LABEL-EFFICIENT SEMANTIC SEGMENTATION WITH DIFFUSION MODELS

用扩散模型进行标签有效的语义分割

摘要

去噪扩散概率模型最近受到了很多研究的关注，因为它们的性能优于其他方法，如GANs，并且目前提供了最先进的生成性能。扩散模型的优越性能使它们在一些应用中成为有吸引力的工具，包括绘画、超级分辨率和语义编辑。在本文中，我们证明了扩散模型也可以作为语义分割的工具，特别是在标记数据稀缺的情况下。特别是，对于几个预训练的扩散模型，我们研究了执行反向扩散过程的马尔科夫步骤的网络的中间激活。我们表明，这些激活有效地捕捉了输入图像的语义信息，并且似乎是分割问题的优秀像素级表征。基于这些观察，我们描述了一种简单的分割方法，即使只提供少量的训练图像也能发挥作用。在相同数量的人工监督下，我们的方法在几个数据集上的表现明显优于现有的替代方法。这个项目的源代码是公开的。

代码链接：<https://github.com/yandex-research/DDPM-segmentation>

1 引言

去噪扩散概率模型（扩散模型）（Sohl-Dickstein等，2015；Ho等，2020）最近在单个样本的真实性和多样性方面都优于其他方法来模拟自然图像的分布（Dhariwal & Nichol, 2021）。扩散模型的这些优势在一些应用中被成功利用，如着色（Song等人，2021年）、绘画（Song等人，2021年）、超级分辨率（Saharia等人，2021年；Li等人，2021b）和语义编辑（Meng等人，2021年），与GANs相比，扩散模型经常取得更令人瞩目的结果。然而，到目前为止，扩散模型还没有被作为有效的图像表征的来源来利用，以解决鉴别性的计算机视觉问题。虽然之前的文献已经证明了各种生成范式，如GANs（Donahue & Simonyan，2019）或自回归模型（Chen et al.，2020a），可以用来提取常见视觉任务的表征，但目前还不清楚扩散模型是否也能作为表征学习者。在本文中，我们在语义分割的背景下对这个问题提供了肯定的答案。特别是，我们研究了来自U-Net网络的中间激活，该网络近似于扩散模型中反向扩散过程的马尔可夫步骤。直观地说，这个网络学会了对其输入进行去噪，而且不清楚为什么中间激活应该捕捉高级视觉问题所需的语义信息。然而，我们表明，在某些扩散步骤上，这些激活确实捕获了这样的信息，因此，有可能被用作下游任务的图像表示。鉴于这些观察，我们提出了一个简单的语义分割方法，该方法利用了这些表征，即使只提供几张标记的图像也能成功地工作。在几个数据集上，我们表明我们基于扩散模型的分割方法在相同数量的监督下优于现有的基线。总而言之，我们的论文的贡献是。  
1. 我们研究了由最先进的扩散模型学习的表征，并表明它们捕捉到了对下游视觉任务有价值的高层次语义信息。  
2. 我们设计了一个简单的语义分割方法，利用这些表征，在几张照片的操作点上优于其他方法。  
3. 我们在相同的数据集上将基于扩散模型的表征与基于GAN的表征进行比较，并证明前者在语义分割方面的优势。

2 相关工作   
在本节中，我们简要介绍与我们工作相关的现有研究思路。  
**扩散模型**（Sohl-Dickstein等人，2015；Ho等人，2020）是一类生成模型，它通过马尔可夫链的端点来近似真实图像的分布，而马尔可夫链来源于一个简单的参数分布，通常是一个标准高斯。每个马尔可夫步骤由一个深度神经网络建模，该网络有效地学习用一个已知的高斯核来反转扩散过程。Ho等人强调了扩散模型和分数匹配的等同性（Song & Ermon，2019；2020），表明它们是通过迭代去噪过程将简单的已知分布逐渐转化为目标分布的两个不同视角。最近的工作（Nichol, 2021; Dhariwal & Nichol, 2021）开发了更强大的模型架构以及不同的高级目标，这使得扩散模型在生成质量和多样性方面 "战胜 "了GANs。扩散模型已被广泛应用于多个领域，包括图像着色（Song等人，2021）、超分辨率（Saharia等人，2021；Li等人，2021b）、画中画（Song等人，2021）和语义编辑（Meng等人，2021）。在我们的工作中，我们证明人们也可以成功地将其用于语义分割。  
**用生成模型进行图像分割**是目前一个活跃的研究方向，然而，现有的方法主要是基于GANs的。第一线工作（Voynov & Babenko, 2020; Voynov et al., 2021; Melas-Kyriazi et al., 2021）是基于最先进的GANs的潜在空间有对应于影响前景/背景像素不同的方向的证据，这可以产生合成数据来训练分割模型。然而，这些方法目前只能进行二元分割，尚不清楚它们是否可以用于语义分割的一般设置中。第二类工作（Zhang等人，2021；Tritrong等人，2021；Xu，2021；Galeev等人，2020）与我们的研究更为相关，因为它们是基于GANs中获得的中间表示。特别是，（Zhang等人，2021）提出的方法在这些表征上训练了一个像素类预测模型，并证实了其标签效率。在实验部分，我们将(Zhang et al., 2021)的方法与我们基于扩散模型的方法进行了比较，并展示了我们解决方案的几个独特优势。  
**来自生成模型的表征用于鉴别性任务**。生成模型的使用，作为表示学习者，在全局预测（Donahue和Simonyan，2019；Chen等人，2020a）和密集预测问题（Zhang等人，2021；Tritrong等人，2021；Xu，2021；Xu等人，2021）已被广泛研究。虽然以前的工作强调了这些表征的实际优势，如分布外的鲁棒性（Li等人，2021a），但与其他无监督的方法相比，如基于对比学习（Chen等人，2020b），生成模型作为表征学习者受到的关注较少。主要原因可能是在复杂多样的数据集上训练一个高质量的生成模型的难度。然而，鉴于最近扩散模型在Imagenet上的成功（Deng等人，2009），我们可以预期这个方向在未来会吸引更多的关注。

3 扩散模型的表征

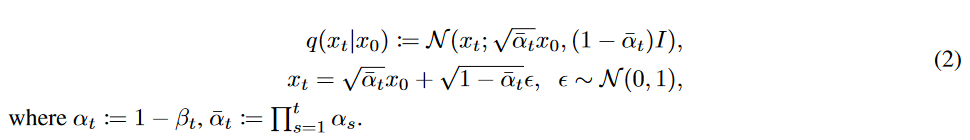
在下一节中，我们研究了由扩散模型学习的图像表征。首先，我们对扩散模型框架做一个简要的概述。然后，我们描述如何用扩散模型提取特征，并研究这些特征可能捕捉到什么样的语义信息。

**背景（扩散模型是啥）**。扩散模型将噪声xT∼N（0，I）转化为样本x0，方法是将xT逐渐去噪为较少噪声的样本xt。从形式上看，我们得到了一个前向扩散过程。



对于一些固定的方差表β1, . . . , βt。

重要的是，可以直接从数据x0中获得噪声样本xt。



预训练的扩散模型近似于一个反向过程。



在实践中，噪声预测器网络θ(xt, t)不是预测方程（3）中分布的平均值，而是预测步骤t的噪声成分；然后平均值是这个噪声成分和xt的线性组合。协方差预测器Σθ(xt, t)可以是一组固定的标量协方差，也可以是学习的（后者被证明可以提高模型质量（Nichol, 2021））。去噪模型θ(xt, t)通常由UNet架构的不同变体提供参数（Ronneberger等人，2015），在我们的实验中，我们研究了（Dhariwal & Nichol, 2021）中提出的最先进的模型。  
**提取表征**。对于一个给定的真实图像x0∈RH×W×3，人们可以从噪声预测网络θ（xt，t）中计算出T组激活张量。图1展示了一个时间步长t的整体方案。

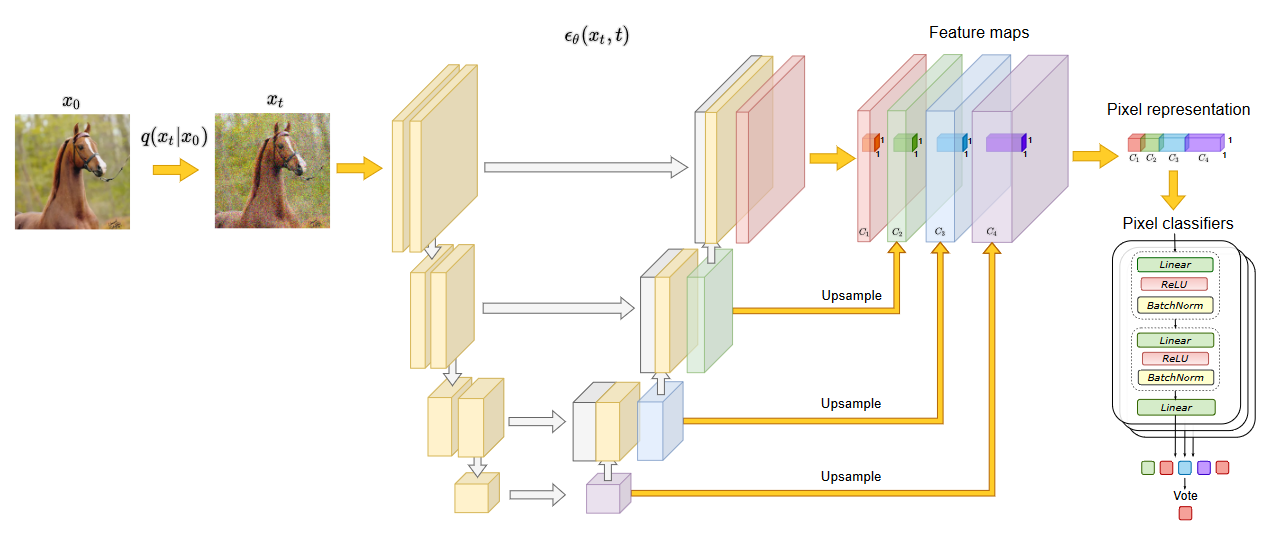
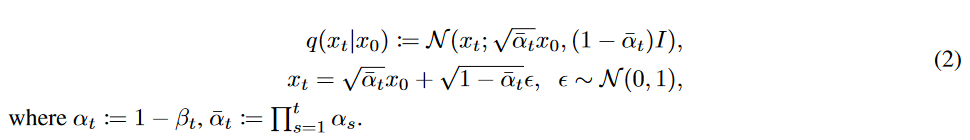


图1：提议的方法概述。(1) 根据q(xt|x0)加入噪声，x0→xt。(2) 从噪声预测器θ(xt, t)中提取特征图。(3) 通过将特征图上采样到图像分辨率，并将其连接起来，收集像素级表示。(4) 使用像素级的特征向量来训练一个MLP的集合，以预测每个像素的类别标签。

首先，根据公式（2），



我们通过添加高斯噪声来破坏x0。噪声的xt被用作θ(xt, t)的输入，其参数由UNet模型确定。然后，UNet的中间激活被双线性插值上采样到H×W。这允许将它们视为x0的像素级表示。

3.1 表征分析

我们分析了噪声预测器θ(xt, t)对不同t产生的表征。我们考虑在LSUN-Horse和FFHQ-256数据集上训练的最先进的扩散模型检查点。

**噪声预测器的中间激活捕捉了语义信息**。在这个实验中，我们从LSUN-Horse和FFHQ数据集中抽取一些图像，并将每个像素分别手动分配到21个和34个语义类别中的一个。我们的目标是了解由扩散模型产生的像素级表示是否有效地捕捉了语义信息。为此，我们训练一个多层感知器（MLP），从18个UNet解码器块之一在特定扩散步骤t上产生的特征中预测像素语义标签。MLPs在20幅图像上进行训练，并在20幅保留图像上进行评估。预测性能是以平均IoU来衡量的。

图2显示了不同区块和扩散步骤t的预测性能的演变。块的编号是从深到浅的。图2显示，噪声预测器θ(xt, t)产生的特征的可分辨性在不同的区块和扩散步骤中是不同的。特别是，对应于反向扩散过程的后期步骤的特征通常能更有效地捕捉语义信息。相比之下，与早期步骤相对应的特征通常是没有信息的。在不同的区块中，UNet解码器中间的层所产生的特征似乎在所有的扩散步骤上都是最有信息的。另外，我们根据注释数据集中的平均面积，分别考虑小尺寸和大尺寸的语义类别。然后，我们在不同的UNet块和扩散步骤中独立评估这些类别的平均IoU。LSUN-Horse的结果见图3。正如预期的那样，在反向过程中，大尺寸物体的预测性能开始提前增长。较浅的区块对较小的物体来说信息量更大，而较深的区块对较大的物体来说则更大。在这两种情况下，最具鉴别力的特征仍然对应于中间的块。（深层的block对大物体更敏感，浅层的对小物体更敏感）。图2意味着，对于某些UNet块和扩散步骤，类似的基于扩散模型的表示方法对应于相同语义的像素。图4显示了FFHQ检查点从扩散步骤{50、200、400、600、800}上的块{6、8、10、12}中提取的特征所形成的k-means聚类（k=5），并证实了聚类可以跨越连贯的语义对象和对象-部分。

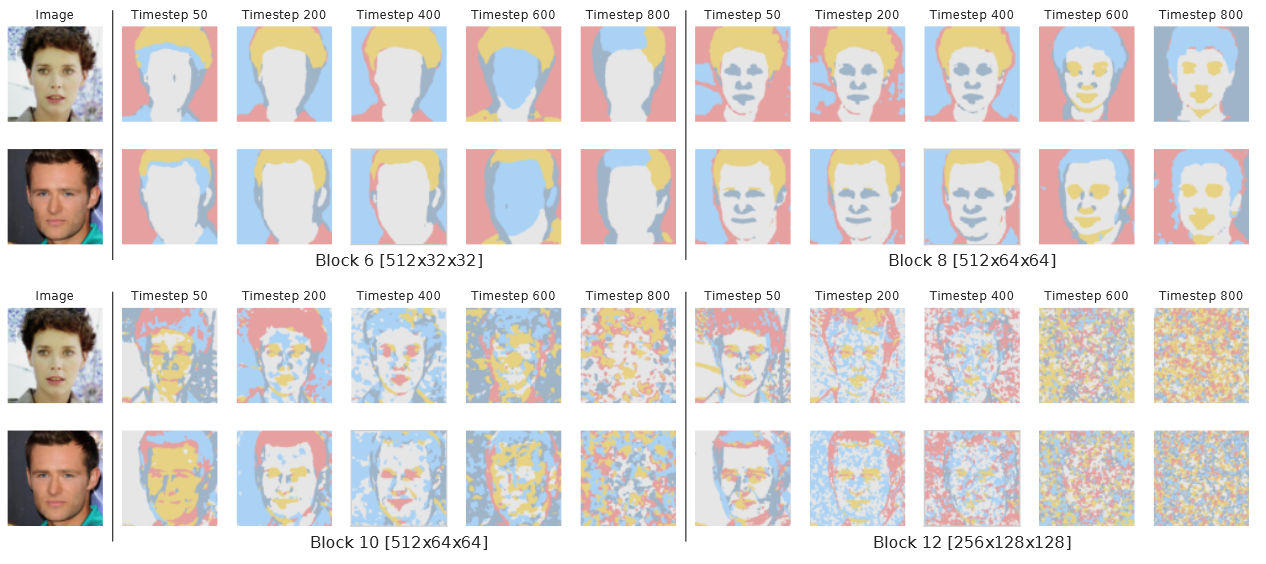
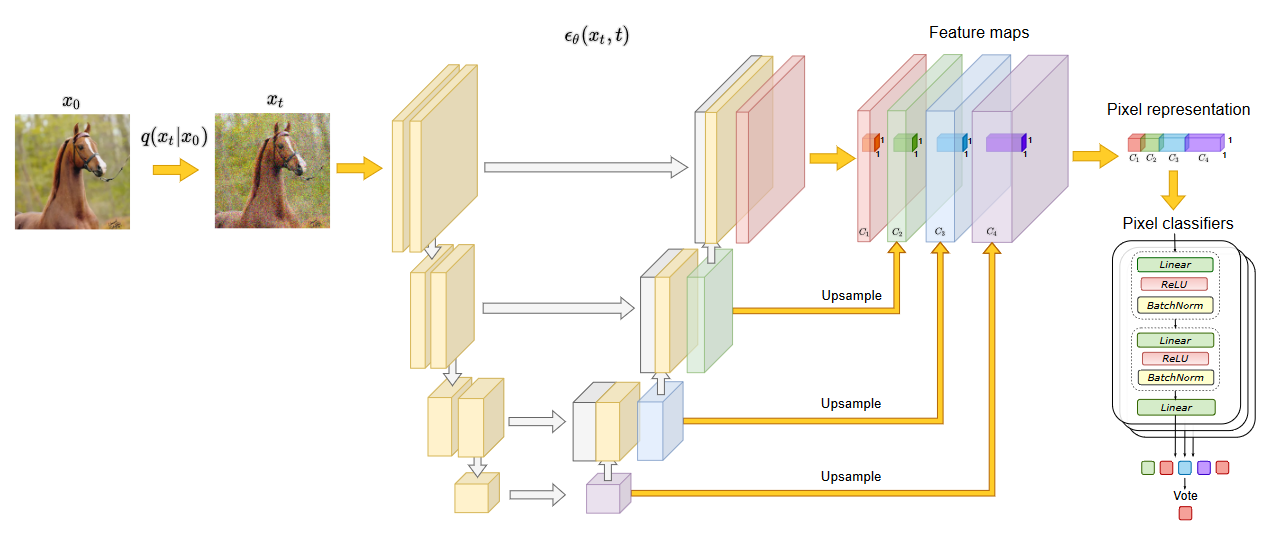


图4：由UNet解码器块{6，8，10，12}在扩散步骤{50，200，400，600，800}上提取的特征形成的k-means聚类的例子（k=5）。来自中间块的聚类在空间上跨越了连贯的语义对象和部分。

在B=6的区块中，特征对应于粗略的语义掩码。在另一个极端，来自B=12的特征可以区分细粒度的人脸部分，但对粗粒度的碎片表现出较少的语义性。在不同的扩散步骤中，最有意义的特征对应于后来的特征。我们把这种行为归因于这样一个事实：在反向过程的早期步骤中，扩散模型样本的整体结构还没有出现，因此，在这个阶段几乎不可能预测分割掩码。图4中的掩码从质量上证实了这种直觉。对于t=800，掩码很难反映实际图像的内容，而对于较小的t值，掩码和图像在语义上是一致的。  
3.2 基于扩散模型的少数语义分割表征   
上面观察到的中间扩散模型激活的潜在有效性意味着它们被用作密集预测任务的图像表示。图1示意性地展示了我们用于图像分割的整体方法，它利用了这些表征的可辨别性。



更详细地说，我们考虑了一个几张照片的半监督设置，当大量的未标记的图像{X1, .... , XN } ⊂RH×W×3的特定领域，并且只对n个训练图像{X1, . . . , Xn} ⊂RH×W ×3的基础真理K类语义掩码{Y1, . . . , Yn} ⊂RH×W ×{1,...,K}被提供。

第一步，我们在整个{X1, ......, XN }上以无监督的方式训练一个扩散模型。然后，这个扩散模型被用来提取标签图像的像素级表示，使用UNet块的子集和扩散步骤t。在这项工作中，我们使用UNet解码器的中间块B={5，6，7，8，12}的表示，以及反向扩散过程的后期步骤t={50，150，250}。这些区块和时间步骤是由第3.1节的见解所激发的，但有意地没有为每个数据集进行调整。

虽然特定时间步骤的特征提取是随机的，但我们固定所有时间步骤t的噪声，并在第4.1节中消减。从所有区块B和步骤t中提取的表征被上采样为图像大小并连接起来，形成训练图像所有像素的特征向量。像素级表征的总维度为8448。然后，按照(Zhang et al., 2021)，我们在这些特征向量上训练一个独立的多层感知器（MLPs）集合，其目的是预测训练图像中每个像素的语义标签。我们采用（Zhang et al., 2021）的集合配置和训练设置，并在我们的实验中对所有其他方法加以利用，详见附录C。为了分割测试图像，我们提取其基于扩散模型的像素表示，并使用它们来预测集合的像素标签。最后的预测是通过多数投票获得的。

5 结论

本文证明了扩散模型可以作为鉴别性计算机视觉问题的表示学习者。与GANs相比，扩散模型允许直接计算真实图像的这些表征，人们不需要学习额外的编码器，将图像映射到潜在空间。这种扩散模型的优势和卓越的生成质量为少数照片的语义分割任务提供了最先进的性能。基于扩散模型的分割的显著限制是要求在手头的数据集上训练高质量的扩散模型，这对于复杂的领域，如ImageNet或MSCOCO，可能是一个挑战。然而，考虑到扩散模型的快速研究进展，我们预计它们将在最近的将来达到这些里程碑，从而扩大相应表示法的适用范围。

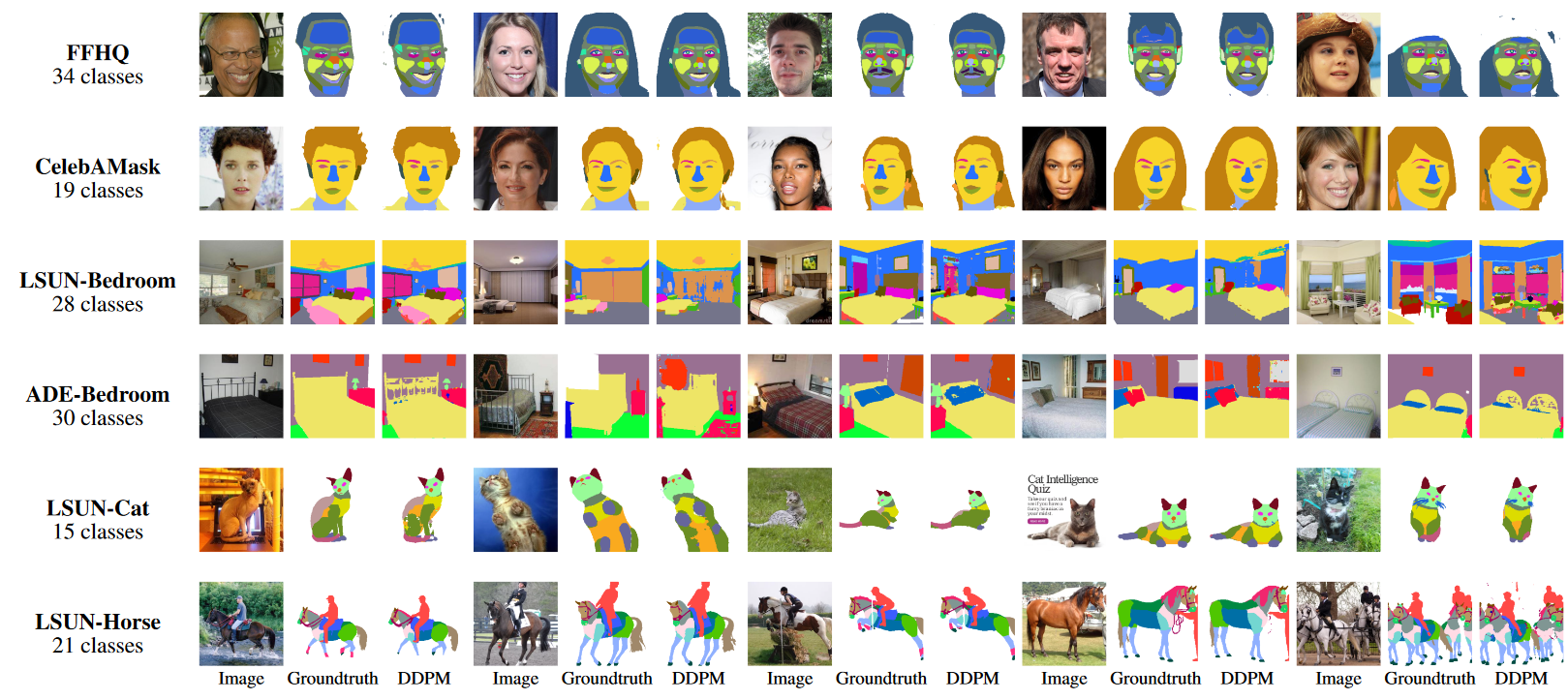


图5：我们的方法在测试图像上预测的分割掩码的例子，以及根据事实注释的掩码。