arXiv:2302.03531v2 [cs.CV] 2 Sep 2024

用于场景理解的结构化生成模型

Christopher K. I. Williams

信息学院，爱丁堡大学，英国爱丁堡 Crichton 街 10 号，EH8 9AB。

贡献作者：c.k.i.williams@ed.ac.uk；

摘要

本文主张使用 *structured generative models*（结构化生成模型）来理解静态场景。这需要从输入图像（或一组多视图图像）中重建三维场景，从而根据实例化对象的模型对图像内容作出因果解释；每个对象都有自己的类型、形状、外观和姿态，同时还涉及场景光照和相机参数等全局变量。这种方法还需要能够描述场景中物体共现关系及相互关系的场景模型。结构化生成模型方法的优点在于其具有组合性和生成性，因此具备可解释性和可编辑性。

为了推进结构化生成模型的研究，我们需要构建物体和场景模型，并采用相应方法进行推断。我们首先回顾物体模型，其中包括“实体”（形状定义明确的物体类别）和“物质”（空间范围不规则的类别）。随后，我们将回顾 *scene models* 用于描述物体间相互关系的模型。对场景生成模型而言，最具挑战性的问题或许是 *inference* 如何根据单张或多张图像输入，推断物体的形状、光照和相机参数，以及场景中的相互关系。最后，我们将讨论推进结构化生成模型研究亟待解决的问题。

关键词：结构化生成模型、生成模型、组合性、场景理解

1 引言

本文旨在推动使用结构化生成模型 （SGM）来理解静态场景。这些模型置于计算机视觉的经典框架之中：在该框架下，根据一幅或多幅输入图像重建三维场景。在这种情况下，图像内容通过实例化物体的模型得到因果解释；每个物体都有自己的类型、形状、外观和姿态，并伴有场景光照和相机参数等全局变量。这种方法还需要场景模型 来描述场景中物体的共现关系及相互关系。因为

这类模型能够生成 （或重建、解释）场景，而且它们具有结构化 特征（即由多个物体及其关系构成），因此我们称之为结构化生成模型。

这种重建框架也称为分析合成 (Grenander, 1978年），或视觉即逆向图形学（*VIG*）(Kulkarni、Whitney、Kohli 和 Tenenbaum，2015；Moreno、Williams、Nash 和 Kohli，2016）。它可以追溯到20世纪60年代早期的“积木世界”研究（Roberts, 1963年）。这一研究方向的其他早期工作包括 Hanson 和 Riseman(1978年）的 VISIONS 系统，以及Ohta、Kanade 和 Sakai(1978年）用于户外场景

理解。例如，VISIONS 系统使用了多个层次的分析（如物体、表面），并将图像特定的解析结果与关于场景的通用知识相映射。

VIG框架的替代方案要么是判别方法，即给定输入图像，预测某个目标量或多个目标量；要么是非结构化生成模型。判别方法通常用于解决特定任务，例如物体检测或语义分割，而这些任务通常在图像空间中定义。这些任务通常被设置为监督学习任务，因此需要标注数据。目前，深度神经网络（DNN）是此类任务中占主导地位的首选方法。DNN通常具有很高的准确率，但作为以判别方式训练的模型，它们有时会严重失效，产生荒谬的错误，[1] 却没有任何有效迹象表明其结果不可靠。这方面的一个例子是模型在对抗样本上的性能失效（例如参见Szegedy等人 2020年）。此外，由于判别模型是在特定数据集上训练的，当面对相同任务但数据集统计特性不同的新数据集时，它们通常表现不佳（分布偏移）。对特定任务评估的关注意味着，无需利用多个任务的预测结果，就无法构建关于输入图像所对应三维世界的一致理解；例如，Zamir等人（2020年）阐述了这一点。

借助非结构化生成模型，图像由一组潜变量生成，并未显式地对物体及其相互作用进行建模。例如，Radford、Metz和Chintala（2016）利用生成对抗网络（GANs）生成了卧室场景图像。这种方法用一个单一的潜在向量表示整个场景，而该表示并未在不同物体之间解耦。较新的图像生成模型也采用了类似的向量表示，例如基于Transformer解码器的DALL-E（Ramesh等人，2021年）和Gemini（Gemini团队，2023年），以及扩散模型，例如Rombach、Blattmann、Lorenz、Esser和Ommer（2022）提出的模型和GLIDE（Nichol等人，2022）。（这些较新的模型可以根据文本的向量表示或其他输入模态进行条件化。）尽管这些模型生成的结果令人印象深刻，但潜在表示的非结构化特性意味着，难以对其进行编辑以实现对场景的特定修改（例如，更改床罩的颜色），因为这种表示不可解释。

需要明确的是，我们并不是反对在计算机视觉中使用深度神经网络。然而，对于结构化生成模型而言，深度神经网络可以用于特定的建模和推断任务（如下文所述），而不是充当一个庞大的黑箱。与监督式深度神经网络不同，结构化生成模型提供了一种连贯的场景表示，可支持包括新任务在内的多种任务。使用场景模型可以将自上而下的信息流（例如有关场景类型的信息）引入物体检测和图像分割等任务中。场景模型的结构化特征意味着其表示是可解释的，从而可以编辑场景，例如添加物体，或改变场景中物体的属性。场景模型还可以具有组合式结构，即由更基本的子结构以层次化方式构成。正如 Yuille 和 Liu（2021年）所论述的，这应当使场景生成模型能够处理可能图像的组合爆炸，而无需指数级规模的训练数据。下文将进一步讨论这些问题，尤其是在第 1.3 节中。

图 1 展示了如何根据三维物体、相机姿态和光照来解释输入图像，从而生成重建图像。有时，构建完整的三维版本可能过于繁琐，因此我们可以考虑分层的“2.1D”模型，其中物体（人）由描述其形状和外观的“精灵”模型表示，并包含背景，如图 2所示。这些图层具有深度排序，因此可以解释遮挡。

尽管为简化起见，本文将研究范围限定于静态 场景理解，但结构化生成模型方法也可以应用于动态场景。在这种情况下，物体、相机和光照都可能随时间变化（即移动、变换等），同时还需要对物体间交互进行动态建模。

本文的目标是：推广结构化生成模型（SGM）的观点；回顾物体与场景建模以及基于场景生成模型的推断相关工作；

[1] “荒谬的错误”这一表述借用了 Daniel Kahneman 在 2021年12月 NeurIPS 会议上的演讲内容。

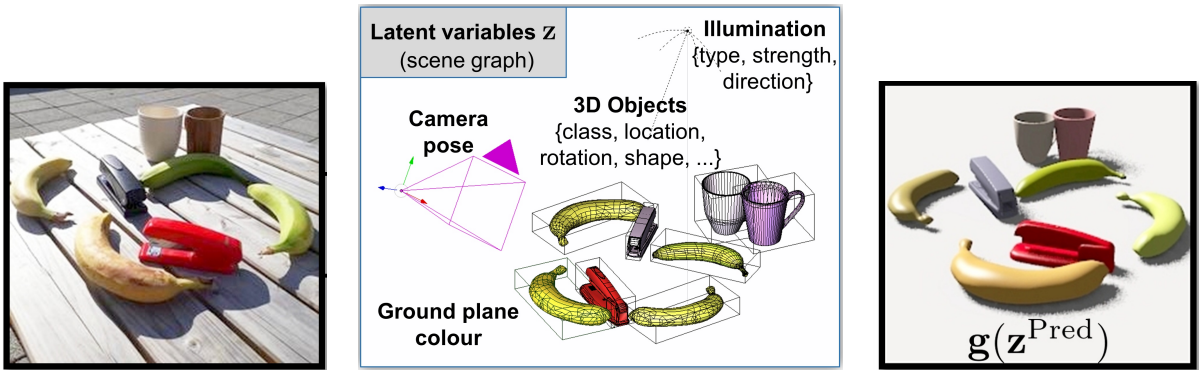


图1 该 输入图像（左）通过三维物体、相机姿态和光照进行解释，从而生成其 e
重建结果 图像（右）。图像来自Romaszko、Williams 和 Winn（2020年）。

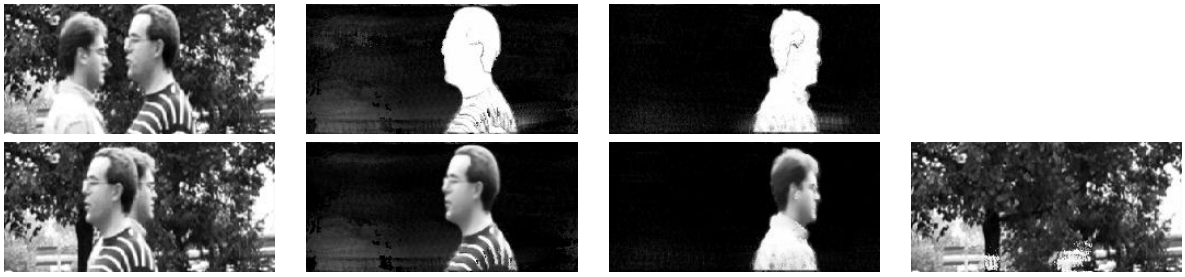


图2 最左侧面板展示了视频中的两帧画面，画面中两个人在背景前相向走过。第二个面板展示了学习得到的第一个精灵的掩码（上方）和外观。第三个面板展示了第二个精灵的相同内容。最右侧面板展示了学习得到的背景。图像来自Williams 和 Titsias（2004年）。

并指出仍待进一步研究的空白与未决问题。本文结构如下：第 1.1 节讨论与场景理解相关的丰富多样的任务。第 1.2 节介绍生成模型的一般优势，第 1.3 节讨论结构化生成模型的优缺点。第2节介绍物体建模，包括第2.1节中的“实体”（形状定义明确的物体类别），其中包括基于部件的模型，以及第 2.2节中的“物质”（空间范围不定形的类别）。第 3 节介绍物体间关系模型，重点讨论室内场景。在定义了物体和场景的模型后，第 4 节讨论如何根据单幅或多幅输入图像对结构化生成模型进行推断。第 5 节讨论推进结构化生成模型研究议程所需解决的问题，包括数据集和基准。

请注意，本文涉及的范围非常广，并不旨在为每个主题提供全面的参考文献。事实上，这些必需主题涵盖了计算机视觉教科书中的大量内容。本文旨在通过具有代表性的研究成果，展示各个主题。

1.1 场景理解

在更普遍地讨论场景理解之前，我们先来看一下图3中的两个示例图像，看看我们能从中提取出什么信息。这些并不是特别复杂的场景——我们很容易找到包含更多物体和关系的图像。先看户外场景，即图3（a）——我们可以判断，这里既不是人口密集的“市中心”区域，也不是非常偏远的农村地区。我们可以识别出一些物体：一个由6头牛组成的小畜群（其中一头可能是小牛，大部分身体隐藏在图像中央附近的白棕相间奶牛后面）；5辆机动车辆（其中一辆被部分遮挡）；一座建筑物、一些路灯杆、一些树木以及一些道路标志。我们对物体的关注，可能使我们忽略了图像中很大一部分属于“物质”类别，例如路面



(a) (b)

图3 来自 PASCAL VOC 2008数据集的两张示例图像：2008 001062 和 2008 000043。

和草地。奶牛在草地上，这很合理，因为它们可以在那里吃草，但不会在路面上吃草。

第二个场景，图3(b)，是一间餐厅的室内场景。我们可以看到一张桌子、6把椅子（设计相同或相似）、桌上的5张桌垫、一个时钟和一盏灯具。房间里有一扇外门、一扇窗户、一个凹入区域，左侧还有一个大型木制柜子（？）。抛光木地板能映出一些倒影，墙壁呈黄绿色。仔细观察还能发现一些较小的细节，例如电灯开关、一个低位散热器、椅子上的一些物体（其中可能包括一张儿童增高椅），以及右下角的一部分室内植物。透过窗户还能看到一些室外栏杆，这表明外门外的区域有一段向上的楼梯。

那么，什么是场景理解？其中一部分是识别可见的物体和物质，但这还不止于此。对于每个物体，我们希望了解其类别、形状及其相对于场景的姿态，以及它由何种材料制成。我们还希望理解相机参数（例如观察者视点）和光照；这将有助于解释遮挡和阴影。例如，在图3(a)中，给定牛的阴影，可以推断太阳的方向。这种三维表示能够提出反事实问题，例如预测移除或添加一个物体后图像将如何变化，或者从一个新方向观察时图像将如何变化（新视角合成，NVS）。它还支持交互场景中的智能体进行交互，例如尝试将牛赶成畜群。

场景理解还包括识别场景类型（例如餐厅），这会使我们对应该（以及不应该）出现哪些物体产生预期。此外，场景理解还涉及物体之间的空间、功能和语义关系。[2] 例如，在图 3(a) 中，我们可能会惊讶地发现，牛没有被围栏与道路隔开以减少碰撞，但这可能取决于拍摄图像的位置所遵循的规范。而在图 3(b) 中，关于餐厅的知识意味着我们很可能会预期桌子上摆放的桌垫数量与椅子数量相同——事实上，仔细观察图像会发现，“缺失”的桌垫被放在了右侧椅子的座位上。

场景理解的一种丰富方法是回答问题。在视觉问答（VQA；例如，参见 Antol等，2015）中，系统需要用文本回答有关图像的问题。GPT-4（OpenAI，2023年）等近期系统能够生成令人印象深刻的输出，尽管它们也会信心十足地犯错。不过，对于某些问题，这或许并非最佳回答方式，例如“图像中心附近的黑色奶牛身上有哪些像素？”。PASCAL VOC 挑战赛（Everingham、Van Gool、Williams、Winn 和 Zisserman，2010）提出了三个问题：对于分类，问题是“图像中是否存在类别为 X 的物体？”。对于

[2]请参阅https://ps.is.mpg.de/researchfields/semantic-scene-understanding。

目标检测，任务是预测图像中每个类别为 X 的物体的边界框。对于 图像分割，任务是为每个像素标注一个已知类别标签或背景。对于后两项任务，用文本回答并不是最自然的方式。Gemini（Gemini团队，2023年）等近期系统能够生成多模态输出（例如文本和图像），这有助于克服文本式视觉问答的局限性。然而，值得注意的是，这类系统使用的是纠缠的向量化潜在状态，而非结构化潜在状态，因此存在上文所述的局限性。

“一图胜千言”这句格言也表明，文本并非描述场景最高效的方式，尤其是考虑到自然语言的歧义性。相反，场景生成模型为场景提供了一种特定领域语言。OpenAI 的文本到图像系统 DALL-E（Ramesh等人，2021年）能够根据“几个人坐在一张木质长凳上”这样的提示生成令人印象深刻的结果，甚至也能响应“一个由手风琴制成的貘”这样古怪的提示（见论文中的图 3 和图 2(a)）。然而，据报道，要求生成三个以上的物体、使用否定表达或指定数字，可能会导致错误；物体的特征也可能出现在错误的物体上（Marcus、Davis 和 Aaronson，2022）。问题之一在于，从网络上获取的配对文本和图像中，文本可能无法充分描述物体及其空间关系等。在 DALL-E 中，文本输入会被压缩为向量表征，随后用于生成图像。因此，DALL-E 及类似系统不太可能为平面艺术家和动画师提供足够的控制权；他们或许大体上满意系统的输出，但可能仍希望进行调整和编辑。我们认为，要实现这一点，就需要采用前文所倡导的基于物体的表征和场景模型。

替代场景表征

上文将由单个潜变量向量构成的非结构化生成模型与结构化生成模型进行了对比。下文将讨论另外两种场景表示。

在语义场景补全 (SSC) 中，例如参见Song 等人 (2017)，目标是预测以下两者：视锥内体素的占据状态和物体类别标签[3]，包括已观测表面上的体素以及遮挡区域中的体素。这种表示具有结构化生成模型（SGM）的一些特征，因为它描述了场景的几何结构，因此适用于新视角合成。然而，与结构化生成模型（SGM）相比，基于体素的类别标签是一种信息较为贫乏的表示，因为它没有明确表示物体。对已标注体素进行后处理，大致可以实现这一点，但这种方法无法分割彼此接触且属于同一类别的物体（例如，放在床上的两个枕头）。此外，类别标签几乎无法提供有关物体外观的信息。

对基于体素的表示建立生成分布的模型非常具有挑战性，因为它将各个物体的形状、它们之间的关系以及场景的几何结构都整合到一个数据结构中。我们认为，与其试图在体素关系层面对此进行整体建模，不如将物体模型、它们之间的相互关系及几何结构组合 起来。

将条件生成模型组合起来，以实现对场景更精细的控制，这一思路已得到研究，例如Du 等人（2023年）。他们展示了如何使用 AND、OR 和 NOT 运算符组合多个扩散模型。作者举例说明，如何生成一幅图像，描绘“晴朗天气以外的某一天，一匹马在沙滩上或草原上”。这句话被解析为“一匹马” AND（“沙滩” OR “草原”）AND（NOT“晴朗”），并使用了四个条件扩散模型，分别用于马、沙滩、草原和晴朗场景。这种方法具备场景生成模型的一些理想特性，因为它是组合式的，不过在这里，组合是在图像空间中完成的。但请注意，场景中的物体并没有明确表示，因此，例如，这种方法无法直接控制和编辑上述示例中马的姿态、形状和外观。此外，相机参数也不可用，因此这里无法进行新视角合成。

[3]包括自由空间的空标签。



图4 图像修复任务：将绿色矩形区域抹去后进行修复，图像取自 PASCAL VOC 2008数据集中的 2008 000959。

1.2 生成模型的一般优势

我们首先定义一些符号：$\mathbf{x}$ 表示观测数据，例如一幅图像。生成模型 $p_\theta(\mathbf{x})$ 定义了图像上的概率分布；这里θ 表示模型的参数。对于场景生成模型，我们的模型用潜变量$\mathbf{z}$来定义。单个物体的潜变量可以分解为，例如$\mathbf{z} = (\mathbf{z}^s、\mathbf{z}^t、\mathbf{z}^p)$，分别表示形状、纹理（包括颜色）和姿态。如果有 K 个物体，它们各自可以具有潜变量（LV）$\mathbf{z}_1$，...，$\mathbf{z}_K$；令$\mathbf{z}_0$ 表示全局变量（例如相机参数和光照）。$\mathbf{z}$ 中还可以包含额外的潜在结构，用于建模场景中物体间的相互关系。我们有

$$p_\theta(\mathbf{x}) = \int p(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z})\, d\mathbf{z}, \tag{1}$$

其中，$p(\mathbf{z})$ 是潜变量的先验分布，而 $p_\theta(\mathbf{x}|\mathbf{z})$ 则将潜变量描述渲染为图像。[4]

生成模型有几个优点，如下所述。

[4]在式 1 中，对于离散潜变量，积分应解释为求和。

模式合成（无条件生成）：

生成模型允许从模型中进行采样，并将采样结果与原始输入数据进行比较。这会很有帮助，尤其是有助于找出模型未能捕捉输入数据某些方面的情形，从而修订模型。模型批判的方法模型批判(例如，Seth、Murray 和 Williams，2019)有助于检测差异；也可以训练判别器来完成这一任务，正如生成对抗网络的训练中所采用的方法。

在某些情况下，这类合成数据（即从模型中采样得到的数据）本身就很有用——例如，在医疗保健领域，出于隐私问题，人们可能不愿公开数据。但如果能够创建出仿照底层数据分布的合成数据，就能让研究更加自由地开展。不过，必须谨慎确保合成数据并非只是“记住”了部分或全部训练数据，正如van den Burg 和 Williams（2021年）所讨论的。

插补与恢复（条件生成）：

如果输入数据$\mathbf{x}$ 被拆分为观测数据 $\mathbf{x}^o$和缺失数据 $\mathbf{x}^m$，那么插补任务就是根据 $\mathbf{x}^o$预测缺失数据。概率模型还会生成 $p(\mathbf{x}^m|\mathbf{x}^o)$ 的概率分布，从而能够量化不确定性。如果图像的一部分缺失，插补任务也可称为图像修复；示例见图 4 。在此，简单的方法可能只是根据车身和柏油路面的红色与黑色纹理进行修补，但关于汽车的知识会让模型预测出该位置有一个车轮，而且它很可能与可见的前轮风格一致。

观测数据也可能是底层数据的含噪或退化版本；在这种情况下，若拥有底层数据和噪声过程的模型，就可以对数据进行概率恢复。

异常检测：
检测不符合学习到的模型 $p(\mathbf{x})$ 的数据点可能会很有帮助。这项任务称为异常检测、新颖性检测或分布外（*OOD*）检测。例如，自动化系统可以检测

运行状态是否发生了变化，从而提醒需要关注或重新训练。一种描述该任务的方式，是将其视为在 $p(\mathbf{x})$ 与广义的“杂物捕捉器”模型之间进行分类$p_0(\mathbf{x})$，例如 Quinn、Williams 和 McIn-tosh (2009) 所采用的方法。如果某个数据点 $\mathbf{x}$ 在$p_0(\mathbf{x})$ 下的可能性更高，那么相对于$p(\mathbf{x})$，它就可以被分类为离群值。

异常可能非常隐蔽。在北美街景图像中，消防栓和邮箱都是常见的街道设施，但消防栓位于邮箱顶部的情况并不常见——这个上下文异常的例子出自Biederman、Mezzanotte 和 Rabinowitz(1982，图1）。

数据压缩：
概率模型 $p(\mathbf{x})$ 可用于压缩数据。给定真实的数据分布 $p(\mathbf{x})$，香农信源编码定理将为 $\mathbf{x}$ 分配长度为 $l(\mathbf{x}) = -\log_2 p(\mathbf{x})$ 比特的编码。因此，期望码长为$-\int p(\mathbf{x}) \log_2 p(\mathbf{x}) \, d\mathbf{x} = H(p)$，即 p 的熵。实际中可以采用例如算术编码等方法进行此类数据压缩，参见 MacKay（2003，第6.2节）。

在实践中，我们可能不知道真实分布 $p(\mathbf{x})$，但有一个替代模型 $q(\mathbf{x})$。在这种情况下，我们需要付出代价，即使用的比特数的期望值。设在 $q(\mathbf{x})$ 下编码时的期望码长记为 L_q。那么

$$
\begin{aligned}
L_q &= -\int p(\mathbf{x}) \log_2 q(\mathbf{x}) \, d\mathbf{x} \\
&= -\int p(\mathbf{x}) \log_2 \left(\frac{q(\mathbf{x})}{p(\mathbf{x})} p(\mathbf{x}) \right) \, d\mathbf{x} \\
&= H(p) + D_{KL}(p \parallel q) \geq H(p). \qquad (2)
\end{aligned}
$$

因此，额外的期望码长由 Kullback-Leibler（KL）项给出；当 $q = p$ 时，该项当然会降为零。这促使我们最小化 KL 散度以生成更好的编码，或者等价地，最大化模型 $q(\mathbf{x})$ 的期望对数似然 $\int p(\mathbf{x}) \log q(\mathbf{x}) \, d\mathbf{x}$（因为熵项是固定的）。

1.3 结构化生成模型的优缺点

下面我们将结构化生成模型（SGM）与判别模型以及非结构化生成模型进行对比。

✓ 结构化生成模型提供连贯的场景表示，而非仅仅针对一组彼此无关的任务给出输出预测。这种表示可用于多种任务，包括此前未曾训练过的新任务（迁移学习）。

✓ 结构化生成模型具有组合式特性。这意味着，在学习某种特定对象类型时，我们无需同时考虑其他对象类型。这应该能提高数据效率。

如果存在对象间关系，就可以在层次模型中将这些关系与单个物体的变化分别建模。这样便能进行课程学习（Bengio、Louradour、Collobert 和 Weston，2009）：先利用类别条件数据，专注于对单个物体类别进行建模，然后再引入场景内关系。这些优势在非结构化生成模型中并不存在。

✓ 正如 Yuille 和 Liu (2021年，第 7 节）所指出的，视觉问题存在组合爆炸。如我们所见，单个物体在对象类型、姿态、形状和外观方面有许多自由度。而场景中多个物体的出现又进一步倍增了这些可能性。组合性假定场景是由更基本的子结构以层次化方式组合而成的，这与语言的“无限生产力”类似。通过利用这种组合结构，场景生成模型应能泛化到组合情形，而无需大量呈指数级增长的训练数据。

✓结构化生成模型（SGM）的表示是可编辑的，例如，可以改变光照方向，或添加／移除物体。用因果语言来说，编辑 SGM 会对场景表示进行干预。
✓ （结构化）生成模型可以通过无监督方式或弱监督方式进行训练。判别模型通常需要可能成本高昂或难度较大的人工标注，不过

也有人探索了使用较弱监督的方法（例如，参见 Shi、Caesar 和 Ferrari 2017）。

阻碍完全无监督学习的一个因素是Locatello 等人（2019）指出：“如果模型和数据都没有归纳偏置，那么无监督地学习解缠表示从根本上是不可能的。”因此，假设我们希望解缠某个物体类别的形状、纹理和姿态等因素，就需要在模型中引入归纳偏置。不过，正如第2.1节所述，常见的物体模型确实具有这类归纳偏置，因此这不应成为场景生成模型面临的问题。

✓结构化生成模型具有可解释性。这种结构化方法与许多深度生成模型形成对比：后者能够学习丰富的数据模型，但采用的是单一的黑箱架构，既不可解释，也不易编辑。结构化生成模型会识别出由特定物体解释的图像区域，并有可能提供更详细的部件级对应关系。结构化表示还支持遮挡推理等其他功能。

✗ 判别模型不太容易受到建模局限的影响，因为它们直接针对给定任务进行“端到端”优化，而不是构建一个可用于多种不同任务推断的通用模型。

✗结构化生成模型框架可能需要耗费大量资源的推断过程，才能推断整个场景的潜变量。我们将在下文第 4节讨论如何缓解这些问题。

SGM 方法可能很有帮助的一个例子是：当某个物体被严重遮挡，但场景上下文有助于重建它时。请看图。5(a)，其中最靠后的红色椅子被严重遮挡。[5] 了解场景中围绕桌子摆放的椅子通常采用相同设计这一信息，有助于对这把严重被遮挡的椅子做出准确预测。可以通过对最后方的椅子进行实例分割来评估这一点；要正确地将地毯上可见的金属框架部件归属于这把椅子，很可能需要了解椅子的设计。其他评估方式还可以包括输出物体的三维模型，或预测场景从新视点看起来会是什么样子，如图所示。5(b)。图像修复是一项相关任务，如图所示。4。在这种情况下，实际上存在一个合成遮挡物（即掩码）；最自然的评估方式是预测被掩码遮挡的区域。

结构化生成模型方法还可用于执行场景编辑，例如移除物体或更改其属性，或调整光照。此时，可以通过收集相关扰动下的视图来评估结果。另一项可能的任务是，仅给定场景中一部分物体，补全三维场景（例如，参见M. Li 等人 2019）。这是一项缺失数据插补任务，类似于图像修复，但有所不同，因为它要插补的是三维场景，而不仅仅是图像。要成功完成这项任务，场景模型需要指定适合该场景类型的新物体，并准确确定它们相对于已有物体的位置。

完整推断是否过度了？

SGM 方法的主要优势在于，它提供了一种统一的表示，由此可以处理许多不同的任务或问题。与针对不同任务分别训练不同模型、却缺乏这种底层结构可能产生的结果相比，这种方法能够以三维（或 2.1D）世界的形式，对输入图像形成连贯的理解。

如果我们只关注一项任务，通常就会显得有些小题大做。但在现实世界中行动，并不只需要完成一项任务。Land、Mennie 和 Rusted (1999) 对泡茶的例子很好地说明了这一点。泡茶这一目标可以分解为若干子目标，例如“把水壶放上去”“泡茶”“准备杯子”。把水壶放上去这一子目标又包括“给水壶装水”，而这又需要“找到水壶”“拿起水壶”“取下盖子”“拿到水槽边”“找到并打开水龙头”“把水壶移到水流下”“水满后关掉水龙头”“盖上盖子”以及“拿回厨房台面”。所需的视觉任务包括物体检测；对物体的姿态和形状进行估计，以便操纵它们；以及监测某些变量的状态（例如水壶中的水位）。在陌生厨房中泡茶，还需要运用有关厨房常见布局的知识，并且可能

[5]此示例的灵感来自Hueting 等人 (2018)。



(a)



(b)

图 5 从两个视点拍摄的办公室场景图像。

还需要了解不同类型的水龙头机构。请注意，“找到并打开水龙头”和“关掉水龙头”等子目标可以重复用于其他任务，例如煮咖啡或洗碗。

2 物体模型

视觉场景可能非常复杂。构成场景的组成部分可以分为“实体”和“物质”。实体是形状定义明确的物体类别（如人或汽车），而物质则对应空间范围不定形的类别，例如草地和天空（参见例如 Sun、Kim、Kohli 和 Savarese，2014）。我们首先在第 2.1 节聚焦于对实体建模的方法，然后在第 2.2 节转而讨论对物质建模。

2.1 实体建模：多因素模型

这里我们考虑对一类视觉物体（例如茶壶或汽车）进行建模。这些物体的形状和纹理（例如由颜色或表面上更复杂的图案描述）可以有所不同。我们还可以改变相机相对于物体的位置和朝向（即姿态），以及光照；我们将这些称为渲染变量。因此，形状、纹理和渲染各自构成独立因素，它们组合起来产生观测结果。包含多个独立因素的模型称为**多因素**模型。下面我们先举几个多因素模型的例子，然后在第2.1.1 节重点介绍基于部件的模型（它们是多因素模型中的一个特殊类别）。

示例：***Blanz*** 和 ***Vetter*** 的人脸可变形模型。

Blanz 和 Vetter（1999）提出了一个早期的三维多因素模型。[6] 考虑面部上的 n个位置；$\mathbf{s}_i$ 记录位置 i的（x、y、z）坐标，$\mathbf{t}_i$ 则记录同一位置的红、绿、蓝颜色值（反照率）。这些测量值由激光扫描仪获取。将各个向量连接起来，得到长度为 $3n$的形状向量 $\mathbf{s} = (\mathbf{s}_1, \mathbf{s}2, \ldots, \mathbf{s}_n)$；同样，也有一个长度为 $3n$的纹理向量 $\mathbf{t}$。Blanz 和 Vetter（1999）在每张人脸上使用了约 70,000 个位置，并收集了 200 名受试者的数据。预处理用于消除人脸之间的全局三维变换，并采用光流法对各个位置进行配准。

给定每个对象的形状向量和纹理向量，可以构建概率主成分分析（PPCA）模型，以捕捉形状和纹理的变化，其潜变量分别为 $\mathbf{z}^s$ 和 $\mathbf{z}^t$。另一种做法是使用一个共同的 $\mathbf{z}$ 来表示形状和纹理变化，例如，在应用 PPCA 之前，将每个示例的$\mathbf{s}$ 向量与 $\mathbf{t}$ 向量拼接起来。这样便可以对嘴巴形状改变以形成微笑时也可能露出牙齿这一现象进行建模，因为这两种变化是相关的。

上述描述对三维中的形状和纹理变化进行了建模。该模型根据相机的内参数和外参数，在几何上变换到像平面中。

[6] 以下描述部分基于Prince （2012年）以及原始论文。

每个像素的颜色由冯氏着色模型决定（例如参见 Szeliski 2021年，第 2.2.2 节）。该模型考虑了定向光源产生的漫反射和镜面反射，以及环境光照。将所有渲染变量记为 $\mathbf{z}^r$，通过优化 $\mathbf{z}^s$、$\mathbf{z}^t$ 和 $\mathbf{z}^r$，使观测像素与预测像素之间的误差度量最小，从而将整体模型拟合到新面部。

Blanz 和 Vetter 的模型1999年是三维的。这类三维模型也曾用于例如大脑区域建模（Babalola、Cootes、Twining、Petrovic 和 Taylor，2008）。Cootes、Taylor 及其合作者早期的一些工作首先开发了二维主动形状模型，使用由地标点定义形状的概率主成分分析模型（Cootes、Taylor、Cooper 和 Graham，1995），随后又开发了主动外观模型，将纹理也纳入考虑（Cootes、Edwards 和 Taylor，1998年）。

示例：***CodeNeRF*** 模型为物体类别构建解耦的神经辐射场。

Blanz 和 Vetter 的模型采用线性方法（PPCA）对形状和纹理引起的变化进行建模。为了使这种方法奏效，需要仔细对齐数据。随着深度学习的兴起，自然会想到能否利用更强大的非线性模型。我们首先从Milden- hall 等人 2020年提出的单个物体神经辐射场（NeRF）表示开始，然后像 Jang and Agapito 2021年的 CodeNeRF 模型那样，添加潜变量来对形状和纹理变化进行建模。

NeRF 的输入是一个三维位置 $\mathbf{x} = (x, y, z)$ 和一个由三维笛卡尔单位向量 $\mathbf{d}$ 定义的观察方向，输出该位置和方向处的体密度 $\sigma(\mathbf{x})$ 以及发射颜色 $\mathbf{c}(\mathbf{x}, \mathbf{d}) = (r, g, b)$。这些值由一个神经网络 F_θ 得到：$(\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}\ \sigma)$，其权重为 θ。正如 Mildenhall 等人（2020年，第 4 节）所述，“体密度 $\sigma(\mathbf{x})$ 可以解释为光线在位置 $\mathbf{x}$ 处终止于一个无穷小粒子的微分概率”。这意味着，在近界限 t_n 和远界限 t_f 内，相机光线 $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ 上观测到的期望颜色 $C(\mathbf{r})$ 由[7]给出

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})\ dt, \qquad (3)$$

其中

$$T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))\ ds\right). \qquad (4)$$

这里，$T(t)$表示从 t_n 到 t沿光线的透射率，即光线从 t_n 出发到达 t且未碰到任何其他粒子的概率。求取 $C(\mathbf{r})$的过程称为体渲染，且可微。通过沿光线进行若干次采样，可近似计算 $C(\mathbf{r})$。

对于单个物体，可以通过最小化多个不同视图（已知相机位姿和内参）中一组光线位置处观测颜色与预测颜色之间的误差，获得 NeRF 表示。这有助于实现新视角合成，即预测从一个新视角拍摄的图像。然而，对于一个物体类别，为每个物体设置潜变量$\mathbf{z}^s$，$\mathbf{z}^t$ 是合理的，CodeNeRF（Jang &Agapito，2021年）便采用了这种做法。这里，神经网络经过扩展，可映射$F_\theta : (\mathbf{x}, \mathbf{d}, \mathbf{z}^s, \mathbf{z}^t) \rightarrow (\mathbf{c}\ \sigma)$。此时，网络权重 θ 的优化问题是在某个物体类别的所有训练样本上进行的；对于给定样本的每组视图，则估计其形状和纹理潜变量。（同时还对潜变量施加与 $|\mathbf{z}^s|^2 + |\mathbf{z}^t|^2$ 成正比的正则化惩罚项，对应于高斯先验。）图6展示了 CodeNeRF 泛化到新形状／纹理／姿态组合的能力。

NeRF 研究还带来了许多其他近期发展。例如Zhang 等人（2021年）将 NeRF 扩展为提取表面（而非体积）表示，进而求解空间变化反射率和环境光照。这样便可以在任意环境光照下渲染物体的新视角，并编辑物体的材质属性。

[7]请注意，在本节中，$\mathbf{x}$ 表示一个三维位置，而不是我们通常所说的输入图像。此外，t 有多重含义，既表示沿光线$\mathbf{r}(t)$的参数化，也表示纹理上标。

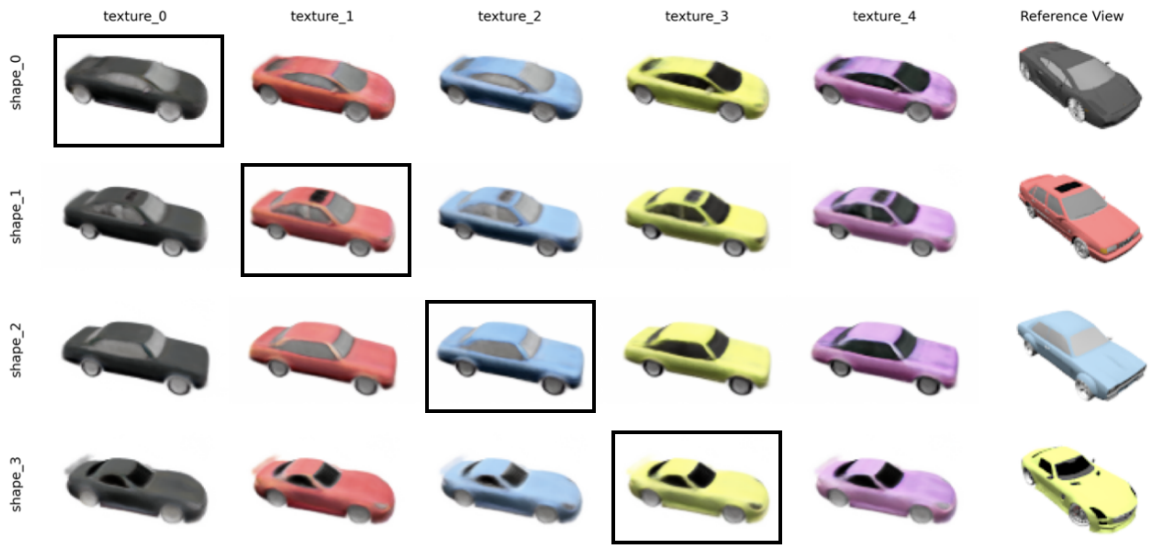


图 6 展示了 CodeNeRF 执行新形状、纹理和姿态合成的能力。4 张带框图像对应于使用与参考视图相对应的形状和纹理编码进行渲染的结果。其他结果则展示了在新视点下，将形状编码与纹理编码交叉组合后得到的渲染图像。图片摘自 Jang and Agapito (2021年) 中的图 9，依据 CC BY 4.0 许可协议授权。

其他多因素模型。
有时可以使用多线性模型来处理多个变化因素。具有两个潜在因素 $\mathbf{z}^1$ 和 $\mathbf{z}^2$ 的多线性模型表示为$x_i = \sum_{j,k} w_{ijk} z_j^1 z_k^2$，其中 w_{ijk} 是一个参数的三阶张量。Tenenbaum 和 Freeman（2000年）使用双线性模型分离风格和内容，例如不同字母的字体（风格）与字母本身（内容）。Vasilescu 和 Terzopoulos（2002年）使用四因素模型来涵盖不同的面部几何结构（人物）、表情、头部姿态和光照条件。

多因素模型的另一个例子是变换不变聚类 (Frey& Jojic，2003)。他们研究了这样一种情况：一个离散因子将形状/外观变化建模为混合模型，而一个连续因子则源于图像中物体的“干扰性”平移和旋转。

2.1.1 基于部件的模型

基于部件的模型是计算机视觉中的一个古老思想。例如，Fischler andElschlager（1973）提出了一种称为图像结构的模型，其中物体由多个按可变形构型排列的部件表示。部件对之间通过类似弹簧的连接来表示形变。Biederman （1987年）倡导在计算机视觉中采用基于部件的方法，并将其称为基于成分的识别。近来，Felzenszwalb、Gir- shick、McAllester 和 Ramanan （2009）使用经判别式训练的基于部件的模型，在物体识别方面取得了（当时的）最先进水平。

Ross 和 Zemel 描述了基于部件的方法的一些优点 (2006 年)，即：

- 如果部分部件可见，则可以识别部分被遮挡的物体；
- 基于部件的方法是对人体等高度关节化物体的多变性进行建模的好方法；
- 姿态变化对部件的影响可能小于对整个物体外观的影响。
- 可以用不同的新方式重新组合已知部件，从而对新的物体类别进行建模。

基于部件的模型与知觉组织或知觉分组（例如，参见Palmer 1999，第6章）有着相似的目标，即将部件组织成一个整体，但通常知觉组织被视为一种通用过程，例如，将边缘分组成轮廓，或将

相似的像素分组成区域，而非利用关于特定物体类别的知识。

下面给出四个人脸基于部件的模型示例。

示例：人脸的基于部件的模型。
基于部件的模型的一个常见应用是人脸。在英国，警方会在调查犯罪时使用“PhotoFit”系统，在目击证人的协助下生成嫌疑人的图像。例如，可参见 Graham Pike教授的“PhotoFit Me”项目。[8]该系统将面部分解为眼睛、鼻子、嘴巴、下颌和头发等部件。

Ross 和 Zemel (2006年) 提出了两种模型，用于学习这种基于部件的分解。我们结合人脸图像建模来介绍这两种模型。[9] 第一种模型——多原因向量量化（MCVQ）——具有 K 多项式因子，每个因子选择一个给定部件的外观。每个部件的“掩码”以概率方式学习，因此像素 j 由部件 k 以多项式概率 π_{jk}解释，其中 $\sum_k \pi_{jk} = 1$。图 7展示了学习到的模型。左下方的图显示掩码概率，右下方则显示每个向量量化对应的 10 个均值，并乘以相关掩码。注意左下方图中，掩码如何标识眼睛、鼻子和下巴等区域。

第二种模型称为多重成因因子分析（MCFA）——它与 MCVQ 类似，但现在每个部件的外观由因子分析器而非离散选择来建模。掩码模型与 MCVQ 相同。Nazabal、Tsagkas 和 Williams(2023年) 使用 MCFA 模型处理人脸，并添加了一个更高层级的因子分析模型，以关联各部件的因子分析器。他们使用的数据集（来自 PhotoFit Me）在性别（女性/男性）和种族（Black/亚洲人/白人）方面均保持平衡，因此高层级的因子分析器能够对各部件之间的规律进行建模，例如肤色方面的规律。

MCVQ 和 MCFA 模型对掩码概率 π_{jk} 使用了一个简单模型。但假设我们正在对汽车侧视图的部件分解进行建模，例如将其分解为车轮、车身和车窗。不同的汽车风格会产生截然不同的掩码模式，而这些模式可以使用潜变量模型进行建模。例如，Eslami 和 Williams (2011) 使用指数族因子分析对掩码模式进行了建模，而Eslami 和 Williams (2012) 使用了多项式形状玻尔兹曼机。

示例：基于部件的时钟模型。
图 8改编自 Zhu 和 Mumford(2006年) 的研究。该图描述了一个用于钟表的与或图；例如，时钟的指针、表盘和数字组件由一个与节点连接，但每个组件中都存在由或节点编码的备选项。或节点上的特定选择会产生一个解析图。[10] 其中一个可能的解析图用深色箭头表示，对应于顶部的图像。因此可以看到，指针（2 根而非 3 根）、数字（阿拉伯数字而非罗马数字）以及表盘形状（圆形，且只包含外圈）都做出了或选择。

与或结构类似于我们使用 MCVQ 处理人脸时所见的结构；其中包含若干部件（与），并且每个部件都有一组离散的选择（或）。不过，与或图通常更为强大，因为它可以具有层次结构。例如，在时钟模型中，可以选择框架部件的外环、内环和中央环全部存在或缺失。

示例：关节式人体的基于部件的模型。

关节式物体的一个典型例子是人体，它可以分解为躯干、头部、手臂和腿。手臂和腿还可以进一步分解；例如，手臂由上臂、前臂和手组成，而手又可以进一步分解为手指和拇指。鉴于人类虚拟形象在电影和游戏行业中的重要性，围绕这一主题已经开展了大量研究。这里我们关注 M. Loper、Mahmood、Romero、Pons-Moll 和 Black (2015年提出的蒙皮多人线性模型（SMPL）。需要考虑两个变化因素。第一个是人体的姿态，

[8]参见 https://www.open.edu/openlearn/PhotoFitMe。[9]所用数据已按位置和尺度对齐，因此在学习过程中无需对这些“干扰因素”建模。

[10]参见第 3.3 节，以进一步讨论文法和与或图。

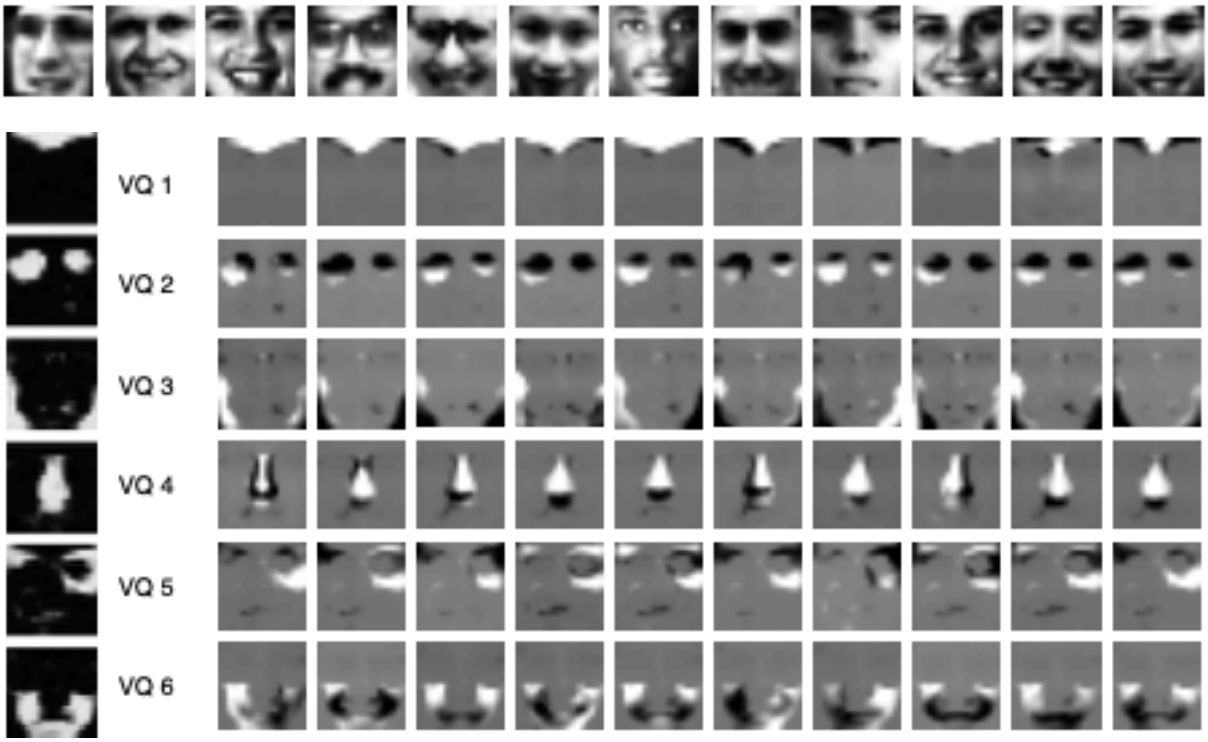


图 7 顶部一行：Ross 和 Zemel 模型的训练图像样本。下方几行：由 MCVQ 学习得到的基于部件的模型。左侧的图显示掩码，即每个像素选择给定向量量化的概率。右侧是每个向量量化对应的 10 个均值，并与左侧所示的掩码相乘。图经 Ross 和 Zemel (2006年) 授权转载，已获 R. S. Zemel 许可。

它由运动学树上的关节角度定义。第二个是特定个体的身体形状。人体模型由一个包含约 6890 个三维顶点的网格定义。每个顶点 i 都被赋予权重 w_{ki}，以表示部位 k 对顶点 i的影响程度。

身体形状采用线性基进行建模，与第2.1节讨论的 Blanz 和 Vetter 模型类似。车身的姿态由运动学树中一个部件相对于其父部件的相对旋转的轴角表示确定。模型的另一个重要组成部分是姿态混合形状，它们会根据姿态调整顶点位置，但在应用姿态变换之前进行。在 SMPL 中，姿态混合形状是部件旋转矩阵各元素的线性函数。姿态混合形状用于抵消关节处可能出现的不自然形变。将中性姿态下的形状和姿态混合效果结合后，顶点i 的最终预测位置是通过对顶点在部件 i 下经 k变换后的位置进行加权求和得到的（权重为 w_{ki}）。

示例：胶囊

另一种基于部件的模型称为“胶囊”。这一术语由 Hinton、Krizhevsky 和 Wang （2011）提出，后续发展包括Sabour、Frosst 和 Hinton （2017）、Hinton、Sabour 和 Frosst （2018）以及 Kosiorek、Sabour、Teh 和 Hinton （2019）。De Sousa Ribeiro、Duarte、Everett、Leontidis 和 Shah （2022）近期发表了一篇关于胶囊的综述论文。“胶囊”一词指一种视觉实体，它既输出该实体存在的概率，也输出一组该实体的实例化参数。虽然通常以推断的方式描述胶囊，即信息流从部件传递到物体，但下文将遵循Nazabal 等人（2023年）对生成式胶囊模型的阐述。

考虑一个由 N 个部件 $\{\mathbf{p}_n\}_{n=1}^N$组成的物体模板 T。每个部件$\mathbf{p}_n$ 都由其类别、姿态、形状和纹理进行描述。该模板 T 具有关联的潜变量向量$\mathbf{z}$，它会影响各部件的姿态、形状和纹理。例如，对于一个二维模板，$\mathbf{z}$ 中的一部分可能定义

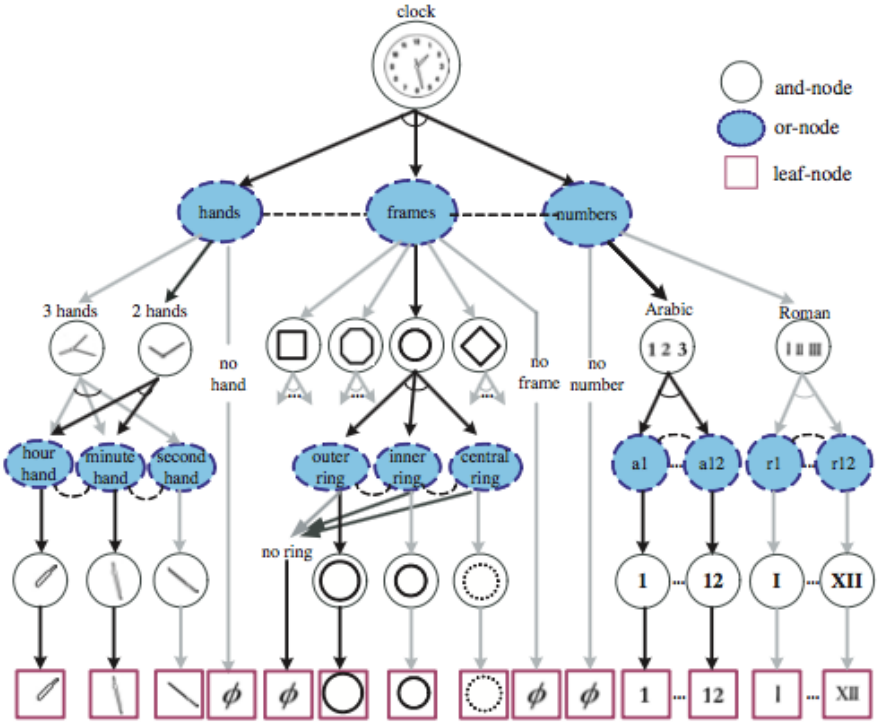


图 8 时钟物体类别的与或图。与节点的子节点之间的虚线表示关系和约束。经许可转载自 Zhu 和 Mumford (2006年，图 6.1)，版权归 now publishers inc. 所有。 ©2006。

模板的相似变换参数（包括平移 $\mathbf{t}$、旋转 θ 和缩放 s ）。物体与部件之间的几何变换可以用关于 $\mathbf{t}$、 $s\cos\theta$和 $s\sin\theta$的线性变换来描述。 $\mathbf{z}$ 的其他部分可以对部件之间的形状和纹理相关性进行建模。将观测到的部件与模板部件进行匹配的方法见第4节。

尽管我们在此描述的是物体-部件关系，但胶囊可以构成层次结构，从而能够表示物体间关系等。

2.2 物质建模：视觉纹理

Forsyth 和 Ponce (2003，第 164 页）对纹理作了如下论述：

> 纹理是一种普遍存在、容易识别却难以定义的现象。通常，某种效果是否被称为纹理，取决于观察它的尺度。

占据图像大部分区域的叶片是一个物体，但树木的枝叶则是一种纹理。大量小物体的视图通常最适合被视为纹理。示例包括草地、枝叶、灌木、卵石和头发。许多表面带有有序图案，看起来像是由大量小物体组成。示例包括豹或猎豹等动物身上的斑点、老虎或斑马等动物身上的条纹，以及树皮、木材和皮肤上的图案。纹理往往呈现重复性：大致相同的局部块会反复出现，不过可能会受到观察变换的扭曲。

纹理可分为规则纹理和随机纹理，不过两者之间也存在程度上的差异，例如近规则纹理或近随机纹理。规则纹理的例子包括砖砌结构、地砖图案和编织结构。随机纹理的例子包括云、木纹和枝叶。下文我们将主要关注随机纹理。

我们将纹理合成 的目标定义为：给定一个（较小的）训练样本，生成任意大小的纹理区域。对于规则纹理，应当可以提取重复元素，并将其平铺至目标区域，但这种方法并不适用于随机纹理。因此，我们的目标是从训练样本中学习纹理的生成模型。一种常用的模型类型是基于能量的模型 （基于能量的模型）。令$\mathbf{x}(k)$ 表示图像中以位置 k 为中心的图像块（例如方形图像块）。专家场 （FoE）的能量定义为

$$E_{FoE}(\mathbf{x}) = \sum_k \sum_j \phi_j(\mathbf{w}_j \cdot \mathbf{x}_{(k)}), \qquad (5)$$

其中，$\mathbf{w}j$ 是一个大小与图像块相同的滤波器，而 $\phi_j()$ 是某个函数。随后通过玻尔兹曼分布定义图像上的概率分布，即：

$$p_{FoE}(\mathbf{x}) = \frac{1}{Z(W)} \exp(-E_{FoE}(\mathbf{x})). \qquad (6)$$

这里， $Z(W)$ 是用于归一化配分函数， $p_{FoE}(\mathbf{x})$。专家场这一术语由Roth and Black (2005) 提出，是对 HintonHinton (2002) 提出的专家乘积构造的推广，能够处理任意尺寸的图像。

通常，直接从基于能量的模型中抽取采样点并不容易；一种标准方法是构造一个平衡分布为所需玻尔兹曼分布的马尔可夫链。更多细节请参见，例如，Murphy(2012年，第 24 章)。

一种简单的选择是将函数定义为 ϕ_j 关于$\mathbf{x}(k)$ 的二次型。如果该二次型正定，那么 $p_{FoE}(\mathbf{x})$将定义良好且可归一化，并构成高斯马尔可夫随机场（GMRF），例如参见Rasmussen 和 Williams (2006年，B.5 节）。规则网格上的平稳高斯马尔可夫随机场可以通过傅里叶分析（ Rozanov，1977）从功率谱的角度进行分析。然而，众所周知，仅基于功率谱（或等价地，通过维纳-辛钦定理基于二阶统计量）的图像模型并不充分。例如，Galerne、Gousseau 和 Morel(2011)展示了一段瓦片屋顶。随机化其傅里叶变换的相位，同时保持功率谱不变，会导致图像模糊，因为生成锐利边缘所需的相位对齐不再出现。

高斯随机场模型也可以通过最大熵原理（例如，参见Cover 和 Thomas 1991，第 12 章），基于分布的二阶（协方差）约束得到。但另一种方法是考虑一组滤波器，并施加约束，使最大熵模型对每个滤波器都匹配观测到的直方图。由此得到 Zhu、Wu 和 MumfordZhu、Wu 和 Mumford(1998年) 提出的 FRAME（滤波器、随机场与最大熵）模型。RothandBlack (2005) 讨论了 FRAME 模型，并指出该方法因使用离散滤波器直方图而变得复杂。因此，他们提出了专家场模型。

Kivinen 和 Williams （2012年）定义了一种不同的基于能量的模型，其能量函数为

$$E_{Tm}(\mathbf{x}) = \frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{a})^T(\mathbf{x}-\mathbf{a}) - \sum_{k,j} \log[1 + \exp(b_j + \sigma^{-1}(\mathbf{w}_j \cdot \mathbf{x}_{(k)})]. \qquad (7)$$

该模型受到 H. Lee、Grosse、Ranganath 和 Ng （2009）提出的卷积受限玻尔兹曼机的启发，但它没有使用卷积，而是采用平铺卷积 ，如 Ranzato、Mnih 和 Hinton（2010）所述。他们指出，卷积权重共享会带来问题，因为同一滤波器对应的相邻潜变量之间存在很强的相关性。在平铺卷积策略中，每个滤波器以自身的副本无重叠地平铺图像（即步幅等于滤波器直径）。不过，不同滤波器之间会相互重叠，以避免产生平铺伪影。能量函数E_{Tm} 由两项组成。第一项对应一个均值为 $\mathbf{a}$ 的简单球形高斯分布。第二项则通过对受限玻尔兹曼机的隐单元进行解析积分消去而得到（如 Kivinen 和 Williams 2012 的公式 1 所示）。该模型被称为“Tm”，以表明它采用平铺卷积，并且建模的是均值而非协方差（通过 $\{\mathbf{w}_j\}$ s 和 $\mathbf{a}$）。

Kivinen 和 Williams （2012年）指出，要对多种纹理建模，可以保持滤波器组固定

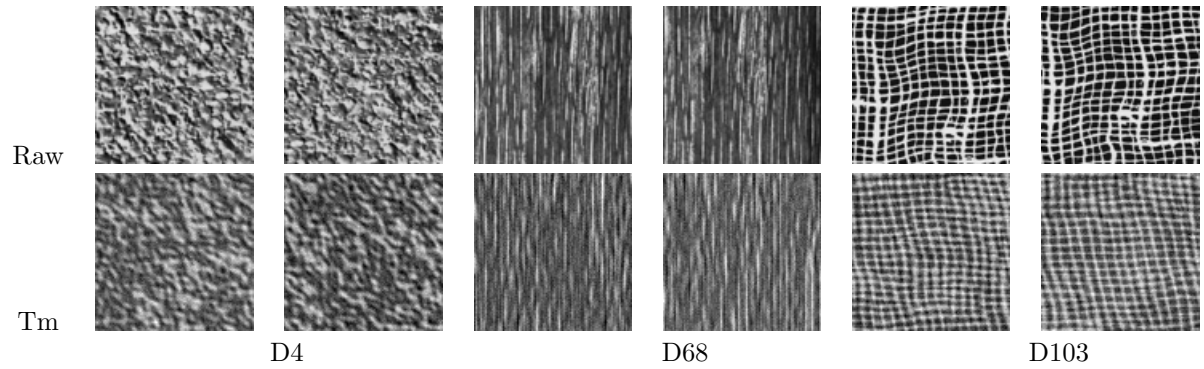


图9 顶行展示了三种纹理（标记为 D4、D68 和 D103）各自的两个样例图像块。底行展示了从针对这些特定纹理训练的 Tm 模型中抽取的两个样例。图引自 Kivinen 和 Williams (2012年，图4）。

，但针对每种纹理调整偏置——他们将此称为“Multi-Tm”模型。图9 展示了三种不同纹理的结果，每种纹理均有两幅原始数据图像和采样结果图像。原始数据来自 98×98从 Brodatz纹理图集中裁剪出的图像块（Brodatz，1966）。

基于能量的模型的另一种替代方案是自回归模型，其中预测的像素值 x_{ij} 在位置（i，j）取决于一个上下文向量；通常在按光栅扫描顺序处理时，该向量涵盖左侧和上方的像素。 Efros 和 Leung (1999) 采用这种方法进行基于样例的纹理合成，从一个纹理源训练样本开始。对于目标位置（i，j），在源纹理中找出与当前上下文区域相似的邻域，然后随机选取其中一个（选取依据是它与上下文区域的匹配程度）。对于填补空洞，可以采用沿边缘逐步向内扫描的“剥洋葱”策略，而不是按光栅扫描顺序。请注意，这是一种非参数建模方法——它不构造关于 $p(x_{ij}|$上下文）的参数化模型，而是从源纹理中提取相关区域。

自回归模型不一定基于样例。PixelCNN（ van den Oord 等人，2016）和 PixelCNN+ （ Salimans、Karpathy、Chen、Kingma 和 Bulatov，2017）是近年来颇具代表性的自回归模型，它们使用卷积层提取特征，并通过掩码确保只访问有效的上下文区域，从而预测 p（$x_{ij}|$上下文）。

上文讨论了在二维平面上生成平面纹理。纹理可以应用（映射）到表面上；这称为纹理映射（例如，参见 Szeliski 2021年）。

3 场景模型

本节介绍描述场景生成模型各个方面的先前研究。我首先介绍“以物体为中心的模型”，这类模型将场景组织为相互独立的物体；随后讨论基于自回归模型、能量模型和层次模型来描述物体之间关系的场景模型。

一种非常简单的场景模型，是随机选择物体并将它们放入场景中。可以将其概括为“图像的独立成分是物体”。其中一个例子是Williams 和 Titsias (2004年) 的二维精灵图研究，如图 2 所示。在这里，模型学会了背景和两个人，但从先验分布来看，它会将人物随机放置在图像中的位置。较新的一个例子是 Greff 等人提出的 IODINE（Iterative Object Decomposition Inference NEtwork，迭代对象分解推理网络）Greff etal. (2019)。该模型采用“以物体为中心”的表示，由 K 潜变量向量 $\mathbf{z}_1, \ldots, \mathbf{z}_K$ 组成，每个物体对应一个向量。IODINE 在二维精灵图数据上进行了演示，也用于处理 CLEVR 数据集中的三维场景图像。该数据集包含球体、立方体等几何物体，它们的位置和材质属性均为随机。IODINE 中每个物体对应的 $\mathbf{z}_k$并未分解为形状、纹理和姿态成分（如第 2.1 节所述），而是一个将这些因素纠缠在一起的单一向量。

然而，正如句子并非由随机的词语序列组成一样，[11] 视觉场景也不是由随机的物体集合构成的——物体之间存在共现关系和相互关系。例如，场景类型（如厨房、客厅、城市街道、乡村田野）与观察到的物体种类之间存在相关性。此外，场景中的物体之间还会发生物质-物质、实体-物质和实体-实体交互（例如，参见Heitz 和 Koller 2008）。实体-物质交互的例子包括汽车通常出现在道路上，奶牛通常出现在草地上。实体-实体交互的一个例子是，餐椅通常围绕餐桌摆放。另一个实体-实体交互的例子是为粗略的场景布局添加细节，例如在餐桌上摆放桌垫、餐具和陶瓷餐具。[12] 读者会发现，第2.1.1节所述的基于部件的模型与场景模型之间存在相似之处。不过，基于部件的模型通常受到更多限制，部件数量固定；而场景模型中的物体数量可变，物体之间的关系也更宽泛。场景关系还可能跨越较远的距离——一个经典例子是电视与供人观看电视的沙发之间的关系，两者需要保持舒适的距离。

下面我们将重点关注室内场景的模型，这反映了研究文献中的关注重点。但也存在室外场景，包括乡村和城市环境中的场景。对于山地或海岸线场景，最自然的做法可能是考虑侵蚀对景观的雕刻作用，例如河谷的形成。在农田中，会有由人类建造的田地边界（受景观限制），以及农作物或牲畜。在城市环境中，可以使用文法类型的模型生成建筑立面，然后用人、汽车和街道设施等其他物体“装饰”街道建筑。

下面我们依次介绍自回归模型、基于能量的模型和层次模型。请注意，自回归模型和基于能量的模型不是潜变量模型，而层次模型是潜变量模型。本节假设三维数据可用，例如带有类别标签的三维定向包围盒（OBB）。

[11]是随机的非单词句子序列！[12]Ritchie (2019) 术语“虚拟布景”。

3.1 自回归模型

我们以 Ritchie、Wang 和 Lin（2019）的工作为例。他们描述了这样一个过程：物体逐个被添加到房间布局中，直到作出停止的决定。模型首先提取房间的俯视平面图（用于定义放置物体的有效区域）。随后，它将该平面图输入由四个模块组成的序列，这些模块分别用于：（i）决定添加哪个物体（如果有的话）；（ii）指定物体应位于何处；（iii）确定其方向；以及（iv）确定其物理尺寸。添加物体后，系统会更新房间布局表示，将该物体纳入其中，然后再次执行步骤（i）至（iv）。这些模块通过卷积神经网络实现。

令 m 有序物体表示为$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$。每个$\mathbf{x}_i$均由物体类别标签、位姿特征（位置和方向）、形状特征以及纹理特征[13]组成，因此$\mathbf{x}_i = (\mathbf{x}_i^c, \mathbf{x}_i^p, \mathbf{x}_i^s, \mathbf{x}_i^t)$。于是，在自回归模型下，我们有

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) = p(\mathbf{x}_1) \prod_{i=2}^{m} p(\mathbf{x}_i | \mathbf{x}_{<i}), \quad (8)$$

其中 $\mathbf{x}_{<i}$ 表示序列 $\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}$。这适用于从模型生成场景。如果给定物体的某种排列，也可以用它来计算 $p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$。然而，当我们观察到一组无序物体 $X = \{\mathbf{x}_i\}_{i=1}^m$时，我们应该对所有可能的排列求和，以便

$$p(X) = \sum_{\pi \in \Pi} p(\pi) p(\mathbf{x}_{\pi(1)}) \prod_{i=2}^{m} p(\mathbf{x}_{\pi(i)} | \mathbf{x}_{\pi(<i)}), \quad (9)$$

其中 π 表示一个特定的排列，而 Π是所有排列的集合。排列的先验分布 $p(\pi)$可以设为均匀分布，即 $1/(m!)$。事实上，Ritchie 等人(2019)确实在训练物体类别标签模块时，依据类别重要性度量对物体进行排序。该度量取决于类别的平均尺寸与其出现频率的乘积。这意味着，在卧室场景中，床这类大型物体会首先出现，其他物体则围绕它布置。

[13]Ritchie 等人(2019) 不使用纹理特征，但通常情况下它们可能存在。

Ritchie 等人提出的模型(2019)相对简单，但使用包含已添加物体的平面图表示后，就可以直接利用概率链式法则对上下文进行建模，而无需采用马尔可夫模型之类的简化方法。[14] 我们可以追问：将自回归过程作为场景生成模型是否合理？先添加大型物体（例如卧室场景中的床）并非不合理。但人们在布置卧室时，会按照固定顺序依次摆放所有物体类别，这似乎不太可能。

还需注意，自回归模型不会显式地将共同出现的物体归为一组，例如床和床头柜，或为方便观看而组合摆放的电视和沙发。相比之下，层次模型（见下文）应该能够识别出这类结构。

在 Ritchie 等人 (2019) 的工作基础上，Paschalidou 等人(2021年) 开发了一种使用 Transformer 的自回归模型（称为 ATISS）。与公式 9 中对序列的所有排列求和不同，作者会随机打乱物体的顺序，并训练 Transformer 根据已有物体预测下一个物体。为处理可能预测出多种不同物体的情况，他们在物体特征上使用混合模型。他们展示了 ATISS 相较于早期工作的若干优势，例如在场景补全方面，尤其是在序列中较早出现的物体（如卧室中的床）被省略时。

3.2 基于能量的模型

假设我们已经生成了一些要放入房间的物体。接下来需要对它们进行布置，使其满足一系列空间、功能和语义关系或约束。基本思路是定义一个能量函数，用来衡量布局对这些约束的符合程度，然后按照公式 6，使用玻尔兹曼分布定义概率分布。Yu 等人 (2011) 采用了这种方法，通过在配置空间中进行模拟退火搜索，自动优化家具布局。

[14]不过请注意，平面图表示会舍弃物体的添加顺序。

Yu 等人 (2011)提出的方法适用于物体集合固定的情况。 Yeh、Yang、Watson、Goodman 和 Hanrahan(2012年) 将这一思路扩展到“开放世界”，使物体数量可以变化。他们采用可逆跳跃马尔可夫链蒙特卡罗（马尔可夫链蒙特卡罗）方法在该空间中进行搜索。

基于能量的模型普遍存在一个问题：通常需要运行马尔可夫链蒙特卡罗链很多次迭代，才能从平衡分布中抽取采样点。这一限制也适用于上述方法，这意味着基于能量的模型难以扩展到更大的场景。

3.3 层次模型

构建层次模型的一种自然方式是采用**基于文法的方法**。利用文法进行模式分析由来已久，例如可参见 K. S. Fu(1982) 关于句法模式识别的研究。在场景理解方面，或许最值得一提的是朱松纯及其合作者的工作。朱松纯和 Mumford 的长篇论文《图像的随机文法》(2006年) 是一篇重要的参考文献。上文第 2.1.1 节讨论了 Zhu 和 Mumford 于2006年提出的时钟部件模型。

上下文无关文法（CFG）由终结符、非终结符、开始符号和产生式规则定义。受Liu 等人（2014）的启发，我们以卧室场景为例。这里，终端符号表示观测到的物体，例如床垫、床头柜、枕头等。如上所述，完整描述不仅涉及类别标签，还包括姿态、形状和纹理信息。此处的非终结符对应物体的分组；例如，我们可以有睡眠区 →床床头柜组。通常会有一个或两个床头柜（通常取决于床是单人床还是双人床），因此产生式“床头柜组 → 床头柜”和“床头柜组 → 床头柜 床头柜”都是有效的。图Liu 等人（2014）中的卧室场景文法示例如图10所示。

概率上下文无关文法（PCFG）会为规则左部具有相同非终端节点的产生式添加一个概率分布。Charniak(1993，第 5 章）对上下文无关文法和概率上下文无关文法作了很好的概述。上下文无关文法通常用一组产生式

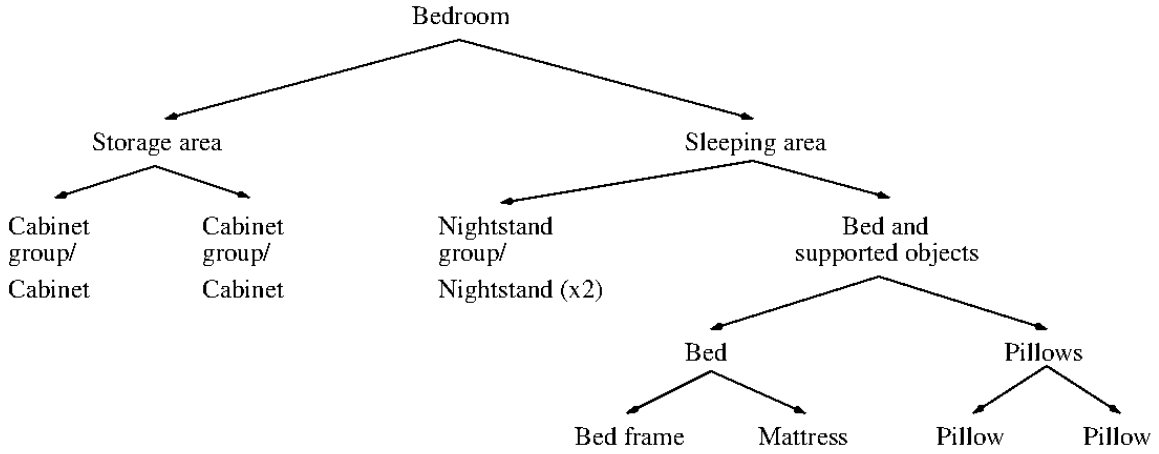


图10 一种层次　卧室场景的文法，改绘自Liu　等人(2014)。

规则来描述，但也可以用与或图来描述；Hall (1973）证明了二者的等价性。此处，或关系作用于左部具有相同非终端节点的产生式；与关系则作用于右部包含多个符号的产生式，因为这些符号都必须生成。正如Zhu 和 Mumford (2006年，第 6.1 节）所述，“与或图嵌入了整个图像文法，并包含所有有效的解析图”。

我们已经看到，终结符通常包括类别标签、姿态、形状和纹理信息。这意味着，支配一组终结符的非终结符需要包含潜变量，以生成它们之间适当的相关性。例如，产生式「睡眠区　→床床头柜组」需要指定床的尺寸（单人床或双人床），并且还需要将这一信息传递给床头柜组变量，以便它确定生成一个还是两个床头柜（以及它们相应的位置）。在上下文无关文法中，姿态和尺寸信息可以相对于更高层级的变量来定义。因此，例如，如果已经定义了睡眠区的尺寸和位置，那么相对于睡眠区来确定床的位置就是合理的。编码这些信息的一种可能替代方案，是添加水平关系结构，以编码节点之间的上下文信息或约束；这一做法由Zhu 和 Mumford (2006年，第 4 章）和 Jin and Geman(2006年）提出。

使用上下文无关文法来建模视觉场景的一个问题是，很难从数据中学习这类文法。即使在自然语言处理（自然语言处理）中，大多数成功的文法学习也使用经过标注的“树库”数据，而非未标注的句子。类似地，研究者也利用标注数据学习图像的一些层次模型，例如Yao、Yang 和 Zhu(2007年）以及Liu 等人 (2014）。然而，最近针对自然语言处理数据，Kim、Dyer 和 Rush(2019）发现，除了概率上下文无关文法结构之外，还可以加入句子级连续向量潜变量（从而构成“复合”概率上下文无关文法），并在解析任务中取得最先进水平的结果。

虽然文法是生成层次结构的一种方式，但并非唯一方式。我们可以不使用离散值非终结符，而是在节点中使用连续值潜向量。考虑从单个潜在向量 $\mathbf{z}$ 开始。第一个二元分裂可以生成如下

$$[\mathbf{z}_1, \mathbf{z}_2] = \mathbf{f}(\mathbf{z}), \tag{10}$$

其中 $\mathbf{z}_1$ 和$\mathbf{z}_2$ 是两个子节点的潜在向量，而 $\mathbf{f}$ 是一个非线性的向量值函数，可以简单到 $[\tanh(W_1\mathbf{z} + \mathbf{b}_1), \tanh(W_2\mathbf{z} + \mathbf{b}_2)]$，也可以是更深层的神经网络。对于每个生成的潜在向量，都会应用一个二元节点分类器，以判断它是终端节点（用于生成物体），还是可以进一步拆分的非终端节点。这种生成结构由 Pollack (1990) 描述，是递归变分

自编码器将在下文介绍。连续状态的优点在于，它非常适合表达姿态、形状和纹理变化。请注意，Nazabal 等人提出的用于人脸的层次化 MCFA 模型Nazabal 等人 (2023年)（见第 2.1.1 节）是这种类型的简单单层模型，但其具有 K子节点而非仅两个，且不含 tanh 非线性。

M. Li 等人 (2019) 开发了用于室内场景的生成式递归自编码器（GRAINS）模型，这是一种递归变分自编码器。该模型以三维空间中的无结构物体集合为输入，并计算出潜在场景表示$\mathbf{z}$，再将其解码以生成物体的层次结构，随后可对其进行渲染。具体而言，GRAINS 模型将潜在向量 $\mathbf{z}$ 解码为五个变量，分别表示房间的地板和四面墙。每个变量要么是终端节点（对应一个物体），要么是非终端节点，具体由节点分类神经网络判定。非终端节点有若干种，分别对应支撑、环绕和共现关系。终端节点则解码为物体，由定向包围盒、位置、方向和语义标签指定。树结构编码器首先将场景中的物体聚类到最近的墙壁，再利用非终端节点为每个聚类构建子树。随后，树结构神经网络将场景编码为潜在向量 $\mathbf{z}$。

较新的递归变分自动编码器模型是 Gao 等人提出的场景层次图网络（SceneHGN）Gao 等人 (2023年）。该模型摆脱了对四面墙的依赖，构建了一个层次化的场景表示，其中包含房间、功能区域、物体和物体部件等层级。它还采用了另一组非终端节点。

需要注意的是，递归变分自编码器所使用的编码器结构不必与解码器结构完全镜像；它只需输出潜在向量 $\mathbf{z}$。GRAINS 和 SceneHGN 中使用的启发式树构建方法可以与文法进行对比，后者会对所有解析树求和。与对树进行完整搜索/求和相比，这种构建方法可以显著降低推断的复杂度。

4 推断

本节的任务是从单个 RGB 或 RGBD 图像，或从一组具有已知视点的图像（多视图数据）中进行推断，并假设物体模型和关系模型已经学习完毕。因此，我们关注的是$p(\mathbf{z}|\mathbf{x})$。

我们首先介绍物体和全局变量的推断，然后介绍基于场景模型的推断。最后，我们讨论模型存在缺陷时的推断，即输入图像无法被精确重建的情况。

物体和全局变量的推断。
基本的推断任务包括：(i) 检测每个物体并确定其类别标签、形状、纹理和姿态信息；(ii) 检测并描述背景区域；以及 (iii) 确定光照和相机的全局潜变量。如果我们将每个潜变量的取值离散化并对其进行搜索，那么即使是单个物体的推断问题也已十分棘手（参见例如 Yuille 和 Liu 2021年第 7.1 节），而考虑多个物体和全局变量时则会出现组合爆炸。应对组合爆炸最直接的方法是近似搜索。一种此类方法是马尔可夫链蒙特卡罗（MCMC），其目标是构建一条马尔可夫链，使其平衡分布能够从潜变量的目标后验分布中采样。关于 MCMC 的讨论，参见例如 Murphy (2012年，第 24 章）。状态空间中的提议移动通常是通用的，并未利用问题的结构，因而导致混合速度缓慢。一种更先进方法的例子是数据驱动马尔可夫链蒙特卡罗（DDMCMC，Tu 和 Zhu 2002年），该方法允许将物体检测结果候选集等提议纳入有效的 MCMC 算法。

另一种近似方法是使用变分推断，例如参见 Jordan、Ghahramani 和 Jaakkola(1999)，其目标是用变分分布近似后验分布 $p(\mathbf{z}|\mathbf{x})$，即 $q(\mathbf{z})$。变分自编码器（Kingma 和 Welling, 2014；Rezende、Mohamed 和 Wierstra, 2014）使用摊销变分推断，根据输入数据预测潜变量的分布。这通过一个“编码器”网络（也称为

“识别模型”，参见Dayan、Hinton、Neal和 Zemel，1995）。如果图像中只有一个关注物体，这种摊销推断相对直接；但若存在多个物体，则会更加复杂，例如这是由置换对称性导致的。在 IODINE（Greff 等人，2019）中，图像到物体潜变量的前馈预测会经过迭代细化，以考虑解释消除等影响。在Locatello 等人(2020年）提出的槽注意力模型中，从输入到潜变量的映射采用迭代注意力机制，使它们相互竞争，以解释图像中的物体。Eslami 等人提出的注意、推断、重复（AIR）模型(2016）采用另一种顺序方法，一次识别一个物体。

马尔可夫链蒙特卡罗和变分推断（VI）方法可用于表示给定数据时潜在表示$\mathbf{z}$ 的不确定性，数据为$\mathbf{x}$。例如，（部分）遮挡可能导致这种不确定性，并可能产生多模态后验分布，对应于对输入图像的不同解释或“解析结果”。变分推断方法的一个局限是，如果假定的 $q(\mathbf{z})$形式是单模态的，就可能无法很好地捕捉这种多模态性。如Murphy(2012年，第21.2.2节）所述，当变分分布为单模态时，变分推断中使用的库尔贝克-莱布勒散度 $KL(q \parallel p)$往往会选取后验分布的一个众数，从而低估不确定性。原则上，马尔可夫链蒙特卡罗方法可以从多模态后验分布中采样，但在实践中可能会陷入某个众数，除非采用退火等特殊措施（参见例如Murphy 2012，第24.6节）。如 $q(\mathbf{z})$的不确定性的一种降低方式是融合多个视角的信息，如N. Li、Eastwood 和 Fisher (2020年）所示。

上述模型（如 IODINE 和 AIR）均采用无监督方式进行训练。不过，过去二十年来物体检测器的发展意味着，我们可以针对汽车、菠萝等已知类别训练物体检测器。其输出可能是一个边界框（BBox），也可能是一个感兴趣区域。这种方法由Izadinia、Shan 和 Setiz (2017) 在其 IM2CAD 系统中采用。该系统以房间图像作为输入，将房间几何结构估计为一个三维框，检测物体，预测每个物体对应的潜在变量，并将物体放置在场景中，然后通过渲染场景并将其与输入图像进行比较，来优化物体的位置。

如果输入数据是一组多视图图像，就可以采用几何计算机视觉技术。可以使用同时定位与建图（SLAM）或运动恢复结构（SfM）来估计相机位姿和稀疏点云表示，如 COLMAP（Sch¨onberger&Frahm，2016）。这些技术本身不会将数据分割成物体，但可以结合针对每张图像预测的目标检测边界框和掩码，生成三维边界框，例如R¨unz 等人 (2020，第 5.1 节）。物体分割完成后，就可以估计其形状、纹理和姿态潜变量，例如参见 R¨unz 等人(2020，第 5.2–5.3 节）。

近年来一项重要进展是可微渲染，它可以优化估计出的潜在状态，以改善渲染场景与输入图像之间的匹配。早期的一种可微渲染器是 *OpenDR*，由 M.M. Loper 和 Black (2014) 提出。Kato 等人 (2020年) 综述了针对网格、体素、点云和隐式曲面表示提出的各种方法。

场景模型的推断。

上文讨论了物体和相机／光照全局变量的实例化；现在我们来考虑关系结构的推断。像 ATISS 这样的自回归变换器模型（Paschalidou 等人，2021年）能够在 M 时间内计算模型下$O(M^2$个物体的联合概率（例如，参见Murphy 2023年，第 22.4 节）。训练中使用的随机排列有效地使其能够对集合而非序列进行建模。

对于层次化模型，情况会更加复杂，因为可能需要动态构建解析图。其中一个主要问题是，将一组物体和部件检测结果匹配到类似与或图这样的场景结构。首先，以物体的胶囊模型为例，正如第2.1.1节所述。该模型由若干部件组成，这些部件之间具有特定的几何关系。（类似的论述也适用于场景中物体的分组，例如

电脑显示器-键盘，但这里我们将使用“物体及其部件”这一术语。）

现在，假设我们检测到一组部件 $\{\mathbf{x}_m\}_{m=1}^M$，任务是将它们与一组模板 $\{T_k\}_{k=1}^K$ 匹配。令 $w_{mnk} \in \{0,1\}$表示观测 $\mathbf{x}_m$ 是否与模板 k的部件 n 匹配。 w可以视为一个二元矩阵 W ，其索引为 m 和n, k。描述这一指派问题的一种方式，是考虑观测部件与模板部件之间的有效匹配，并用置换矩阵来定义这些匹配。如果 M 不等于模型部件的总数 $N = \sum_{k=1}^K N_k$，则可以添加虚拟行或列，使问题成为方阵，例如在某些部件缺失的情况下。由于精确计算需要考虑所有可能的排列，其规模会随 M呈指数增长，Nazabal 等人（2023年）因此对每个模板的匹配变量 W 和潜变量 $\mathbf{z}_k$ 采用变分推断。每次迭代的复杂度为 $O(M^2)$，这是因为需要计算实值变量 r_{mnk} ，以近似在满足行和列概率之和为 1 的均值场假设下， W 的后验概率。这实现了一致性路由过程，正如 Sabour 等人（2017）和 Hinton 等人（2018）所讨论的那样，但它是从变分推断方程推导而来，而非一种临时设定的目标函数。与一致性路由不同的另一种方法是采用随机样本一致性方法（ RANSAC，Fischler 和 Bolles，1981），用最少数量的部件实例化一个物体模板，然后通过在预测位置寻找其余部件来验证该模板。Kosiorek 等人（2019）采用了另一种推断方法：自编码器架构预测每个模板的潜变量，并利用集合变换器架构（J. Lee 等人，2019）处理数量任意的观测 M。

关于层次模型，图像解析算法的讨论示例包括：Zhu 和 Mumford (2006年，第 8 节）讨论 AND-OR 树，以及Liu 等人 (2014）讨论概率上下文无关文法。对于序列（如文本），概率上下文无关文法模型的推断复杂度为 $O(M^3$，其中句子长度为 M，因为可以利用内外算法（例如参见 Charniak 1993，第 6 章）。然而，如果不具备这种序列排序特性，精确推断的代价会更高，因此可以采用近似方法。

近期一种用于推断的方法是 GFlowNets，例如参见 Hu 等人 (2023年）。其思路是训练一个随机策略，从一组物体 $\mathcal{Z}$（例如完整解析树）的目标分布中进行采样。具体做法是从更大的物体集合 $\mathcal{S}$ （由部分构造的物体组成，例如不完整的解析树）中采样。这些物体通过前向策略 $p_F(s'|s, \mathbf{x})$ 从初始空状态开始逐步构建，直到获得目标分布中的一个有效物体。训练目标是学习p_F，从而生成来自目标分布的有效采样点，此处为 $p(\mathbf{z}|\mathbf{x})$。

如果层次结构是根据递归神经网络定义的，例如 GRAINS（M. Li 等人，2019年），那么就可以构建一个编码器来预测场景潜变量 $\mathbf{z}$，然后将其解码以生成层次结构。这种方法在 GRAINS 中取得了成功，但请注意，在该方法中，输入是一组三维物体，而不是未分割图像。

虽然自然会认为层次模型中的推理信息流是自底向上的，即从部件到物体再到场景，但并不一定非要如此。整体场景类型可能可以通过全局场景描述符（如 场景概要 （Oliva 与 Torralba，2006年）得到很好的表征，而这会为某些物体类别而非其他类别产生自顶向下的预期。例如，Torralba、Murphy 和 Freeman（2010）利用场景概要来预测场景类型（例如海滩场景、街道场景、客厅）。这进而对图像中某一类别物体的垂直位置做出了有用的预测（例如街景中的汽车）[15]。更普遍地说，这表明在场景模型中进行推断时，来自图像的信息既可以自顶向下流动，也可以自底向上流动。

正如 Zhu 和 Huang （2021年，第 1.3.4 节）所指出的，视觉由大量任务驱动，而“这些任务各自只利用解析图的一小部分；解析图会基于相关任务层次，以任务驱动方式实时构建”。因此，考虑对完整解析图进行“惰性推理”或许很有用。[16] 。

[15]在所考察的场景类型中，场景类型通常无法很好地限定物体的水平位置。[16]感谢 Kevin Murphy 提出这个术语。

模型缺陷：
引用乔治·博克斯的话，我们可以说：“所有模型都是错的，但有些模型有用”（Box &Draper，1987年，第 424 页）。许多视觉场景内容丰富、细节繁多，要对其进行完整建模可能并不可行（甚至未必可取）。然而，如果无法让像素误差归零，那么生成模型究竟在多大程度上可以说完整解释了图像？问题在于，误差聚合度量（例如均方误差，MSE）可能由多种情况造成：(i) 物体的姿态略有偏差，但外观和形状正确，导致物体边界周围出现“光晕”状误差；(ii) 姿态和形状正确，但物体的纹理与已知纹理库不匹配，因此物体范围内出现高频误差模式；(iii) 物体形状不正确（可能是推断或建模失败所致），但姿态和纹理正确；或 (iv) 小物体的目标检测出现假阳性或假阴性，同样会产生相同的均方误差。

解决这一问题的一种自然方法，是比较输入图像与预测图像，以及对物体范围等的预测结果。据我所知，在缺少真实标注的情况下，针对视觉算法输出质量进行预测的相关先前研究并不多。Jammalamadaka、Zisserman、Eich-ner、Ferrari 和 Jawahar(2012年) 探讨了评估算法，用于预测人体姿态估计算法的输出是否正确；而 Xia、Zhang、Liu、Shen 和 Yuille(2020年) 则研究了如何预测语义分割中的失败。在重建框架中，几何变量（相机参数、物体姿态和形状）的拟合优度可以通过预测的物体掩膜与真实标注的物体掩膜之间的交并比（IoU）来衡量（例如，Romaszko、Williams 和 Winn，2020年）。因此，可以通过评估算法预测这一交并比指标。这些变量中的误差会进一步影响像素误差。但如果交并比令人满意，那么物体层面的像素误差很可能是由纹理或光照变量中的误差造成的。若要利用评估算法评估此类误差的重要性，将需要对所发生误差的严重程度和类型进行标注。

人们可能会认为，能够估计其预测不确定性的算法有助于解决视觉算法评估问题，而这在一定程度上是正确的。然而，概率模型不确定性的一个局限在于，它是相对于某个模型（或一组固定模型）来评估的。但例如，如果模型的纹理库与输入图像中的纹理不匹配，那么模型的后验不确定性度量就无法很好地刻画真实情况。这种缺陷称为“开放世界”（M-open）情形，与 M-closed 相对，Bernardo 和 Smith（1994 年，§6.1.2）对此有所讨论。解决这一问题的一种方法是进行模型批判，例如 Seth 等人（2019）对此有所讨论。

鉴于视觉场景的复杂性，可能无法在像素层面呈现场景中的所有细节。相反，可以针对图像特征的空间布局建立生成模型，例如由深度神经网络计算出的特征。Wang、Kortylewski 和 Yuille（2021年）的神经网格模型（NeMo）便采用了这种方法。与基于特征的表示相比，重建输入图像更容易解释，但后者可能使建模更容易。NeMo 生成模型的实验结果表明，与标准前馈深度神经网络相比，它在部分遮挡下进行识别、预测此前未见过的姿态等分布外任务上具有更强的鲁棒性。

5 推进结构化生成模型议程

上文我已概述了物体与场景建模、在这些模型中进行推断以及相关问题的关键主题。下文我将进一步讨论最先进水平，以及数据集等方面的问题。

物体建模：
最先进水平似乎已相当不错。在数据方面，Collins、Goael、Deng 等人（2022）的亚马逊伯克利物体数据集（ABO）是一个近期案例，其中收录了规模相当大的模型集合（约 8,000 个物体），具有复杂的几何结构和高分辨率、基于物理的材料，可实现照片级真实感渲染。

场景建模：
与物体相比，场景建模的最先进水平还不够成熟。目前研究主要聚焦于室内场景，对城市和乡村环境中的室外场景建模则探索较少。场景建模的任务也比物体建模更加困难，因为它涉及数量和类型各异的物体，以及物体之间的空间、功能和语义关系。这里仍有许多工作待完成，例如探索自回归模型和层次模型各自的优缺点，并开发新的模型。一个值得关注的挑战是如何从数据中自动学习层次结构。对于层次模型而言，递归变分自编码器尤其有前景，因为它们既能发挥神经网络和分布式表示的能力，又能灵活选择编码器架构，同时解码为基于对象的结构。

这里的一个问题是数据。二维图像识别的发展得益于大规模数据集。目前可用的大型三维数据集非常少，没有任何一个达到例如 ImageNet 那样的规模，而 ImageNet 包含一百多万张图像。值得一提的数据集是 CAD-Estate（Maninis、Popov、Niessner 和 Ferrari，2023年），它由 19,512 个复杂房地产场景的视频构建而成。每个物体都标注了数据库中的一个 CAD模型，并被置于场景的三维坐标系中。以这种方式标注的物体超过 10 万个。当然，获取并标注一个三维场景所需的工作量，远远大于仅仅标注图像中的边界框。计算机图形学当然可以提供丰富的三维真值和标注来源（例如参见 Johnson 等人于 2017 年提出的 CLEVR 数据集 Johnson 等人，2017），还可以提供一种受控方式来测试组合泛化（见其第 4.7 节）。然而，这需要良好的物体模型，尤其是场景模型，以生成逼真的场景布局；此外，还存在一个问题：在此类渲染数据上训练的模型能否很好地迁移到真实场景中。

虽然收集三维数据集颇具挑战，但也可以收集多视图数据，即对给定场景（或其部分）拍摄的多张图像，并且已知相机参数（内参和外参）。例如，可参考 UNC[17] 的主动视觉数据集，以及 Meta 的 Aria数据集[18]。可以通过预测从一个新的测试视角会观察到什么，来测试场景生成模型在多视图数据上的表现。

推断：
场景级推理颇具挑战，因为需要将层次场景模型的各个部分与图像数据相匹配。对文法和与或图进行精确推断的扩展性可能较差，但 GFlowNets（Hu 等人，2023年）等近似方法，以及递归变分自编码器等替代模型，或许前景可期。正如上文所见，推断可能涉及模型中自底向上和自顶向下的信息流。这可能导致复杂的推断过程，让人想起早期视觉模型中的推断过程，例如 VISIONS（Hanson & Riseman，1978年）以及Ohta 等人（1978年）提出的系统。采用“惰性推理”的思路或许可以在一定程度上简化这一过程，即仅根据给定任务激活整个场景表示中的部分内容。至于模型缺陷，要取得进展，关键是收集大量有关最常见缺陷类型的数据，以便解决重要问题。

任务和基准：
目前，大多数计算机视觉任务都在像平面上进行评估。这在一定程度上可能是因为在这种情况下，收集图像并创建标注相对容易。不过，也有一些三维基准测试，例如 KITTI 套件[19] 和 nuScenes[20]。这些基准包括道路场景中的三维目标检测（使用三维边界框）。自动驾驶以及杂乱场景中的料箱拣选任务，都是可能推动三维重建研究方向发展的领域。

为了发挥结构化生成模型的价值，必须定义一组能够利用同一底层表示的任务。第1.3节给出了一些具有挑战性的任务示例；其中包括严重遮挡情况下的物体重建、场景编辑和场景补全任务，这些任务都能

[17] https://www.cs.unc.edu/~ammirato/主动视觉数据集网站/。[18] https://about.meta.com/realitylabs/projectaria/ 数据集/。[19] http://www.cvlibs.net/datasets/kitti/。[20] https://www.nuscenes.org。

利用 SGM 表示中包含的三维和场景级信息。研究界需要重点关注并审慎选择三维/多视图基准和任务，以推动 SGM 方法的发展。

致谢

感谢 David Hogg、Alan Yuille、Oisin Mac Aodha、Antonio Vergari、Siddharth N.、Kevin Murphy、Paul Henderson、Adam Kortylewski、Hakan Bilen、Titas Anciukevicius、Nikolay Malkin 和 Bradley Love 提供宝贵的意见并参与讨论；同时感谢匿名审稿人的意见，这些意见帮助改进了本文。

为实现开放获取，作者已对本次投稿产生的任何作者接受稿版本采用知识共享署名（CC BY）许可协议。

参考文献

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C., Parikh, D.（2015年）。VQA：视觉问答。*2015*年 *IEEE* 计算机视觉国际会议（*ICCV*）（第2425–2433页）。

Babalola, K.O., Cootes, T.F., Twining, C.J., Petrovic, V., Taylor, C.（2008）。使用主动外观模型和局部回归器进行三维脑分割。D. Metaxas、L. Axel、G. Fichtinger 和 G. Sz´ ekely（编），医学图像计算与计算机辅助干预国际会议（*MICCAI 2008*）（第401–408页）。施普林格。（计算机科学讲义，第5241卷）

Bengio, Y., Louradour, J., Collobert, R., Weston, J.（2009）。课程学习。L. Bot- tou & M.Littman（编），第二十六届国际机器学习会议论文集（*ICML 2009*）（第41-48页）。

Bernardo, J.M., & Smith, A.F.M.（1994）。贝叶斯理论。奇切斯特，英国：Wiley。

Biederman, I.（1987）。基于成分的识别：人类图像理解理论。《心理学评论》，*94*，115-147。

Biederman, I., Mezzanotte, R.J.,Rabinowitz, J.C.（1982）。场景感知：检测并判断发生关系违例的物体。《认知心理学》，*14*，143-177。

Blanz, V., & Vetter, T.（1999）。用于合成三维人脸的可变形模型。*SIGGRAPH 99*：第*26*届计算机图形学与交互技术年度会议论文集 （第187-194页）。ACM出版社。

Box, G.E.P., & Draper, N.R.（1987年）。经验模型构建与响应面。John Wiley & Sons。

Brodatz, P.（1966）。纹理：艺术家与设计师的摄影集。纽约：Dover出版社。

Charniak, E. (1993). 统计语言学习. 马萨诸塞州剑桥：麻省理工学院出版社。

Collins, J., Goael, S., Deng, K., 等（2022）。ABO：用于真实世界三维物体理解的数据集与基准。*IEEE*计算机视觉与模式识别会议论文集，*CVPR 2022*。

Cootes, T.F., Edwards, G.J., Taylor, C.J.（1998）。主动外观模型。H. Burkhardt & B. Neumann（编），第五届欧洲计算机视觉会议论文集，*ECCV 1998* （第 II 卷，第 484–498 页）。施普林格。（计算机科学讲义 1407）

Cootes, T.F., Taylor, C.J., Cooper,D.H., Graham, J.（1995）。主动形状模型——其训练与应用。计算机视觉与图像理解，*61(1)*，38–59。

Cover, T.M., & Thomas, J.A.（1991）。信息论基础。John Wiley & Sons。

Dayan, P., Hinton, G.E., Neal, R.M., Zemel, R.S.（1995）。亥姆霍兹机。神经计算，*7(5)*，889-904。

De Sousa Ribeiro，F.，Duarte，K.，Everett，M.，Leontidis，G.，Shah，M.（2022）。胶囊学习：综述。（arXiv:2206.02664）

Du, Y., Durkan, C., Strudel, R., Tenenbaum, J. B., Dieleman, S., Fergus, R., . . . Grathwohl, W. (2023年). 化繁为简、重复利用、循环再用：利用基于能量的扩散模型和马尔可夫链蒙特卡罗进行组合式生成。第*40*届国际机器学习会议论文集 (第202卷). PMLR.

Efros, A.A., & Leung, T.K. (1999). 通过非参数采样进行纹理合成。国际计算机视觉会议论文集（*ICCV 1999*）(第1033-1038页). Eslami, S.M.A., Heess, N., Weber, T., Tassa, Y., Szepesvari, D., Kavukcuoglu, K., Hinton, G.E. (2016). 注意、推断、重复：利用生成模型快速理解场景。D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (编), 神经信息处理系统进展。Eslami, S.M.A., & Williams, C.K.I. (2011). 用于基于部件的物体理解的因子化形状与外观。*2011*年英国机器视觉会议论文集。Eslami, S.M.A., & Williams, C.K.I. (2012). 基于部件的物体分割生成模型。P. Bartlett, F. Pereira, C. Burges, L. Bottou, & K. Weinberger (编),神经信息处理系统进展 *25* (第100-107页). Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A. (2010). PASCAL视觉物体类别（VOC）挑战赛。国际计算机视觉期刊, *88(2)*, 303-338.

Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D. (2009). 基于判别式训练的部件模型物体检测。*IEEE* 模式分析汇刊与机器智能，*32(9)*，1627-1645。

Fischler, M.A., & Bolles, R.C. (1981). 随机采样一致性：一种用于模型拟合的方法，应用于图像分析和自动制图。*ACM*通讯，*24*(6), 381-395.

Fischler, M.A., & Elschlager, R.A. (1973). 图像结构的表示与匹配。*IEEE*计算机汇刊，*22(1)*, 67-92。

Forsyth, D.A., & Ponce, J. (2003). 计算机视觉：现代方法（第二版）。Prentice Hall。

Frey, B.J., & Jojic, N. (2003). 使用 EM 算法进行变换不变聚类。*IEEE* 模式分析与机器智能汇刊，*25(1)*，1-17。

Fu, K.S.（1982）。句法模式识别与应用。Prentice-Hall。

Galerne, B., Gousseau, Y., Morel, J.-M.（2011）。随机相位纹理：理论与综合。*IEEE* 图像处理汇刊，*20(1)*，257-267。

Gao, L., Sun, J.-M., Mo, K., Lai, Y.-K., Guibas, L.J., Yang, J.（2023年）。*SceneHGN*：基于层次图网络的三维室内场景生成与细粒度几何结构。（arXiv:2302.10237）

Gemini团队（2023年）。*Gemini*：高能力多模态模型系列。（Google。arXiv:2312.11805）

Greff, K., Kaufman, R.L., Kabra, R., Watters, N., Burgess, C., Zoran, D., . . . Lerchner, A.（2019年）。使用迭代变分推断进行多对象表征学习。*ICML* （第97卷，第2424–2433页）。

Grenander, U.（1978年）。模式理论讲义：第*2*卷，模式分析。施普林格出版社。Hall, P.A.V.（1973）。与/或图与上下文无关文法之间的等价性。*ACM*通讯，*16(7)*，444-445。Hanson, A.R. 与 Riseman, E.M.（1978年）。VISIONS：用于解释场景的计算机系统。A.R. Hanson 与 E.M. Riseman（编），计算机视觉系统。学术出版社。Heitz, G. 与 Koller, D.（2008）。学习空间上下文：利用物质寻找实体。*2008*年欧洲计算机视觉会议论文集。Hinton, G.E.（2002年）。通过最小化对比散度训练专家乘积模型。神经计算，*14*，1771-1800。Hinton, G.E.、Krizhevsky, A.、Wang, S.D.（2011）。变换自编码器。*ICANN 2011*论文集。Hinton, G.E.、Sabour, S.、Frosst, N.（2018）。采用EM路由的矩阵胶囊。国际学习表征会议。Hu, E.J.、Malkin, N.、Jain, M.、Everett, K.、Graikos, A.、Bengio, Y.（2023年）。用于学习组合式潜变量模型的*GFlowNet-EM*。（arXiv:2302.06576）Hueting, M.、Reddy, P.、Kim, V.G.、Yumer, E.、Carr, N.、Mitra, N.J.（2018）。Seethrough：在严重遮挡的室内场景图像中寻找物体。国际三维视觉会议（*3DV*）论文集。Izadinia, H.、Shan, Q.、Setiz, S.M.（2017）。IM2CAD。*IEEE*计算机视觉与模式识别会议论文集，*CVPR 2017*。

Jammalamadaka, N., Zisserman, A., Eichner, M., Ferrari, V., Jawahar, C. (2012年). 我的算法成功了吗？人体姿态估计器评估工具。A. Fitzgib- bon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (编), 计算机视觉 ——*ECCV 2012* 年。施普林格。（计算机科学讲义 7574 ）

Jang, W., & Agapito, L. (2021年). CodeN- eRF：面向物体类别的解耦神经辐射场。*IEEE/CVF*国际计算机视觉会议论文集（*ICCV*）（第12949-12958 页）。（亦可通过arXiv:2109.01750 获取）

Jin, Y., & Geman, S. (2006年). 概率图像模型中的上下文与层级。*CVPR 2006*年论文集。 IEEE计算机学会。

Johnson, J.、Hariharan, B.、van der Maaten, L.、Fei-Fei, L.、Zitnick, C.L.、Girshick, R.（2017）。CLEVR：用于组合语言和基本视觉推理的诊断数据集。*IEEE*计算机视觉与模式识别会议论文集（第1988–1997页）。

Jordan, M.I.、Ghahramani, Z.、Jaakkola, L.K.、T. S. Saul。（1999）。图模型变分方法简介。机器学习，*37*，183-233。

Kato, H.、Beker, D.、Morariu, M.、Ando, T.、Matsuoka, T.、Kehl, W.、Gaidon, A.（2020）。可微渲染：一项综述。（https://arxiv.org/2006.12057）

Kim, Y., Dyer, C., Rush, A.（2019年）。用于文法归纳的复合概率上下文无关文法。第*57*届计算语言学协会年会论文集（第2369–2385页）。Kingma, D.P., & Welling, M.（2014）。自编码变分贝叶斯。*ICLR*。Kivinen, J.J., & Williams, C.K.I.（2012年）。多纹理玻尔兹曼机。第十五届国际人工智能与统计会议论文集。

Kosiorek, A., Sabour, S., Teh, Y.W., Hinton, G.E.（2019年）。堆叠式胶囊自编码器。H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alch´e Buc, E. Fox, & R. Garnett（编），神经信息处理系统进展第*32*卷。

Kulkarni, T.D., Whitney, W.F., Kohli, P., Tenen- baum, J.（2015年）。深度卷积逆图形网络。C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett（编），神经信息处理系统进展（第 28 卷）。

Land, M., Mennie,N.,Rusted,J.（1999）。视觉和眼球运动在日常生活活动控制中的作用。《知觉》，*28*，1311-1328。

Lee, H., Grosse, R., Ranganath, R., Ng, A.Y.（2009）。用于层次化表征可扩展无监督学习的卷积深度置信网络。L. Bottou & M. Littman（编），第二十六届国际机器学习会议论文集（*ICML 2009*）。

Lee, J., Lee, Y., Kim, J., Kosiorek, A.R., Choi, S., Teh, Y.-W.（2019）。集合变换器：一种基于注意力的置换不变神经网络框架。第三十六届国际机器学习会议论文集（第3744–3753页）。

Li, M., Gadi Patil, A., Xu, K., Chaudhuri, S., Khan, O., Shamir, A., .. . Zhang, H.（2019）。GRAINS：室内场景生成式递归自编码器。*ACM* 图形学汇刊，*37(2)*，1-16。Li, N., Eastwood, C., Fisher, R.（2020年）。从多视图中学习多对象场景以对象为中心的表征。H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin（编），神经信息处理系统进展（第33卷，第5656–5666页）。

Liu, T., Chaudhuri, S., Kim, V., Huang, Q.- X., Mitra, N.J., Funkhouser, T.（2014）。使用概率文法创建一致的场景图。*ACM* 图形学汇刊（*SIGGRAPH Asia* 论文集），*33*（6）。

Locatello, F., Bauer, S., Lucic, M., R¨atsch, G., Gelly, S., Sch¨olkopf, B., Bachem, O.（2019）。质疑无监督学习解缠表示中的常见假设。第三十六届国际机器学习会议（*ICML*）论文集。

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., . . . Kipf, T. (2020). 基于槽位注意力的物体中心学习。神经信息处理系统进展第 *33* 卷 (第 11525–11538 页)。Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M. (2015). SMPL：一种蒙皮多人线性模型。*ACM* 图形学汇刊,*34(6)*。Loper, M.M., & Black, M.J. (2014). OpenDR：一种近似可微渲染器。计算机视觉—*ECCV 2014* (第 154-169 页)。MacKay, D.J.C. (2003). 信息论、推断与学习算法。剑桥：剑桥大学出版社。Maninis, K.-K., Popov, S., Niessner, M., Ferrari, V. (2023). CAD-Estate：RGB 视频中的大规模 CAD 模型标注。*I EEE/CVF* 国际计算机视觉大会（*ICCV*）论文集(第 20132-20142 页)。Marcus, G., Davis, E., Aaronson, S. (2022).对 *DALL-E 2* 的一项非常初步的分析。(arXiv:2204.13807) Mildenhall, B., Srinivasan, P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R. (2020). NeRF：将场景表示为用于视图合成的神经辐射场。A. Vedaldi,

H. Bischof, T. Brox, & J. Frahm (编),计算机视觉——*ECCV 2020*。施普林格。（计算机科学讲义第 12346 卷）

Moreno, P., Williams, C.K.I., Nash, C., Kohli, P. (2016). 使用逆向图形学克服遮挡。H. Gang & H. Jegou (编)，计算机视觉 ——*ECCV 2016* 研讨会论文集第三部分（第170-185页）。施普林格。（计算机科学讲义 9915）

Murphy, K.P. (2012年）。机器学习：概率视角。MIT出版社。

Murphy, K.P. (2023年）。概率机器学习：高级主题。MIT出版社。

Nazabal, A., Tsagkas,N.,Williams, C.K.I. (2023年）。生成式胶囊模型的推断与学习。神经计算，*35(4)*，727-761。

Nichol, A., Dhariwal, P., Ramesh, A., Shya m, P., Mishkin, P., McGrew, B., . . . Chen, M. (2022). GLIDE：利用文本引导扩散模型实现照片级真实感图像生成与编辑。第*39*届国际机器学习会议论文集（第162卷）。PMLR。Ohta, Y., Kanade, T., Sakai, T. (1978年）。一种用于含有子结构物体的场景的分析系统。第*4*届国际模式识别会议（第752-754页）。Oliva, A., & Torralba, A. (2006年）。场景概要的构建：全局图像特征在识别中的作用。脑研究进展，*155*，23-36。

Open AI（2023）。*GPT-4*技术报告。（ arXiv:2303.08774）

Palmer, S.E.（1999）。视觉科学：从光子到现象学。麻省理工学院出版社。

Paschalidou, D., Kar, A., Shugrina, M., Kreis, K., Geiger,A., Fidler, S.（2021年）。ATISS：用于室内场景合成的自回归 Transformer。M. Ranzato, A. Beygelzimer,

Y. Dauphin, P. Liang, & J.W. Vaughan（编），神经信息处理系统进展（第34卷，第12013–12026页）。Curran Associates, Inc.

Pollack, J.B.（1990）。递归分布式表示。《人工智能》，*46*，77-105。Prince, S.J.D.（2012）。计算机视觉：模型、学习与推断。剑桥大学出版社。

Quinn, J.A.、Williams, C.K.I.、McIntosh, N.（2009）。应用于生理状况监测的因子切换线性动态系统。《*IEEE* 模式分析与机器智能汇刊》，*31(9)*，1537-1551。

Radford, A.、Metz, L.、Chintala, S.（2016）。使用深度卷积生成对抗网络进行无监督表征学习。国际学习表征会议（*ICLR 2016*）。Ramesh, A.、Pavlov, M.、Goh, G.、Gray, S.、Voss, C.、Radford, A.、……Sutskever, I.（2021）。零样本文本到图像生成。M. Meila & T. Zhang（编），第*38*届国际机器学习会议论文集 （第139卷，第8821–8831页）。PMLR。Ranzato, M.、Mnih, V.、Hinton, G.E.（2010）。使用门控MRF生成更逼真的图像。J. Lafferty、C. Williams、J. Shawe-Taylor、R. Zemel & A. Culotta（编），神经信息处理系统进展 （第23卷）。Rasmussen, C.E. & Williams, C.K.I.（2006）。机器学习中的高斯过程。马萨诸塞州剑桥：麻省理工学院出版社。Rezende, D.J.、Mohamed, S.、Wierstra, D.（2014）。深度生成模型中的随机反向传播与近似推断。*ICML*。

Ritchie, D.（2019）。三维场景生成模型。（三维结构生成模型学习教程的一部分，发表于 Eurographics 2019，https://3dstructgen.github.io）

Ritchie, D., Wang, K., Lin, Y.-A.（2019）。通过深度卷积生成模型快速、灵活地合成室内场景。*IEEE*计算机视觉与模式识别会议论文集（*CVPR 2019*）。

Roberts, L.G.（1963年）。三维实体的机器感知（技术报告第315号）。MIT林肯实验室。

Romaszko, L., Williams, C.K.I., Winn, J.（2020年）。学习用于场景理解的直接优化。模式识别，*105*，107369。

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.（2022）。使用潜在扩散模型进行高分辨率图像合成。*IEEE*计算机视觉与模式识别会议论文集，*CVPR 2022*。

Ross, D., & Zemel, R.（2006年）。学习基于部件的数据表示。机器学习研究杂志，*7*，2369-2397。

Roth, S., & Black, M.J.（2005）。专家场：学习图像先验分布的框架。*IEEE*计算机视觉与模式识别会议论文集，*CVPR 2005*（第III卷：860-867页）。

Rozanov, J.A.（1977）。马尔可夫随机场与随机偏微分方程。《苏联数学文集》，*32(4)*，515-534。

R¨unz, M., Li, K., Tang, M., Ma, L., Kong, C., Schmidt, T., . .. Newcombe, R. (2020年). FroDO：从目标检测到三维物体。*IEEE*计算机视觉与模式识别会议论文集（*CVPR 2020*）。

Sabour, S., Frosst, N., Hinton, G. (2017). 胶囊之间的动态路由。神经信息处理系统进展（第3856–3866 页）。Salimans, T., Karpathy, A., Chen, X., Kingma, D.P., Bulatov, Y. (2017). PixelCNN++：一种采用离散化逻辑斯蒂混合似然及其他修改的 PixelCNN 实现。国际学习表征会议论文集（*ICLR 2017*）。Sch¨onberger, J.L., & Frahm, J.-M. (2016). 运动恢复结构的重新审视。*IEEE*计算机视觉与模式识别会议论文集（*CVPR 2016*）。

Seth, S., Murray, I., Williams, C.K.I.（2019）。潜在空间中的模型批判。贝叶斯分析，*14(3)*，703-725。

Shi, M., Caesar, H., Ferrari, V.（2017）。使用物体与背景转移进行弱监督目标定位。国际计算机视觉会议论文集（*ICCV 2017*）。

Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.（2017）。从单幅深度图像进行语义场景补全。*IEEE*计算机视觉与模式识别会议论文集，*CVPR 2017*。Sun, M., Kim, B., Kohli, P., Savarese, S.（2014）。通过物体属性交互关联实体与物质。*IEEE*模式分析与机器智能汇刊，*36(7)*，1370-1383。

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. (2013). 神经网络的迷人特性。(arXiv:1312.6199)

Szeliski,R.（2021年）。计算机视觉：算法与应用（第二版）。施普林格。

Tenenbaum, J.B.，& Freeman, W.T.（2000年）。利用双线性模型分离风格与内容。神经计算，*12*，1247-1283。Torralba, A.，Murphy, K.P.，Freeman, W.T.（2010）。利用森林观察树木：运用上下文进行视觉物体检测与定位。*ACM*通讯，*53(3)*，107–114。Tu, Z.，& Zhu, S.-C.（2002）。数据驱动马尔可夫链蒙特卡罗图像分割。*IEEE*模式分析与机器智能汇刊，*24(5)*，657-673。van den Burg, G.J.J.，& Williams, C.K.I.（2021年）。概率深度生成模型中的记忆现象。M. Ranzato、A. Beygelzimer、Y. Dauphin、P. Liang，& J.W. Vaughan（编），神经信息处理系统进展（第34卷，第27916–27928页）。Curran Associates, Inc. van den Oord, A.、Kalchbrenner, N.、Espeholt, L.、kavukcuoglu, k.、Vinyals, O.、Graves, A.（2016）。利用PixelCNN解码器进行条件图像生成。D. Lee、M. Sugiyama、U. Luxburg、I. Guyon，& R. Garnett（编），神经信息处理系统进展。Vasilescu, M.A.O.，& Terzopoulos, D.（2002）。图像集合的多线性分析：张量人脸。A. Heyden、G. Sparr、M. Nielsen，& P. Johansen（编），计算机视觉–*ECCV 2002*。施普林格。（计算机科学讲义2350）Wang, A.、Kortylewski, A.、Yuille, A.（2021年）。NeMo：用于稳健三维姿态估计的对比特征神经网格模型。国际学习表征会议论文集（*ICLR 2021*）。Williams, C.K.I.，& Titsias, M.K.（2004）。利用稳健统计和因子学习对图像中的多个物体进行贪心学习。神经计算，*16*（5），1039-1062。

Xia, Y., Zhang, Y., Liu, F., Shen, W., Yuille, A. (2020). 先合成再比较：检测语义分割中的失败与异常。A. Vedaldi, H. Bischof, T. Brox, & J. Frahm (编), 计算机视觉——*ECCV 2020*。施普林格。（计算机科学讲义 12346）Yao, B., Yang, X., Zhu, S.-C. (2007). 大规模通用真实标注数据库简介：方法、标注工具与基准。*EMMCVPR'07*：第六届国际计算机视觉与模式识别能量最小化方法会议论文集。Yeh, Y.-T., Yang, L., Watson, M., Goodman, N.D., Hanrahan, P. (2012). 使用局部退火可逆跳跃马尔可夫链蒙特卡罗，在约束条件下合成开放世界。*ACM* 图形学汇刊，*31(4)*，1-11。Yu, L.-F., Yeung, S.-K., Tang, C.-K., Terzopoulos, D., Chan, T.F., Osher, S.J. (2011). 安居：家具布置的自动优化。*ACM* 图形学汇刊，*30(4)*。Yuille, A.L., & Liu, C. (2021). 深度网络：它们究竟为视觉做过什么？国际计算机视觉杂志，*129*，781-802。Zamir, A.R., Sax, A., Yeo, T., Kar, O., Cheerla, N., Suri, R., . . . Guibas, L. (2020). 通过跨任务一致性实现稳健学习。*IEEE*计算机视觉与模式识别会议论文集，*CVPR 2020*。Zhang, X., Srinivasan, P.P., Deng, B., Debevec, P., Freeman, W.T., Barron, J.T. (2021). NeR-Factor：未知光照条件下形状与反射率的神经因子分解。*ACM* 图形学汇刊（*SIGGRAPH Asia* 论文集），*40*(6)。

Zhu, S.-C.,& Huang, S. (2021). 计算机视觉：用于解析物体、场景与事件的随机文法。施普林格。
Zhu, S.-C., & Mumford, D. (2006). 图像随机文法。计算机图形学与视觉基础与趋势，*2(4)*，259-362。Zhu, S.-C., Wu, Y., Mumford, D. (1998). 滤波器、随机场与最大熵（FRAME）：迈向纹理建模的统一理论。国际计算机视觉杂志，*27(2)*，107-126。