arXiv:2305.02541v2 [cs.CV] 3 Nov 2023

捕捉缺失细节：使用频率增强变分自编码器进行图像重建

Xinmiao Lin 罗切斯特理
工学院xl3439@rit.edu

Yikang Li
OPPO 美国研究院
yikang.li1@oppo.com

Jenhao Hsiao
OPPO 美国研究院
mark@oppo.com

Chiuman Ho
OPPO 美国研究院
chiuman@oppo.com

Yu Kong 密歇根州
立大学yukong@msu.edu

摘要

广受欢迎的向量量化变分自编码器模型通过学习离散码本来重建图像，但存在一个显著问题：随着压缩率提高，图像重建质量会迅速下降。一个主要原因是，较高的压缩率会导致高频谱上的视觉信号损失更多，而这些信号反映了像素空间中的细节。本文提出一种频率补充模块（*FCM*）架构，以捕获缺失的频率信息，从而提升重建质量。*FCM* 可轻松集成到向量量化变分自编码器结构中，我们将这一新模型称为***F**requancy **A**ugmented **VAE** (**FA-VAE**)*。此外，我们引入动态频谱损失（*DSL*），引导频率补偿模块动态平衡不同频率，以实现最佳重建效果。我们进一步将 *FA-VAE* 扩展到文本到图像合成任务，并提出一种交叉注意力自回归变换器（*CAT*），以从文本中获取更精确的语义属性。我们在多个基准数据集上针对不同压缩率开展了广泛的重建实验，结果表明，与最先进的方法相比，所提出的 *FA-VAE* 能够更忠实地还原细节。*CAT* 也展现出更优的生成质量和更好的图文语义对齐。代码可在 *https://github.com/oppo-us-research/FA-VAE* 获取。

1. 引言

向量量化变分自编码器模型 [6,11,25,29,39,46]通过学习潜在嵌入的离散码本来重建图像。它们凭借可扩展且用途广泛的码本广受欢迎，可广泛应用于许多视觉

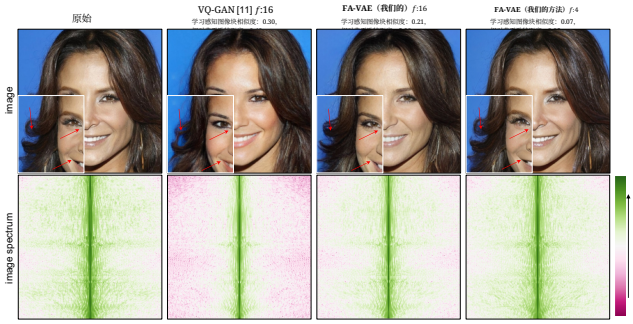


图 1. 图像及其频率图。第 1 行：原始图像和重建图像。第 2 行：图像的频率图，频率沿任意方向远离中心时都会增大。f 表示压缩率。压缩率越高，重建过程中丢失的细节就越多，例如眼睛和嘴巴的形状以及头发纹理（红色箭头所指），这与高频特征的损失相吻合。本文所有频率图均使用相同的色图。rFID [14]和 lpips [16] 的值越低越好；频率值从红色到绿色逐渐增大，放大图像可获得更好的视觉效果。

图像合成 [11,49] 和图像修复 [4,31]等任务。在 VQ-变分自编码器模型中，通常压缩率越高越好，因为它能提高内存效率，并更好地学习连贯的语义结构 [11,40]。

提高压缩率会很快带来一个主要挑战，即重建准确度会受到严重影响。图 1 第 1 行显示，尽管较高压缩率下的重建图像看起来与原始图像一致，但仔细观察便会发现细节存在差异，例如嘴唇的颜色和轮廓。图1 第 2 行显示，频域中也出现了类似的退化：压缩率越高，处于中频和高频频谱范围内的特征就越难以恢复。

造成像素空间与频率空间之间这一差距的原因有很多。自动编码器的卷积特性

使其容易产生谱偏置，倾向于学习低频特征 [22, 36]。当现有方法仅专注于设计损失函数或改进模型架构，以获得更好的语义相似性时，这一挑战会进一步加剧 [11,16,25]，但往往忽视了频率域上的对齐 [12,15]。此外，直观而言，解码器从单个码本嵌入（高压缩）重建图像块，比从多个嵌入（较低压缩）重建更具挑战性。原因在于，前者混合了不完整且多样的频率特征，而后者能够保留不同频率下更细粒度且完整的特征。

受这些发现的启发，我们提出了频率增强变分自编码器（***FA-VAE***）模型，旨在通过更好地对齐原始图像与重建图像的频谱来提升重建质量。具体而言，我们设计了名为频率补全模块（*FCM*）的新模块，并将其嵌入 FA-VAE 解码器的多个层中，使其能够学习为解码器的特征补充缺失频率。

我们观察到，在通过编码器进行压缩的过程中，有价值的中频和高频会与编码器的特征图交织在一起，如图3 第 4 行所示。因此，我们提出了一种名为频谱损失（*SL*）的新损失函数，用于引导频率补偿模块生成缺失特征，使其在频域上与同层编码器特征对齐。由于大多数图像语义都位于低频频谱中 [48], , SL 优先学习低频特征，并随着频率升高逐渐降低权重。

有趣的是，我们发现，尽管采用 SL 后重建效果有所提升，补充后的解码器特征中仍会出现棋盘格模式（图3 第 4 列）。我们推测，这是因为 SL 在对频率加权时，为低频频谱设定了一个确定的范围，却未考虑频率的重要性可能因层而异。因此，我们在 SL 的基础上设计了一种改进的损失函数——动态频谱损失（*DSL*），其中包含一个可学习组件，可动态调整低频频谱的范围，以实现最佳重建。与 SL 相比，DSL 能进一步提升重建质量，同时避免特征中出现不自然的棋盘格伪影（图3第 5 列）。

我们进一步将 FA-VAE 扩展到文本到图像生成任务，并提出了 交叉注意力自回归变换器（*CAT*）模型。我们首先观察到，仅使用一个或少数几个词元嵌入来表示较长文本是一种粗略的表示方式 [8, 27, 33]。因此，CAT 使用所有词元嵌入作为条件，以提供更精确的引导。此外，现有方法通常采用自注意力机制，并且仅在生成开始时嵌入文本条件 [11, 49]。这种机制在自回归生成过程中会产生问题，因为一个图像词元是生一次生成一个词元，因此文本条件与生成词元之间的联系会逐渐减弱。为了解决这一问题，CAT 嵌入了一种交叉注意力机制，使文本条件能够指导每一步的生成过程。

总而言之，我们的工作包括以下贡献：

- 我们提出了一种名为频率增强变分自编码器（*FA-VAE*）的新型架构，通过实现更精确的细节重建来提升图像重建效果。
- 我们提出了一种名为频谱损失（*SL*）的新型损失函数及其增强版本动态频谱损失（*D SL*），用于引导 FA-VAE 中的频率补偿模块（*FCM*）自适应地学习不同的高低频混合比例，以实现最佳重建效果。
- 我们提出了一种新型交叉注意力自回归变换器（*CAT*），用于文本到图像生成。该模型采用更细粒度的文本嵌入作为条件，并结合交叉注意力机制，以实现更好的图文语义对齐。

2. 相关工作与背景

图像重建向量量化变分自编码器（VQ-VAE） [46] 扩展了变分自编码器结构 [23]，并提出使用向量量化（VQ）将图像编码为离散潜在编码。随后，可以训练一个生成模型（例如自回归变换器 [13], ），并将其与 VQ-VAE 的解码器配合，以合成新图像。后续研究 [11,25,39]通过改进生成模型架构或量化效率，进一步提升了重建和生成质量。由于 VQ-VAE 模型在离散潜在空间中运行，因此无法与采用连续潜在空间的其他基于 VAE 的模型 [1, 6, 23, 29, 45, 51]直接进行公平比较。与现有 VQ-VAE 模型不同，我们提出的 FA-VAE 技术从频率角度提升重建效果，并且具有更强的通用性，因为 FCM 可以轻松扩展到采用相同 VQ-VAE 结构的其他神经网络。

很少有其他工作从频率视角进行图像重建。FFL [15] 提出了一种损失函数，用于惩罚高难频率上的差异。在 [12] 中，提出了一种为图像压缩而在连续潜在空间中设计的新模块，并将其整合到编码器和解码器中。据我们所知，FA-VAE 是首个旨在通过频率视角改善离散潜在空间上图像重建的工作。

图像生成 图像生成可以通过基于生成对抗网络的模型实现 [3, 18, 42, 44, 52, 55], ，这类模型能够无条件地，或以文本、掩码等不同输入为条件，从噪声向量合成图像。它

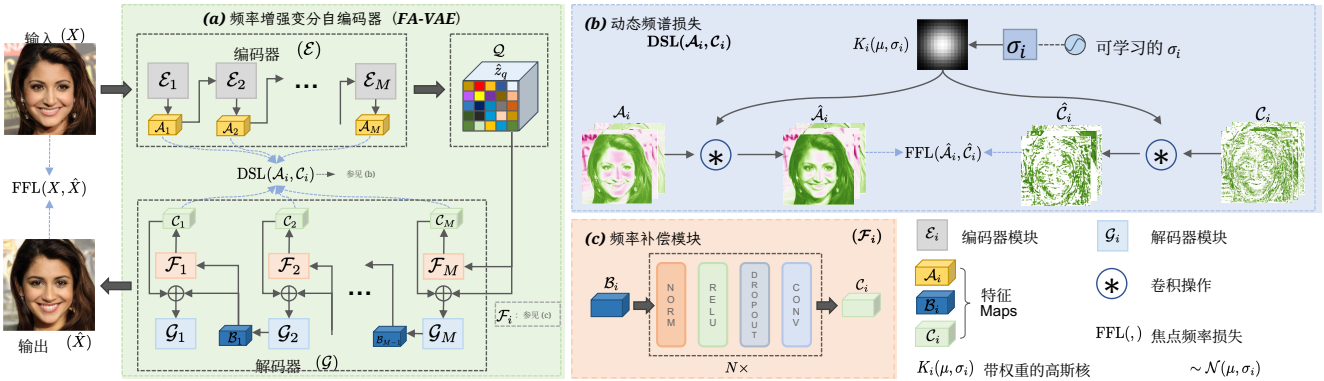


图2。频率增强*VAE*（***FA-VAE***）。(a) 编码器 $\mathcal{E}$ 将图像 X 编码到离散潜在码本空间 $\hat{z}_q$，解码器 $\mathcal{G}$ 利用该空间重建图像 $\hat{X}$。DSL (b) 引导频率补偿模块 (c) 学习利用重要频率的缺失特征补充重建特征，从而提高重建质量。更详细的图示见补充材料。

会使得为单一应用训练一个模型变得繁琐。因此，StyleGAN [21] 将语义属性编码到连续潜在空间中，后续研究 [19, 20,24, 32, 43, 50, 54] 则利用这一空间，根据属性或文本描述生成图像。然而，当属性数量大幅增加时，基于 StyleGAN 的模型无法扩展到大型数据集，因为这需要增大模型规模。相比之下，向量量化变分自编码器模型中的码本 [46]无需增加模型复杂度，便可扩展到大型数据集。基于扩散的模型 [5, 7, 10, 27, 28, 37, 40, 41] 通过反向扩散过程从高斯噪声生成图像，通常需要大量训练和大型数据集。大多数文生图模型通常使用预训练模型（如 CLIP [33] 或 T5-XXL [35]）的一种或少数几种嵌入作为条件。相比之下，我们提出的自回归变换器 CAT 使用所有文本词元的嵌入，在图像生成过程中提供更精细的引导。

向量量化变分自编码器 下面我们将介绍向量量化变分自编码器模型 [46]，因为它是 FA-VAE 模型所采用的骨干网络，该模型如图 2 所示。向量量化变分自编码器 [46] 及相关模 [11, 25, 38, 39] 型由一个编码器组成，该编码器将图像编码为嵌入码本。随后，训练一个解码器，从一组离散的码本嵌入中重建图像。本文采用 VQ-GAN 中的重建损失，其形式如下：

$$\mathcal{L}_{rec} = \|X - \hat{X}\|_1 + \mathcal{L}_{pips}(X - \hat{X}) \tag{1}$$

ˆ其中 X 和 X 分别是原始图像和重建图像。我们还采用了 VQ-GAN 模型的对抗设置；与 VQ-VAE 模型相比，该设置引入了判别器 [39, 46], 更多详情请参阅 [11]。

3. 方法

本文提出的频率增强变分自编码器 ***VAE (FA-VAE)*** 如图2所示。本节将介绍

FA-VAE 通过弥合重建图像与原始图像之间的频谱域差距，改善重建质量。通过将频率补全模块（*FCM*）明确嵌入解码器，动态频谱损失（*DSL*）利用编码器中频率维度上的重要特征，并引导频率补偿模块补充解码器的特征，从而在不同重建阶段实现更好的频率恢复。

3.1. 频率增强变分自编码器（FA-VAE）

设图像为 $X \in \mathbb{R}^{H\times W\times 3}$。码本 $\mathfrak{C}$ 是由 $|\mathfrak{C}|$ 嵌入组成的集合，例如 $\mathfrak{C} = \{\mathfrak{c}_i | i = 1, ..., |\mathfrak{C}|\} \in \mathbb{R}^{n_z}$，而 n_z 是一个码本嵌入的长度。FA-VAE 由编码器 $\mathcal{E}$ 组成，用于将图像 X编码为潜在空间表示，例如 $z = \mathcal{E}(x)$，其中$z \in \mathbb{R}^{h\times w\times n_z}$ 且 $x \in X$；$(h \times w)$ 是编码后表示的分辨率。设 $f = H/h = W/w$ 为下采样因子或压缩率。z 中的每个特征都通过向量量化模块 $\mathcal{Q}$ 使用最近的码本条目进行近似，表示如下：

$$\mathcal{Q}(z) = \hat{z}_q = \underset{c_k\in\mathcal{C}}{\arg\min}\|z_{ij} - \mathfrak{c}_k\| \tag{2}$$

其中，ˆz_q 是量化潜在嵌入，随后由解码器 $\mathcal{G}$ 用于生成重建图像 ˆ$x = \mathcal{G}(\hat{z}_q)$。

3.1.1 频率互补模块 (FCM)

动机 图 1 显示，更高的压缩率会导致更高频频谱上的重建差异更加显著。现有模型 [11, 25, 39]引导重建图像在像素空间和特征空间中与原始图像更加对齐（式 1），但忽略了频谱对齐，也没有为编码器和解码器提供进一步的指导。图 3 第 2 列显示，基线 VQ-GAN 的编码器激活 $\mathcal{A}_1$ [11] 包含丰富的高频特征（第 3 行

& 4)，但解码器的激活主要只能恢复低频特征（第 5 和第 6 行）。

因此，我们提出了频率补偿模块（*FCM*），如图 2 （c）所示，旨在利用编码器激活 $\mathcal{B}_i$ 中的特征补充解码器的特征图，使其包含缺失频率的特征 $\mathcal{A}_i$。FCM $\mathcal{F}_i$由卷积层和激活组成的序列构成。嵌入FCM的解码器 $\mathcal{G}$ 可表示为：

$$\begin{aligned}\mathcal{B}_{M-1} &= \mathcal{G}_M\left(\mathcal{C}_M + \hat{z}_q\right) = \mathcal{G}_M\left(\mathcal{F}_M(\hat{z}_q) + \hat{z}_q\right) \\ &\cdots \\ \mathcal{B}_1 &= \mathcal{G}_2(\mathcal{C}_2 + \mathcal{B}_2) = \mathcal{G}_2(\mathcal{F}_2(\mathcal{B}_2) + \mathcal{B}_2)\end{aligned} \tag{3}$$

同样，编码器 $\mathcal{E}$ 也可以抽象为：

$$z = \mathcal{E}(x) = \mathcal{E}_M\left(\cdots\left(\mathcal{E}_2\left(\mathcal{E}_1(x)\right)\right)\right) \tag{4}$$

其中，$\mathcal{G}_i$ 是解码器的第 i层，而 $\mathcal{A}_i = \mathcal{E}_i(x)$是相应编码器模块的输出或特征图。先前模块的输出 $\mathcal{B}_i$ 由输出 $\mathcal{C}_i = \mathcal{F}_i(\mathcal{B}_{i+1})$补充，这些输出包含从编码器激活 $\mathcal{A}_i$ 中学习到的丰富频率特征。下一节将介绍动态频率损失（DSL）如何引导频率补偿模块（FCMs）的学习，使其专门学习利用缺失频率的特征来补充 $\mathcal{B}_i$。

在论文中，频率特征补偿通过相加实现（见公式 3）。不过，FCM 可以灵活采用任意架构和融合技术，如图 5所示。需要注意的是，只要 $\mathcal{A}_i$ 和 $\mathcal{B}_i$ 具有相同的分辨率，编码器和解码器模块的架构就不受限制。VQ-GAN [11] 及相关模型 [25,46]已经采用了类似的架构，因此 FA-VAE 选择 VQ-GAN 作为骨干网络。

3.1.2 动态频谱损失（DSL）

动机 频谱损失（*SL*）旨在明确引导 FCM 的输出 $\mathcal{C}_i$，使其与编码器的激活更加一致 $\mathcal{A}_i$，因为频谱中包含丰富的高频特征（图3）。此外，为了考虑不同解码器阶段中频率重要性的差异，提出了动态频谱损失（*DSL*），这是 SL 的一种更通用的变体，能够自适应地调整高频频谱所占的权重。因此，每个解码器模块的输出都能融入最关键频率的特征，从而实现精准重建。

背景 编码器模块 $\mathcal{E}_i$ 和 FCM 模块 $\mathcal{F}_i$ 的输出 $\mathcal{A}_i$ $\mathcal{C}_i$ 首先通过离散傅里叶变换（DFT）转换到频域，如下所示：

$$F(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\cdot e^{-i2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \tag{5}$$

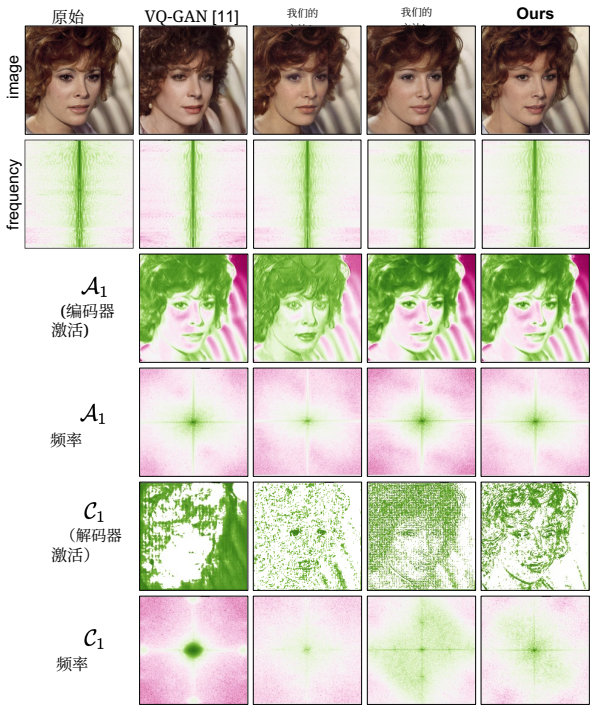


图3. 图像、激活及其频谱图。Ours* 是采用 FFL 的 FA-VAE 模型，[15], Ours‡ 是采用 SL 的 FA-VAE，Ours 则是采用 DSL 的 FA-VAE 模型。DSL 在低频与高频之间实现了更协调的平衡（第5行），并能更准确地重建嘴巴和头发纹理。频谱图取所有通道的平均值绘制，并调整了对比度，以突出高频频谱。

其中 e 和 i 分别为欧拉数和虚数单位。$M \times N$ 是特征图的空间分辨率，并对每个特征图应用傅里叶变换（公式 5）。$f(x,\ y)$ 是每个特征图在 $(x,\ y)$ 处的值。$F(u,\ v)$ 是频率谱中对应 $(u,\ v)$ 坐标处的值。焦点频率损失（FFL）[15]可表示为：

$$\mathrm{FFL}(\mathcal{A}_i, \mathcal{C}_i) = \frac{1}{MN|\mathcal{C}_i|}\sum_{c=0}^{|\mathcal{C}_i|-1}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1} w(u,v)J(u,v), \tag{6}$$

其中 $w(u,\ v) = |F_{\mathcal{A}_i}(u,\ v) - F_{\mathcal{C}_i}(u,\ v)|$ 是各个频率对应的权重。$J(u,\ v) = |F_{\mathcal{A}_i}(u,\ v) - F_{\mathcal{C}_i}(u,\ v)|^2$是基于频率差异的主要误差函数。$|\mathcal{C}_i|$ 是 $\mathcal{A}_i$ 和 $\mathcal{C}_i$ 中特征图的数量。频域的实部和虚部均纳入考虑，更多详情见 [15]。

FFL 的局限性 公式 6 表明，FFL 通过调制项 $w(u,v)$ 引导学习聚焦于最难重建的频率；根据我们的观察，这些是高频（图 1 &3）。然而，这可能并不理想，因为低频特征定义了图像内容，而过分强调高频可能会对学习施加过多约束

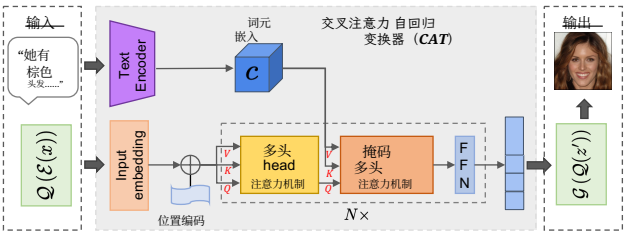


图 4. 交叉注意力自回归变换器（CAT）。

并导致重建效果不理想，见图7 第 1 行第 3 列和表2 第 3 行。此外，图3 第 3 列显示，由简单 FFL 引导的频率补偿模块可提升重建性能（我们的方法*）。然而，解码器激活的频谱图 $\mathcal{C}_1$ 包含过多噪声，这是由于过分强调较高频谱而忽略了较低频率（第 6 行）。

频谱损失（SL）因此，我们提出在权重上应用低通滤波器 $w(u,v)$，以惩罚低频域中更多的不匹配，并逐渐减小高频频谱对应的惩罚权重。因此，令高斯核为$K_i(\mu,\sigma)$，其权重使用均值和标准差初始化 (μ,σ_i)，并作用于特征图上：

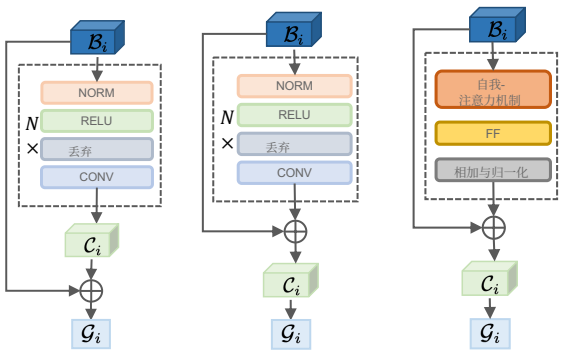
$$(\hat{\mathcal{A}}_i,\hat{\mathcal{C}}_i)=(K_i(\mu,\sigma_i)\star\mathcal{A}_i,K_i(\mu,\sigma_i)\star\mathcal{C}_i). \tag{7}$$

其中，$\star$ 表示卷积操作。然后，频谱损失（*SL*）定义为：

$$\mathrm{SL}(\mathcal{A}_i,\mathcal{C}_i)=\mathrm{FFL}(\hat{\mathcal{A}}_i,\hat{\mathcal{C}}_i). \tag{8}$$

SL 的局限性 到目前为止，所有 $K_i(\mu,\sigma_i)$ 以及所有编码器和解码器模块的 $\mathcal{A}_i$ 和 $\mathcal{C}_i$ 上均应用了固定的高斯滤波器。图3 第 4 列表明，SL（我们的方法‡）改善了低频频谱的重建效果，因此重建效果优于基线模型（表2 第 5 行与第 1 行）。然而，$\mathcal{C}_1$上也出现了棋盘格伪影（第 5 行）。原因之一是，确定性方差 σ_i 假设不同层级的解码器激活需要相同数量的高频特征才能准确重建，这可能会对学习施加过于僵硬的约束。另一个原因可能是，相同的 σ_i 会放大 CNN 中由上采样或步幅操作产生的棋盘格效应，而这些效应通常会被后续的卷积层规避 [2, 30]。通过实验观察，我们发现，由于低频与高频之间的界限效应，在不同层中使用相同的 σ_i 可能会加剧棋盘格效应（见图3 第 4 列第 6 行）。

动态频谱损失（DSL）因此，我们提出优化方差 σ_i，而不是将其设为静态超参数，以适应每个模块 $\mathcal{B}_i$ 所需的不同频率数量。新的动态频谱损失（*DSL*）是频谱损失的一种更通用形式。



(a) FCM (b) 带残差的 FCM (c) 带注意力机制的 FCM

F图 5. 采用不同架构和连接方式的 FCM 。

(SL)使用可学习的 σ_i。请注意，在原始图像和重建图像上使用且不采用高斯滤波器时，DSL 也包含 FFL 这一特殊形式。

因此，FA-VAE 的总重建损失可描述为：

$$\begin{aligned}\mathcal{L}_{rec}=\alpha\mathrm{FFL}(X,\hat{X})+\beta\sum_{i=0}^{M-1}\mathrm{DSL}(\mathcal{A}_i,\mathcal{C}_i)\\+\|X-\hat{X}\|_1+\mathcal{L}_{pips}(X-\hat{X}).\end{aligned} \tag{9}$$

其中，$\mathrm{FFL}(X-\hat{X})$ 是作用于原始重建图像的焦点频率损失。公式9 中的第二项是作用于编码器和 FCM 模块输出的 DSL 损失。前两项损失旨在最小化图像和内部特征图上的频谱差异。第三项和第四项损失作用于像素空间和特征图空间 [11]。α 和 β 是超参数。

σ_i 为模型参数，并按以下方式进行优化：

$$\sigma_i^*,\mathcal{E}^*,\mathcal{G}^*,\mathcal{C}^*=\arg\min_{\sigma_i,\mathcal{E},\mathcal{G},\mathcal{C}}(\mathcal{L}_{rec}+\mathcal{L}_Q) \tag{10}$$

$\mathcal{L}_Q$ 是量化损失，用于最小化码本嵌入与编码器 $\mathcal{E}$ 给出的嵌入之间的差异，详情见 [11]。我们还在量化过程中对码本嵌入使用 L_2 正则化，如 [53] 所示，并使用指数移动平均（EMA）更新码本 [46]，因为这些方法能让量化模块 $\mathcal{Q}$ 的训练更加稳定。

DSL 的优势 DSL 中可学习 σ_i 的优势如图3第 5 列所示，其中重建的激活没有出现第 4 列中的棋盘格伪影。此外，重建的激活 C_1 在频谱上与编码器激活 A_1 更加相似。与基线模型（第 2 列）相比，我们的模型在低频和高频之间展现出更协调的平衡，从而生成更准确的重建图像。

3.2. 交叉注意力自回归变换器（CAT）

我们进一步将 FA-VAE 扩展到文本到图像的图像生成任务，并引入一种新的 交叉注意力自回归

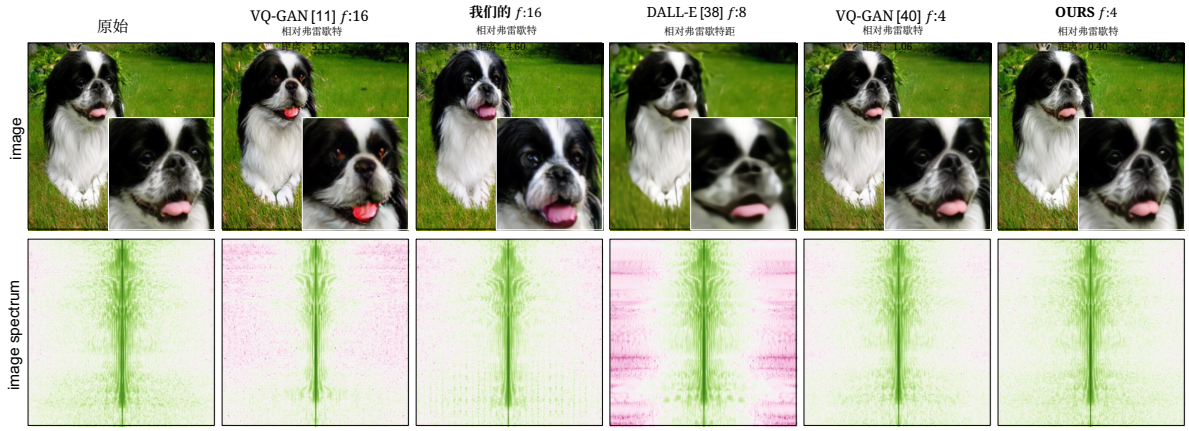


图 6. ImageNet 上的重建 [9], 标签：日本梗。

式变换器（*CAT*），如图 4所示。CAT 使用 CLIP 模型预训练文本编码器生成的文本描述中的所有词元嵌入 [33], ，而现有工作大多使用一个或部分文本嵌入 [8, 38, 49]。此外，CAT 使用一种结合文本词元的交叉注意力机制，在每一步引导图像生成。这种更细粒度的文本条件使生成结果能够更精确地捕捉文本与图像之间语义属性的关系。

$$p(s|c) = \prod_i p(s_i|s_{<i}, c) \tag{11}$$

然后，可以使用 FA-VAE 的解码器 G 将预测的索引 z' 解码为图像。损失函数旨在最大化数据表征的对数似然，

$$\mathcal{L}_{\text{CAT}} = \mathbb{E}_{x\sim p(x)}[-\log p(s)] \tag{12}$$

本文采用 GPT2 模型 [34], ，该模型是 [11, 49], 中的默认设置，并将其用作 CAT 的骨干网络；此外，还在 GPT2 结构中应用了受 [47] 启发的交叉注意力机制。

4. 实验

4.1. 实验细节

本文使用的数据集包括：(1) Multi-Modal CelebA-HQ 数据集 [50]，包含 30,000 张高分辨率名人脸部图像，每张图像都配有十条描述；(2) Flickr-Faces-HQ (FFHQ) 数据集 [21]，包含 70,000 张高分辨率人脸图像；(3) ImageNet [9]，包含来自 1000 个类别的约 100 万张图像。实验在 V-100 GPU 上进行，除非另有说明，结果均在验证集上报告。为与现有模型进行公平比较，所有数据集的训练分辨率均为 256×256。训练设置和超参数的详细信息见补充材料。

模型	数据集	码本大小	$(h \times w)$	相对弗雷歇特距离 ↓
RQ-VAE [25]	FFHQ	2048	(8×8)	5.33
FA-VAE（我们的方法）	FFHQ	2048	(16×16)	4.98
向量量化变分自编码器-2 [39]	ImageNet	512	(64×64) & (32×32)	~ 10（训练）
VQ-GAN[40]	ImageNet	8192	(64×64)	1.06
FA-VAE（我们的方法）	ImageNet	8192	(64×64)	**0.40**
DALL-E [38]	ImageNet	8192	(32×32)	32.01
VQ-GAN [11]	ImageNet	16384	(16×16)	5.15
VQ-GAN [11]	ImageNet	1024	(16×16)	7.94
VQ-GAN [25]	ImageNet	16384	(8×8)	17.95
RQ-VAE† [46]	ImageNet	16384	(8×8)	10.77
RQ-VAE* [25]	ImageNet	16384	(8×8)	4.73
FA-VAE（我们的方法）	ImageNet	16384	(16×16)	**4.60**

表 1. FFHQ 验证数据上的重建结果 [21] 和 ImageNet [9] 分别，更多结果见补充材料——材料。† 表示深度 2，$*$ 表示来自 RQ-VAE [25]的深度 4。

4.2. 图像重建

FFHQ 和 ImageNet 上的重建FFHQ 上重建的实验结果 [21] 和 ImageNet[9] 如表1所示。基线模型的结果取自原始论文。请注意，较小的 h 和 w意味着下采样因子或压缩率f 较大。在这两个数据集上，FA-VAE 模型在不同压缩率下均表现出优于基线模型的重建质量。主要原因在于，FA-VAE 能够成功重建基线模型所忽略的重要中频和高频信息。

图6 进一步支持了我们此前的论断，并清楚地展示了基线模型 VQ-GAN [11], DALL-E [38], 与我们的 FA-VAE 模型在重建质量方面存在明显差异。随着压缩率的提高，图 6 第 2 行显示，VQ-GAN 和 DALL-E 中更多的中频和高频信息被压缩。相比之下，FA-VAE 能够更准确地重建中频和高频信息，从而在图 6 第 1 行中呈现出语义和局部细节均得到改善的优质表征。例如，当压缩率为 16 时，狗的舌头

	消融实验	FFL(X, $\hat{X}$)	SL ($\mathcal{A}_i$, $\mathcal{C}_i$)	FCM $\mathcal{F}$	核 $K_i(\mu, \sigma_i)$	DSL ($\mathcal{A}_i$, $\mathcal{C}_i$)	核大小 μ	初始值 σ_i	成对 σ_i	$\mathcal{L}_1 \downarrow$	$\mathcal{L}_{pips} \downarrow$	相对弗雷歇特距离↓
1	VQ-GAN [11]	✗	✗	✗	✗	✗	✗	✗	✗	0.121	0.30	10.12
2	VQ-GAN + 风格 [53]	✗	✗	✗	✗	✗	✗	✗	✗	0.085	0.23	11.90
3	VQ-GAN [15]	✓	✗	✗	✗	✗	✗	✗	✗	0.114	0.35	30.65
4	不含核的 SL	✓	✗	**CONV**	✗	✗	✗	✗	✗	0.082	0.22	7.04
5		✓	✓	**CONV**	✗	✗	✗	✗	✗	0.082	0.22	7.02
6	含核的 SL	✓	✓	CONV	✓	✗	**9**	**3**	✗	0.080	0.22	7.39
7	FCM架构	✓	✓	**CONV**	✓	✓	9	3	✗	0.081	0.21	5.90
8		✓	✓	**RES**	✓	✓	9	3	✗	0.078	0.21	6.44
9		✓	✓	**ATTN**	✓	✓	9	3	✗	0.089	0.23	7.49
10	DSL 核大小 μ	✓	✓	RES	✓	✓	**3**	3	✓	0.081	0.21	6.66
11		✓	✓	RES	✓	✓	**5**	3	✓	0.081	0.21	7.04
12		✓	✓	RES	✓	✓	**9**	3	✓	0.082	0.22	6.53
13		✓	✓	RES	✓	✓	**11**	3	✓	0.083	0.22	7.48
14		✓	✓	RES	✓	✓	**15**	3	✓	0.083	0.22	6.36

表2. 在 CelebA-HQ 验证数据集上的消融研究 [17], 可视化结果见图 7。加粗的词语用于突出显示发生变化的组件。



图7. CelebA-HQ 上消融研究的重建效果对比。图中的数字表示表2中对应行所采用的设置。SL* 表示不使用高斯核的 SL，DSL** 表示采用非成对 σ的 DSL。默认情况下，FCM 采用卷积架构。

FA-VAE 的重建在颜色和形状方面比 VQ-GAN 更接近原始图像；更多定性结果请参见补充材料。

FCM 和 DSL 的消融研究在表 2中，我们针对 FCM 的不同架构，结合频谱损失（SL）和动态频谱损失（DSL）的不同设置进行了消融研究。图 7 展示了相应的定量结果可视化图。

首先，正如第 3 节所述，表2 第 3 行显示，将所有频率视为同等重要会导致 VQ-GAN 模型在 rFID 和学习感知图像块相似度方面表现不佳（见图7图像 3）。相比之下，表 2第 4 行结果表明，仅使用频率补偿模块就能有所帮助

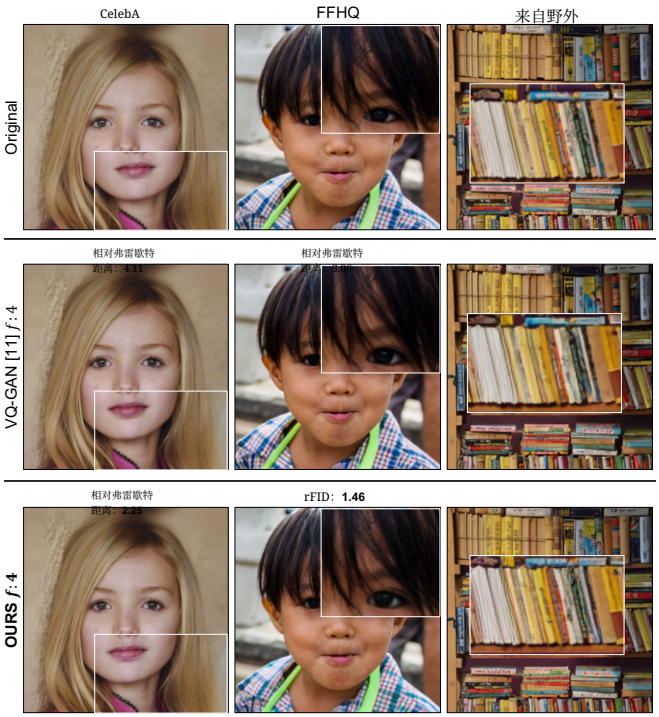


图8. CelebA-HQ 上的零样本重建 [17], FFHQ [21] 以及取自野外的照片，来源于 [11]。这些模型在 ImageNet 上以不同的压缩率进行训练。

FA-VAE 模型能够实现更好的重建，因为 FCMs 中的残差连接更好地保留了信息流。与简单的 SL 结合后，在图像层面（第 5 行）或模块层面（第 6 行）进行重建，可进一步提升效果。尽管由于忽略了解码器块间频率重要性的差异，SL（第 6 行）的重建效果略差，但图7 中的图像 5 表明，与图像 1 所示的基线模型相比，耳环和头发等细节仍能得到进一步增强。



图9. CelebA-HQ-MM 上的文本到图像生成 [17]

接下来，表2 的第 7–9 行比较了如图 5所示的 FCM 架构变化时的性能；请注意，我们没有对 σ_i 使用相同的 $(\mathcal{A}_i, \mathcal{C}_i)$，而是分别为 σ 和 $\mathcal{A}_i$ 使用两个可学习的 $\mathcal{C}_i$。在其他条件相同的情况下，从定量结果来看，卷积架构的重建效果优于残差架构和注意力架构。同样，图 7中的图像 7 在面部辐射度、平滑度、唇色和形状方面比图像 8 和 9 更接近原图。原因在于，残差连接增强了 FCM 的输出，这与第 3.1.1 节提出的、旨在增强解码器特征中高频特征的目的相悖。注意力机制在此处处于劣势，因为解码器和编码器都只使用卷积。

最后，在 DSL 中改变卷积核大小（表2 第 10–14 行）所得的重建定量性能相近；而从定性结果来看，图7 表明，较大的卷积核往往会生成更平滑的重建结果。一个可能的原因是，卷积时会使用更多周围的数值，因此较大的卷积核往往会使特征图更加平滑。因此，在其他实验中，我们选择大小为 3 的卷积核；卷积核大小的影响留待未来研究。

零样本重建为进一步展示 FA-VAE 模型的重建能力，表 3给出了使用在 ImageNet 上训练的模型，在 CelebA-HQ 上评估的零样本重建性能 [17] 和 FFHQ 上的结果 [21] 。 [9]相应的定性结果见图8。总体而言，FA-VAE 展现出出色的迁移能力，能够更忠实地重建细节，例如第 1 列中的光线对比度和第 3 列中的书本细节。另请注意，图像压缩中常用的指标峰值信噪比（PSNR）与相对弗雷歇特距离（rFID）并非完全相关，但我们在此列出该指标以供参考。

4.3. 图像合成

表 4 展示了在 CelebA-HQ-MM 上的文本到图像生成性能 [17]。我们提出的自回归变换器 CAT 比其他基于生成对抗网络的模型（包括 AttnGAN [52], Con-trolGAN [26], ）生成的图像质量更好，而这些模型仅针对图像生成而设计。图 9 显示，CAT 能够根据不同长度的文本输入生成令人满意的图像。其

		CelebA		FFHQ	
预训练模型	f	相对弗雷歇特距离 ↓	峰值信噪比 ↑	相对弗雷歇特距离 ↓	峰值信噪比 ↑
VQ-GAN [40]	16	8.62	**23.40**	6.83	**22.68**
VQ-GAN [11]	4	4.11	31.20	3.06	30.82
FA-VAE（我们的方法）	16	**6.52**	22.59	**6.19**	21.95
FA-VAE（我们的方法）	4	**2.25**	**31.39**	**1.46**	**30.85**

表3. 使用在 Im- ageNet [9]上训练的模型，在 FFHQ [21] 和 CelebA-HQ [17] 验证数据上的零样本重建结果。 f 为下采样因子，$f = \{16, 8, 4\}$ 的码本大小分别为 $\{16384, 16384, 8192\}$ 。

模型	FID ↓
AttnGAN [52]	125.98
ControlGAN [26]	116.32
DM-GAN [55]	131.05
DF-GAN [44]	137.60
TeDigan [50]	106.37
LAFITE [54]	12.54
CAT（我们的方法）	**10.23**

表4. 在 CelebA-HQ MM [50] 上进行文本到图像生成。

在生成过程中，模型能够捕捉“嘴巴张开”和“弓形眉”等属性，因为所有词元的嵌入都用作条件，从而提供了更精确的引导。补充材料中提供了不同数据集上的更多定量和定性结果。

5. 结论

本文提出了频率增强变分自编码器（FA-VAE）模型，旨在通过弥合原始图像与重建图像在频域上的差异来提升重建质量。我们设计了名为频率补全模块（FCM）的新模块，并在新的（动态）频谱损失（（D）SL）指导下，使其学习补全缺失频率对应的重建特征。我们还提出了一种新的交叉注意力自回归变换器（CAT），以在文本到图像生成任务中实现更精确的文本与图像对齐。与最先进的方法相比，FA-VAE 在多个数据集上实现了更好的重建效果。

致谢

本研究部分得到了国家科学基金会资助，项目编号为 2040209。

参考文献

[1] Jyoti Aneja、Alexander G. Schwing、Jan Kautz 和 Arash Vahdat。Ncp-vae：带噪声对比先验的变分自编码器。ArXiv，abs/2010.02917，2020年。2[2] Valerio Biscione 和 Jeffrey S. Bowers。卷积神经网络并不具备平移不变性，但可以学习获得这种不变性。JMLR，2022年。5[3] Andrew Brock、Jeff Donahue 和 Karen Simonyan。用于高保真自然图像合成的大规模生成对抗网络训练。载于 ICLR，2019年。2[4] Huiwen Chang、Han Zhang、Lu Jiang、Ce Liu 和 William T. Freeman。Maskgit：掩码生成式图像变换器。载于 CVPR，2022年。1[5] Ting Chen、Ruixiang Zhang 和 Geoffrey Hinton。模拟比特：使用带自条件的扩散模型生成离散数据，2022年。3[6] Rewon Child。非常深的 {变分自编码器}能够泛化自回归模型，并在图像任务上超越它们。载于 ICLR，2021年。1、2[7] Jooyoung Choi、Jungbeom Lee、Chaehun Shin、Sungwon Kim、Hyunwoo Kim 和 Sungroh Yoon。扩散模型的感知优先训练。载于 CVPR，2022年。3[8] Katherine Crowson、Stella Rose Biderman、Daniel Kornis、Dashiell Stander、Eric Hallahan、Louis Castricato 和 Edward Raff。Vqgan-clip：利用自然语言引导进行开放域图像生成与编辑。载于 ECCV，2022年。2、3、6[9] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大规模分层图像数据库。载于 CVPR，2009年。6、8[10] Prafulla Dhariwal 和 Alexander Nichol。扩散模型在图像合成方面胜过生成对抗网络。载于 NeurIPS，2021年。3[11] Patrick Esser、Robin Rombach 和 Bj¨orn Ommer。驯服变换器以实现高分辨率图像合成。载于 CVPR，2021年。1、2、3、4、5、6、7、8[12] Ge Gao、Pei You、Rong Pan、Shunyuan Han、Yuanyuan Zhang、Yuchao Dai 和 Hojae Lee。通过注意力多尺度反投影与频率分解实现神经图像压缩。载于 ICCV，2021年。2[13] Karol Gregor、Ivo Danihelka、Andriy Mnih、Charles Blundell 和 Daan Wierstra。深度自回归网络。载于 ICML，2014年。2[14]Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。采用双时间尺度更新规则训练的生成对抗网络收敛于局部纳什均衡。载于 NeurIPS，2017年。1[15] Liming Jiang、Bo Dai、Wayne Wu 和 Chen Change Loy。用于图像重建与合成的焦点频率损失。载于 ICCV，2021年。2、4、7[16] Justin Johnson、Alexandre Alahi 和 Li Fei-Fei。用于实时风格迁移与超分辨率的感知损失。载于 ECCV，2016年。1、2[17] Tero Karras、Timo Aila、Samuli Laine 和 Jaakko Lehtinen。通过 GAN渐进式增长提升质量、稳定性与多样性。计算研究储存库，abs/1710.10196，2017年。7、8[18] Tero Karras、Timo Aila、Samuli Laine 和 Jaakko Lehtinen。GAN渐进式增长：提升质量、稳定性和多样性。载于国际学习表征会议，2018年。2[19] Tero Karras、Miika Aittala、Janne Hellsten、Samuli Laine、Jaakko Lehtinen 和 Timo Aila。使用有限数据训练生成对抗网络。载于 NeurIPS，2020年。3[20] Tero Karras、Miika Aittala、Samuli Laine、Erik H¨ark¨onen、Janne Hellsten、Jaakko Lehtinen 和 Timo Aila。无混叠生成对抗网络。载于 NeurIPS，2021年。3[21] Tero Karras、Samuli Laine 和 Timo Aila。生成对抗网络的基于风格的生成器架构。载于 CVPR，2019年。3, 6, 7, 8[22] Soo Ye Kim、Kfir Aberman、Nori Kanazawa、Rahul Garg、Neal Wadhwa、Huiwen Chang、Nikhil Karnad、Munchurl Kim 和 Orly Liba。缩放到图像修复：具有高频细节的图像修复。载于 CVPRW，2022年。2[23] Diederik P Kingma 和 Max Welling。自编码变分贝叶斯，2013年。2[24] Umut Kocasarı、Alara Dirik、Mert Tiftikci 和 Pinar Ya- nardag。StyleMC：基于多通道的快速文本引导图像生成与操控。2022年 IEEE/CVF冬季计算机视觉应用会议（WACV），第3441–3450页，2022年。3[25] Doyup Lee、Chiheon Kim、Saehoon Kim、Minsu Cho 和 Wook-Shin Han。使用残差量化进行自回归图像生成。载于 CVPR，2022年。1, 2, 3, 4, 6[26] Bowen Li、Xiaojuan Qi、Thomas Lukasiewicz 和 Philip H. S. Torr。可控文本到图像生成。载于 NeurIPS，2019年。8[27] Alex Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob McGrew、Ilya Sutskever 和 Mark Chen。GLIDE：利用文本引导扩散模型实现逼真图像生成与编辑。载于 ICML，2022年。2, 3[28] Alexander Quinn Nichol 和 Prafulla Dhariwal。改进的去噪扩散概率模型。载于 ICML，2021年。3[29] Gaurav Parmar、Dacheng Li、Kwonjoon Lee 和 Zhuowen Tu。双重对比生成自编码器。载于 CVPR，2021年。1, 2[30] Yash Patel、Srikar Appalaraju 和 R. Manmatha。深度感知压缩，2019年。5[31] Jialun Peng、Dong Liu、Songcen Xu 和 Houqiang Li。使用层次化向量量化变分自编码器生成多样化结构以实现图像修复。载于 CVPR，2021年。1[32] Stanislav Pidhorskyi、Donald A. Adjeroh 和 Gianfranco Doretto。对抗性潜变量自编码器。载于 CVPR，2020年。3[33]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。通过自然语言监督学习可迁移视觉模型。载于 ICML，2021年。2, 3, 6[34] Alec Radford、Jeff Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。语言模型是无监督的多任务学习者。2019年。6

[35] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。通过统一的文本到文本变换器探索迁移学习的极限。JMLR，21(140):1–67，2020年。3[36] Nasim Rahaman、Aristide Baratin、Devansh Arpit、Felix Dr¨axler、Min Lin、Fred A. Hamprecht、Yoshua Bengio 和 Aaron C. Courville。论神经网络的谱偏置。载于 ICML，2019年。2[37] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。利用 CLIP 潜变量进行分层文本条件图像生成。ArXiv，abs/2204.06125，2022年。3[38] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。零样本文本到图像生成。载于 ICML，2021年。3, 6[39] Ali Razavi、Aaron van den Oord 和 Oriol Vinyals。利用 VQ-VAE-2 生成多样化高保真图像。载于 NeurIPS，2019年。1, 2, 3, 6[40]Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。利用潜在扩散模型进行高分辨率图像合成。载于 CVPR，2022年。1, 3, 6, 8[41]Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily L. Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、Seyedeh Sara Mahdavi、Raphael Gontijo Lopes、Tim Salimans、Jonathan Ho、David J. Fleet 和 Mohammad Norouzi。具有深度语言理解能力的照片级真实感文本到图像扩散模型。ArXiv，abs/2205.11487，2022年。3[42] Edgar Schonfeld、Bernt Schiele 和 Anna Khoreva。一种基于 U-Net 的生成对抗网络判别器。载于 CVPR，2020年。2[43] Jianxin Sun、Qiyao Deng、Qi Li、Muyi Sun、Min Ren 和 Zhenan Sun。AnyFace：自由风格的文本到人脸合成与操控。载于 CVPR，2022年。3[44] Ming Tao、Hao Tang、Fei Wu、Xiao-Yuan Jing、Bing-Kun Bao 和 Changsheng Xu。DF-GAN：一种简单有效的文本到图像合成基线方法。载于 CVPR，2022年。2, 8[45] Arash Vahdat 和 Jan Kautz。NVAE：一种深度分层变分自编码器。载于 NeurIPS，2020年。2[46] Aaron van den Oord、Oriol Vinyals 和 koray kavukcuoglu。神经离散表示学习。载于 NeurIPS，2017年。1, 2, 3, 4, 5, 6[47] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力机制即一切所需。载于 NeurIPS，2017年。6[48] Haohan Wang、Xindi Wu、Zeyi Huang 和 Eric P. Xing。高频分量有助于解释卷积神经网络的泛化能力。载于 CVPR，2020年。2[49] Zihao Wang、Wei Liu、Qian He、Xinglong Wu 和 Zili Yi。CLIP-GEN：利用 CLIP 对文本到图像生成器进行无语言训练。arXiv 预印本 arXiv:2203.00386，2022年。1, 2, 3, 6[50] Weihao Xia、Yujiu Yang、Jing-Hao Xue 和 Baoyuan Wu。TeDigan：文本引导的多样化人脸图像生成与操控。载于 CVPR，2021年。3, 6, 8

[51] Zhisheng Xiao、Karsten Kreis、Jan Kautz 和 Arash Vahdat。Vaebm：变分自编码器与基于能量的模型之间的共生关系。发表于国际学习表征会议，2021年。2[52] Tao Xu、Pengchuan Zhang、Qiuyuan Huang、Han Zhang、Zhe Gan、Xiaolei Huang 和 Xiaodong He。AttnGAN：利用注意力机制生成对抗网络进行细粒度文本到图像生成。发表于*CVPR*，2018年。2，8[53] Jiahui Yu、Xin Li、Jing Yu Koh、Han Zhang、Ruoming Pang、James Qin、Alexander Ku、Yuanzhong Xu、Jason Baldridge 和 Yonghui Wu。采用改进 VQGAN 的矢量量化图像建模。发表于国际学习表征会议，2022年。5，7[54] Yufan Zhou、Ruiyi Zhang、Changyou Chen、Chunyuan Li、Chris Tensmeyer、Tong Yu、Jiuxiang Gu、Jinhui Xu 和 Tong Sun。LAFITE：迈向用于文本到图像生成的无语言训练。*CVPR*，2022年。3，8[55] Minfeng Zhu、Pingbo Pan、Wei Chen 和 Yi Yang。DM-GAN：用于文本到图像合成的动态记忆生成对抗网络。发表于*CVPR*，2019年。2，8