计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.010098] [CSTR: 32024.14.csa.010098]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 潜在空间下扩散模型图像生成①

刘浩南, 陈姚节, 高登科

(武汉科技大学 计算机科学与技术学院, 武汉 430070)
通信作者: 陈姚节, E-mail: chenyaojie@wust.edu.cn

**摘　要**: 本文旨在解决传统图像生成模型在复杂场景下潜在空间表示能力不足及高分辨率图像生成保真度低的问题, 提出一种基于改进矢量变分自编码器 (improved vector quantized variational autoencoder, IVQ-VAE) 和特征融合 Transformer 扩散 (feature-fused Transformer diffusion, FFTD) 模型的双阶段训练框架. IVQ-VAE 通过引入注意力机制、残差块和多重损失函数, 显著提升潜在空间的语义表达能力与生成图像的保真度, 克服了传统编码器在复杂图像特征捕获上的局限性; FFTD 模型基于 Transformer 架构, 结合多分辨率采样和自适应特征融合, 进一步增强模型对复杂图像结构的建模能力. 双阶段训练策略首先预训练 IVQ-VAE 以生成高质量潜在表示, 随后冻结其参数, 利用去噪扩散隐式模型 (DDIM) 训练 FFTD 模型以优化噪声预测和图像生成的过程, 该框架在 CelebA-HQ 和 AFHQ 等数据集上生成图像的细节保真度和视觉质量均有显著提升, 验证了其在高分辨率图像生成中的有效性.

**关键词**: 变分自编码器; 扩散模型; Transformer; 图像生成

引用格式: 刘浩南,陈姚节,高登科.潜在空间下扩散模型图像生成.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/10098.html

## Image Generation Based on Diffusion Model in Latent Space

LIU Hao-Nan, CHEN Yao-Jie, GAO Deng-Ke

(School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430070, China)

**Abstract**: Aiming to address the problems of traditional image generation models, including insufficient latent space representation capability in complex scenes and low fidelity in high-resolution image generation, this study proposes a two-stage training framework based on an improved vector quantized variational autoencoder (IVQ-VAE) and a feature-fused Transformer diffusion (FFTD) model. By introducing the attention mechanism, residual blocks, and multi-component loss function, IVQ-VAE significantly enhances the semantic representation capability of the latent space and fidelity of generated images, overcoming the limitations of traditional encoders in capturing complex image features. Built on a Transformer architecture, FFTD further improves the modeling capacity of complex image structures by integrating multi-resolution sampling and adaptive feature fusion. The two-stage training strategy first pre-trains IVQ-VAE to generate high-quality latent representations, then freezes its parameters, and trains FFTD by employing a denoising diffusion implicit model (DDIM) to optimize the noise prediction and image generation process. This framework achieves significant improvements in detail fidelity and visual quality of the generated images on datasets such as CelebA-HQ and AFHQ, validating its effectiveness in high-resolution image generation.

**Key words**: variational autoencoder; diffusion model; Transformer; image generation

① 收稿时间: 2025-08-27; 修改时间: 2025-09-24; 采用时间: 2025-10-14; csa 在线出版时间: 2026-01-19

## 1 引言

近年来, 生成模型在人工智能领域取得了显著进展, 尤其是在图像生成、音频合成以及其他复杂数据分布建模任务中表现卓越. 图像生成技术在计算机视觉领域中扮演着至关重要的角色, 其应用涵盖图像增强、风格迁移、图像修复和高分辨率重建等多个方面. 随着深度学习技术的迅猛发展, 图像生成模型不断突破, 涌现出流模型[1]、变分自编码器[2]、生成对抗网络[3]和扩散模型[4]等多种创新方法.

然而这些方法各有局限: 流模型的生成速度慢, 且对数据类型有较高要求、降维能力有限; GAN 虽在图像生成表现出色, 但训练稳定性差, 易出现模式崩溃; VAE 生成的图像细节模糊, 难以满足高分辨率需求; 扩散模型虽训练稳定且生成质量高, 但在无条件生成场景中, 因计算成本高和复杂结构建模能力有限, 生成图像保真度受影响.

目前, 基于潜在空间的训练方式正成为图像生成领域的研究热点. 这类模型通过在潜在空间中对数据分布进行建模, 有效压缩高维数据并捕捉关键语义信息, 为生成高质量图像提供了新的可能性. 基于潜在空间的模型, 如自编码器 (autoencoder)[5]及其变体, 通过将输入数据映射到离散或连续的潜在表示, 再结合解码器生成图像, 在降低计算复杂度和提升生成质量方面展现出潜力. 然而, 现有基于潜在空间的训练方法仍面临关键挑战.

(1) 潜在空间的表示能力受限于编码器的特征提取效率和损失函数的设计, 难以充分捕捉复杂场景的全局特征和细粒度细节.

(2) 在高分辨率图像生成中, 潜在空间的解码易导致信息损失, 影响生成图像的保真度和纹理质量.

为应对上述挑战, 本文提出双阶段训练框架, 结合改进矢量量化变分自编码器 (IVQ-VAE) 和特征融合 Transformer 扩散 (FFTD) 模型, 通过增强编码器与解码器的结构以及优化损失函数, 提升潜在空间表示能力和生成图像保真度; 通过 FFTD 模型的自适应融合与多分辨率采样, 增强复杂场景下的建模能力; 该方法充分利用了 IVQ-VAE 的语义压缩能力与 FFTD 模型多尺度特征捕获和信息整合能力, 能够有效解决传统 VQ-VAE 特征提取效率低和生成质量有限的问题, 同时优化扩散过程中的噪声预测, 突破多尺度特征建模和跨层信息整合瓶颈, 在生成图像中保留精细细节和纹理, 为生成高质量、高分辨率的图像提供了新的解决方案.

## 2 相关工作

本文的核心贡献在于提出双阶段训练框架, 通过增强潜在空间语义表示、优化复杂结构建模能力、提升采样效率, 解决传统模型潜在空间表示不足、高分辨率保真度低、复杂结构建模弱、采样效率低的问题. 以下从生成模型 4 大主流方向, 结合本文核心改进, 梳理相关工作并明确本文方法的针对性研究.

### 2.1 流模型

流模型通过一系列可逆可微分的变换将复杂数据分布映射到简单分布, 核心优势是密度估计精准, 但在潜在空间语义建模与高分辨率采样效率上存在显著局限. Kingma 等[6]提出的 Glow 模型采用可逆的 1×1 卷积, 实现了图像的可控生成, 但由于缺乏潜在空间结构化的语义编码, 导致高分辨率图像处理速度慢, 且潜在空间特征冗余. Hoogeboom 等[7]对 Glow 进行改进, 将 1×1 可逆卷积扩展为 $d\times d$ 可逆卷积, 该方法虽提升了灵活性, 但未解决语义建模不足的问题. Xu 等[8]研发的泊松流模型通过泊松方程加速推理, 但仅针对采样速度进行优化, 仍无法避免高分辨率图像细节丢失问题.

### 2.2 变分自编码器

变分自编码器因其能学习结构化潜在表示而成为生成模型的重要分支, 但传统方法在复杂特征捕获能力、细节保真度与长距离依赖建模上存在明显不足. Tomczak 等[9]首次提出 VAE, 通过编码器-解码器框架结合重建损失和 KL 散度损失建模数据分布, 但由于仅依赖像素级损失, 导致传统 VAE 生成的图像模糊. van den Oord 等[10]随后提出了 VQ-VAE, 通过离散化潜在空间显著提升重建质量, 但编码器仍为简单的卷积堆叠, 无法有效捕捉复杂图像的多层级特征. 后续研究进一步推进 VAE 的优化, Razavi 等[11]提出的 VQ-VAE2 通过分层多尺度的潜在映射机制提升分辨率, Child 等[12]提出的 VDVAE 将 VAE 与深度嵌入学习相结合, 使用多层神经网络实现数据的非线性降维. Hazami 等[13]提出 Efficient-VDVAE, 通过改进 VDVAE 使模型收敛速度提升、内存负载减少的同时增强训练稳定性, 但均缺乏全局特征整合机制. Vahdat 等[14]提出的 NVAE 通过层次潜在结构增强模型的表达能力, NCP-VAE[15]和 DC-VAE[16]分别引入归一化流和解耦策略,

进一步提升样本多样性和视觉质量, 然而, 并没有解决“深层网络梯度消失”与“感知层面保真度低”的问题.

### 2.3 生成对抗网络

生成对抗网络自 Goodfellow 等[17]提出以来, 因其生成高保真图像的能力而成为图像生成领域的核心技术, 但训练不稳定与模式崩溃是其固有缺陷. Radford 等[18]提出的 DCGAN 将 CNN 与 GAN 结合, 有效捕捉图像轮廓的语义属性, 但高分辨率生成时易出现模式崩溃. Karras 等[19–22]提出 PGGAN 与 StyleGAN 通过逐步提升分辨率、嵌入潜在向量, 显著提升图像的质量, 但仍受限于对抗训练的不稳定性, 导致人脸、动物等样本属性的控制难度大. Esser 等[23]提出 VQGAN 结合矢量量化和 GAN, 改进了潜在空间建模能力, StyleALAE[24]和 StyleNeRF[25]采用自适应潜在编码和三维感知生成技术, 进一步提升图像生成的多样性和一致性, 但都未摆脱“超参数敏感、训练易发散”的问题.

### 2.4 扩散模型

扩散模型凭借训练稳定性和高生成质量成为近年研究热点, 但其在无条件生成场景中存在两大局限: 一是复杂结构建模能力弱, 传统模型采用单一分辨率处理, 难以捕捉图像的多尺度细节; 二是采样效率低, 计算成本高. Ho 等[26]提出去噪扩散概率模型 (DDPM) 通过构造马尔可夫链逐步向数据添加高斯噪声, 并通过逆向去噪过程重建数据, 但反向推理的 1 000 步采样耗时久, 且对复杂纹理的建模能力不足. Nichol 等[27]提出改进版本实现竞争性对数似然, 在反向扩散过程中学习方差使采样效率提升多个数量级, 但该方法未优化特征建模架构; Dhariwal 等[28]通过分类器指导增强条件图像合成的保真度与多样性, Wang 等[29]将扩散框架用于 GAN 训练, 引入自适应扩散过程及与扩散时间步长相关的鉴别器和生成器, Song 等[30]提出去噪扩散隐式模型 (DDIM) 改进 DDPM, 通过非马尔可夫逆向过程大幅减少采样步骤, 同时保持生成质量, 但复杂结构建模仍依赖单一分辨率, 无法有效对全局-局部特征进行有效整合. CDM[31]通过条件建模增强样本多样性, 但仅适用于类条件场景, 无条件生成时保真度仍受限.

## 3 本文方法

本文提出基于双阶段训练的图像生成框架, 结合改进的矢量量化变分自编码器 (IVQ-VAE) 和特征融合 Transformer 扩散 (feature-fused Transformer diffusion, FFTD) 模型实现高质量图像生成. 方法分为 3 个阶段: 1) IVQ-VAE 预训练, 构建高质量潜在表示; 2) 冻结 IVQ-VAE 参数, 在潜在空间中训练 FFTD 模型, 优化噪声预测; 3) 基于 DDIM 采样生成图像. 整体架构如图 1 所示.

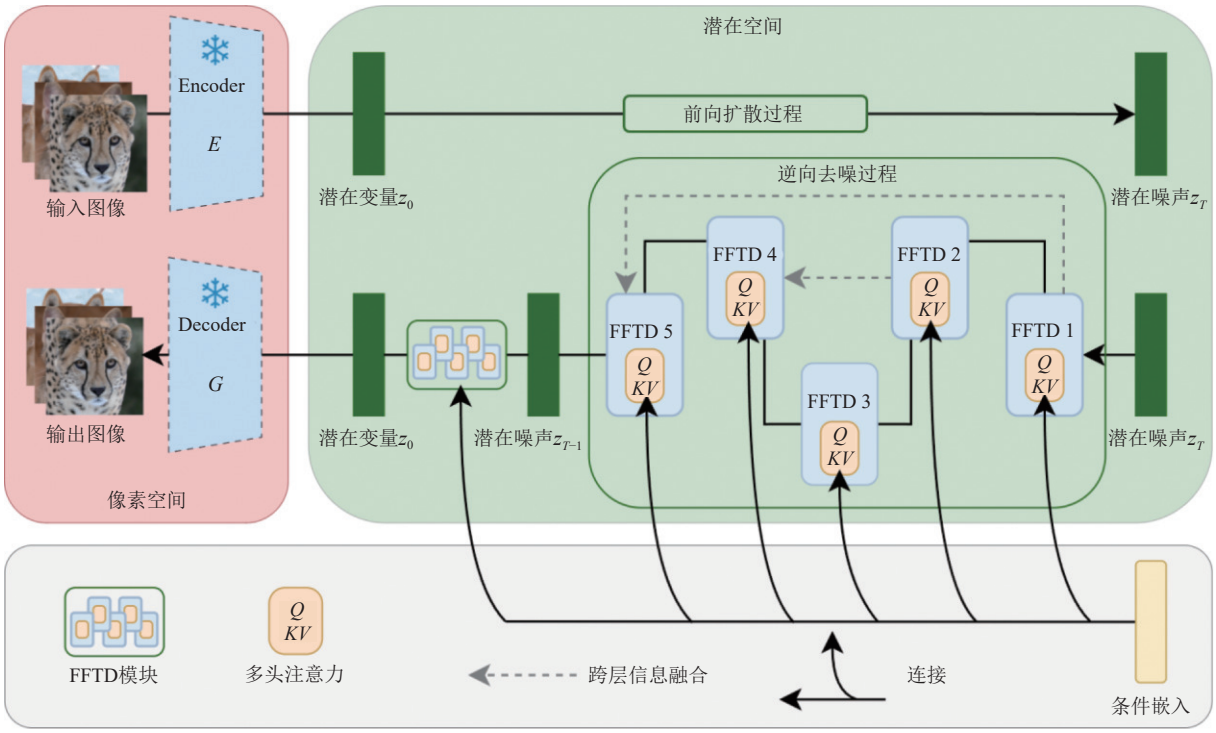


图 1 基于 IVQ-VAE 和 FFTD 模型的生成图像方法

### 3.1 改进矢量量化变分自编码器 (IVQ-VAE)

IVQ-VAE 基于 VQ-VAE 改进, 结构如图 2 所示, 通过引入注意力机制、残差连接、多分辨率采样和混合损失函数, 显著提升潜在空间表示能力和生成图像质量. 在特定分辨率层引入注意力机制以捕获全局上下文信息, 确保编码器输出的低维潜在表示能有效保留输入图像的关键语义特征, 并通过解码器重建高质量图像.

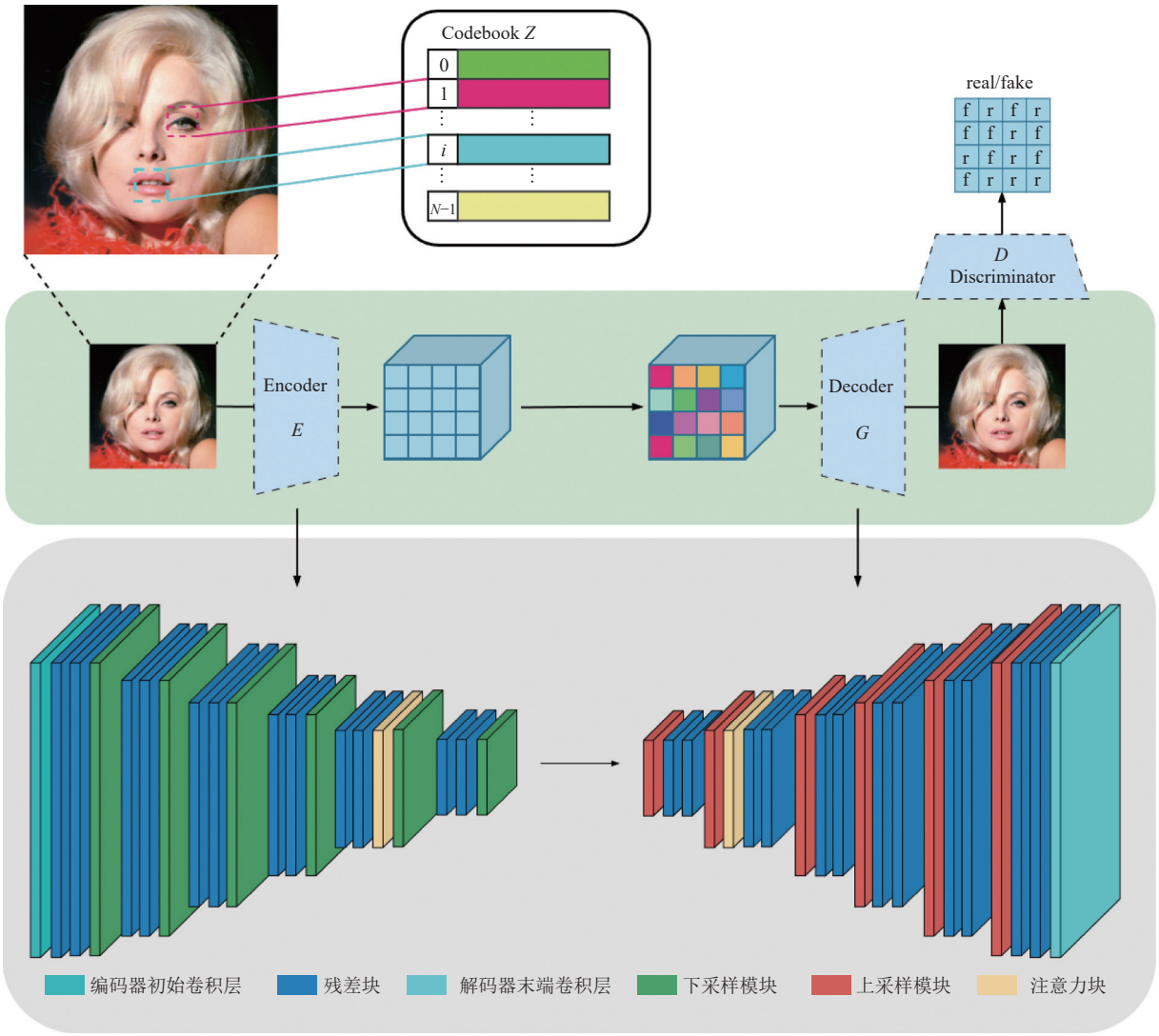


图 2 IVQ-VAE 结构

算法 1 展示了 IVQ-VAE 预训练具体流程, 包括初始化参数与码本、迭代训练 (编码、量化、解码、计算损失、更新参数), 最终输出训练好的编码器、解码器和码本, 为后续生成高质量潜在表示奠定基础.

算法 1. IVQ-VAE 预训练 (阶段 1)

1. 初始化:
IVQ-VAE 参数$\{\phi,\psi\}$和码本$Z$
2. 迭代训练:
对每个批次$x\sim\mathcal{D}$:
编码: $z=E_\phi(x)$
量化: $z_q=Quantize\langle z,Z\rangle$
解码: $\hat{x}=G_\psi(z_q)$
计算损失: $\mathcal{L}_{\text{total}}=\mathcal{L}_{\text{rec}}+\mathcal{L}_P+\lambda_d\mathcal{L}_G+\lambda_{\text{vq}}\mathcal{L}_{\text{vq}}$
更新参数: 梯度下降优化$\{\phi,\psi,Z\}$
若步数大于 250k, 更新判别器参数: $\nabla_{\theta_d}\mathcal{L}_d$
3. 输出: 固定训练好的$E_\phi,G_\psi,Z$

#### 3.1.1 多尺度编码器与解码器设计

IVQ-VAE 能逐步提取从低级到高级的特征, 相较于传统编码器和解码器的简单卷积堆叠, 可捕捉更丰富的多尺度信息. 残差块通过跳跃连接缓解深层网络

梯度消失问题, 增强模型对复杂图像结构的建模能力, 显著提升特征提取能力.

编码器由一系列残差块、下采样模块和注意力模块组成, 逐步将输入图像压缩为低分辨率潜在表示. 其输入为 $x \in \mathbb{R}^{C\times H\times W}$ ($C$为通道数, $H\times W$为空间分辨率), 输出为$z \in \mathbb{R}^{C_z\times H_z\times W_z}$ (其中$H_z = H/f$, $W_z = W/f$, 其中$f$是压缩倍数).

编码器初始卷积层采用步幅为 1 的 3×3 卷积核将输入通道数从$C$映射到初始特征通道数$C_h$:

$$h_0 = Conv_{3\times 3}(x; C \to C_h) \tag{1}$$

残差块由层归一化 (GroupNorm)、*Swish* 激活函数和 3×3 卷积构成. 编码器包含多个阶段, 每个阶段的通道数由通道乘数$ch_{\text{mult}}$控制, 输出通道数为$ch\cdot m_i$, 其中$m_i$为对应阶段的乘数. 每个阶段包含两个残差块, 定义为:

$$h_{i+1} = h_i + Conv_{3\times 3}(Swish(GroupNorm(h_i))) \tag{2}$$

其中, *Swish* 激活函数定义为: $Swish(x) = x\cdot\sigma(x)$, $\sigma(x) = (1+\mathrm{e}^{-x})^{-1}$. 残差块通过分组归一化和 *Swish* 激活函数增强了训练稳定性和非线性表达能力.

在除最后一个阶段外的每个阶段后, 应用下采样模块 (DownSample), 将空间分辨率减半, 通过步幅为 2 的 3×3 卷积实现空间分辨率减半的下采样:

$$h_{i+1} = Conv_{3\times 3, stride=2}(h_i) \tag{3}$$

经过多次下采样, 将图片从高分辨率的清晰图像压缩成低分辨率的潜在表示.

最终, 末端卷积层将特征映射至嵌入维度$C_z$:

$$z = Conv_{1\times 1}(h_{\text{last}}; C_h \to C_z) \tag{4}$$

解码器与编码器对称, 采用上采样模块和残差块, 逐步将潜在表示$z_q \in \mathbb{R}^{C_z\times H_z\times W_z}$重构为输出图像$\hat{x} \in \mathbb{R}^{C\times H\times W}$.

初始卷积部分将量化表示映射至高维特征空间:

$$h_0 = Conv_{1\times 1}\left(z_q; C_z \to C_h\cdot m_{\text{last}}\right) \tag{5}$$

其中, $m_{\text{last}}$为通道倍数列表末项. 按照通道乘数的逆序$\overline{ch}_{\text{mult}}$, 每个阶段包含两个残差块, 结构与编码器相同.

在除最后一个阶段外的每个阶段后, 应用上采样模块 (UpSample), 将空间分辨率翻倍, 通过最近邻插值 (interpolate) 和步幅为 2 的 3×3 卷积实现分辨率翻倍的上采样:

$$h_{i+1} = Conv_{3\times 3}(Interpolate(h_i, scale = 2)) \tag{6}$$

最终输出层使用卷积与 tanh 激活生成重建图像:

$$\hat{x} = \tanh(Conv_{3\times 3}(h_{\text{last}}; C_h \to C)) \tag{7}$$

3.1.2 注意力机制优化

IVQ-VAE 在编码器和解码器的特定分辨率层 (16×16) 引入注意力模块 (AttnBlock), 增强全局上下文捕获能力, 确保生成的潜在表示能够保留关键全局和局部特征.

在指定分辨率层引入多头自注意力模块, 增强全局特征建模能力:

$$Q = Conv_{1\times 1}(h),\ K = Conv_{1\times 1}(h),\ V = Conv_{1\times 1}(h) \tag{8}$$

$$W = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{C_h}}\right) \tag{9}$$

$$h_{\text{attn}} = h + Conv_{1\times 1}(W\cdot V) \tag{10}$$

3.1.3 混合损失函数设计

为优化重建保真度与视觉质量, IVQ-VAE 引入混合损失函数, 结合像素级重建损失、感知损失[32]和对抗损失[17], 在训练编码器和解码器的过程中实现多目标优化.

像素级重建损失($\mathcal{L}_{\text{rec}}$): 通过 L1 范数约束重构图像与输入图像的像素级差异:

$$\mathcal{L}_{\text{rec}} = \|\hat{x} - x\|_1 \tag{11}$$

感知损失($\mathcal{L}_P$): 基于预训练 VGG 网络的特征空间相似性度量:

$$\mathcal{L}_P = LPIPS(\hat{x}, x) = \sum_l \omega_l \|\phi_l(\hat{x}) - \phi_l(x)\|_2^2 \tag{12}$$

其中, $\phi_l$为 VGG 网络第$l$层特征, $\omega_l$为$l$层权重.

对抗损失: 引入判别器$Dis$提升生成图像真实性.

生成器损失定义为:

$$\mathcal{L}_G = -\mathbb{E}[Dis(\hat{x})] \tag{13}$$

判别器损失定义为:

$$\mathcal{L}_{Dis} = \frac{1}{2}(\mathbb{E}[1 - Dis(x)] + \mathbb{E}[1 + Dis(\hat{x})]) \tag{14}$$

其中, $x$为真实图像, $\hat{x}$为生成图像.

总损失函数定义为:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_P + \lambda_d\mathcal{L}_G + \lambda_{\text{vq}}\mathcal{L}_{\text{vq}} \tag{15}$$

其中, $\mathcal{L}_{\text{vq}}$是量化损失, 用于缩小编码器输出特征与码本 (Codebook) 向量的差异, 同时约束码本向量的更新方向, 避免量化过程中信息丢失. $\lambda_d$和$\lambda_{\text{vq}}$是对应损失项的权重系数, 且$\lambda_d + \lambda_{\text{vq}} = 1$.

### 3.2 特征融合 Transformer 扩散 (FFTD) 模型

基于特征融合的 Transformer 扩散 (FFTD) 模型是对传统 Transformer[28]架构的改进, 通过添加多分辨率采样和自适应特征融合, 优化了复杂图像生成任务中的噪声预测能力. FFTD 模型结合多尺度特征提取和跨层特征融合策略, 增强对图像全局和局部特征的建模能力, 从而生成更高质量的图像. 算法 2 展示了 FFTD 模型在潜在空间的训练流程, 首先冻结 IVQ-VAE 参数, 再构建潜在表示, 初始化 FFTD 模型参数后, 通过加噪、去噪、计算损失和更新参数完成扩散过程训练, 以优化噪声预测能力.

算法 2. FFTD 潜在空间训练算法 (阶段 2)

1. 冻结 IVQ-VAE 参数
2. 潜在表示构建
对 $x\sim\mathcal{D}$, 计算离散潜变量 $z_q=Quantize\langle E_\phi(x),Z\rangle$
3. 初始化
随机初始化 FFTD 模型参数 $\theta$
4. 扩散过程训练
对每个 $x\sim\mathcal{D}$, 随机采样 $t\sim Uniform(1,T)$
加噪: 根据 DDIM 正向过程计算 $z_t$
去噪: FFTD 模型预测噪声 $\epsilon_\theta=G_\theta(z_t,t,y)$
计算损失: $\mathcal{L}_{\text{diff}}=\|\epsilon-\epsilon_\theta\|_2^2$
更新参数: 梯度下降优化 $\theta$

#### 3.2.1 FFTD 整体架构

FFTD 模型基于 Transformer 架构, 如图 3 所示, 结合时间嵌入、标签嵌入和图像分块嵌入, 其中图像分块嵌入 (patchify) 采用 1×1 像素分块尺寸, 将潜在空间中的每一个特征点视为一个图像块, 通过 FFTD 块进行特征处理. FFTD 模型的核心包括以下两个方面.

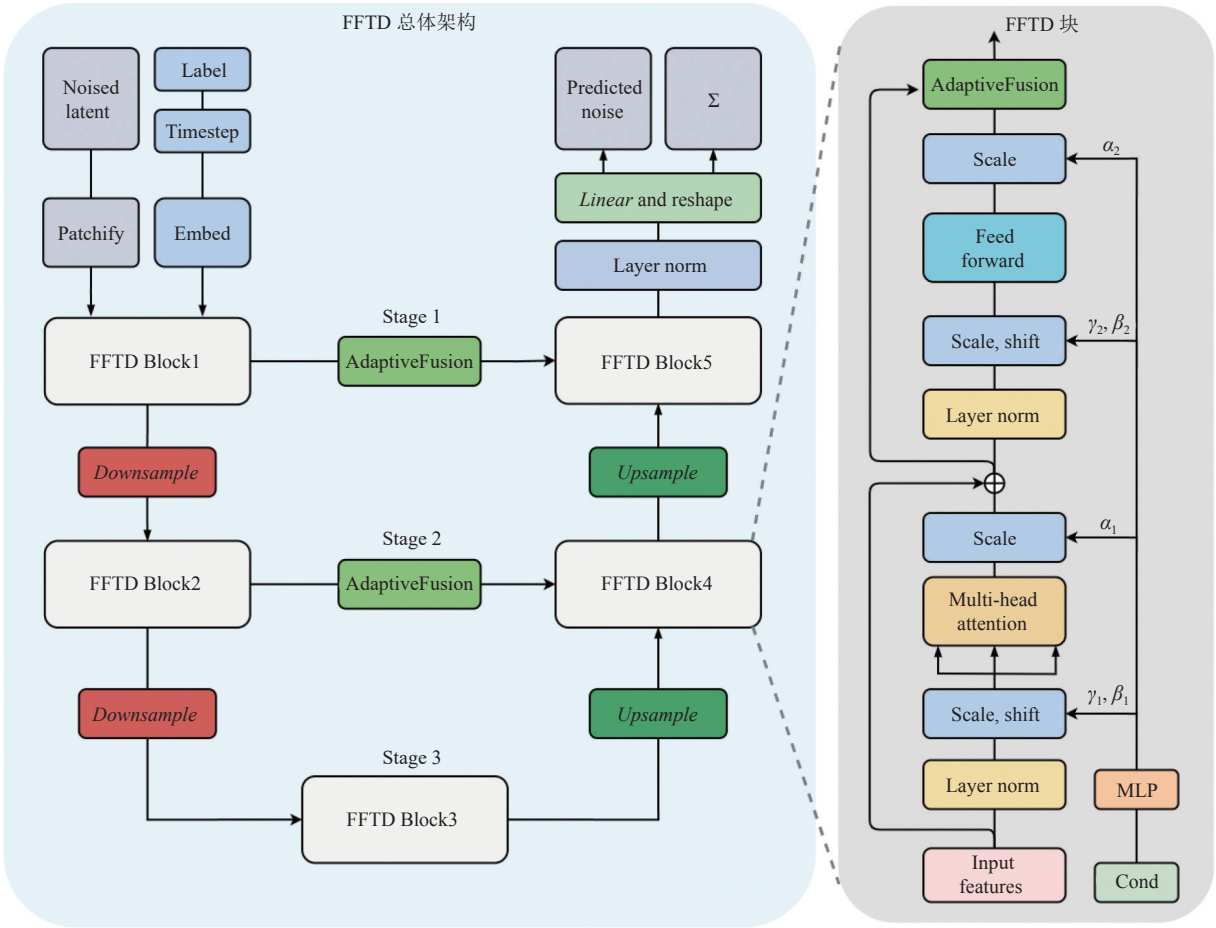


图 3 FFTD 模型架构和 FFTD 模块

多分辨率采样策略: 通过引入上采样和下采样模块, FFTD 模型能够在不同分辨率下提取和整合特征, 增强对多尺度信息的建模能力.

自适应特征融合: 通过自适应融合模块 (AdaptiveFusion), FFTD 模型实现了跨层特征的动态融合, 显著提升了特征表达的丰富性.

3.2.2 多分辨率采样机制

传统 Transformer 模型通常在单一分辨率下处理特征, 难以有效捕捉图像的多尺度信息. FFTD 模型采用多分辨率采样的策略: 在块与块之间设计了下采样 (*Downsample*) 和上采样 (*Upsample*) 模块, 构建一个多分辨率处理流水线, 解决多尺度特征捕获上的局限.

下采样模块通过步幅为 2 的卷积操作将特征图的空间分辨率减半, 定义如下:

$$Downsample(x) = Conv2d(x) \tag{16}$$

其中, $x \in \mathbb{R}^{B\times C\times H\times W}$, $Conv2d$表示二维卷积, 采用 kernel 为 3 的卷积核大小, stride 为 2 的步幅, padding 为 1 的填充, 输出特征图为$x \in \mathbb{R}^{B\times C\times H/2\times W/2}$.

上采样模块通过最近邻插值 (nearest interpolation) 和卷积操作将特征图分辨率翻倍, 定义如下:

$$Upsample(x) = Conv2d(Interpolate(x)) \tag{17}$$

其中, 输出特征图为$x \in \mathbb{R}^{B\times C\times 2H\times 2W}$.

3.2.3 自适应特征融合模块

自适应特征融合模块 (AdaptiveFusion) 能够动态融合 FFTD 块内、不同 FFTD 块间的输出特征, 用来增强跨层信息整合能力, 其结构如图 4 所示. 与传统 Transformer 通过简单的残差连接进行特征传递不同, AdaptiveFusion 结合了通道注意力和空间注意力机制, 动态调整特征权重以优化特征表达.

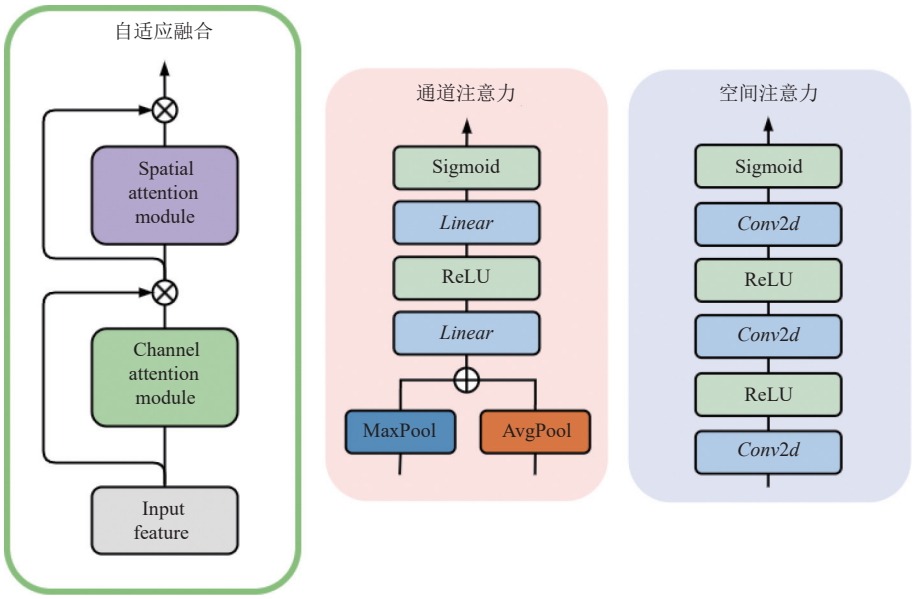


图 4 自适应融合模块

自适应特征融合模块的输入为两个特征向量$x_1$, $x_2 \in \mathbb{R}^{B\times L\times D}$, 输出为融合特征$x_{\text{fused}} \in \mathbb{R}^{B\times L\times D}$. 该模块在 FFTD 模型中有 3 种融合方式, 如图 3 所示, 在 FFTD Block1 与 FFTD Block5 的融合链路中, $x_1$为 FFTD Block1 输出的原始潜在分辨率 (4×4) 全局结构特征, $x_2$为 FFTD Block5 输出的经上采样恢复至原始分辨率的局部细节特征, 以全局结构约束局部细节, 避免高分辨率恢复时的结构错位与细节失真, 保障全局结构一致性; 在 FFTD Block2 与 FFTD Block4 的融合链路中, $x_1$为 FFTD Block2 输出的 1/2 潜在分辨率 (2×2) 中层结构特征, $x_2$为 FFTD Block4 输出的 1/2 潜在分辨率中层细节特征, 强化中层特征连贯性, 衔接全局与局部信息, 避免多尺度特征处理的信息断层; 在 FFTD 块内, $x_1$为多头注意力层的输出, $x_2$为前馈网络层的输出, 这种融合方式能够整合长距离捕捉与局部非线性优化能力, 提升块内特征表达能力, 增强局部建模精度. 其计算过程如下.

通过平均池化和最大池化提取通道特征, 计算通道注意力权重.

$$\begin{cases} p_{\text{avg}} = Mean(x_1 \,\|\, x_2, dim = 1) \\ p_{\max} = Mean(x_1 \,\|\, x_2, dim = 1) \end{cases} \tag{18}$$

$$p_{\text{channel}} = Concat\left(p_{\text{avg}}, p_{\max}\right) \tag{19}$$

$$w_{\text{channel}} = \sigma(Linear_2(Linear_1(p_{\text{channel}}))) \tag{20}$$

其中, $\|$表示沿通道维度的拼接, $\sigma$为 Sigmoid 激活函数, $Linear_1 : \mathbb{R}^{2D} \to \mathbb{R}^{D/16}$, $Linear_2 : \mathbb{R}^{D/16} \to \mathbb{R}^{D}$.

将一维序列重塑为二维特征图, 计算空间注意力权重.

$$x_{\text{spatial}} = Reshape(x_1 \parallel x_2; B, 2D, \sqrt{L}, \sqrt{L}) \tag{21}$$

$$w_{\text{spatial}} = \sigma\left(Conv_3\left(Conv_2\left(Conv_1\left(x_{\text{spatial}}\right)\right)\right)\right) \tag{22}$$

其中, $Conv_1 : \mathbb{R}^{2D} \to \mathbb{R}^{D/4}$, $Conv_2 : \mathbb{R}^{D/4} \to \mathbb{R}^{D/4}$, $Conv_3 : \mathbb{R}^{D/4} \to \mathbb{R}^1$, 卷积核大小为 3.

结合通道和空间注意力权重, 以及残差连接进行特征融合.

$$x_{\text{channel}} = w_{\text{channel}} \cdot LayerNorm(x_1) \tag{23}$$

$$x_{\text{spatial}} = w_{\text{spatial}} \cdot LayerNorm(x_2) \tag{24}$$

$$x_{\text{fused}} = x_{\text{channel}} + x_{\text{spatial}} + s \cdot Linear(x_1 \parallel x_2) \tag{25}$$

其中, $s$为可学习的残差比例因子, $Linear : \mathbb{R}^{2D} \to \mathbb{R}^{D}$.

### 3.3 去噪概率扩散隐式模型加速采样

扩散模型的核心思想是通过前向加噪过程将数据分布逐步转化为高斯噪声分布, 再通过逆向去噪过程从噪声中逐步重建数据. DDIM[4]在 DDPM[26]的基础上优化了逆向过程, 使其成为非马尔可夫过程, 从而允许更灵活的采样策略. 算法 3 展示了基于 DDIM 图像生成流程, 包括从标准高斯分布采样潜变量, 通过 FFTD 模型和 DDIM 采样器迭代去噪, 最后解码得到图像, 实现高效的高质量图像生成.

算法 3. DDIM 采样算法 (阶段 3)

1. 采样潜变量:
从标准高斯分布采样$z_T \sim \mathcal{N}(0,I)$.
2. 迭代去噪:
使用训练好的 FFTD 模型 $G_\theta$和 DDIM 采样器, 从$t$=$T$到$t$=1:
预测噪声$\epsilon_\theta = G_\theta(z_t, t, y)$
更新$z_{t-1} = \sqrt{\alpha_{t-1}}\hat{z}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\epsilon_\theta(z_t,t) + \sigma_t\epsilon_t$
$\hat{z}_0$是预测初始数据, $\sigma_t$是控制随机性的参数, $\epsilon_t \sim \mathcal{N}(0,I)$
3. 解码图像:
$\hat{x} = D_\psi(Quantize\langle z_0, Z\rangle)$

#### 3.3.1 扩散过程与潜在空间结合

前向过程定义为马尔可夫链: 对于 IVQ-VAE 生成的初始潜在表示$z_0$, 通过$T$个时间步逐步添加高斯噪声, 最终将其转化为高斯噪声分布. 前向过程的条件分布为:

$$q(z_t|z_{t-1}) = \mathcal{N}\left(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t I\right) \tag{26}$$

其中, $\beta_t \in (0,1)$是时间步$t$的噪声调度参数, 表示噪声强度; 训练采样器的过程中定义了一个从$\beta_1$到$\beta_T$的线性调度序列.

逆向过程为非马尔可夫链, 允许从任意时间步$t$的潜变量$z_t$推断$z_{t-1}$, 与 FFTD 模型的特征融合能力深度协同: FFTD 模型捕获潜变量在不同尺度下的结构信息, DDIM 则利用这些信息优化噪声预测精度. 逆向过程中, FFTD 模型的噪声预测网络$\epsilon_\theta(z_t,t)$被训练以预测前向过程中添加的噪声$\epsilon$, 其核心是最小化损失函数:

$$\mathcal{L} = \mathbb{E}_{z_0,t,\epsilon}\left[\epsilon - \epsilon_\theta(z_t,t)_2^2\right] \tag{27}$$

$$z_t = \sqrt{\alpha_t}z_0 + \sqrt{1-\alpha_t}\epsilon,\ \epsilon \sim \mathcal{N}(0,I) \tag{28}$$

其中, 累积乘积$\alpha_t = \prod_{s=1}^{t}(1-\beta_t)$表示从初始数据到时间步$t$的信号保留比例. $z_t$是初始潜变量$z_0$与噪声的加权组合.

#### 3.3.2 高效采样解码生成图像

DDIM 通过隐式采样减少所需的时间步数. 采样公式基于以下形式:

$$z_{t-1} = \sqrt{\alpha_{t-1}}\hat{z}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\epsilon_\theta(z_t,t) + \sigma_t\epsilon_t \tag{29}$$

其中, $\hat{z}_0$是模型预测的初始数据, $\sigma_t$是控制随机性的参数, $\epsilon_t \sim \mathcal{N}(0,I)$. 训练采样器时通过设置$\eta$控制$\sigma_t$, 当$\eta = 0$时, 采样过程变为确定性过程, 从而显著加速采样.

采样完成后, 得到的潜变量$z_0$将通过解码器, 最终重建为图像:

$$\hat{x} = D_\psi(Quantize\langle z_0, Z\rangle) \tag{30}$$

## 4 实验

### 4.1 数据集

为了充分验证本文方法在图像生成上的有效性, 实验选用 CelebA-HQ[20]、AFHQ[33]两个标准数据集及 AFHQ 的 3 个子类别数据集 (AFHQ-Dog, AFHQ-Cat 和 AFHQ-Wild).

(1) CelebA-HQ 数据集, 从 CelebA 中选取 30 000 张名人图片, 使用 PGGAN[19]提升图像的质量, 分辨率提高到 1024×1024, 丰富人脸的细节, 是 CelebA 的高质量版本.

(2) AFHQ 数据集包含动物面部图像, 从猫、狗和野生动物 3 个类别中各选取 5 000 张样本, 分辨率为 512×512, 动物面部有复杂纹理.

实验数据集采用中心裁剪, 使用 256×256 像素的图像进行训练.

### 4.2 实验设置

本文在 Windows 10 操作系统、PyTorch 2.5.0 环境下, 使用 GPU 为 12 GB 显存的 NVIDIA RTX4070 Super 进行实验. 使用 Python 3.9 编程语言, 实验参数

设置: 学习率$1\times10^{-4}$; BatchSize 为 1; 扩散周期为 500; 训练总步数为$4\times10^{6}$; 码本空间设置为 16 384; 潜在空间设置为 4; EMA 设置为 0.999 9; 采用 Adam 优化器; dropout 在小型数据集上的设置为 0.1, 在通用数据集上设置为 0. 同时, 对于生成图像的评估, 使用 Fréchet inception distance (*FID*) 值[34], Fréchet 距离是一种度量两个分布之间距离的方法, 它考虑到了两个分布的均值和协方差矩阵, 可以更好地描述两个分布之间的差异, *FID* 值越低, 则生成的图片越接近真实图片样本, 生成的质量越好. 评估阶段加载训练好的 FFTD 模型和 IVQ-VAE 模型, 生成 50 000 张图像, 计算 *FID* 如下:

$$FID = \|m - m_\omega\|_2^2 + \mathrm{tr}\left(C + C_\omega - 2(CC_\omega)^{1/2}\right) \tag{31}$$

其中, $m$和$C$分别为真实图像特征向量的均值和协方差矩阵, $m_\omega$和$C_\omega$分别为生成图像特征向量的均值和协方差矩阵.

### 4.3 对比实验

为了充分评估本文方法的有效性, 首先进行实验确定最优压缩倍率、图像分块大小、采样步数. 再与基线方法及代表性方法在标准数据集上对比. 所有实验均使用 50 000 条生成样本的 *FID* 值.

#### 4.3.1 潜在空间压缩倍率对比实验

潜在空间压缩倍率作为 IVQ-VAE 模型的核心参数, 直接决定了编码器对输入图像特征的压缩效率与潜在表示的信息保留能力. 为明确其对生成质量的影响, 实验以 CelebA-HQ 数据集为对象, 测试了压缩倍率$f\in\{8,16,32,64\}$ (对应潜在分辨率 32×32、16×16、8×8、4×4) 下的重建 L1 损失、*LPIPS* 损失及重建 *FID* 值, 并结合可视化结果展开多维度分析.

由表 1 可知, 随着压缩倍率从 8 提升至 64, 模型的重建 L1 损失从 0.082 降至 0.062, *LPIPS* 损失从 0.215 降至 0.132, 重建 *FID* 值从无法有效计算 (Nan) 逐步优化至 18.93, 这一变化源于 IVQ-VAE 多尺度编码器的结构优势: 低压缩倍率 (*f* 为 8 和 16) 下, 编码器对图像的压缩程度较低, 潜在空间维度较高, 虽能保留部分原始图像特征, 但受限于特征提取效率, 无法充分整合全局语义与局部细节, 导致像素级差异 (L1 损失) 和感知层面差异 (*LPIPS* 损失) 较大; 而高压缩倍率 (*f* 为 32 和 64) 下, 编码器通过残差块与注意力机制的协同作用, 可逐步剥离冗余信息, 将关键特征浓缩至低维度潜在空间, 使潜在表示更精准地匹配原始图像的语义分布, 进而降低各类损失指标.

表 1 不同压缩倍率下的损失与重建 *FID* 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 |
| 128 | 2×2 | — | — | — |

从图 5 的可视化结果来看, 不同压缩倍率的重建效果差异直观反映了潜在空间质量的变化. 当 *f*=8 时, 生成图像细节模糊, 人脸的五官轮廓、发丝纹理等关键特征几乎无法辨认, 表明此时潜在空间未能有效捕获图像核心信息; *f*=16 虽能呈现较明显的人脸形态, 但眼部细节缺失、肤色过渡不均, 仍存在显著的特征丢失问题; *f*=32 时, 人脸特征清晰度大幅提升, 五官比例趋于合理, 仅在眉毛、胡须等精细纹理处存在轻微模糊; *f*=64 则进一步优化细节表现, 不仅五官轮廓清晰可辨, 甚至能还原睫毛密度、面部光影等高频纹理, 生成效果更接近真实图像.

(a) *f*=8 (b) *f*=16 (c) *f*=32 (d) *f*=64

图 5 潜在空间压缩倍率对比试验结果

值得注意的是, 压缩倍率并非越高越好. 实验中 *f*=128 因潜在分辨率降至 2×2, 导致模型参数量激增、计算负担超出硬件承载能力, 最终无法完成训练. 这表明压缩倍率的提升需平衡信息保留能力与计算效率: 过低的倍率会导致潜在空间冗余, 增加后续 FFTD 模型的噪声预测难度; 过高的倍率则会超出硬件算力限

制, 还会导致潜在表示丢失关键语义信息. 综合定量指标与实验可行性, $f$=64 成为最优选择, 既能通过高压缩倍率实现 18.93 的低 *FID* 值, 保证重建质量, 又能将潜在分辨率控制在 4×4 的合理范围, 避免计算资源浪费, 为后续 FFTD 模型在潜在空间的高效训练奠定基础.

4.3.2 图像分块对比实验

为探究图像分块大小 (patchsize) 对 FFTD 模型性能和计算效率的影响, 本实验在 CelebA 数据集上进行测试, 分块大小分别设置为 1×1、2×2、4×4, 对比分析不同方案下的模型训练时间及生成图像 *FID* 值.

如表 2 所示, 当分块大小为 1×1 时, *FID* 值达到最低, 生成质量最高, 显著优于其他分块方法. 这是因为 1×1 分块将潜在空间中的每个特征点视为独立图像块, 能够完整保留潜变量的细粒度局部信息, 综合生成质量与计算效率, 本实验选择 1×1 作为最优图像分块大小, 该方案需承担略高的参数量与训练时间, 但能最大限度保留潜在空间的细节信息, 为 FFTD 模型输出高保真图像提供基础.

表 2 不同分块大小对比实验

| 分块大小 | 分块数量 | 模型参数量 (M) | 训练时间 (h)* | *FID* |
|---|---|---|---|---|
| 1×1 | 16 | 14.98 | 5.7 | 9.64 |
| 2×2 | 4 | 13.73 | 5.3 | 14.38 |
| 4×4 | 1 | 12.76 | 5.2 | 19.85 |

注: *表示每$4\times10^5$步的训练时间

4.3.3 采样步数对比实验

扩散模型的采样步数直接影响去噪迭代的充分性与采样效率, 为确定 DDIM 采样器的最优步数, 本实验在 CelebA-HQ 与 AFHQ 数据集上测试不同 DDIM 采样步数 (50、100、500 步) 和 DDPM 采样步数 (1 000 步) 的生成 *FID* 值与单张图像采样时间, 实验结果如表 3 所示.

表 3 不同 DDIM 采样步数对比实验

| 采样步数 | *FID* | | 采样时间 (ms/张) |
|---|---|---|---|
| | CelebA-HQ | AFHQ | |
| 50 | 37.98 | 59.84 | 851 |
| 100 | 16.83 | 18.46 | 1 140 |
| 200 | 14.37 | 15.41 | 1 431 |
| 500 | 9.64 | 10.25 | 4 026 |
| 1 000 | 14.96 | 12.31 | 10 386 |

随着采样步数从 50 增加至 500, CelebA-HQ 数据集的 *FID* 值从 37.98 降至 9.64, AFHQ 数据集从 59.84 降至 10.25, 降幅分别达 74.5% 与 82.7%, 这一结果表明, 更多的采样步数为 FFTD 模型提供了更充分的逆向去噪迭代空间, 使其能够更精准地预测潜在空间中的噪声分布, 但步数增加也会导致采样时间延长. 实验结果显示, DDIM 的 500 步采样在生成质量和采样速率上均优于 DDPM 的 1 000 步采样, 因此本实验确定 500 步 DDIM 为最优采样策略, 可在生成质量与计算效率间实现最佳平衡.

4.3.4 基线方法对比实验

本文选用的基线方法为 DDPM[26], 其具体设置: 采用 U-Net 作为主干网络, 包含 4 个下采样阶段和 4 个上采样阶段, 每个阶段含 2 个残差块, 注意力模块仅在 16×16 分辨率层引入; 参数设置: 学习率$2\times10^{-4}$, BatchSize=16, 总训练步数$4\times10^5$步, EMA 衰减率为 0.999 9, 噪声调度$\beta_1=1\times10^{-4}$, $\beta_2=0.02$, 采样步数为 1 000 时采用原始 DDPM 调度, 500 步采用线性插值噪声调度, 均使用$\varepsilon$预测模式; 数据处理与本文方法一致, 图像调整为 256×256, 归一化至[−1, 1].

与基线方法 DDPM 的对比实验中, 将 DDIM 采样步数设置为 500 步, 选取基线方法 1 000 步和 500 步的 *FID* 值的结果进行对比, 实验结果如表 4 所示.

表 4 不同数据集下与基线对比实验结果

| 方法 | 步数 | CelebA-HQ | AFHQ | Dog | Cat | Wild |
|---|---|---|---|---|---|---|
| Baseline | 1 000 | 14.96 | 12.31 | 18.29 | 8.15 | 6.64 |
| Baseline | 500 | 18.55 | 15.62 | 23.70 | 13.45 | 8.98 |
| Ours | 500 | 9.64 | 10.25 | 15.56 | 7.38 | 4.21 |

根据表 4 可知, 本文方法在 CelebA-HQ、AFHQ 及 AFHQ 的 3 个子类别上的 *FID* 值均显著低于基线方法, 表明在相同或更少采样步数下, 生成图像质量更优, 尤其在小数据集上改进明显.

4.3.5 标准数据集对比实验

表 5、表 6 是本文方法在标准数据集 CelebA-HQ 和 AFHQ 上的对比结果, 实验选取 FLOW、VAE、GAN、Diffusion 等类型的相关方法进行对比, 在 *FID* 值的比较中, 本文方法在 CelebA-HQ 上达到了 9.64, 在 AFHQ 上达到了 10.25, 优于 VAE 和 FLOW 模型, 接近大多数 GAN 类模型, 在两个标准数据集上均展现出较强的竞争力, 验证了其在高质量图像生成中的有效性.

### 4.4 消融实验

4.4.1 IVQ-VAE 核心组件消融实验

为了验证 IVQ-VAE 中核心组件对样本质量的提升作用, 本文设置了 5 组消融实验, 由表 7 可知, 当注意力机制、残差块和混合损失均不使用时, 无法生成

有效样本; 仅使用单一组件时, *FID* 值较高, 图像质量较差; 而三者同时使用时, *FID* 值最低, 表明注意力机制、残差块和混合损失的协同作用能显著提升 IVQ-VAE 的性能, 有效提高生成样本质量.

表 5 标准数据集 CelebA-HQ (256×256) 对比实验结果

| 模型类别 | 方法 | *FID* |
|---|---|---|
| FLOW | Glow[6] | 68.93 |
| VAE | VAE[9] | 97.07 |
| | VQ-VAE[10] | 68.20 |
| | NVAE[14] | 40.26 |
| | NCP-VAE[15] | 24.79 |
| | DC-VAE[16] | 15.81 |
| | VAEBM[35] | 20.38 |
| | D2C[36] | 18.74 |
| | DiffuseVAE[37] | 11.28 |
| GAN | PGGAN[19] | 8.0 |
| | StyleALAE[24] | 19.21 |
| | VQGAN[23] | 10.20 |
| Diffusion | LSGM[27] | 7.22 |
| | SDE[27] | 7.25 |
| | DDMI[27] | 7.25 |
| | Ours | 9.64 |

### 4.4.2 FFTD 模块特征融合机制消融实验

为了进一步验证 FFTD 模型中自适应融合模块和多分辨率采样的作用, 设置了 4 组消融实验, 表 8 展示了消融实验结果, 在两个数据集上, 采用自适应特征融合或者多分辨率采样策略的 *FID* 值均有所降低. 图 6 为 FFTD 模块消融生成图像的对比, 直观呈现了在不同模块配置下模型生成图像的效果差异.

表 6 标准数据集 AFHQ (256×256) 对比实验结果

| 模型类别 | 方法 | *FID* |
|---|---|---|
| GAN | HoloGAN[38] | 77.98 |
| | GRAF[38] | 121.04 |
| | GIRAFFE[38] | 30.98 |
| | π-GAN[25] | 46.87 |
| | StyleSDF[38] | 12.80 |
| | StyleNeRF[25] | 14.00 |
| | StyleGAN2[21] | 11.37 |
| Diffusion | CDM[31] | 24.20 |
| | Ours | 10.25 |

表 7 IVQ-VAE 消融实验结果

| 注意力机制 | 残差块 | 混合损失 | CelebA-HQ | AFHQ |
|---|---|---|---|---|
| × | × | × | Nan | Nan |
| √ | × | × | 97.26 | 108.31 |
| × | √ | × | 68.31 | 75.80 |
| × | × | √ | 36.57 | 48.84 |
| √ | √ | √ | 9.64 | 10.25 |

表 8 FFTD 模块消融实验结果

| 自适应融合 | 多分辨率采样 | CelebA-HQ | AFHQ |
|---|---|---|---|
| × | × | 64.96 | 72.31 |
| √ | × | 31.37 | 43.21 |
| × | √ | 24.52 | 31.68 |
| √ | √ | 9.64 | 10.25 |


图 6 FFTD 模块消融实验生成图像对比结果

由图 6 可知, 当 FFTD 模块仅采用简单连接, *FID* 值较高, 生成图像质量最差, 仅使用多分辨率采样生成

的图像特征混乱, 仅采用自适应融合的图像有了准确的特征但图片存在模糊问题, 而自适应融合和多分辨率采样同时使用时, 达到了最低的 *FID* 值, 充分验证了自适应融合模块和多分辨率采样策略在提升 FFTD 模型生成图像质量方面的重要作用, 且两者协同工作时能最大程度增强模型对复杂图像结构的建模能力, 优化生成效果.

#### 4.4.3 融合方式消融实验

为验证 FFTD 模块中不同特征融合方式对图像生成质量、模型复杂度及推理效率的影响, 本实验设计 3 种融合方案: 1) 简单 Add: 对特征进行直接相加; 2) 简单 Concat: 沿通道维度拼接特征后直接输入后续层; 3) 自适应融合: 结合通道注意力与空间注意力动态调整特征权重. 实验使用 CelebA-HQ 与 AFHQ 两个数据集, 对比不同方案的模型参数量、训练时间及 *FID* 值, 结果如表 9 所示.

由实验可知, 相比简单 Add 与简单 Concat, 自适应融合方式虽增加模型参数量与训练时间, 但能最大程度发挥 FFTD 模型的多尺度特征整合能力, 在两个数据集上均取得最低 *FID* 值, 是实现高质量图像生成的核心组件.

表 9 融合方式消融实验结果

| 融合方式 | 模型参数量 (M) | 训练时间 (h)* | CelebA-HQ | AFHQ |
|---|---|---|---|---|
| 简单Add | 9.54 | 3.8 | 64.96 | 72.31 |
| 简单Concat | 11.49 | 4.6 | 48.53 | 54.12 |
| 自适应融合 | 14.98 | 5.7 | 9.64 | 10.25 |

注: *表示每$4\times10^5$步的训练时间

### 4.5 样本展示

如图 7 所示, (a) 是使用 AFHQ 数据集生成的动物面部样本, (b) 是使用 CelebA-HQ 数据集生成的人脸样本, 生成的动物面部与人脸都有着清晰的细节与纹理, 如 AFHQ 样本中动物的毛发层次、CelebA-HQ 样本中人脸的睫毛密度与面部光影过渡均接近真实样本, 验证了模型在核心区域上生成的有效性.


图 7 模型生成样本展示

图 8 展示了模型在复杂场景生成中存在局限性的典型案例, 从视觉效果可见, 模型生成动物面部、人脸仍能保持一定的细节完整性, 如动物的口鼻轮廓、人脸的五官比例基本合理, 但背景区域存在显著缺陷: AFHQ 动物样本的背景出现无意义的纹理混乱, 部分样本的背景也存在明显割裂感, 缺乏自然的环境过渡; CelebA-HQ 人脸样本的背景多呈现为均匀的低精度模糊区域, 仅能保留简单的颜色基调, 无法还原高保真的背景语义信息, 甚至部分样本出现背景像素粘连.

## 5 结束语

本文提出的基于 IVQ-VAE 和 FFTD 模型的双阶段框架, 通过增强潜在空间表示与优化扩散过程, 在 CelebA-HQ、AFHQ 数据集上实验结果表明, 本文方法相比多种主流模型有着较为明显的提升, 同时一系列的消融实验也证明该模型的改进组件有着良好效果. 然而, 该方法生成的图像背景部分仍然存在着模糊的问题, 其核心原因在于: IVQ-VAE 在高压缩倍率 (最优 $f$=64) 下优先保留面部、毛发等核心语义, 背景冗余信息被过度剔除; FFTD 模型多分辨率采样与自适应融合模块聚焦核心区域多尺度建模, 背景特征因权重分配偏低而弱表达; 混合损失函数侧重优化主要部分的保真度, 对背景约束不足. 未来将研究更有效的图像关键信息提取方法, 采用更为明确的特征处理优先级策略,

兼顾背景特征的保留与优化, 以进一步提升生成图像的整体真实性与场景一致性.


图 8 复杂场景生成中存在局限性的样本展示

## 参考文献


1 胡铭菲, 左信, 刘建伟. 深度生成模型综述. 自动化学报, 2022, 48(1): 40–74.

2 翟正利, 梁振明, 周炜, 等. 变分自编码器模型综述. 计算机工程与应用, 2019, 55(3): 1–9.

3 陈佛计, 朱枫, 吴清潇, 等. 生成对抗网络及其在图像生成中的应用研究综述. 计算机学报, 2021, 44(2): 347–369.

4 刘泽润, 尹宇飞, 薛文灏, 等. 基于扩散模型的条件引导图像生成综述. 浙江大学学报 (理学版), 2023, 50(6): 651–667.

5 来杰, 王晓丹, 向前, 等. 自编码器及其应用综述. 通信学报, 2021, 42(9): 218–230.

6 Kingma DP, Dhariwal P. Glow: Generative flow with invertible 1×1 convolutions. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 10236–10245.

7 Hoogeboom E, Van Den Berg R, Welling M. Emerging convolutions for generative normalizing flows. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 2771–2780.

8 Xu YL, Liu ZM, Tegmark M, *et al*. Poisson flow generative models. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 16782–16795.

9 Tomczak J, Welling M. VAE with a VampPrior. Proceedings of the 21st International Conference on Artificial Intelligence and Statistics. Playa Blanca: PMLR, 2018. 1214–1223.

10 van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6309–6318.

11 Razavi A, van den Oord A, Vinyals O. Generating diverse high-fidelity images with VQ-VAE-2. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 1331.

12 Child R. Very deep VAEs generalize autoregressive models and can outperform them on images. arXiv:2011.10650, 2020.

13 Hazami L, Mama R, Thurairatnam R. Efficient-VDVAE: Less is more. arXiv:2203.13751, 2022.

14 Vahdat A, Kautz J. NVAE: A deep hierarchical variational autoencoder. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1650.

15 Aneja J, Schwing A, Kautz J, *et al*. NCP-VAE: Variational autoencoders with noise contrastive priors. arXiv:2010.02917, 2020.

16 Parmar G, Li DC, Lee K, *et al*. Dual contradistinctive generative autoencoder. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 823–832.

17 Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139–144. [doi: 10.1145/3422622]

18 Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434, 2015.

19 Karras T, Aila T, Laine S, *et al*. Progressive growing of GANs for improved quality, stability, and variation. arXiv:1710.10196, 2017.

20 Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. Proceedings of the 2019 IEEE/CVF Conference On Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4396–4405.

21 Karras T, Laine S, Aittala M, *et al*. Analyzing and improving the image quality of StyleGAN. Proceedings of the 2020

IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8107–8116.

22 Karras T, Aittala M, Laine S, *et al*. Alias-free generative adversarial networks. Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2021. 852–863.

23 Esser P, Rombach R, Ommer B. Taming Transformers for high-resolution image synthesis. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12868–12878.

24 Pidhorskyi S, Adjeroh DA, Doretto G. Adversarial latent autoencoders. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 14092–14101.

25 Gu JT, Liu LJ, Wang P, *et al*. StyleNeRF: A style-based 3D-aware generator for high-resolution image synthesis. arXiv:2110.08985, 2021.

26 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 574.

27 Nichol AQ, Dhariwal P. Improved denoising diffusion probabilistic models. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 8162–8171.

28 Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2021. 672.

29 Wang ZD, Zheng HJ, He PC, *et al*. Diffusion-GAN: Training GANs with diffusion. arXiv:2206.02262, 2022.

30 Song JM, Meng CL, Ermon S. Denoising diffusion implicit models. arXiv:2010.02502, 2020.

31 Ho J, Saharia C, Chan W, *et al*. Cascaded diffusion models for high fidelity image generation. The Journal of Machine Learning Research, 2022, 23(1): 47.

32 Johnson J, Alahi A, Li FF. Perceptual losses for real-time style transfer and super-resolution. Proceedings of the 14th European Conference on Computer Vision (ECCV 2016). Amsterdam: Springer, 2016. 694–711.

33 Karras T, Aittala M, Hellsten J, *et al*. Training generative adversarial networks with limited data. Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver: Morgan Kaufmann Publishers, 2020. 12104–12114.

34 Heusel M, Ramsauer H, Unterthiner T, *et al*. GANs trained by a two time-scale update rule converge to a local nash equilibrium. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6629–6640.

35 Xiao ZS, Kreis K, Kautz J, *et al*. VAEBM: A symbiosis between variational autoencoders and energy-based models. arXiv:2010.00654, 2020.

36 Sinha A, Song JM, Meng CL, *et al*. D2C: Diffusion-decoding models for few-shot conditional generation. Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2021. 12533–12548.

37 Pandey K, Mukherjee A, Rai P, *et al*. DiffuseVAE: Efficient, controllable and high-fidelity generation from low-dimensional latents. arXiv:2201.00308, 2022.

38 Or-El R, Luo X, Shan MY, *et al*. StyleSDF: High-resolution 3D-consistent image and geometry generation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 13493–13503.

(校对责编: 张重毅)