大规模生成图像模型输出的定性评估

Yannick Assogba、Adam Pearce 和 Madison Elliott

摘要——评估生成图像模型仍然是一项难题。这是因为模型输出具有很高的维度，既要表示训练数据又不能简单复制训练数据的任务颇具挑战，而且现有指标无法充分对应人类感知，也无法捕捉我们期望这些模型具备的所有特性。因此，对模型输出进行定性评估是模型开发和研究发表实践中的重要环节。现有工具对定量评估的支持仍然不足，因为它们难以帮助用户有条理地探索模型潜在空间中的大量样例。为了解决这一问题，我们推出了 Ravel，这是一种可视分析系统，能够对数十万张图像规模的模型输出进行定性评估。Ravel 可帮助用户发现模式崩溃等现象，并找出模型未能捕捉的训练数据区域。用户可以借助它评估生成图像的质量和多样性，并将其与真实图像或作为基线的另一模型的输出进行比较。本文介绍了三个案例研究，展示借助 Ravel 得出的关键见解，并以一项领域专家用户研究作为支持。

索引词—信息可视化、图像生成、机器学习

arXiv:2301.04518v1 [cs.HC] 11 Jan 2023

1 引言

生成式图像模型是一类基于神经网络的模型，旨在生成能够忠实刻画目标图像分布的**新颖**、**高质量**且**多样化**的图像。人们设计了各种架构和训练方法来学习此类模型，例如生成对抗网络（GANs） [1], 、变分自编码器（VAEs） [2],、流模型 [3]和扩散模型 [4]。

评估这些模型仍然很困难 [5]。所使用输出架构的高维性使得模型输出的似然估计十分困难，在某些情况下甚至无法实现。研究还表明，平均对数似然等指标并不总是与人类对样本质量的感知判断相关 [6]。此外，虽然我们希望模型能够很好地捕获目标分布，但不希望它们生成训练集中的实际图像（这一问题通常被称为记忆）。

围绕生成图像模型的文献中已经出现了许多指标 [7] [8],，其中 Fr´echet Inception 距离（FID） [9] 最为常用。然而，人们发现 FID 存在一些问题，因此发展出了精确率和召回率等更细粒度的指标 [10] [11]。

FID 等单一数值指标对于推动该领域的发展是必要的，但无法涵盖人们对这些模型所期望的全部质量。因此，人工视觉检查通常在生成图像建模方面的评估和成果传播中发挥着关键作用。然而，借助现有的评估工具，实践者通常只能查看这些模型输出空间中很小的一部分，规模大约为几十到几百张图像（例如， [12], [13],[14], [15], [16]）。

我们对领域专家的访谈证实，人工评估是从业者工作流程中的关键环节。一些专家依赖众包评估人员进行人工评估，但他们也意识到，这种方式往往成本高昂或耗时，因此只会在评估的最后阶段采用。即便会进行评估也是如此，因此他们在模型开发过程中主要依赖小规模定性评估。

与此同时，该领域的专家担心发表的结果存在挑选性呈现的问题，但在这些结果偶尔随学术论文一同发表时，他们通常没有办法全面探索模型输出。

为满足这些需求，我们构建了一个名为 *Ravel* 的系统，使用户能够对模型输出进行视觉检查，其规模最高可达典型用户工作流程的千倍。我们展示了该系统在规模从 50k 到 120k 张图像不等的数据集上的应用。这些数据集的规模与生成式图像模型标准定量评估所用的数据集相当。

我们的主要贡献包括：

- 一种可视分析系统，支持生成式图像模型的多种评估任务（例如评估质量和多样性、发现模式崩溃或模型输出中的缺失），且不依赖于模型架构及其内部机制。
- 对大型生成图像模型数据集进行交互式探索，通过聚类以及细粒度的簇指标可视化来引导定性评估。
- 一种用户界面，通过由语义上有意义的嵌入空间驱动的视觉比较，支持用户推理图像分布之间的差异，并提出关于模型行为的假设。

2 背景

2.1 生成式图像模型

过去几年中，生成图像模型的能力大幅提升。自从最初的 GAN 论文 [1] 打破了人脸建模的僵局以来，如今我们已有 BigGAN [12], StyleGAN [13],GLOW [14],向量量化变分自编码器 [15], CDM [16] 等系统，它们采用了多种多样的模型架构和训练算法，并能够在广泛的领域中生成非常逼真的图像。

• *Yannick Assogba*、*Adam Pearce* 和 *Madison Elliot* 均来自谷歌研究院。电子邮箱：{*yassogba,adampearce,madisone*}*@google.com*

2.2 评估生成图像模型的定量指标

本节将概述研究文献中最常引用的指标：

- **Fr´echet Inception Distance** [9]：使用经过预训练的 InceptionV3 分类器 [17]为真实图像和生成图像生成嵌入，然后使用统计度量比较这两类图像的嵌入分布。FID 是文献中最受欢迎的指标。要得出准确估计，它需要大量样本（通常至少需要 5 万张生成图像），且无法检测对训练集的记忆。Karras 等人 [18] 指出，基于 ImageNet 的卷积神经网络（如 InceptionV3）[19] 存在纹理偏差，这意味着由此类网络衍生出的指标无法捕捉图像质量的所有方面。

- **Inception 分数** [20]：使用经过预训练的 Inception 分类器来衡量：*a)* 每张生成图像与单个 ImageNet 类别的匹配程度，以及*b)* 整组生成图像对所有 ImageNet 类别的覆盖是否均匀。与 FID 类似，Inception 分数需要相当多的图像，且无法用于检测记忆。它也无法衡量类内多样性或检测模式崩溃。有关此指标问题的详细讨论，请参阅 Barrat 和 Sharma [21] 。**Inception 分数和 FID** 都是旨在同时反映图像质量和多样性的标量分数，因此无法揭示模型是否通过牺牲其中一项来提升另一项，从而获得更高的分数。

- **精确率与召回率指标**：为区分图像质量和多样性的测量，Sajjadi 等人[10] 和 Kynk¨a¨anniemi 等人 [11] 提出了精确率和召回率指标，以分别进行独立测量。广义而言，精确率对应于样本质量，而召回率 对应于样本分布相对于目标分布的覆盖程度——即多样性。

这些指标通常在整个数据集上计算，因此粒度较低。即使这些指标表明某个模型表现更好或更差，它们也无法指出生成图像分布中的哪些部分出现了改进或退化。Ravel 提高了这些指标的粒度，提供了一种查找在某些指标上得分较低的特定图像簇的方法

2.3 生成式图像模型输出的定性评估/可视化

由于上述定量指标存在局限性，研究人员还依靠对模型输出进行视觉检查来评估模型性能。视觉检查通常在训练期间进行，以监控训练过程是否立即失败；训练完成后也会进行视觉检查，以评估模型的整体质量。我们在第 6.1 节与领域专家共同验证并详细阐述了这一工作流程和策略。尽管模型通常会输出数十万张图像，但我们的用户研究发现，研究人员在分析过程中只能检查其中很小一部分图像（约数百张）。

视觉效果出色的样本对于在学术会议上成功发表模型进展至关重要。生成式图像模型领域的相关论文通常会包含数十到数百张图像 [12] [13] [18] [22]。这仅仅展现了这些模型通常接受训练以生成的多种图像中的极少一部分。我们的用户研究还发现，人们普遍认为作者会“挑选”最好的图像纳入论文发表。与论文一同发布未经筛选的输出图像大型数据集的作者相对较少。

目前还没有专门打造的界面，能够方便地浏览或检查这些内容。例如， [13] [18]eachpublish 会将 100koutputimages 发布到公开可访问的 Google 云端硬盘中。这些图像被整理到 1000 个子文件夹中，以便在文件数量庞大的情况下提升界面的可用性。要查看这些输出，用户必须逐个点击文件夹，或者自行使用界面来浏览图像。

3 相关工作

Borji [7] [8]汇总了用于自动评估生成模型的许多指标。Ravel 采用了 [10] 和 [11], 中的精确率和召回率指标，但**未**提出任何新指标。因此，我们主要将这项工作置于有关**定性评估**和**界面**的研究背景下，以探索生成模型输出。

众包工作者评估

Denton 等人 [23]使用一小组志愿人工标注员，通过询问他们能否区分真实图像与生成图像来估计质量。Zhou 等人 [24] 改进并扩展了这一技术，要求来自亚马逊 Mechanical Turk 的众包工作者对真实图像与生成图像进行心理物理判断。虽然这些方法能够很好地将评估扩展到更大的数据集规模，但与研究人员进行人工检查相比，它们耗时更长、成本也更高。因此，这些方法通常会被保留到评估流程的后期阶段使用。此外，它们往往侧重于可在单幅图像上进行评估的指标（例如图像质量），而不是语料库层面的属性（例如图像多样性）。

相比之下，Ravel 旨在支持在使用外部评分者之前、开发流水线的更早阶段进行研究人员评估，并允许研究人员在同一界面中评估质量和多样性。它介于对训练动态的初步监控（以确保模型正在收敛）与通常在模型发布临近时进行的大规模人工评分评估之间。

解释模型内部机制

Bau 等人 [25] 探索了在生成对抗网络中寻找可解释单元（神经元），并将消融这些神经元所产生的因果效应可视化。他们的方法依赖于访问模型内部机制，并需要一个预训练的目标分割网络来找出场景中的对象，从而建立神经元激活与网络输出之间的因果关系。 [26] Bau 等人使用预训练的分割网络，比较生成图像中发现的对象与一组真实图像中发现的对象的分布。这提供了一种衡量模型输出多样性的指标，考察的是预训练网络能够分割的对象。作者还提出了一种方法（层反演），用于训练能够将真实图像近似反演到模型潜在空间中的网络，以观察网络在缺失对象的位置生成了什么。

尽管这些方法对于更好地理解驱动模型行为的内部机制至关重要，但它们通常针对特定的模型架构。Ravel 将模型视为黑盒，因此具有与模型架构无关的特性。Ravel 确实使用预训练网络来计算图像的向量表示，但对预训练网络所接受训练以执行的最终任务不太敏感。例如，

可以使用待研究模型的嵌入，或任何学习到了具有语义价值的特征的模型（例如 InceptionV3）。

模型输出的在线探索Wh
像的多种方式探索，从而能够在模型
比较两个或
样图

他们引入了多种可视化方法，用于考察模型针对用于测试模型行为的特定输入图像 的表现。在后续工作中，White & Loh [28] 引入了一种基于电子表格隐喻的新型可视化界面，使用户能够在潜在空间中使用几何操作，以交互方式查询这些模型，从而探索其输出。

虽然在线方法使用户能够探索特定假设，但由于图像生成速度较慢，这些方法通常只能支持相对较小的探索空间。Ravel 专注于对生成图像进行离线分析，因此能够考察规模大得多的数据集，并可作为生成假设的手段，与在线方法互为补充，以便开展后续更具针对性的研究。

基于嵌入的可视化
许多研究对大型高维数据集的嵌入空间进行了可视化 [29] [30] [31] [32]。

Liu 等人 [33] 提出了一种可视分析系统，利用变分自编码器的编码器 学习到的潜在空间，探索现有图像数据集中的变化。这项任务在概念上与 Ravel 支持的功能相似，但我们并不尝试学习新的潜在空间，因为我们希望聚焦于生成器的输出及其潜在空间（而非固定的输入数据）。

Ravel 以这些较早的嵌入空间可视化研究为基础，并引入聚类，使这些空间更易于浏览。我们着重通过数据集比较来为输出空间的探索提供参照，并满足图像生成应用场景的需求。

Xiang 等人 [34] 提出了一种可视分析系统，用于纠正大型数据集中的标注错误。他们将图像嵌入的 t-SNE 投影可视化，并按照标签进行颜色编码，以突出显示标注错误的数据点。相比之下，Ravel 旨在支持无条件生成场景，其中生成图像（以及真实图像）没有标签。因此，Ravel 不依赖标签来提供参照，而是支持对数据集进行比较评估，从而以真实图像的分布为参照，评估生成图像。

4 方法与系统设计

Ravel 致力于支持用户将模型输出与基线图像集合进行比较时进行视觉检查，基线通常是真实的“真实值”图像。根据我们的访谈（第 6.1 节），我们了解到，实践者会对模型生成的图像样本进行可视化分析，但通常只涉及少量图像。我们的目标是支持并扩展这些工作流程，使其能够更系统地对从几十或几百张图像到数万或数十万张图像进行视觉检查。为了便于对大量图像进行比较，我们采用了四个关键概念：

1) **图像嵌入**：神经图像嵌入为图像提供了语义丰富的向量空间，可用于计算成对图像之间的相似度分数 [35]。它们已成为生成式图像模型评估流程中的标准（第 4.1.2 节），而**我们希望在 Ravel 中利用研究人员对基于嵌入的指标的熟悉程度**。计算图像之间语义相似度的能力，使我们能够将相似的生成图像归为一组**并**将这些图像与相似的真实图像进行比较。默认情况下，我们计算 FID 等指标所用的标准 InceptionV3 嵌入，但也可以添加来自其他模型的嵌入，包括预训练模型或受评估模型（如果这些模型可用）。

2) **图像聚类**：为了实现对数十万张图像的可视化探索，我们需要将它们分组为数量较少且有意义的组。我们使用图像无监督聚类来减少用户需要考虑的顶层项目 数量，例如，将 10 万张图像减少到 1000、500 甚至 250 个簇。我们使用 K 均值聚类，相关详情将在下文的流程细节 部分介绍

3) **簇指标**：将顶层元素的数量缩减到便于管理的程度后，我们希望为用户提供提示，帮助他们判断**哪些簇最值得优先探索**，并**搭建可重复的工作流程，以引导探索**。我们针对每个簇计算指标，从而能够按照可预测的顺序对簇进行排序。我们计算的许多指标旨在揭示某个簇中生成数据与基线数据之间的差异，其他指标则展示簇本身的一般属性。有关所计算指标的更多详情，请参见下文的流程细节 部分。按照这些指标对簇进行排序，既有助于发现离群簇，也有助于分析生成式图像空间中不同部分的属性（例如，查看哪些类型的输出图像精确率较低）

4) **交互式可视化**：最后，我们提供响应迅速的交互式可视化，用于呈现图像、簇及相关簇指标。由于所有数据均由我们离线计算，用户界面的响应速度很快，能够让用户快速、自由地探索大量数据点。

4.1 流程细节

4.1.1 聚类

Ravel 使用 Scikit-learn 中 K 均值聚类的实现对图像嵌入进行聚类。选择 K 均值作为起始聚类算法颇具吸引力，因为它是一个广为人知的算法，且只有一个容易理解的超参数，即簇的数量。这使用户能够直接为这一超参数指定他们认为适合其数据集的值。此外，与谱聚类等具有类似超参数选项的其他算法相比，K 均值在样本数量和所选簇数量增加时仍能良好扩展 [36]。K 均值通常也会生成大小较为均匀的簇，这有助于以统一的方式展示其内容。

所有聚类方法都会受到簇分配不完善问题的影响，尤其是在簇的边界处。不过，Ravel 通过根据簇在潜在空间中的位置来可视化簇，在一定程度上缓解了这一问题。这样，用户便能快速发现并检查彼此相近的簇，并判断它们是否实际上属于同一个语义簇。更多详情请参阅下文的降维部分（4.1.3）。

4.1.2 簇指标

我们为每个簇计算多个指标。多个指标提供了不同的视角，帮助我们考察这些簇。例如，人们可能会关注

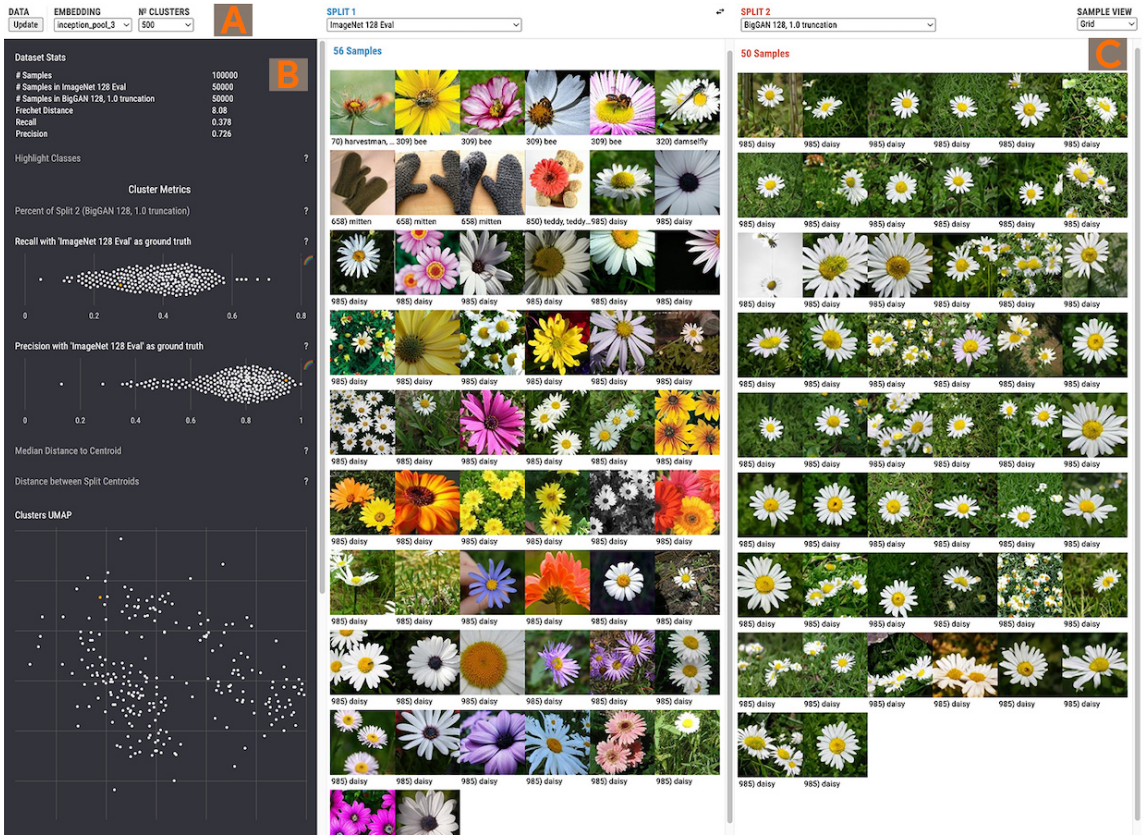

图 1. Ravel界面主要由以下部分组成：A）数据集和视图选项。B）汇总图表和关联的簇图。C）并排显示的图像网格，用于直观比较各簇。此视图展示了一个簇，左侧为真实图像，右侧为生成图像。

召回率或精确率较低的情况，或者簇排列紧密或分布较为分散的情况。

簇中的图像可能来自生成数据或基线/真实值。在用户界面中，我们将这些数据源称为数据划分，通常会在左侧显示基线数据，不过用户可以交互式地更改此设置。许多指标旨在揭示当前正在探索的两个数据划分之间的差异。

- **划分 2 的百分比：** 簇中的图像来自右侧所示数据划分（通常为生成数据）的比例。
- **召回率：** 召回率是一项指标，用于描述真实值 数据划分中的样本有多少比例在另一个数据划分中也有对应支持。我们按照 [11] 中的定义计算召回率，但会汇总每个簇中各图像的数值。比较生成图像与真实图像时，较高的召回率意味着模型生成的图像具有与真实数据同样丰富的多样性。
- **精确率：** 精确率是一个指标，用于描述生成数据中有多少比例的样本在真实值中有对应支持。如上所述，我们按照 [11] 中的定义计算精确率，但会汇总每个簇中各图像的数值。在将生成图像与真实图像进行比较时，较高的精确率意味着图像质量/真实感更好。
- **数据划分质心之间的距离：** 衡量簇中属于各数据划分的样本质心之间的距离。数值越大，表明两个数据划分中的数据在视觉上差异越大；数值越小，则表明左侧和右侧数据划分在视觉上越相似。
- **到质心的距离中位数：** 簇中样本到该簇质心的距离中位数。数值越小，表示簇越紧凑，样本之间越相似；数值越大，则表明簇中的图像种类越丰富。

4.1.3 降维

除了通过上述指标对簇进行可视化外，我们还会在嵌入空间中可视化聚类簇质心的位置。由于这些嵌入空间具有很高的维度，我们需要使用降维方法将其可视化为二维图像。我们使用 UMAP 算法[37]进行这一投影。其他降维方法包括 PCA [38] 和 t-SNE [39]。据报道，与 t-SNE 相比，UMAP 能更好地保留原始空间的全局结构，并且随着维度数量增加和数据集规模扩大，其计算效率都显著更高 [37] [40]。我们使用的 InceptionV3 嵌入有 2048 个维度，因此计算效率高的方法尤为有吸引力。虽然 UMAP 不像 PCA 这样的线性方法那样容易直接解释，但它更善于捕捉嵌入空间中编码的图像之间某些复杂的非线性关系。

4.2 用户界面

该用户界面是一个基于浏览器的应用程序。我们将介绍用户界面的设计，并在案例研究部分进一步讨论这些功能的实用性。

图 1 展示了 Ravel界面，分为 3 个主要部分：

4.2.1 (A)：数据集与视图控制
用户可以选择要使用的嵌入、簇的数量，以及要在左侧或右侧显示的数据划分。

4.2.2 (B)：图表
用户界面的图表部分主要有三种数据展示方式。

汇总统计

首先是一个静态表格，展示整个数据集的汇总统计和指标。其中包括整个数据集中的样本数量、每个数据划分中的样本数量，以及数据集级别的指标，例如 Fr´echet 距离[1],召回率和精确率 [11]。

簇指标图

汇总图表下方是一系列蜂群图，展示所计算的簇级指标（详情请参见上文的簇指标部分）。每个蜂群图都以**一个点表示一个簇**，并绘制各簇在该指标上的分布。这些图表支持交互操作，选择单个点即可在样本查看器中显示该簇的图像。

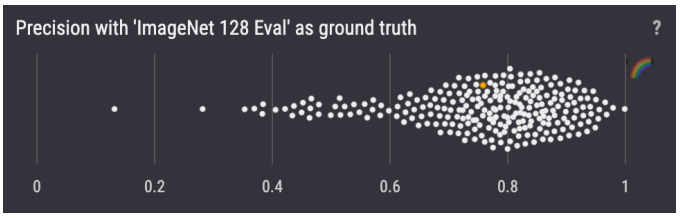

图 2. 展示簇精确率得分分布的蜂群图。每个点代表一个簇，当前选中的点以橙色显示。点击 ? 图标可查看指标说明。

点击图表旁的彩虹图标开关，也可以应用颜色编码。此颜色编码会应用于所有图表，从而便于比较不同指标（图 3）。



图 3. 按**精确率**得分着色的召回率和精确率蜂群图。可通过召回率图中的位置编码，以及精确率图应用的颜色编码，识别召回率高、精确率低的簇。为节省显示空间，我们不显示图例，因为彩虹图标开启的图表会直接呈现数值高低与色相之间的关系，而且相较于具体数值，相对比较更为重要。

UMAP图

1. 使用 Inceptionv3 嵌入时，这是 Fr´echet Inception 距离 (FID)

蜂群图下方有两张图，展示了二维 UMAP 投影：a) 所有簇，以及 b) 当前选定聚类簇中的图像（见图 4）。尤其是聚类簇 UMAP 视图，用户可以按视觉相似性而非指标得分浏览各个簇。将鼠标悬停在样本 UMAP 图上的点上，即可预览单张图像

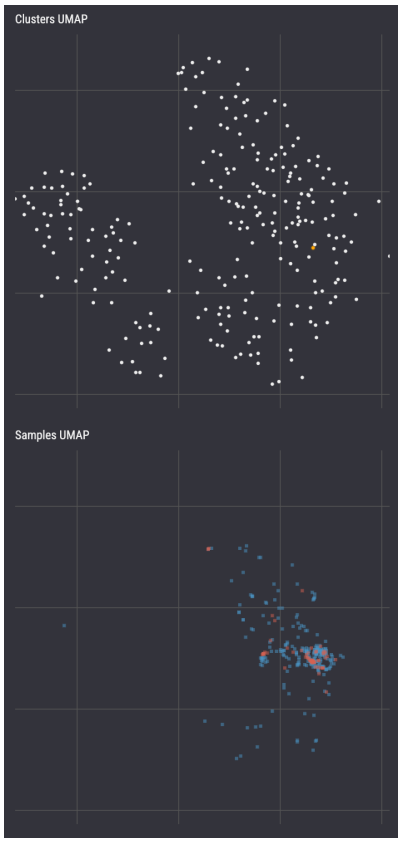

图4。上：嵌入空间中**聚类簇质心**的UMAP投影。下：当前选定聚类簇的UMAP投影，真实样本以蓝色显示，生成样本以红色显示

高亮类别I

数据集包含类别标签，Ravel 将显示一个搜索界面，允许用户搜索并选择数据集中的任意数量的匹配类别。

如果选择了一个或多个类别，Ravel 会在簇图（5）中淡化那些不包含这些类别图像的簇。

所有簇图的交互都相互关联，使用户能够查看某个簇在多个指标或嵌入空间中的位置。

点击图表标题即可将其折叠，这样就能为用户认为对其分析最有用的图表腾出更多空间。

4.2.3 样本查看器
图表右侧是样本查看器。它由一对可滚动视图组成，显示选定簇中按数据划分分类的图像缩略图。如果图像关联了类别/标签，这些信息会显示在缩略图下方，且同一类别的图像会归为一组；否则，图像将按其在流水线中的处理顺序显示（实际上是随机排列）。用户可以点击图像以查看较大的图像，并查看任何其他元数据。

图像缩略图的最大尺寸为 150x150px，最小尺寸为 100x100px，具体取决于

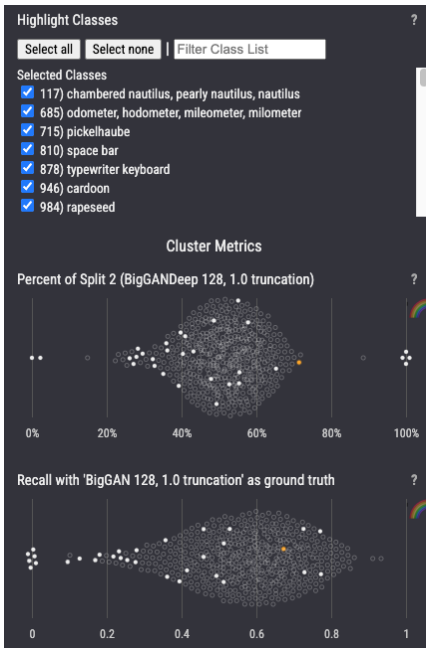

图5. 在所有簇图中，至少包含一张所选类别图像的簇会突出显示。



图6. 基于StyleGAN2的模型难以对带有脸部彩绘和其他色彩鲜艳、遮挡面部的配饰的面部图像进行建模，反而会生成这些伪影。

浏览器窗口的大小。在 3840 x 2160分辨率显示器上[2]，通常一屏可以显示每个数据划分中的 72 张图像，总计 144 张。不过，一个明显的局限是，如果一个簇中的图像数量超过这个数，用户就无法一次看到所有图像，必须滚动浏览，才能更好地了解该簇。一种缓解方法是选择更多的簇进行查看。这样会形成更小、更细粒度的簇，但需要浏览的内容也会更多。

5 案例研究

本节通过一系列案例研究，说明 Ravel 的设计特性如何支持用户探索。这些案例研究展示了作者使用该工具时所发现的内容。我们的 3 个用例是：1）单一领域（人脸）的无条件图像生成；2）基于 ImageNet 数据的类别条件生成；3）分析生成模型在另一项任务中的下游应用——本例中为超分辨率。

虽然 Ravel 支持使用多种嵌入进行聚类和探索，但本文主要关注*FID* 分数及其他指标中使用的相同 *Inceptionv3* 嵌入，因为它们已成为生成模型领域事实上的标准，且可公开获取。

5.0.1 无条件图像生成

无条件生成指的是使用训练好的模型生成图像，除从随机分布中抽取的输入向量（通常称为“噪声”向量或“Z”向量）外，不施加其他“条件信息”。在本案例研究中，我们使用一个基于 FFHQ 数据集训练的 StyleGAN2架构实现 [3] [18] 生成 60k 张图像。这些图像

2. 此分辨率是默认的*4K* 分辨率，广泛可用。它也接近 2021 款 14 英寸 MacBook Pro 的原生分辨率（3024 x 1964）

3. 我们想明确说明，我们使用的是独立训练的 StyleGAN2，而不是原作者发布的 StyleGAN2 权重。

这些图像由从均值为零、标准差为一的随机分布中抽取的 Z 向量生成。我们加载这些图像以及训练集中的 67,542 张图像，总计 127,542 张图像。

该模型生成的图像中，是否存在不属于输入分布的图像？

分析人员在浏览器中打开Ravel界面，并将簇数设为250，左侧显示训练数据中的图像，右侧显示生成图像。她知道精确率较低通常意味着图像不够逼真，于是开始查看精确率图表中较低端的簇，并注意到精确率最低的10个簇中，每个簇都包含存在明显视觉缺陷的生成图像。将这些图像与真实数据进行对照，让她了解模型在各个类别中难以捕捉哪些特征。例如，她注意到模型难以处理复杂背景、包含多张面孔的肖像，或带有手、麦克风等遮挡物的肖像。另一个显而易见的问题是，模型难以生成带有脸部彩绘的面孔，如图6所示。将这些失真的图像与训练数据中的真实图像并列展示，使她得以推测模型为何难以处理这类图像；她尤其注意到，这类图像在训练数据中相对少见。

在检查 10个簇时，她快速浏览了约 4234 张生成的图像和 2718 张真实图像。

5.0.2 类别条件图像生成

类别条件生成是指经过训练、能够针对一组固定的不同类别生成输出图像的模型。在这里，我们直接比较两个使用相同数据和任务进行训练的模型的输出：即从 ImageNet数据集的 1000 个类别生成图像，分辨率为 128x128px [41]。我们使用 Lucic 等人论文《GAN是否生而平等？一项大规模研究》[42]中的模型实现。

我们搭建了一个 Ravel 实例，其中包含由 BigGAN 128 和 BigGAN-deep 生成的 5 万张图像 128 [12], , 以及来自

ImageNet[4]验证集的 5 万张图像。每个模型的输入向量均按照[42]，中的描述，从正态随机分布中抽取，且不对输入向量进行截断。

使用这些模型展示了一种工作流程：研究人员试图确定模型的某个变体或改进后的模型架构，与某个基线模型相比，行为有何不同。这是机器学习中常见的工作流程，通常通过报告 FID 等关键指标的表现来完成。我们展示了 Ravel 如何补充传统指标，以发现模型行为的具体差异。

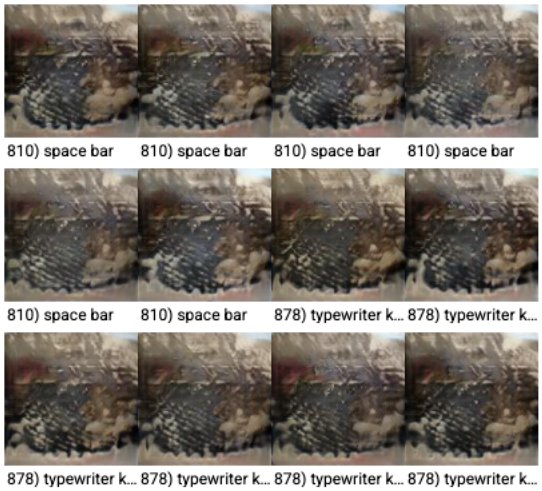

图 7. BigGAN-deep 模型中“空格键”和“打字机键盘”类别发生模式崩溃。该簇中的 100 张图像看起来完全相同，因为这两个类别都崩溃成了相同的输出。



图 8. BigGAN 中“空格键”和“打字机键盘”类别的输出多样得多。该模型在这些类别上没有出现模式崩溃

BigGAN 与 *BigGAN-deep* 的输出多样性相比如何？

我们的分析人员打开 Ravel，并将簇的数量设为 500，左侧数据划分显示 BigGAN 的输出，右侧数据划分显示 BigGAN-deep 的输出。

他立刻注意到，召回率图表中有多个簇的得分为 0。他点击在其中一个簇上进行操作后，他发现其中仅包含来自 BigGAN-deep 的图像。该模型生成的全部 100 张图像都属于两个类别，几乎完全相同，而且视觉质量较低（见图 7）。这是模式崩溃 [43]的典型案例；模型未能学习这些类别的真实分布，并将这些类别的所有输出“崩溃”为一张图像。在本例中，两个类别都崩溃为同一张输出图像。分析人员点击其他召回率为零的簇，发现 BigGAN-deep 在以下类别中也存在类似的模式崩溃：“空格键”、“打字机键盘”、“尖顶盔”、“油菜籽”、“刺苞菜蓟”、“鹦鹉螺、珍珠鹦鹉螺、鹦鹉螺”以及“里程表”。

使用 Ravel 5 的“高亮类别”功能后，分析人员能够找到所有包含这些类别图像的簇，并确认 BigGAN 模型没有在这些类别上出现模式崩溃（图 8）。

那么，对于两个模型在生成效果上都还不错的类别（即两个模型都没有出现模式崩溃等严重故障），情况又如何呢？

分析人员现在决定将簇的数量设为 1000，以提高簇的粒度。他启用按颜色选项，然后查看**精确率图表**，再通过**召回率图表**寻找召回率较高但精确率相对较低（不过没有上一节那么低）的簇。查看召回率较高的簇，有助于找到其分布至少存在部分重叠的簇；同时关注精确率，则可以发现质量差异。这些簇在召回率图表右侧显示为浅黄色圆点。他随机选择了一个，发现这是一个海胆簇。两个生成器生成的输出质量都很高，但视觉检查表明，BigGAN-deep 生成的输出更多样（图 9）。

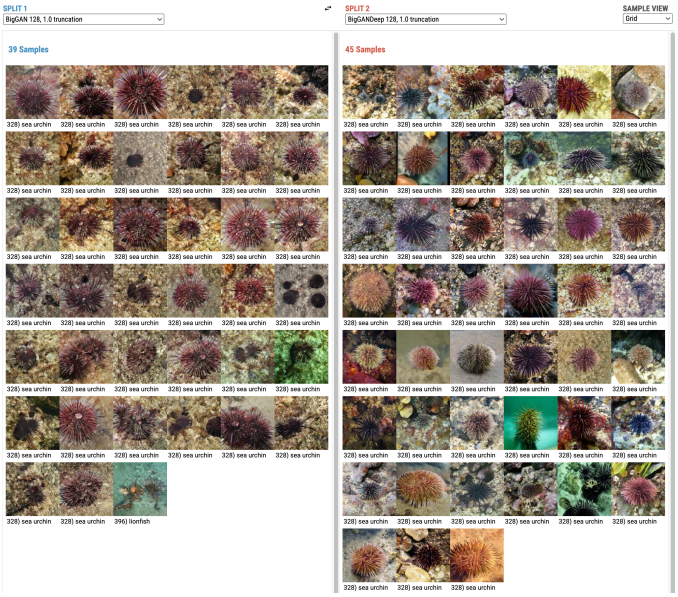

图 9. 两个生成器生成的一簇海胆样本，两者都能生成高质量图像，但右侧的生成器生成的结果更加多样。右侧图像呈现出更加丰富多样的颜色、纹理和姿势。

按视觉相似性进行聚类，而不是按标签分组，其优势在于更容易发现类别之间视觉语义的重叠或泄漏。使用与上述相同的方法，我们的分析人员选择了另一个簇，其中包含相对逼真的抹布图像（图10）。随后，他突出显示所有包含抹布的簇。现在，他可以

4. 从 https://www.tensorflow.org/datasets/ catalog/imagenet2012 获取的数据

逐一检查所有这些簇。他在（图11和图12）中展示的发现表明，其他类别，即“手帕”和“门垫”，正在“泄漏”到每个生成器对抹布的建模方式中。

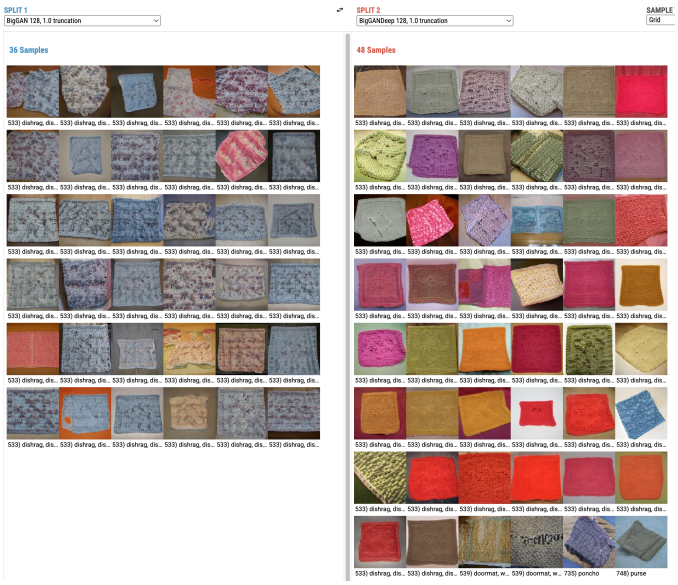

图 10. 左侧是 BigGAN、右侧是 BigGAN-deep 生成的一簇“抹布”样本。两者都能生成看起来较为逼真的抹布，但右侧的生成器生成的图像具有更多样的纹理和颜色。



图 11. BigGAN 的一个簇中有少量抹布，但有许多色彩鲜艳的手帕，其图案与主要抹布簇中的图案相似。这表明这两个概念的视觉表征之间存在一定的泄漏。



图 12. BigGAN-deep 的一个簇中有少量抹布，但有许多门垫。这为许多 BigGAN-deep 生成的抹布具有非常粗糙的纹理提供了一种解释。

5.0.3 生成模型的下游应用：超分辨率

在最后一个案例研究中，我们考察了 PULSE（通过潜在空间探索进行照片上采样）算法 [22],——一种超分辨率算法。该算法在生成式图像模型（本例中为 StyleGAN2）的流形上进行搜索，从而创建与低分辨率输入图像相对应的逼真上采样图像。该模型在发布后受到严厉批评，因为用户很快发现，模型在对具有非高加索人种特征的人物图像进行上采样时存在缺陷 [44]。我们使用这一例子来说明，对输出流形进行更广泛的探索，如何帮助研究人员和

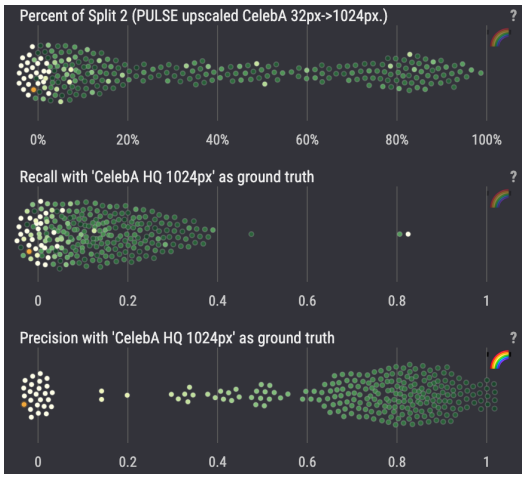

图 13. 按簇精确率着色的簇指标图表。精确率图表左侧是一组精确率很低的簇，其中大多只包含来自真实数据的图像；其中许多簇的召回率也很低，这表明算法难以为它们生成任何输出。

使用这类模型的从业者在将其应用于下游应用时了解其弱点，并在发布前降低风险。

在此场景中，我们使用 Karras 等人提供的原始 CelebA-HQ 数据集 [13]。该数据集包含 30,000 张名人面部肖像图像，分辨率为 1024x1024px。我们将 CelebA-HQ 图像缩小至 32x32px，再使用 PULSE 算法 [5] 将其上采样回 1020x1024px。我们使用作者在模型发布版 6 中提供的默认设置调用 PULSE，不过，为了提高 PULSE 针对给定输入生成结果的概率，我们将运行该算法的步数从 100 步增加到 200 步。[7]

随后，我们使用 Ravel 比较原始的 1024x1024px 图像和经 PULSE 上采样的图像。总计有 30,000 张原始 CelebA-HQ 图像，以及 22,092 张由 PULSE 输出的图像。7,908 张图像在经过 200 步 PULSE 算法处理后仍未能生成任何输出。

PULSE 难以为哪些类型的图像生成任何输出？

我们的分析人员将簇数设为 250，并将原始的 CelabA-HQ 图像放在左侧的数据划分中，随即注意到精确率图表最底端有一组簇。她发现，许多簇的召回率也很低，而且其中大多只包含来自 真实值的图像（见图 13）。她点击了一些低召回率、低精确率的簇，观察到算法难以生成任何输出的几类图像。这些图像包括：人们戴着帽子或其他头饰的图像；人物面前有麦克风或手的图像；脸部有大量脸部彩绘或化妆的图像；以及人们戴着太阳镜的图像。每个簇都包含 60 至 120 张图像，图 14 展示了其中的一些样本。

PULSE 难以为哪些类型的图像生成高质量输出？

5. 这一缩放因子（24 倍）处于原作者在评估中使用的缩放因子范围（8 倍至 64 倍）之内。 6. https://github.com/adamian98/pulse 7. 我们这样做是因为 PULSE 并非总能为输入图像生成结果，而我们发现，增加允许的步数后，能生成某种结果的输入图像数量显著增加。



图 14. 每一行包含来自不同簇的样本，这些簇中只有真实数据的图像，没有算法输出的图像。这些例子展示了算法难以放大的图像。

她继续从左向右扫描图表，探索精确率较低的簇。精确率最低的簇之一包含戴眼镜（也有少数戴太阳镜）的人物真实图像；然而，从算法输出中，她看到了一些质量较差（即不真实）的图像，这些图像很可能是将主体戴太阳镜的图像放大而来（图 15）。她进一步完善了关于算法如何处理戴太阳镜肖像的假设，并得出结论："如果一个人戴着太阳镜，算法往往无法生成任何图像；即使生成了，输出也不真实"。



图 15. 来自一个簇的样本，这些样本是戴着太阳镜的人物放大图像。

PULSE 擅长上采样哪些类型的图像？精确率图表也可用于查看 PULSE 擅长处理哪些图像。分析人员浏览分布右侧的簇时发现，高质量输出大多是肤色较浅、看起来相对年轻或处于中年的面孔。

尽管她无法获取关于肤色或年龄等敏感特征的定量性能指标，但通过人工检查，她发现了模型在人口统计类别方面的一些失效模式和优势。

6 领域专家用户研究

我们通过一项分两个阶段开展的研究对 Ravel 进行了评估：第一阶段确定领域专家的模型评估目标、工作流程和现有工具，第二阶段考察 Ravel 用户界面的可用性和实用性。参与者从一家大型科技公司的全职员工便利样本中招募，且在性别认同（n_{female} = 1、n_{Male} = 5）、产品领域（5 个不同团队）和办公地点（n_{Israel} = 2、$n_{UnitedStates}$ = 4）方面具有多样性。第一阶段（n = 4）包括四名目前从事生成模型训练和评估的专家研究科学家及软件工程师；第二阶段（n = 5）包括第一阶段的三名用户，以及另外两名具有生成模型工作经验的用户。两个评估阶段均通过远程主持的 Google Meet 视频通话与用户交流，并让他们共享屏幕，展示他们如何与 Ravel 互动。

6.1 第一阶段：评估的当前状态 ion

第一阶段旨在了解当前生成式图像模型输出评估的现状。研究采用半结构化访谈，询问参与者对生成模型的熟悉程度及日常工作、模型评估的目标与动机，以及当前的工作流程和实践。四位用户目前都从事生成模型相关工作，熟悉 StyleGAN 及其变体、BigGAN 及其变体，以及其他基于 ImageNet 和 FFHQ 训练的模型。他们使用的架构以及将生成模型应用于的任务各不相同。本文报告用户中最常见的实践。

6.1.1 评估目标与工作流程

四位参与者都表示，发表模型性能或改进成果是其工作的主要目标。四位参与者也都认为，评估图像输出质量对于理解模型性能至关重要。

四位参与者都表示会结合使用定量和定性评估方法。总体而言，他们的工作流程可归纳为以下常见模式：训练模型，同时进行有限的视觉检查以进行合理性核验→ 检查指标 → 继续训练 → 重新检查指标，直至达到预设阈值 →对图像输出进行定性视觉检查。

两位参与者表示，他们认为当前确定图像质量的"最佳实践"是人工定性评估，通常来自众包研究。重要的是，四位参与者也都强调了在模型评估中普遍依赖 FID 分数及其局限性。一位用户表示，"FID 分数远胜于 Inception 分数和其他指标"，但也承认"评估生成图像并没有黄金标准"。所有用户都表示，他们对将 FID 分数作为主要评估指标并不满意，而且都曾遇到或读到过一些案例，认为 FID 分数与人工判断不一致；Zhou 等人的研究也支持这一观点。 [24]

6.1.2 评估工具

四位用户都提到，他们使用 TensorBoard [45] 或 Colab[8]（一种类似于 Jupyter Notebook [46]的计算笔记本环境）来检查模型输出。用户指出，Colab 的主要优势是灵活性。Colab 的灵活性支持定制分析，我们将在 6.2.5 节中详细介绍。其主要局限在于难以在不同项目间复用代码和共享结果，尤其是与非技术类合作者，或未直接参与模型训练的利益相关者共享时。用户指出，TensorBoard 的主要优势是在训练过程中易于使用，可以快速将指标可视化，并查看训练不同阶段的样本输出。用户提到，TensorBoard 的主要局限在于加载大量图像时延迟较高，而且与 Colab 这类开放式工具相比，定制化程度较低。

8. https://colab.research.google.com/

6.1.3 定性视觉评估任务
据报告，判定图像质量是最重要的任务。对所有四位用户而言，判定图像的多样性也很重要，其中一位用户提到，检查是否出现模式崩溃是其评估流程中的明确环节。所有用户都表示，进行更细致、更定制化的图像生成很重要，包括查看特定类别中的样本，或嵌入空间中彼此接近的样本。一位用户谈到，检查潜向量的不同截断程度，以便对输出的真实感和美学做出判断。用户表示，他们只会查看少量图像。一位用户报告称，在模型训练流程的每个评估步骤中，会查看约 64 张图像，且从不超过 100 张。另一位用户解释说，与团队合作进行评估时，他们通常会生成并检查约 50 张图像。然而，独自工作时，这位用户表示至少会检查 200 张图像，并指出一次要让团队对超过 50 张图像达成共识很困难。

我们从这部分研究中确定了四类关键评估任务：图像多样性、图像质量、模式崩溃，以及一个“综合”类别，涵盖对类别和样本的其他探索。

6.2 第二阶段：观察 Ravel 的使用方式
本阶段旨在确定如何使用 Ravel 用户界面完成第一阶段确定的关键评估任务。确认参与后，我们向用户发送了一份简短的幻灯片和一段介绍用户界面组件的视频，并请他们在研究环节开始前查看。每场研究环节持续一小时，包括围绕任务对工具进行探索，以及回答半结构式访谈问题。

6.2.1 任务一：多样性
我们要求用户尝试完成的第一个任务，是判断 BigGAN 或 BigGAN-deep 在海胆图像多样性方面表现得更好；该设置与第 5.0.2 节所述的 Ravel 实例相仿。此实例展示了两个模型生成的 128x128 像素图像，每个模型有 50,000 张图像，且每个模型的每个类别包含 50 张图像。用户界面的分辨率设置为让每位用户看到相同数量的图像：每个数据划分中每行 5 张图像。

用户看到的是一个 Ravel 实例，其中在 高亮类别菜单中选中了“海胆”类别，左侧数据划分和右侧数据划分分别显示 BigGAN 和 BigGAN-deep 的结果。初始加载时，该实例显示了包含最多海胆图像的簇：BigGAN 中有 39 只（共 50 只），BigGAN-deep 中有 45 只（共 50 只）。

所有用户都先关注每个数据划分中的图像。没有用户在第一步查看指标或询问指标。这凸显了 Ravel 直观呈现评估图像多样性所需部分最重要信息的重要性：在感兴趣的簇中醒目地展示样本图像网格。三位用户立即注意到每个数据划分中的样本数量，并密切关注哪个模型在感兴趣的簇中拥有更多样本。最常见的流程是检查单张图像，并数数或“目测”每个模型中具有不同背景颜色、海胆姿态和海胆颜色的图像数量（所有用户都提到了这三项特征）。三位用户浏览了指标图表，并点击了其他突出显示的含有海胆的簇。五位用户都认定在海胆图像多样性方面，BigGAN-deep 的表现更好，其中一位用户明确表示，这证实了他们先前的预期。这表明用户使用 Ravel 完成这一关键任务的方式具有一致性，也展现了在定性视觉评估中实现决策趋同和工作流程操作化的潜力。值得注意的是，Ravel 帮助一位用户确认了他们先前对 BigGAN-deep 多样性表现的预期，不过这也可能成为影响其评估过程的潜在偏见因素。

6.2.2 任务二：质量
在第二项任务中，用户需要判断 BigGAN 或 BigGAN-deep 在金毛寻回犬图像质量方面表现更好。用户看到的 Ravel 实例与任务一相同，但这次在高亮类别菜单中选中了“金毛寻回犬”类别，并选中了显示大多数金毛寻回犬图像的簇：BigGAN 数据划分中有 46 张（共 50 张），BigGAN-deep 数据划分中也有 46 张（共 50 张）。

所有用户都立即注意到两个模型生成的图像中存在许多伪影。所有用户都点击了单张图像，并指出了具体问题，例如狗鼻子的形状、腿的数量，以及每个样本中狗的形态和姿势是否准确。最常见的工作流程是数一数或“目测”每个数据划分中的伪影数量，以做出初步判断；但在任务的其余部分，不同用户的评估工作流程明显更加多样。五位用户表示，这项任务比多样性任务更难，而一位用户表示更容易。一些用户指出，在他们通常的工作流程中，每个模型的 FID 和 Inception 分数是做出这一判断的重要依据。五位用户中有四位认定，就金毛寻回犬图像的质量而言，BigGAN-deep 的表现也更好；另有一位用户没有明确判断这项任务。再一次，Ravel 使所有用户在主要的图像质量评估策略上保持一致，但在做出这一初步决定后，各自的探索方式有所不同。一些用户通过查看召回率和精确率图表来验证自己的视觉判断，但其他用户只是浏览图表，并未对任务形成进一步看法。

6.2.3 任务三：模式崩溃
在质量任务之后，我们询问用户是否熟悉“模式崩溃”这一术语，以及他们是否会用它来评估生成模型输出。五位用户中有四位熟悉模式崩溃，并表示在模型评估过程中发现这一问题很有帮助；另有一位用户不熟悉这个术语。随后，我们请熟悉模式崩溃的用户使用 Ravel，判断任一模型是否在任何类别中发生了模式崩溃。

由于这项任务不局限于单一类别，因此更加开放，也更难让用户作出判断。由于没有人指导用户如何完成任务或作出判断，不同用户的工作流程差异很大。几位用户提到，在第一项任务中观察到 BigGAN-deep 的多样性更高后，他们原以为 BigGAN 会出现更严重的模式崩溃。四位发现了模式崩溃具体实例（例如图6）的用户，是通过选择召回率图表中数值最小的簇来找到的。五位用户中有四位在探索过程中曾查看图7中展示的簇样本，但其中一位没有将其识别为模式崩溃，另有一位完全没有选择该簇。尽管如此，

这是一个很有希望的观察结果，也进一步证明了 Ravel 的易用性，以及召回率可视化的独特实用价值。多数人一致认为，召回率图表可用于识别模式崩溃；研究还发现，模式崩溃在 BigGAN-deep 中更常见。

6.2.4 任务四：进一步探索类别、样本和特征

在最终任务中，用户需要查看 Ravel 的一个实例：右侧数据划分显示基于StyleGAN2的模型生成的图像，左侧数据划分显示其真实训练数据 FFHQ 中的图像。用户被告知可以自由探索该界面，既可以重复进行质量和多样性评估，也可以尝试自己感兴趣的新任务。

这项任务完全没有引导，但大多数用户都从评估基于StyleGAN2的模型的图像质量开始。用户通常会选择精确率分布两端的簇。一位用户观察到，点击一个精确率较高的簇后，会看到“典型的 FFHQ 图像……图像质量很好，没有遮挡，人物面朝镜头，展示的是直发，而模型的输出与这些图像非常相似。”五位用户中有四位发现并明确指出，基于StyleGAN2的模型难以生成脸部彩绘的人物图像。他们点击精确率较低的簇，注意到许多生成图像上存在伪影，从而得出了这一发现。这四位用户在发现模型这一新问题时都表示惊讶。这表明，即使没有具体提示，许多用户也会通过 Ravel 做出相同类型的发现，并遵循相同类型的评估流程。我们的一位参与者曾参与训练我们所用基于StyleGAN2的模型的团队，并发现该模型无法生成戴着某种款式的毛绒冬帽的人脸，而这种帽子在真实数据中相当常见。他表示，自己此前并不知道模型存在这一局限，因此能够发现这一点让他感到惊喜。

其他常见任务是对簇和嵌入空间进行“语义探索”。五位用户中有四位查看了 FFHQ 数据划分中的图像，以判断这些簇在语义上是否有意义，并探索 FFHQ 的多样性。两位用户思考嵌入空间是否能有效捕捉人类会归为一类的事物。这些用户考察了样本 UMAP 图表中样本之间的邻近关系，并由此认定 ImageNet 特征空间并不适合人脸聚类，因为他们无法在某些簇中发现相邻样本之间稳定的视觉相似性。

6.2.5 专家反馈讨论

总体而言，用户对该工具印象深刻。五位用户都认为 Ravel 能融入他们当前的评估流程，而且对发表生成模型研究成果的研究人员尤其有用。此处，我们总结了第二阶段反思环节中的其他反馈。

两位用户表示，他们的探索使他们开始怀疑 InceptionV3 特征是否适用于人脸肖像生成评估。在反思中，其中一位用户指出，Ravel 可用于了解特征空间本身，这对生成图像模型评估具有广泛的应用价值。

在发现 BigGAN-deep 中存在模式崩溃后，一位用户表示，Ravel 可用于探查最先进的模型，了解它们无法生成哪些类别的图像，这或许能指出研究人员可重点改进的系统性失效模式。

一位用户指出，与使用 Colab 相比，Ravel 能帮助他们的团队更快地得出有关模型性能的结论，尤其是在理解图像的多样性方面。他们解释说，使用 Colab 需要开发定制的可视化工具，并手动计算每个簇的 FID；而在 Ravel 用户界面中，同类有用信息则可以直接获取。两位用户强调，Ravel 有助于将研究人员评估质量和多样性的方式落到实处。其中一位解释说，Ravel 可以成为“推动结果采用标准用户界面/流程的力量”，并认为这样一来，“与其他团队成员或非技术人员分享结果会更容易……甚至可以分享给没时间运行我的代码的经理”。这让我们确信，像 Ravel 这样的定制专用界面确有用武之地。虽然它们不如 Colab 灵活，但更适合研究人员执行常见的评估任务。

总体而言，第二阶段的专家反馈积极而热烈，几位用户表示期待继续探索 Ravel，并将其纳入自己的工作流程。

7 局限性与未来工作

我们的定性用户研究由一家公司的少量领域专家参与者完成，因此可能无法在外部效度上代表所有研究人员使用生成图像模型的体验。研究环节对每位用户也限制在一小时，每项评估任务的时间不超过 15 分钟。若能延长工具的使用时间，或许可以开展更丰富、更多样的交互并获得新的发现。

两位用户希望查看更高分辨率的图像，另有两位用户在检查图像是否存在伪影时，希望查看图像的原始分辨率；这并非 Ravel 本身固有的问题，但却是未来设计中一项重要的设计可供性。两位用户希望在查看每个数据划分中的图像后能够为其做标记；启用此功能将有助于支持我们观察到参与者在完成任务时采用的基于“计数”的工作流程。

使用类别条件模型时，用户最初表示，他们更希望在同一视图中查看某个类别的所有图像（即按类别标签进行聚类），以便对该类别作出判断。然而，经过进一步探索，他们指出，将某个标签的“离群”图像与其他最相似的图像放在一起查看也很有用。这表明，对于类别条件模型，这两种工作流程都很重要，应由 Ravel 这类工具予以支持。

作者还指出，在观察用户使用该工具时，我们发现了系统的一些局限：

用户并不总能理解簇的“含义”（即，无法构建合理解释，说明一组图像为何会聚在一起）。这是聚类等无监督方法普遍存在的问题，但我们发现，用户往往会尝试为每个簇附上某种具有语义意义的描述，以此作为比较的依据。

探索如何“描述”簇或概括簇之间的差异，可能是有助于用户理解的一项重要未来工作。我们还认为，探索其他聚类方法，尤其是层次方法，可能是生成不同粒度簇的特别有吸引力的途径，从而帮助人们理解数据集中的各个群组。

在介绍样本查看器（第 4.2.3 节）时，我们指出了一项局限：如果簇中包含大量图像，就无法在一个屏幕上显示所有图像。虽然一种缓解方法是增加簇的数量，以缩小每个簇的规模，但我们认为，未来的工作可以探索更好的方法，让人们在一个视图中直观把握整个簇。或许可以借鉴 Activation Atlas [47], 中介绍的技术，在簇内将非常相似的图像组成图像堆叠；也可以采用其他子采样或排序方法，确保在单个屏幕上展示一个簇中尽可能丰富多样的内容。

Ravel 无法直接支持的一项用户任务是检测记忆。有用户指出，在界面中加入实时最近邻搜索，可能会让它适用于这项任务。

8 结论

我们介绍了 Ravel，这是一款可视分析工具，能够帮助研究人员对生成式模型输出进行大规模定性评估。我们的主要贡献包括：

- 一个可视分析系统，支持多种评估任务（例如评估质量和多样性、发现模式崩溃或模型输出中的缺失），适用于生成式图像模型，且不受模型架构和内部机制的限制。
- 对大型生成图像模型数据集进行交互式探索，借助聚类以及细粒度簇指标可视化来指导定性评估。
- 一种用户界面，利用语义上有意义的嵌入空间驱动可视比较，以支持对图像分布差异的推理，并提出有关模型行为的假设。

我们研究中的专家用户能够持续获得关于模型行为的见解，包括发现模型未能捕捉真实数据分布的部分，例如基于StyleGAN2的模型中的面部彩绘或某些类型的头饰，以及BigGAN-deep中的模式崩溃。这类见解无法仅通过查看 FID 等定量指标获得。

我们的研究参与者证实了我们的假设：单一数值指标不足以全面衡量模型性能。除了以更高粒度考察指标外，Ravel 还允许用户以更细的粒度探索指标，从而揭示模型输出中召回率或精确率等指标未能反映生成图像问题的部分。用户推测，无论是在我们的工具中，还是在该领域主要采用的指标中所使用的底层 InceptionV3 嵌入，可能都无法关注到某些人类一眼就能发现的视觉伪影。我们认为，未来沿这一方向开展的研究有望帮助人们更好地理解嵌入空间本身的局限，以及这些局限如何影响指标和使用这些指标的工作流程。

致谢

作者感谢 Mario Lucic、Marvin Ritter、Ben Poole、Han Zhang、Chitwan Saharia、James Wexler 和 Lucas Dixon 在本研究中提供的帮助与反馈。我们也感谢参与者提供的反馈。

参考文献

[1] I. Goodfellow、J. Pouget-Abadie、M. Mirza、B. Xu、D. Warde-Farley、S. Ozair、A. Courville 和 Y. Bengio，“生成对抗网络”，载于《神经信息处理系统进展》，第 2014 [ ] 27, 卷。在线。可获取：https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html [2] D. P. Kingma 和 M. Welling，“自编码变分贝叶斯”，2013年12月。 [在线]。可获取：https://arxiv.org/abs/ 1312.6114v10[3] L. Dinh、D. Krueger 和 Y. Bengio，“NICE：非线性独立成分估计”，arXiv:1410.*8516 [cs],* 2015年4月，arXiv：1410.8516。 [在线]。可获取：http://arxiv.org/abs/ 1410.8516[4] J. Sohl-Dickstein、E. Weiss、N. Maheswaranathan 和 S. Ganguli，“使用非平衡热力学进行深度无监督学习”，载于第32届国际机器学习大会论文集，2015年6月，第2256–2265页，ISSN：1938-7228。[在线]。可获取：https://proceedings.mlr.press/v37/sohl-dickstein15.html [5] A. Odena，“关于生成对抗网络的开放问题”，Distill，第4卷，第4期，文章e18，2019年4月。 [在线]。可获取：https://distill.pub/2019/gan-open-problems [6] L. Theis、A. v. d. Oord 和 M. Bethge，“生成模型评估札记”，arXiv:1511.*01844 [*cs，stat*],* 2016年4月，arXiv：1511 01844 [ ] 1511 01844 。在线。[7] 可获取：http://arxiv.org/ 。A. Borji，“GAN评估指标的利弊”，arXiv:1802.*03446 [*cs*],* 2018年10月。[8] ——，“GAN评估指标的利弊：新进展”，arXiv:2103.*09396 [*cs*],*2021年3月。[9] M. Heusel、H. Ramsauer、T. Unterthiner、B. Nessler 和 S. Hochreiter，“采用双时间尺度更新规则训练的GAN收敛到局部纳什均衡”，载于《神经信息处理系统进展》，第30卷，2017年。[在线]。可获取：https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html [10] M. S. M. Sajjadi、O. Bachem、M. Lucic、O. Bousquet 和 S. Gelly，“通过精确率和召回率评估生成模型”，载于《神经信息处理系统进展》，第31卷，2018年。[在线]。可获取：https://proceedings.neurips.cc/paper/2018/hash/f7696a9b362ac5a51c3dc8f098b73923-Abstract.html [11] T. Kynk¨a¨anniemi、T. Karras、S. Laine、J. Lehtinen 和 T. Aila，“用于评估生成模型的改进精确率和召回率指标”，arXiv:1904.*06991 [*cs，stat*],* 2019年10月，arXiv：1904.06991。 [ ] 1904 06991 在线。[12] 可获取：http://arxiv.org/abs/ 。A. Brock、J. Donahue 和 K. Simonyan，“用于高保真自然图像合成的大规模GAN训练”，arXiv:1809.*11096 [*cs，stat*],* 2019年2月，arXiv：1809.11096。 [在线]。可获取：http://arxiv.org/abs/1809 11096 。T. Karras、S. Laine 和 T. Aila，“基于风格的生成对抗网络生成器架构”，arXiv:1812.*04948 [*cs，stat*],* 2019年3月，arXiv：1812.04948。 [在线]。可获取：http://arxiv.org/abs/1812 04948 。D. P. Kingma 和 P. Dhariwal，“GLOW：具有可逆1×1卷积的生成流”，arXiv:1807.*03039 [*cs，stat*],* 2018年7月，arXiv：1807.03039。 [在线]。可获取：http://arxiv.org/abs/1807. 03039[15] A. v. d. Oord、O. Vinyals 和 K. Kavukcuoglu，“神经离散表示学习”，arXiv:1711.*00937 [*cs*],* 2018年5月，arXiv： 1711 00937 [ ] 1711 00937 。[16] 在线。可获取：http://arxiv.org/ 。J. Ho、C. Saharia、W. Chan、D. J. Fleet、M. Norouzi 和 T. Salimans，“用于高保真图像生成的级联扩散模型”，arXiv:2106.*15282 [*cs*],* 2021年12月，arXiv：2106.15282。 [ ] 2106 15282 在线。可获取：[17] http://arxiv.org/ 。C. Szegedy、V. Vanhoucke、S. Ioffe、J. Shlens 和 Z. Wojna，“重新思考计算机视觉中的Inception架构”，arXiv:1512.*00567 [*cs*],* 2015年12月。[18] T. Karras、S. Laine、M. Aittala、J. Hellsten、J. Lehtinen 和 T. Aila，“分析和改进StyleGAN的图像质量”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第8110–8119页。[19] R. Geirhos、P. Rubisch、C. Michaelis、M. Bethge、F. A. Wichmann 和 W. Brendel，“经过ImageNet训练的卷积神经网络偏向纹理；提高形状偏差可改善准确率和鲁棒性”，arXiv:1811.*12231 [*cs，q-bio，stat*],* 2019年1月。[20] T. Salimans、I. Goodfellow、W. Zaremba、V. Cheung、A. Radford、X. Chen 和 X. Chen，“改进的GAN训练技术”，载于《神经信息处理系统进展》，第29卷，2016年。

[在线]。可获取：https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html[21] S. Barratt 和 R. Sharma，“Inception分数札记”，arXiv:1801.*01973 [*cs, stat*]*, 2018年6月，arXiv：1801.01973。 [在线]。可获取：http://arxiv.org/abs/ 1801 01973 。S. [22] Menon、A. Damian、S. Hu、N. Ravi 和 C. Rudin，“PULSE：通过生成模型的潜在空间探索进行自监督照片上采样”，载于2020年IEEE/CVF计算机视觉与模式识别会议（CVPR），美国华盛顿州西雅图，2020年6月，第2434–2442页。[23] E. Denton、S. Chintala、A. Szlam 和 R. Fergus，“使用拉普拉斯金字塔对抗网络的深度生成图像模型”，arXiv:1506.*05751 [*cs*]*, 2015年6月。[24] S. Zhou、M. Gordon、R. Krishna、A. Narcomey、L. F. Fei-Fei 和 M. Bernstein，“HYPE：生成模型的人类视觉感知评估基准”，第13页。[25] D. Bau、J.-Y. Zhu、H. Strobelt、B. Zhou、J. B. Tenenbaum、W. T. Freeman 和 A. Torralba，“GAN剖析：可视化并理解生成对抗网络”，arXiv: *1811 10597* 。*[ ]*, 2018 cs，12月。[26] D. Bau、J.-Y. Zhu、J. Wulff、W. Peebles、H. Strobelt、B. Zhou 和 A. Torralba，“观察GAN无法生成的内容”，arXiv:1910.*11626 [*cs，eess*]*, 2019年10月。[27] T. White，“生成网络采样”，arXiv:1609.*04468 [*cs，stat*]*, 2016年12月。[28] T. White 和 B. Loh，“SpaceSheet：使用电子表格界面导航概念空间”，第8页。[29] D. Smilkov、N. Thorat、C. Nicholson、E. Reif、F. B. Vi´egas 和 M. Wattenberg，“嵌入投影器：嵌入的交互式可视化与解释”，arXiv:1611.*05469 [*cs，stat*]*, 2016年11月。[30] K. Ho，“字体地图”，https://experiments.withgoogle.com/font- map。[31] C. Diagne 和 S. Doury，“策展人表格”，https://experiments.withgoogle.com/curator-table。[32] K. McDonald、M. Tan 和 Y. Mann，“鸟鸣”，https://experiments.withgoogle.com/bird-sounds。[33] Y. Liu、E. Jun、Q. Li 和 J. Heer，“潜在空间制图：向量空间嵌入的可视化分析”，《计算机图形学论坛》，第38卷，第3期，第67–78页，2019年6月。[34] S. Xiang、X. Ye、J. Xia、J. Wu、Y. Chen 和 S. Liu，“错误标记训练数据的交互式纠正”，载于2019年IEEE可视分析科学与技术会议（VAST），2019年10月，第57–68页。[35] R. Zhang、P. Isola、A. A. Efros、E. Shechtman 和 O. Wang，“深度特征作为感知度量的非凡有效性”，2018年，第586–595页。 [在线]。可获取：https://openaccess.thecvf.com/content cvpr 2018/html/ Zhang The Unreasonable Effectiveness CVPR 2018 paper.html [36] “Scikit-learn聚类概览”。[在线]。可获取：https://scikit-learn.org/stable/modules/clustering.html#overview-of-clustering-methods [37] L. McInnes、J. Healy、N. Saul 和 L. Grossberger，“UMAP：统一流形近似与投影”，《开放源代码软件期刊》，第3卷，第29期，文章861，2018年。[38] K. P. F.R.S.，“论空间中点集系统的最佳拟合直线与平面”，《伦敦、爱丁堡和都柏林哲学杂志与科学期刊》，第2卷，第11期，第559–572页，1901年。[39] L. Van der Maaten 和 G. Hinton，“使用t-SNE可视化数据”，《机器学习研究期刊》，第9卷，第11期，2008年。[40] A. Coenen 和 A. Pearce，“理解UMAP”，2019年。 [在线]。可获取：https://pair-code.github.io/understanding-umap/ [41] O. Russakovsky、J. Deng、H. Su、J. Krause、S. Satheesh、S. Ma、Z. Huang、A. Karpathy、A. Khosla、M. Bernstein、A. C. Berg 和 L. Fei-Fei，“ImageNet大规模视觉识别挑战赛”，《国际计算机视觉杂志》（IJCV），第115卷，第3期，第211–252页，2015年。[42] M. Lucic、K. Kurach、M. Michalski、S. Gelly 和 O. Bousquet，“GAN是否生而平等？一项大规模研究”，载于《神经信息处理系统进展》，第2018 [ ] 31, 卷。在线。可获取：https://proceedings.neurips.cc/paper/2018/hash/e46de7e1bcaaced9a54f1e9d0d2f800d-Abstract.html[43] I. Goodfellow，“NIPS 2016教程：生成对抗网络”，arXiv:1701.*00160 [*cs*]*, 2017年4月。[44] “PULSE模型的经验与讨论”，https://thegradient.pub/pulse-lessons/，2020年6月。[45] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Good- fellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow：异构分布式系统上的大规模机器学习，”*arXiv:1603.04467 [cs],* 2016年3月。[46] T. Kluyver, B. Ragan-Kelley, F. P´erez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing，“Jupyter 笔记本——一种用于可复现计算工作流的出版格式”，载于《学术出版中的定位与权力：参与者、代理人与议程》，F. Loizides 和 B. Schmidt 编。IOS Press，2016年，第87–90页。[47] S. Carter, Z. Armstrong, L. Schubert, I. Johnson, and C. Olah，“Activation Atlas”，*Distill*，第4卷，第3期，e15页，2019年3月。 [在线]。网址：https://distill.pub/2019/activation-atlas