流形假设下的架构再思考：JiT 扩散模型的改进与机理探究

萧德雄

2025403474 / 计算机系

xdx25@tsinghua.edu.cn

摘　要： Li 和 He 提出的 JiT (Just image Transformers) 架构基于流形假设，通过直接预测干净图像 (x-prediction)，验证了简单的线性层配合 ViT 即可有效处理高维像素数据。然而，JiT 的极简线性 Patch Embedding 可能不足以充分捕捉自然图像高度卷曲的非线性流形结构。本文首先在 Embedding 层引入 SiLU 激活函数，构建非线性瓶颈以增强对低维流形嵌入的拟合能力。

进一步地，本文深入探讨了骨干网络 (Backbone) 中流形约束 (降维) 与计算容量 (升维) 的本质矛盾。通过将 Transformer Block 内部替换为瓶颈结构的对比实验，本文揭示了一个关键的精度-多样性权衡 (Precision-Recall Trade-off)：显式的降维压缩虽然能有效过滤非流形噪声，从而显著提升生成图像的保真度 (Precision) 与 FID 指标；但这种严苛的流形约束同时也限制了模型对高熵随机偏差的建模能力，导致生成样本的多样性 (Recall) 下降。

此外，针对 JiT 缺乏语义约束的问题，本文引入了时间 (time) 与旋转 (rotation) 预测的自监督辅助损失。在 ImageNet 256×256 数据集上的实验表明，非线性 Embedding 与自监督信号有效提升了 FID 指标，而 Block 层的瓶颈化实验则从反面论证了“计算容量”在扩散模型骨干网络中的必要性。

关键词： 计算机视觉；扩散模型；JiT；非线性流形；瓶颈结构；高维数据拟合；自监督学习

1　引言

扩散生成模型 (Diffusion Models) 近年来在内容生成领域取得了显著成就。例如，Stable Diffusion [14] 通过将扩散过程迁移至潜在空间 (Latent Space)，在大幅降低计算成本的同时实现了高质量的文本到图像生成，成为 AIGC 领域的里程碑式工作。此外，随着应用场景向视频领域拓展，针对扩散模型推理速度慢的瓶颈，清华大学朱军团队提出的 TurboDiffusion [19] 通过稀疏注意力机制与时间步蒸馏等技术，在保持生成质量的前提下实现了 100 至 200 倍的端到端推理加速，为实时视频生成奠定了基础。

Li 和 He 在最近的工作 JiT [11] 中提出了“回归基础”的观点，指出通过简单的 Vision Transformer (ViT) [4] 配合 v-loss 与 x-prediction 策略，可以直接在高维像素空间进行高效高精度生成。JiT 的核心论点建立在流形假设 (Manifold Assumption) 之上 [1]：自然数据位于高维像素空间的低维流形上，噪声则散布于整个高维空间。因此，作者认为简单的线性映射足以提取有效的流形特征。

尽管 JiT 成功验证了线性层的有效性，但在深度学习的架构设计中，如何平衡“表征的几何映射”与“去噪的计算容量”仍是一个未被充分探讨的问题。首先，自然图像的流形通常高度卷曲且非线性 [17]，JiT 极简的线性 Patch Embedding 可能难以实现从高维观测空间到低维流形的完美映射。其次，JiT Block 沿用了被广泛证明有效的 SwiGLU 结构 [15]，这是一种典型的维度扩展 (Expansion) 设计，旨在利用门控机制增强模型的表达能力，通过升维解耦特征。然而，这似乎与 JiT 所强调的“专注于低维流形”的直觉相悖——如果目标是恢复低维数据，在网络内部强制进行维度压缩 (Compression) 的瓶颈结构是否在理论上更具优势？此外，作为自包含模型，JiT 摒弃了除类别外的所有语义引导，这可能导致模型在去噪初期缺乏对图像全局几何与语义结构的感知。

针对上述理论分歧与设计空间，本文在 JiT 基线模型上进行了以下探索与改进：

- 架构层面的流形与计算权衡探究：本文试图解构 Transformer 中“降维”与“升维”的作用机制。一方面，在 Patch Embedding 阶段引入非线性激活，增强对流形的几何映射能力；另一方面，尝试将 Transformer Block 中的 MLP 替换为“压缩-恢复”式的瓶颈结构，通过对比实验探究在骨干网络中，显式的流形约束 (降维) 与计算容量 (升维) 对对抗高熵噪声的异同影响。
- 损失函数层面的语义增强：为了弥补像素级去噪任务在语义理解上的不足，本文引入自监督学习 (SSL) 信号 [5]，通过联合训练旋转角度预测和时间步预测任务，辅助模型建立对图像全局结构及噪声水平的鲁棒感知。

2 相关工作

2.1 JiT 与流形假设

JiT [11] 强调了 x-prediction (直接预测原图) 相比 ϵ-prediction (预测噪声) 在高维空间的优势。其核心设计摒弃了 Tokenizer 和预训练 VAE，直接在像素空间应用 Vision Transformer。这一设计深受流形假设 (Manifold Assumption) [1] 的启发：自然数据位于高维空间的低维流形上，而噪声则散布于整个高维空间。

然而，经典的流形学习研究指出，自然图像的流形通常是高度卷曲且非线性的 [6, 17]。JiT 原文中采用纯线性的 Patch Embedding 虽然极简，但可能难以将这种非线性流形完美“展开”至 Transformer 的输入空间。本文通过引入非线性层来探究这一映射过程的优化空间。

2.2 表征学习中的维度：压缩与扩展

深度神经网络中关于维度的处理存在两种截然不同的范式，这也构成了本文架构探索的理论基础。

一方面是“维度扩展 (Expansion)”。标准的 Transformer 架构 [18] 及其变体 (如 SwiGLU [15]) 在前馈网络 (FFN) 中广泛采用“升维-激活-降维”的结构。这一设计遵循科弗定理 (Cover's Theorem) [2]，即通过将低维特征投影至高维空间来增加特征的线性可分性，从而提供足够的计算容量 (Capacity) 来处理复杂的去噪任务。

另一方面是“维度压缩 (Compression)”。以 MAE [8] 为代表的掩码建模方法，通过极高比例的掩码构建了严苛的信息瓶颈 (Information Bottleneck)，迫使模型学习紧凑的推理特征。本文试图在 JiT 的骨干网络中引入类似的显式瓶颈结构，以此对比“升维计算”与“降维流形约束”在扩散生成任务中的有效性。

2.3 自监督辅助任务

自监督学习 (Self-Supervised Learning) 通过设计代理任务 (Pretext Tasks) 挖掘数据的内在结构，常用于提升无标签数据的表征质量。预测图像旋转 [5] 是一种经典的无监督特征学习方法，它要求模型具备理解物体姿态和全局类别的能力。在扩散模型仅依赖像素重构损失 (x-loss) 的情况下，引入此类几何与语义相关的辅助任务，有望弥补纯生成式目标在语义引导上的不足。

3 方法

3.1 JiTBlock MLP 的瓶颈化替换

原版 JiT 的 Transformer Block 的 MLP 使用了 SwiGLU FFN [15]，其隐藏层维度通常直接扩展为输入维度的 $\frac{8}{3}$ 倍或 4 倍。为了分析 Transformer Block 的 MLP 是否满足流形假设，我们将 MLP 替换为带有显式降维瓶颈的三层非线性结构。

具体实现为：首先构建一个“硬瓶颈”，将特征维度从 H 压缩至 $H/2$，经过 GELU [9] 激活后，再扩展至标准的 MLP 隐藏层维度 (D_{mlp}) 以保证后续的非线性映射容量，最后恢复回 H。代码逻辑如下：

```python
if args is not None and getattr(args, 'use_nonlinear', False):
    self.mlp = nn.Sequential(
        nn.Linear(hidden_size, hidden_size // 2),
        nn.GELU(),
        nn.Linear(hidden_size // 2, mlp_hidden_dim),
        nn.SiLU(),
        nn.Linear(mlp_hidden_dim, hidden_size)
    )
else:
    self.mlp = SwiGLUFFN(hidden_size, mlp_hidden_dim, drop=proj_drop)
```

Listing 1: MLP 瓶颈层替换逻辑

这种设计旨在测试一个关键假设：在进入高维计算空间 (D_{mlp}) 之前，如果强制特征先通过一个低维关卡 ($H/2$)，是否能像流形学习预测的那样有效过滤非流形噪声；亦或是这种早期的信息压缩会破坏模型对抗高熵噪声所需的计算路径，导致表达能力不足。

3.2 非线性 Patch Embedding

JiT 原论文强调使用线性 Patch Embedding(由两个线性层组成)。我们认为，在进入 Transformer 之前引入非线性，可以构建一个微型的“特征提取 Stem”，有助于更快地锁定局部纹理。我们在原有的两个线性层之间插入了 SiLU 激活函数：

$$E(x) = W_2(\text{SiLU}(W_1(x))) \tag{1}$$

3.3 多任务自监督损失

为了在不依赖预训练分类器的情况下增强语义理解，本文分别探究了引入自监督辅助任务对扩散模型的影响。我们进行了两组独立的实验，分别将时间步预测和旋转预测作为辅助目标加入训练。总损失函数统一定义为：

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda \mathcal{L}_{\text{aux}} \tag{2}$$

其中，$\mathcal{L}_{\text{aux}}$ 根据实验设置分别为时间步预测损失 $\mathcal{L}_{\text{time}}$ 或旋转预测损失 $\mathcal{L}_{\text{rot}}$。

时间步预测 (Time Prediction)：该任务旨在增强模型对噪声水平的感知。我们将 Transformer 的输出特征 (全局平均池化后) 送入一个线性预测头，回归预测当前的归一化时间步 t。

Exp.	Patch Embed	MLP Type	Aux. Loss	**FID ↓**	**IS ↑**	**KID ↓**	**Prec. ↑**	**Rec. ↑**
Baseline	Linear	SwiGLU	-	119.4591	10.2573	0.1132	0.2314	0.1310
Exp 1	Non-linear	Bottleneck	-	**106.2162**	9.9534	**0.0873**	**0.3056**	0.1316
Exp 2	Non-linear	Bottleneck	Time	115.6307	9.3687	0.1029	0.2440	0.0980
Exp 3	Non-linear	SwiGLU	-	112.5951	10.2405	0.1008	0.2617	0.1406
Exp 4	Non-linear	SwiGLU	Time	109.9101	**10.5540**	0.1013	0.2814	**0.1422**
Exp 5	Non-linear	SwiGLU	Rotation	195.2785	3.5111	0.1831	0.0008	0.1004

表 1: 对非线性 Patch Embedding，JiTBlock 瓶颈，自监督进行消融实验 (Ablation Study)。数据基于 50,000 张验证集图像。

旋转预测 (Rotation Prediction)：该任务旨在增强模型的全局语义理解。我们在训练时对输入 Batch 中的图像进行随机旋转 $k \times 90^\circ$ ($k \in \{0, 1, 2, 3\}$)，并要求模型通过一个辅助分类头预测旋转角度 k。

4 实验结果与讨论

4.1 实验设置

我们基于 JiT 的官方实现 [12] 进行改进。在 ImageNet 256×256 数据集上进行训练，分辨率为 256，Patch Size 为 16。基线模型为 JiT-B/16。受限于计算资源，为了在有限的算力预算下快速验证架构改进的有效性，我们对原论文的训练配置进行了以下调整：

- 采样器与步数调整：原论文采用 50 步的 Heun 采样器 (二阶 ODE 求解器，NFE=50)。为了显著降低推理与验证的时间成本，我们将采样器替换为 25 步的 Euler 采样器 (一阶 ODE 求解器，NFE=25)。虽然 Euler 方法的离散化误差略高于 Heun [10]，但它将单次推理的计算量减少了 75%，更适合快速迭代。
- 梯度累计 (Gradient Accumulation)：受限于显存容量，我们无法直接进行 Batch Size 为 1024 的训练。我们采用梯度累计技术 [7]，将多个微批次 (Micro-batch) 的梯度进行累加，以在数学上等效模拟原论文中的大批次优化效果，保证训练动态的一致性。
- EMA 衰减率调整：我们将模型参数的指数移动平均 (EMA) [13] 衰减率从默认的 0.9999 调整为 0.999。由于我们的总训练步数较少，较高的衰减率会导致 EMA 模型更新过于滞后；降低该参数有助于 EMA 权重更快地追踪当前模型的优化状态，加速收敛。
- 数据集缩减：为了在有限时间内完成消融实验，我们构建了 ImageNet [3] 的随机子集进行训练，采样比例设为 0.01(即仅使用 1% 的训练数据)。
- 增加 Epoch：因为上述更改导致了模型收敛速度变慢，我们将 Epoch 从原先的 600 增加到了 1000 以尽量让所有测试模型接近收敛。

4.2 结果与输出

图 1 展示了几个模型实际生成的表现。图 2 展示了引入时间自监督的成功输出。

表 1 展示了不同模块对最终 FID，IS，KID，Precision，Recall 的影响。

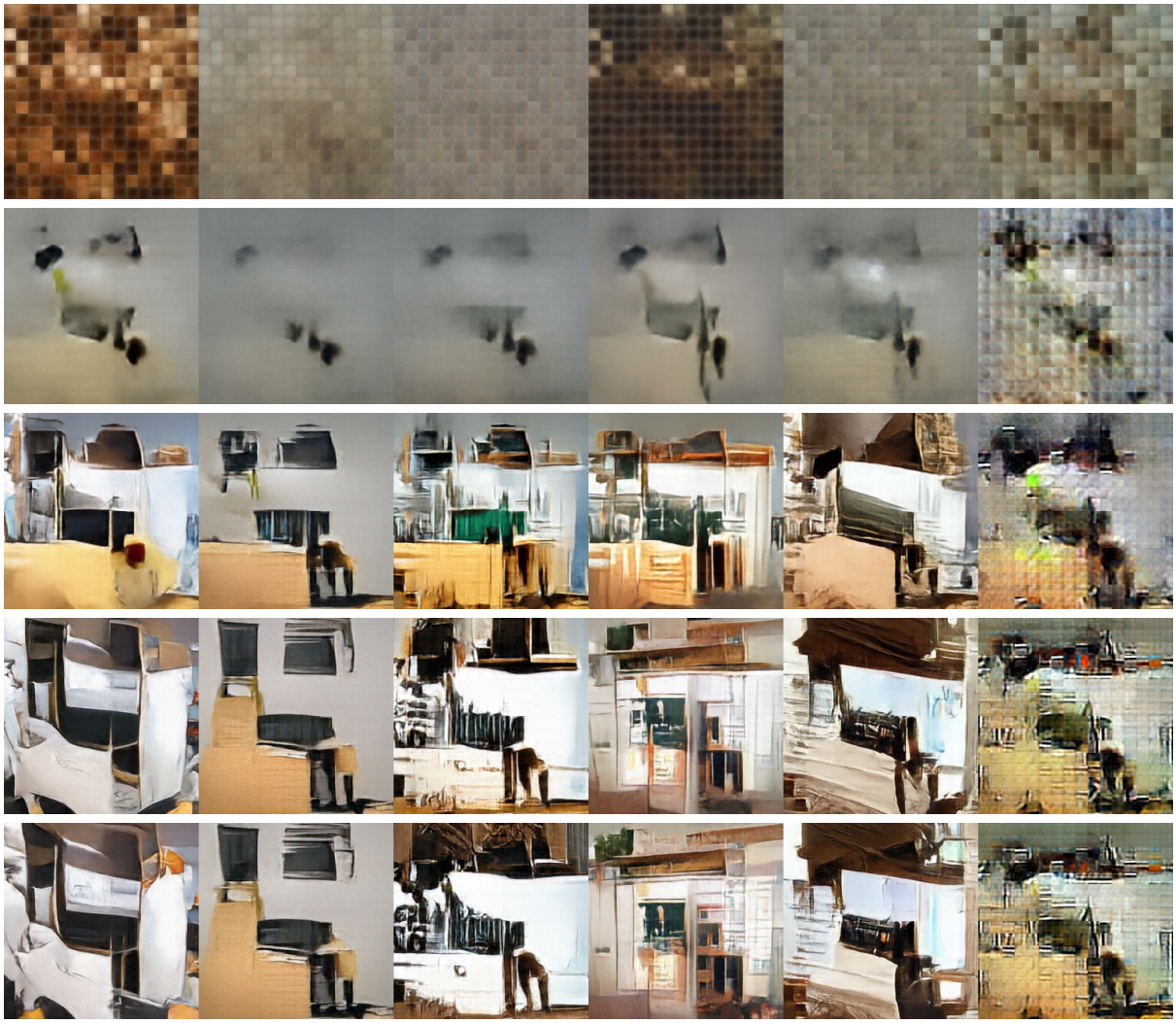

图 1: 训练过程中的生成样本演变。从上至下分别为 Epoch 100, 300, 600, 900, 999。从左至右依次为：Baseline, Exp 1, Exp 2, Exp 3, Exp 4, Exp 5 (失败)。可以看出 Exp 1 和 Exp 4 收敛最快，而 Exp 5 出现了明显的语义崩坏。

4.3 分析与讨论

4.3.1 非线性映射对流形展开的必要性

对比 Baseline (线性 Embedding) 与 Exp 3 (非线性 Embedding + SwiGLU)，我们发现引入简单的 SiLU 激活函数将 FID 从 119.46 显著降低至 112.60。这一结果强有力地支持了本文关于“流形非线性”的假设。

从几何视角来看，原始 JiT 的线性 Patch Embedding 仅能对输入空间进行仿射变换 (旋转、缩放、剪切)。然而，自然图像流形通常是高度卷曲的 (例如 Swiss Roll 结构)。一个纯线性的入口层试图将卷曲的高维像素数据直接投影到 Transformer 的隐空间，势必会造成流形结构的重叠或拓扑断裂。我们在 Embedding 层引入的非线性激活函数 (SiLU) 实际上赋予了模型在进入骨干网络之前对输入空间进行“局部弯曲”和“流形展开”的能力。这种早期的非线性特征提取充当了一个轻量级的 Stem，使得后续的 Transformer Block 能在更平坦、更解耦的特征空间上进行去噪，从而显著提升了整体生成质量。

图 2: 训练过程中的生成样本演变。从上至下分别为 Epoch 100, 300, 600, 900, 999。从左至右依次为：Baseline, Exp 1 , Exp 2 (失败), Exp 3, Exp 4 (最佳), Exp 5 (失败)。可以看出对于这幅图像 Exp 1 画面右边出现了崩坏，Exp 3 画面整体失真，但有时间自监督的 Exp 2 和 4 看起来要好一些。

4.3.2　自监督任务的冲突与分布偏移

Exp 5 的失败揭示了判别式自监督与生成式任务之间潜在的目标冲突。我们认为失败主要归因于两点。

第一，ImageNet 数据集中的图像大多具有固定的重力方向 (Upright)。在训练中对图像进行随机旋转并强制模型识别旋转角度，虽然迫使模型学习了语义特征，但也隐含地向模型输入了“旋转后的图像也是合法数据”这一错误先验。这导致扩散模型在去噪时可能会生成任意方向的图像，从而破坏了生成分布与验证集分布 (正向图像) 的一致性，导致出现数据分布偏移，FID 恶化。

第二，旋转预测要求提取的是高度抽象的、对纹理不敏感的语义特征；而像素级扩散生成 (尤其是基于 x-prediction) 需要极其精确的纹理和边缘信息。在有限的模型容量下，强力的旋转分类损失可能主导了梯度的更新方向，导致模型过度关注高级语义而忽略了底层的图像重建细节，引发语义崩坏。

4.3.3 时间的校准作用

与旋转预测不同，时间步预测 (Exp 2 和 Exp 4) 在大多数指标上带来了正面收益或保持了良好的权衡。Exp 4 相比于同构的 Exp 3，FID 进一步从 112.60 优化至 109.91。

我们认为，显式的时间预测充当了噪声水平校准器 (Noise Level Calibration)。在扩散模型中，输入 x_t 是信号与噪声的线性组合，其信噪比随 t 剧烈变化。虽然 Transformer 通过 Time Embedding 接收了时间信息，但在深层网络中，这种全局条件往往会被稀释。通过增加辅助损失 $\mathcal{L}_{\text{time}}$，我们强制骨干网络的全局特征必须包含足以恢复出 t 的信息。这确保了模型在推理的每一步都对当前的噪声强度保持“清醒”的感知——在 t 较大时专注于恢复低频结构，在 t 较小时专注于修补高频纹理。这种显式的对齐机制减少了去噪过程中的错位 (例如在需要构图的阶段去纠结噪点)，从而提升了生成的连贯性。

4.3.4 模型对“信号结构”与“噪声熵”的偏好权重

如表 1 所示，我们的实验输出揭示了一个性能倒挂现象：采用“压缩-恢复”瓶颈结构的 Exp 1 取得了最低的 FID (106.2) 和 KID，以及最高的 Precision。这意味着该模型生成的图像质量极高，且紧贴真实图像流形的核心区域，产生的“废片”很少；相反，采用标准升维结构的 Exp 3 虽然 FID 略高，但在 IS 和 Recall 指标上显著优于 Bottleneck。这意味着 SwiGLU 生成的样本覆盖了更广泛的数据分布，保留了更多的多样性。

结合扩散过程的数学形式 $z_t = t \cdot x + (1-t) \cdot \epsilon$，我们对上述现象提出如下理论解释：

Bottleneck 结构强制特征通过一个低维关卡 ($H/2$)。由于信号 x 是低维且结构化的，而噪声 ϵ 是高维且高熵的，这种物理上的“挤压”迫使模型必须学会区分信号与噪声——只有符合低维流形结构的特征 x 才能顺利通过瓶颈，而代表随机性的高频噪声 ϵ 则被滤除。这种机制充当了强效的去噪正则化。因此，Exp 1 能够生成非常清晰、结构合理的图像 (高 Precision)，但代价是那些位于流形边缘、难以被低维特征概括的稀有样本也被一同过滤掉了，导致生成多样性不足 (低 Recall)。

而 SwiGLU 的升维设计提供了巨大的参数空间 (计算容量)。在对抗高熵噪声 ϵ 时，高容量网络不仅仅是在“去除”噪声，它实际上有能力“建模”噪声的复杂分布。噪声 ϵ 充斥在整个高维空间，代表了生成过程中的随机性与不确定性。SwiGLU 能够容纳这种高熵信息，将其转化为生成结果中的多样化变体。因此，Exp 3 能够覆盖更广阔的图像分布 (高 Recall)，但由于保留了过多的“随机性”和“边缘特征”，其生成的平均质量 (FID/Precision) 反而不如经过严格过滤的 Bottleneck 结构纯粹。

所以，尽管我们可以看到引入非线性层的 Patch Embedding 层从各个指标来看 (除了 IS 有微差) 都是基本优于线性层的 Patching Embedding 的，但 Block 层中 Bottleneck 与 SwiGLU 更多不是孰优孰劣的相比，而是一种保真度和多样性的权衡。在有限数据下，Bottleneck 通过牺牲多样性换取了极高的生成保真度。

5 结论与未来工作

本文通过对 JiT 架构的解构与重组，深入探讨了流形假设在扩散模型中的具体表现形式。实验结果表明，在 Patch Embedding 阶段引入非线性激活能显著改善模型对图像流形的拟合能力。

此外，瓶颈结构 (Bottleneck) 通过显式的降维压缩，强制模型忽略无法被低维流形解释的高熵噪声与随机偏差。这种“提纯”过程显著提升了生成的保真度 (Precision) 和 FID，但也抑制了样本的多样性；而扩展结构 (如 SwiGLU) 提供了处理高维噪声所需的计算容量。这种容量使得模型能够保留数据分布中的长尾部分和随机变化，从而显著提升了多样性 (Recall)，但牺牲了部分生成的纯净度。这表明，架构设计本质上是在调节模型对“信号结构”与“噪声熵”的偏好权重。

未来的工作将集中在探究如何打破这种“精度-多样性”的权衡：是否能够在仅拟合低维数据流形的同时，尽可能地保留数据的多样性；又或者，如果数据本身的多样性是高维特征，那么这是否代表了数据流形的最小维度就是高维。例如，我们计划探索一种集成架构，同时训练一个高容量模型与一个有瓶颈层的低容量模型。通过融合的方法让最终输出在保持低容量模型高 Recall 的同时实现高容量模型的高 Precision。

6 可复现性声明

本报告的所有实验代码均基于 PyTorch 框架。我们复用了 JiT 官方代码库 [12] 的核心组件，并在此基础上修改了 ‘JiTBlock‘ 类和 ‘PatchEmbed‘ 类。相关代码均已公开在 Github [16]：https://github.com/a-little-bear/JiT。

参考文献

[1] 2006, *Semi-Supervised Learning*, Cambridge, MA: MIT Press.

[2] T. M. Cover, 1965, “Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition”, *IEEE transactions on electronic computers*, **(3)**: 326–334.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, 2009, “Imagenet: a large-scale hierarchical image database”, *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, IEEE.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold *et al.*, 2020, “An image is worth 16x16 words: transformers for image recognition at scale”, *arXiv preprint arXiv:2010.11929*:

[5] S. Gidaris, P. Singh and N. Komodakis, 2018, “Unsupervised representation learning by predicting image rotations”, *International Conference on Learning Representations*.

[6] I. Goodfellow, Y. Bengio and A. Courville, 2016, *Deep learning*, MIT press, Chapter 5.11.3: Manifold Learning.

[7] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia and K. He, 2017, “Accurate, large minibatch sgd: training imagenet in 1 hour”, *arXiv preprint arXiv:1706.02677*:

[8] K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick, 2022, “Masked autoencoders are scalable vision learners”, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009.

[9] D. Hendrycks and K. Gimpel, 2016, “Gaussian error linear units (gelus)”, *arXiv preprint arXiv:1606.08415*:

[10] T. Karras, M. Aittala, T. Aila and S. Laine, 2022, “Elucidating the design space of diffusion-based generative models”, *Advances in Neural Information Processing Systems*, **35**, pp. 26565–26577.

[11] T. Li and K. He, 2025, “Back to basics: let denoising generative models denoise”, *arXiv preprint arXiv:2511.13720*: arXiv:2511.13720v1.

[12] T. Li and K. He，“Official PyTorch Implementation of JiT”，*GitHub repository*，2025 年，GitHub.

[13] B. T. Polyak and A. B. Juditsky, 1992, “Acceleration of stochastic approximation by averaging”, *SIAM journal on control and optimization*, **30(4)**: 838–855.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, 2022, “High-resolution image synthesis with latent diffusion models”, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695.

[15] N. Shazeer, 2020, “Glu variants improve transformer”, *arXiv preprint arXiv:2002.05202*:

[16] J. Siu，“Improved JiT with Non-linear Bottlenecks and Self-Supervised Learning”，*GitHub repository*，2025 年，GitHub.

[17] J. B. Tenenbaum, V. De Silva and J. C. Langford, 2000, “A global geometric framework for nonlinear dimensionality reduction”, *Science*, **290(5500)**: 2319–2323.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . Kaiser and I. Polosukhin, 2017, “Attention is all you need”, *Advances in neural information processing systems*, **30**.

[19] J. Zhang, H. Chen, Z. Zhang, Y. Wei, H. Feng, Z. Dong, J. Zhang, Q. Xiao and J. Zhu, 2025, “Turbodiffusion: accelerating video diffusion models by 100-200 times”, *arXiv preprint arXiv:2512.16093*: