基于Diffusion-Mamba和尺度不变损失的渐进式图像生成方法

李 豪,郝文宁*,邹世辰,谢晓宇
(中国人民解放军陆军工程大学指挥控制工程学院,江苏南京 210007)

摘 要: 扩散模型在图像生成领域由于精度高而受到了广泛关注,其骨干网络经历了从U-Net到Transformer的演变. 然而,由于Transformer的运算量与序列长度的平方成正比这一特性,导致扩散模型在处理高分辨率图像时存在计算复杂度高的问题. 为了解决这一问题,本文提出一种基于Diffusion-Mamba和尺度不变损失的渐进式图像生成方法. 该方法利用多方向扫描机制和轻量级局部结构增强模块融合了Mamba的高效特性以及扩散模型的建模能力,并通过渐进式级联扩散过程实现了从低分辨率图像向高分辨率图像的高效转换. 此外,设计基于对比学习的尺度不变损失函数,通过最大化同一目标在不同分辨率下的互信息,实现了跨尺度特征表示的对齐与增强. 在ImageNet(FID = 1.67)数据集上的实验结果表明:本文方法取得了综合精度的提高,充分验证了该方法的有效性和高效性.

关键词: 图像生成;扩散模型;状态空间模型;对比学习;尺度不变损失
基金项目: 国防工业技术发展项目(No.JCKY2020601B018)
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2025)09-3384-13
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250308

Progressive Image Synthesis Method Based on Diffusion-Mamba and Scale-Invariant Loss

LI Hao, HAO Wen-ning*, ZOU Shi-chen, XIE Xiao-yu
(*College of Command and Control Engineering*, *Army Engineering University of PLA*, *Nanjing*, *Jiangsu* 210007, *China*)

Abstract: Diffusion models have garnered significant attention in the field of image generation due to their high precision. The backbone networks of these models have evolved from U-Net to Transformer architectures. However, the computational complexity of Transformer-based models scales quadratically with sequence length, posing a substantial challenge for generating high-resolution images. To address this issue, we propose a novel progressive image synthesis method based on Diffusion-Mamba and scale-invariant loss. Our method leverages the efficient characteristics of Mamba and the powerful modeling capabilities of diffusion models by integrating multi-directional scanning mechanisms and lightweight local structure enhancement modules. It achieves an efficient transformation from low-resolution images to high-resolution images through a progressive cascaded diffusion process, significantly reducing computational complexity. Furthermore, we design a contrastive learning-based scale-invariant loss function that maximizes the mutual information of the same target across different resolutions, thereby aligning and enhancing cross-scale feature representations. Experimental results on the ImageNet (FID = 1.67) dataset demonstrate that our proposed method achieves comprehensive improvements in accuracy, effectively validating its efficacy and efficiency.

Key words: image synthesis; diffusion model; state space model; contrastive learning; scale-invariant loss
Foundation Item(s): Defense Industrial Technology Development Program (No.JCKY2020601B018)

收稿日期:2025-04-21;录用日期:2025-09-16;责任编辑:覃怀银
*通讯作者:郝文宁

1　引言

近年来,扩散模型(Diffusion Model,DM)在图像生成[1~3]任务中取得了显著进展,挑战了生成对抗网络[4](Generative Adversarial Network,GAN)长期以来的主导地位.DM受非平衡热力学[5]的启发,通过对数据施加渐进式噪声扰动,再逆向训练模型从随机噪声中重构出与原始数据分布相契合的高质量合成数据.得益于Transformer[6]的高效性和泛化性,DM的骨干网络经历了从以U-Net[7]为代表的卷积神经网络(Convolutional Neural Network,CNN)到Transformer[8]的演变.基于Transformer的DM首先将图像在隐空间编码成特征图(Latent Feature Map,LFM),再将LFM划分为多个小块(patch),每个patch随即被编码为独立的特征向量(token).后续过程中,Transformer逐步去除嵌入在图像token中的噪声成分,从而实现了由随机噪声至清晰图像的蜕变.然而,Transformer内部自注意力机制的局限性导致模型计算复杂度与图像token数量呈二次方增长的关系,这对于高分辨率图像的生成构成了严峻的计算负担,成为当前亟待解决的关键技术难题.

近期,状态空间模型[9](State Space Model,SSM)因实现了线性的计算复杂度而受到广泛关注,其受卡尔曼滤波器[10]的启发,有机结合了循环神经网络(Recurrent Neural Network,RNN)与CNN,显著降低了计算成本.其中,Mamba[11]在自然语言处理、计算机视觉和语音处理等领域实现了与Transformer相当的建模能力,成为当前的研究热点.Mamba通过时变参数保留关键数据,并引入一种简洁而高效的选择机制,使得模型有效地筛选出无关信息.在此基础上,Mamba提出一种硬件感知算法,采用并行扫描的方式替代传统的卷积操作,提升了计算速度.因此,本文将Mamba作为新的骨干网络引入扩散模型,探索其在高分辨率图像生成方面的应用.然而,将Mamba与DM结合用于高分辨率图像生成面临着一系列难题,其核心难点在于Mamba的因果序列建模特性与图像固有的二维结构之间存在显著差异,Mamba最初的设计理念是针对一维信号的因果序列建模,而这一设计理念难以直接迁移至二维图像的建模任务中.Vision Mamba[12]将二维图像数据通过光栅扫描顺序(Raster-Scan Order,RSO)转换为一维序列,但这一做法却将每个token的感受野限定在RSO既定的token中,从而限制了模型的整体感受野.更为关键的是,尽管一行末尾与下一行首部在RSO中相邻,但在空间上却缺乏连续性,这进一步加剧了模型对图像局部特征的捕捉难度,且相邻像素的拓扑关系在序列化过程中被破坏.此外,尽管Mamba在推理效率上展现出显著优势,但在高分辨率图像上训练基于Mamba的扩散模型仍然需要消耗大量计算资源.为应对这一难题,LDM(Latent Diffusion Model)[7]在潜空间中训练扩散模型,并将所得结果反向映射至像素空间.此方法虽然提升了训练效率,但也不可避免地引入了低级伪影,降低了模型精度.

为了克服上述问题,本文提出一种基于Diffusion-Mamba和尺度不变损失的渐进式图像生成方法,称之为PIS-DM.该方法直接在像素空间将图像分割为patch并编码成token特征,然后将Mamba作为扩散过程的骨干网络对token特征进行去噪.为避免token之间形成单向因果关系,本文利用图像的空间结构设计了循环扫描模块,通过交替执行8个不同的扫描方向确保每个token拥有全局的感受野.同时,PIS-DM在网络的输入与输出层分别引入3×3的深度卷积层,以提升生成图像的局部连贯性.此外,本文在网络的浅层特征与深层特征之间通过长跳跃连接搭建起信息传递的通道,有效地将低层次信息传递至高层次特征,提升扩散模型在像素级别预测目标的性能.另一方面,为进一步提升模型对尺度变化的鲁棒性,本文基于对比学习设计了一种尺度不变损失函数,通过最大化不同分辨率下骨干网络特征的互信息,实现了多分辨率特征表示的一致性对齐.在高分辨率图像生成阶段,PIS-DM采用由粗到细的级联扩散策略,以上采样的低分辨率图像作为起点进行细化,提升了高分辨率图像的生成效率.

本文的主要贡献和创新点如下:

(1)设计一个由粗到细的图像生成级联网络架构,通过解耦级联网络中扩散过程与骨干网络,提升了高分辨率图像生成的训练和推理速度.

(2)联合扩散模型和Mamba提出一种渐进式图像生成方法,利用图像的空间结构设计了循环扫描模块,提高了Mamba对二维图像的建模能力.

(3)引入基于对比学习的尺度不变损失函数,通过最大化同一目标不同分辨率的互信息,增强了多分辨率下特征表示的一致性.在ImageNet数据集上对提出的方法进行了实验验证,实验结果表明:本文提出的方法取得了综合精度的提升.

2　相关工作

2.1　扩散模型

DM的概念最初源于统计物理学,用于描述粒子从高浓度区域向低浓度区域移动的过程[5].2015年,Sohl-Dickstein等人[5]受非平衡热力学的启发,首次提出了一种通过迭代正向扩散过程破坏数据分布结构,并构建逆向去噪过程学习机制的生成模型.这一方法通过构造可逆马尔可夫链实现了数据分布重建,展现出高度的灵活性和易处理性,迅速引起了广泛关注.2020年,Ho等人[13]提出去噪扩散概率模型(Denoising Diffusion

Probabilistic Models, DDPM), 使用U-Net骨干网络近似得分函数, 取得了较好的效果, 但推理速度慢且训练成本高. 为解决这一问题, 2022年, Rombach等人[7]提出潜在扩散模型, 该模型在预训练的变分自编码器(Variational Auto-Encoder, VAE)潜在空间中进行操作, 降低了计算复杂度, 推动了高分辨率图像生成的发展. 近年来, 随着Transformer在计算机视觉任务中的广泛应用, 研究者开始探索其在扩散模型中的应用. 2022年, Yang等人[14]提出GenViT, 首次证明了Vison Transformer(ViT)能够用于图像生成, 但其性能劣于U-Net. 2023年, Peebles等人[15]受到ViT的启发, 提出了一种名为Diffusion Transformers(DiT)的扩散模型, 并与LDM结合在ImageNet数据集上取得了较好的生成性能. 同年, Bao等人[8]提出的U-ViT通过加入长跳跃连接和卷积层将U-Net和ViT融合为统一的骨干网络, 为低层次特征提供快捷传递途径, 从而确保了扩散模型训练过程的稳定性. 2024年, Hatamizadeh等人[16]提出了DiffiT, 通过引入时间相关多头自注意力机制, 实现了对去噪过程的细粒度控制, 进一步提高了高分辨率图像生成的速度. 同年, RDM[17]通过级联扩散策略, 将低分辨率与高分辨率生成过程解耦, 实现了跨分辨率的统一扩散框架. 尽管如此, Transformer在处理大量token时的效率瓶颈仍然限制了其在高分辨率图像生成领域的广泛应用.

2.2 状态空间模型

SSM的输入和输出均为一维序列, 输入序列的token与前一隐藏状态以及模型权重进行线性变换, 递归映射至隐藏状态, 并产生输出. 近期, Mamba[11]因其强大的建模能力而备受关注. Mamba的每个模块由选择扫描单元、一维因果卷积和归一化层三部分组成, 其通过输入的模型权重实现数据选择功能, 构建了一个通用的语言模型架构. 2024年, Yan等人[18]提出DiffuSSM, 首次将扩散模型中的注意力机制替换为SSM, 该方法无需借助全局压缩技巧, 在扩散过程中保持了高分辨率图像的细节完整性. Fei等人[19]提出的DiS将所有输入的patch统一串联为离散的token, 使扩散模型继承了SSM的泛化性, 提高了高分辨率图像的生成能力. Hu等人[20]结合图像的归纳偏置, 利用空间连续性设计了一个即插即用的ZigMa模块, 通过随机插值网络在高分辨率图像生成上取得了显著成效. Park等人[21]引入了一个混合模型MambaFormer, 该方法将Mamba与多头注意力模块相结合, 提升了SSM的上下文学习(In-Context Learning, ICL)能力. Teng等人[22]提出的DiM通过将SSM引入扩散模型, 替代传统Transformer中的注意力机制, 降低了高分辨率图像生成的计算复杂度. 与上述方法不同, 本文提出一种基于Mamba的扩散模型, 该模型能够处理超过10 000个token的图像序列, 从而充分挖掘了Mamba在长序列处理上的潜力. 同时, 该模型融入了多个新颖模块来处理空间先验信息, 验证了基于Mamba的扩散模型在不同分辨率下的微调能力.

3 基础知识

3.1 DM

DM通过渐进式噪声扰动和逆向去噪过程, 实现了从随机噪声到高质量合成数据的精确重建, 其训练过程包含前向(扩散)过程与逆向(去噪)过程两个阶段. 在前向过程中, 从数据分布$q(\boldsymbol{X})$采样原始图像$\boldsymbol{X}_0$, 模型向原始数据逐步添加噪声, 直至数据转变为纯粹的随机噪声. 前向过程可以形式化为$\boldsymbol{X}_{t-1}$到$\boldsymbol{X}_t$的马尔可夫链:

$$q\left(\boldsymbol{X}_t\middle|\boldsymbol{X}_{t-1}\right)=\mathcal{N}\left(\boldsymbol{X}_t;\ \sqrt{1-\beta_t}\,\boldsymbol{X}_{t-1},\beta_t\boldsymbol{I}\right) \tag{1}$$

其中, $\mathcal{N}\left(\boldsymbol{X};\mu,\sigma\right)$代表均值为$\mu$、方差为$\sigma$的高斯分布; β_t表示噪声调度. 根据高斯核的性质, 对于任意步t, 可以从原始图像$\boldsymbol{X}_0$直接得到$\boldsymbol{X}_t$:

$$\begin{aligned}q\left(\boldsymbol{X}_{1:T}\middle|\boldsymbol{X}_0\right)&=\prod_{t=1}^{T}q\left(\boldsymbol{X}_t\middle|\boldsymbol{X}_{t-1}\right)\\&=\mathcal{N}\left(\boldsymbol{X}_t;\ \sqrt{\overline{\alpha_t}}\,\boldsymbol{X}_{t-1},\sqrt{1-\overline{\alpha_t}}\,\boldsymbol{I}\right)\end{aligned} \tag{2}$$

其中, $\boldsymbol{X}_{1:T}$表示从时间步$t=1$至$t=T$间形成的受噪声污染图像序列, $\alpha_t=1-\beta_t$, $\overline{\alpha_t}=\prod_{t=1}^{T}\alpha_t$. 因此, $\boldsymbol{X}_t=\sqrt{\overline{\alpha_t}}\,\boldsymbol{X}_0+\sqrt{1-\overline{\alpha_t}}\,\varepsilon$, $\varepsilon\sim\mathcal{N}(0,\boldsymbol{I})$. 当$t=T$时, $\alpha_T\approx0$:

$$\begin{aligned}q\left(\boldsymbol{X}_T\right)&=\int q\left(\boldsymbol{X}_T\middle|\boldsymbol{X}_0\right)q\left(\boldsymbol{X}_0\right)\mathrm{d}\boldsymbol{X}_0\\&\approx\mathcal{N}\left(\boldsymbol{X}_T;\ 0,\boldsymbol{I}\right)\end{aligned} \tag{3}$$

3.2 Mamba

Mamba通过隐藏状态$h(t)\in\mathbb{R}^N$将输入信号$x(t)\in\mathbb{R}^N$映射为输出信号$y(t)\in\mathbb{R}^N$, 其过程可以形式化为常微分方程:

$$h'(t)=\boldsymbol{A}h(t)+\boldsymbol{B}x(t) \tag{4}$$

$$y(t)=\boldsymbol{C}h(t) \tag{5}$$

其中, $h'(i)$是$h(i)$的导数; $\boldsymbol{A}\in\mathbb{R}^{N\times N}$、$\boldsymbol{B}\in\mathbb{R}^{N\times1}$、$\boldsymbol{C}\in\mathbb{R}^{1\times N}$代表状态转移矩阵、输入矩阵、输出矩阵. 为了与真实场景的数据相适应, 状态空间模型利用零阶保持(Zero-Order Hold, ZOH)假设进行离散化, 因此上述常微分方程可以进行迭代求解:

$$h_t=\overline{\boldsymbol{A}}\,h_{t-1}+\overline{\boldsymbol{B}}\,x_t \tag{6}$$

$$y_t=\boldsymbol{C}h_t \tag{7}$$

其中, $\overline{\boldsymbol{B}}=(\Delta\boldsymbol{A})^{-1}\left(\exp(\Delta\boldsymbol{A})-\boldsymbol{I}\right)\cdot\Delta\boldsymbol{B}$和$\overline{\boldsymbol{A}}=\exp(\Delta\boldsymbol{A})$是离散化参数, $\Delta=\left[t_{i-1},t_i\right]$为离散化时间步长. 为了提高计算效率和可扩展性, 式(6)和式(7)中的迭代过程可以表示为卷积形式:

$$y = x * \overline{K} \tag{8}$$

其中，$\overline{K} = \left(C\overline{B}, C\overline{AB}, \cdots, C\overline{A}^{k}\overline{B}, \cdots\right)$作为SSM的卷积核，$N$是输入序列$x(i)$的长度，*表示卷积操作.

4　基于Diffusion-Mamba和尺度不变损失的渐进式图像生成方法

如图1所示，PIS-DM分为低分辨率图像生成和高分辨率图像重构两个阶段，每个阶段将带有时间步t、类别c和零填充的噪声图像patch作为输入，输出预测图像. 具体工作流程如下：首先，输入噪声被分割成patch并通过3×3深度可分离卷积层注入局部信息；其次，patch、时间步t、类别c和零填充通过嵌入层转换为高维特征向量token，并利用循环扫描模块将token展平成序列；再次，输入序列利用Mamba进行去噪，同时在浅层和深层特征之间通过长跳跃连接促进低级信息向高级特征的传递；最后，将得到的噪声序列恢复空间结构，与噪声输入相结合得到预测的图像.

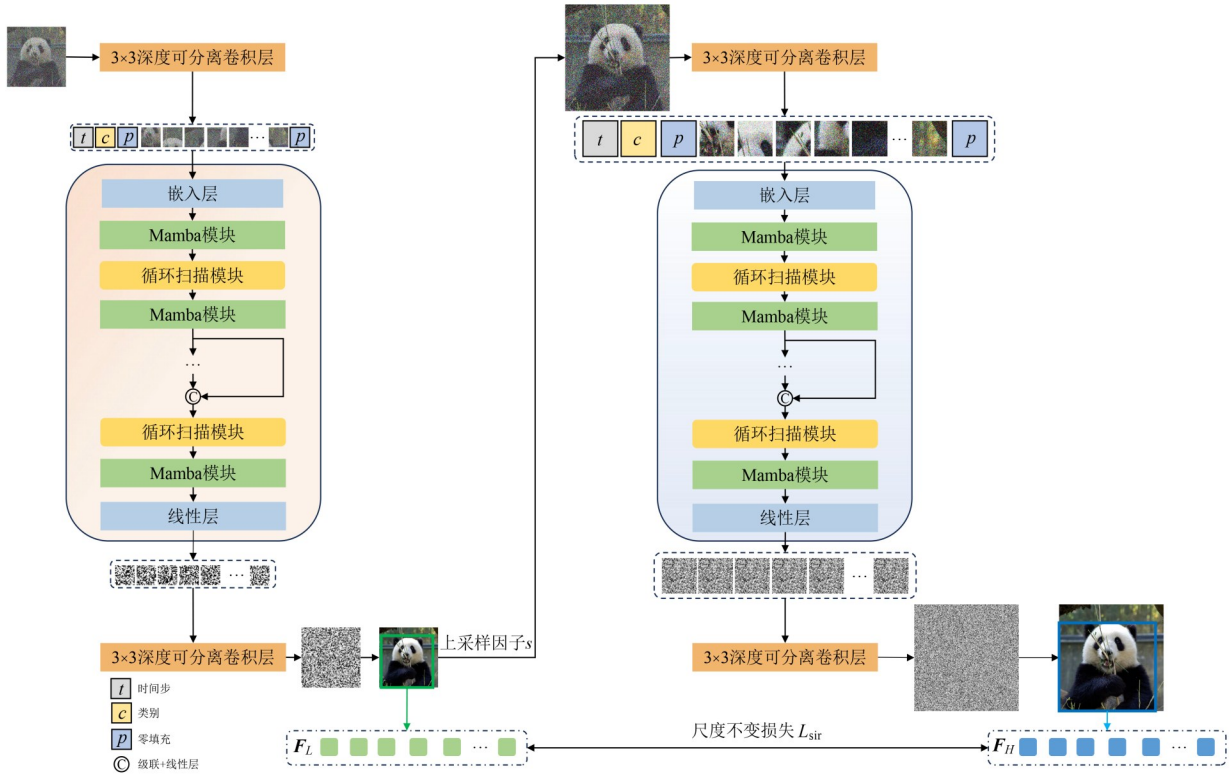

图1　PIS-DM总体框架和工作流程

4.1　循环扫描模块

并行关联扫描操作是Mamba的核心组成部分，旨在解决由选择机制引起的计算问题，加速训练过程并降低内存需求. 通过利用SSM的线性特性，Mamba在硬件层面设计了内核融合和重新计算策略来实现这一目标. 然而，Mamba的单向序列建模范式限制了其对图像和视频等多维数据的全面学习能力. 若仅沿单一方向扫描patch，则会导致patch的感受野呈现单向性和局限性. 例如，左上角首先被扫描的patch永远无法聚合来自其他patch的信息. 为了解决这一问题，研究者们提出了多种高效的扫描方法[22]，以提升模型性能并促进Mamba的训练过程. 如图2所示，这些扫描机制可归纳为四类：双向扫描、交叉扫描、连续扫描和跳跃扫描. 双向扫描通过同时使用前向和后向SSM处理输入，增强了模型的空间感知能力，但增加了计算复杂度. 交叉扫描通过多路径遍历提高了局部和全局信息的整合效率，但需要更多计算资源来处理复杂的路径组合. 连续扫描通过扫描相邻token来保持输入序列的连续性，有助于提高视觉任务的表现，但会限制对远距离依赖关系的建模. 跳跃扫描通过并行处理和跳过部分patch来减少计算时间和资源消耗，但可能影响全局特征的完整性.

全局感受野对于PIS-DM在图像生成任务[23~25]中有效捕捉图像中的空间结构至关重要. 为了使每个patch都能拥有全局感受野，PIS-DM在不同的Mamba模块中采用了多样化的扫描模式. 如图3所示，在前8个模块中，PIS-DM采用由内至外的扫描方式，即从图像中心向边缘逐块扫描. 这种由内至外的扫描方式能够更好地捕获中心区域的详细信息，并逐步扩展到外围，从而确保每个块都能获得全局上下文信息，由于物体通

(*a*) 双向扫描和交叉扫描

(*b*) 连续扫描

(*c*) 跳跃扫描

图2 扫描机制

常位于图像中心区域,这种扫描方式能够有效提取物体的结构特征并区分前景与背景. 在后续模块中,PIS-DM反转序列顺序,采用由外至内的扫描方式,进一步增强模型对图像边缘信息的敏感性,并促进全局特征的均匀分布. 通过循环遍历所有扫描模式,PIS-DM不仅充分利用了不同扫描模式的优势,还避免了单一扫描模式带来的局限性.

循环扫描模块不仅结合了多种扫描模式,还在不同模块间动态切换扫描方向,从而在保证局部细节的同时,提升模型对全局空间结构的感知能力. 此外,通过在不同模块中交替使用正向和反向扫描,PIS-DM不仅能够保持计算效率,还能增强模型对多层次特征的建模能力. 特别地,PIS-DM通过在图像生成任务中引入这种循环扫描机制,能够更有效地捕捉和生成图像中的复杂空间结构,从而提高生成图像的质量和多样性.

4.2 轻量级局部结构增强模块

为了保持Mamba架构高效性并增强局部结构的感知能力,PIS-DM在网络的起始和末端添加了轻量级局部结构增强模块. 具体来说,PIS-DM在Mamba网络的输入端和输出端分别添加了两层3×3深度可分离卷积层. 这种设计在不显著增加计算负担的情况下,有效增强了模型对局部结构的感知能力. 深度可分离卷积层能够捕捉图像中的局部特征,确保在进入和离开Mamba模块时图像的局部连续性得到有效加强,从而弥补了Mamba单向序列建模范式的局限性. 这种设计使得PIS-DM能够在全局感受野的基础上更好地捕捉和利用局部细节,进一步提高了生成图像的质量. 此外,为了进一步增强模型对序列边界信息的处理能力,PIS-DM在每个序列的开头和结尾增加了一个可学习的零填充token. 这些token通过学习能够更好地处理序列边界效应,从而在生成图像时提供更自然的过渡效果. 这种设计不仅提高了模型对边界信息的敏感性,还增强了模型在处理边缘区域时的鲁棒性. 从理论角度来看,PIS-DM的轻量级局部结构增强模块不仅解决了传统方法中局部结构感知不足和边界处理不自然的问题,还提升了模型的生成质量.

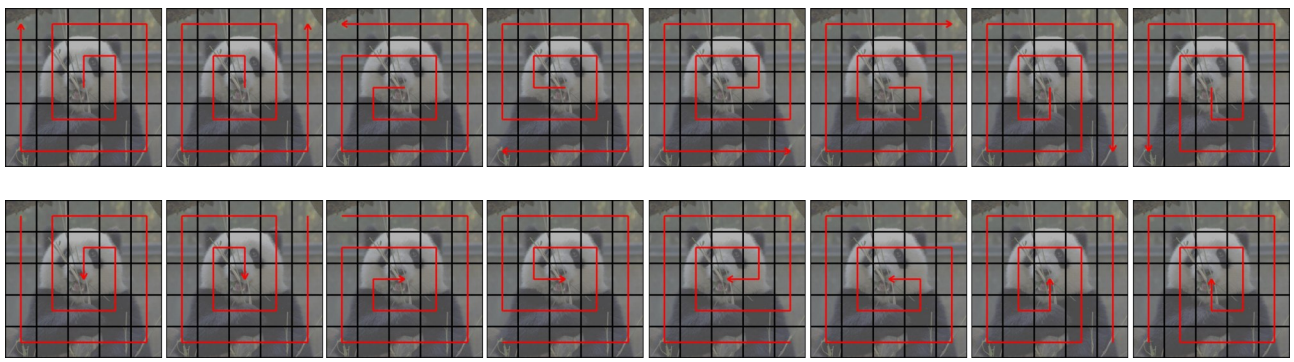

图3 PIS-DM循环扫描模块

4.3 由粗到细的图像生成级联网络

扩散模型需要噪声调度来控制每一步各向同性高斯噪声的数量. 噪声调度的设置对性能有很大的影响,目前大多数模型都遵循线性或余弦时间表. 然而,理想的噪声调度应该是分辨率相关的,直接使用为低分辨率(如32×32或64×64)设计的调度训练高分辨率模型会导致性能次优. 这是因为随着图像尺寸的增加,数据中的冗余信息(如附近像素之间的相关性)也随之增加. 同时,噪声独立添加到每个像素中,使得在高分辨率下更容易恢复原始信号. 由RDM[17]可知:对于高分辨率的图像进行加噪,其加噪后的结果在低频处有更高的信噪比. 一张图像的低频包含了主要的纹理信息,而高频主要为图像的边缘信息. 换言之,图像的低频信息与图像的“语义”有着紧密的联系. 对于高分辨率的

图像,加入同种噪声,其污染程度相较于低分辨率的更轻,这就造成了高分辨率图像下扩散模型训练与采样过程的不一致. 这种训练推理不匹配会在采样过程中逐步累积,从而导致性能下降.

与端到端的方法相比,级联方法在训练效率和噪声调度方面具有显著优势. 首先,级联方法允许为每个阶段调整模型大小和架构,从而找到最有效的组合;其次,低分辨率条件的存在使得早期采样步骤变得容易,因此常见的噪声调度可以作为超分辨率模型的可行基线;最后,级联方法可以利用低分辨率样本的知识,同时保持生成高分辨率图像的能力. 为此,PIS-DM利用由粗到细的图像生成级联网络生成高分辨率图像. 在高分辨率图像生成阶段,模型并非从纯噪声状态出发,而是直接将生成的低分辨率图像$\boldsymbol{X}_L$利用上采样因子s进行缩放:

$$\boldsymbol{X}_t'=\sqrt{\overline{\alpha}_t}\,s\boldsymbol{X}_L+\sqrt{1-\overline{\alpha}_t}\,\varepsilon \tag{9}$$

其中,上采样操作采用最近邻插值法,将低分辨率图像的每个像素复制到高分辨率图像对应的$s\times s$网格区域中,以确保高分辨率图像的初始生成效率. 同时,时间步长是扩散模型最关键的属性之一,针对扩散模型采样早期阶段中样本带噪程度大的问题,PIS-DM在训练过程中将均匀分布采样的时间步,通过一个凸函数映射到更大的一个值,从而对干净数据进行更加有针对性的破坏:

$$\hat{t}=\left[1-\left(\frac{t}{T}\right)^n\right]\times T \tag{10}$$

其中,$\hat{t}$表示重采样时间步长,$t\in U(0,T)$,n是控制时间步重采样大小的超参数.

相比于普通的级联扩散模型,PIS-DM有以下3个优点:一是计算复杂度低,PIS-DM在高分辨率阶段跳过了低频信息的重新生成,减少了训练和采样步骤的数量. 该方法通过由粗到细的图像生成级联网络,直接从低分辨率图像逐步细化至高分辨率图像,避免了从纯噪声状态出发的冗余计算,这种设计不仅提高了训练和推理效率,还降低了整体计算资源的消耗. 二是结构简单,PIS-DM摒弃了条件增强技巧,从而省去了与低分辨率条件相关的交叉注意力计算开销,使得模型结构更加简洁高效. 三是生成精度高,PIS-DM作为一个马尔可夫去噪过程,能够在高分辨率阶段纠正低分辨率图像中的伪影. 由于其逐层精细化的特点,PIS-DM可以在每个阶段对图像进行局部和全局优化,确保生成图像的质量. 特别是在处理高分辨率图像时,PIS-DM能够有效捕捉并纠正低分辨率阶段可能产生的伪影,提升最终生成图像的清晰度和一致性.

与端到端模型相比,PIS-DM在调整模型大小和利用更多低分辨率数据方面更具灵活性. 端到端模型通常需要一次性处理整个图像生成过程,导致模型参数量大、计算复杂度高. 而PIS-DM通过分阶段生成的方式,可以根据实际需求灵活调整模型大小,并充分利用低分辨率数据进行预训练和特征提取,从而提高模型的泛化能力和鲁棒性.

4.4　尺度不变损失函数

互信息(Mutual Information,MI)是一种量化两个变量之间依赖关系的度量[26],随机变量$\boldsymbol{a}\in V$与$\boldsymbol{b}\in W$之间的MI可以表示为

$$\begin{aligned}I(\boldsymbol{a},\boldsymbol{b})&=\mathbb{E}_{p(\boldsymbol{a},\boldsymbol{b})}\left[\log\frac{p(\boldsymbol{a},\boldsymbol{b})}{p(\boldsymbol{a})p(\boldsymbol{b})}\right]\\&=\mathrm{D}_{\mathrm{KL}}\left(p(\boldsymbol{a},\boldsymbol{b})\|\,p(\boldsymbol{a})p(\boldsymbol{b})\right)\end{aligned} \tag{11}$$

其中,$p(\boldsymbol{a},\boldsymbol{b})$表示联合概率分布,$p(\boldsymbol{a})$和$p(\boldsymbol{b})$表示边缘分布,$\mathrm{D}_{\mathrm{KL}}$表示KL散度[27]. 在实际应用中,由于无法获取样本数据的底层分布[28],直接估计两个变量之间的MI变得十分困难. 为此,研究者们通常借助神经网络将MI的估计问题转化为参数优化问题,通过估计观测样本互信息的上界或下界来近似计算变量间的MI. 受AVTrack[29]的启发,本文采用基于Jensen-Shannon散度(Jensen-Shannon Divergence,JSD)的Deep InfoMax互信息估计器[30]来进行MI的估计:

$$\begin{aligned}\hat{I}_{\Theta}^{(\mathrm{JSD})}(\boldsymbol{a},\boldsymbol{b})=&\mathbb{E}_{p(\boldsymbol{a},\boldsymbol{b})}\left[-\alpha\left(-T_{\Theta}(\boldsymbol{a},\boldsymbol{b})\right)\right]\\&-\mathbb{E}_{p(\boldsymbol{a})p(\boldsymbol{b})}\left[\alpha\left(T_{\Theta}(\boldsymbol{a},\boldsymbol{b})\right)\right]\end{aligned} \tag{12}$$

其中,$T_{\Theta}:X\times Y\to\mathbb{R}$表示参数为$\Theta$的神经网络,$\alpha(z)=\ln\left(1+\mathrm{e}^{z}\right)$表示Softplus函数. Deep InfoMax通过最大化局部与全局特征表示之间的互信息,显著提升了模型的学习能力. 本文通过最大化同一目标在不同分辨率下骨干网络特征的互信息,来学习具有尺度不变性的特征表示,尺度不变损失函数L_{sir}定义如下:

$$L_{\mathrm{sir}}=-\hat{I}_{\Theta}^{(\mathrm{JSD})}(\boldsymbol{F}_L,\boldsymbol{F}_H) \tag{13}$$

其中,$\boldsymbol{F}_L$表示低分辨率图像生成网络提取的目标特征,$\boldsymbol{F}_H$表示高分辨率图像重构网络提取的目标特征.

本文提出的尺度不变损失函数不仅关注局部细节和全局结构之间的相互关系,还进一步扩展到不同分辨率下的特征对齐问题. 通过这种方式,模型能够在保持高分辨率细节的同时,获得对尺度变化鲁棒的特征表示. 这种策略不仅增强了模型对多尺度信息的理解能力,还在实际应用中展现出更强的泛化性能. 另一方面,在推理阶段不涉及尺度不变特征表示的学习过程,因此不会增加额外的计算成本.

5　网络训练与图像生成

图像生成过程在分布$p\left(\boldsymbol{X}_T\right)=\mathcal{N}\left(\boldsymbol{X}_T;\,0,\boldsymbol{I}\right)$中通过可学习的核$p_{\theta}\left(\boldsymbol{X}_{t-1}\middle|\,\boldsymbol{X}_t\right)$生成$p_{\theta}\left(\boldsymbol{X}_0\right)$,可学习的高斯核

p_θ可以表示为

$$p_\theta\left(\boldsymbol{X}_{t-1}|\boldsymbol{X}_t\right)=\mathcal{N}\left(\boldsymbol{X}_{t-1};\mu_\theta\left(\boldsymbol{X}_t,t\right),\sigma_\theta\left(\boldsymbol{X}_t,t\right)\boldsymbol{I}\right) \tag{14}$$

其中,均值$\mu_\theta(\cdot)$和方差$\sigma_\theta(\cdot)$是模型可学习的参数,训练过程旨在通过优化这些参数,来最大化$\boldsymbol{X}_0$对数似然函数的变分下界:

$$\max_\theta\left(\begin{array}{c}\sum_t D_{\mathrm{KL}}\left(q\left(\boldsymbol{X}_{t-1}|\boldsymbol{X}_t,\boldsymbol{X}_0\right)\middle\|\, p_\theta\left(\boldsymbol{X}_{t-1}|\boldsymbol{X}_t\right)\right)\\ -\log p_\theta\left(\boldsymbol{X}_0|\boldsymbol{X}_1\right)\end{array}\right) \tag{15}$$

为了简化训练流程,本文利用贝叶斯规则对$q\left(\boldsymbol{X}_{t-1}|\boldsymbol{X}_t,\boldsymbol{X}_0\right)$进行重参数化:

$$q\left(\boldsymbol{X}_{t-1}|\boldsymbol{X}_t,\boldsymbol{X}_0\right)=\mathcal{N}\left(\boldsymbol{X}_{t-1};\tilde{\boldsymbol{\mu}}_t\left(\boldsymbol{X}_t,\boldsymbol{X}_0\right),\tilde{\beta}_t\boldsymbol{I}\right) \tag{16}$$

其中,$\tilde{\beta}_k=\dfrac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_k$,$\alpha_t=1-\beta_t$,$\tilde{\mu}_t\left(\boldsymbol{X}_t,\boldsymbol{X}_0\right)=\dfrac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{X}_0+\dfrac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t}\boldsymbol{X}_t$,$\bar{\alpha}_t=\prod_{t=1}^{T}\alpha_t$.

损失函数的设计旨在促使模型生成接近真实数据分布的样本并学习尺度不变的特征表示,预测噪声$\varepsilon_\theta\left(\boldsymbol{X}_t\right)$与真实的高斯噪声$\varepsilon_t$之间的均方误差与尺度不变损失的加权和:

$$L=\mathbb{E}_{t\sim\mathcal{U}(1,T),\boldsymbol{X}_0\sim q\left(\boldsymbol{X}_0\right),\varepsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[\lambda(t)\left\|\varepsilon_\theta\left(\boldsymbol{X}_t,t\right)-\varepsilon\right\|_2^2\right]+\kappa L_{\mathrm{sir}} \tag{17}$$

其中,$\lambda(t)=\dfrac{{\beta_t}^2}{2\sigma_T^2\alpha_t\left(1-\bar{\alpha}_t\right)}$是改变噪声时刻表的权重,$\kappa=0.4$是超参数. 给定噪声数据$\boldsymbol{X}_T$和时间步长$T$,通过所学得的均值$\mu_\theta(\cdot)$和方差$\sigma_\theta(\cdot)$来重构出原初的图像信息:

$$\begin{aligned}p_\theta\left(\boldsymbol{X}_{t-1}|\boldsymbol{X}_t\right)&=\mathcal{N}\left(\boldsymbol{X}_{t-1};\mu_\theta\left(\boldsymbol{X}_t,t\right),\sigma_\theta\left(\boldsymbol{X}_t,t\right)\boldsymbol{I}\right)\\&\sim\frac{1}{\sqrt{\alpha_k}}\left(\boldsymbol{X}_t\right)-\frac{\beta_k}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta\left(\boldsymbol{X}_t,t\right)\\&\quad+\sigma_\theta\left(\boldsymbol{X}_t,t\right)z\end{aligned} \tag{18}$$

其中,$z\sim\mathcal{N}(0,\boldsymbol{I})$,$\beta_k\approx\sigma_\theta^2\left(\boldsymbol{X}_k,k\right)$.

6 实验与分析

6.1 实验环境、数据集及评价指标

实验训练硬件平台使用配置为6块显存为40 GB的NVIDIA A100 GPU服务器,操作系统为Ubuntu 22.04,软件环境为Python 3.9、PyTorch 1.8、CUDA 11.1和cuDNN 8.0. 本文利用ImageNet[31]数据集对提出的图像生成方法进行实验验证. ImageNet是一个广泛使用的图像分类和生成任务数据集,包含1 281 167张图像,涵盖1 000个不同的类别,为模型提供了丰富的多样性和挑战性. 为了全面评估所提出方法的性能,本文采用了多种评价指标,包括Fréchet Inception Distance(FID)[32]、Spatial Fréchet Inception Distance(sFID)[33]、Inception Score(IS)[34]、精确率(Precision)[35]和召回率(Recall)[35]. FID测量生成图像与真实图像在特征空间中的分布差异,通过预训练的Inception模型提取图像特征并计算生成图像特征分布与真实图像特征分布之间的距离,FID值越低表示生成图像的质量越高. sFID作为FID的一种改进版本,使用空间特征而非标准池化特征来计算分布距离,这种改进能够更好地捕捉图像的空间关系. IS用于评估生成图像的质量和多样性,通过预训练的Inception模型对生成的图像进行分类并计算每个图像属于各个类别的概率分布,较高的IS值表明生成图像具有更高的清晰度和多样性. Precision指在模型预测为正样本的实例中,真正为正样本的比例,在生成模型中Precision关注生成图像的保真度,即生成图像与真实图像的相似程度. Recall指在所有真正的正样本中,被模型正确预测为正样本的比例,在生成模型中Recall关注生成图像的多样性,即模型能否生成多样化且真实的图像. FID和sFID主要用于评估生成图像与真实图像在特征空间中的相似性,其中FID更侧重于整体分布的相似性,而sFID则更注重图像的高层结构,IS主要评估生成图像的清晰度和多样性,Precision和Recall从不同角度评估生成模型的性能,前者关注生成图像的保真度,后者关注生成图像的多样性.

6.2 实现细节

在实验中,本文使用AdamW优化器,其权重衰减系数设置为0.03,betas参数设置为(0.99,0.999). 学习率设定为2×10^{-4},patch大小为2×2,batch size为256,迭代次数(iteration)为500 K,上采样因子s为4,并使用随机翻转进行数据增强. 本文遵循DiM[22]中的设置,使用DPM-Solver作为采样器,并将采样步数设置为50. 指数移动平均(Exponential Moving Average,EMA)的衰减率设定为0.999 9,以保证模型参数的稳定性和平滑性. 此外,本文利用各阶段最后一层Mamba模块输出的特征向量进行尺度不变损失计算,从而增强模型在多尺度下的特征表示能力,Mamba模块的配置与文献[11]保持一致,以确保方法的有效性和可靠性.

6.3 消融实验

为了验证PIS-DM中各个组件的重要性,本文在ImageNet数据集上进行消融实验,实验结果如表1所示. 第一行展示了PIS-DM最佳性能模型的结果,其他行则分别对应于去除某些关键组件后的模型性能. 首先,循环扫描模块(Cyclic Scanning Module,CSM)通过交替执行8个不同的扫描方向,确保每个token拥有全局的感受野,从而有效解决了Mamba应用于二维图像

建模时的难题. 实验结果显示，去除该模块后，模型在生成图像时的细节捕捉能力下降，FID得分明显升高，表明该模块对提升模型性能至关重要. 其次，由粗到细的图像生成级联网络（Cascade Network，CN）通过解耦扩散过程与骨干网络，提升了高分辨率图像生成的质量. 实验结果表明，去除该级联网络会导致生成图像的质量下降，生成效果明显不如完整模型. 再次，尺度不变损失（Scale-Invariant Loss，SIL）通过最大化同一目标在不同分辨率下的互信息，实现了特征表示的有效对齐. 实验结果表明，去除该损失函数会导致生成图像的质量变差，显示出该损失函数对提升模型鲁棒性和一致性的关键作用. 此外，零填充（Padding Tokens，PT）通过引入可学习的边界token，为序列首尾提供了额外的上下文信息，缓解了Mamba因果序列建模中的边界伪影问题. 实验结果表明，去除零填充后，模型的FID值从1.67上升至1.71，验证了零填充对生成质量的提升作用. 最后，轻量级局部结构增强模块（Local Structure Enhancement Module，LSEM）通过引入3 × 3深度卷积层，增强了生成图像的局部连贯性. 尽管该模块对整体性能的影响相对较小，但在细节处理和局部特征捕捉方面仍发挥了重要作用.

表1　在ImageNet数据集上的消融实验

CSM	CN	PT	LSEM	SIL	FID ↓
✔	✔	✔	✔	✔	1.67
	✔	✔	✔	✔	1.98
✔		✔	✔	✔	2.26
✔	✔		✔	✔	1.71
✔	✔	✔		✔	1.75
✔	✔	✔	✔		2.11

为了进一步验证循环扫描模块在PIS-DM中的有效性，本文对不同的扫描模式进行了消融实验，实验结果如表2所示. 从实验结果可以看出，使用多种扫描模式比使用单一的扫描模式具有更低的FID. 这一结果表明，通过交替执行多个不同的扫描方向，可以为Mamba模块提供更好的初始化，从而增强模型对二维图像空间结构的理解能力. 当仅使用单一扫描模式时，生成图像的FID得分相对较高. 这主要是因为单一扫描模式限制了每个token的感受野，导致模型难以捕捉全局的空间关系. 特别是在处理复杂场景和细节丰富的高分辨率图像时，单一扫描模式容易产生伪影和不连贯的特征表示. 相比之下，交替执行8个不同的扫描方向降低了生成图像的FID得分. 这种多方向扫描机制不仅扩展了每个token的感受野，还使得Mamba模块能够更全面地捕捉图像的空间结构信息. 通过这种方式，模型能够在生成过程中更好地保留关键数据，并有效地筛选出无关信息，从而提升生成图像的质量和一致性. 此外，交替执行多个不同的扫描方向不仅有助于提高生成图像的质量，还能为后续的扩散过程提供更好的初始化条件. 通过这种方式，PIS-DM能够在高分辨率阶段纠正低分辨率图像中的伪影，进一步提升生成图像的视觉质量和一致性.

表2　在ImageNet数据集上的扫描模式消融实验

扫描模式	FID ↓
①	1.97
①②	1.93
①②③	1.90
①②③④	1.75
①②③④⑤	1.74
①②③④⑤⑥	1.73
①②③④⑤⑥⑦	1.71
①②③④⑤⑥⑦⑧	1.67

6.4　实验结果与分析

表3展示了PIS-DM在ImageNet数据集上的实验结果，并报告了PIS-DM使用无分类指导器[36]（Classifier-Free Guidance，CFG）的性能指标. 为了确保实验的可比性和一致性，本文将分辨率设置为256 × 256，并在相同实验条件下对DiM[22]进行了复现（标记为DiM*）. 从表3可以看出：PIS-DM在ImageNet数据集上的各项指标均处于领先地位. 与DiM（FID=2.21）和RDM（FID=1.87）相比，PIS-DM的FID值分别降低了0.54和0.20，表明其能够在保持高质量生成的同时，展现出良好的多样性覆盖能力. PIS-DM通过循环扫描模块和尺度不变损失函数的联合优化，不仅继承了DiM和RDM的高效特性，还在局部细节生成和多尺度一致性方面实现了性能提升.

同时，与BigGAN[37]、StyleGAN-XL[38]等基于GAN的图像生成方法相比，PIS-DM借助渐进式生成策略解决了在隐空间出现的模糊和伪影问题，保持了生成图像的高度一致性. 与ADM[39]、LDM[7]、DiT[15]、MDT[40]、U-ViT[8]等基于扩散模型的图像生成方法相比，PIS-DM利用循环扫描模块，提升了生成图像的多样性. 与CDM[41]等基于级联扩散模型的图像生成方法相比，PIS-DM采用Mamba模块避免了冗余计算，提高了生成图像的质量. 与DiffuSSM[18]等基于状态空间模型的图像生成方法相比，PIS-DM通过尺度不变损失，增强了模型对不同分辨率图像的鲁棒性.

另外，本文在512 × 512分辨率下还与DiffuSSM[18]、DiM[22]等基于Mamba的图像生成方法进行了对比分析，实验结果如表4所示. 实验结果表明：PIS-DM在FID、sFID和IS指标上均优于DiffuSSM和DiM，表明其在生成图像的质量与视觉多样性上更具优势. 同时，PIS-DM的Precision和Recall值略高于DiffuSSM和DiM，

说明其生成图像更贴近真实数据分布. 这一优势源于PIS-DM通过渐进式生成策略和尺度不变损失对局部细节与全局一致性的优化,进一步验证了PIS-DM在高分辨率场景中的有效性.

表3 在ImageNet数据集上的对比实验(256×256)

模型	来源	FID ↓	sFID ↓	IS ↑	Precision ↑	Recall ↑
BigGAN-deep[37]	ICLR2019	6.95	7.36	171.40	0.87	0.28
ADM[39]	NeurIPS2021	10.94	6.02	100.98	0.69	0.63
ADM-U/G[39]	NeurIPS2021	3.94	6.14	215.84	0.83	0.53
StyleGAN-XL[38]	SIGGRAPH 2022	2.30	4.02	265.12	0.78	0.53
LDM-4[7]	CVPR2022	10.56	—	103.49	0.71	0.62
LDM-4-G (CFG=1.50)[7]	CVPR2022	3.60	—	247.67	0.87	0.48
DiT-XL/2[15]	ICCV2023	9.62	6.85	121.50	0.67	0.67
DiT-XL/2-G (CFG=1.50)[15]	ICCV2023	2.27	4.60	278.24	0.83	0.57
MDT-XL/2[40]	ICCV2023	6.23	5.23	143.02	0.71	0.65
MDT-XL/2-G (dynamic CFG)[40]	ICCV2023	1.79	4.57	283.01	0.81	0.61
MDT-XL/2-G (CFG=1.325)[40]	ICCV2023	2.26	4.28	246.06	0.81	0.59
CDM[41]	JMLR2022	4.88	—	158.71	—	—
DiffuSSM-XL[18]	CVPR2024	9.07	5.52	118.32	0.69	0.64
DiffuSSM-XL-G[18]	CVPR2024	2.28	4.49	259.13	0.86	0.56
U-ViT-L/2[8]	CVPR2023	3.52	—	—	—	—
U-ViT-H/2[8]	CVPR2023	2.29	—	—	—	—
DiM-Huge*(CFG=4.0)[22]	ARXIV2024	2.21	4.46	263.17	0.85	0.59
RDM[17]	ICLR2024	5.27	4.39	153.43	0.75	0.62
RDM (CFG=3.50)[17]	ICLR2024	1.99	3.99	260.45	0.81	0.58
RDM (CFG=3.50) + class-balance[17]	ICLR2024	1.87	3.97	278.75	0.81	0.59
PIS-DM	Ours	3.95	4.36	223.68	0.71	0.59
PIS-DM(CFG=4.0)	Ours	1.67	3.52	284.65	0.86	0.66

表4 在ImageNet数据集上的对比实验(512×512)

模型	来源	FID ↓	sFID ↓	IS ↑	Precision ↑	Recall ↑
DiffuSSM-XL-G[18]	CVPR2024	3.41	5.84	255.06	0.85	0.46
DiM-Huge* (CFG=4.0)[22]	ARXIV2024	3.94	6.16	235.75	0.84	0.41
PIS-DM(CFG=4.0)	Ours	3.12	5.15	260.73	0.85	0.48

图4展示了PIS-DM在ImageNet数据集上生成的样本图像. 从图4可以看出,PIS-DM不仅能够生成高质量的图像,还能有效捕捉到不同类别的特征,生成的图像在视觉上具有较高的清晰度和多样性. 此外,生成的图像在不同分辨率下保持了一致的高层结构,验证了尺度不变损失函数的有效性.

6.5 计算复杂度分析

为了评估Transformer和Mamba在图像生成任务中的计算复杂度和速度,本文在配备一张NVIDIA A100 GPU的服务器上进行了对比实验,实验结果如表5所示. 为了清晰展示速度差异,表格中显示的速度为每毫秒迭代次数的对数值. 此外,本文通过调整PIS-DM中模块的数量,使其与U-ViT拥有相同的参数量,以确保公平比较. 从实验结果可以看出:当分辨率低于1 024×1 024时,PIS-DM模型的处理速度相较于基于Transformer的模型稍慢;当分辨率大于1 280×1 280时,由于Mamba具有线性的计算复杂度,PIS-DM能够以更快的速度生成质量更高的图像,这进一步验证了PIS-DM在高分辨率图像生成任务中的优越性能和高效性.

6.6 讨论

综合精度、鲁棒性和速度3个性能指标,由上述实验可以看出:与现有图像生成方法相比,本文方法具有较高的性能. 其主要原因在于:一方面,PIS-DM通过由粗到细的图像生成级联网络和Mamba架构,降低了计算复杂度. 这种设计不仅减少了不必要的计算开销,还确保了生成过程的高效性,从而在处理大规模数据集时具备优势. 另一方面,尺度不变损失函数和轻量级局部结构增强模块的应用进一步增强了模型的鲁棒性和多样性覆盖能力. 尺度不变损失函数通过最大化同一目标在不同分辨率下的互信息,实现了特征表示的有

(*a*) 256×256

(*b*) 512×512

(*c*) 1 024×1 024

(*d*) 1 536×1 536

图4　PIS-DM在ImageNet数据集上图像生成结果

表5 在ImageNet数据集上的推理速度对比实验

模型	参数	256×256	512×512	1 024×1 024	1 280×1 280	1 536×1 536	2 048×2 048
PIS-DM	0.9 B	1.4	1.3	1.1	0.9	0.7	0.6
U-ViT	0.9 B	1.9	1.7	1.2	0.9	0.6	0.2

效对齐,确保生成图像在多分辨率任务中的一致性和稳定性. 轻量级局部结构增强模块则通过引入3×3深度卷积层,有效提升了生成图像的局部连贯性和细节清晰度. 这些设计提高了PIS-DM处理复杂多类别场景的鲁棒性,确保其能够生成高质量且多样化的图像. 此外,PIS-DM中的循环扫描模块通过交替执行8个不同的扫描方向,确保每个token拥有全局的感受野,提升了生成图像的细节丰富度和视觉一致性.

在高分辨率生成阶段,PIS-DM采用由粗到细的级联策略,通过低分辨率图像的上采样和加噪逐步细化生成结果. 然而,该策略可能导致高频信息的异常增加. 如图5所示,高频信息的异常增加往往出现在复杂纹理或密集结构区域,表明模型在极端场景下的细节生成能力仍需优化. 高频信息的增加可能与低分辨率阶段的伪影、上采样操作对细节误差的放大,以及高分辨率阶段对细节修正的负担相关. 此外,尺度不变损失虽能增强跨分辨率的语义一致性,但对像素级高频细节的约束有限,难以完全消除上采样伪影. 为缓解这一问题,下一步将研究如何更好地利用多尺度特征信息,通过跨分辨率特征交互减少低分辨率伪影的传递,以进一步增强高分辨率阶段的细节控制能力以及模型对复杂图像内容的生成能力.

图5 PIS-DM在ImageNet数据集上的失败案例

7 结论

针对扩散模型在高分辨率图像生成中计算复杂度高的问题,本文提出一种基于Diffusion-Mamba和尺度不变损失的渐进式图像生成方法PIS-DM. 通过构建多阶段级联扩散架构,并结合多方向扫描机制和轻量级局部结构增强模块,PIS-DM实现了二维图像特征的全局-局部协同建模,同时提升了模型在高分辨率图像生成任务中的性能. 此外,PIS-DM通过引入基于对比学习的尺度不变损失函数,最大化同一目标在不同分辨率下的互信息,确保特征表示的精准对齐,增强了模型对尺度变化的鲁棒性. 实验结果表明,PIS-DM在ImageNet数据集上生成的图像质量和计算效率均优于现有方法,特别是在高分辨率图像生成任务中表现较优.

参考文献

[1] 何琨, 佘计思, 张子君, 等. 基于引导扩散模型的自然对抗补丁生成方法[J]. 电子学报, 2024, 52(2): 564-573.
HE K, SHE J S, ZHANG Z J, et al. A guided diffusion-based approach to natural adversarial patch generation[J]. Acta Electronica Sinica, 2024, 52(2): 564-573. (in Chinese)

[2] 牛玉贞, 张凌昕, 兰杰, 等. 基于分频式生成对抗网络的非成对水下图像增强[J]. 电子学报, 2025, 53(2): 527-544.
NIU Y Z, ZHANG L X, LAN J, et al. FD-GAN: Frequency-decomposed generative adversarial network for unpaired underwater image enhancement[J]. Acta Electronica Sinica, 2025, 53(2): 527-544. (in Chinese)

[3] 罗会兰, 敖阳, 袁璞. 一种生成对抗网络用于图像修复的方法[J]. 电子学报, 2020, 48(10): 1891-1898.
LUO H L, AO Y, YUAN P. Image inpainting using generative adversarial networks[J]. Acta Electronica Sinica, 2020, 48(10): 1891-1898. (in Chinese)

[4] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial Nets[C]//The 27th Advances in Neural Information Processing Systems. New York: ACM, 2014: 2672-2680.

[5] SOHL-DICKSTEIN J, WEISS E A, MAHESWARANATHAN N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//The 32nd International Conference on Machine Learning. Cambridge: PMLR, 2015: 2246-2255.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//The 31st Advances in Neural Information Processing Systems. New York: ACM, 2017: 5998-6008.

[7] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 10674-10685.

[8] BAO F, NIE S, XUE K W, et al. All are worth words: A ViT backbone for diffusion models[C]//2023 IEEE/CVF

Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 22669-22679.

[9] SMITH J T H. Advancing Sequence Modeling with Deep State Space Methods[D]. Stanford: Stanford University, 2024.

[10] KALMAN R E. A new approach to linear filtering and prediction problems[J]. Journal of Basic Engineering, 1960, 82(1): 35-45.

[11] DAO T, GU A. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality[EB/OL]. (2024-05-31)[2025-09-10]. https://arXiv.org/abs/2405.21060.

[12] ZHU L H, LIAO B C, ZHANG Q, et al. Vision Mamba: Efficient visual representation learning with bidirectional state space model[EB/OL]. (2024-11-14) [2025-09-10]. https://arXiv.org/abs/2401.09417.

[13] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//The 34th Advances in Neural Information Processing Systems. New York: ACM, 2020: 6840-6851.

[14] YANG X L, SHIH S M, FU Y L, et al. Your ViT is secretly a hybrid discriminative-generative diffusion model[EB/OL]. (2022-08-16)[2025-09-10]. https://arXiv.org/abs/2208.07791.

[15] PEEBLES W, XIE S N. Scalable diffusion models with transformers[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2024: 4172-4182.

[16] HATAMIZADEH A, SONG J M, LIU G L, et al. DiffiT: Diffusion vision transformers for image generation[C]// Computer Vision - ECCV 2024. Cham: Springer, 2025: 37-55.

[17] TENG J Y, ZHENG W D, DING M, et al. Relay diffusion: Unifying diffusion process across resolutions for image synthesis[EB/OL]. (2023-09-04)[2025-09-09]. https://arXiv.org/abs/2309.03350.

[18] YAN J N, GU J T, RUSH A M. Diffusion models without attention[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 8239-8249.

[19] FAN M, YU C, HUANG J. Scalable diffusion models with state space backbone[EB/OL]. (2024-03-28) [2025-09-10]. https://arxiv.org/abs/2402.05608.

[20] HU V T, BAUMANN S A, GUI M, et al. ZigMa: A DiT-style zigzag mamba diffusion model[C]//Computer Vision - ECCV 2024. Cham: Springer, 2025: 148-166.

[21] PARK J, PARK J, XIONG Z Y, et al. Can Mamba learn how to learn a comparative study on in-context learning tasks[C]//Proceedings of the 41st International Conference on Machine Learning. New York: ACM, 2024: 39793-39812.

[22] TENG Y, WU Y, SHI H, et al. DiM: Diffusion mamba for efficient high-resolution image synthesis[EB/OL]. (2024-07-10)[2025-09-09]. https://arXiv.org/abs/2405.14224.

[23] 刘少鹏, 赵慧民, 洪佳明, 等. 面向医学图像生成的鲁棒条件生成对抗网络[J]. 电子学报, 2023, 51(2): 427-437.
LIU S P, ZHAO H M, HONG J M, et al. Medical image synthesis using robust conditional GAN[J]. Acta Electronica Sinica, 2023, 51(2): 427-437. (in Chinese)

[24] 马宾, 王一利, 徐健, 等. 基于双向生成对抗网络的图像感知哈希算法[J]. 电子学报, 2023, 51(5): 1405-1412.
MA B, WANG Y L, XU J, et al. An image perceptual hash algorithm based on bidirectional generative adversarial network[J]. Acta Electronica Sinica, 2023, 51(5): 1405-1412. (in Chinese)

[25] 黄欣研, 刘芳, 鲍骞月, 等. 基于多任务学习和身份约束的生成对抗网络人脸校正识别方法[J]. 电子学报, 2023, 51(10): 2936-2949.
HUANG X Y, LIU F, BAO Q Y, et al. Multi-task learning and identity-constrained generative adversarial network for face frontalization and recognition[J]. Acta Electronica Sinica, 2023, 51(10): 2936-2949. (in Chinese)

[26] SHANNON C E. A mathematical theory of communication[J]. The Bell System Technical Journal, 1948, 27(3): 379-423.

[27] MACKAY D J C. Information Theory, Inference, and Learning Algorithms[M]. Cambridge: Cambridge University Press, 2003: 1-628.

[28] POOLE B, OZAIR S, VAN DEN OORD A, et al. On variational bounds of mutual information[C]//The 36th International Conference on Machine Learning. Cambridge: PMLR, 2019: 2412-2421.

[29] LI Y X, LIU M Y, WU Y, et al. Learning adaptive and view-invariant vision transformer for real-time UAV tracking[EB/OL]. (2025-08-150)[2025-09-09]. https://arxiv.org/abs/2412.20002.

[30] HJELM R D, FEDOROV A, LAVOIE-MARCHILDON S, et al. Learning deep representations by mutual information estimation and maximization[EB/OL]. (2019-02-22) [2025-09-09]. https://arXiv.org/abs/1808.06670.

[31] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//2009 IEEE

Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009: 248-255.

[32] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6629-6640.

[33] NASH C, MENICK J, DIELEMAN S. Generating images with sparse representations[EB/OL]. (2021-03-05)[2025-09-09]. https://arxiv.org/abs/2103.03841.

[34] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training GANs[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM, 2016: 2234-2242.

[35] KYNKÄÄNNIEMI T, KARRAS T, LAINE S, et al. Improved precision and recall metric for assessing generative models[C]//Proceedings of the 32th Advances in Neural Information Processing Systems. New York: ACM, 2019: 3929-3938

[36] HO J, SALIMANS T. Classifier-free diffusion guidance [EB/OL]. (2022-07-26)[2025-09-09]. https://arxiv.org/abs/2207.12598.

[37] BROCK A, DONAHUE J, SIMONYAN K. Large scale GAN training for high fidelity natural image synthesis[EB/OL]. (2019-02-25)[2025-09-09]. https://arXiv.org/abs/1809.11096.

[38] SAUER A, SCHWARZ K, GEIGER A. StyleGAN-XL: Scaling StyleGAN to large diverse datasets[C]//ACM SIGGRAPH 2022 Conference Proceedings. New York: ACM, 2022: 1-10.

[39] DHARIWAL P, NICHOL A. Diffusion models beat GANs on image synthesis[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: ACM, 2021: 8780-8794.

[40] GAO S H, ZHOU P, CHENG M M, et al. Masked diffusion Transformer is a strong image synthesizer[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2024: 23107-23116.

[41] HO J, SAHARIA C, CHAN W, et al. Cascaded diffusion models for high fidelity image generation[J]. Journal of Machine Learning Research, 2022, 23(1): 2249-2281.

作者简介

李 豪 男,1994年8月出生于山西省晋城市. 现为中国人民解放军陆军工程大学指挥控制工程学院博士研究生. 主要研究方向为计算机视觉、机器学习及其应用.

E-mail: lihao@aeu.edu.cn

郝文宁 男,1971年5月出生于山西省运城市. 2014年博士毕业于中国人民解放军理工大学,现为中国人民解放军陆军工程大学指挥控制工程学院教授、博士生导师. 主要研究方向为数据挖掘、机器学习及其应用.

E-mail: hwnbox@aeu.edu.cn

邹世辰 男,1988年8月出生于黑龙江省哈尔滨市. 2017年博士毕业于哈尔滨工业大学,现为中国人民解放军陆军工程大学指挥控制工程学院讲师. 主要研究方向为数据科学与大数据技术、机器学习及其应用.

E-mail: zoushichen@aeu.edu.cn

谢晓宇 女,1989年6月出生于江苏省徐州市. 2014年硕士毕业于中国矿业大学,现为中国人民解放军陆军工程大学指挥控制工程学院讲师. 主要研究方向为智能教育、机器学习及其应用.

E-mail: 15335198852@163.com