计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(8):80−92 [doi: 10.15888/j.cnki.csa.009920] [CSTR: 32024.14.csa.009920]

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

基于以物体为中心扩散的组成式场景建模①

沈知萌[1,2], 黄尹璇[1,2]

[1](复旦大学 计算机科学技术学院, 上海 200433)
[2](复旦大学 上海市智能信息处理重点实验室, 上海 200433)
通信作者: 沈知萌, E-mail: zmshen22@m.fudan.edu.cn

摘　要: 以物体为中心的学习方法旨在以组成式的方式对场景进行解析与建模, 并提取场景中物体的表示. 早期以物体为中心的学习方法通常使用简单的像素混合解码器来建模场景. 然而, 这些方法在处理复杂的合成数据集和真实世界数据集时通常表现不佳. 相比之下, 最近的一些以物体为中心的学习方法已经开始尝试使用结构更为复杂的解码器 (例如自回归 Transformer 和扩散模型) 来更有效地提取物体表示并建模场景. 尽管这些近期的方法相比于早期的方法具有更好的效果, 但这些方法采用的非组成式建模方法与人类的直觉相悖, 且它们无法根据物体的表示生成对应的物体图像. 为了解决这个问题, 本文提出了以物体为中心的扩散 (object-centric diffusion, OCD) 模型, OCD 使用一种改进的扩散模型作为解码器, 在重构场景的过程中分别生成物体的外观和掩码, 从而在保证模型效果的同时实现图像的组成式建模. 大量的实验证明, OCD 在多种数据集 (包括两个合成数据集和两个真实世界数据集) 上的图像分割和生成任务中表现出色, 证明了其普适性和有效性.

关键词: 以物体为中心的学习; 无监督学习; 组成式场景建模; 扩散模型; 生成模型

引用格式: 沈知萌,黄尹璇.基于以物体为中心扩散的组成式场景建模.计算机系统应用,2025,34(8):80–92. http://www.c-s-a.org.cn/1003-3254/9920.html

Compositional Scene Modeling with Object-centric Diffusion

SHEN Zhi-Meng[1,2], HUANG Yin-Xuan[1,2]

[1](School of Computer Science, Fudan University, Shanghai 200433, China)
[2](Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433, China)

Abstract: Object-centric learning methods aim to parse and model scenes in a compositional way while extracting representations of objects within those scenes. Early object-centric approaches typically employ simple pixel-mixing decoders for scene modeling. However, these methods often perform poorly when handling complex synthetic datasets and real-world datasets. In contrast, recent object-centric learning methods have begun experimenting with more complex decoders, such as autoregressive Transformers and diffusion models, to extract object representations and model scenes more effectively. Despite the improved performance of these newer methods over earlier ones, their non-compositional modeling approaches contradict human intuition and fail to generate corresponding object images given object representations. To address this issue, the proposed object-centric diffusion (OCD) model employs an improved diffusion model as a decoder. OCD generates the appearance and masks of objects separately during the scene reconstruction process, achieving true compositional modeling while maintaining model performance. Extensive experiments demonstrate that OCD excels in image segmentation and generation tasks across various datasets, including two synthetic and two real-world datasets, proving its versatility and effectiveness.

Key words: object-centric learning (OCL); unsupervised learning; compositional scene modeling; diffusion model; generative model

① 基金项目: 上海市科学技术委员会项目 (22511105000); 上海市类脑芯片与片上智能系统研发与转化功能型平台 (17DZ2260900)
收稿时间: 2024-12-23; 修改时间: 2025-02-12; 采用时间: 2025-02-24; csa 在线出版时间: 2025-06-13
CNKI 网络首发时间: 2025-06-16

近年来, 随着算力的大幅度提升与大量先进的人工智能算法的提出, 计算机视觉技术在各类工业生产与日常生活场景中都得到了广泛的应用. 例如: SAM[1]等分割大模型能够准确地对各类真实场景图像进行分割, 并且允许用户与系统进行交互来进一步提高分割的准确性; 一系列生成式人工智能系统[2–4]可以基于各类条件生成逼真的图片, 并且被广泛应用于各类插画绘制、游戏制作等内容创作领域中.

然而, 现有的这些最先进的方法大部分在训练的时候都需要海量的标注数据: SAM 从各个国家和地区的互联网中提取了数千万张图像以及十几亿掩码标注来作为训练数据集, 而在训练各类图像生成模型则需要各类文本或者图像标注来进行训练, 在绝大部分有标注互联网数据已经被用于模型训练的当下, 这无疑限制了人工智能的进一步发展. 相比之下, 人类在获得相似的场景解析或者场景构建的能力时则不需要如此大量的标注数据, 甚至对于没有语言能力的婴幼儿而言, 他们也不需要任何的外界提示信息即可对所处的环境进行初步的理解. 这种能力的一个关键因素在于人类以组成式的方式理解并构建场景: 人类在解析或者构建场景的时候, 并不会直接处理整个场景, 而是会以组成式的方式处理场景中的各个物体. 在理解或者构建完毕场景中的所有物体后, 实际上也就完成了对整个场景的理解或者构建. 相比直接处理整个场景, 以组成式的方式处理场景中的各个物体会在效率上会更有优势, 并且能够很好地应对由于物体不同所导致的组合爆炸问题.

为了使机器能够获得与人类类似的以组成式的方式处理场景的能力, 计算机视觉研究中发展出了以物体为中心的表示学习 (object-centric learning, OCL)[5]这一研究领域. 这一领域的方法大多使用组成式的自编码器, 从场景中提取各个物体的表示, 并以重构损失作为监督信号. 由于模型的训练通常是在弱监督或者无监督条件下进行的, 这一研究领域能够很大程度上缓解现有大模型过度依赖有标注数据的问题.

得益于 OCL 方法能够在无监督条件下提取物体表示的同时还能够得到场景中物体的形状、大小、外观等信息, 它们已经在诸如无监督目标跟踪[6]、组成式图像生成[7,8]、机器人环境感知[9]等多个计算机视觉相关的下游任务中得到了广泛的应用. 在具体应用过程中, 部分下游模型直接以预训练的 OCL 方法提取的物体表示作为模型的输入在指定任务中训练, 另一些模型下游则在网络中引入 OCL 组成式建模的结构, 从而直接在指定任务上进行端到端训练.

早期的 OCL 方法 (例如槽注意力 (slot attention)[10]模型和改进版生成式场景推断与采样网络 (GENESIS-V2)[11]), 通常采用像素混合解码器, 该解码器分别生成物体的掩码以及外观, 并使用掩码作为混合权重来组合外观以进行图像重构. 由于这种简单的解码器在复杂场景中表现不佳[12,13], 最近的 OCL 方法专注于探索更强大的解码器结构, 以提高模型处理复杂自然数据集的能力. 在这些方法中, 槽注意力变换器模型 (slot attention Transformer, SLATE)[12]和面向视频的槽注意力变换器模型 (slot attention Transformer for videos, STEVE)[14]使用自回归变换器 (auto-regressive Transformer) 作为解码器, 而潜在插槽扩散模型 (latent slot diffusion, LSD)[7]和基于扩散模型的以物体为中心的生成模型 (SlotDiffusion)[8]则使用扩散模型 (diffusion model, DM)[15,16]作为解码器. 这些方法在分割和生成性能上相较于早期方法取得了更好的效果. 然而, 这些近期的模型在生成过程中并未明确建模物体的掩码, 而是将所有的物体表示一起输入进解码器中, 这与人们想象或创造场景的方式相悖. 实验发现, 这类方法无法根据物体的表示生成与之对应的单个物体的图像. 以 LSD 为例, 如图 1 所示, 第 1 行展示了完整的输入图像以及使用在编码过程中通过 LSD 获取的注意力掩码进行覆盖后的输入图像, 第 2 行展示了重构图像以及使用与注意力掩码相对应的单个物体表示生成的图像. 尽管 LSD 在分割和重构整个图像方面具有相对合理的能力, 但它会从单一物体表示中生成许多预料之外的物体, 这限制了模型的泛化能力, 并且违背了 OCL 的初衷. 一种可行的解决方案是将混合解码器模块与扩散模型结合起来, 从而使模型能够像人类一样想象或重构复杂的自然场景.

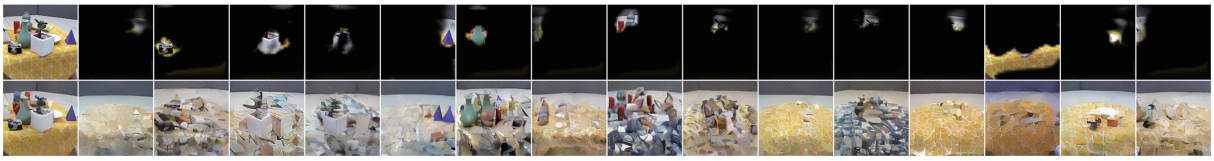


图 1 对比方法 LSD 的单物体生成性能

本文提出一种以物体为中心的扩散 (object-centric diffusion, OCD) 模型, 以解决最近方法所面临的问题. 受到 GENESIS-V2 的启发, OCD 使用实例染色折棍过程编码器来提取物体表示, 而不是广泛使用的 Slot Attention 编码器, 这是因为 GENESIS-V2 提取物体表示的方式更加简单直接, 能够在一定程度上解决不同物体表示之间的耦合问题, 并且符合本文组成式建模的初衷. 在 OCD 的解码器方面, 本文提出了一种混合扩散解码器, 该解码器在解码过程中专门建模了物体的掩码, 并以组成式的方式重构图像潜在空间. 这种设计增强了模型的鲁棒性, 并且更符合人类与世界交互的方式. 在两个合成数据集和两个真实世界数据集上进行了一系列的实验表明, OCD 在无监督分割和图像生成方面能够得到与类似的以物体为中心的学习方法相比更好或者相近的结果. 此外, 本文还展示了 OCD 可以通过在生成过程中选择物体表示来编辑图像. 值得注意的是, OCD 还能够根据物体的表示生成单个物体的图像, 而这是其他类似方法难以实现的.

1 相关工作

1.1 以物体为中心的表示学习

大多数的早期模型都基于以变分自编码器 (variational autoenoder, VAE)[17]为代表的生成模型进行改进, 并以最大化证据似然下界 (evidence lower bound, ELBO) 为目标, 其中物体的表示被建模为隐变量. 一些方法并行推断所有物体的表示: 神经期望最大化算法 (neural expectation maximization, N-EM)[18]通过空间混合模型建模场景, 并提取物体的表示. 可学习深度先验模型 (learnable deep prior, LDP)[19]首次提出学习掩码的先验分布, 并在 LDP 中专门建模遮挡, 以便模型能够学习物体的完整形状. 迭代物体分解推断网络 (iterative object decomposition inference network, IODINE)[20]通过摊销变分推断并行推断表示. 基于空间并行注意力的成分提取 (spatially parallel attention and component extraction, SPACE)[21]模型专门建模了背景, 并且并行推断前景物体.

提取物体表示的另一种方式是顺序推断: 注意力-推断-循环 (attend-infer-repeat, AIR)[22]网络通过循环神经网络 (repeat neural network, RNN) 模块迭代推断物体的表示, 每个物体的表示又可以被解耦为 3 个部分: 存在性、位置以及外观. 针对无穷数量遮挡物体的生成式模型 (generative modeling of infinite occluded objects)[23]通过摊销变分推断方法来推断物体的表示, 并通过长短神经记忆网络 (long short-term memory, LSTM) 模块迭代生成整个场景. 空间一致的注意力-推断-循环 (spatially invariant attend-infer-repeat, SPAIR)[24]网络, 如 AIR 一样推断表示, 但与 AIR 的不同之处在于, 它将整个场景分为几个部分, 每个部分仅检测其中心位于该部分内的物体. 多物体网络 (multi-object network, MONet)[25]提取物体的注意力掩码并以此推断物体的表示. 生成式场景推断与采样 (generative scene inference and sampling, GENESIS)[26]网络分别建模物体的形状和外观表示, 并且这些形状表示相互依赖. GENESIS-V2[11]通过实例折棍染色过程提取物体的形状, 而不是像 GENESIS 那样通过一个解码器对形状表示进行解码来提取形状.

最近的方法更注重将物体表示提取为嵌入, 而不是隐变量. 尽管这些方法不是生成模型, 但它们在更复杂的数据集上取得了令人印象深刻的结果. 与早期方法相比, 最近的方法通常使用注意力机制提取物体表示. 槽注意力 (slot attention)[10]通过图像与物体表示之间的交叉注意力提取物体表示, 并迭代更新这些表示. 然后, 这些表示被用于像素混合解码器来重构整个场景. 双层优化查询槽注意力 (bi-level optimized query slot attention, BO-QSA)[27]模型学习物体的初始表示, 并根据双层优化策略优化槽注意力模型. SLATE[12]将图像编码为潜在表示, 并在潜在表示空间应用槽注意力, 但与槽注意力不同的是, SLATE 使用自回归变换器解码器. LSD[7]使用了预训练的图像编解码器, 将输入图像转换为潜在表示并添加噪声. 然后以槽注意力提取的表示作为去噪特征图的输入. 基于无标签知识蒸馏的真实场景槽注意力 (DINO and slot attention using real-world data, DINOSAUR)[13]模型采用预训练的视觉变换器 (vision Transformer, ViT) 模型提取图像潜在表示, 并使用槽注意力模型重构这些图像潜在表示.

1.2 扩散模型

扩散模型[15,16]可以被视为一种层次化的 VAE[28]. 它首先逐渐向图像中添加噪声, 直到其与标准高斯噪声相似, 然后训练模型根据这些含有噪声的图像来预测其中添加的噪声. 在生成过程中, 通过反复使用训练好的模型对从标准高斯分布种采样的噪声样本进行逐步去噪来生成图像. 自从去噪扩散概率模型 (denoising

diffusion probabilistic model, DDPM)[15]提出以来, 扩散模型在图像生成[2]和图像修复[29]等领域一直表现出色. 如 LDM[30]这样的模型也尝试通过自然语言、分割图和物体位置等多种条件来控制图像的生成过程. LSD 和 SlotDiffusion 借鉴 LDM 将图像转换到潜在空间的思想, 以物体表示为条件训练条件扩散模型, 首次成功地将扩散模型应用到 OCL 这一领域中.

一些近期图像生成的研究[31,32]致力于通过组成式扩散模型生成完整的图像. 需要明确的是, 这些方法只能以组成式的方式生成图像, 而无法从图像中提取物体表示. 与 OCL 不同, 这些模型学习到的表示可能代表的是颜色和光等抽象的概念. 此外, 与大多数 OCL 方法不同, 这些方法无法生成或推断物体的掩码.

2 基于以物体为中心扩散的组成式场景建模算法

本节中将首先介绍每个部分的结构, 然后介绍训练过程中所使用的损失函数. OCD 主要由 3 个模块组成: 用于推断物体表示的实例折棍染色过程编码器、用于提取图像潜在表示和添加噪声的加噪模块, 以及利用物体表示进行去噪的组成式解码器, 完整的模型结构如图 2 所示. 实例折棍染色过程编码器接收输入图像, 并提取图像中的物体表示. 加噪模块提取图像特征并对特征图进行加噪, 得到带噪声特征图. 混合扩散解码器以物体表示为条件, 对带噪声特征图进行去噪, 得到预测的无噪声特征图, 并将其与真实的无噪声特征图计算重构损失. 另外, 混合扩散解码器输出的图像分割结果还将被用于评价 OCD 的图像分割性能. 给定一张图像 $\boldsymbol{x}$, OCD 首先使用实例折棍染色过程编码器提取 K 个物体表示 $\boldsymbol{s}_i$. 这些物体表示将被当作混合扩散解码器中的去噪条件. 在加噪模块中, 图像 $\boldsymbol{x}$ 将被输入到预训练的图像编码器中以获取图像潜在表示 $\boldsymbol{s}_i$, 然后添加噪声以得到含有噪声的图像潜在表示 $\boldsymbol{z}_t$. 在混合扩散解码器中, OCD 以一个 U-Net 网络 $g_\theta^{\text{U-Net}}$ 作为去噪网络. 该网络以物体表示作为输入的条件, 通过物体与图像表示之间的交叉注意机制生成物体掩码 $\boldsymbol{m}_i^{\text{gen}}$ 和预测的无噪声的图像潜在表示 $\hat{\boldsymbol{z}}_{0,i}$. 然后, 解码器通过 Softmax 操作对掩码 $\boldsymbol{m}_i^{\text{gen}}$ 进行归一化得到 $\hat{\boldsymbol{m}}_i^{\text{gen}}$. 最后, 解码器以 $\hat{\boldsymbol{m}}_i^{\text{gen}}$ 为权重对 $\hat{\boldsymbol{z}}_{0,i}$ 进行加权求和, 以获得无噪声图像潜在表示的最终重构 $\hat{\boldsymbol{z}}$. 另外, 我们将生成的物体掩码 $\hat{\boldsymbol{m}}_i^{\text{gen}}$ 作为 OCD 对输入图像的分割结果, 并以该分割结果作为衡量模型性能的一个关键指标.

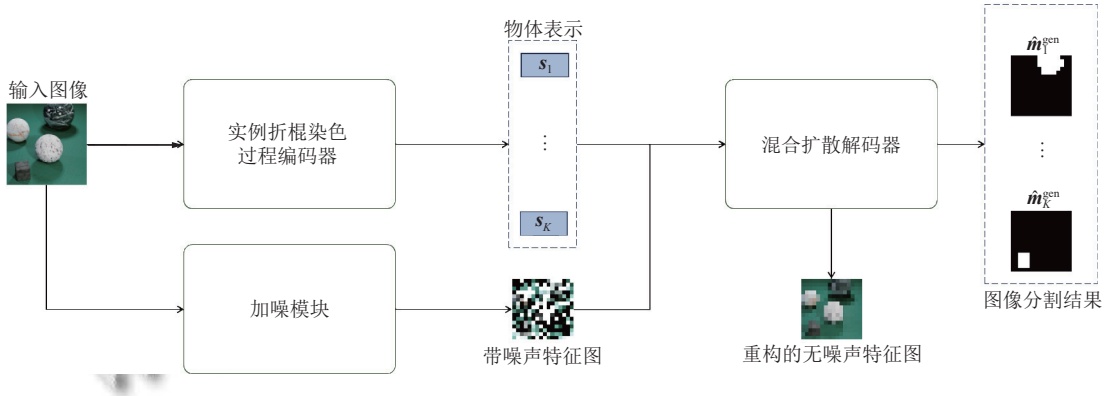


图 2 OCD 的主要结构

2.1 实例折棍染色过程编码器

OCD 的编码器主要基于 GENESIS-V2 中的实例折棍染色过程编码器[11]进行改进. 如图 3 所示, 给定一个输入图像 $\boldsymbol{x} \in R^{C\times H\times W}$, 它首先通过一个 U-Net 主干网络 $f_\theta^{\text{U-Net}}$ 转换为特征图 $\boldsymbol{x}^{\text{feat}} \in R^{C_{\text{feat}}\times H_{\text{feat}}\times W_{\text{feat}}}$, 以得到不同尺度的原始图像信息. 接着, 编码器通过一个折棍染色过程对特征图中不同位置的特征进行聚类 (具体的算法如算法 1 所示), 并从特征图中获得 K 个推断掩码 $\boldsymbol{m}_i^{\text{inf}} \in [0,1]^{1\times H_{\text{feat}}\times W_{\text{feat}}}$ $(i=1,\cdots,K)$. 这里 K 表示物体的数量, 并被设定为超参数. 为了简洁起见, 后续讨论中将省略物体索引 i 的范围. 然后, 对于每个物体, 实例折棍染色过程编码器以它们对应的推断掩码作为权重, 对特征图 $\boldsymbol{x}^{\text{feat}}$ 进行加权平均, 以获得每个推断掩码的物体表示 $\boldsymbol{s}_i \in R^{D_{\text{obj}}}$.

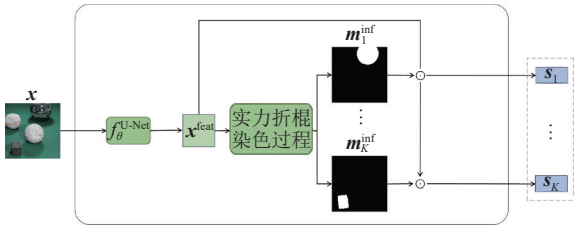


图 3 实例折棍染色过程编码器

算法 1. 实例折棍染色过程算法

1) 输入特征图 $\boldsymbol{x}^{\text{feat}}$ 以及物体数量 K, 初始化推断掩码 $\boldsymbol{m}^{\text{inf}}=\varnothing$, 掩码区域 $\boldsymbol{d}=\mathbf{1}^{H_{\text{feat}}\times W_{\text{feat}}}$, 随机种子 $c\sim U(\mathbf{0},\mathbf{1})\in[0,1]^{H_{\text{feat}}\times W_{\text{feat}}}$, 初始掩码编号 $k=1$.

2) 将当前掩码区域 $\boldsymbol{d}$ 与随机种子相乘, 并选取相乘值最大的位置为聚类中心 (i,j).

3) 计算特征图所有位置的特征到该聚类中心的距离并归一化, 得到距离 d.

4) 将该距离 α 与掩码区域 $\boldsymbol{d}$ 相乘, 得到第 k 个物体推断掩码 $\boldsymbol{m}_k^{\text{inf}}$.

5) 更新剩下的掩码区域 $\boldsymbol{d}=\boldsymbol{d}\odot(1-\alpha)$, 更新掩码编号 $k=k+1$, 返回第 2) 步.

原始 GENESIS-V2 中的实例折棍染色过程编码器将 $\boldsymbol{s}_i$ 视为物体隐变量后验分布的参数, 并从该后验分布中进行采样以获取物体表示. 而 OCD 则将 $\boldsymbol{s}_i$ 本身视为物体表示, 不进行采样, 以增强其稳定性. 与槽注意力编码器[10]相比, OCD 的编码器采用更简单的方式直接提取物体表示, 而无需竞争性注意力机制以及 RNN 的循环更新机制, 这种结构使其更适合与第 2.3 节中描述的混合扩散解码器以及第 2.4 节中描述的掩码损失结合使用.

2.2 加噪模块

在传统的以物体为中心的学习方法中, 解码器尝试使用物体表示作为输入来重构输入图像. 相比之下, 在最近的方法中, 解码器可能用于重构潜在表示 (如 SLATE[12]和 DINOSAUR[13]), 甚至预测在潜在表示中添加的噪声 (如 LSD[7]或 SlotDiffusion[8]). OCD 与这些最近的方法紧密相关. 与 LSD 和 LDM[30]类似, 如图 4 所示, OCD 首先将输入图像转换为潜在表示, 并在潜在表示空间上训练一个扩散去噪模型. OCD 与 LSD 之间的主要区别在于模型采用物体来表示去噪的方式.

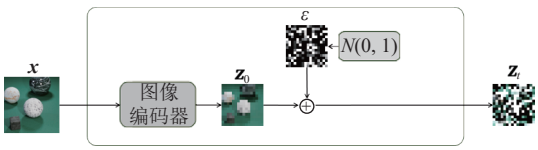


图 4 加噪模块

(1) 图像编解码器

由于在隐式空间中的操作更高效且更可能提取图像特征, OCD 主要遵循 LSD 和 LDM 的方式将图像转换为潜在表示. 给定一个图像 $\boldsymbol{x}\in R^{C\times H\times W}$, OCD 首先通过预训练的图像编码器将其转换为潜在表示 $\mathbf{z}_0\in R^{C_{\text{AE}}\times H_{\text{AE}}\times W_{\text{AE}}}$. 在生成过程中, 当 OCD 获得潜在表示的重构 $\mathbf{z}'_0\in R^{C_{\text{AE}}\times H_{\text{AE}}\times W_{\text{AE}}}$ 后, 会将其输入到预训练的图像解码器中, 以获得重构图像 $\hat{\boldsymbol{x}}\in R^{C\times H\times W}$.

(2) 隐式条件扩散模型

OCD 主要遵循标准的 LDM 来构建隐式条件扩散模型. 给定作为条件的物体表示集合 $S=\{\boldsymbol{s}_1,\cdots,\boldsymbol{s}_K\}$, 图像潜在表示 $\mathbf{z}_0$ 的分布可以描述为:

$$p(\mathbf{z}_0\mid S)=\int p(\mathbf{z}_{0:T}\mid S)\,\mathrm{d}\mathbf{z}_{1:T} \tag{1}$$

其中, 联合分布 $p(\mathbf{z}_{0:T}\mid S)$ 被建模为一个马尔可夫链, 如下所示:

$$p(\mathbf{z}_{0:T}\mid S)=p(\mathbf{z}_T)\prod_{t=T,\cdots,1}p(\mathbf{z}_{t-1}\mid \mathbf{z}_t,S) \tag{2}$$

$$p(\mathbf{z}_T)=N(\mathbf{0},\boldsymbol{I}) \tag{3}$$

$$p(\mathbf{z}_{t-1}\mid \mathbf{z}_t,S)=N(\boldsymbol{\mu}_\theta(\mathbf{z}_t,t,S),\beta_t\boldsymbol{I}) \tag{4}$$

其中, 正态分布的方差 $\beta_t\boldsymbol{I}$ 是一个随时间 t 递增的方差序列. 为了预测正态分布的均值 $\boldsymbol{\mu}_\theta$, OCD 直接预测不含噪声的图像潜在表示 $\mathbf{z}_0$, 而不是像 DDPM 那样预测噪声, 这样可以提高训练的稳定性, 即:

$$\boldsymbol{\mu}_\theta(\mathbf{z}_t,t,S)=\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{z}}_0+\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{z}_t \tag{5}$$

其中, $\alpha_t=1-\beta_t$, $\bar{\alpha}_t=\prod_{i=1}^{t}\alpha_i$, 而 $\hat{\mathbf{z}}_0=g_\theta(\mathbf{z}_t,t,S)$ 是 OCD 混合扩散解码器的输出.

在生成过程中, OCD 首先从标准高斯分布中采样得到噪声 $\mathbf{z}'_T$, 然后从条件分布 $p(\mathbf{z}_{t-1}\mid \mathbf{z}_t,S)$ 中逐步采样得到 $\mathbf{z}'_{T-1},\mathbf{z}'_{T-2},\cdots$, 并最终得到 $\mathbf{z}'_0$, 这可以看作是图像潜在表示的重构. 值得一提的是, 这里采样得到的 $\mathbf{z}'_0$ 与混合扩散解码器输出的 $\hat{\mathbf{z}}_0$ 是不同的. 后者仅用于去噪的中间步骤, 不能视为最终的重构结果.

在训练过程中, OCD 主要遵循 LDM 的方式. 其首先从 $\{1,\cdots,T\}$ 中采样一个时间步 t, 并采样一个标准的高斯噪声 ε, 然后通过以下公式得到带噪声的图像潜在表示 $\mathbf{z}_t$:

$$\mathbf{z}_t=\sqrt{\bar{\alpha}_t}\mathbf{z}_0+\sqrt{1-\bar{\alpha}_t}\varepsilon \tag{6}$$

相应的损失函数可以描述为:

$$L_{\text{Diff}} = \|\boldsymbol{z}_0 - g_\theta(\boldsymbol{z}_t, t, S)\|^2 \tag{7}$$

2.3 混合扩散解码器

由于编码器提取的物体表示包含图像潜在信息, OCD 可以方便地以这些物体表示为条件, 利用其中的信息在$g_\theta(\boldsymbol{z}_t, t, S)$中预测$\boldsymbol{z}_0$. 当前常用的条件去噪网络通常采用单个 U-Net. 这种网络将所有条件和带噪声的图像潜在表示一起输入进单个 U-Net 中, 而其中所有条件通过在特定 U-Net 层中的交叉注意力模块引导去噪过程. 这种设计适用于自然语言和复数个物体表示等条件, 在这些情况下, 条件之间存在竞争机制. 然而, 当使用单个条件进行去噪时, 单个 U-Net 可能会失效. 这是因为当条件数量为 1 时, U-Net 中的交叉注意力机制中没有条件之间的竞争, 并且会忽略图像的信息, 而这与训练期间的条件数量大于 1 时的计算过程显著不同, 从而限制了模型的泛化能力, 并最终使得 LSD 无法充分利用单个物体槽中的信息进行生成.

相比之下, 为了充分提取每个物体表示 $\boldsymbol{s}_i$ 的信息, 本文提出了一种改进的混合扩散解码器. 如图 5 所示, OCD 采用一个共享的 U-Net 网络 $g_\theta^{\text{U-Net}}$, 该 U-Net 网络仅接收单个物体表示 $\boldsymbol{s}_i$, 带噪声的图像潜在表示$\boldsymbol{z}_t$, 以及时间步 t. 与 LSD 一样, OCD 在 U-Net 的不同层之间使用单个物体表示与图像特征图之间的交叉注意力机制, 由于 OCD 只使用单个物体表示作为条件, 因此这种交叉注意力机制其实最终会退化为一个以物体表示为输入的简单的线性层. 尽管这种设计在某些层会忽略图像信息, 但由于 OCD 的预测目标是预测无噪声的$\boldsymbol{z}_0$, 而不是像 LSD 那样预测噪声ε, 因此图像信息的重要性会在一定程度上被削弱, 而这反而迫使物体表示包含更多的图像信息, 从而提高了它们的质量. 另一方面, 这种设计能够使得物体表示的数量对输出结果质量的影响相对较小, 这是因为无论物体表示数量为多少, 交叉注意力机制都会退化为线性层, 因此生成过程与训练过程中 OCD 的计算过程是类似的.

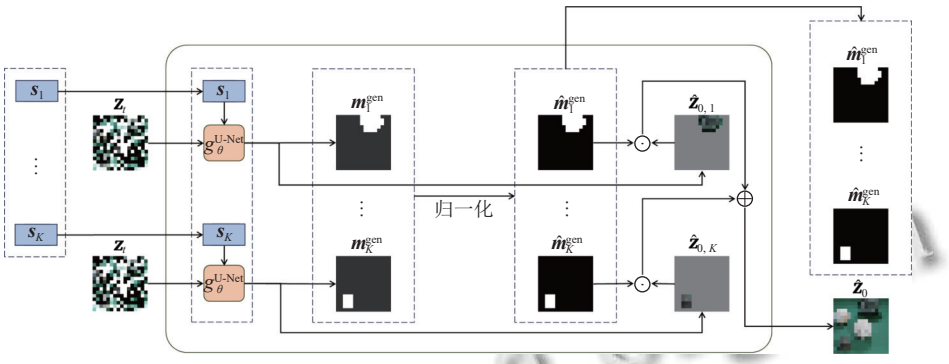


图 5 混合扩散解码器

为了整合每个物体表示对应的预测结果, $g_\theta^{\text{U-Net}}$ 需要计算一个额外的、未归一化的生成掩码$\boldsymbol{m}_i^{\text{gen}} \in [0,1]^{1\times H_{\text{AE}}\times W_{\text{AE}}}$, 还要计算每个物体外观的预测值$\hat{\boldsymbol{z}}_{0,i} \in R^{C_{\text{AE}}\times H_{\text{AE}}\times W_{\text{AE}}}$. 最后, OCD 以归一化后的生成掩码$\hat{\boldsymbol{m}}_i^{\text{gen}}$为权重对预测的物体外观$\hat{\boldsymbol{z}}_{0,i}$进行加权求和, 最终得到不含噪声的图像潜在表示$\hat{\boldsymbol{z}}_0$:

$$[\boldsymbol{m}_i^{\text{gen}}, \hat{\boldsymbol{z}}_{0,i}] = g_\theta^{\text{U-Net}}(\boldsymbol{z}_t, condition = \boldsymbol{s}_i, t) \tag{8}$$

$$\hat{\boldsymbol{m}}_i^{\text{gen}} = \underset{i=1,\cdots,K}{Softmax}(\boldsymbol{m}_i^{\text{gen}}) \tag{9}$$

$$\hat{\boldsymbol{z}}_0 = \sum_{i=1}^{K} \hat{\boldsymbol{m}}_i^{\text{gen}} \hat{\boldsymbol{z}}_{0,i} \tag{10}$$

值得一提的是, U-Net 网络$g_\theta^{\text{U-Net}}$的结构几乎与经典的用于条件图像生成的 U-Net 相同, 唯一的区别在于它输出一个带有额外通道的特征图, 用于表示生成掩码$\boldsymbol{m}_i^{\text{gen}}$.

2.4 损失函数

由于 OCD 实际上仍然是一种扩散模型, 因此在损失函数的设计上, 仍然遵循与 LDM 类似的设置, 期望 OCD 能够在给定物体表示的情况下对带噪声图像潜在表示进行去噪. 然而, 正如第 2.2 节所提到的那样, 在实验中, 我们发现直接预测噪声可能会导致训练的不稳定, 因而使用式 (7) 预测不含噪声的图像潜在表示.

由于专门建模了生成掩码, 因此 OCD 紧密遵循 GENESIS-V2, 并添加了额外的掩码损失函数, 以鼓励生成掩码与推断掩码相似, 从而进一步提高物体表示

的质量. 总的损失函数可以描述为:

$$L = L_{\text{Diff}} + \eta KL(\boldsymbol{m}^{\text{inf}} || nograd(\hat{\boldsymbol{m}}^{\text{gen\_s}})) \tag{11}$$

其中, L_{Diff} 的计算公式如式 (7) 所示, η 是一个在训练过程中变化的超参数, KL 函数用于计算两个多项分布之间的 KL 散度, 这两个多项分布分别以 $\boldsymbol{m}^{\text{inf}}$ 和 $\hat{\boldsymbol{m}}^{\text{gen\_s}}$ 作为权重. $\hat{\boldsymbol{m}}^{\text{gen\_s}}$ 由 $\hat{\boldsymbol{m}}^{\text{gen}}$ 上采样而来, 这是为了确保用于计算 KL 散度的两个分布具有相同的维度. "$nograd()$"表示我们将切断 KL 散度损失经由 $\hat{\boldsymbol{m}}^{\text{gen\_s}}$ 的反向传播以提高训练的稳定性. 假设 $\boldsymbol{m}^{\text{inf}} = [m_{1,1}^{\text{inf}}, \cdots, m_{1,K}^{\text{inf}}, \cdots, m_{N,1}^{\text{inf}}, \cdots, m_{N,K}^{\text{inf}}]$, $\hat{\boldsymbol{m}}^{\text{gen\_s}} = [\hat{m}_{1,1}^{\text{gen\_s}}, \cdots, \hat{m}_{1,K}^{\text{gen\_s}}, \cdots, \hat{m}_{N,1}^{\text{gen\_s}}, \cdots, \hat{m}_{N,K}^{\text{gen\_s}}]$, 其中 K 代表场景中物体的数量, N 代表图像中总的像素个数, 并且 $\sum_{j=1}^{K} m_{i,j}^{\text{inf}} = \sum_{j=1}^{K} \hat{m}_{i,j}^{\text{gen\_s}} = 1$, $\forall i \in \{1, \cdots, N\}$, 即生成与推断掩码中每一个位置上各个物体的权重加和为 1, 则 KL 散度项的表达式为:

$$KL(\boldsymbol{m}^{\text{inf}} || nograd(\hat{\boldsymbol{m}}^{\text{gen\_s}})) = \sum_{i=1}^{N}\sum_{j=1}^{K} m_{i,j}^{\text{inf}} \log\left(\frac{m_{i,j}^{\text{inf}}}{\hat{m}_{i,j}^{\text{gen\_s}}}\right) \tag{12}$$

即这一项可以被解析地计算与求导, 这样, 我们即可根据式 (11) 通过反向传播来对 OCD 进行训练.

3 实验分析

3.1 实验环境

本文实验环境为 Ubuntu 20.04 操作系统; GPU 为 GeForce GTX 4090, 11 GB; CPU 为 Intel Xeon Gold 6133 2.50 GHz; 深度学习框架为 PyTorch 1.13.1 框架.

3.2 数据集

本文采用 4 个数据集来评估 OCD 的性能. 在这 4 个数据集中, ClevrTex[33]和 MOVi-C[34]为合成数据集, 而 OCTScenes-A 和 OCTScenes-B[35]则为真实世界数据集. 所有数据集的图像大小均被设置为 128×128.

在两个合成数据集中, ClevrTex 包含由多个具有简单形状 (如球体和立方体) 和复杂纹理的物体组成的单视角场景, 并且具有纹理复杂的背景. MOVi-C 是一个视频数据集, 不过在本文的实验中则被当作单视角图像来处理. 与 ClevrTex 相比, MOVi-C 具有更复杂的物体和更为自然的背景. 两个以物体为中心的真实世界数据集 OCTScenes-A (OCT-A) 和 OCTScenes-B (OCT-B) 都包含多视角场景数据, 其中物体是静态的, 并被放置在桌子上. 这两个数据集中的场景也将被视为单视角图片. OCT-A 包含简单的单色物体, 而 OCT-B 则具有更复杂的物体, 且桌子具有不同的复杂纹理. 相比 OCT-A, OCT-B 还包含更多的物体.

3.3 对比方法

本文主要将 OCD 与 3 个方法进行比较: LSD[7]、GENESIS-V2[11]和 SLATE[12]. LSD 是唯一具有扩散模型框架的单视角 OCL 模型, 并且是目前具备生成能力的最先进的模型, 和 OCD 相比, LSD 并没有使用组成式的扩散解码器. 选择 GENESIS-V2, 是因为 OCD 主要采用了与其相似的实例染色折棍过程编码器, 但 GENESIS-V2 并没有使用扩散模型作为解码器, 而是使用了简单的空间广播解码器. 选择 SLATE 则是为了比较扩散模型解码器和自回归 Transformer 解码器的表现, 和 OCD 相比, SLATE 使用槽注意力编码器, 并使用自回归 Transformer 解码器. 在评价过程中, 我们对于所有模型都预先设置相同的物体数量, 在 ClevrTex、MOVi-C、OCT-A 和 OCT-B 上的物体数量分别设置为 11、11、8 和 15.

3.4 评价指标

在评估无监督分割性能时, 本文主要根据前景物体的调整兰德指数 (adjusted Rand index-foreground, $ARI\text{-}FG$) 和平均交并比 (mean intersection over union, $mIoU$) 对不同的模型进行比较. 在衡量模型生成能力时, 本文主要使用感知分数 (inception score, IS).

(1) 前景物体的调整兰德指数 ($ARI\text{-}FG$): 该指标主要用于衡量聚类结果与真实类别的相似程度. 该指标的计算公式如下:

$$ARI\text{-}FG = \frac{\sum_{ij} C(n_{ij},2) - \left[\sum_i C(a_i,2)\sum_j C(b_j,2)\right] \Big/ C(n,2)}{\frac{1}{2}\left[\sum_i C(a_i,2) + \sum_j C(b_j,2)\right] - \left[\sum_i C(a_i,2)\sum_j C(b_j,2)\right] \Big/ C(n,2)} \tag{13}$$

其中, n_{ij} 表示模型预测属于第 i 个物体而实际属于第 j 个物体的像素数量. a_j 表示模型预测第 i 个物体的像素数量, b_j 表示实际第 j 个物体的像素数量. n 表示图片中所有的像素数量, C 表示组合数运算. 本文仅对前

景像素计算该指标, 而不考虑背景. 在对测试集中的所有图像计算 *ARI-FG* 指标之后, 我们对其进行平均, 得到结果.

(2) 平均交并比 (*mIoU*): 和 *ARI-FG* 指标不同, *mIoU* 考虑图片中所有的物体以及背景, *mIoU* 的计算方式如下:

$$mIoU = \frac{\sum_i \boldsymbol{A}_i \cap \boldsymbol{B}_i}{\sum_i \boldsymbol{A}_i \cup \boldsymbol{B}_i} \tag{14}$$

其中, $\boldsymbol{A}_i$ 与 $\boldsymbol{B}_i$ 表示模型预测第 i 个物体的区域以及与预测的第 i 个物体区域最相似的实际的物体区域. 和 *ARI-FG* 的计算类似, 我们对所有图片上的结果进行平均后得到比较结果.

(3) 感知分数 (*IS*): 感知分数同时衡量模型生成图片的真实性和多样性, *IS* 的计算公式如下:

$$IS = \exp(E_{\boldsymbol{x}\sim p(\boldsymbol{x})}(KL(P(\boldsymbol{y}|\boldsymbol{x})\|P(\boldsymbol{y})))) \tag{15}$$

其中, *KL* 表示两个分布之间的 *KL* 散度. $P(\boldsymbol{y}|\boldsymbol{x})$ 与 $P(\boldsymbol{y})$ 分别表示单个图像属于各个类别的多项分布以及所有图像中所有类别形成的多项分布. 为了得到 $P(\boldsymbol{y}|\boldsymbol{x})$ 与 $P(\boldsymbol{y})$, 我们将模型生成的图像输入进一个预训练的分类器模型来得到对单个图像属于各个类别的多项分布 $P(\boldsymbol{y}|\boldsymbol{x})$ 的估计. 随后, 我们对每 10 张图片进行平均来得到对图像中所有类别形成的多项分布 $P(\boldsymbol{y})$ 的估计. 为了充分衡量模型的生成能力, 我们对所有方法都采样了 1 000 张图像, 并以 10 张图片为一个单位计算 *IS* 指标, 最后对所有的 *IS* 指标结果进行平均来得到表格中的结果.

3.5 无监督图像分割

由于 LSD 和 SLATE 重构的是图像潜在表示而非图像, 因此在测试时, 本文将这两个模型的注意力掩码上采样到图像大小. 对于 OCD, 本文使用在去噪过程中计算得到的生成掩码, 然后将其上采样到图像大小. 对于 GENESIS-V2 (GEN-V2), 本文简单使用其原始大小的生成掩码. 数值结果和可视化结果见表 1 和图 6.

表 1 无监督分割与生成的对比结果

数据集	方法	*ARI-FG* (%)	*mIoU* (%)	*IS*
ClevrTex	OCD	**77.9**	**40.5**	3.26
	LSD[7]	64.2	29.0	**4.56**
	GEN-V2[11]	25.3	15.4	2.42
	SLATE[12]	38.9	21.5	2.09
MOVi-C	OCD	51.2	24.6	**4.29**
	LSD[7]	**51.7**	**24.7**	4.20
	GEN-V2[11]	17.6	8.3	1.45
	SLATE[12]	43.2	16.5	3.03
OCT-A	OCD	84.6	34.7	**2.94**
	LSD[7]	29.9	12.8	1.94
	GEN-V2[11]	**90.9**	**62.6**	2.51
	SLATE[12]	46.6	22.2	2.33
OCT-B	OCD	**75.6**	**35.0**	**3.19**
	LSD[7]	65.4	28.2	2.46
	GEN-V2[11]	64.1	27.7	2.64
	SLATE[12]	41.3	22.4	2.68

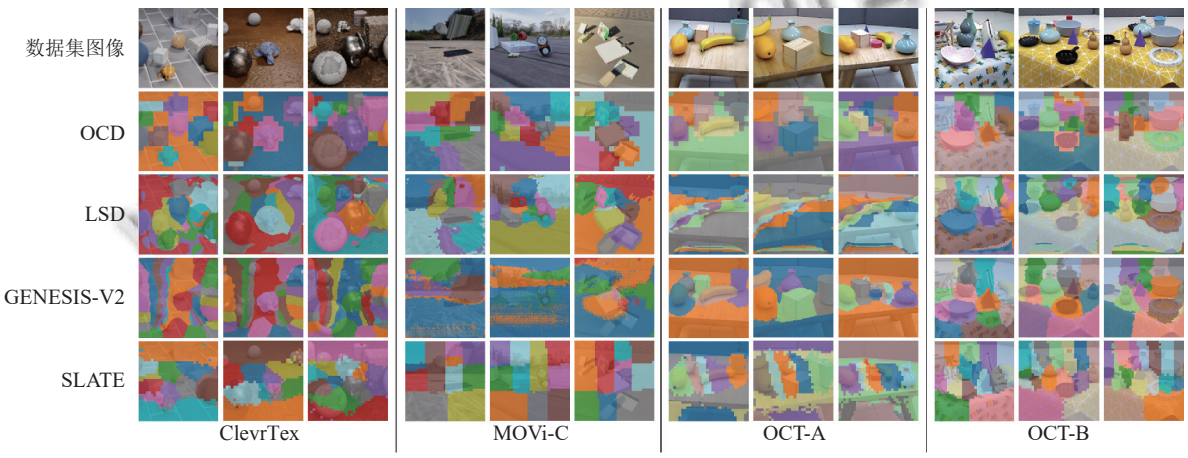


图 6 OCD、LSD、GENESIS-V2 以及 SLATE 的分割结果

ARI-FG 与 *mIoU* 的数值结果显示, 相比于最近的方法 LSD 和 SLATE, OCD 在除 MOVi-C 外的所有数据集中, 都实现了最佳的分割性能. 而在 MOVi-C 中, OCD 与 LSD 的性能相当. 这证明了 OCD 有效性与普遍适用性. 对于 GENESIS-V2, 可以看到在简单数据集 OCT-A 中, 其分割性能最强, 但在其他更复杂的数据集

中表现不佳, 这表明在处理更复杂的数据集时, 扩散去噪解码器相比于原始像素混合解码器更具优势.

3.6 无监督图像生成

为了衡量各个方法的图像生成能力, 我们首先需要使用各个方法生成若干张图片. GENESIS-V2 可以通过解码从其高斯先验中采样的物体表示来生成图像. 然而, 本文发现从 GENESIS-V2 的先验中采样会导致较差的图像生成结果. 为了获得更好的生成结果和公平性, 本文对所有方法采用相同的采样策略: 首先从测试数据集图像中收集编码的物体表示, 以获取物体表示集合 S_{test}, 然后随机选择 K 个物体表示并将其输入进解码器以生成完整图像. 与 LSD 不同, 本文省略了在获得 S_{test} 后的所有后处理步骤, 包括聚类. 这是因为相似的表示无法同时从同一聚类中采样, 从而可能会降低生成图像的多样性. *IS* 的数值结果和样本见表 1 和图 7.

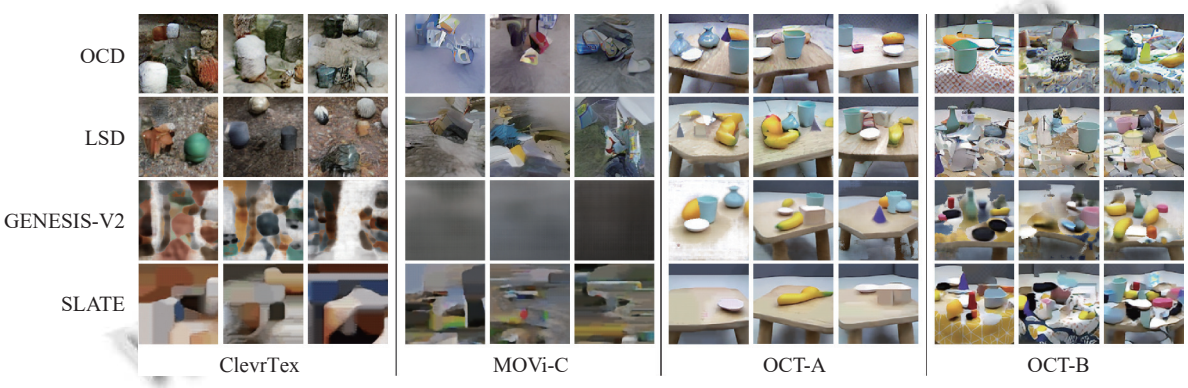


图 7 OCD、LSD、GENESIS-V2 以及 SLATE 的生成结果

结果显示, 除了 ClevrTex 之外, OCD 在所有数据集上达到了最佳的生成结果. OCD 在 ClevrTex 数据集上的生成效果不如 LSD 的原因在于, 如果生成过程中没有采样到背景的表示 (由于一张图片中会有多个物体, 而仅有一个背景, 因此这种情况是很可能发生的), 通过交叉注意力机制, LSD 可能会将一些物体的表示解码为背景, 而生成的背景则会与这些物体具有相同的纹理, 从而在保持图像真实性的同时增加了图像的多样性. 相比之下, OCD 如果没有提供背景表示, 生成的图片中背景则是普通的灰色背景 (见第 3.8 节), 而这实际上与人类的直觉更为一致. 因此, 在处理背景和物体纹理明显不同的数据集 (如 MOVi-C、OCT-A 和 OCT-B) 时, LSD 在 ClevrTex 数据集上的这一特性不复存在, 最终使得 OCD 的生成能力优于 LSD.

3.7 可控组成式生成

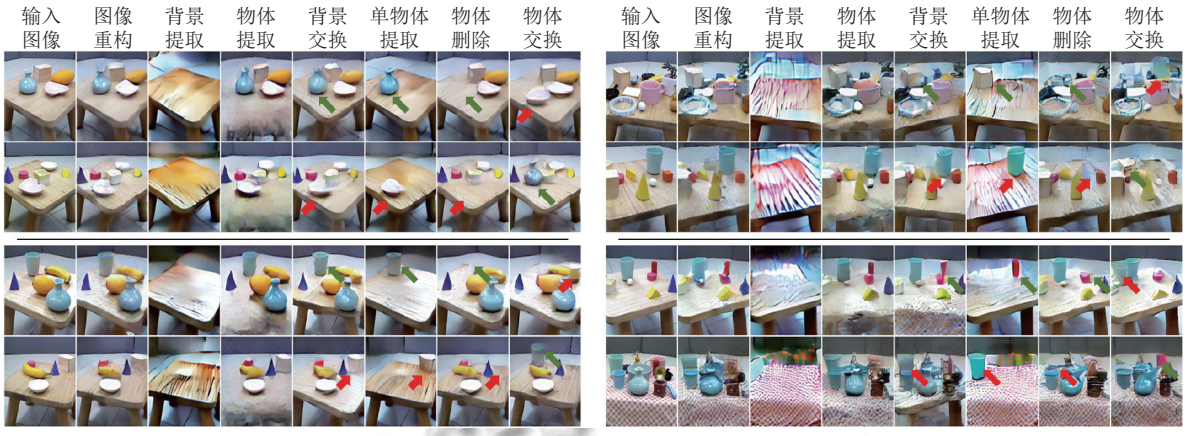
为了展示 OCD 组成式生成的能力, 本文遵循 LSD 的方式, 使用来自不同图像的表示来生成完整图像. 实验首先随机选择两张测试数据集中的图像, 并提取物体或背景的表示, 将具有最大生成掩码的表示视为背景表示, 其余的表示视为前景物体表示. 在生成过程中, OCD 会交换两张图像的背景表示或随机选择的物体表示. 本文还尝试了其他的编辑方式, 包括背景提取、物体提取、单物体提取、物体删除等. 如图 8 所示, OCD 主要在 OCT-A 和 OCT-B 这两个数据集上进行了评估, 其中被交换的物体分别用红色或者绿色的箭头所指示. 结果表明 OCD 提取的物体表示可以被重复使用以生成不同的图像. 这证明了其在组合生成任务中的灵活性和有效性. 另外, 根据调查, 本文是首个在真实的以物体为中心的数据集中衡量可控生成能力的工作.

3.8 单个物体生成

另一项生成实验是使用单个物体表示生成完整的图像. 本文在 OCT-A 和 OCT-B 上评估 OCD 及 LSD. 如图 9 所示, 即使未提供背景表示, OCD 也能使用单个物体表示生成完整图像. 如第 2.3 节所述, 这是因为物体表示的数量对 OCD 中混合扩散解码器的影响较小, 相比之下, LSD 的解码器在处理单个物体表示时会展现出与训练时完全不同的行为.

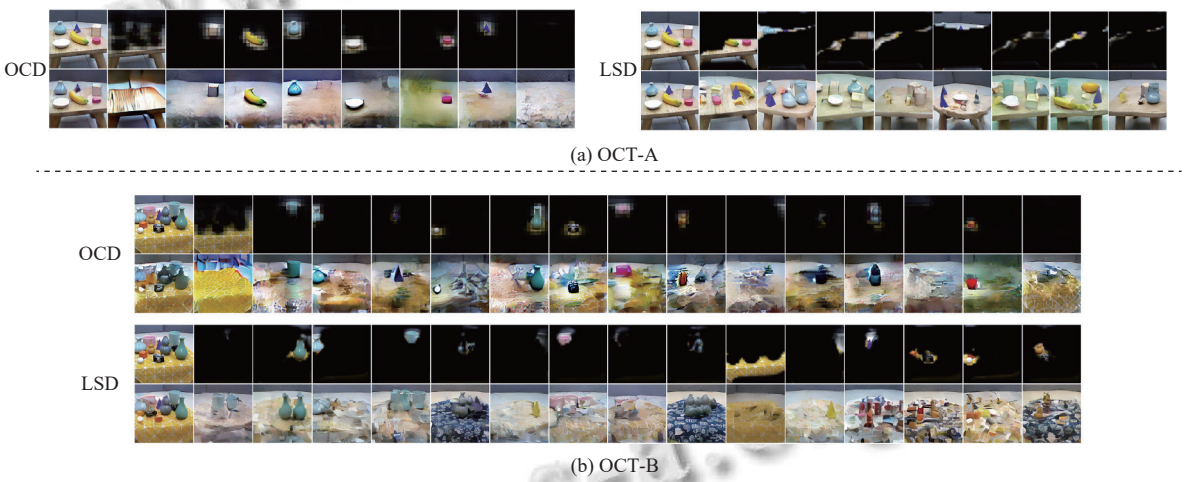
一个有趣的结果是, 对于某些表示, 即使 OCD 的生成掩码几乎为空, 它们仍然可以生成有意义的图像, 这意味着它们的掩码之所以为空, 仅是因为在重构过程中生成结果可能无效. 相比之下, 如果 LSD 的单个物体表示不对应于背景, 它在无法正常生成与表示相对应的图像. 即使只提供一个物体表示, 它也可能生成多个物

体. 特别是在 OCT-A 中, 尽管 LSD 未能对整个场景进行分割, 但在给定其学习到的槽时, 仍能生成有意义的图像. 本文将这一现象归因于 LSD 在去噪阶段强迫物体表示与背景表示在交叉注意力中竞争. 因此, 当背景表示被省略时, LSD 无法确定单个物体表示在整个图像中的位置, 并导致其生成多个不相关的物体.



(a) OCT-A (b) OCT-B

图 8 OCD 的可控组成式生成效果



(a) OCT-A

(b) OCT-B

图 9 OCD 和对比方法 LSD 的单物体生成结果

3.9 消融实验

本文通过将每个模块和训练策略替换为其他模块和策略或直接省略它们来展示 OCD 中每个部分的重要性. 结果见表 2, 其中“SA 编码器”表示将实例染色折棍过程编码器更改为 Slot Attention 编码器; “LSD 解码器”表示将混合扩散解码器更改为 LSD 解码器, 此时所有物体表示将输入到单个 U-Net 中, 并通过交叉注意力机制来对图像进行去噪; “无掩码 *KL*”表示在训练过程中省略了损失函数中推断掩码与生成掩码之间的 *KL* 散度.

结果表明, 本文的模型设计和损失函数是合理的, 并且能够提高分割性能. Slot Attention 编码器在复杂数据集 MOVi-C 和 OCT-B 的训练初期出现了崩溃情况, 这表明 Slot Attention 编码器不适合 OCD. 本文认为这很大程度上是由掩码一致性损失 $KL(\boldsymbol{m}^{\text{inf}}\|nograd(\hat{\boldsymbol{m}}^{\text{gen\_s}}))$ 导致的. 在 Slot Attention 编码器中, 由于物体的表示是随机初始化并根据推断掩码进行多轮更新的, 因此推断掩码与最终编码得到的物体表示之间并没有显式关系. 如果 OCD 使用掩码一致性损失, 则可能会导致在梯度回传时产生包括训练崩溃在内的不可预见的结果. 相比之下, 实例染色折棍过程编码器中推断掩码与物体表示之间的关系是更为显式的, 因而该损失项可以以一种更加直接的方式对物体表示的提取过程产生影响.

在 ClevrTex 中, 训练时不使用掩码 *KL* 散度可能导致模型学习到较差的分割效果, 而这可能反而会增加生成图像的多样性, 原因与第 3.8 节中描述的类似. 另外, LSD 解码器通过当前主流的图像生成模型中的交叉注意力模块来捕获物体之间的关系, 这种方式可能更有助于图像生成.

表 2 消融实验结果

数据集	模型	*ARI-FG* (%)	*mIoU* (%)	*IS*
ClevrTex	完整OCD	**77.9**	**40.5**	3.26
	SA编码器	22.1	12.6	3.72
	LSD解码器	28.2	22.0	3.79
	无掩码*KL*	67.7	34.0	**3.97**
MOVi-C	完整OCD	**51.2**	**24.6**	**4.29**
	SA编码器	—	—	—
	LSD解码器	43.1	15.6	4.19
	无掩码*KL*	48.7	18.6	3.69
OCT-A	完整OCD	**84.6**	**34.7**	**2.94**
	SA编码器	17.1	8.4	2.61
	LSD解码器	46.1	11.3	2.83
	无掩码*KL*	83.3	31.1	2.93
OCT-B	完整OCD	**75.6**	**35.0**	3.19
	SA编码器	—	—	—
	LSD解码器	40.7	16.5	**3.44**
	无掩码*KL*	71.1	32.3	2.81

4 总结与展望

本文介绍了一种全新的以物体为中心的扩散模型 OCD, 并设计了一种新型的混合扩散解码器, 该解码器能够在生成过程中生成物体掩码, 并以组成式的方式预测图像潜在表示. 在 4 个数据集上的实验表明, OCD 在分割和生成方面能够得到优于或与最先进方法相当的结果. 本文还证明了 OCD 能够以组成式的方式生成图像. 特别地, OCD 还能够根据单个物体表示生成包含该物体的图像, 而这一点是同类的对比方法所无法做到的.

本文未来的研究方向主要包含以下两个方面.

(1) 本文尚未探究 OCD 在面对更复杂的真实世界数据集 (如 COCO[36]) 时的效果, 因为这些数据集中几乎没有像以物体为中心的数据集那样的重复物体特征. 与当前主流目标检测方法所处的情况类似, OCL 在这类开放世界数据集中分割与生成的潜力仍然有待发掘[37]. 另外, 本文尚未探索在更复杂的真实世界数据集中进行可控组合生成的能力, 而这在监督学习领域仍然是一个具有挑战性的问题. 本文工作的下一步可能是探索在更自然的数据集中进行组成式建模. 尽管目前一些方法尝试在自然数据集中学习物体表示, 但它们可能没有生成能力, 或者没有以组合方式生成图像. 探索在无监督情况下提取物体表示的可能性, 并利用这些表示以组合方式生成图像, 仍然是一个具有挑战性但有意义的研究领域.

(2) 本文目前仅在单视角图片数据集上验证了 OCD. 而多视角数据集或者视频数据集与单视角数据集相比, 包含更多不同物体之间的关系信息以及同一物体的一致性信息, 可能能够进一步挖掘 OCD 的能力. 参考近期的一些多视角姿态估计方法[38]以及深度估计方法[39], 将来的工作可以进一步尝试整合这些不同帧的信息, 以及进一步使用深度图, 分割图等辅助信息, 来提高 OCD 的效果.

参考文献

1 Kirillov A, Mintun E, Ravi N, *et al*. Segment anything. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 3992–4003. [doi: 10.1109/ICCV51070.2023.00371]

2 Ramesh A, Dhariwal P, Nichol A, *et al*. Hierarchical text-conditional image generation with CLIP latents. arXiv: 2204.06125, 2022.

3 Nichol AQ, Dhariwal P, Ramesh A, *et al*. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. Proceedings of the 2022 International Conference on Machine Learning. Baltimore: PMLR, 2022. 16784–16804.

4 Zhang LM, Rao AY, Agrawala M. Adding conditional control to text-to-image diffusion models. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 3813–3824. [doi: 10.1109/ICCV51070.2023.00355]

5 Yuan JY, Chen TL, Li B, *et al*. Compositional scene representation learning via reconstruction: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 11540–11560. [doi: 10.1109/TPAMI.2023.3286184]

6 He Z, Li J, Liu DX, *et al*. Tracking by animation: Unsupervised learning of multi-object attentive trackers. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1318–1327. [doi: 10.1109/CVPR.2019.00141]

7 Jiang JD, Deng F, Singh G, *et al*. Object-centric slot diffusion. Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: ACM, 2023. 375.
8 Wu ZY, Hu JY, Lu WY, *et al*. SlotDiffusion: Object-centric generative modeling with diffusion models. Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: ACM, 2023. 2216.
9 Zhu YF, Joshi A, Stone P, *et al*. VIOLA: Imitation learning for vision-based manipulation with object proposal priors. Proceedings of the 6th Conference on Robot Learning. Auckland: PMLR, 2023. 1199–1210.
10 Locatello F, Weissenborn D, Unterthiner T, *et al*. Object-centric learning with slot attention. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: ACM, 2020. 967.
11 Engelcke M, Parker Jones O, Posner I. GENESIS-V2: Inferring unordered object representations without iterative refinement. Proceedings of the 35th International Conference on Neural Information Processing Systems. ACM, 2021. 618.
12 Singh G, Deng F, Ahn S. Illiterate DALL-E learns to compose. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.
13 Seitzer M, Horn M, Zadaianchuk A, *et al*. Bridging the gap to real-world object-centric learning. Proceedings of the 11th International Conference on Learning Representations. Kigali: OpenReview.net, 2023.
14 Singh G, Wu YF, Ahn S. Simple unsupervised object-centric learning for complex and naturalistic videos. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: ACM, 2022. 1322.
15 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: ACM, 2020. 574.
16 Sohl-Dickstein J, Weiss E A, Maheswaranathan N, *et al*. Deep unsupervised learning using nonequilibrium thermodynamics. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille: ACM, 2015. 2256–2265.
17 Kingma DP, Welling M. Auto-encoding variational Bayes. Proceedings of the 2nd International Conference on Learning Representations. Banff: OpenReview.net, 2013.
18 Greff K, Van Steenkiste S, Schmidhuber J. Neural expectation maximization. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6694–6704.
19 Yuan JY, Li B, Xue XY. Spatial mixture models with learnable deep priors for perceptual grouping. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2019. 9135–9142. [doi: 10.1609/aaai.v33i01.33019135]
20 Greff K, Kaufman RL, Kabra R, *et al*. Multi-object representation learning with iterative variational inference. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 2424–2433.
21 Lin ZX, Wu YF, Peri SV, *et al*. SPACE: Unsupervised object-oriented scene representation via spatial attention and decomposition. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: OpenReview.net, 2020.
22 Eslami SMA, Heess N, Weber T, *et al*. Attend, infer, repeat: Fast scene understanding with generative models. Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona: ACM, 2016. 3233–3241.
23 Yuan JY, Li B, Xue XY. Generative modeling of infinite occluded objects for compositional scene representation. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 7222–7231.
24 Crawford E, Pineau J. Spatially invariant unsupervised object detection with convolutional neural networks. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2019. 3412–3420. [doi: 10.1609/aaai.v33i01.33013412]
25 Burgess CP, Matthey L, Watters N, *et al*. MONet: Unsupervised scene decomposition and representation. arXiv:1901.11390, 2019.
26 Engelcke M, Kosiorek AR, Jones OP, *et al*. GENESIS: Generative scene inference and sampling with object-centric latent representations. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: OpenReview.net, 2019.
27 Jia BX, Liu Y, Huang SY. Improving object-centric learning with query optimization. Proceedings of the the 11th International Conference on Learning Representations. Kigali: OpenReview.net, 2023.
28 Luo C. Understanding diffusion models: A unified perspective. arXiv:2208.11970, 2022.
29 Lugmayr A, Danelljan M, Romero A, *et al*. Repaint: Inpainting using denoising diffusion probabilistic models. Proceedings of the 2022 IEEE/CVF Conference on Computer

Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11451–11461. [doi: 10.1109/CVPR52688.2022.01117]

30 Rombach R, Blattmann A, Lorenz D, *et al*. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 10674–10685. [doi: 10.1109/CVPR52688.2022.01042]

31 Liu N, Li S, Du YL, *et al*. Compositional visual generation with composable diffusion models. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 423–439. [doi: 10.1007/978-3-031-19790-1_26]

32 Liu N, Du YL, Li S, *et al*. Unsupervised compositional concepts discovery with text-to-image generative models. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 2085–2095. [doi: 10.1109/ICCV51070.2023.00199]

33 Karazija L, Laina I, Rupprecht C. ClevrTex: A texture-rich benchmark for unsupervised multi-object segmentation. Proceedings of the 2021 Neural Information Processing Systems Track on Datasets and Benchmarks. NeurIPS, 2021.

34 Greff K, Belletti F, Beyer L, *et al*. Kubric: A scalable dataset generator. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 3739–3751. [doi: 10.1109/CVPR52688.2022.00373]

35 Huang YX, Chen TL, Shen ZM, *et al*. OCTScenes: A versatile real-world dataset of tabletop scenes for object-centric learning. arXiv:2306.09682, 2023.

36 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

37 聂晖, 王瑞平, 陈熙霖. 开放世界物体识别与检测系统: 现状、挑战与展望. 计算机研究与发展, 2024, 61(9): 2128–2141. [doi: 10.7544/issn1000-1239.202440054]

38 徐梓雄, 郭璠, 王宗雨, 等. 基于多视角学习策略的手部姿态估计. 计算机系统应用, 2023, 32(10): 22–33. [doi: 10.15888/j.cnki.csa.009291]

39 陈国军, 付云鹏, 于丽香, 等. 自适应多尺度特征融合的单目图像深度估计. 计算机系统应用, 2024, 33(7): 121–128. [doi: 10.15888/j.cnki.csa.009587]

(校对责编: 张重毅)