用于混合现实的三维场景生成技术

3D Scene Generation for Mixed Reality

江海燕/JIANG Haiyan[1]，东野啸诺/DONGYE Xiaonuo[1]，王涌天/WANG Yongtian[1,2]
（1. 北京市混合现实与新型显示工程技术研究中心，北京理工大学光电学院，中国 北京 100081；
2. 北理工郑州智能科技研究院，中国 郑州 450000）
(1. Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China;
2. Zhengzhou Academy of Intelligent Technology, Zhengzhou 450000, China)

DOI：10.12142/ZTETJ.2024S1007
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20240724.1130.014.html
网络出版日期：2024-07-25
收稿日期：2023-11-26

摘要：在混合现实系统中，三维场景作为虚拟空间的关键构成要素，其高效生成方法一直是本领域的研究热点。人工智能辅助内容生成技术的发展，为该问题的解决提供了新的思路。综述性的归纳与总结了近年来三维场景生成的各项技术方法，以及混合现实场景下三维场景生成的现状，并对其发展趋势进行了分析与展望。

关键词：三维场景生成；混合现实；人工智能

Abstract: Mixed reality, as a typical form of a multimodal digital information system, aims to seamlessly merge virtual information with real-world information. It is one of the key technologies for the next-generation Internet. In mixed reality systems, the efficient generation of three-dimensional scenes, as the core element of virtual space, has been a key research focus in this field. In recent years, the development of artificial intelligence-assisted content generation method has introduced novel approaches to addressing this issue. This paper provides a synthesis and summary of various methods for three-dimensional scene generation in recent years and offers an analysis and outlook on their development trends.

Keywords: 3D scene generation; mixed reality; AI

引用格式：江海燕，东野啸诺，王涌天．用于混合现实的三维场景生成技术 [J]. 中兴通讯技术，2024, 30(S1): 43-53. DOI: 10.12142/ZTETJ.2024S1007
Citation: JIANG H Y, DONGYE X N, WANG Y T. 3D scene generation for mixed reality [J]. ZTE technology journal, 2024, 30(S1): 43-53. DOI: 10.12142/ZTETJ.2024S1007

混合现实技术致力于实现真实世界、虚拟世界和参与者三者之间的无缝融合，其最终目的是实现自然逼真的虚实融合人机交互。该技术既解决了虚拟现实中用户因无法看到真实环境导致行动受限的问题，也通过叠加虚拟信息的方式扩展了物理世界的边界，在医疗康复、教学培训、航空航天、娱乐休闲等领域具有广阔的应用前景。

三维场景的生成是混合现实场景实现高沉浸、自由交互的前提，也是该方向的研究重点。近年来，随着头戴显示器、立体投影等硬件设备的不断成熟，使得人们对三维场景生成的需求不断提升。同时，人工智能技术的发展，尤其是智能生成技术的快速发展，为这一领域注入了新的活力。

目前，三维场景自动生成的技术方法主要包括自回归神经网络、语法过程建模、图推理、自注意力模型等。随着多模态生成模型的发展，还可以通过自然语言条件约束、扩散模型来控制场景的生成。

具体到混合现实环境中，由于用户所处的物理环境与虚拟环境相互影响，三维场景的生成不仅需要考虑虚拟场景的需求，也要考虑用户以及周围物理环境的因素。针对用户因素，一些方法会通过用户参与的交互过程来控制场景的生成，实现更符合用户意图的混合现实环境。这类人在环的半自动生成方法允许用户实时控制和选择虚拟物体，具有很好的可控性。然而，由于其人在环的特性，场景生成速度较为缓慢。针对物理环境因素，一些方法通过实时动态监测物理环境，并将提取到的物理信息用于虚拟环境的生成，实现具有物理环境特征的实时三维场景生成。

基金项目：国家自然科学基金重点项目（62332003）；长沙市2022年科技重大专项（kh2301019）

1 通用三维场景生成技术

三维场景生成通常使用计算机图形学技术和计算机视觉技术，自动生成逼真的三维室内外环境。这些场景可用于虚拟现实、游戏开发、电影制作等多个领域。

从整体思路上来说，当前的主要研究倾向于将预先创建好的单体模型通过某种方式实现自动布局以构成所需要的三维场景。从方法上来说，三维场景生成的主要技术手段包括自回归神经网络、语法过程建模、图推理、自注意力模型以及扩散模型等。在这些技术手段之上，设计者可以通过加入场景生成的先验知识，对生成的三维模型进行风格化创建，实现条件场景的生成。

1.1 基于自回归神经网络的方法

基于自回归神经网络的方法常用于生成具有高真实感的室内场景。这种方法一般会利用网络来学习输入数据中物体出现及相互关联关系的概率分布，然后用其生成与训练数据相似的场景。

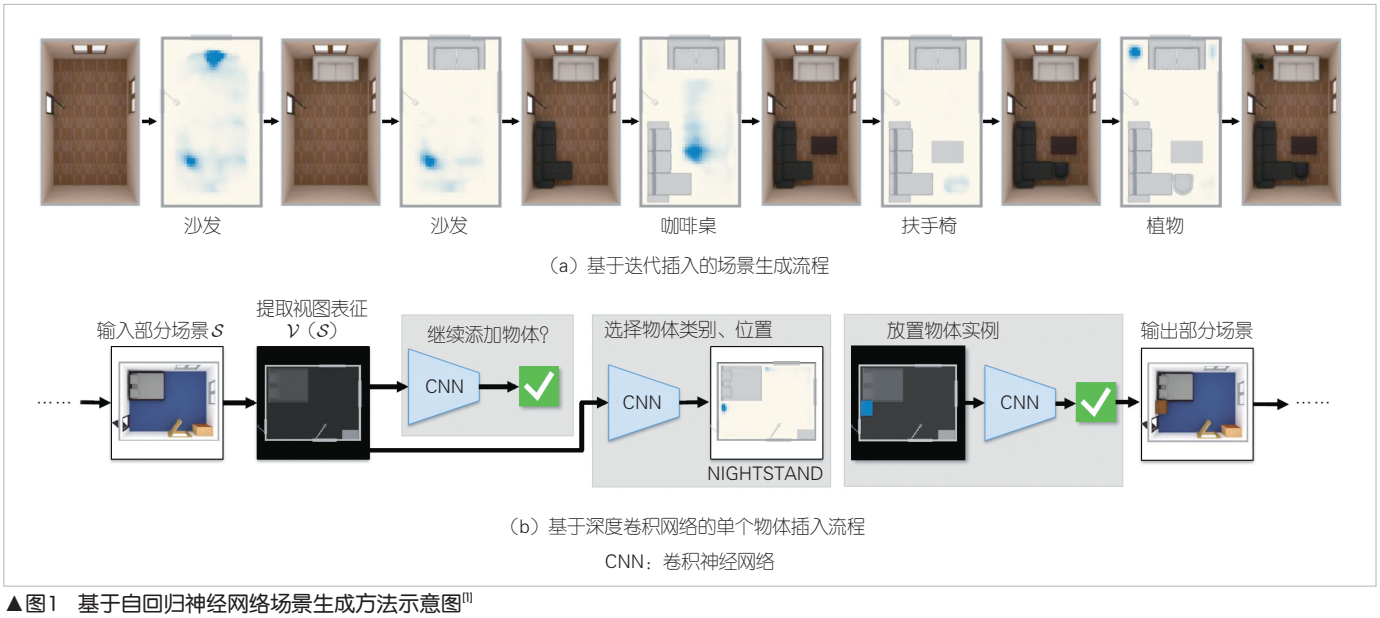
2018年，WANG等[1]提出了基于图像的自回归深度卷积网络的室内场景合成方法，实现从零开始迭代生成包含多个物体的房间，如图1所示。该方法仅将房间矩形轮廓作为输入，采用基于正交自上而下视图表示，通过卷积神经网络实现单个物体的添加。但在该方法中，一个模型只能针对特定房间类型（如卧室、客厅、办公室），进行生成，并且忽略了房间的层次关系、对象之间的功能关系、物体的大小等。此外，该方法是基于局部进行推理，难以用于推理物体在全局坐标中的位置。

基于上述工作，2019年，RITCHIE等[2]采用自上而下的平面图像作为输入，并加入房间的几何信息（如天花板、墙等），使用自回归深度卷积神经网络实现快速的场景生成。相比于上述方法，该方法通过使用单独的神经网络模块预测对象的类别、位置、方向和大小，实现了部分场景的自动补全以及完整场景的合成，并且生成单个场景的平均速度达到1.858 s，而上述方法生成单个场景耗时约240 s。但同上述方法一样，该方法仍然忽略了场景的层次结构，并且难以生成具有风格一致性的场景。

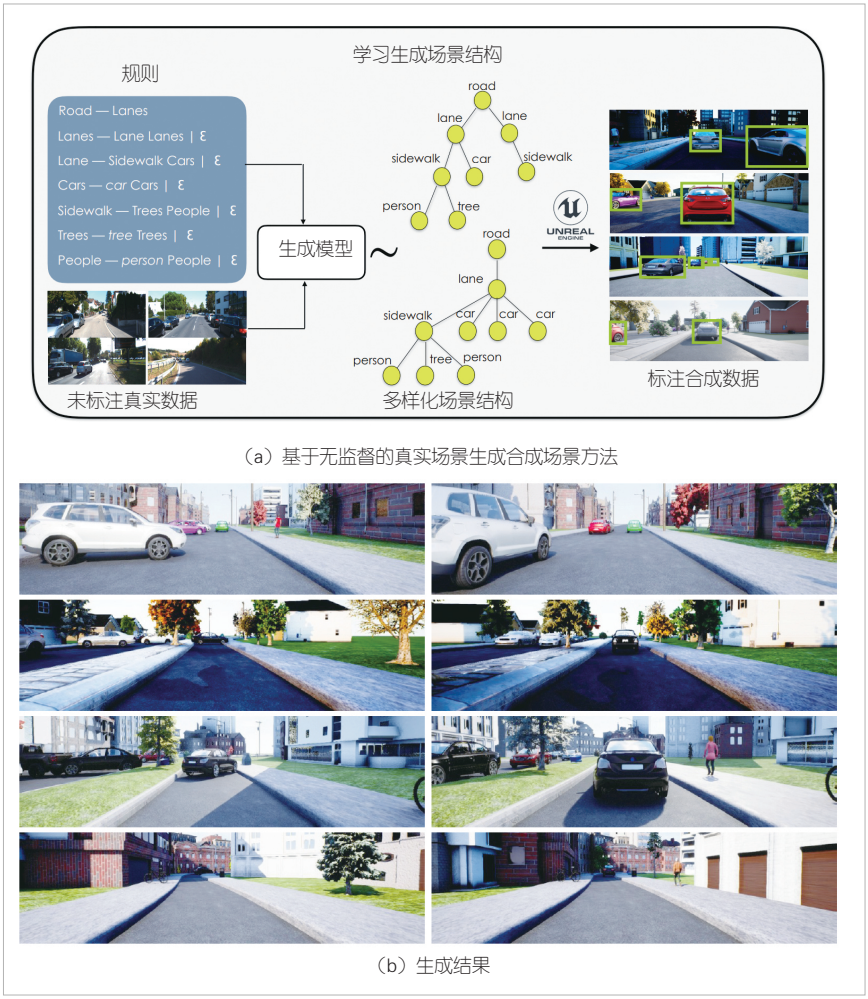
1.2 基于语法的过程建模方法

基于语法的过程建模方法可用于生成具有较为明确层次感的场景,通过从给定的概率场景语法图中采样，逐步添加或者修改物体属性从而实现场景的生成。

2019年，KAR等相继提出了Meta-Sim[3]以及Meta-sim2[4]算法。Meta-Sim将场景解析为概率场景语法，通过神经网络从语法中采样场景结构，并实现场景中物体位置、姿态以及其他属性的修改，实现更符合真实关联关系的场景生成，如图2所示。Meta-Sim算法可以灵活地调整合成内容的结构和外观。基于这个特性，该算法可以通过优化生成更符合下游任务的场景数据集。但是由于Meta-Sim依赖语法获取场景结构，其可以生成的场景仍然受到限制。基于Meta-Sim以非监督学习的方式实现过程建模的场景生成，Meta-sim2比较真实空间合生成场景的特征空间离散度，通过强化学习方式学习给定的概率场景语法顺序采样规则，得到更符合真实场景分布的生成场景。



▲图1 基于自回归神经网络场景生成方法示意图[1]



▲图2　基于语法的过程建模方法示意图[4]

2020年，PURKAIT等提出SG-VAE算法[5]。该算法通过基于语法的自编码器，学习不同对象类别的形状和位置等参数，从而实现紧凑而准确的场景布局。SG-VAE算法从训练数据中，提取“物体是否可以同时存在”的信息，并由此进行推理。随后该算法将推理后形成的生成规范用于自动构建语法信息，并通过增强解析树来表示场景，以确保生成的场景始终符合正确的语义信息。基于语法的过程建模方法也可以用于超大规模的场景生成任务中。早在2001年，PARISH和MÜLLER就提出了一个模拟城市的系统[7]。该系统将土地划分为地块，为各地块分配的建筑物创建适当的几何形状，并可以连接各个地块生成一个公路网络和街道系统。基于生成的三维几何信息，该方法在几何信息上添加额外的纹理，以赋予建筑物更多的细节。生成城市的过程中，利用图像地图作为输入数据，控制道路和建筑的分布和形状。在道路网络生成环节中，该系统分别使用高速公路和街道进行区域划分。在建筑生成方面，该系统使用另一个参数的随机化系统来生成建筑的几何信息。每个建筑都是由一个任意的地面轮廓经过变换和挤压而成，形成摩天大楼、商业建筑和住宅房屋等不同种类的建筑，并分别由分区规则和图像地图来控制其生成。2006年，PARISH等[6]提出了一种形状语法，用于生成具有高视觉质量和几何细节的建筑外壳。该方法不仅可以高效地创建大规模的城市模型，还可以表达复杂的屋顶和立面结构。此外，该方法还分析了从简单的体积形状生成复杂的建筑外壳所面临的问题，并提出了两种重要的机制：遮挡查询和吸附查询，用以处理形状之间的相互作用和冲突。

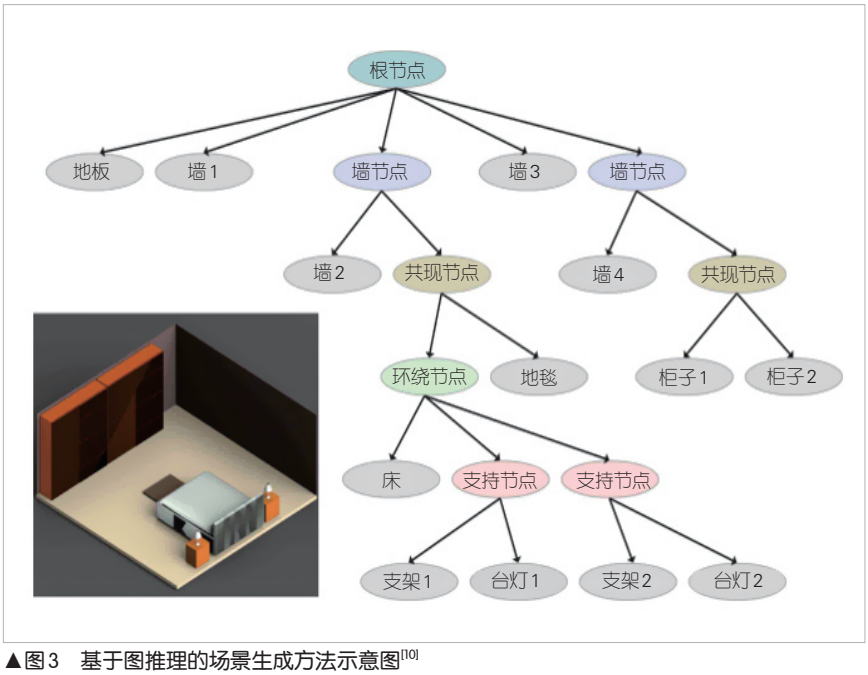
1.3 基于图推理的方法

基于图推理的方法利用图结构来表示和推理室内场景中物体之间的关系。图结构不仅可以描述物体的类别、位置、朝向等属性，也可以描述物体之间的相对距离、方向、对齐、支撑等关系。基于图推理的方法可以根据输入的房间形状和大小，或者参考样例场景，生成符合物体约束关系的室内场景。根据场景输入是否存在参考样例，可以将室内场景分为无样例生成算法和有样例生成算法。无样例算法通常从大规模数据中总结规则来生成场景，而有样例算法则基于文本、图像等输入，要求生成场景与输入在一定程度上匹配，属于有条件的场景生成任务。

无样例生成算法中，对象之间的排列常由向量表示，而这种抽象表示容易忽略几何细节，因此德克萨斯大学的ZHANG等[8]于2020年提出一种三维对象排列表示方法。三维对象排列表示基于对象的大小和形状属性对对象的位置和方向进行建模。通过与投影二维图像表示组合来训练三维场景生成器，同时兼具了二维场景生成和三维场景生成的优点。此外，该表示可以使用基于数据驱动的方法，通过分析数据集中对象的“共现”来提取先验。但是，高频率的“共现”并不一定代表强空间关系。鉴于上述问题，来自清华大学的研究团队[9]通过完全空间随机性测试来衡量对象之间空间关联的强度，并基于能够准确表示离散布局模式的样本提取复杂先验。该方法通过将输入对象划分为不相交的组，然后基于豪斯多夫度量进行布局优化，最终实现算法加速和置信度增强。

此外，无样例生成算法中更为常用的方法是对图形结构的编码。2019年，LI等[10]基于空间关系为房间中的每个家具构建树形的层次结构；对给定的室内场景，对象和对象组分别由叶节点和内部节点表示，将空间接近的物体视作相关的对象，并将对象间关系分类为支撑、环绕和共存三种。该方法首先将场景空间重构为一个具有各种对象关系和相应场景层次的物理模型。该研究以一个卧室为例，通过支撑关系将两对床头柜和台灯独立融合，然后通过环绕关系与床融合，如图3所示。利用上述结构，他们训练了一个变分递归自编码器，该编码器在编码阶段执行场景对象分组，在解码期间执行场景生成。同样地，WANG等[11]将面向对象和面向空间的范式结合在一起，提出了新的布局概念框架。该框架通过关系图表示场景的规划，将对象编码视为节点，将对象之间的空间–语义关系编码作为边。在规划阶段，采用深度图卷积生成模型对关系图进行综合；在实例化阶段，基于图像的卷积网络模块被用来指导搜索过程，以与关系图一致的方式将对象放置到场景中。

在有样例生成算法上，2020年LUO等[12]提出了一种三维场景布局网络，将变分自编码器与图卷积网络相结合，该网络能够基于相同输入的图结构生成不同的场景。在测试时，从学习分布中采样潜在编码，并与场景图一起发送给解码器以生成场景布局。在训练过程中，编码器将地面真实场景布局和场景图转换成一个分布，从中采样和解码潜在编码。而KESHAVARZI等[13]另辟蹊径，尝试将虚拟环境中的物体迁移到真实场景中，提出了一种基于场景图的生成式新框架。该框架允许从现有的场景中预测虚拟物体的位置和方向，从而实现基于环境感知的场景增强。ZHOU等[14]设计了一种神经信息传递方法，以预测给定场景中特定位置对象类型的概率分布。该方法将场景建模为图模型，其中节点表示场景中的现有对象，边表示对象之间的结构关系，通过图结构学习对象间的关系。



▲图3 基于图推理的场景生成方法示意图[10]

此外，以设计者输入的文字提示作为条件，可以生成具有特定语义信息的三维场景。斯坦福大学的研究团队[15]在该方向上进行了一系列的早期探索。他们首先将设计者提示的文本信息在符号化的语义空间进行解析，再通过解析出的语义信息生成相应的三维场景。这一工作包括收集带有自然语言描述的三维场景数据集，对场景生成文本进行语义解析，学习先验语义推断三维物体的隐含空间约束和通过深度相机学习人–物体间几何分布的概率模型等。他们早期的工作主要是将文本信息解析成三维对象和场景之间的符号化连接关系，并且通过数据集的收集获取不同场景类型中物体出现的统计数据作为隐性约束，从而生成可信的三维场景。在该团队的研究成果基础之上，更多的研究人员对语言条件驱动的三维场景生成方法进行了探索。例如，西蒙弗雷泽大学的研究团队提出了一种由语言驱动的三维场景建模方法[16]。该方法首先从三维场景数据库中通过用户的语言输入进行子场景检索，随后将检索到的子场景与当前环境合成新的三维场景。在每次用户编辑时，输入的语句都会转化为语义场景图，使用图对齐的方法从三维场景数据库中检索合适的子场景。检索到子场景后，根据输入文本和场景上下文，用附加对象对场景进行增强。然后，将增强后的子场景与当前场景进行语义对齐。最后，将增强的子场景拼接到当前场景中，合成一个新的场景。

除了利用设计者给出的条件进行三维场景生成，另一种方法是从人体的姿势、动作、移动进行场景生成。这种方法通过捕捉人体与环境物体交互过程的动作，估计场景内物体的位置、种类等内容，并联合推理。这种方式与自顶向下的设计逻辑相反，在给定大量的三维人体运动数据的基础上，生成与运动学信息相符的三维场景。为了采集人体的接触信息和运动信息，研究人员首先通过深度相机、惯性传感器或光学动作捕捉装置对三维人体运动数据进行获取。这些数据中蕴含着丰富的人与环境的交互信息，因此在场景生成上具有

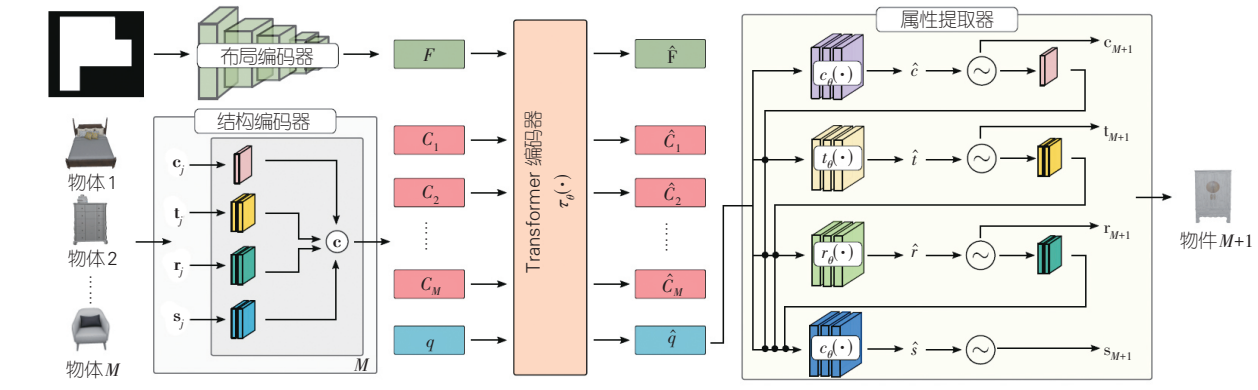
高度的可行性与合理性[17]。在获得了大量的用户姿势和动作信息之后，该方法在采集到的数据集上进行预测，估计被接触物体的种类，并使用空间、图来表示室内场景的空间属性，根据场景内物体的交互功能进行编码，形成马尔可夫随机场。随后，该方法从室内数据集场景中学习其数据分布，使用蒙特卡洛方法对马尔可夫链上的结点进行采样，形成三维场景。这种方法考虑到了物体的交互功能，因此在满足了视觉准确性的同时，在室内场景布局的功能性和自然性上具有优势。

1.4 基于自注意力模型的方法

自注意力机制允许模型在处理一个序列时考虑序列中每个元素与其他所有元素的关系。在三维场景生成中，使用自注意力模型可以更好的学习物体之间的关系，并将这些关系编码为排列表示。

PASCHALIDOU 等[18]将场景生成视为无序的对象集生成，提出了一个自动化生成三维室内场景的方法。该方法可以根据房间类型和形状，在不需要人为添加规则或关系图注释情况下，基于自注意力模型以自回归的方式生成具有逼真外观和3D一致性的家具布局，如图4所示。此外，该方法为用户提供对象约束，允许用户将任意类别和数目的家具固定在场景中特定位置。此后，PARA 等[19]针对该方法进行改进，提出了一种新的编码器–解码器架构的家具布局生成方式。该方法使用自回归布局生成器生成具有任意条件信息的布局，不仅允许仅输入对象的单个属性，而且可以对生成场景的细粒度进行控制，进一步增加了生成的逼真度和灵活度。在对象和场景的生成序列上，慕尼黑工业大学的WANG 等[20]提出了一个基于数据驱动的室内场景生成方法。该方法将场景生成问题转化为对象及其属性的序列生成问题，通过隐式学习对象关系，并使用自注意力模型解码器的交叉注意机制来构建条件模型，进而避免了生成结果对手工注释的依赖。

除此之外，该技术还可用于条件场景生成，通过人体在场景中的运动进行场景合成。自注意力机制模型输入可以仅依赖于场景中的三维人体姿态轨迹[21]。该方法所述的系统包括两个模块：接触预测模块和场景合成模块。接触预测模块利用现有的人–物体交互数据集来学习从人体到接触对象语义标签的映射，通过结合时域的上下文信息来增强标签预测在时间上的一致性。在生成估计的语义接触点后，场景合成模块首先根据语义可供性和物理可供性搜索适合接触点的对象，然后借鉴人体的运动推测出交互物体，用与人类没有接触的其他物体填充场景。2023 年，YI 等[22]提出了基于“人体运动”作为输入的三维室内场景生成方法。该方法基于人类与环境的交互，包含了场景对象的位置信息这一特性来推



（a）生成方法流程图

（b）3个不同平面图生成的场景结果

▲图4 基于自注意力模型的场景生成方法示意图[18]

断室内场景，并在此基础上使用自回归的自注意力模型结构，在给定人类运动序列的情况下，预测与人类接触的家具，从而生成合理的室内场景。

1.5 基于扩散模型的方法

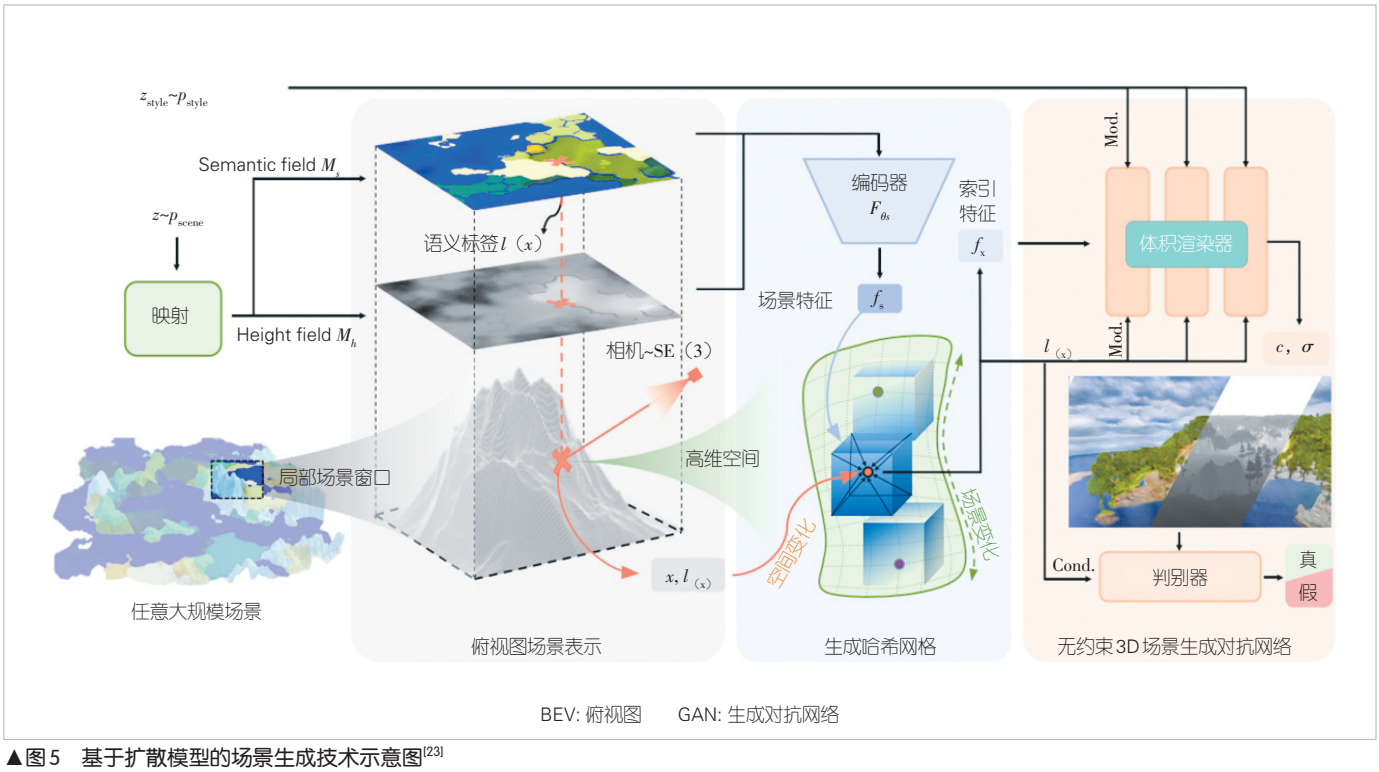
基于扩散模型的方法是一种利用随机微分方程来平滑地扰乱数据分布，将原始数据分布转化到已知的先验分布，然后通过学习逆向随机微分方程来从先验分布生成新的数据的方法。

基于扩散模型的场景生成方法主要通过随机噪声合成大规模的三维场景。最近出现的SceneDreamer就是一个典型的从随机噪声中生成三维场景的模型[23]，能够合成大规模的三维场景，并获得逼真的渲染效果。如图5所示，SceneDreamer提出了高效而富有表现力的鸟瞰视图场景表示方法、新颖的生成式神经哈希网格和基于风格的体积渲染器；给定从哈希网格中采样的潜在特征，渲染器即可通过风格调制的体积渲染来将其生成为逼真的三维视图。在此基础之上，CityDreamer模型[24]则可用于生成无限组合的三维城市场景。该模型将建筑实例和其他背景对象的生成分开，以处理生成建筑的多样性，并在此基础上构建了OSM数据集和Google Earth数据集，以提高生成的三维城市在其布局和外观的真实性。CityDreamer生成城市场景的方法可以分为三步：首先使用无边界城市布局生成技术生成城市场景，随后生成城市背景和建筑实例，最后将城市布局与建筑示例进行图像融合。

在风格化条件场景生成方面，设计者通过一个现有的场景和一个风格文本提示生成一个与文本相符、几何形状和纹理一致的新三维场景。此类方法首先生成三维场景的纹理，然后对网格纹理和几何图形进行联合优化，从网格中心开始，使用扩散模型更新未改变的区域，以确保生成的纹理与场景风格一致。

2023年，香港科技大学的研究团队将场景的位置布局和外观生成两阶段进行了拆分[25]。在位置布局阶段，采用了基于文本的条件扩散模型。该方法允许用户对生成的场景进行灵活的编辑，以生成高可信度、高纹理的优质三维场景。

在产业界方面，苹果公司[26]通过将真实三维场景的序列图像映射到一个完全分离的辐射场中进行潜在编码，之后通过大量的可变换视图在这些潜在编码上学习生成模型。该生成模型在训练时可引入图片、文本提示等不同的条件变量，以生成和这些条件变量一致的辐射场。Meta公司提出了一种基于语义信息的一致性风格室内场景生成方法[27]。该方法主要训练一个自回归模型，在推理过程中，将场景内已生成的物体作为条件，在每一步输出一个包括新物体及其位置信息的预测，以此实现风格一致且类别多样的场景。



BEV: 俯视图　GAN: 生成对抗网络

▲图5　基于扩散模型的场景生成技术示意图[23]

2 混合现实环境中的三维场景生成方法

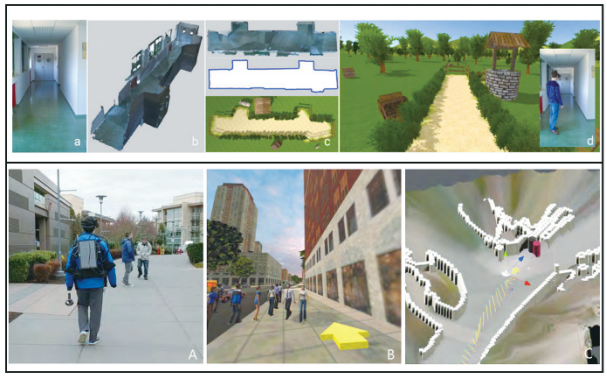
混合现实环境中，用户所处的实时环境与虚拟环境往往处于相互对应、相互映射的关系条件下。混合现实环境中的三维场景生成技术需要考虑真实环境的实时变化，通过环境感知将提取的真实环境信息用于虚拟场景的生成。这种情况下，往往需要通过虚实融合的手段，将真实场景中的物体映射为虚拟物体，并实现虚实配准。当前，完成这一任务有两种主要途径（如图6所示），一种是通过人在环对物体进行选择，实现物体间的映射；另一种方法是通过环境识别，自动将环境内物体进行映射。

2.1 基于人在环的场景生成方法

人在环场景生成方法采用半自动的方式，通过设计者在三维环境中实时编辑和操作虚拟物体，生成具有高合理性和高质量的三维场景。

早期的研究工作主要通过人为定义的配对方式，将真实物体替换为虚拟对象。例如，通过设计者的交互选择实现场景内的物体替换：系统向用户呈现一系列可以选择的物体，随后用户通过三维控制器点选所需的虚拟物体，并将虚拟物体和真实对象进行匹配，进而实现人在环的场景生成。

（a）人在环场景生成方法示意图[28]



（b）自动场景生成方法示意图[29]

▲图6 混合现实环境中的两种三维场景生成技术

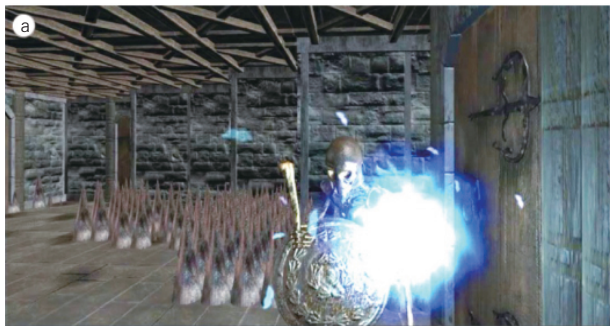
2021年，普渡大学的研究团队提出了一种端到端的系统设计框架：VRFromX[28]。该框架允许用户从真实世界的扫描中创作交互式三维场景。首先，用户使用手持式3D扫描仪扫描真实世界场景，形成点云图像；然后通过人工智能辅助在模型库中检索，并基于用户的交互选择，将点云对象替换为相应的虚拟模型；随后定义虚拟对象的功能和虚拟对象间的逻辑联系。该方法所设计的虚拟场景可以允许用户在场景中进行培训、交互等任务。

2.2 自动场景生成方法

人在环的场景生成方法在场景生成方面依赖于设计者的多次参与，生成效率比较低。为此，研究人员们希望探索一种基于环境感知的自动场景生成方法，实现从真实环境到虚拟环境的自动映射。

2017年，麻省理工学院的团队提出了一种以真实环境为输入，自动生成高沉浸感虚拟场景的系统[29]。该系统首先捕捉室内三维场景，检测家具和墙壁等障碍物，并绘制可行走区域地图。随后，系统将检测到的物体与虚拟对象配对，通过真实的物体向用户提供触觉反馈。该方法允许用户在任意大小和形状的室内空间中自动生成虚拟世界。在此基础上，他们又进行了系统优化[30]，允许用户通过自身虚拟代理与现实世界的物体进行交互，从而获得完整的触觉反馈，并将这些触觉反馈融入到三维场景的语义生成中。

此外，谷歌公司提出了一种通过真实环境生成虚拟环境的方法[31]。该方法将三维场景生成问题转化为一个同时满足多个约束的优化问题，约束条件包括场景几何信息、场景内物体、语义信息和物理约束。该系统首先将物体逐个生成的流程看作是在一个马尔可夫链上进行蒙特卡罗采样，然后将采样到的物体放置在一个二维俯视图上，以推断虚拟世界的环境布局。然后，使用满足几何、语义、物理等条件约束的多种物体和角色对模型进行填充。最后，将这些约束联合并使用一组离散变换，基于半定松弛的最新技术进行全局优化。微软公司则提出了一个可以在室外移动使用的虚拟现实系统[32]，如图7所示。该系统允许用户在真实世界中自由移动，并通过头戴显示设备完全沉浸在大型虚拟环境中。该系统首先预设了一个虚拟的环境，并通过视觉引导用户在虚拟世界中行走。在用户行走的同时，系统的跟踪器融合了GPS定位、光学跟踪和深度摄像机的画面，在现实环境中定位用



（a）实时动态生成虚拟现实场景

（b）对应用户真实环境

▲图7 自动场景映射生成技术叙事对应示意图[32]

户位置，并将用户重定向到目的地。同时，将场景实时感知到的用户行进路径上的障碍物映射为虚拟对象，并显示到用户头戴显示中。此后，微软公司的另一项自动生成虚拟场景的工作[33]以深度相机进行环境的识别和分割，实时检测并提取可行走区域或障碍物，并将可行走区域映射为实例化的预创建虚拟房间，障碍物则映射为阻挡用户前进的物体，进而提供长时间、高沉浸感的用户交互体验。

3 结论与展望

目前，虽然已经使用包括变分自编码器、生成对抗网络、自回归神经网络、图模型以及扩散模型等手段，面向混合现实的三维场景生成技术仍然需要较多的人工参与以及人为标注工作。以真实场景为输入，自动映射为风格化的虚拟场景在混合现实环境中仍然是一大挑战。

未来的研究工作可以专注于以下4个方面：1）专用数据集的构建问题：基于人工智能的生成方法虽然降低了场景生成的难度，增加了场景的丰富度，但目前仍然缺乏足够的数据集。而且目前较多的数据集是基于室内场景构建的，缺乏室外场景数据集，尤其是具有细节的室外物体数据集。2）场景物体多样性问题：生成方法大多基于已有虚拟物体模型生成布局，导致场景中的虚拟物体相对单一。未来可以利用人工智能技术直接生成差异化的虚拟物体。3）用户参与性问题：后续算法需要进一步简化用户的控制难度，同时提升用户对生成场景细节的精准控制，最终在两者间达到更好的平衡。4）环境交互问题：虽然已有方法考虑了部分环境的交互性，但大多局限于障碍物的识别。未来应更多地考虑用户与真实环境中不同物体的交互，将更多的物体功能融合到混合现实环境中。

参考文献

[1] RITCHIE D, WANG K, LIN Y. Fast and flexible indoor scene synthesis via deep convolutional generative models [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 6175 – 6183. DOI: 10.1109/CVPR.2019.00634

[2] WANG K, SAVVA M, CHANG A X, et al. Deep convolutional priors for indoor scene synthesis [J]. ACM transactions on graphics, 2018, 37(4): 1 – 14. DOI: 10.1145/3197517.3201362

[3] KAR A, PRAKASH A, LIU M Y, et al. Meta-sim: learning to generate synthetic datasets [C]//Proc. IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 4550 – 4559. DOI: 10.1109/ICCV.2019.00465

[4] DEVARANJAN J, KAR A, FIDLER S. Meta-Sim2: unsupervised learning of scene structure for synthetic data generation [C]// Computer Vision ECCV 2020. Springer, 2020: 715 – 733. DOI: 10.1007/978-3-030-58520-4_42

[5] PURKAIT P, ZACH C, REID I. SG-VAE: scene grammar variational autoencoder to generate new indoor scenes [C]//Computer Vision ECCV 2020. Springer, 2020: 155 – 171. DOI: 10.1007/978-3-030-58586-0_10

[6] MÜLLER P, WONKA P, HAEGLER S, et al. Procedural modeling of buildings [M]//ACM SIGGRAPH 2006 Papers. 2006: 614-623

[7] PARISH Y I H, MÜLLER P. Procedural modeling of cities [C]//Proc. 28th annual conference on Computer graphics and interactive techniques. ACM, 2001: 301 – 308. DOI: 10.1145/383259.383292

[8] ZHANG Z W, YANG Z P, MA C Y, et al. Deep generative modeling for scene synthesis via hybrid representations [J]. ACM transactions on graphics, 2020, 39(2): 1 – 21. DOI: 10.1145/3381866

[9] ZHANG S H, ZHANG S K, XIE W Y, et al. Fast 3D indoor scene synthesis with discrete and exact layout pattern extraction [EB/OL].(2020-02-05)[2024-03-15]. http://arxiv.org/abs/2002.00328

[10] LI M Y, PATIL A G, XU K, et al. GRAINS: generative recursive autoencoders for indoor scenes [J]. ACM transactions on graphics, 2019, 38(2): 1 – 16. DOI: 10.1145/3303766

[11] WANG K, LIN Y A, WEISSMANN B, et al. Planit: planning and instantiating indoor scenes with relation graph and spatial prior networks [J]. ACM transactions on graphics, 2019, 38(4): 1 – 15. DOI: 10.1145/3306346.3322941

[12] LUO A, ZHANG Z T, WU J J, et al. End-to-end optimization of scene layout [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 3753 – 3762. DOI: 10.1109/CVPR42600.2020.00381

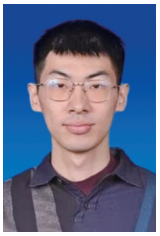
[13] KESHAVARZI M, PARIKH A, ZHAI X Y, et al. SceneGen: generative contextual scene augmentation using scene graph priors [EB/OL]. (2020-09-30)[2024-03-15]. http://arxiv.org/abs/2009.12395

[14] ZHOU Y, WHILE Z, KALOGERAKIS E. SceneGraphNet: neural message passing for 3D indoor scene augmentation [C]//Proc.

IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 7383 – 7391. DOI: 10.1109/ICCV.2019.00748
[15] CHANG A, SAVVA M, MANNING C D. Learning spatial knowledge for text to 3D scene generation [C]//Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2014: 2028 – 2038. DOI: 10.3115/v1/d14–1217
[16] MA R, PATIL A G, FISHER M, et al. Language–driven synthesis of 3D scenes from scene databases [J]. ACM transactions on graphics, 2018, 37(6): 1 – 16. DOI: 10.1145/3272127.3275035
[17] QI S Y, ZHU Y X, HUANG S Y, et al. Human–centric indoor scene synthesis using stochastic grammar [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 5899 – 5908. DOI: 10.1109/CVPR.2018.00618
[18] PASCHALIDOU D, KAR A, SHUGRINA M, et al. ATISS: autoregressive transformers for indoor scene synthesis [EB/OL]. (2021–10–07)[2024–03–20]. http://arxiv.org/abs/2110.03675
[19] PARA W R, GUERRERO P, MITRA N, et al. COFS: controllable furniture layout synthesis [C]//Proc. Special Interest Group on Computer Graphics and Interactive Techniques Conference. ACM, 2023: 1 – 11. DOI: 10.1145/3588432.3591561
[20] WANG X P, YESHWANTH C, NIEßNER M. SceneFormer: indoor scene generation with transformers [C]//Proc. International Conference on 3D Vision (3DV). IEEE, 2021: 106 – 115. DOI: 10.1109/3DV53792.2021.00021
[21] YE S F, WANG Y X, LI J M, et al. Scene synthesis from human motion [C]//Proc. SIGGRAPH Asia 2022 Conference. ACM, 2022: 1 – 9. DOI: 10.1145/3550469.3555426
[22] YI H W, HUANG C H P, TRIPATHI S, et al. MIME: human–aware 3D scene generation [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 12965 – 12976. DOI: 10.1109/CVPR52729.2023.01246
[23] CHEN Z X, WANG G C, LIU Z W. SceneDreamer: unbounded 3D scene generation from 2D image collections [EB/OL]. (2023–12–07)[2024–03–20]. http://arxiv.org/abs/2302.01330
[24] XIE H Z, CHEN Z X, HONG F Z, et al. CityDreamer: compositional generative model of unbounded 3D cities [EB/OL]. (2023–09–01)[2024–03–20]. https://arxiv.org/abs/2309.00610
[25] FANG C, DONG Y, LUO K M, et al. Ctrl–room: controllable text–to–3D room meshes generation with layout constraints [EB/OL]. (2023–10–05)[2024–03–20]. http://arxiv.org/abs/2310.03602
[26] BAUTISTA M A, GUO P S, ABNAR S, et al. Gaudi: a neural architect for immersive 3D scene generation [EB/OL]. (2023–07–27)[2024–03–20]. https://arxiv.org/abs/2207.13751
[27] XIONG W H, OĞUZ B, GUPTA A, et al. Simple local attentions remain competitive for long–context tasks [EB/OL]. (2021–12–14)[2024–03–20]. https://arxiv.org/abs/2112.07210
[28] IPSITA A, LI H, DUAN R L, et al. VRFromX: from scanned reality to interactive virtual experience with human–in–the–loop [C]// Proc. Extended Abstracts of 2021 CHI Conference on Human Factors in Computing Systems. ACM, 2021: 1 – 7. DOI: 10.1145/3411763.3451747
[29] SRA M, GARRIDO–JURADO S, MAES P. Oasis: procedurally generated social virtual spaces from 3D scanned real spaces [J]. IEEE transactions on visualization and computer graphics, 2018, 24(12): 3174 – 3187. DOI: 10.1109/TVCG.2017.2762691
[30] SRA M, GARRIDO–JURADO S, SCHMANDT C, et al. Procedurally generated virtual reality from 3D reconstructed physical space [C]//Proc. 22nd ACM Conference on Virtual Reality Software and Technology. ACM, 2016: 191 – 200. DOI: 10.1145/2993369.2993372
[31] SHAPIRA L, FREEDMAN D. Reality skins: creating immersive and tactile virtual environments [C]//Proc. IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, 2016: 115 – 124. DOI: 10.1109/ISMAR.2016.23
[32] CHENG L P, OFEK E, HOLZ C, et al. VRoamer: generating on–the–fly VR experiences while walking inside large, unknown real–world building environments [C]//Proc. IEEE Conference on Virtual Reality and 3D User Interfaces (VR). IEEE, 2019: 359 – 366. DOI: 10.1109/VR.2019.8798074
[33] YANG J J, HOLZ C, OFEK E, et al. DreamWalker: Substituting real–world walking experiences with a virtual reality [C]//Proc. 32nd Annual ACM Symposium on User Interface Software and Technology. ACM, 2019: 1093 – 1107. DOI: 10.1145/3332165.3347875

作 者 简 介

江海燕，北京理工大学在读博士研究生；研究方向为混合现实、人机交互与人工智能；曾获第七届中国国际“互联网+”大学生创新创业大赛银奖；发表论文20篇，申请专利18项（已授权7项）。



东野啸诺，北京理工大学在读博士研究生；主要研究方向为虚拟现实、具身智能、多模态人机交互等。

王涌天，北京理工大学教授、博士生导师，北京市混合现实与新型显示工程技术研究中心主任，“长江学者”、国家杰出青年科学基金获得者；长期在技术光学、虚拟现实和增强现实领域从事教学和科研工作，主要研究方向为光学系统设计和CAD、新型三维显示、虚拟现实和增强现实、医学图像处理等；获得国家技术发明奖和国家科技进步奖各1项，省部级和国家一级学会/协会科技奖励10余项；出版专著4部，发表论文320余篇，授权发明专利200余项，主持制定虚拟现实和增强现实领域首批国家标准6项。