

IGG: 由变形空间中的测地动力学指导的图像生成

Nian Wu¹, Nivetha Jayakumar¹, Jiarui Xing¹, and Miaomiao Zhang^{1,2}

¹ Department of Electrical and Computer Engineering, University of Virginia, USA

² Department of Computer Science, University of Virginia, USA

摘要 生成模型最近在图像生成和编辑任务中获得了越来越多的关注。然而，它们通常缺乏与对象几何的直接联系，在计算解剖学、生物学和机器人技术等敏感领域这至关重要。本文提出了一种基于测地线动力学 (IGG) 的图像生成新框架。我们的 IGG 模型包括两个关键组件：(i) 一个高效的自动编码器，明确学习图像变换在潜在空间中的测地线路径；以及 (ii) 一个潜在的测地线扩散模型，捕捉文本指令条件下的测地线变形的潜在表示分布。通过利用测地线路径，我们的方法确保平滑、保持拓扑结构且可解释的变形，在保持几何一致性的同时捕获图像结构中的复杂变化。我们在植物生长数据和脑磁共振成像 (MRI) 上验证了提出的 IGG 模型。实验结果表明，与最先进的图像生成/编辑模型相比，IGG 在生成具有保留对象拓扑结构和减少伪影的真实、高质量图像方面表现出色。我们的代码可以在 <https://github.com/nellie689/IGG> 公开获取。

1 介绍

生成扩散模型近年来因其能够捕捉复杂的数据分布并产生高质量和逼真的样本而受到了广泛关注 [7,29]。在这些模型中，语言-图像生成模型 [34,36] 利用文本提示来引导扩散过程，使用户可以对图像进行高度定制化且具有语境意识的编辑。在医学成像中，这类模型被广泛用于创建多样化的数据集，改进下游任务如图像分类 [45,18]、分割 [1,10] 和对象识别 [12]。此外，它们在图像到图像翻译方面表现出色，解决了从低质量数据合成高质量医学扫描的关键挑战 [9,25] 以及模拟缺失模态的问题 [50,31]。凭借潜在提高疾病诊断准确性并支持个性化治疗计划的能力，生成扩散模型在推动医疗保健中的实际临床应用方面具有巨大潜力。

尽管取得了一定的成功，现有的扩散模型主要集中在操纵从图像中提取的强度和纹理特征上，通常忽略了在编辑或生成过程中与对象几何结构的直接联系 [30,24,13,20]。这种限制可能导致不现实或生物学上不可信的样本，对需要准确保留几何结构的下游任务构成风险 [4]。最近的研究进展引入了

简单的几何约束条件，如形状定位 [32,11]、边界条件 [26] 和 3D 形状先验知识 [48,17] 到建模过程中。然而，这些方法缺乏精细的结构保护，在计算解剖学、机器人技术以及医学成像等敏感领域中这是非常需要的。当前扩散模型的另一个重要但尚未充分研究的限制是它们无法提供解释性指标或图像对象拓扑完整性的定量度量。常用的评估指标包括用于评估真实数据和生成数据之间分布相似性的 Fréchet Inception Distance (FID) [14]，用于衡量多样性和质量的 Inception Score (IS) [37]，以及用于评估像素级保真度的 Structural Similarity Index (SSIM) [41]。虽然这些指标能有效地评估视觉质量，但它们没有考虑到物体几何形状和拓扑结构的保存问题，在结构敏感的应用中引发了关于其可靠性和可信度的问题。

为了克服这些限制，我们引入了一个新的框架，基于测地线动力学的图像生成 (IGG)，首次通过沿着变形空间中的随机测地线在给定模板上进行形变来根据文本指令生成图像。与当前的方法 [8,16,21] 不同，我们的 IGG 通过学习微分同胚变换（即一对一、平滑且可逆的映射）将每个生成的图像视为模板或参考图像的变形变异，从而确保拓扑一致性。在扩散过程中将明确学习到一条动态测地线，产生自然变化并提供解释性的度量来量化拓扑变化。我们提出的 IGG 模型包含两个关键组件：一个由测地线动力学指导的微分同胚变换潜在表示学习框架，以及一个基于用户定义文本指令条件化的潜在几何扩散模型。总结来说，IGG 的贡献有三点：

- (i) 开发一种新型图像生成模型，通过图像变形的测地线学习与强大的条件扩散过程相结合来融入几何原理。
- (ii) 启用量化指标来评估和确保生成样本的拓扑一致性。
- (iii) 建立一个全新的范式，用于生成解剖结构完整且逼真的图像样本，这可能推动需要结构保真度和精度领域的应用发展。

我们展示了 IGG 在一系列真实世界数据集上的有效性，包括 [38] 芜菁生长数据和 [23] 脑部 MRI。我们将模型的性能与最先进的文本引导生成模型进行了比较。实验结果表明，IGG 在生成具有良好保存拓扑结构的图像样本方面取得了显著改进。

2 背景：变形空间中的测地线

在本节中，我们简要回顾了变形空间中测地线的基本概念，这提供了一种原则性的方式来建模图像之间平滑自然的变形，同时保持结构完整

性 [28]。假设一个通用类别的对象（例如，人类的大脑、心脏或肺）可以被描述为给定模板的变形变体，在这种情况下，该类别的描述符可以通过沿着**测地线**—具有最小能量变换的最短路径 [3,19] 将模板变形到其他图像中自然产生。理论上，通过强制产生的变换为微分同胚（即可微、双射映射且逆映射也是可微的）[5,2,28]，可以保持变形模板的所有拓扑性质。违反这些约束会导致图像伪影，例如撕裂、交叉或自我穿透。

令 $\text{Diff}^\infty(\Omega)$ 表示在 d 维环面域 $\Omega = \mathbb{R}^d/\mathbb{Z}^d$ 上的光滑微分同胚空间。微分同胚的空间的切空间是 Ω 上的光滑向量场空间 $V = \mathfrak{X}^\infty(T\Omega)$ 。考虑一个时变速度场 $\{v_t\} : [0, \tau] \rightarrow V$ ，我们可以通过求解

$$\frac{d\phi_t}{dt} = v_t(\phi_t), t \in [0, 1], \quad (1)$$

来生成成对图像之间的微分同胚 $\{\phi_t\}$ ，其中 ϕ_0 是恒等映射， ϕ_1 是目标变换。

测地线最小化泛函 $\int_0^1 (\mathcal{L}v_t, v_t) dt$ [5]，其中 $\mathcal{L} : V \rightarrow V^*$ 是一个对称正定的微分算子，将切向量 $v(t) \in V$ 映射到其对偶空间作为动量向量 $m(t) \in V^*$ 。我们通常写为 $m(t) = \mathcal{L}v(t)$ ，或 $v(t) = \mathcal{K}m(t)$ ，其中 $\mathcal{K}$ 是 $\mathcal{L}$ 的逆算子。在本文中，我们采用常用的拉普拉斯算子 $\mathcal{L} = (-\alpha\Delta + \text{Id})^3$ ，其中 α 是一个控制变换场平滑度的加权参数，而 Id 是单位矩阵。 $(\cdot, \cdot)$ 是一个对偶配对，类似于向量之间的内积。根据一个著名的测地线射程算法 [40]，上述泛函的最小值通过求解带有给定初始条件的欧拉-庞加莱微分 (EPDiff) 方程 [40,47] 唯一确定。也就是说，对于 $\forall v_0 \in V$ ，可以通过向前射击 EPDiff 方程来计算形变映射空间中的测地线路径 $t \mapsto \phi_t$

$$\frac{\partial v_t}{\partial t} = -K [(Dv_t)^T m_t + Dm_t v_t + m_t \text{div} v_t], \quad (2)$$

其中 D 表示雅可比矩阵， div 是散度算子。

从图像中推导测地线。考虑源图像 S 和目标图像 T ，它们在域 $\Omega(S(x), T(x) : \Omega \rightarrow \mathbb{R})$ 中定义。测地变换的优化可以表述为在初始速度上最小化一个能量函数，且满足 EPDiff 方程，即

$$E(v_0) = (\mathcal{L}v_0, v_0) + \lambda \text{Dist}(S(\phi_1), T) \text{ s.t. Eq. (1) \& (2)}.$$

这里， $\text{Dist}(\cdot, \cdot)$ 是一个度量图像间差异的距离函数，而 λ 是一个正的加权参数。本文中，我们将使用常用的强度平方和差 [5,42,49]。

3 我们的方法：IGG

本节介绍了我们提出的模型 IGG，该模型首次在变形空间中根据文本指令生成图像。动态测地线将在扩散过程中被显式学习，产生自然变换并提供可解释的度量以量化拓扑变化。我们的 IGG 模型由两个主要组件组成：(i) 基于自编码器的可变形注册网络，该网络学习受测地动力学启发的微分同胚变换的潜在表示；以及 (ii) 一个潜歧行波扩散模型，该模型捕获基于用户定义文本输入条件下的连续变形特征的潜在分布。我们的模型概述如图 1 所示。

3.1 网络架构

学习变形空间中测地线的潜在表示。对于与文本指令 $\{s^n, f^n, \Lambda^n\}_{n=1}^N$ 关联的 N 模板和目标图像，我们采用基于自编码器的可变形配准网络来学习由速度场 $\{v_t\}$ 参数化的测地变换的潜在表示，即在模板 s^n 和目标 f^n 之间的变换。该网络的编码器， $\mathcal{E}_{\theta_v}$ ，将输入数据映射到速度场序列的潜在表示， $\{z_{v_t}\}$ 。受最近的工作 NeurEPDiff [43] 启发，该工作开发了一种神经算子用于学习 EPDiff 方程的映射函数，我们在模型 IGG 中集成了 NeurEPDiff 模块， $\mathcal{G}_{\theta_r}$ ，以在潜在变形空间中学习测地线。更具体地说，从一个潜在初始速度 z_{v_0} 开

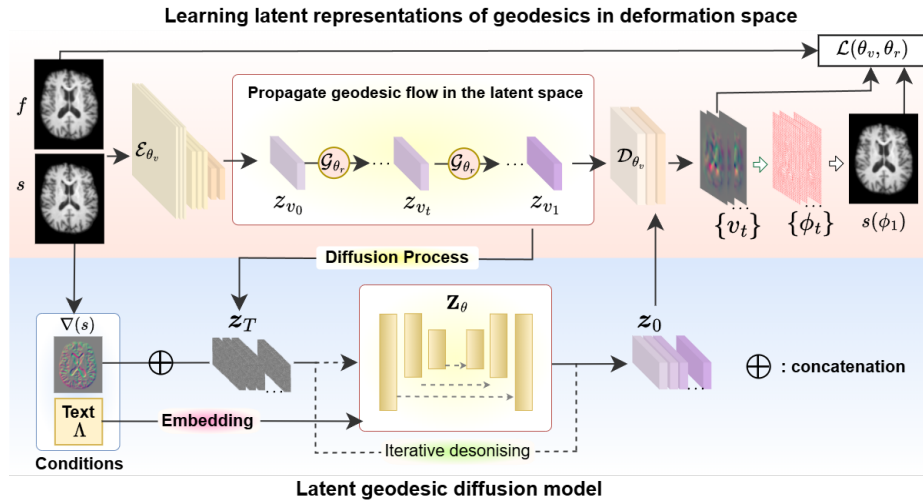


图 1. 提出的 IGG 模型概述。

始, 操作符 $\mathcal{G}_{\theta_r}$ 沿着测地路径迭代传播它, $z_{v_0} \mapsto \cdots z_{v_t} \mapsto \cdots z_{v_1}$ 。在每个时间点 t , 采用多层神经网络来模拟从 z_{v_t} 到 $z_{v_{t+1}}$ 的测地线映射函数。每隐藏层结合局部线性变换 W^j 和全局卷积核 $\mathcal{H}^j$ 来提取全局和局部表示。此外, 使用一个平滑非线性激活函数 $\sigma(\cdot)$ 来鼓励测地路径的平滑演化。该过程定义为

$$z_{v_{t+1}} := \sigma(W^J, \mathcal{H}^J) \circ \cdots \circ \sigma(W^2, \mathcal{H}^2) \circ \sigma(W^1, \mathcal{H}^1, z_{v_t}),$$

其中 $\circ$ 表示网络操作的组合。解码器 $\mathcal{D}_{\theta_v}$ 然后将这些潜在表示投影回全维图像空间, 以生成速度场的测地线流 $v_0 \mapsto \cdots v_t \mapsto \cdots v_1$, 并根据方程 1 所述将其积分以生成相应的变形 $\phi_0 \mapsto \cdots \phi_t \mapsto \cdots \phi_1$ 。

网络损失结合了无监督配准损失和由 EPDiff 方程的数值解通过欧拉积分引导的测地线损失的贡献。这些解表示为 $\{\hat{v}_t^n\}, t \in (0, 1]$, 同时学习初始速度 $\hat{v}_0^n$ 。为了简化符号, 我们在公式化损失函数之前定义 $\hat{\mathbf{v}}^n \triangleq \{\hat{v}_t^n\}$ 为

$$\begin{aligned} \mathcal{L}(\theta_v, \theta_r) = & \sum_{n=1}^N \lambda \| (s^n(\phi_1^n(\theta_v)) - f^n) \|_2^2 + \frac{1}{2} (\mathcal{L}v_0^n(\theta_v), v_0^n(\theta_v)) \\ & + \eta \| \mathcal{D}_{\theta_v}(\mathcal{G}_{\theta_r}(z_{v_0})) - \hat{\mathbf{v}}^n \|_2^2 + \text{Reg}(\theta_r, \theta_v), \quad s.t. \text{ Eq. (1)}, \end{aligned} \quad (3)$$

其中 λ 和 η 是正权重参数, 用于平衡图像匹配项和测地线损失之间的关系, 而 $\text{Reg}(\cdot)$ 是一个网络正则化项。

潜在测地线扩散模型。预训练了测地线学习网络之后, 我们现在准备引入一个潜在几何扩散模块来模拟图像变换的潜在测地流分布, $\{z_{v_t}^n\}$ 。为了简化表示, 我们定义 $\mathbf{z}^n := \{z_{v_t}^n\}_{t=0}^1$ 为第 n 个主题, 其中所有时间相关的成分 $\{z_{v_t}^n\}$ 都被连接起来形成一个统一的表示。与之前在图像空间 [8,16,21] 中执行扩散过程的方法不同, 我们的潜在几何扩散是在变形空间中进行操作。具体来说, 我们的扩散过程直接采样保持拓扑同胚变换的测地线, 然后将这些测地线应用于输入模板以生成最终输出。

几何条件在潜在测地线空间中的应用。遵循视频扩散模型的原则 [16], 我们通过逐步添加噪声 $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ 到初始的潜在表示 $\mathbf{z}_0^n \sim p(\mathbf{z}_0^n)$, $\mathbf{z}_0^n \leftarrow \mathbf{z}^n$ 来进行前向扩散过程, 这一过程共进行 T 步。前向扩散过程导致了一个分布 $q(\mathbf{z}_\tau^n | \mathbf{z}_0^n)$, 其中 $\tau \in [1, \cdots, T]$ [16]。

在反向扩散过程中, 我们首先在潜在测地线空间中引入了一种新颖的几何条件。这使得我们的模型能够更有效地适应特定上下文; 因此根据观察

到的模板图像和提供的文本指令来提高其预测性能。更具体地说，我们的几何条件是通过拼接三个组件实现的：图像嵌入 - 下采样的模板图像梯度 $\{\nabla s^n\}$ ；文本嵌入，使用 CLIP 文本编码器 [33] 编码， $\{A^n\}$ ；以及测地线变形嵌入，由潜在速度场表示， $\{z_\tau^n\}$ 。反向扩散过程随后涉及采样去噪的潜在测地线流的速度场 z_0^n ，从第 τ 个时间步长的噪声状态 z_τ^n 开始，并逐步去除噪声以重构原始潜在表示。该过程用数学表达式表示为

$$p_\theta(z_{\tau-1}^n | z_\tau^n, \nabla s^n, A^n, \tau) = \mathcal{N}(z_{\tau-1}^n; \mu_\theta(z_\tau^n, \nabla s^n, A^n, \tau), \Sigma_\theta(z_\tau^n, \nabla s^n, A^n, \tau)), \quad (4)$$

其中 μ_θ [16] 和 Σ_θ 分别表示逆扩散过程中每一步高斯分布的均值和方差。参数 μ_θ [16] 和 Σ_θ 是学习函数，逆过程实际上使用了一个去噪网络 $\mathbf{Z}_\theta$ 实现。在我们的实验中，我们采用了一个 3D UNet 架构 [35] 作为 $\mathbf{Z}_\theta$ 的主干网络，该网络预测要从 z_τ^n 中移除的噪声成分以重构 $z_{\tau-1}^n$ 。

我们提出的潜在测地线扩散模块的网络损失被定义为

$$\mathcal{L}_\theta = \frac{1}{N} \sum_{n=1}^N \|\epsilon_\tau^n - \mathbf{Z}_\theta(z_\tau^n, \nabla s^n, A^n, \tau)\|_2^2 + \text{reg}(\theta), \tau \in \mathcal{U}[1, T], \quad (5)$$

其中 $\text{reg}(\cdot)$ 是网络参数 θ 的正则化， $\mathcal{U}$ 表示均匀分布。

为了在训练后的潜在空间中平衡样本质量和速度场多样性之间的权衡，我们优化了条件扩散模型和无条件扩散模型的组合。这是通过利用尺度为 δ_I 的图像条件引导、尺度为 δ_T 的文本条件引导以及零条件 $\emptyset$ 来实现的。遵循分类器自由引导方法 [15] 中的类似原则，我们首先联合训练一个条件和无条件的 IGG 模型。然后将这些模型得出的分数估计结合起来预测噪声 $\hat{\mathbf{Z}}$ 。

$$\begin{aligned} \hat{\mathbf{Z}}_\theta(z_\tau^n, \nabla s^n, A^n, \tau) &= \mathbf{Z}_\theta(z_\tau^n, \emptyset, \emptyset, \tau) + \delta_I \cdot [\mathbf{Z}_\theta(z_\tau^n, \nabla s^n, \emptyset, \tau) - \mathbf{Z}_\theta(z_\tau^n, \emptyset, \emptyset, \tau)] \\ &\quad + \delta_T \cdot [\mathbf{Z}_\theta(z_\tau^n, \nabla s^n, A^n, \tau) - \mathbf{Z}_\theta(z_\tau^n, \nabla s^n, \emptyset, \tau)] \\ &= (1 - \delta_I) \cdot \mathbf{Z}_\theta(z_\tau^n, \emptyset, \emptyset, \tau) + (\delta_I - \delta_T) \cdot \mathbf{Z}_\theta(z_\tau^n, \nabla s^n, \emptyset, \tau) \\ &\quad + \delta_T \cdot \mathbf{Z}_\theta(z_\tau^n, \nabla s^n, A^n, \tau). \end{aligned} \quad (6)$$

3.2 网络优化

我们首先优化测地变换表示学习的损失（公式 (3)），然后是潜在测地扩散模型（公式 (5)）。IGG 的训练过程总结在算法 1 中。在测试过程中，给定模板图像和文本指令条件下采样潜在速度场的测地线的过程概述在算法 2 中。

Algorithm 1 IGG 培训	Algorithm 2 IGG 采样
1: 输入: 数据 $\{s^n, f^n, \Lambda^n\}_{n=1}^{N_{train}}$	1: 输入: 数据 $\{s^n, \Lambda^n\}_{n=1}^{N_{test}}$
2: 预训练 $\mathcal{E}_{\theta_v}, \mathcal{E}_{\theta_r}, \mathcal{D}_{\theta_v}$ 通过公式 3。	2: 初始化 $z_T^n \sim \mathcal{N}(0, \mathbf{I})$
3: repeat	3: for $\tau = T, \dots, 1$ do
4: $z_0^n \leftarrow \mathcal{G}_{\theta_r}(\mathcal{E}_{\theta_v}(s^n, f^n))$	4: $\hat{z}_{\tau-1}^n \leftarrow p_{\theta}(z_{\tau-1}^n z_{\tau}^n, \nabla s^n, \Lambda^n, \tau)$
5: 样本 $\tau \sim \mathcal{U}(1, T), \epsilon \sim \mathcal{N}(0, \mathbf{I})$	5: end for
6: $z_{\tau}^n \leftarrow q(z_{\tau}^n z_0^n)$	6: $\{\hat{v}^n\} = \mathcal{D}_{\theta_v}(\hat{z}_0^n)$
7: 对等式取梯度下降步)	7: $\{\phi^n\} \leftarrow \{\hat{v}^n\}$ 通过式. 1
8: until converge	8: return $\{\phi^n\}$

4 实验评估

我们使用一套多样化的现实世界图像数据集来评估所提出的 IGG 框架，这些数据集捕捉了随时间变化的可变形形状变化。我们首先通过将学习到的测地线潜在表示与通过欧拉积分获得的 EPDiff 方程数值解进行比较，来评估其质量。

接下来，我们评估由采样测地线引导的给定文本指令变形生成图像的质量。这些结果与合成图像序列/视频的状态-of-艺术生成模型进行了比较，这些模型使用公开提供的训练代码或微调选项进行训练，包括视频扩散模型 (VDM) [16]，CogVideoX [46] 和 DynamiCrafter [44]。请注意，所有基线都是使用沿 EPDiff 方程数值解导出的测地变换生成的变形图像序列进行训练的。这种方法确保了结果独立于 IGG 框架第一模块预测的学习测地线，提供了稳定且无偏的评估。

4.1 数据集

芜菁甘蓝植物。我们包含了来自公开数据存储库 [38] 的 300 帧 RGB-D 标签图，代表五株芝麻菜植物。这些标签图涵盖了从芽中长出并随时间增长的不同叶子，其中植物的生长情况在 228-236 小时之间进行了监测。所有数据帧均重采样至 128^2 并通过仿射变换进行预对齐。

纵向脑部 MRI。我们共包括了来自公开影像研究系列 (OASIS-3) 数据集的 2618 个 T1 加权纵向 (时间序列) 脑部 MRI 图像 [23]。本实验旨在使用包含不同时间间隔扫描的纵向数据来验证我们的方法，这些数据涵盖了健康受试者和阿尔茨海默病 (AD) 患者，年龄范围为 60-90 岁。鉴于许多现有的带有

文本指令的图像生成/编辑方法主要关注 2D 自然图像 [8,27]，我们特别利用从该 3D 脑数据中提取的 2D 扫描图与最先进的基线进行比较。所有 MRI 均经过预处理，包括调整大小至 128^2 ，具有各向同性的体素为 1mm^2 ，去除颅骨、强度归一化、偏置场校正以及使用仿射变换进行预对齐。

文本指令。我们的文本条件包括模板图像的描述，随后是对其应用的具体编辑或进展的细节。对于脑部 MRI，这包括年龄、性别和性别的生物变量。相比之下，对于植物，它主要关注生长时间线。所有信息均直接来源于数据存储库中的原始数据集。

4.2 实验设计与实现细节

评估学习到的测地线潜在表示。我们通过将学习到的测地线与使用欧拉积分求解 EPDiff 方程得到的数值解进行比较，评估模型预测测地线动力学的能力。评估侧重于两种方法之间速度场、变换和变形图像中的误差。在每个时间积分步骤中计算定量指标，如速度场的平均绝对误差。

评估生成的样本。我们将由 IGG 生成的样本质量与三个基线模型进行比较：VDM [16]，CogVideoX [46] 和 DynamiCrafter [44]。特别地，我们首先使用标准图像合成指标评估每个时间步长的单独合成图像，包括 Fréchet Inception Distance (FID) [14] 和 Kernel Inception Distance (KID) [6] 来测量生成图像和真实图像之间的分布相似性。此外，我们使用结构相似性指数 (SSIM) [41] 来评估结构、对比度和亮度的相似性。然后，通过计算 Fréchet 视频距离 (FVD) [39] 并提供定性的视觉比较来评估生成的视频序列的连续性和平滑度。最后，我们使用 Inception 分数 (IS) [37] 来评估生成样本的感知质量。

评估对象的几何和拓扑保真度。为了展示我们的 IGG 模型在保持精细几何结构和拓扑方面的效率，我们首先通过使用 IGG 产生的采样测地流变形来变形一个模板图像生成图像序列，并将这些结果与所有基准模型进行比较。其次，我们展示了变形对应的雅可比行列式 (DetJac) 图。DetJac 值揭示了体积变化的重要模式：1 的值表示没有体积变化， $\text{DetJac} < 1$ 反映体积收缩，而 $\text{DetJac} > 1$ 意味着体积膨胀。DetJac 值低于零表明变换场中的伪影或奇点，突出显示未能保持拓扑完整性。请注意，只有我们的 IGG 模型能够使这

种度量量化生成样本中的拓扑变化。

评估模型预测的可靠性。我们评估了 IGG 在学习几何形状随时间变化的测地线变形时的可靠性和置信度，通过计算单个采样图像序列沿测地路径的像素级均值和标准差。我们将置信区间可视化以突出显示生长模式中 95% 确定性的区域。这些区间由生成样本的均值定义，1000 的像素级边界为均值上下两个标准差。

参数设置。我们将所有数据集分为 80%、10% 和 10%，用于训练、验证和测试。所有实验均在 NVIDIA A100 GPU 上进行。对于地测学习的自编码器训练，我们交替更新编码器-解码器和潜在的地测神经算子，并最终在每个阶段联合训练两个组件 2000 轮。我们将权重衰减设置为 $1e^{-4}$ ，批量大小设置为 64，学习率设置为 $5e^{-4}$ 。对于训练潜在测地线扩散模型，我们将学习率设置为 $1e^{-4}$ ，epochs 数量设置为 6000，批量大小设置为 36，扩散步数设置为 500。所有模型均使用 Adam 优化器进行训练 [22]。

4.3 实验结果

图 2 的左面板可视化了 IGG 自编码器预测的变换、速度场和变形图像，并与 EPDiff 方程的数值解进行了比较。这些方法显示出高度的相似性。图 2 的右面板展示了真实数值解与 IGG 的潜在表示学习模块在每个时间步长之间差异的定量结果。这些结果共同突出了 IGG 有效学习类似于真实数值解的测地映射函数的能力。

图 3 比较了我们的模型 IGG 生成的图像（连同变形相关的 DetJac 地图）与基线生成模型。可视化图像和 IGG 的 DetJac 值表明，我们的模型有效地保持了生成图像中对象的拓扑结构，并且很好地捕捉了几何形状随时间的变化过程。相比之下，基线生成的样本无法维持各种结构的几何完整性。例如，在植物生长的图像中，叶子似乎不自然地合并或重叠，破坏了它们的生物拓扑结构。同样，基线模型对海马体的部分预测不准确，导致生成的区域与原始脑部结构不符。

表 1 报告了我们模型 IGG 和基线生成的图像的质量和多样性评估指标。虽然所有模型都实现了相似的 IS，但 IGG 显示出显著较低的 FVD、FID 和 KID 分数以及较高的 SSIM 分数。这表明尽管所有模型都能生成多样化的图像（从而获得良好的 IS），基线模型缺乏现实特征或与真实图像适当分布

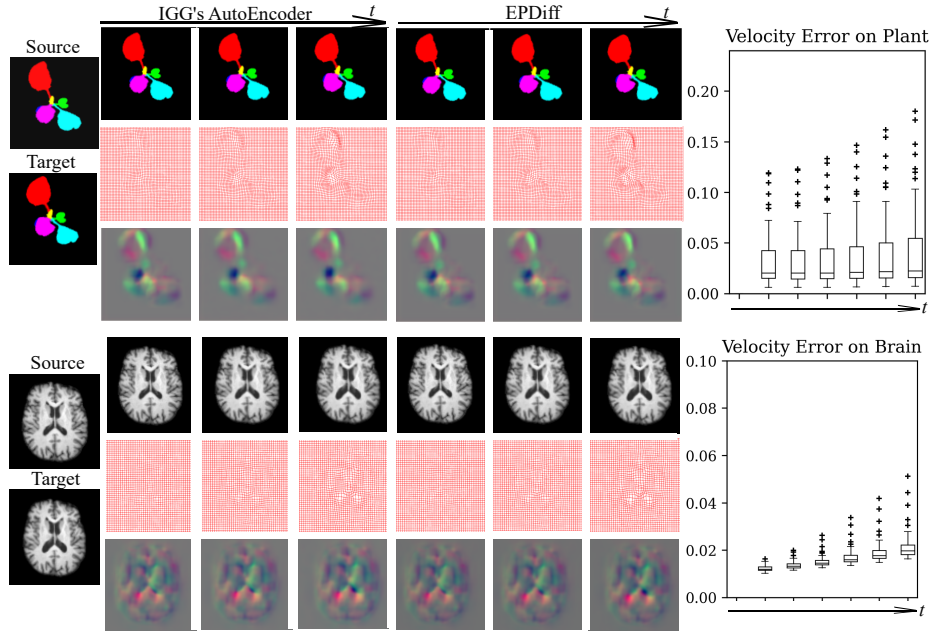


图 2. IGG 预测的测地线与 EPDiff 方程实际数值解的比较。左：时间线上预测变形图像、变形和速度的可视化。右：预测速度随时间的平均绝对误差与 EPDiff 的数值积分比较。

对齐的能力。观察到 IGG 得分大约高出 10 倍，证明了其生成更逼真图像的能力。

图 4 可视化了由 IGG 生成的植物生长和大脑进展图像的示例置信度图。这表明我们的模型能够有效地捕捉植物随时间的增长模式，主要关注叶片的边界。大脑图像的置信度图突出了脑室中扩张的模式，展示了该模型捕捉动态变化的能力。

5 结论与讨论

本文介绍了 IGG，一种新颖的图像生成方法，它通过在变形空间中学习到的随机测地线来变形给定的模板，受到文本指令的引导。与当前专注于操纵图像强度和纹理的生成模型不同，IGG 在扩散过程中明确学习图像变换的动态测地线。这使得自然、可解释的转换成为可能，并提供了评估生成样本中的拓扑变化的数量化指标。我们的主要贡献包括一个受测地线启发的网络，用于学习时间依赖的微分同胚变换的潜在表示，以及一个潜几何扩散模

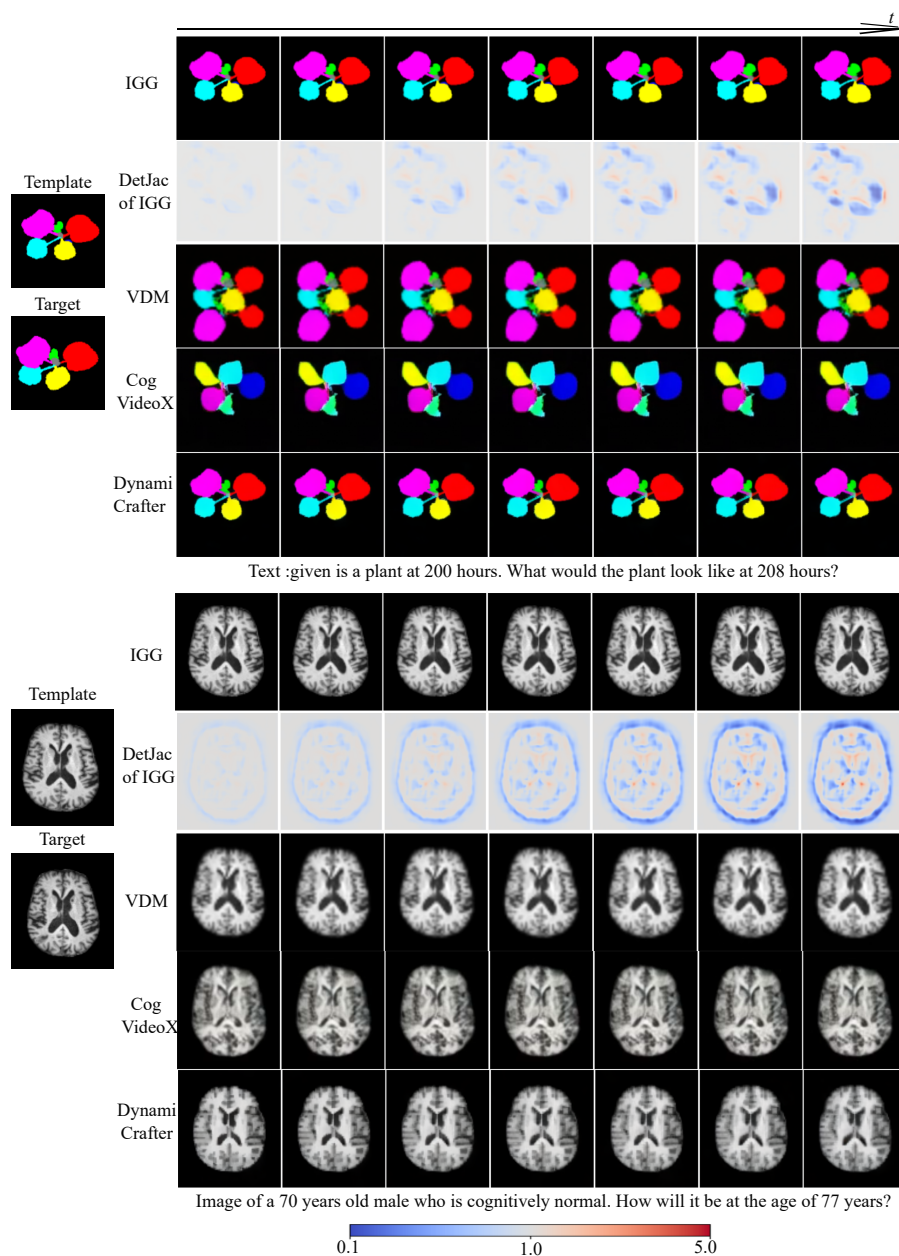


图3. IGG（带有相应的 DetJac）生成的图像与所有基线模型在不同时间点上的比较。给定输入模板图像和文本指令，所有模型都会生成目标图像样本。真实的“目标”以及输入模板图像提供在面板左侧以供参考。

表 1. 各种数据集上所有方法的性能指标对比。

Dataset	Model	FVD ↓	FID ↓	KID ↓	SSIM ↑	IS ↑
Plant	IGG	48.29	9.23	0.007 ± 0.02	0.98 ± 0.02	1.06 ± 1.2%
	VDM [16]	357.56	71.79	0.23 ± 0.03	0.23 ± 0.10	1.05 ± 1.0%
	CogVideoX [46]	471.49	153.93	0.63 ± 0.07	0.69 ± 0.19	1.05 ± 0.6%
	DynamiCrafter [44]	466.70	80.78	0.41 ± 0.04	0.89 ± 0.04	1.03 ± 0.3%
Brain	IGG	26.39	6.23	0.006 ± 0.008	0.97 ± 0.03	1.02 ± 0.08%
	VDM [16]	247.84	69.23	0.46 ± 0.03	0.89 ± 0.02	1.02 ± 0.07%
	CogVideoX [46]	302.20	114.50	0.46 ± 0.04	0.69 ± 0.27	1.06 ± 2.8%
	DynamiCrafter [44]	288.62	142.40	1.05 ± 0.03	0.88 ± 0.03	1.03 ± 0.20%

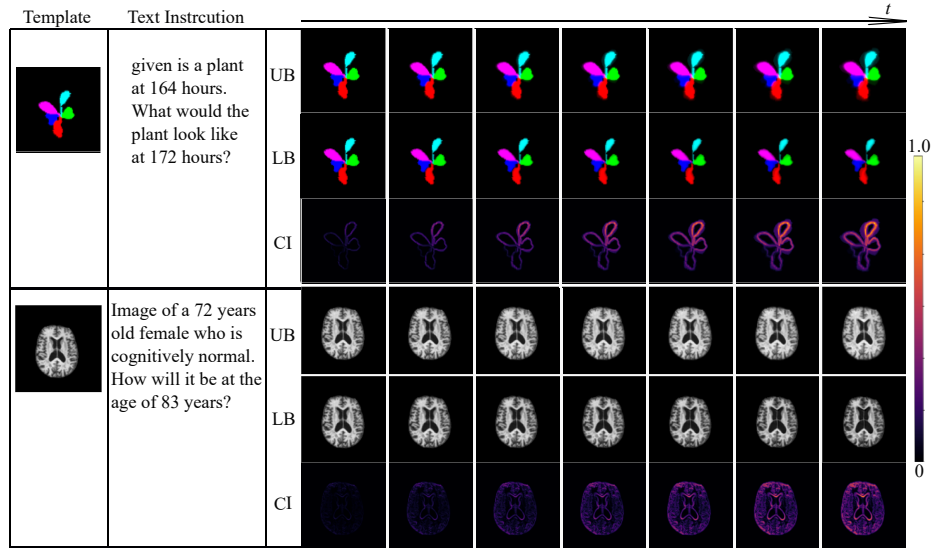


图 4. 从左到右：带有文本指令的输入模板图像，随后是展示下限 (LB)、上限 (UB) 和置信区间 (CI) 的信心图，这些代表了基于我们的 IGG 模型生成的约 1000 个样本在不同时间范围内的理想生长模式的 95% 区域。

型，该模型根据文本输入条件捕捉序列变形特征。实验结果表明，与最先进的方法相比，IGG 在样本保真度和多样性方面有显著提高。

我们的 IGG 模型通过整合几何原理，提供了拓扑评估工具，并建立了一个生成解剖学一致图像样本的框架。未来的研究将进一步 (i) 利用 IGG 的输出来惠及下游任务，如医学领域的图像分类、分割和对象识别；以及 (ii) 彻底验证并扩展其在各种应用中的实用性。

致谢。此项工作得到了美国国家科学基金会职业奖 2239977 的支持。

参考文献

1. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv:2112.00390 (2021)
2. Arnold, V.: Sur la géométrie différentielle des groupes de lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. In: Annales de l'institut Fourier. vol. 16, pp. 319–361 (1966)
3. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical image analysis **12**(1), 26–41 (2008)
4. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. arXiv:2304.08466 (2023)
5. Beg, M.F., Miller, M.I., Trouné, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International journal of computer vision **61**(2), 139–157 (2005)
6. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv:1801.01401 (2018)
7. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis (2019)
8. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
9. Chung, H., Ye, J.C.: Score-based diffusion models for accelerated mri. Medical image analysis **80**, 102479 (2022)
10. Fernandez, V., Pinaya, W.H.L., Borges, P., Tudosiu, P.D., Graham, M.S., Vercauteren, T., Cardoso, M.J.: Can segmentation models be trained with fully synthetically generated data? In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 79–90. Springer (2022)
11. Gupta, S., Samaras, D., Chen, C.: Topodiffusionnet: A topology-aware diffusion model. arXiv:2410.16646 (2024)
12. Hamamci, I.E., Er, S., Simsar, E., Sekuboyina, A., Gundogar, M., Stadlinger, B., Mehl, A., Menze, B.: Diffusion-based hierarchical multi-label object detection to analyze panoramic dental x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 389–399. Springer (2023)

13. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control (2022)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
15. Ho, J., Salimans, T.: Classifier-free diffusion guidance. *arXiv:2207.12598* (2022)
16. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. *Advances in Neural Information Processing Systems* **35**, 8633–8646 (2022)
17. Hu, J., Fei, B., Xu, B., Hou, F., Yang, W., Wang, S., Lei, N., Qian, C., He, Y.: Topology-aware latent diffusion for 3d shape generation. *arXiv:2401.17603* (2024)
18. Ijishakin, A., Abdulaal, A., Hadjivasiliou, A., Martin, S., Cole, J.: Interpretable alzheimer’s disease classification via a contrastive diffusion autoencoder. *arXiv:2306.03022* (2023)
19. Joshi, S., Davis, B., Jomier, M., Gerig, G.: Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage* **23**, S151–S160 (2004)
20. Kavar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6007–6017 (2023)
21. Kim, B., Han, I., Ye, J.C.: Diffusemorph: Unsupervised deformable image registration using diffusion model. In: *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*. pp. 347–364. Springer (2022)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014)
23. LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., et al.: Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medrxiv* pp. 2019–12 (2019)
24. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems* **36** (2024)
25. Mao, Y., Jiang, L., Chen, X., Li, C.: Disc-diff: Disentangled conditional diffusion model for multi-contrast mri super-resolution. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 387–397. Springer (2023)

26. Mazé, F., Ahmed, F.: Diffusion models beat gans on topology optimization. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 9108–9116 (2023)
27. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv:2108.01073 (2021)
28. Miller, M.I., Trouné, A., Younes, L.: On the metrics and euler–lagrange equations of computational anatomy. *Annual Review of Biomedical Engineering* **4**(1), 375–405 (2002)
29. Mukhopadhyay, S., Gwilliam, M., Agarwal, V., Padmanabhan, N., Swaminathan, A., Hegde, S., Zhou, T., Shrivastava, A.: Diffusion models beat gans on image classification (2023)
30. Nguyen, T., Li, Y., Ojha, U., Lee, Y.J.: Visual instruction inversion: Image editing via image prompting. *Advances in Neural Information Processing Systems* **36** (2024)
31. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Öztürk, Ş., Güngör, A., Çukur, T.: Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging* (2023)
32. Patashnik, O., Garibi, D., Azuri, I., Averbuch-Elor, H., Cohen-Or, D.: Localizing object-level shape variations with text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23051–23061 (2023)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
34. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022)
35. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
36. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding (2022)
37. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. *Advances in neural information processing systems* **29** (2016)

38. Uchiyama, H., Sakurai, S., Mishima, M., Arita, D., Okayasu, T., Shimada, A., Taniguchi, R.i.: An easy-to-setup 3d phenotyping platform for komatsuna dataset. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 2038–2045 (2017)
39. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv:1812.01717 (2018)
40. Vialard, F.X., Risser, L., Rueckert, D., Cotter, C.J.: Diffeomorphic 3d image registration via geodesic shooting using an efficient adjoint calculation. *International Journal of Computer Vision* **97**(2), 229–241 (2012)
41. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
42. Wu, N., Xing, J., Zhang, M.: Tlrn: Temporal latent residual networks for large deformation image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 728–738. Springer (2024)
43. Wu, N., Zhang, M.: Neurepdiff: Neural operators to predict geodesics in deformation spaces. In: International Conference on Information Processing in Medical Imaging. pp. 588–600. Springer (2023)
44. Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Liu, G., Wang, X., Shan, Y., Wong, T.T.: Dynamicrafter: Animating open-domain images with video diffusion priors. In: European Conference on Computer Vision. pp. 399–417. Springer (2025)
45. Yang, Y., Fu, H., Aviles-Rivero, A.I., Schönlieb, C.B., Zhu, L.: Diffmic: Dual-guidance diffusion network for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 95–105. Springer (2023)
46. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv:2408.06072 (2024)
47. Younes, L., Arrate, F., Miller, M.I.: Evolutions equations in computational anatomy. *NeuroImage* **45**(1), S40–S50 (2009)
48. Yu, Z., Dou, Z., Long, X., Lin, C., Li, Z., Liu, Y., Müller, N., Komura, T., Habermann, M., Theobalt, C., et al.: Surf-d: Generating high-quality surfaces of arbitrary topologies using diffusion models. In: European Conference on Computer Vision. pp. 419–438. Springer (2025)

49. Zhang, M., Liao, R., Dalca, A.V., Turk, E.A., Luo, J., Grant, P.E., Golland, P.: Frequency diffeomorphisms for efficient image registration. In: International Conference on Information Processing in Medical Imaging. pp. 559–570. Springer (2017)
50. Zhu, L., Xue, Z., Jin, Z., Liu, X., He, J., Liu, Z., Yu, L.: Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 592–601. Springer (2023)