,160 个三维补丁时，当超过 $k = 7$，在 16GB 图形处理器上会出现内存溢出错误（图 12 展示了这种增长情况）。由于人们可能希望获得更大的 k 跳邻域，同时又不想增加加速器内存或创建更多、更小的补丁，未来的工作可以考虑使用最近邻算法创建边，以降低节点度。或者，采用一种根据生成的补丁所含节点数量而非子域空间范围来进行平衡的划分方法，也可能有所帮助。最后，也可以通过违反条件 1 来解决这一问题（即将 k 设为小于消息传递步数 m）：我们在 CylinderFlow 上进行的初步实验表明，只要子域中的大多数节点都能访问其真实的 m 跳邻居，使用较小的 k 时性能依然稳定。我们预计，进一步研究这些问题将有助于提升网格物理图神经网络代理模拟器的可扩展性和性能。

[3] 我们注意到，高阶积分的内存和计算开销可以通过减少消息传递步数来解决：例如，采用 15 个消息传递步数的前向欧拉法，其内存成本和性能大致与采用 7 个消息传递步数的 Heun 二阶方法相当。

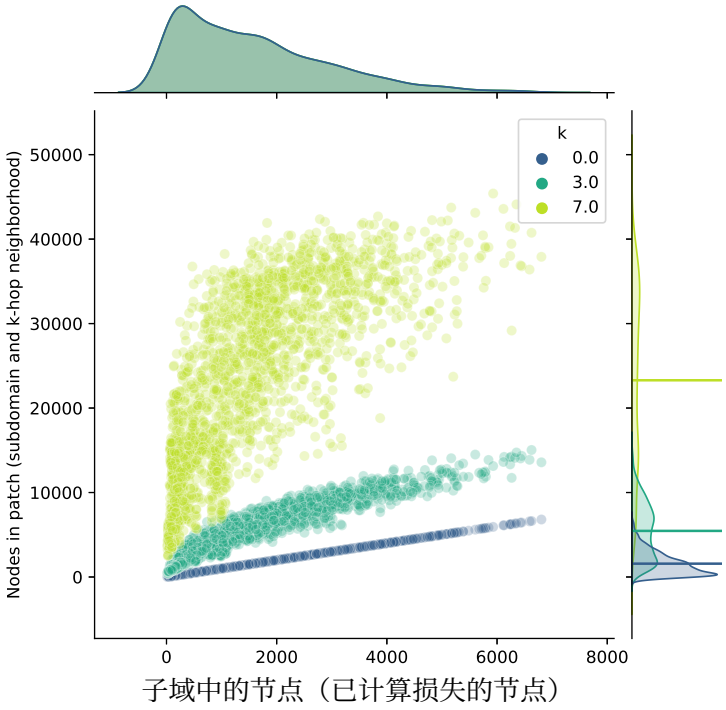


图 12：**对于三维 C3FD，增大 k 会导致补丁越来越大。** 每个补丁都包含其子域的 k 跳邻域。对于每个 k，右侧的水平条表示补丁中节点的平均数量，该数量会随 k 非线性增长。因此，随着消息传递步数 m增加，要满足第 2.2 节中的条件 1（$m \leq k$）需要越来越多的资源。