第 11 卷 第 1 期
2022 年 1 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 11 No. 1
Jan. 2022

引文格式：
杨淼, 陈宝权. 室内场景生成算法综述 [J]. 集成技术, 2022, 11(1): 40-51.
Yang M, Chen BQ. A survey of indoor scene generation algorithms [J]. Journal of Integration Technology, 2022, 11(1): 40-51.

室内场景生成算法综述

杨 淼[1] 陈宝权[2*]

[1]（山东大学计算机科学与技术学院 青岛 266237）
[2]（北京大学前沿计算研究中心 北京 100871）

摘 要 室内场景生成任务是近年来热门的研究课题。它不仅能为计算机视觉任务提供天然带有标注的室内场景数据集，帮助其更好地理解场景，还能应用到诸多现实场景中，如机器人导航等。室内场景布局的多样性使得场景生成成为一项非常具有挑战性的任务。该文梳理了近年来在室内场景生成算法领域中的研究进展，从场景输入、场景上下文关系、场景表达方式、场景生成方式以及家具摆放顺序对生成算法进行总结分类，并以无样例的基于物体关系的生成方式、无样例的基于人类活动的生成方式以及基于样例和物体关系的生成方式 3 个分支对室内场景生成算法的发展以及优缺点进行分析。此外，该文还总结了现有算法的不足，并指出了室内场景生成算法未来可以尝试的方向。

关键词 室内场景生成；场景理解；三维模型生成；布局生成
中图分类号 TP 391.7 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20210928001

A Survey of Indoor Scene Generation Algorithms

YANG Miao[1] CHEN Baoquan[2*]

[1](*Computer Science and Technology College of Shandong University*, *Qingdao* 266237, *China*)
[2](*Center on Frontiers of Computing Studies of Peking University*, *Beijing* 100871, *China*)
[*]*Corresponding Author*: *baoquan@pku.edu.cn*

Abstract The indoor scene generation task is an important research topic in recent years. It can not only provide a natural annotated indoor scene dataset for computer vision tasks to help better understand the scene, but also can be applied to many real scenes such as robot navigation. The diversity of indoor scene layouts makes scene generation a very challenging task. This paper reviews the recent research progress in the field of indoor scene generation, summarizes and classifies the generation algorithms in terms of scene input, scene generation method, scene representation, scene generation order, and scene context relationship. The three categories of the generation algorithms including sample-free generation method based on object relationship, sample-free generation method based on human activities, and sample-based object relationship

收稿日期：2021-09-28 **修回日期**：2021-11-19
作者简介：杨淼，硕士研究生，研究方向为计算机图形学；陈宝权（通讯作者），教授，研究方向为计算机图形学，E-mail：baoquan@pku.edu.cn。

based on object relationship are analyzed with advantages and disadvantages. In addition, this article also summarizes the limitations of the existing algorithms and points out the direction that can be explored in the field of indoor scene generation in the future.

Keywords　scene generation; scene understanding; shape assembly; layout arrangement

1　引　言

近年来，虚拟室内场景被广泛应用于虚拟现实、增强现实、开放式游戏以及机器人领域，然而，室内场景的设计需要耗费大量时间且室内场景建模也需要复杂的场景设计工具。因此，自动化室内场景生成任务成为了研究人员关注的热点，并得到了快速发展。

室内场景生成任务，其目的就是将家具摆放在具有固定大小、结构的三维空间中，并满足现实室内场景中的功能约束及物理约束。其中，家具在三维空间中的属性信息可由位置、朝向、类别来描述。室内场景生成的本质是确定摆放的家具类别以及家具排列在三维空间中的方式。与室内场景生成任务类似，三维模型生成过程是确定模型部件的排列方式[1-4]，户型图生成则是将房间进行排列[5-7]，所以户型图、三维模型以及室内场景生成任务的解决方案可以互相借鉴或组合使用。三维模型由于其固定的功能特性，使得模型部件之间具有相对固定的布局规律，不具有太多的变化。与三维模型生成相比，室内场景生成有以下 3 个难点：

(1) 室内场景中家具的摆放方式有很大的自由度，同样的家具可能有多种合理的布局，所以家具之间的关系相对更加复杂。

(2) 三维模型有固定的计算机可以理解的表达方式，如点云、体素、网格等。如何将室内场景抽象为计算机可以理解的场景表达方式是一项挑战。

(3) 室内场景生成还需要考虑更多特殊的约束，如过道连通性、视野开阔性等。

主流的室内场景生成算法，使用物体之间的上下文关系对场景布局进行结构化，少量基于人类活动与物体上下文关系来表达场景布局的室内场景生成算法组成了新的分支。无参考信息的室内场景生成任务是非常具有挑战性的，由此，出现了一系列基于样例的场景生成算法。本文从上述 3 个难点出发，将场景生成任务按照无样例的基于物体关系的生成方式、无样例的基于人类活动的生成方式以及基于样例和物体关系的生成方式 3 个分支，对场景生成算法进行描述和分析。图 1 展示了整个场景生成过程中所涉及的算法框

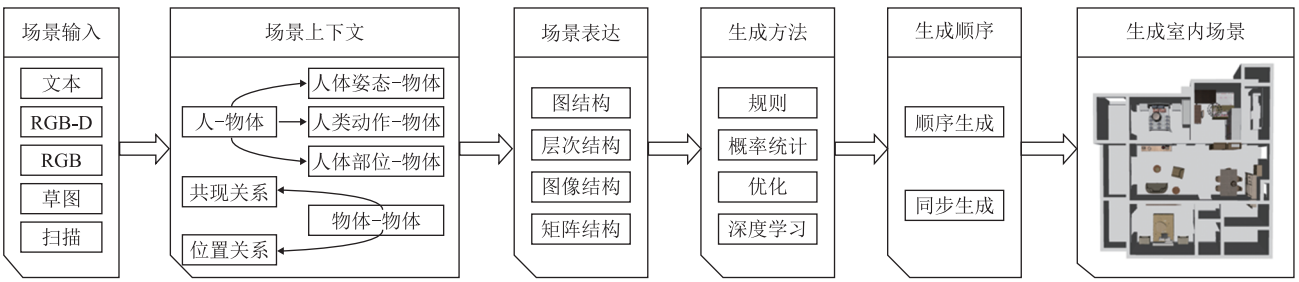

图 1　场景生成算法框架

Fig. 1　The framework of scene generation algorithms

架，其中，室内场景效果图来自 3D-FRONT 数据集[8]。

2 室内场景数据集

基于室内场景图像的图像检测、图像分割、本征分解等计算机视觉相关的任务已经得到了深入研究，这些研究致力于使计算机能够像人类一样更好地理解室内场景。然而，在视觉领域中对应不同任务的标注数据是非常难获得的，因此，该领域迫切需要天然带有标注的室内场景数据集来模拟真实室内场景，从而减轻标注数据的负担。

最早出现的室内场景数据集是由 Handa 等[9]提出的 SceneNet 数据集，该数据集仅提供少量带有三维模型的室内场景。Song 等[10]提出了被广泛使用的大规模室内场景数据集 SUNCG，然而，该室内场景是由业余设计师设计的，所以与真实场景存在一定的差距。与三维模型组合而来的合成室内场景数据集[9-10]不同，Dai 等[11]提出了一个基于真实场景扫描且包含丰富注释的 RGB-D 扫描图像数据集 ScanNet。Li 等[12]提出的 InteriorNet 数据集使用了更多高质量的计算机辅助设计(Computer Aided Design，CAD)模型，并请专业设计师据此设计了近 2 000 万个室内场景，同时渲染出了更接近照片效果的室内场景图像，只是其并不公开对应的三维模型，仅提供图像供研究使用。不同于前述的数据集(都不包含对室内场景结构的真实标注)，Mo 等[13]提出的 StructureNet 数据集提供了由专业设计师设计的带有场景结构信息标注的室内场景，可以为诸如房间结构预测等任务提供更可靠的标注数据。之后，Fu 等[8]提供了 3D-FRONT 室内场景数据集，该数据集是由家装领域用户使用的真实室内场景数据构成的，其中有将近一半的房间场景被设计师认作具有一定设计理念的高质量场景。由 Roberts 等[14]提出的 Hypersim 数据集在提供三维模型的同时，也提供了带有实例和语义分割标注的渲染图像，以及图像解耦之后的表示图像，是目前标注信息最完善的室内场景数据集。

3 室内场景生成算法的分类

室内场景生成算法发展至今已有许多出色的研究成果。本文从已有算法中，抽象出了 5 个分类标准，并从不同角度对现有的算法进行归纳总结，分析比较现有算法的优劣，用于帮助读者更好地理解室内场景生成算法的发展现状，具体如表 1 室内场景生成算法分类所示。

3.1 基于场景输入的分类

根据场景输入是否存在参考样例，可以将室内场景生成算法分为无样例的场景生成算法和基于样例的场景生成算法。无样例的场景生成算法往往是从大规模室内场景数据集中总结规则[15]、抽象能量函数[16-17]，或者将布局规律融入概率统计[18-19]、深度学习先验[20-22]，进而从无到有地生成合理的室内场景。基于样例的场景生成算法有文本[23]、草图[24]、图像[25]、三维信息[26]等输入形式，要求场景的生成结果与输入一定程度上匹配，属于有条件的场景生成任务。

在场景生成的实际应用中，往往需要加入用户的喜好，所以基于样例的生成算法能够更好地与人交互，更具有应用前景。但是，在需要大量多样的虚拟室内场景时，无样例的场景生成算法更具优势。

3.2 基于场景上下文关系的分类

根据场景上下文信息建模的方式不同，可将室内场景生成算法分为基于物体与物体之间的关系[17,21]和基于人类与物体之间的关系。大多数室内场景生成算法考虑的都是物体与物体之间的关系，这种关系可用来确定家具摆放的空间位置关

表 1　室内场景生成算法分类

Table 1　The classification of scene generation algorithms

	时间	场景上下文	场景表达	生成顺序	场景输入	生成算法
Xu 等[15]	2002	O	G	Seq	N	C
Yu 等[16]	2011	O	G	Syn	N	C
Merrell 等[17]	2011	O	G	Syn	N	C
Fisher 等[18]	2012	O	G	Seq	N	C
Kermani 等[31]	2016	O	G	Seq	N	C
Liu 等[32]	2014	O	H	Seq	N	C
Henderson 等[33]	2017	O	G	Seq	N	C
Wang 等[20]	2018	O	I	Seq	N	L
Ritchie 等[34]	2019	O	I	Seq	N	L
Li 等[21]	2019	O	H	Seq	N	L
Zhang 等[22]	2020	O	M	Syn	N	L
Yang 等[35]	2021	O	M	Syn	N	L
Zhou 等[29]	2019	O	G	Syn	N	L
Wang 等[36]	2019	O	G/I	Seq	N	L
Luo 等[30]	2020	O	G	Syn	N	L
Dhamo 等[37]	2021	O	G	Syn	N	L
Wang 等[38]	2020	O	M	Seq	N	L
Jiang 等[27]	2012	P	G	Seq	N	C
Fisher 等[26]	2015	P	G	Seq	N	C
Ma 等[19]	2016	P	G	Seq	N	C
Savva 等[28]	2016	P	G	Seq	N	C
Qi 等[40]	2018	P	H	Seq	N	C
Fu 等[41]	2017	P	G	Seq	N	C
Seversky 等[42]	2006	O	G	Seq	T	C
Coyne 等[43]	2001	O	G	Seq	T	C
Chang 等[44-45]	2014	O	G	Seq	T	C
Ma 等[23]	2018	O	G	Seq	T	C
Shin 等[48]	2007	O	G	Seq	S	C
Xu 等[24]	2013	O	G	Seq	S	C
Nie 等[25]	2020	O	I	Syn	I	L
Huang 等[51]	2018	O	I	Syn	I	C
Xiao 等[52]	2021	O	I/G	Syn	I	L
Zhang 等[53]	2021	O	I/G	Syn	I	L
Chen 等[54]	2014	O	G	Seq	D	C
Avetisyan 等[56]	2020	O	G	Syn	D	L
Hampali 等[55]	2021	O	H	Syn	D	C

注：场景上下文关系中，O 代表物体-物体关系，P 代表人-物体关系；场景表达中，G 代表图结构，H 代表层次结构，I 代表图像结构，M 代表矩阵结构；生成顺序中，Seq 代表顺序生成，Syn 代表同步生成；场景输入中，N 代表无参考，T 代表文本输入，S 代表草图输入，I 代表图像输入，D 代表三维信息输入；生成算法中，C 代表传统算法，L 代表深度学习算法。

系和判断家具类别的共现关系，还有少部分采用隐式的方式学习场景布局的上下文信息，如采用神经网络的注意力机制或利用 CNN、DNN 网络参数自动学习。考虑到物体的摆放与人类的活动息息相关，近年来出现了基于人类与物体之间的关系进行建模的算法，其主要包含人体姿态-物体[27]、人体动作-物体[19,26]以及人体部位-物体[28]3 种形式。

设计师在利用计算机软件进行室内场景设计的过程中，未将人可能在的区域或动作输入计算机中，所以缺少包含人类行为的虚拟室内场景数据集。另外，在不考虑复杂度的情况下，基于物体与物体之间的关系建模是最易实现的，未来的算法会更加倾向于基于这种关系的建模方式。然而，现有的算法仍避免不了人为定义如支撑、环绕等物体与物体之间的关系，采用注意力机制学习物体与物体之间的关系能更好地解决该问题。

3.3　基于场景表达方式的分类

室内场景生成算法根据其表达方式的不同主要分为图结构[29]、层次结构[21]、图像结构[20]和矩阵结构[22]。图结构是由节点集合和边集合构成的，其优点是灵活直观，可以在任意两个对象之间添加联系，所以大部分算法都采用图结构的方式进行表达。层次结构是由一组具有父子关系的节点组成，每一个子节点都只有一个父节点。一般将整个场景作为根节点，家具或家具部件作为叶子节点，节点之间存在指向性关系。按照层次递进的方式生成室内场景比较符合设计师设计场景时的设计思路。考虑到家具一般排列在二维空间中，有研究人员采用俯视图的方式来表达场景，图像结构的表达方式可以更加直观地将家具之间的位置关系体现到坐标上。基于矩阵的表达方式，首先将每个家具节点的属性用向量来表示，然后将所有家具的向量整合为一个可以代表整个场景布局信

息的矩阵，矩阵的表达形式最为简单，不需要定义物体之间的上下文关系，但是在生成过程中同样无法体现家具之间的关系，可解释性不强。

目前，深度学习是主要的场景生成算法，而图卷积神经网络的出现使得图结构表示的场景也能利用深度学习的优势，自动学习场景布局规律。同时，图结构本身能灵活增加、删除节点的优势也能使深度学习算法更好地与人交互。所以利用图卷积神经网络进行深度学习的图结构表达场景会更具优势。

3.4 基于场景生成方式的分类

根据场景生成方式的不同，可将室内场景生成算法主要分为传统算法和深度学习算法。在深度学习算法出现以前，为了确定布局参数，传统方法采用基于规则[15]、优化[16-17]以及概率统计[18-19]的传统方式对场景的布局规律进行建模。在大规模室内场景数据集出现以后，深度学习算法由于其强大的特征学习和提取能力成为室内场景布局算法的主要手段。传统算法需要耗费大量的人工成本对室内场景布局规律进行抽象，并且在生成过程中耗时相对较长。而深度学习算法则通过端到端的生成神经网络[20-22,29]，可以自动学习布局规律并且快速生成布局，成为目前主流的场景生成算法。

3.5 基于家具摆放顺序的分类

根据场景中家具摆放的顺序，可以将室内场景生成算法分为顺次迭代[15,21]和同步生成[29-30]两种方式。顺次迭代的缺点是后序摆放的物体由于顺序性无法对前序摆放的物体产生影响，而前序摆放的物体也无法预测后续摆放的物体类别，但其优点是如果发现放不下的物体可以舍弃，所以只要算法正确就一定能够生成合理的场景。同步生成的优点是所有家具在摆放时都可以考虑其他所有家具的摆放，缺点是摆放的过程中不能舍弃家具，可能会有不合理的场景出现。

4 室内场景生成算法

室内场景布局生成的目标是确定家具的大小、类别及摆放在三维空间中的位置和朝向。目前，主流的场景生成算法是在无样例约束的条件下，基于物体关系对场景建模后进行生成。本节对无样例的基于物体关系的场景生成算法，按照传统算法和深度学习算法两大类进行梳理，并概述了场景表达方式和场景生成顺序。在上述主流算法的基础上，还衍生出了基于人类活动的无样例场景建模的场景生成方式，以及基于物体关系和样例的场景生成方式。

本节按照无样例的基于物体关系的场景生成算法、无样例的基于人类活动的场景生成算法以及基于样例和物体关系的场景生成算法 3 大类对场景生成算法进行具体地梳理和分析，用于帮助读者更好地了解场景生成算法的发展趋势。

4.1 无样例的基于物体关系的场景生成算法

4.1.1 传统生成方法

早期用于研究自动室内场景生成的传统方法主要分为 3 种：基于规则的场景生成、基于能量优化的场景生成以及基于概率统计的场景生成。传统方法需要充分发挥科研人员的概括和抽象能力，利用有限的知识和能力抽象出可能的场景布局规律，并用算法来表示。

Xu 等[15]最早提出基于规则并以顺序迭代的方式将家具逐个摆放到室内场景中的场景生成算法。在摆放过程中，该算法根据可放置的平面、平面的支撑能力以及摆放物体间的距离来考虑家具可能摆放的位置，同时将现实场景中家具之间的语义相关性融入到规则中使得家具摆放更加合理。此外，该算法还增加了物体的非互穿性、稳定性及物体间的摩擦等物理约束，避免出现不和谐的场景摆放。

Yu[16]和 Merrell[17]等提出用以布局信息为变量的能量函数来表示规则，通过优化能量函数的

方式生成合理的布局。家具之间的语义相关性可以概括为层次关系、空间位置关系以及成对关系，其可作为场景上下文信息融入能量方程[16]。也可以考虑功能性摆放规则以及视觉性摆放规则，来约束家具的摆放位置[17]。两种算法从解决问题的不同角度分别提出了不同的优化能量函数方式。由于搜索空间过大，前述算法采用了模拟退火算法，逐步确定家具的摆放后得到一个合理的室内场景。后者则考虑到由于布局的多样性会使得场景有多种合理的布局，对应的能量函数也具有多峰值，所以该算法采用了马尔科夫蒙特卡洛采样的方式，对能量函数进行优化。

Fisher 等[18]提出一种基于高斯混合模型和贝叶斯模型的概率模型来学习场景的布局先验，并以此为基础生成与用户给定示例场景相似的新室内场景。该模型能够根据成对物体在空间位置中的共现概率，预测可摆放的家具类别以及场景空间中最有可能摆放该家具的位置。为了增加新场景布局的多样性，该论文还提出了一种基于场景上下文信息的聚类算法，并以此提供可互相替换的家具类别。除了考虑成对的对象关系以外，Kermani 等[31]还采用了涉及两个以上对象的关系表示场景的上下文。与前述只考虑局部家具之间的上下文关系不同，Liu 等[32]利用给定的大规模室内场景数据集构造了一个具有层次结构的语法概率模型，用其来概括整个室内场景的上下文关系。这种利用数据集学到的层次关系，将其作为一个整体进行摆放更能增加布局的合理性。Henderson 等[13]也是按照主要物体、小物体、天花板物体、墙面物体这样层次递进的顺序进行家具摆放。

4.1.2 深度学习生成方法

深度学习是一种基于数据进行表征学习的方法。随着大规模室内场景数据集的出现，采用深度学习的方式进行场景生成成为了可能。传统生成方法需要人工定义具体的规则、优化函数或密度函数，而深度学习可以直接利用具有特殊结构的神经网络来隐式地学习表达这些信息，避免人工定义的复杂性。除此之外，采用深度学习的方式从大规模室内场景数据集中学习到的场景先验能够很好地补充人有限的室内场景设计经验。

Wang[20]和 Ritchie[34]等提出一种基于卷积神经网络的场景生成模型，能够快速灵活地生成场景。虽然室内场景存在于三维空间，但重力决定了大多数对象在二维平面上进行布局，所以该模型将俯视图表示的场景作为模型输入，利用不同的卷积神经网络对家具的类别、位置、朝向以及大小进行预测，并以迭代的方式将家具逐个添加到场景中。将场景图表示为俯视图可以实现像素级别的细粒度的推理过程，并且可以利用专门为图像理解而开发的卷积神经网络学习复杂的室内场景结构。

与图像平铺表示的室内场景不同，Li 等[21]认为室内场景的结构本质上是分层的，提出将室内场景抽象为包含支撑关系、共现关系、环绕关系以及墙依靠关系的分层树结构。首先，循环神经网络根据抽象出的分层树结构，由叶子节点开始自下而上地对家具属性以及与其他家具的相对位置关系进行编码；然后，从根节点向下解码出每个家具节点的类别、大小、朝向等场景布局信息；最后，以变分自编码器的方式进行训练，由随机采样的噪声进行生成。

与分层树结构类似，Zhang 等[22]也采用了变分自编码器的生成网络结构，该结构将场景中所有物体特征排列成固定大小矩阵作为室内场景的表达方式，将参数化矩阵输入稀疏连接的前向神经网络中学习室内场景粗粒度的全局布局信息，同时利用有向距离场将场景映射到二维空间中学习细粒度的局部布局信息。Yang 等[35]也采用了矩阵的表达形式，除了包含单个物体信息的生成网络外，该研究还利用生成网络对物体的相对属性进行编码，最后利用贝叶斯方法结合单

独对象属性、相对属性以及参数先验对最终布局进行优化。

随着图卷积神经网络的出现和发展，图结构表达的室内场景可以利用深度学习的方式编码场景先验，且图结构是最直观表达物体上下文关系的场景表达方式。消息传递网络是实现图卷积网络常用的框架，Zhou 等[29]提出了一种利用消息传递网络对场景图表示的室内场景进行场景增强的方法，并利用从大量的室内场景数据集中抽象出的支撑、环绕、靠近、共现关系及消息传递的注意力机制，让模型更加集中于最相关的场景上下文对新物体进行预测。Wang 等[36]将场景生成拆分成两个步骤，首先，在决策模型的框架下，利用图卷积神经网络生成场景图表示场景布局规划；然后，利用卷积神经网络将图结构中每个节点代表的家具实例化到具体的空间位置中。Luo 等[30]将图卷积神经网络和条件变分自编码器结合起来，提出了一种端到端生成室内场景的算法。Dhamo 等[37]在图卷积神经网络和条件变分自编码器的基础上，通过在解码器之前加入增强后的与原始图像有差异的场景进行训练，能够根据人的偏好改变场景图之后，进行室内场景生成。

以往的场景生成方式会对家具之间的关系进行假设，Wang 等[38]提出将场景抽象成一个对象属性的序列，将场景生成任务转化为序列生成任务，采用 Transformer 结构[39]生成场景，并通过 Transformer 中的注意力机制隐式学习家具物体之间的关系。

4.2 无样例的基于人类活动的场景生成算法

真实场景中物体之间的关系复杂多样，很难从中提取出重要的场景上下文关系进行建模。现实中场景的布局往往是按照功能区划分，并且和人的活动密切相关，因此可以通过建模人类活动与物体之间的关系，对场景布局进行解析。

通过人体姿态可以预测接下来的动作倾向，而动作是人和物体之间产生交互的中介，所以人体姿态和物体之间存在一定的上下文关系。Jiang 等[27]通过构造概率密度函数来学习物体的支撑性、可达性以及易用性与人体 6 种标准姿态的上下文关系。在场景生成过程中，模型首先根据已有物体推断可能的人体姿态以及位置，然后再以人体姿态为中心，从概率密度函数中采样找到可能的下一个物体的摆放位置。

Fisher 等[26]直接利用动作建模人和物体之间的联系，并且以此为依据生成与给定粗糙扫描场景在功能和几何属性方面都类似的新场景。该研究从扫描场景中提取带有功能区的几何场景模版后，将虚拟人代理放在对应的功能区。虚拟人代理根据场景功能确定凝视、触摸、背部支撑以及臀部支撑等与物体交互的动作，将物体摆放成与动作相关的排列。

Ma 等[19]研究了一种由动作驱动的场景生成框架，该框架通过模拟被人类动作改变的物体放置确定场景布局。首先，利用标注图像对动作模型进行学习，其中，每种类型的动作结合了一个或多个人体姿势、一个或多个物体类别和表示人-物体以及物体-物体之间的空间位置关系信息；然后，通过采样动作序列生成场景。与其他算法不同的是，由于一个动作可能涉及多个人体姿势和物体，该框架能在确定动作后同时触发一系列物体的放置，而且场景中所有动作之间具有某种顺序关系，使得整个场景的生成更具一致性。

不同于其他研究关注固定的人体姿态，Savva 等[28]将人体姿态用动作属性控制，允许更高自由度的人体姿态与场景的交互。该研究建立能反映人体部位与人附近物体联系的人体姿态属性，从大规模数据集中构建概率模型，将人体姿态估计融入场景生成任务，以生成更合理的场景。Qi 等[40]提出用一种与空间属性相关的与或图来表示室内场景，在终端节点上将与人类活动相关的上下文关系编码进马尔可夫随机场，然后

以采样的方式生成新的场景。Fu 等[41]的研究则是在给定一个空场景以及部分家具类别的情况下，通过人类活动与物体之间的联系，在给定家具的基础上进行类别扩充，构造完整的场景功能区。

4.3　基于样例和物体关系的场景生成算法

自动化室内场景生成的目的是减轻设计布局所耗费的时间精力，然而，在某些应用上仍需要一定程度的与人交互，才能生成符合用户需求的室内场景布局。除此之外，完全无参考的室内场景生成任务需要学习所有可能的场景布局，较难实现，而基于样例的方式大大减少了布局多样性，使场景生成任务更加简单。本节将对文本输入、草图输入、图像输入以及三维信息输入几种不同的基于样例的场景生成算法进行介绍。

4.3.1 文本输入

利用自然语言描述来获取场景布局是一种较为简便的方式，自然语言作为人们日常表达思维的方式在描述场景时不需要任何培训。Seversky[42]和 Coyne[43]等较早地提出了语言驱动的场景生成方式，该方法通过自然语言详细描述对象及空间位置的关系，实现自然语言到场景的映射，这种方式可以生成符合语言表达的室内场景，但限制了用户表达场景的自由度和多样性，只能生成固定的场景布局。Chang 等[44-45]提出将自然语言解析成一个包含要摆放物体以及物体之间如何排列的场景模版后，根据数据集中学习到的物体之间的空间位置先验，将隐含的物体之间的位置关系扩充进来。Ma 等[23]提出的算法不仅考虑了物体之间的空间位置关系，还对物体成对出现的概率进行建模，因此，该方法不仅支持隐式位置关系扩充，也支持隐式的物体类别扩充，这种利用从数据集中提取出的隐式、通用的布局规律对场景增强的方式，可以让用户不必像以前的大多数方法一样提供明确的布局信息。Chang 等[46]尝试将描述内容与物体关联起来，以便找到更适合文本描述的三维模型，还将规则转化成基于文本的交互式场景编辑操作，并开发出给用户使用的 UI 界面[47]。

4.3.2 草图输入

草图对于用户而言也是一种表达场景布局的简易方式，建模师会根据室内设计师预先绘制的概念草图创建对应的三维室内场景。现有的三维场景设计工具需要建模师多次重复模型寻找以及模型摆放两个步骤将家具逐个摆放到室内场景中。在给定草图进行场景生成的自动化算法中，Shin 等[48]也采用了类似的流程，首先，从草图中识别出单个物体；然后，根据物体的视觉特征去模型库中找到对应的三维模型；最后，摆放到三维空间中。将三维模型拆分成部件，完成部件级别的模型寻找以及摆放，也能实现模型草图到三维模型的生成[49-50]，但是单个物体的检索和摆放往往会产生歧义，为此，Xu 等[24]提出对数据集中提取的具有共现关系和空间位置关系的家具组合进行提取，从而实现多个物体的协同检索和放置，该方法极大地减少了用户干预。

4.3.3 图像输入

手机是人们日常携带且具有摄像功能的设备，得到一张场景图像只需要按下拍照按钮，所以通过 RGB 图像生成场景也是用户可选的方案，并得到了广泛的研究和发展。Huang 等[51]提出利用能够表征场景功能和几何约束联合分布的整体场景语法来表示三维场景的结构，利用蒙特卡洛方法找到与真实场景最相似的由场景三维信息渲染得到的场景。Nie 等[25]将图像重建分为场景布局结构估计、物体检测及网格重建 3 个子任务，本质上还是检测要摆放的物体，根据相机姿态投影到三维空间中。该研究通过注意力机制对所有物体的卷积特征进行加权求和，将上下文信息融入到物体三维空间预估过程中。Xiao 等[52]采用了更复杂的图卷积神经网络，通过消息传递融入上下文信息。Zhang 等[53]结合了以上两种思

路，利用 Nie 等[25]提出的注意力机制得到初始位置后，再利用 Xiao 等[52]提出的图卷积神经网络进一步优化场景布局。

4.3.4 三维信息输入

用二维 RGB 图像进行三维场景生成较易出现偏差，而基于深度信息构建的 RGB-D 场景或扫描场景带来的三维场景信息更加明确，但是 RGB-D 图像往往充满噪声，所以 Chen 等[54]提出将在数据库中学习到的物体上下文关系用于约束重建，确保重建家具与扫描家具之间的语义相似性。Hampali 等[55]采用了蒙特卡洛方法，搜寻 RGB-D 图像中可能的家具集合信息，以最小化重建场景与真实场景的差异。Fisher 等[26]利用人-物体的上下文关系生成和具有噪声的扫描场景相似的新场景。Avetisyan 等[56]则是在检测出扫描场景中的物体后，利用物体-物体的上下文关系对家具的摆放位置进行优化。

5 总结与未来展望

本文对场景生成算法做了一个较为全面的分析和描述，分别从基于规则、概率统计、优化函数的传统场景生成方式到基于图卷积神经网络、深度神经网络、卷积神经网络的深度学习生成方法，从基于物体-物体之间上下文关系的算法到基于人-物体的上下文关系的算法，从无样例的生成模型到基于样例的生成模型，从矩阵结构、层次结构、图像结构到图结构的场景表达方式，从顺序生成到同步生成的生成顺序分析了各个场景生成算法的优缺点，以及近年来的发展。

目前，室内场景生成算法仍然存在问题。主流的深度学习方法虽然能学习到一定的场景先验知识，但仍然需要预定义一些空间位置关系和共现关系来辅助算法进行场景理解，且预定义的关系所能表达的场景上下文关系有限。神经网络中的注意力机制[25,29,38]可以很好地解决这个问题，但注意力机制仅能表示物体之间联系的紧密程度，不包含任何语义信息。所以将语义关系预测融入场景生成是未来研究的方向之一。

最直观且目前最有发展前景的场景表达方式是图结构和图像结构的表达方式。图结构表达方式可以忽略家具本身在三维空间中的位置，在任意节点之间构造联系，但该表达方式的节点之间不具备明确顺序。而图像结构由于其本身排列在二维空间坐标系中，所以图像表示的场景能自然地捕捉到家具之间的位置关系。因而将图结构和图像结构结合进行场景预估也是一个值得研究的课题。现有的算法[36]虽将图和图像相结合但是采用的是两步走的策略，未来可以尝试训练一个端到端的网络将两者结合起来。

参 考 文 献

[1] Mo K, Guerrero P, Yi L, et al. StructureNet: hierarchical graph networks for 3D shape generation [J]. arXiv Preprint, arXiv: 1908.00575, 2019.

[2] Wu R, Zhuang Y, Xu K, et al. PQ-NET: agenerative part Seq2Seq network for 3D shapes [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 829-838.

[3] Li J, Xu K, Chaudhuri S, et al. GRASS: generative recursive autoencoders for shape structures [J]. ACM Transactions on Graphics (TOG), 2017, 36(4): 1-14.

[4] Mo K, Wang H, Yan X, et al. PT2PC: learning to generate 3D point cloud shapes from part tree conditions [C] // European Conference on Computer Vision, 2020: 683-701.

[5] Merrell P, Schkufza E, Koltun V. Computer-generated residential building layouts [C]. International Conference on Computer Graphics and Interactive Techniques, 2010: 181.

[6] Ma C, Vining N, Lefebvre S, et al. Game level layout from design specification [J]. Computer

Graphics Forum: Journal of the European Association for Computer Graphics, 2014, 33(2): 95-104.

[7] Nauata N, Chang KH, Cheng CY, et al. House-GAN: relational generative adversarial networks for graph-constrained house layout generation [C] // European Conference on Computer Vision, 2020: 162-177.

[8] Fu H, Cai B, Gao L, et al. 3D-FRONT: 3D Furnished Rooms with layOuts and semaNTics [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10933-10942.

[9] Handa A, Pătrăucean V, Stent S, et al. SceneNet: an annotated model generator for indoor scene understanding [C] // 2016 IEEE International Conference on Robotics and Automation, 2016: 5737-5743.

[10] Song S, Yu F, Zeng A, et al. Semantic scene completion from a single depth image [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1746-1754.

[11] Dai A, Chang AX, Savva M, et al. ScanNet: richly-annotated 3D reconstructions of indoor scenes [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5828-5839.

[12] Li W, Saeedi S, McCormac J, et al. InteriorNet: mega-scale multi-sensor photo-realistic indoor scenes dataset [J]. arXiv Preprint, arXiv: 1809.00716, 2018.

[13] Mo K, Guerrero P, Yi L, et al. StructureNet: hierarchical graph networks for 3D shape generation [J]. arXiv Preprint, arXiv: 1908.00575, 2019.

[14] Roberts M, Ramapuram J, Ranjan A, et al. Hypersim: a photorealistic synthetic dataset for holistic indoor scene understanding [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10912-10922.

[15] Xu K, Stewart J, Fiume E. Constraint-based automatic placement for scene composition [C] // Graphics Interface, 2002, 25-34.

[16] Yu LF, Yeung SK, Tang CK, et al. Make it home: automatic optimization of furniture arrangement [J]. ACM Transactions on Graphics (TOG)-Proceedings of ACM SIGGRAPH 2011, 2011, 30(4).

[17] Merrell P, Schkufza E, Li Z, et al. Interactive furniture layout using interior design guidelines [J]. ACM Transactions on Graphics (TOG), 2011, 30(4): 1-10.

[18] Fisher M, Ritchie D, Savva M, et al. Example-based synthesis of 3D object arrangements [J]. ACM Transactions on Graphics (TOG), 2012, 31(6): 1-11.

[19] Ma R, Li H, Zou C, et al. Action-driven 3D indoor scene evolution [J]. ACM Transactions on Graphics, 2016, 35(6): 1-13.

[20] Wang K, Savva M, Chang AX, et al. Deep convolutional priors for indoor scene synthesis [J]. ACM Transactions on Graphics (TOG), 2018, 37(4): 1-14.

[21] Li M, Patil AG, Xu K, et al. GRAINS: generative recursive autoencoders for indoor scenes [J]. ACM Transactions on Graphics (TOG), 2019, 38(2): 1-16.

[22] Zhang Z, Yang Z, Ma C, et al. Deep generative modeling for scene synthesis via hybrid representations [J]. ACM Transactions on Graphics (TOG), 2020, 39(2): 1-21.

[23] Ma R, Patil AG, Fisher M, et al. Language-driven synthesis of 3D scenes from scene databases [J]. ACM Transactions on Graphics (TOG), 2018, 37(6): 1-16.

[24] Xu K, Chen K, Fu H, et al. Sketch2Scene: sketch-based co-retrieval and co-placement of 3D models [J]. ACM Transactions on Graphics (TOG), 2013, 32(4): 1-15.

[25] Nie Y, Han X, Guo S, et al. Total3DUnderstanding: joint layout, object pose and mesh reconstruction for indoor scenes from a single image [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 55-64.

[26] Fisher M, Savva M, Li Y, et al. Activity-centric scene synthesis for functional 3D scene modeling [J]. ACM Transactions on Graphics (TOG), 2015, 34(6): 1-13.

[27] Jiang Y, Lim M, Saxena A. Learning object arrangements in 3D scenes using human context [J]. arXiv Preprint, arXiv: 1206.6462, 2012.

[28] Savva M, Chang AX, Hanrahan P, et al. PiGraphs: learning interaction snapshots from observations [J]. ACM Transactions on Graphics (TOG), 2016, 35(4): 1-12.

[29] Zhou Y, While Z, Kalogerakis E. SceneGraphNet: neural message passing for 3D indoor scene augmentation [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 7384-7392.

[30] Luo A, Zhang Z, Wu J, et al. End-to-end optimization of scene layout [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 3754-3763.

[31] Kermani ZS, Liao Z, Tan P, et al. Learning 3D scene synthesis from annotated RGB-D images [J]. Computer Graphics Forum, 2016, 35(5): 197-206.

[32] Liu T, Chaudhuri S, Kim VG, et al. Creating consistent scene graphs using a probabilistic grammar [J]. ACM Transactions on Graphics (TOG), 2014, 33(6): 1-12.

[33] Henderson P, Subr K, Ferrari V. Automatic generation of constrained furniture layouts [J]. arXiv Preprint, arXiv: 1711.10939, 2017.

[34] Ritchie D, Wang K, Lin Y. Fast and flexible indoor scene synthesis via deep convolutional generative models [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 6182-6190.

[35] Yang H, Zhang Z, Yan S, et al. Scene synthesis via uncertainty-driven attribute Synchronization [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 5630-5640.

[36] 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 (TOG), 2019, 38(4): 1-15.

[37] Dhamo H, Manhardt F, Navab N, et al. Graph-to-3D: end-to-end generation and manipulation of 3D scenes using scene graphs [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 16352-16361.

[38] Wang X, Yeshwanth C, Nießner M. SceneFormer: indoor scene generation with transformers [J]. arXiv Preprint, arXiv: 2012.09793, 2020.

[39] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] // Advances in Neural Information Processing Systems, 2017: 5998-6008.

[40] Qi S, Zhu Y, Huang S, et al. Human-centric indoor scene synthesis using stochastic grammar [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5899-5908.

[41] Fu Q, Chen X, Wang X, et al. Adaptive synthesis of indoor scenes via activity-associated object relation graphs [J]. ACM Transactions on Graphics (TOG), 2017, 36(6): 1-13.

[42] Seversky LM, Yin L. Real-time automatic 3D scene generation from natural language voice and text descriptions [C] // Proceedings of the 14th ACM International Conference on Multimedia, 2006: 61-64.

[43] Coyne B, Sproat R. WordsEye: an automatic text-to-scene conversion system [C] // Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, 2001: 487-496.

[44] Chang A, Savva M, Manning CD. Interactive learning of spatial knowledge for text to 3D scene generation [C] // Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 2014: 14-21.

[45] Chang A, Savva M, Manning CD. Learning spatial knowledge for text to 3D scene generation [C] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014:

2028-2038.

[46] Chang A, Monroe W, Savva M, et al. Text to 3D, scene generation with rich lexical grounding [J]. arXiv Preprint, arXiv: 1505.06289, 2015.

[47] Chang AX, Eric M, Savva M, et al. SceneSeer: 3D scene design with natural language [J]. arXiv Preprint, arXiv: 1703.00050, 2017.

[48] Shin H J, Igarashi T. Magic canvas: interactive design of a 3-D scene prototype from freehand sketches [C] // Proceedings of Graphics Interface 2007, 2007: 63-70.

[49] Xie X, Xu K, Mitra NJ, et al. Sketch-to-design: context-based part assembly [J]. Computer Graphics Forum, 2013, 32(8): 233-245.

[50] Lee J, Funkhouser TA. Sketch-based search and composition of 3D models [C] // EUROGRAPHICS Workshop on Sketch-Based Interfaces and Modeling, 2008: 97-104.

[51] Huang S, Qi S, Zhu Y, et al. Holistic 3D scene parsing and reconstruction from a single RGB image [C] // Proceedings of the European Conference on Computer Vision, 2018: 187-203.

[52] Xiao J, Wang R, Chen X. Holistic Pose Graph: modeling geometric structure among objects in a scene using graph inference for 3D object prediction [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 12717-12726.

[53] Zhang C, Cui Z, Zhang Y, et al. Holistic 3D scene understanding from a single image with implicit representation [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 8833-8842.

[54] Chen K, Lai YK, Wu YX, et al. Automatic semantic modeling of indoor scenes from low-quality RGB-D data using contextual information [J]. ACM Transactions on Graphics, 2014, 33(6).

[55] Hampali S, Stekovic S, Sarkar SD, et al. Monte Carlo Scene Search for 3D scene understanding [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13804-13813.

[56] Avetisyan A, Khanova T, Choy C, et al. SceneCAD: predicting object alignments and layouts in RGB-D scans [C] // Computer Vision and Pattern Recognition, 2020: 596-612.