

Structured3D：大型逼真数据集 用于结构化 3D 建模

贾政^{1,2*}^f, 张俊飞^{1*}, 李静², 唐睿¹,
高盛华^{2,3}, 和周子涵⁴

¹KooLab, 酷家乐 ²上海科技大学
³上海市智能视觉与影像工程研究中心
⁴宾夕法尼亚州立大学
<https://structured3d-dataset.org>

抽象的。最近，人们越来越关注开发基于学习的方法来检测和利用显着的半全局或全局结构，例如连接点、线、平面、长方体、光滑表面和所有类型的对称性，用于 3D 场景建模和理解。然而，地面实况注释通常是通过人工获得的，由于大量的 3D 结构实例，这对于此类任务尤其具有挑战性且效率低下。例如，线段）和其他因素，如视点和遮挡。在本文中，我们提出了一个新的合成数据集 Structured3D，旨在为广泛的结构化 3D 建模任务提供具有丰富 3D 结构注释的大型逼真图像。我们利用专业室内设计的可用性并自动从中提取 3D 结构。我们使用行业领先的渲染引擎生成高质量图像。我们使用我们的合成数据集与真实图像相结合来训练深度网络以进行房间布局估计，并在基准数据集上展示改进的性能。

关键词：数据集 · 3D 结构 · 逼真的渲染

1 简介

从图像和视频等 2D 感官数据推断 3D 信息长期以来一直是计算机视觉的核心研究课题。构建 3D 模型的传统方法通常依赖于检测、匹配和三角测量局部图像特征（例如、补丁、超像素、边缘和 SIFT 特征）。尽管在过去的几十年中取得了重大进展，但这些方法仍然存在一些基本问题。特别是局部特征检测对场景外观等大量因素很敏感（例如、无纹理区域和重复图案）、光照条件和遮挡。此外，嘈杂的基于点云的 3D 模型通常无法满足现实世界应用中对高级 3D 理解日益增长的需求。

* : 同等贡献。

^f: 贾政在酷家乐酷实验室实习时完成了部分工作。

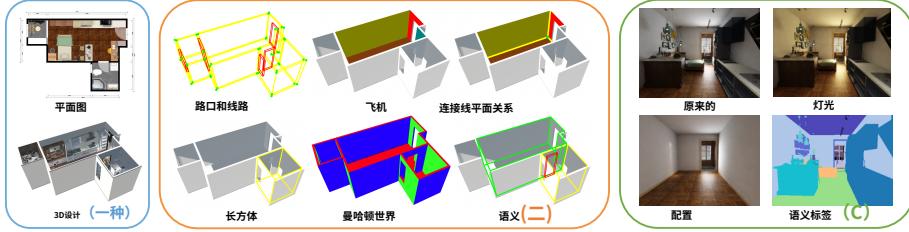


图 1: Structured3D 数据集。来自大量房屋设计 (一种) 由专业设计师创建, 我们自动提取各种ground truth 3D结构注释 (b)并生成照片般逼真的 2D 图像 (C) 。

在感知 3D 场景时, 人类在使用显着的全局结构 (如线条、轮廓、平面、光滑表面、对称性和重复图案) 方面非常有效。因此, 如果重建算法可以利用这种全局信息, 自然期望该算法能够获得更准确的结果。然而, 传统上, 从嘈杂的局部图像特征中可靠地检测这种全局结构在计算上具有挑战性。最近, 基于深度学习的方法在直接从图像中检测各种形式的结构方面显示出了可喜的结果, 包括线条 [12,40], 飞机 [19,35,16,36], 长方体 [10], 平面图 [17,18], 房间布局[14,41,26], 抽象的 3D 形状 [28,32], 以及光滑的表面 [11]。

随着深度学习方法的快速发展, 需要大量
大量准确注释的数据。为了训练提出的神经
网络, 大多数先前的工作收集他们自己的图像集并手动标记它们感兴趣的结构。这
种策略有几个缺点。第一的, 由于手动标记和验证所有结构实例的繁琐过程 (例如
, 线段) 在每个图像中, 现有数据集通常具有有限的大小和场景多样性。并且注释
也可能包含错误。第二, 由于每项研究主要关注一种类型的结构, 因此这些数据集
都没有标记多种类型的结构。因此, 现有方法无法利用不同类型结构之间的关系
(例如, 线和平面), 就像人类为有效、高效和稳健的 3D 重建所做的那样。

在本文中, 我们提出了一个具有丰富 3D 结构注释的大型合成数据集和室内人造
环境的逼真的 2D 渲染图 (图 1)。¹ 我们数据集设计的核心是 3D 结构的统一
表示, 它使我们能够有效地捕获场景中的多种类型的 3D 结构。具体来说, 所提出
的表示将任何结构视为关系之中几何基元. 例如, “线框”结构编码线段之间的入射
和相交关系, 而“长方体”结构编码其平面之间的旋转和反射对称关系。使用我们
的“原始 + 关系”表示, 可以轻松地为各种半全局和全局结构推导出基本事实注释
(例如, 线条, 线框, 平面, 规则形状,

表 1：具有结构注释的数据集概述。†：实际数字没有明确给出并且难以估计，因为这些数据集包含来自互联网 (LSUN Room Layout, PanoContext) 或多个来源 (LayoutNet) 的图像。*：数据集在发布时在线不可用。

数据集	# 场景	# 房间	# 帧	注释结构
平面RCNN [16] 线框 [12] 场景	-	-	100,000	飞机
城市 3D [40] SUN 原语 [34]	-	-	5,462	线框 (2D)
LSUN 房间布局 [39]	230	-	23,000	线框 (3D)
PanoContext [37] 布局网 [41]	-	-	785	长方体, 其他基元
MatterportLayout [42] 光栅到矢量 [17]	-	不适用 †	5,394	长方体布局
	-	不适用 †	500 (全景)	长方体布局
	-	不适用 †	1,071 (全景)	长方体布局
	-	不适用 †	2,295 (RGB-D全景)	曼哈顿布局
	870	-	-	平面图
结构化3D	3,500	21,835	196,515	“原始+关系”

平面图和房间布局），并在未来的数据驱动方法中利用它们的关系（例如，由场景中的平面相交形成的线框）。

为了创建一个大型数据集，旨在促进对结构化 3D 场景理解的数据驱动方法的研究，我们利用了专业室内设计和数百万生产级 3D 对象模型的可用性——所有这些都带有精细的几何细节和高分辨率纹理（图 1（一种））。我们首先使用计算机程序从原始房屋设计文件中自动提取有关 3D 结构的信息。如图所示。1(b)，我们的数据集包含丰富的 3D 房间结构注释，包括各种几何图元和关系。进一步生成逼真的 2D 图像（图 1）。1(c))，我们利用行业领先的渲染引擎来模拟光照条件。目前，我们的数据集包含 3,500 个场景中 21,835 个房间的 196k 多张图像 (IE, 房屋)。

为了展示所提出的 Structured3D 数据集的有用性和独特性，我们在数据集的子集上训练用于房间布局估计的深度网络。我们表明，在合成数据和真实数据上训练的模型优于仅在真实数据上训练的模型。此外，本着[27,8]，我们展示了数据集中的多模态注释如何使域适应任务受益。

综上所述，主要贡献这篇论文是：

- 我们创建了 Structured3D 数据集，其中包含 3,500 个场景中 21,835 个房间的丰富地面实况 3D 结构注释，以及超过 196k 个房间的逼真 2D 渲染。
- 我们引入了统一的“原始+关系”表示。这种表示使我们能够有效地捕获各种半全局或全局 3D 结构及其相互关系。
- 我们通过使用它来训练用于房间布局估计的深度网络并在公共基准测试中展示改进的性能来验证我们的数据集的有用性。

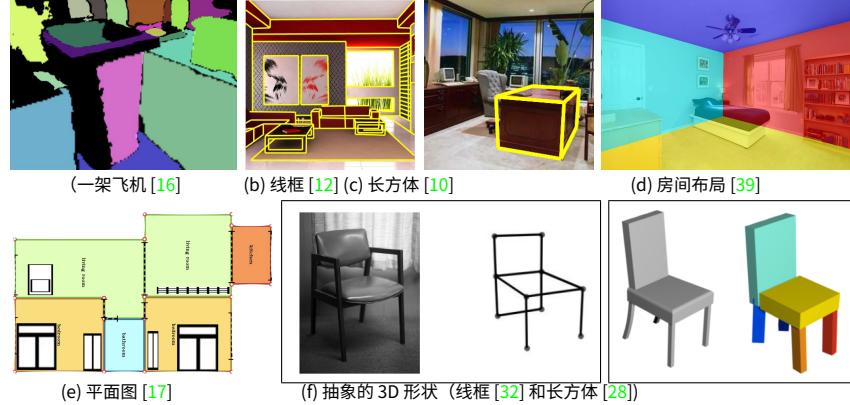


图 2：现有数据集中结构的示例注释。参考编号表示插图最初来自的纸张。

2 相关工作

数据集。桌子1总结了结构化3D场景建模的现有数据集。此外，[28,32]为数据集提供单个对象的结构化表示。我们在图1中显示了这些数据集中的示例注释。2. 请注意，大多数数据集中的基本事实注释都是手动标记的。这是所有这些数据集大小有限的主要原因之一，IE，包含不超过几千张图像。一个例外是[16]，它采用多模型拟合算法从ScanNet数据集中的3D扫描中自动提取平面[9]。但是这种算法对数据噪声和异常值很敏感，因此会在注释中引入错误（图1）。2（一种）。与我们的工作类似，SceneCity 3D [40]还包含从CAD模型中自动提取的基本事实的合成图像。但是场景的数量限制在230个。此外，这些数据集都没有超过一种标记的结构，尽管不同类型的结构之间通常有很强的关系。例如，从图中的线框。2(b)人类可以很容易地识别其他类型的结构，例如平面和长方体。我们的新数据集旨在弥合训练机器学习模型以实现人类水平的整体3D场景理解所需的内容与现有数据集提供的内容之间的差距。

请注意，我们的数据集与其他流行的大规模3D数据集非常不同，例如NYU v2 [23]，太阳RGB-D [24]、2D-3D-S [4,3]，扫描网络 [9] 和 Matterport3D [6]，其中地面实况3D信息以点云或网格的格式存储。这些数据集缺乏半全局或全局结构的地面实况注释。虽然理论上可以通过将结构检测算法应用于点云或网格来提取3D结构（例如，从ScanNet中提取平面，如[16]），检测结果往往是嘈杂的，甚至包含错误。此外，对于某些类型的结构

表 2：3D 场景数据集的比较。 \dagger ：网格是通过 3D 重建算法获得的。应用符号：O（对象检测）、U（场景理解）、S（图像合成）、M（结构化 3D 建模）。

数据集	场景设计类型	3D 注释	2D 渲染	应用
纽约大学 v2 [23]	真实的	原始 RGB-D	真实图像	欧
太阳RGB-D [24]	真实的	原始 RGB-D	真实图像	欧
2D-3D-S [4,3] 扫描网 [9]	真实的	网 \dagger	真实图像	欧
Matterport3D [6]	真实的	网 \dagger	真实图像	欧
孙克 [25]	业余	网	不适用	欧
SceneNet RGB-D [20] 室内网 [15]	随机的	网	逼真的照片	欧
	专业的	不适用	逼真的照片	欧斯
结构化3D	专业的	3D 结构	逼真的照片	欧斯曼

像线框和房间布局一样，如何从原始传感器数据中可靠地检测它们仍然是计算机视觉领域的一个活跃研究课题。

近年来，合成数据集在深度神经网络的成功训练中发挥了重要作用。室内场景理解的著名例子包括 SUNCG [25]，SceneNet RGB-D [20] 和内部网 [15]。这些数据集在场景多样性和帧数方面超过了真实数据集。但就像它们的真实对应物一样，这些数据集缺乏基本事实结构注释。一些合成数据集的另一个问题是 3D 模型和 2D 渲染的真实度。[38] 表明基于物理的渲染可以提高各种室内场景理解任务的性能。为了确保我们数据集的质量，我们使用了由专业设计师和最先进的工业渲染引擎创建的 3D 房间模型。桌子 2 总结了 3D 场景数据集的差异。

房间布局估计。房间布局估计旨在重建室内场景的封闭结构，包括墙壁、地板和天花板。现有的公共数据集（例如，全景上下文 [37] 和 LayoutNet [41]）假设一个简单的盒形布局。全景上下文 [37] 从 SUN360 数据集中收集了大约 500 张全景图 [33]，布局网 [41] 扩展布局注释以包含 2D-3D-S [3]。最近，MatterportLayout [42] 从 Matterport3D 收集 2,295 张 RGB-D 全景图 [6] 并将注释扩展到曼哈顿布局。我们注意到，这些真实数据集中的所有房间布局都是由人工手动标记的。由于房间结构可能被家具和其他物体遮挡，因此人类推断的“基本事实”可能与实际布局不一致。在我们的数据集中，所有真实的 3D 注释都是从原始房屋设计文件中自动提取的。

3 3D 结构的统一表示

我们数据集的主要目标是为地面实况 3D 结构提供丰富的注释。一种天真的方法是以与现有作品相同的格式生成和存储不同类型的 3D 注释，例如 [12]，飞机

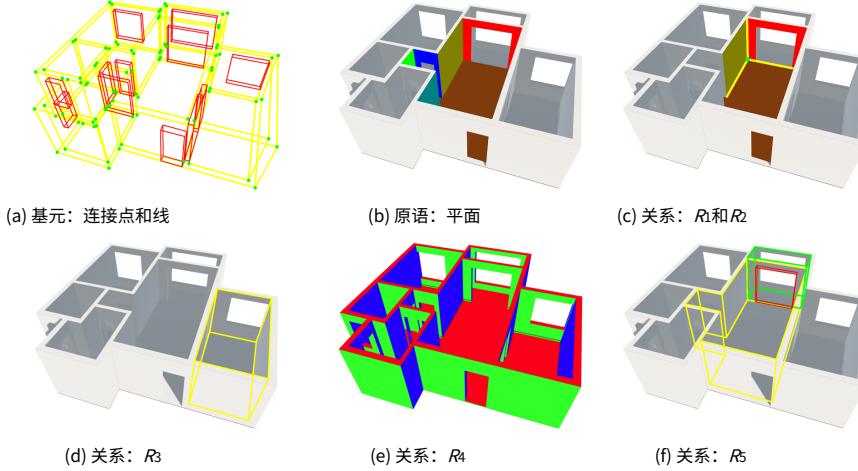


图 3：我们数据集中的地面实况 3D 结构注释由基元和关系表示。(一种)：路口和线路。(b):飞机。我们突出显示单间中的飞机。(C)：平面线和线交点关系。我们突出显示一个交界处，三条线在交界处相交，平面在每条线上相交。(d):长方体。我们突出显示一个长方体实例。(e):曼哈顿世界。我们使用不同的颜色来表示与不同方向对齐的平面。(F)：语义对象。我们突出了一个“房间”、一个“阳台”和连接它们的“门”。

如 [16]，平面图如 [17]，等等。但这会导致大量冗余。例如，人造环境中的平面通常由许多线段界定，这些线段是线框的一部分。更糟糕的是，通过分别表示线框和平面，它们之间的关系丢失了。在本文中，我们提出了一个统一的表示，以便在保持相互关系的同时最小化冗余。我们展示了文献中最常见的结构类型（例如，平面，长方体，线框，房间布局和平面图）可以从我们的表示中得出。

我们对结构的表示很大程度上受到了 Witkin 和 Tenenbaum 早期工作的启发 [31]，将结构表征为“一种形状、图案或配置，在空间和时间间隔内复制或延续，几乎没有变化”。因此，为了描述任何结构，我们需要指定：(i) 什么模式正在继续或复制（例如，补丁，边缘或纹理描述符），以及 (ii) 其复制或延续的域。在本文中，我们称前者原语和后者关系。

3.1 “原始+关系” 表示

我们现在展示如何使用统一的表示来描述人造环境。为了便于说明，我们假设场景中的所有对象都可以建模

通过分段平面表面。但是我们的表示可以很容易地扩展到更一般的表面。我们的表示图如图所示。[3](#).

原语。通常，人造场景具有以下几何图元：

- 平面 P：我们将场景建模为平面的集合 $P = \{p_1, p_2, \dots\}$ 。每个平面由其参数描述 $p = \{n, d\}$ ，在哪里 n 和 d 分别表示表面法线和到原点的距离。
- L 行：当两个平面在 3D 空间中相交时，将创建一条线。我们用大号 $=\{l_1, l_2, \dots\}$ 表示场景中所有 3D 线的集合。
- 连接点 X：当两条线在 3D 空间中相交时，就形成了一个交汇点。我们用 $X = \{x_1, x_2, \dots\}$ 来表示所有连接点的集合。

关系。接下来，我们定义几何图元之间的一些常见类型的关系：

- 平面线关系 (R_1)：我们使用矩阵 W_1 记录平面之间的所有入射和相交关系磷和线 L。具体来说， i,j -第一个条目 W_1 如果是 1 —— 开启 p_i ，否则为 0 。请注意，当且仅当对应的条目 W_1

$$W_1$$

是非零的。

- 线点关系 (R_2)：同样，我们使用矩阵 W_2 记录所有线之间的关联和相交关系大号，并指出 X。具体来说， i,j -第一个条目 W_2 如果是 1 —— 开启 l_i ，否则为 0 。请注意，当且仅当

对应的条目 W_2 是非零的。

- 长方体 (R_3)：长方体是平面基元的特殊排列，沿 x、y 和 z 轴具有旋转和反射对称性。对应的对称群是二面体群 D_{2H} 。

- 曼哈顿世界 (R_4)：这是一种特殊类型的 3D 结构，常用于室内外场景建模。它可以被视为一个分组关系，其中所有平面图元可以分为三类，磷₁,磷₂, 和磷₃,P =

$\overset{3}{\underset{i=1}{\text{磷}_i}}$ 此外，每个类由

单个法线向量 n_i ，这样 $n_i = -n_j$ ， $i \neq j$ 。

- 语义对象 (R_5)：语义信息对于许多 3D 计算机视觉任务至关重要。它可以被视为另一种类型分组关系，其中每个语义对象实例对应于上面定义的一个或多个原语。例如，每个“墙”、“天花板”或“地板”实例都与一个平面图元相关联；每个“椅子”实例都与一组多个平面图元相关联。此外，这样的分组是分层的。例如，我们可以进一步将一层、一层天花板和多面墙组合成一个“客厅”实例。而“门”或“窗”是连接两个房间（或一个房间和外部空间）的开口。

请注意，关系不是相互排斥的，因为原语可以属于相同类型或不同类型的多个关系实例。例如，一个平面图元可以由两个长方体共享，

同时属于曼哈顿世界模型中的三类之一。

讨论。我们上面讨论的原语和关系只是一些最常见的例子。它们绝不是详尽无遗的。例如，我们的表示可以很容易地扩展到包括其他基元，例如参数曲面。除了长方体，人造环境中还有许多其他类型的规则或对称形状，其中类型对应不同的对称组。

我们对 3D 结构的表示也与语义场景理解中的图表示有关 [13,2,30]。由于这些图侧重于语义，因此几何以简化的方式表示为 (i) 6D 对象姿势和 (ii) 粗略、离散的空间关系，例如“支持”、“前”、“后”和“相邻”。相比之下，我们的表示侧重于使用细粒度图元对场景几何进行建模 (I/E 、连接点、线和平面) 和关系（根据拓扑和规律性）。因此，它与先前工作中的场景图高度互补。直观地说，它可用于几何分析和合成任务，就像场景图用于语义场景理解一样。

3.2 与现有模型的关系

给定我们包含原语的表示 $\Phi = \{\text{磷}, \text{大号}, X\}$ 和关系 $R = \{R_1, R_2, \dots\}$ ，我们展示了文献中常见的几种类型的 3D 结构是如何从中推导出来的。我们再次请读者参考图 2 这些结构的插图。

飞机：文献中的大量研究将场景建模为 3D 平面的集合，其中每个平面由其参数和边界表示。要生成这样的模型，我们只需使用平面图元 P 。对于每个 $p \in \Phi$ ，我们通过使用矩阵进一步获得它的边界 W_1 在 R_1 找到所有的行 大号与 p 。

线框：线框由线条组成大号和交汇点磷，以及它们的发生和交叉关系 (R_2)。

长方体：该模型与 R_3 。

曼哈顿布局：曼哈顿房间布局模型包括定义的“房间” R_5 这也满足曼哈顿世界假设 (R_4)。

平面图：平面图是由一组线段和语义标签组成的二维向量表示（例如，房间类型）。为了获得这样的向量表示，我们可以识别大号和连接点 X 它位于“地板”上（定义见 R_5 ）。为了进一步获得语义房间标签，我们可以投影所有“房间”、“门”和“窗户”（定义见 R_5 ）到这一层。

抽象的 3D 形状：除了房间结构外，我们的表示还可以应用于单个 3D 对象模型，以创建线框或长方体形式的抽象，如上所述。

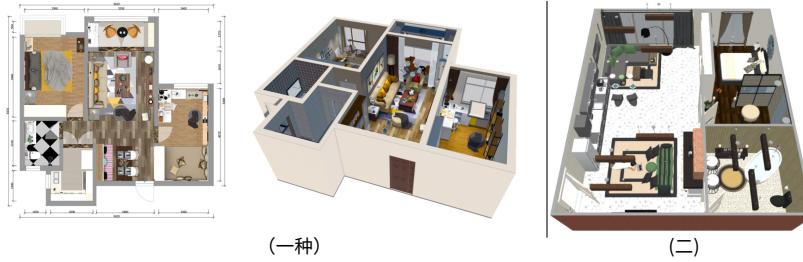


图 4：3D 房屋设计的比较。(一种)：我们数据库中的 3D 模型由专业设计师使用来自世界领先制造商的高品质家具模型创建。大多数设计都在实际生产中使用。

(b)：SUNCG 数据集中的 3D 模型 [25] 是使用 Planner 5D 创建的 [1]，业余室内设计的在线工具。

4 Structured3D 数据集

我们的统一表示使我们能够为结构化 3D 建模编码一组丰富的几何基元和关系。通过这种表示，我们的最终目标是构建一个数据集，可用于训练机器以实现人类对 3D 环境的理解。

作为实现这一目标的第一步，在本节中，我们将描述我们正在努力创建一个大规模的室内场景数据集，其中包括 (i) 场景的地面真实 3D 结构注释和 (ii) 场景的逼真 2D 渲染。请注意，在这项工作中，我们只专注于在房间结构上提取地面实况注释。我们计划在未来扩展我们的数据集以包括单个家具模型的 3D 结构注释。

在下文中，我们描述了创建数据集的一般过程。我们向读者推荐补充材料以获取更多详细信息，包括数据集统计和示例注释。

4.1 结构化 3D 模型的提取

为了提取“原始+关系”场景表示，我们利用专业设计师手工制作的大型房屋设计数据库。示例设计如图所示。**4 (一种)**。设计的所有信息都以行业标准格式存储在数据库中，因此有关几何形状的规范（例如，每面墙的精确尺寸），纹理和材料，以及功能（例如，墙属于哪个房间）的所有对象都可以很容易地检索到。

从数据库中，我们选择了 21,835 个房间的 3,500 个房屋设计。我们创建了一个计算机程序来自动提取与房间结构相关的所有几何图元，其中包括天花板、地板、墙壁和开口（门窗）。鉴于这些实体的精确测量和相关信息，可以直接生成所有平面、线和交汇点，以及它们的关系 (R_1 和 R_2)。

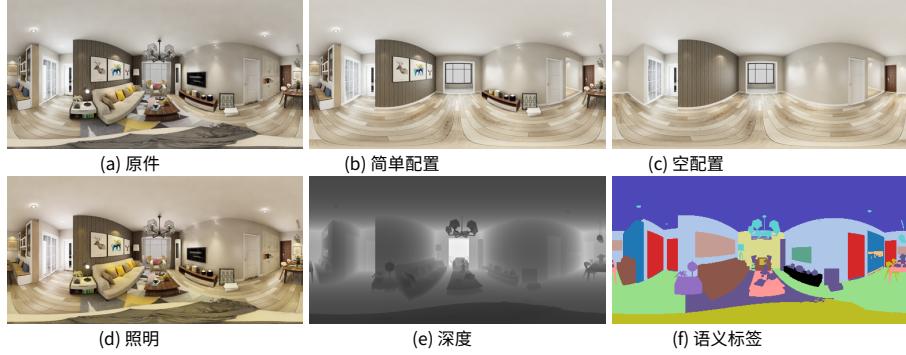


图 5：我们渲染的全景图像示例。

由于测量是高度准确且无噪音的，其他类型的关系，如曼哈顿世界 (R_3) 和长方体 (R_4) 也可以通过对基元进行聚类来轻松获得，然后进行几何验证过程。最后，包含语义信息 (R_5) 到我们的表示中，我们将专业设计师提供的相关标签映射到我们表示中的几何图元。如图。3 显示了提取的几何图元和关系的示例。

4.2 逼真的 2D 渲染

为确保我们 2D 渲染的质量，我们的渲染引擎是与一家专门从事室内设计渲染的公司合作开发的。我们的引擎使用众所周知的光线追踪方法 [21]，一种近似真实全局照明 (GI) 的蒙特卡罗方法，用于 RGB 渲染。其他地面实况图像由英特尔 Embree 之上的定制路径跟踪器渲染器获得 [29]，一个用于 x86 CPU 的光线追踪内核的开源集合。

每个房间均由专业设计师手工打造，拥有来自世界领先制造商的超过 100 万件 CAD 家具模型。这些高分辨率家具模型以真实世界的尺寸进行测量，并用于实际生产。还提供了默认照明设置。如图。4 将我们数据库中的 3D 模型与 SUNCG 中的模型进行比较 [25]，它们是使用 Planner 5D 创建的 [1]，业余室内设计的在线工具。

在渲染时，将全景或针孔相机放置在房间内未被物体占据的随机位置。我们使用 512×1024 分辨率的全景图和 720×1280 用于透视图像。如图。5 显示了我们的引擎渲染的示例全景图。对于每个房间，我们通过移除部分或全部家具来生成不同的配置（完整的、简单的和空的）。我们还修改了照明设置以生成具有不同温度的图像。对于每个图像，我们的数据集还包括深度图和语义掩码。如图。6 说明了我们数据集的照片真实程度，我们将渲染图像与设计指导的真实装饰照片进行比较。



图 6：照片般逼真的渲染与真实世界的装饰。第一列和第三列是渲染图像。

4.3 用例

由于我们数据集的独特特征，我们设想它在方法论和应用方面都有助于计算机视觉研究。

方法。由于我们的数据集包含多种类型的 3D 结构注释以及地面实况标签（例如语义图、深度图和 3D 对象边界框），它使研究人员能够为各种视觉任务设计新颖的多模态或多任务方法。作为一个例子，我们在第 5 也就是说，通过利用我们数据集中的多模态注释，我们可以提高域适应框架中现有房间布局估计方法的性能。

应用程序。我们的数据集还有助于研究许多问题和应用。例如，如表所示 1，所有用于房间布局估计的公开数据集都仅限于简单的长方体房间。我们的数据集是第一个提供一般（非长方体）房间布局注释的数据集。作为另一个例子，用于平面图重建的现有数据集[18,7] 包含大约 100-150 个场景，而我们的数据集包含 3,500 个场景。

另一个将从我们的数据集受益的主要研究方向是图像合成。使用逼真的渲染引擎，我们能够生成给定的图像任何场景配置和视点。这些图像可用作包括图像修复在内的任务的基本事实（例如，在移除某些家具时完成图像）和新颖的视图合成。

最后，我们想强调我们的数据集在扩展能力方面的潜力。正如我们之前提到的，统一表示使我们能够在数据集中包含许多其他类型的结构。至于 2D 渲染，根据应用程序，我们可以轻松模拟不同的效果，例如光照条件、鱼眼和新颖的相机设计、运动模糊和成像噪声。此外，数据集可以扩展为包括用于视觉 SLAM 等应用的视频 [5]。

5 个实验

5.1 实验设置

为了展示我们数据集的优势，我们使用它来训练深度神经网络以进行房间布局估计，这是结构化 3D 建模中的一项重要任务。

表 3：房间布局统计。 \neq : MatterportLayout 是唯一具有非长方体布局注释的其他数据集，但在发布时不可用。

# 角落	4	5	6	7	8	9	10+	全部的
Matterport布局 \neq	1211	0	501	0	309	0	274	2295
结构化3D	13743	52	3727	30	1575	17	2691	21835

真实数据集。我们使用与 LayoutNet 相同的数据集 [41]。数据集由来自 PanoContext [37] 和 2D- 3D-S [3]，包括 818 个训练图像、79 个验证图像和 166 个测试图像。请注意，这两个数据集都仅提供长方体布局注释。

我们的 Structured3D 数据集。在本实验中，我们使用具有原始照明和完整配置的全景图子集。每个全景图对应于我们数据集中的不同房间。我们在表中的数据集中显示了不同房间布局的统计数据3. 由于当前的真实数据集仅包含长方体布局注释 ($/E$, 4 个角)，我们在数据集中选择了 12k 个具有长方体布局的全景图像。我们将图像分成 10k 用于训练、1k 用于验证和 1k 用于测试。

评估指标。下列的 [41,26]，我们采用三个标准度量：(i) 3D IoU：预测的 3D 布局和地面实况之间的交集，(ii) 角误差 (CE)：归一化预测角点和地面实况之间的距离，以及 (iii) 像素误差 (PE)：预测平面类和地面实况之间的像素误差。

基线。我们选择了两种最近的基于 CNN 的方法，LayoutNet [41,42]¹ 和 Horizo nNet [26]²，基于它们的性能和源代码可用性。LayoutNet 使用 CNN 从全景图和消失线中预测角概率图和边界图，然后根据网络预测优化布局参数。Horizo nNet 将房间布局表示为三个 1D 向量， $/E$ ，楼-墙、顶-墙的边界位置，墙-墙边界的存。它训练 CNN 直接预测三个 1D 向量。在本文中，我们遵循各自方法的默认训练设置。具体培训流程见补充材料。

5.2 实验结果

增强真实数据集。在这个实验中，我们以四种不同的方式训练 LayoutNet 和 Horizo nNet：(i) 仅在我们的合成数据集上训练 (“s”)，(ii) 仅在真实数据集上进行训练 (“r”)，(iii) 使用平衡梯度贡献 (BGC) 在合成和真实数据集上进行训练 [22] (“s + r”)，(iv) 在我们的合成数据集上进行预训练，然后在真实数据集上进行微调 (“s \rightarrow r”)。我们采用 LayoutNet 的训练集作为本次实验的真实数据集。结果如表所示4. 可以看到，增强真实

¹<https://github.com/zouchuhang/LayoutNetv2>

²<https://github.com/sunset1995/Horizo nNet>

表 4：不同训练方案下的定量评价。最好和次优的结果分别用粗体和下划线表示。

方法	配置。	全景语境			2D-3D-S		
		3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓
布局网 [41,42]	s	75.64	1.31	4.10	57.18	2.28	7.55
	r	84.15	0.64	<u>1.80</u>	83.39	0.74	2.39
	s + r	84.96	0.61	1.75	<u>83.66</u>	<u>0.71</u>	2.31
	s → r	<u>84.77</u>	0.63	1.89	84.04	0.66	2.08
地平线网 [26]	s	75.89	1.13	3.15	67.66	1.18	3.94
	r	83.42	0.73	2.09	84.33	0.64	2.04
	s + r	<u>84.45</u>	<u>0.70</u>	<u>1.89</u>	<u>84.36</u>	0.59	1.90
	s → r	85.27	0.66	1.86	86.01	<u>0.61</u>	1.84

表 5：在预训练中使用不同的合成数据大小进行定量评估。最好和次优的结果分别用粗体和下划线表示。

方法	合成的数据大小	全景语境			2D-3D-S		
		3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓
布局网 [41,42]	1k	83.81	<u>0.66</u>	1.99	83.57	0.72	2.31
	5k	<u>84.47</u>	0.67	<u>1.97</u>	84.55	<u>0.69</u>	<u>2.21</u>
	10k	84.77	0.63	1.89	<u>84.04</u>	0.66	2.08
地平线网 [26]	1k	83.77	0.74	2.11	85.19	<u>0.63</u>	2.01
	5k	<u>84.13</u>	<u>0.73</u>	<u>2.07</u>	86.35	0.61	<u>1.87</u>
	10k	85.27	0.66	1.86	<u>86.01</u>	0.61	1.84

包含我们合成数据的数据集提高了两个网络的性能。我们向读者推荐补充材料以获得更多定性结果。

性能与合成数据大小。我们进一步研究了预训练中使用的合成图像数量与真实数据集的准确性之间的关系。我们采样 1k、5k 和 10k 合成图像进行预训练，然后在真实数据集上微调模型。结果如表所示 5. 正如预期的那样，使用更多的合成数据通常会提高性能。

领域适应。领域适应技术（例如, [27]）已被证明在将在合成数据上学习的模型直接应用于真实环境时有效地缩小性能差距。在这个实验中，我们不假设可以访问真实数据集中的地面实况布局标签。我们采用 LayoutNet 作为任务网络，分别使用 PanoContext 和 2D- 3D-S。我们应用鉴别器网络来对齐两个域的 LayoutNet 的输出特征。灵感来自 [8]，我们通过向 LayoutNet 添加另一个解码器分支来进一步利用数据集中的多模态注释进行深度预测。我们将边界、角点和深度预测连接起来作为判别器网络的输入。结果如表所示 6. 通过合并附加信息，IE，深度图，我们进一步提高了两个数据集的性能。这说明了在我们的数据集中包含多种类型的基本事实的优势。

真实数据集的局限性。由于人为错误，实际数据集中的注释并不总是与实际房间布局一致。在左图中

表 6：域适应结果。NA：非自适应基线。+DA：对齐布局估计输出。+深度：对齐布局估计和深度输出。真实：在目标域中训练。

方法	全景语境			2D-3D-S		
	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓	3D IoU (%) ↑行政长官 (%) ↓市盈率 (%) ↓
不适用	75.64	1.31	4.10	57.18	2.28	7.55
+ 达	76.91	1.19	3.64	70.08	1.36	4.66
+ 深度	78.34	1.03	2.99	72.99	1.24	3.60
真实的	81.76	0.95	2.58	81.82	0.96	3.13



图 7：真实数据集的局限性。剩下：PanoContext 数据集。对：2D-3D-S 数据集。蓝线是地面实况布局，绿线是预测。

图的 7，房间是非长方体布局，但地面实况布局被标记为长方体形状。在右图中，前墙未标记为地面实况。这些例子说明了使用真实数据集作为基准的局限性。我们通过从原始设计文件自动生成基本事实来避免数据集中的此类错误。

六，结论

在本文中，我们介绍了 Structured3D，这是一个大型合成数据集，具有 21,835 个房间的丰富地面实况 3D 结构注释和超过 196k 的照片般逼真的 2D 渲染。在我们数据集的许多潜在用例中，我们进一步证明了它在增强真实数据和促进房间布局估计任务的域适应方面的好处。

我们认为这项工作是构建智能机器的重要且令人兴奋的一步，可以实现人类水平的整体 3D 场景理解。未来，我们将继续向数据集中添加更多场景和对象的 3D 结构注释，并探索使用数据集的新方法来推进结构化 3D 建模和理解技术。

确认。感谢酷家乐提供房屋设计数据库和渲染引擎。特别感谢酷家乐的叶青和吴奇对数据渲染的帮助。这项工作得到了国家重点研发计划 (#2018AAA0100704) 和国家自然科学基金 (#61932020) 的部分支持。周子涵得到了 NSF 奖 #1815491 的支持。

参考

1. 规划师 5d。 <https://planner5d.com>^{9,10}
2. Armeni, I., He, ZY, Gwak, J., Zamir, AR, Fischer, M., Malik, J., Savarese, S.: 3d 场景图：统一语义、3d 空间和相机的结构. 在：ICCV。第 5664–5673 页 (2019 年) ⁸
3. Armeni, I., Sax, A., Zamir, AR, Savarese, S.: 用于室内场景理解的联合 2d-3d 语义数据。心电图绝对/1702.01105 (2017)^{4,5,12}
4. Armeni, I., Sener, O., Zamir, AR, Jiang, H., Brilakis, IK, Fischer, M., Savarese, S.: 大型室内空间的 3d 语义解析. 在：CVPR。第 1534–1543 页 (2016 年) ^{4,5}
5. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, ID, Leonard, JJ: 同步定位和映射的过去、现在和未来：迈向稳健感知时代。IEEE Trans。机器人技术32(6), 1309–1332 (2016)¹¹
6. Chang, AX, Dai, A., Funkhouser, TA, Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d：从 RGB 学习-D 室内环境中的数据。在：3DV。第 667–676 页 (2017 年) ^{4,5}
7. Chen, J., Liu, C., Wu, J., Furukawa, Y.: Floor-sp：按连续房间最短路径计算平面图的逆 CAD。在：ICCV。第 2661–2670 页 (2019 年) ¹¹
8. Chen, Y., Li, W., Chen, X., Van Gool, L.: 从合成数据中学习语义分割：一种几何引导的输入输出适应方法。在：CVPR。第 1841–1850 页 (2019 年) ^{3,13}
9. Dai, A., Chang, AX, Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet：带丰富注释的室内场景 3D 重建。在：CVPR。第 5828–5839 页 (2017 年) ^{4,5}
10. Dwibedi, D., Malisiewicz, T., Badrinarayanan, V., Rabinovich, A.: 深度长方体检测：超越二维边界框。心电图绝对/1611.10010 (2016)^{2,4}
11. Groueix, T., Fisher, M., Kim, VG, Russell, B., Aubry, M.: 学习 3D 表面生成的纸浆方法。在：CVPR。第 216–224 页 (2018 年) ²
12. Huang, K., Wang, Y., Zhou, Z., Ding, T., Gao, S., Ma, Y.: 学习解析人造环境图像中的线框。在：CVPR。第 626–635 页 (2018 年) ^{2,3,4,5}
13. Huang, S., Qi, S., Zhu, Y., Xiao, Y., Xu, Y., Zhu, SC: 从单个 rgb 图像进行整体 3d 场景解析和重建。在：ECCV。第 194–211 页 (2018 年) ⁸
14. Lee, C., Badrinarayanan, V., Malisiewicz, T., Rabinovich, A.: Roomnet：端到端房间布局估计。在：ICCV。第 4875–4884 页 (2017 年) ²
15. Li, W., Saeedi, S., McCormac, J., Clark, R., Tzoumanikas, D., Ye, Q., Huang, Y., Tang, R., Leutenegger, S.: Interiornet：超大规模多传感器照片般逼真的室内场景数据集。在：BMVC。页。77 (2018)⁵
16. Liu, C., Kim, K., Gu, J., Furukawa, Y., Kautz, J.: Planercnn：从单个图像进行 3d 平面检测和重建。在：CVPR。第 4450–4459 页 (2019 年) ^{2,3,4,6}
17. Liu, C., Wu, J., Kohli, P., Furukawa, Y.: 光栅到矢量：重新审视平面图转换。在：ICCV。第 2214–2222 页 (2017 年) ^{2,3,4,6}
18. Liu, C., Wu, J., Furukawa, Y.: Floornet：从 3D 扫描重建平面图的统一框架。在：ECCV。第 203–219 页 (2018 年) ^{2,11}
19. Liu, C., Yang, J., Ceylan, D., Yumer, E., Furukawa, Y.: Planenet：从单个 rgb 图像进行分段平面重建。在：CVPR。第 2579–2588 页 (2018 年) ²

20. McCormac, J., Handa, A., Leutenegger, S., Davison, AJ: Scenenet RGB-D: 5m 合成图像能否在室内分割上击败通用 imagenet 预训练?
在: ICCV。第 2697–2706 页 (2017 年) 5
21. Purcell, TJ, Buck, I., Mark, WR, Hanrahan, P.: 可编程图形硬件上的光线追踪。ACM 翻译。图形。21(3), 703–712 (2002) 10
22. Ros, G., Stent, S., Alcantarilla, PF, Watanabe, T.: 训练用于道路场景语义分割的约束反卷积网络。心电图绝对/1604.01545 (2016) 12
23. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: 室内分割和支持从 RGBD 图像推断。
在: ECCV。第 746–760 页 (2012 年) 4,5
24. Song, S., Lichtenberg, SP, Xiao, J.: SUN RGB-D: RGB-D 场景理解基准套件。在:
CVPR。第 567–576 页 (2015 年) 4,5
25. Song, S., Yu, F., Zeng, A., Chang, AX, Savva, M., Funkhouser, TA: 来自单个深度图像的语义场景完成。在: CVPR。第 1746–1754 页 (2017 年) 5,9,10
26. Sun, C., Hsiao, CW, Sun, M., Chen, HT: Horizo nnet: 具有一维表示和全景拉伸数据增强的学习室布局。在: CVPR。第 1047–1056 页 (2019 年) 2,12,13
27. Tsai, YH, Hung, WC, Schulter, S., Sohn, K., Yang, MH, Chandraker, M.: 学习为语义分割调整结构化输出空间。在: CVPR。第 7472–7481 页 (2018 年) 3,13
28. Tulsiani, S., Su, H., Guibas, LJ, Efros, AA, Malik, J.: 通过组装体积基元学习形状抽象。
在: CVPR。第 2635–2643 页 (2017 年) 2,4
29. Wald, I., Woop, S., Benthin, C., Johnson, GS, Ernst, M.: Embree: 高效 CPU 光线追踪的内核框架。ACM 翻译。图形。33(4), 143:1–143:8 (2014) 10
30. Wang, K., Lin, YA, Weissmann, B., Savva, M., Chang, AX, Ritchie, D.: Planit: 使用关系图和空间先验网络规划和实例化室内场景。ACM 翻译。图形。38(4) (2019) 8
31. Witkin, AP, Tenenbaum, JM: 论结构在视觉中的作用。在: Beck, J., Hope, B., Rosenfeld, A. (编辑) 人类和机器视觉, 第 481–543 页。学术出版社 (1983) 6
32. Wu, J., Xue, T., Lim, JJ, Tian, Y., Tenenbaum, JB, Torralba, A., Freeman, WT: 用于以观众为中心的线框建模的 3d 解释器网络。IJCV 126(9), 1009–1026 (2018) 2,4
33. Xiao, J., Ehinger, KA, Oliva, A., Torralba, A.: 使用全景位置表示识别场景视点。在:
CVPR。第 2695–2702 页 (2012 年) 5
34. Xiao, J., Russell, B., Torralba, A.: 在单视图图像中定位 3d 长方体。在: NeurIPS。第
746–754 页 (2012 年) 3
35. Yang, F., Zhou, Z.: 通过卷积神经网络从单个图像中恢复 3d 平面。在: ECCV。第
87–103 页 (2018 年) 2
36. Yu, Z., Zheng, J., Lian, D., Zhou, Z., Gao, S.: 通过关联嵌入的单图像分段平面 3d 重建。
在: CVPR。第 1029–1037 页 (2019 年) 2
37. Zhang, Y., Song, S., Tan, P., Xiao, J.: Panocontext: 用于全景场景理解的全房间 3d 上下文模型。在: ECCV。第 668–686 页 (2014 年) 3,5,12
38. Zhang, Y., Song, S., Yumer, E., Savva, M., Lee, JY, Jin, H., Funkhouser, T.: 使用卷积神
经网络进行室内场景理解的基于物理的渲染。在: CVPR。第 5287–5295 页 (2017 年) 5
39. Zhang, Y., Yu, F., Song, S., Xu, P., Seff, A., Xiao, J.: 大规模场景理解挑战: 房间布局估计
(2016) 3,4

40. Zhou, Y., Qi, H., Zhai, S., Sun, Q., Chen, Z., Wei, LY, Ma, Y.: 学习从单个图像重建 3d 曼哈顿线框。在：ICCV。第 7698-7707 页（2019 年）[2,3,4](#)
41. Zou, C., Colburn, A., Shan, Q., Hoiem, D.: Layoutnet：从单个 RGB 图像重建 3d 房间布局。在：CVPR。第 2051-2059 页（2018 年）[2,3,5,12,13](#)
42. Zou, C., Su, J., Peng, C., Colburn, A., Shan, Q., Wonka, P., Chu, H., Hoiem, D.: 从单个 360 度图像重建 3d 曼哈顿房间布局。心电图 绝对/1910.04099 (2019)[3,5,12,13](#)