生成多样化的高保真图像
使用 VQ-VAE-2

Ali Razavi
DeepMind
alirazavi@google.com

Aäron van den Oord
DeepMind
avdnoord@google.com

Oriol Vinyals
DeepMind
vinyals@google.com

摘 要

我们探索使用矢量量化变分自动编码器（VQ-VAE）模型进行大规模图像生成。为此，我们扩展和增强 VQ-VAE 中使用的自回归先验，以生成比之前更高的相干性和保真度的合成样本。我们使用简单的前馈编码器和解码器网络，使我们的模型成为编码和/或解码速度至关重要的应用的有吸引力的候选者。另外，VQ-VAE 要求仅在压缩的潜在空间中采样自回归模型，其比像素空间中的采样快一个数量级，尤其对于大图像。我们证明了 VQ-VAE 的多尺度分层组织，与潜编码相比具有强大的先验，能够生成质量可与多层数据集（如 ImageNet）上最先进的生成对抗网络相媲美的样本，而不会遭受 GAN 已知的缺点，如模式崩溃和缺乏多样性。

1 介绍

深度生成模型在过去几年中有显著改善[4,24,22]。 这在一定程度上要归功于架构创新以及计算方面的进步，这些都可以在数据量和模型大小方面对其进行更大规模的训练。 这些模型生成的样本很难在没有仔细检查的情况下与实际数据区分开来，它们的应用范围从超分辨率[20]到域编辑[40]，艺术操作[32]，或文本到语音和音乐生成[22]。

图 1：来自在 ImageNet 上训练的两级模型的类条件 256x256 图像样本。

我们区分了两种主要类型的生成模型：基于似然度的模型，包括 VAE [15,28]，基于流的[8,27,9,16]以及自回归模型[19,35]和隐性生成模型

* 同等的贡献。

例如 Generative Adversarial Networks（GANs）[11]。这些模型中的每一个都提供了一些权衡，例如样本质量，多样性，速度等。

GAN 利用生成器神经网络优化 minimax 目标，通过将随机噪声映射到图像上来生成图像，并通过将其样本分类为真实或假造的来定义生成器损失函数的判别器。更大规模的 GAN 模型现在可以生成高质量和高分辨率的图像[4,13]。然而，众所周知，来自这些模型的样本不能完全捕获真实分布的多样性。此外，GAN 在评估方面具有挑战性，并且尚未存在用于评估过度拟合的测试集的令人满意的泛化度量。对于模型比较和选择，研究人员使用图像样本或图像质量的代理度量，如初始分数（IS）[30]和 Fréchet 初始距离（FID）[12]。

相反，基于似然度的方法优化训练数据的负对数似然（NLL）。该目标允许模型比较和测量甚至包括看不见的数据。此外，由于模型为训练集中的所有样本分配的概率最大化，因此基于似然度的模型原则上涵盖数据的所有模式，并且不会遇到模式崩溃和 GAN 中缺乏多样性的问题 。尽管有这些优点，但直接最大化像素空间中的似然度可能具有挑战性。首先，像素空间中的 NLL 并不总是很好地衡量样本质量[33]，并且不能可靠地用于在不同的模型类之间进行比较。例如，这些模型没有内在的动力来关注全局结构。通过引入诸如多尺度[34,35,26,21]的归纳偏差或通过对图像中的显性位平面建模来减轻这些问题中的一些[17,16]。

在本文中，我们使用有损压缩的思想来减轻生成模型对可忽略信息的建模。实际上，诸如 JPEG [39]之类的技术已经表明，通常可以在不显著改变感知图像质量的情况下移除超过 80％的数据。如[37]所提出的，我们通过矢量量化自动编码器的中间表示将图像压缩成离散的潜在空间。这些表示比原始图像小 30 倍，但仍允许解码器以很小的失真重建图像。这些离散表示的先验可以用具有自注意力[38]的现有技术 PixelCNN [35,36]建模，称为 PixelSnail [6]。当从该先验中采样时，解码图像也表现出相同的高质量和重建的一致性（参见图 1）。此外，该生成模型在离散潜在空间上的训练和采样也比直接应用于像素时快 30 倍，使我们能够在更高分辨率的图像上进行训练。最后，在这项工作中使用的编码器和解码器保留了原始 VQ-VAE 的简单性和速度，这意味着所提出的方法对于需要快速，低开销编码和解码大图像的情况是有吸引力的解决方案。

2 背景

2.1 矢量量化变分自动编码器

VQ-VAE 模型[37]可以更好地理解为通信系统。它包括将观测值映射到一系列离散潜变量的编码器，以及从这些离散变量重建观测值的解码器。 编码器和解码器都使用共享码本（codebook）。更正式地，编码器是从输入空间 x 到向量 E(x)的非线性映射。然后基于其与码本$e_k, k \in 1 \dots K$中的原型矢量的距离来量化该矢量，使得每个向量 E(x)被码本中最近的原型向量的索引替换，并且被发送到解码器（注意该过程可能是有损的）。

$$\text{Quantize}(E(\mathbf{x})) = \mathbf{e}_k \quad \text{where } k = \arg\min_j ||E(\mathbf{x}) - \mathbf{e}_j|| \tag{1}$$

解码器将接收到的索引映射回码本中的相应矢量，从中通过另一个非线性函数重建数据。 为了学习这些映射，重构误差的梯度然后通过解码器反向传播，并使用直通梯度估计器反馈到编码器。

VQ-VAE 模型在其目标函数中包含两个附加项，以将码本的向量空间与编码器的输出对齐。第一个是仅适用于码本变量的码本损失，其使所选择的码本 e 接近编码器的输出 E(x)。 第二个是承诺

损失（commitment loss）仅适用于编码器权重，它鼓励编码器的输出保持接近所选择的码本矢量，以防止它从一个码矢量到另一个码矢量过于频繁地波动。 总体目标在等式 2 中描述，其中 e 是训练样本 x 的量化代码，E 是编码器函数，D 是解码器函数。 运算符 sg 指的是阻止梯度流入其自变量的停止梯度运算，并且β是超参数，其控制“阻力”以改变对应于编码器输出的编码。

$$\mathcal{L}(\mathbf{x}, D(\mathbf{e})) = ||\mathbf{x} - D(\mathbf{e})||_2^2 + ||sg[E(\mathbf{x})] - \mathbf{e}||_2^2 + \beta||sg[\mathbf{e}] - E(\mathbf{x})||_2^2 \quad (2)$$

正如[37]中提出的，我们使用码本的指数移动平均更新，作为码本损失的替代（等式 2 中的第二个损失项）：

$$N_i^{(t)} := N_i^{(t-1)} * \gamma + n_i^{(t)}(1-\gamma), \quad m_i^{(t)} := m_i^{(t-1)} * \gamma + \sum_j^{n_i^{(t)}} E(x)_{i,j}^{(t)}(1-\gamma), \quad e_i^{(t)} := \frac{m_i^{(t)}}{N_i^{(t)}}$$

其中$n_i^{(t)}$是将被量化为码本项e_i的小批量中的 E(x)中的向量数，并且是具有 0 到 1 之间的值的衰减参数 γ。我们使用默认值 γ = 0.99 在我们所有的实验中。 我们在 Sonnet 库（见底部 2、3）中使用已发布的 VQ-VAE 实现。

2.2 PixelCNN 系列自回归模型

深度自回归模型是常见的概率模型，可以在几种数据模态下实现最先进的密度估计结果[24,6,23,22]。 这些模型背后的主要思想是利用概率链规则将输入空间上的联合概率分布分解为数据的每个维度的条件分布的乘积，给定所有先前维度的某些预定义顺序：$p_\theta(\mathbf{x}) = \prod_{i=0}^{n} p_\theta(x_i|\mathbf{x}_{<i})$。每个条件概率由深度神经网络参数化，深度神经网络的结构根据数据所需的归纳偏差来选择。

3 方法

所提出的方法遵循两阶段方法：首先，我们训练分层 VQ-VAE（参见图 2a）以将图像编码到离散的潜在空间上，然后我们在由包含所有数据的离散潜在空间上拟合强大的先验 PixelCNN。

Algorithm 1 VQ-VAE training (stage 1)

Require: Functions E_{top} , E_{bottom}, D, $\mathbf{x}$ (batch of training images)

1: $\mathbf{h}_{top} \leftarrow E_{top}(\mathbf{x})$

▷ quantize with top codebook eq 1

2: $\mathbf{e}_{top} \leftarrow Quantize(\mathbf{h}_{top})$

3: $\mathbf{h}_{bottom} \leftarrow E_{bottom}(\mathbf{x}, \mathbf{e}_{top})$

▷ quantize with bottom codebook eq 1

4: $\mathbf{e}_{bottom} \leftarrow Quantize(\mathbf{h}_{bottom})$

5: $\hat{\mathbf{x}} \leftarrow D(\mathbf{e}_{top}, \mathbf{e}_{bottom})$

▷ Loss according to eq 2

6: $\theta \leftarrow Update(\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}))$

Algorithm 2 Prior training (stage 2)

1: $\mathbf{T}_{top}, \mathbf{T}_{bottom} \leftarrow \emptyset$ ▷ training set

2: **for** $\mathbf{x} \in$ training set **do**

3: $\mathbf{e}_{top} \leftarrow Quantize(E_{top}(\mathbf{x}))$

4: $\mathbf{e}_{bottom} \leftarrow Quantize(E_{bottom}(\mathbf{x}, \mathbf{e}_{top}))$

5: $\mathbf{T}_{top} \leftarrow \mathbf{T}_{top} \cup \mathbf{e}_{top}$

6: $\mathbf{T}_{bottom} \leftarrow \mathbf{T}_{bottom} \cup \mathbf{e}_{bottom}$

7: **end for**

8: $p_{top} = \texttt{TrainPixelCNN}(\mathbf{T}_{top})$

9: $p_{bottom} = \texttt{TrainCondPixelCNN}(\mathbf{T}_{bottom}, \mathbf{T}_{top})$

▷ Sampling procedure

10: **while** true **do**

11: $\mathbf{e}_{top} \sim p_{top}$

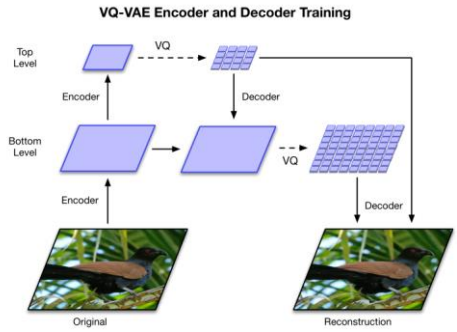
12: $\mathbf{e}_{bottom} \sim p_{bottom}(\mathbf{e}_{top})$

13: $\mathbf{x} \leftarrow D(\mathbf{e}_{top}, \mathbf{e}_{bottom})$

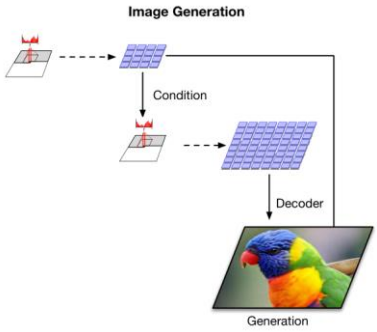
14: **end while**

[2] https://github.com/deepmind/sonnet/blob/master/sonnet/python/modules/nets/vqvae.py

[3] https://github.com/deepmind/sonnet/blob/master/sonnet/examples/vqvae_example.ipynb



(a) 我们的分层 VQ-VAE 架构概述。 编码器和解码器由深度神经网络组成。 模型的输入是 256×256 图像，其被压缩为分别为底部和顶部水平的尺寸为 64×64 和 32×32 的量化潜在编码图。 解码器从两个潜在编码图重建图像。



（b）多阶段图像生成。 顶级 PixelCNN 先验以类标签为条件，底层 PixelCNN 以类标签以及第一级编码为条件。 由于前馈解码器，潜码到像素之间的映射很快。（使用此模型生成带有鹦鹉的示例图像）。

图 2: VQ-VAE 架构

图 3：具有三个潜在编码图（顶部，中部，底部）的分层 VQ-VAE 的重建。 最右边的图像是原始图像。 每个潜在编码图都会为重建添加额外的细节。 这些潜在编码图分别比原始图像（分别）小约 3072x，768x，192x 倍。

3.1 第 1 阶段：学习分层潜在编码

与原始 VQ-VAE 相反，在这项工作中，我们使用矢量量化编码的层次结构来模拟大图像。这背后的主要动机是将局部信息（例如纹理）与诸如对象的形状和几何之类的全局信息分开建模。因此，可以调整每个级别上的先前模型以捕获该级别中存在的特定相关性。我们的多尺度分层编码器的结构如图 2a 所示，顶部潜在编码模拟全局信息，底部潜在编码以顶部潜在编码为条件，负责表示局部细节（见图 3）。我们注意到如果我们没有以顶部潜在编码为底部潜在编码的条件，那么顶部潜空间需要对像素中的每个细节进行编码。因此，我们允许层次结构中的每个级别分别依赖于像素，这鼓励在每个潜在编码图中编码补充信息，这有助于减少解码器中的重建误差。有关详细信息，请参阅算法 1。

对于 256×256 图像，我们使用两级潜在层次结构。如图 2a 所示，编码器网络首先将图像变换和下采样 4 倍到 64×64 表示，其被量化为我们的底层潜在图。然后，另一堆残余块进一步按比例缩小表示两倍，在量化之后产生顶层 32×32 潜在图。解码器类似地是前馈网络，其将量化的潜在层级的所有级别作为输入。它由一些残余块组成，然后是一些跨步的转置卷积，以将编码上采样回原始图像大小。

3.2 第二阶段：通过潜在编码学习推广

为了进一步压缩图像，并且能够从阶段 1 中学习的模型中进行采样，我们先了解潜在的编码。使用来自训练数据的神经网络拟合先前分布已成为常见做法，因为它可以显著改善潜变量模型的性能[5]。该程序还减少了后边缘和先前边界之间的差距。因此，在测试时从学习的先验中采样的潜在变量接近解码器网络在训练期间观察到的导致更多相干输出的变量。从信息理论的角度来看，在学习后验之前拟合的过程可以被认为是对潜在空间的无损压缩，通过对潜在变量进行重新编码，其分布更接近于它们的真实分布，从而导致比特率更接近香农的熵。因此，学习先验的真熵与负对数似然之间的差距越小，可以从解码潜在样本中得到更真实的图像样本。

在 VQ-VAE 框架中，这个辅助先验在后来的第二阶段用强大的自回归神经网络建模，例如 PixelCNN。顶部潜编码图的先验负责全局结构信息。因此，我们在[6,23]中配备了多头自我关注层，因此它可以从更大的感受野中获益，以捕捉图像中相距很远的空间位置的相关性。相反，用于编码本地信息的潜在底层的条件先验模型将以更大的分辨率操作。由于存储器限制，在顶级先验中使用自注意层是不实际的。对于这个先验的本地信息，我们因此发现使用大型调节堆栈（来自顶部先验）产生良好的性能（参见图 2b）。分层分解也允许我们训练更大的模型：我们分别训练每个先验的模型，从而利用硬件加速器上的所有可用计算和内存。有关详细信息，请参阅算法 3。

我们的顶级先验网络模型有 32×32 个潜在变量。 PixelCNN 的残余门控卷积层每五层散布着因果多头注意力（causal multi-headed attention）。为了使模型正则化，我们在每个残余块之后合并损失以及在每个注意力矩阵的 logits 上的损失。我们发现在 PixelCNN 堆栈顶部添加由 1×1 卷积组成的深度残余网络可以进一步提高似然度，而不会减慢训练速度或增加内存占用。我们的底层条件先验在潜在的 64×64 空间维度上运行。就所需的存储器和计算成本而言，这是非常昂贵的。如前所述，在层次结构的这个级别编码的信息大多对应于局部特征，这些特征不需要大的感知域，因为它们以顶层先验为条件。因此，我们使用功能较弱的网络而没有注意力层。我们还发现使用深度残留调节堆栈显著有助于此级别。

3.3 基于拒绝抽样的分类器的交易差异

与 GAN 不同，用最大似然目标训练的概率模型被迫模拟所有训练数据分布。这是因为 MLE 目标可以表示为数据和模型分布之间的前向 KL-散度，如果训练数据中的示例被指定为零权重，则将其驱动到无穷大。虽然数据分布中所有模式的覆盖范围是这些模型的吸引人的特性，但是任务比对抗建模要困难得多，因为基于似然度的模型需要适合数据中存在的所有模式。此外，来自自回归模型的祖先采样实际上可以引起可能在长序列上累积的错误并导致质量降低的样本。最近的 GAN 框架[4,1]提出了选择样本的自动程序，以权衡多样性和质量。在这项工作中，我们还提出了一种自动化方法，用于根据我们的样本越接近真实数据流形的直觉来交换样本的多样性和质量，通过预先训练将它们分类到正确的类别标签的可能性越大。具体来说，我们使用在 ImageNet 上训练的分类器网络，根据分类器分配给正确类的概率，从我们的模型中对样本进行评分。

4 相关工作

我们工作的基础是[37]的 VQ-VAE 框架。 我们之前的网络基于 Gated PixelCNN [36]，增强了自注意机制[38]，如[6]中所提出的

BigGAN [4]目前在 FID 和 Inception 评分方面是最先进的，可以生成高质量的高分辨率图像。BigGAN 的改进主要来自于结合架构设计上的进步，例如自我关注，更好的稳定方法，在 TPU 上扩展模型以及将样本多样性与样本质量进行权衡的机制。在我们的工作中，我们还研究了如何添加一些这样的元素，特别是自我关注和计算规模，确实也提高了 VQ-VAE 模型样本的质量。

最近的工作也被提出用于产生高分辨率图像，其中基于似然度的模型包括[21]的 Subscale Pixel Networks。类似于[26]中引入的并行多尺度模型，SPN 对空间维度进行了划分，但与[26]不同，SPN 没有做出相应的独立性假设，因此它通过密度估计性能和样本质量来交换采样速度。 。

已经在例如提出了分层潜在变量[28]。特别是对于 VQ-VAE，[7]使用潜在编码的层次结构来使用 WaveNet 解码器进行建模和生成音乐。然而，编码的细节与我们的不同之处在于：在我们的工作中，层次结构的底层并不专门细化顶级编码的信息，但是它们在每个级别提取补充信息，如 Sect. 3.1。此外，由于我们使用简单的前馈解码器并优化像素中的均方误差，因此我们的模型不会受到[7]中详述的层次崩溃问题的影响，因而无需缓解。与我们的工作同时，[10]扩展[7]以生成高分辨率图像。我们工作的主要区别是在像素空间中使用自回归解码器。相反，由于 Sect.3 中详述的原因，我们在压缩潜在空间中仅使用自回归模型作为先验，这简化了模型并大大提高了采样速度。另外，我们的方法与[10]之间也存在与上述[7]相同的差异。

先前已经针对 GAN [1]以及 VAE [3]探索了通过拒绝抽样来提高样本质量，其将学习的拒绝抽样方法与先验结合起来以减少与后验聚合的差距。

5 实验

客观评估和比较生成模型，特别是模型家族，仍然是一个挑战[33]。 当前的图像生成模型权衡样本质量和多样性（或精度与召回率[29]）。 在本节中，我们提供了在 ImageNet 256×256 上训练的模型的定量和定性结果。样本质量确实很高且很尖锐，跨越几个代表性类别，如图 5 中提供的类条件样本所示。我们提供的样本来自我们的模型与 BigGAN-deep [4]，图 5 中最先进的 GAN 模型的样本（见底部 4）。从这些并排比较中可以看出，VQ-VAE 能够提供具有可靠保真度和更高多样性的样本。

5.1 建模高分辨率人脸图像

为了进一步评估我们的多尺度方法捕获数据中极长距离依赖的有效性，我们在 1024×1024 分辨率下在 FFHQ 数据集[14]上训练三级分层模型。该数据集由 70000 张高品质人物肖像组成，在性别，肤色，年龄，姿势和服装方面具有相当大的多样性。尽管与 ImageNet 相比，建模面部通常被认为不那么困难，但是在如此高的分辨率下，还存在独特的建模挑战，可以以有趣的方式探测生成模型。例如，面部中存在的对称性需要能够捕获长距离依赖性的模型：具有受限的感受野的模型可以分别为每只眼睛选择合理的颜色，但是可能错过距离几百个像素的两只眼睛之间的强相关性，进而相互之间，产生具有不匹配的眼睛颜色的人脸肖像。

[4] 样本取自 TensorFlow 中心的 BigGAN colab 笔记本: https://tfhub.dev/deepmind/biggan-deep-256/1

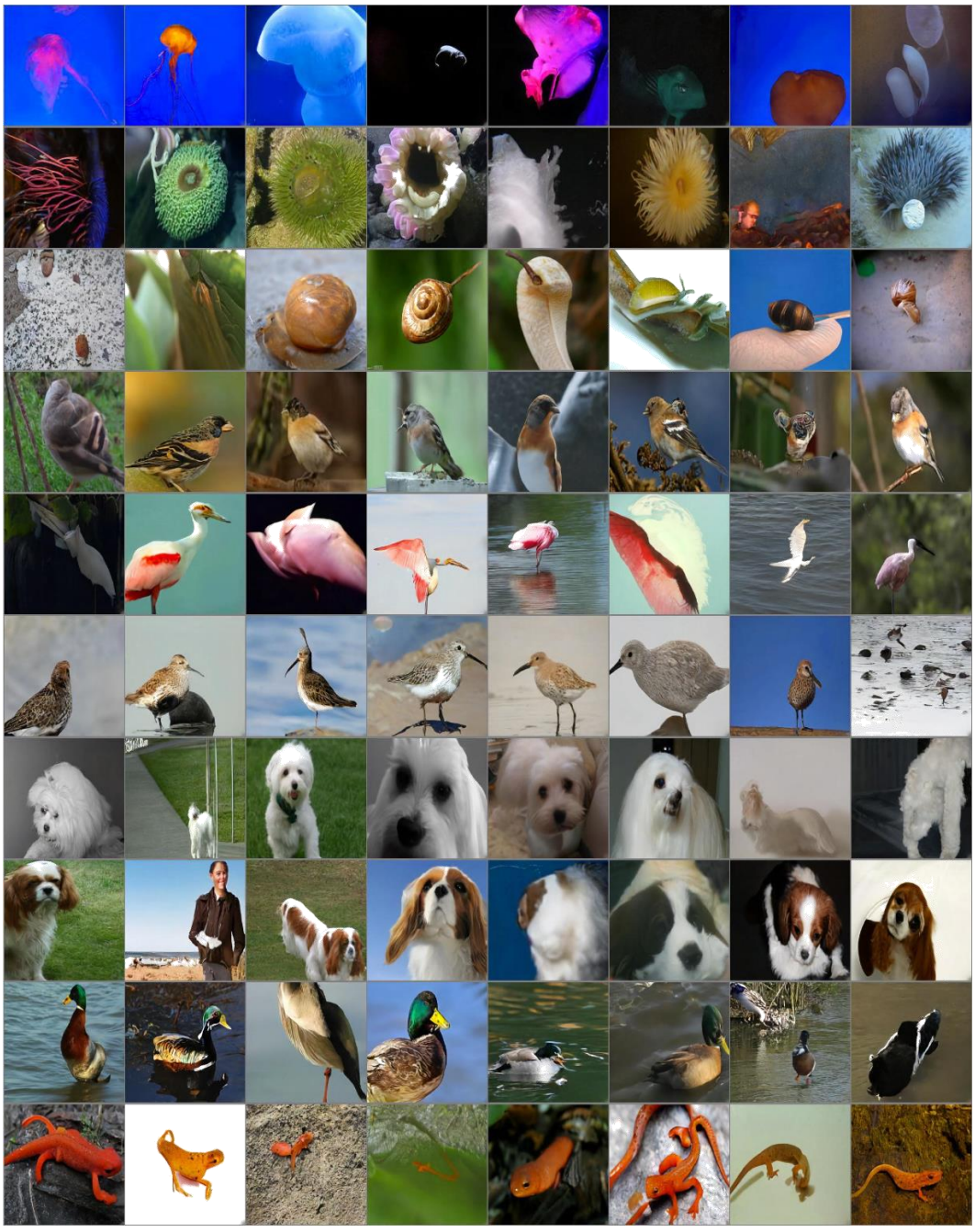
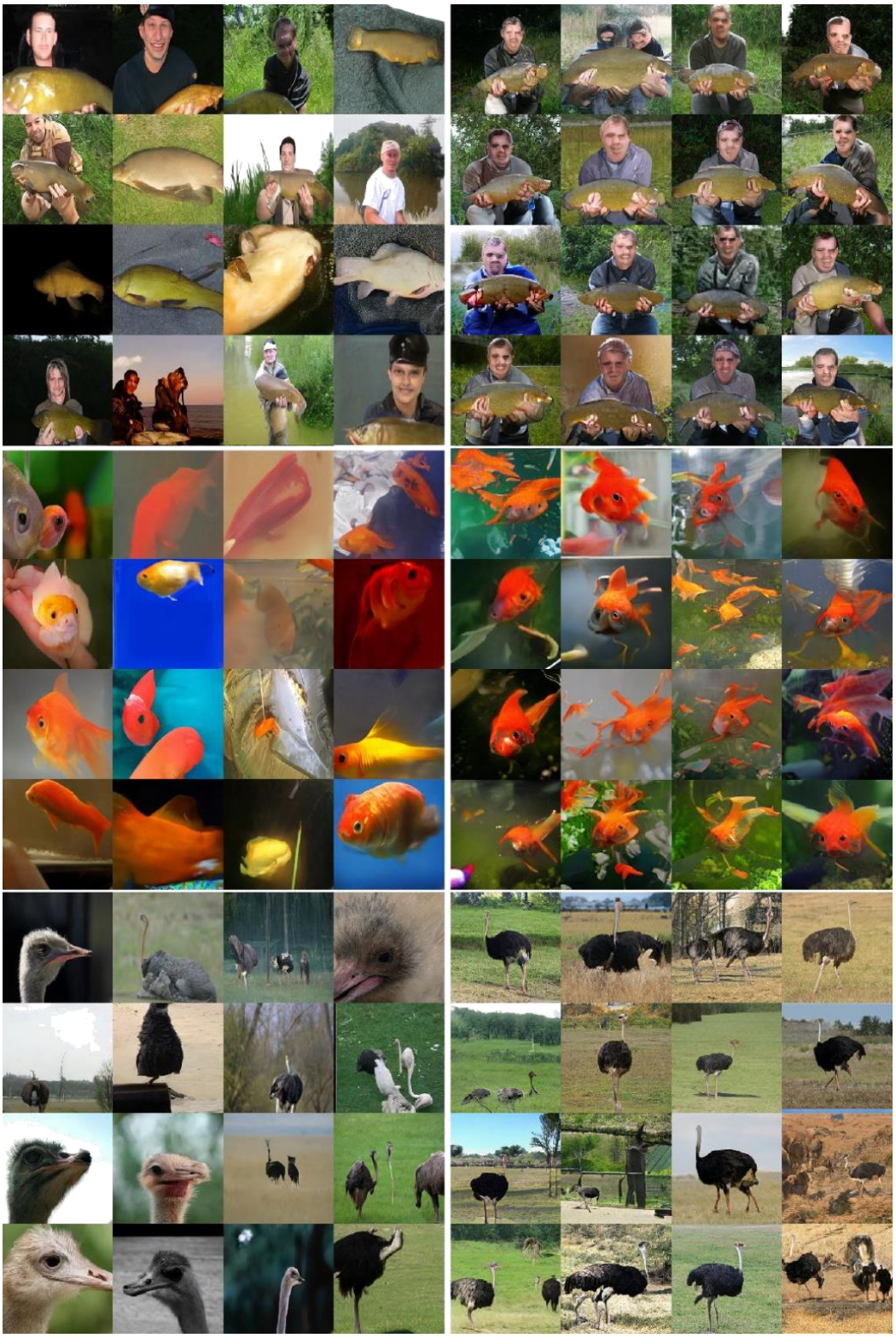


图 4：类条件随机样本。 排在第一行的是：108 海葵，109 脑珊瑚，114 蛞蝓，11 金翅雀，130 火烈鸟，141 红脚鹬，154 哈巴狗，157 蝴蝶犬，97 龙，28 斑点蝾螈。



VQ-VAE (提出) BigGAN deep

图 5：提议方法和 BigGan Deep for Tinca-Tinca（第 1 个 ImageNet 类）和 Ostrich（第 10 个 ImageNet 类）的样本多样性比较。 以截断水平 1.0 拍摄 BigGAN 样品，以产生其最大多样性。 有几种样本，例如鱼的顶视图或不同种类的姿势，例如 BigGAN 样本中没有的近距离鸵鸟。 请放大 pdf 版本以获取更多详细信息，并参阅补充材料以了解更多类别的多样性比较。

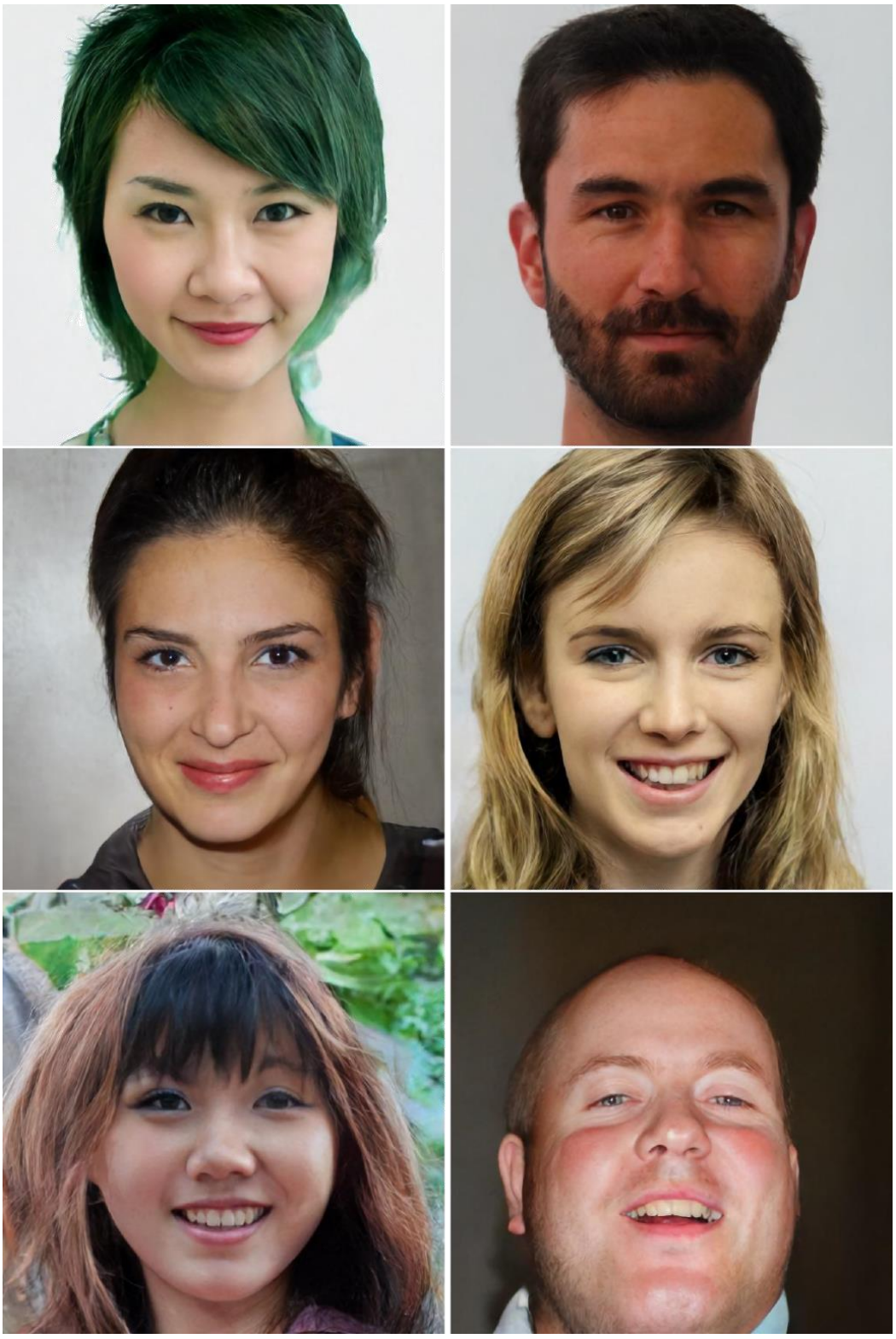


图 6：来自在 FFHQ-1024×1024 上训练的三级分层模型的代表性样本。该模型生成逼真的面部，其重视远距离依赖性，例如匹配眼睛颜色或对称面部特征，同时覆盖数据集的较低密度模式（例如，绿头发）。 有关更多样品，请参阅补充材料，包括全分辨率样品。

5.2 定量评估

在本节中，我们基于旨在测量样品质量和多样性的若干指标报告我们的定量评估结果。

5.2.1 负对数似然和重构误差

使用基于似然度的生成模型的主要动机之一是测试集和训练集上的负对数似然（NLL）给出了一般化的客观测量，并允许我们监测过度拟合。我们强调其他常用的性能指标，如 FID 和 Inception Score，完全忽略了泛化问题；简单记忆训练数据的模型可以获得这些指标的完美分数。同样的问题也适用于最近提出的一些指标，如 Precision-Recall [29,18]和分类准确度分数[25]。这些基于样本的指标仅提供样本质量和多样性的代理，但忽略了对保持图像的推广。

图 1 中报告的顶部和底部先验的 NLL 值接近于训练和验证水平，表明这些网络都不适合。我们注意到这些 NLL 值仅在使用相同预训练 VQ-VAE 编码器和解码器的先验模型之间相当。

	Train NLL	Validation NLL	Train MSE	Validation MSE
Top prior	3.40	3.41	-	-
Bottom prior	3.45	3.45	-	-
VQ Decoder	-	-	0.0047	0.0050

表 1：通过编码训练和验证集分别测量的顶部和底部先验的训练和验证负对数似然（NLL），以及训练和验证集的均方误差。 NLL 和 MSE 的微小差异表明既不是先验网络也不是 VQ-VAE 过度拟合。

5.2.2 精度 - 召回率度量标准

建议使用精度和召回率度量作为 FID 和初始分数的替代方案，用于评估 GAN 的性能[29,18]。这些指标旨在明确量化覆盖（召回率）和质量（精度）之间的权衡。 我们将模型中的样本与 BigGAN-deep 的样本进行比较，使用改进版本的精度-召回率测量，使用[18]中概述的相同程序对 ImageNet 中的所有 1000 个类进行比较。

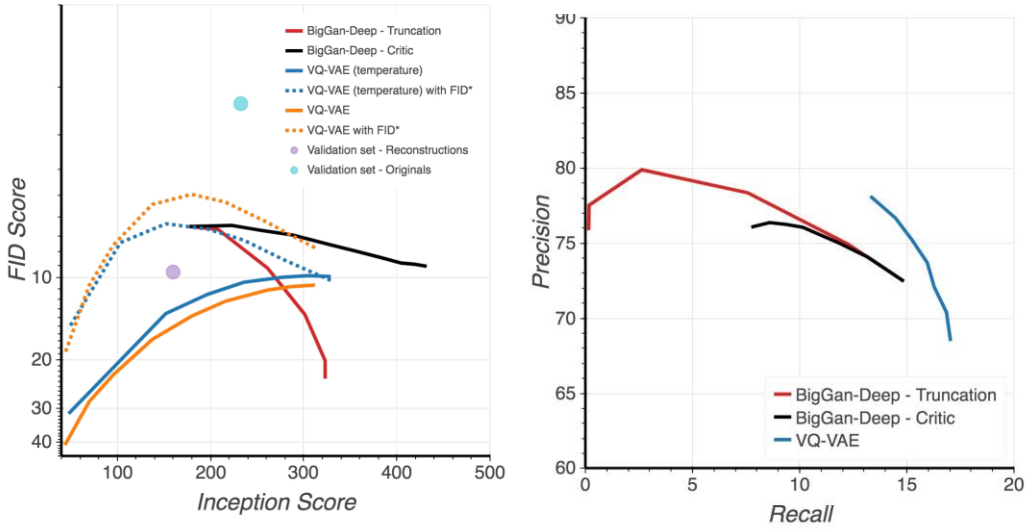
图 7b 显示了 VQ-VAE 和 BigGan 的精度-召回率结果，其中基于分类器的拒绝采样（'critic'见 3.3 节）针对不同的拒绝率和不同截断水平的 BigGan-deep 结果。 VQ-VAE 导致精度略低，但召回率更高。

5.3 分类准确度分数

我们还使用最近提出的分类准确度分数（CAS）[25]来评估我们的方法，这需要仅对来自候选模型的样本训练 ImageNet 分类器，然后从测试集中评估其对真实图像的分类准确度，从而测量样本质量和多样性。 表 2 中报告了我们对该度量的评估结果。对于 VQ-VAE，ImageNet 分类器仅针对样本进行训练，样本缺少高频信号，噪声等（由于压缩）。评估测试图像的 VQ-VAE 重建上的分类器关闭“域隔阂”（domain gap）并改善 CAS 得分而无需重新训练分类器。

	Top-1 Accuracy	Top-5 Accuracy
BigGAN deep	42.65	65.92
VQ-VAE	54.83	77.59
VQ-VAE after reconstructing	58.74	80.98
Real data	73.09	91.47

表 2：真实数据集，BigGAN-deep 和我们的模型的分类准确度分数（CAS）[25]。



（a）初始分数[31]（IS）和 Fréchet 初始距离分数（FID）[12]。

（b）精度 - 召回率指标[29,18]。

图 7：使用 FID / IS 和精度 - 召回率进行多样性 - 质量权衡的定量评估。

5.3.1 FID 和初始分数

比较 GAN 的两个最常见的指标是初始分数[31]和 Fréchet 初始距离（FID）[12]。虽然这些指标有几个缺点[2,29,18]，并且增强的指标（例如前一节中提供的指标）可能证明更有用，但我们在图 7a 中报告了我们的结果。我们使用基于分类器的拒绝抽样作为一种权衡质量多样性的方法（第 3.3 节）。对于 VQ-VAE，这改善了 IS 和 FID 分数，FID 从大约 30 到 10。对于 BigGan-deep，拒绝采样（称为 critic）比 BigGAN 论文[4]中提出的截断方法更好。我们观察到初始分类器对 VQ-VAE 重建中引入的轻微模糊或其他扰动非常敏感，如简单地压缩原件时由 FID 是~10 而不是~2。出于这个原因，我们还计算 VQ-VAE 样本和重建（我们表示为 FID*）之间的 FID，表明初始网络统计数据比 FID 建议的更接近真实图像数据。

6 结论

我们提出了一种使用 VQ-VAE 生成各种高分辨率图像的简单方法，该方法具有强大的自回归模型。 我们的编码器和解码器架构与原始 VQ-VAE 一样保持简单和轻便，唯一的区别是我们使用分层多尺度潜在编码图来提高分辨率。 我们最好的有条件样本的保真度与最先进的生成对抗网络竞争，在几个类别中具有更广泛的多样性，使我们的方法与 GAN 的已知限制形成对比。尽管如此，样本质量和多样性的具体测量仍处于起步阶段，目视检查仍然是必要的。最后，我们相信我们的实验证明潜在空间中的自回归建模是学习大规模生成模型的一个简单而有效的目标。

致谢

我们要感谢 Suman Ravuri，Jeff Donahue，Sander Dieleman，Jeffrey Defauw，Danilo J. Rezende，Karen Simonyan 和 Andy Brock 的帮助和反馈。

参考文献

[1] Samaneh Azadi, Catherine Olsson, Trevor Darrell, Ian Goodfellow, and Augustus Odena. Discriminator rejection sampling. In International Conference on Learning Representations, 2019.

[2] Shane Barratt and Rishi Sharma. A note on the inception score. arXiv preprint arXiv:1801.01973, 2018.

[3] M. Bauer and A. Mnih. Resampled priors for variational autoencoders. In 22nd International Conference on Artificial Intelligence and Statistics, April 2019.

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In International Conference on Learning Representations, 2019.

[5] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational Lossy Autoencoder. In Iclr, pages 1–14, nov 2016.

[6] Xi Chen, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel. PixelSNAIL: An Improved Autoregres-sive Generative Model. pages 12–17, 2017.

[7] Sander Dieleman, Aaron van den Oord, and Karen Simonyan. The challenge of realistic music generation: modelling raw audio at scale. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and
R. Garnett, editors, Advances in Neural Information Processing Systems 31, pages 7989–7999. Curran Associates, Inc., 2018.

[8] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. arXiv preprint arXiv:1410.8516, 2014.

[9] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. arXiv preprint arXiv:1605.08803, 2016.

[10] Jeffrey De Fauw, Sander Dieleman, and Karen Simonyan. Hierarchical autoregressive image models with auxiliary decoders. CoRR, abs/1903.04933, 2019.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg,
S.Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 6626–6637. Curran Associates, Inc., 2017.

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. arXiv preprint arXiv:1812.04948, 2018.

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. arXiv preprint arXiv:1812.04948, 2018.

[15] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. CoRR, abs/1312.6114, 2013.

[16] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In
Advances in Neural Information Processing Systems, pages 10236–10245, 2018.

[17] Alexander Kolesnikov and Christoph H Lampert. Pixelcnn models with auxiliary variables for natural image modeling. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1905–1914. JMLR. org, 2017.

[18] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. CoRR, abs/1904.06991, 2019.

[19] Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 29–37, 2011.

[20] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. CoRR, abs/1609.04802, 2016.

[21] Jacob Menick and Nal Kalchbrenner. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. In International Conference on Learning Representations, 2019.

[22] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.

[23] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image Transformer. 2018.

[24] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[25] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. arXiv preprint arXiv:1905.10887, 2019.

[26] Scott Reed, Aäron van den Oord, Nal Kalchbrenner, Sergio Gómez Colmenarejo, Ziyu Wang, Yutian Chen, Dan Belov, and Nando de Freitas. Parallel multiscale autoregressive density estimation. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 2912–2921. JMLR. org, 2017.

[27] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. arXiv preprint arXiv:1505.05770, 2015.

[28] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approxi-mate Inference in Deep Generative Models. 32, 2014.

[29] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In Advances in Neural Information Processing Systems, pages 5234–5243, 2018.

[30] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Advances in neural information processing systems, pages 2234–2242, 2016.

[31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 2234–2242. Curran Associates, Inc., 2016.

[32] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. CoRR, abs/1611.02200, 2016.

[33] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In International Conference on Learning Representations, Apr 2016.

[34] Lucas Theis and Matthias Bethge. Generative image modeling using spatial lstms. In Advances in Neural Information Processing Systems, pages 1927–1935, 2015.

[35] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. CoRR, abs/1601.06759, 2016.

[36] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel Recurrent Neural Networks. In
International Conference on Machine Learning, volume 48, pages 1747–1756, 2016.

[37] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. CoRR, abs/1711.00937, 2017.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. (Nips), 2017.

[39] Gregory K Wallace. The jpeg still picture compression standard. IEEE transactions on consumer electronics, 38(1):xviii–xxxiv, 1992.

[40] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Computer Vision (ICCV), 2017 IEEE International Conference on, 2017.

A 架构细节和超参数

A.1 PixelCNN 先验网络

	Top-Prior (32×32)	Bottom-Prior (64×64)
Input size	32×32	64×64
Batch size	1024	512
Hidden units	512	512
Residual units	2048	1024
Layers	20	20
Attention layers	4	0
Attention heads	8	-
Conv Filter size	5	5
Dropout	0.1	0.1
Output stack layers	20	-
Conditioning stack residual blocks	-	20
Training steps	1600000	754000

表 3：用于 Imagenet-256 实验的自回归先验网络的超参数。

	Top-Prior	Mid-Prior	Bottom-Prior
Input Size	32×32	64×64	128×128
Batch size	1024	512	256
hidden units	512	512	512
residual units	2048	1024	1024
layers	20	20	10
Attention layers	4	1	0
Attention heads	8	-	
Conv Filter size	5	5	5
Dropout	0.5	0.3	0.25
Output stack layers	0	0	0
Conditioning stack residual blocks	-	8	8
Training steps	237000	57400	270000

表 4：用于 FFHQ-1024 实验的自回归先验网络的超参数。

A.2 VQ-VAE 编码器和解码器

	ImageNet	FFHQ
Input size	256×256	1024×1024
Latent layers	32×32, 64×64	32×32, 64×64, 128×128
β (commitment loss coefficient)	0.25	0.25
Batch size	128	128
Hidden units	128	128
Residual units	64	64
Layers	2	2
Codebook size	512	512
Codebook dimension	64	64
Encoder conv filter size	3	3
Upsampling conv filter size	4	4
Training steps	2207444	304741

表 5：用于 ImageNet-256 和 FFHQ-1024 实验的 VQ-VAE 编码器和解码器的超参数。

B 更多的样本

请按照以下链接访问我们论文的完整版本，无需有损压缩即可呈现，其中包括其他样本。
https://drive.google.com/file/d/1H2nr_Cu7OK18tRemsWn_6o5DGMNYentM/view?usp=sharing