基于潜空间的扩散模型图像生成算法改进与性能分析

杨民青

江南机电设计研究所 贵州 贵阳 550009

【摘　要】：为解决现有生成模型潜在空间表征不足、高分辨率生成质量欠佳等问题，提出基于改进矢量量化变分自编码器（IVQ-VAE）与特征融合 Transformer 扩散（FFTD）模型的双阶段图像生成框架。IVQ-VAE 通过多尺度残差模块与混合损失函数优化潜在表征，FFTD 模型融合注意力机制与多分辨率特征提取提升噪声预测精度，采用 DDIM 采样加速推理。在 CelebA-HQ、AFHQ 数据集上的实验表明，该框架 FID 值最低达 9.64，生成质量优于主流模型，验证了方法的有效性。

【关键词】：图像生成；潜在空间；扩散模型；矢量量化；特征融合

DOI:10.12417/3041-0630.26.06.027

引言

图像生成领域中，潜在空间表征能力与扩散去噪精度直接影响生成效果，现有 VAE、GAN、流模型及扩散模型存在特征拟合不足、多尺度信息利用不充分、训练稳定性差等问题，高分辨率图像生成仍存在细节缺失与结构畸变，相关模块融合方式仍需进一步优化以提升整体性能。

1 相关工作

1.1 变分自编码器

VAE 基于编码器-解码器与变分推断，将图像映射至连续潜在空间，通过重建损失与 KL 散度完成分布建模，但像素级约束导致细节模糊。VQ-VAE 引入离散码本提升结构表达，VQ-VAE2 通过多尺度增强重建能力，仍缺乏跨层交互。VDVAE 及其改进提升表达与效率，但存在梯度与全局语义不足问题，整体与扩散模型适配性较弱。

1.2 扩散模型

扩散模型通过加噪与去噪实现生成，DDPM 稳定但计算开销大，DDIM 提升采样效率。分类器引导与自适应策略改善质量但增加复杂度。潜在扩散结合 VAE 降维，但编码与去噪特征不匹配，多尺度融合不足，导致高分辨率生成细节与结构受限[1]。

1.3 生成对抗网络

GAN 通过对抗训练提升图像质量，DCGAN 实现基础生成但不稳定。PGGAN 与 StyleGAN 支持高分辨率生成但训练敏感。VQGAN 结合离散表示提升建模能力，但训练易失衡。整体存在模式崩溃、可控性弱及与扩散模型潜在空间不兼容问题。

1.4 流模型

流模型通过可逆变换实现精确建模，Glow 等方法具备理论优势，但语义表达不足且计算复杂。可逆性限制结构设计，高分辨率下冗余明显，与扩散模型融合时存在维度不匹配与信息损失问题。

2 方法

2.1 改进矢量量化变分自编码器（IVQ-VAE）

IVQ-VAE 基于 VQ-VAE 重构编解码器，引入多尺度残差与多头注意力，实现高效潜在映射与量化表示。编码器采用卷积与残差结构逐步下采样，解码器对称恢复并结合注意力增强全局建模。损失函数由重建、感知、对抗与量化损失构成，其中量化损失定义为

$$L_{vq} = \| z_e - z_q \|_2^2$$

式中，z_e为编码器输出特征，z_q为码本量化后的特征向量，该损失用于约束编码特征与码本向量的一致性。模型通过预训练与联合优化提升潜在表达稳定性与重建质量。

2.2 特征融合 Transformer 扩散（FFTD）模型

FFTD 基于 Transformer，在潜在空间构建多分辨率特征提取与融合机制，提升噪声预测精度。模型以 IVQ-VAE 潜在向量为输入，通过多级下采样与上采样结合残差与注意力结构，实现跨层特征交互与全局建模[2]。引入通道与空间注意力提升特征表达能力，并通过跳跃连接融合多尺度信息。训练阶段最小化噪声预测误差

$$L_{diff} = \| \varepsilon - \varepsilon_\theta(z_t, t) \|_2^2$$

式中，ε 为前向扩散施加的高斯噪声，$\varepsilon_\theta(z_t, t)$ 为模型在时间步 t 对噪声潜变量z_t的噪声预测输出，训练阶段冻结 IVQ-VAE 参数，仅优化 FFTD，以提升复杂结构与细节建模能力。

2.3 双阶段训练与 DDIM 潜在空间采样推理

整体采用双阶段训练与 DDIM 采样策略。第一阶段对 IVQ-VAE 进行预训练，实现高分辨率图像到潜在空间压缩；第二阶段基于 FFTD 在潜在空间进行扩散建模，通过噪声预测优化模型参数。推理阶段采用 DDIM 从高斯噪声逐步去噪生成潜在表示，采样更新为：

$$z_{t-1}=\sqrt{\bar{\alpha}_{t-1}}\hat{z}_0+\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\,\varepsilon_\theta(z_t,t)+\sigma_t\varepsilon_t$$

式中，$\hat{z}_0$为模型预测的初始潜在向量，$\bar{\alpha}_t$为噪声累积乘积系数，σ_t为随机项控制系数，ε_t 为高斯噪声。最终将潜在结果输入 IVQ-VAE 解码器生成高分辨率图像。该策略降低训练复杂度并提升采样效率。

3 实验

3.1 数据集

实验采用 CelebA-HQ、AFHQ 以及 AFHQ-Dog、AFHQ-Cat、AFHQ-Wild 三个子数据集完成模型性能验证，CelebA-HQ 选取 30000 张名人图像，经 PGGAN 处理后分辨率为 1024×1024，AFHQ 包含猫、狗、野生动物三类动物面部图像，每类 5000 张样本，分辨率为 512×512，所有数据集图像均执行中心裁剪操作，采用 256×256 像素尺寸完成模型训练。

3.2 实验设置

实验运行环境为 Windows10 系统，基于 PyTorch2.5.0 框架搭建，搭载显存 12GB 的 NVIDIARTX4070Super 显卡，使用 Python3.9 完成代码实现，模型训练学习率设置为 1×10^{-4}，批次大小 BatchSize 为 1，扩散周期设置为 500，总训练步数为 4×10^6，码本空间维度为 16384，潜在空间维度设为 4，EMA 系数为 0.9999，优化器采用 Adam，小型数据集 dropout 系数设为 0.1，通用数据集 dropout 系数设为 0，模型性能采用 FID 指标完成量化评估，FID 计算式为

$$\mathrm{FID}=\|m_r-m_g\|_2^2+\mathrm{tr}(C_r+C_g-2\sqrt{C_rC_g})$$

其中m_r、C_r为真实图像特征均值与协方差，m_g、C_g为生成图像特征均值与协方差，实验生成 50000 张图像完成 FID 统计。

3.3 对比实验

对比实验包含潜在空间压缩倍率、图像分块大小、采样步数、基线方法与标准数据集五类对比验证，在标准数据集对比中，本文方法优于 Glow、VAE、VQ-VAE 等 FLOW 与 VAE 类模型，FID 表现接近主流 GAN 与 Diffusion 模型。见表 1；

表 1 不同潜在空间压缩倍率下模型重建指标对比

压缩倍率	潜在分辨率	重建 L1 损失	LPIPS 损失	重建 FID
8	32×32	0.082	0.215	Nan
16	16×16	0.075	0.183	87.36
32	8×8	0.068	0.157	48.67
64	4×4	0.062	0.132	18.93

3.4 消融实验

消融实验分析 FFTD 特征融合方式对生成质量与效率的影响，比较 Add、Concat 与自适应融合三种方案，在 CelebA-HQ 与 AFHQ 上评估参数量、训练时间与 FID。Add 参数 9.54M、3.8h，FID 为 64.96/72.31；Concat 参数 11.49M、4.6h，FID 为 48.53/54.12；自适应融合参数 14.98M、5.7h，FID 降至 9.64/10.25。结果表明，自适应融合虽增加开销，但显著提升多尺度特征整合能力，实现最佳生成质量[3]。

4 结语

通过双阶段框架优化潜在空间表征与扩散去噪过程，在标准数据集上取得优异 FID 指标，验证了改进组件的有效性。但生成图像背景存在细节缺失问题，后续将聚焦关键信息提取与特征处理优先级策略，平衡核心区域与背景建模，进一步提升生成图像的整体真实性与场景一致性。

参考文献：

[1] 刘浩南,陈姚节,高登科.潜在空间下扩散模型图像生成[J].计算机系统应用,2026,35(03):170-183.

[2] 侯哲晓,李弼程,蔡炳炎,等.基于改进扩散模型的高质量图像生成方法[J].计算机科学,2025,52(S1):461-469.

[3] 操伟业.基于生成对抗网络的潜在空间语义表达算法研究[D].南京邮电大学,2022.