arXiv:2301.01424v1 [cs.GR] 4 Jan 2023

从人体运动合成场景

Sifan Ye*斯坦福大学 美利坚合众国
sifan.ye
@cs.stanford.edu

Yixing Wang*斯坦福大学 美利坚合众国
yiw998
@stanford.edu

Jiaman Li 斯坦福大学 美利坚合众国
ji
amanli@stanford.edu

Dennis Park 丰田研究院 美利坚合众国
de
nnis.park@tri.global

C. Karen Liu 斯坦福大学 美利坚合众国
karenliu
@cs.stanford.edu

Huazhe Xu†斯坦福大学 美利坚合众国
huazhexu@stanford.edu

Jiajun Wu†斯坦福大学 美利坚合众国
jiajunwu
@cs.stanford.edu



(a) 人体运动输入　(b) 带语义标签的合成场景　(c) 带纹理的合成场景

图 1：给定一段人体运动序列，SUMMON 可合成在物理上合理、语义上恰当的物体。

摘要

尽管大规模捕捉人体在多样且复杂场景中的运动极具价值，但其成本通常被认为高得令人难以承受。与此同时，人体运动本身蕴含着关于人所处并与之交互的场景的丰富信息。例如，坐着的人暗示场景中存在一把椅子，而其腿部姿势还能进一步反映椅子的姿态。本文提出基于人体运动合成多样、语义合理且物理上可信的场景。我们的框架**场**景合成源自**人体运动姿态**(SUMMON)包括两个步骤。首先，它使用我们新提出的接触预测器 ContactFormer，从人体运动中获取时间上连贯的接触标签。随后，SUMMON 基于这些预测选择交互物体并优化物理合理性损失；此外，它还会在场景中布置与人不发生交互的物体。实验结果表明，SUMMON 能够合成可行、可信且多样的场景，并有望为研究社区生成大量人-场景交互数据。

CCS 概念

• **计算方法** → **运动处理**；**形状推断**；**场景理解**。

关键词

场景合成、运动分析、活动理解

ACM 参考格式：

Sifan Ye、Yixing Wang、Jiaman Li、Dennis Park、C. Karen Liu、Huazhe Xu 和 Jiajun Wu。2022年。从人体运动合成场景。载于亚洲计算机图形学国际会议 *2022* 会议论文集（*SA '22* 会议论文集），*2022*年*12*月*6*日至*9*日，大邱，大韩民国。ACM，纽约州纽约市，9 页。https://doi. org/10.1145/3550469.3555426

1 引言

在三维场景中采集、建模并合成逼真的人体运动，对于虚拟现实、游戏角色动画和人机交互等一系列应用至关重要。为了促进该领域的研究，人们整理了大量数据集 [Hassan 等，2019；Mahmood 等，2019；Savva 等， 2016] 用于采集人体运动。例如，Bhatnagar 等人 [2022] 收集了人操纵物体时的轨迹。PROX-E 数据集 [Zhang 等，2020a] 包含了人体与场景网格的接触信息。然而，构建同时标注多样人体运动和丰富三维场景的高质量大规模数据集仍然充满挑战。这主要是因为当前的数据采集流程依赖昂贵的设备，例如动作捕捉系统、结构光相机和三维扫描仪，因此只能在实验室环境中进行；而实验室的物理空间有限，场景

* 和† 表示共同贡献。https://sites.google.com/stanford.edu/summon

特此授权免费制作本作品部分或全部内容的数字或纸质副本，用于个人或课堂用途，前提是副本不得为牟利或商业利益而制作或分发，且副本须载有本声明以及首页的完整引文。本作品中第三方组件的版权必须得到尊重。如需用于其他用途，请联系所有者/作者。*SA '22* 会议论文集，*2022*年*12*月*6*日至*9*日，大邱，大韩民国© 2022年版权归所有者/作者所有。ACM ISBN 978-1-4503-9470-3/22/12。https://doi.org/10.1145/3550469.3555426

多样性也因此受到限制。受近期三维人体姿态建模及其与环境接触建模进展的启发，我们旨在通过探索一种新的可能性来应对这些挑战：我们能否仅根据人体运动学习合成场景？如果成功，我们的系统还将拥有数据采集之外的诸多潜在应用，例如根据艺术家在虚拟现实中的动作，为虚拟环境的创作提供建议。

近期研究提出根据人体轨迹和学习到的房间先验来估计房间布局 [Nie 等人 2022]。然而，重建的布局只考虑了语义，而未考虑可供性。Yi 等人 [2022] 提出从视觉输入中重建场景物体，然后利用人-场景交互（HSIs）进一步提升场景的可行性。尽管这种方法能够生成物理上合理的重建结果，但它需要额外的视觉输入，因此重建出的场景受到限制。

我们提出了基于人体运动的**场**景合成（**人**体**运**动**合成**，SUMMON）方法，如图 1 所示。该方法仅依据 3D人体姿态轨迹预测场景中可行的物体放置位置。SUMMON 由两个模块组成：人-场景接触预测模块和场景合成模块。人-场景接触预测模块名为 ContactFormer，它利用现有的人-场景交互数据，学习从人体顶点到与其接触的物体语义标签的映射。ContactFormer 在 Hassan 等，2021b] 提出的先前方法 [基础上，引入时间线索，以提高标签预测在时间上的一致性。获得估计的语义接触点后，场景合成模块首先根据物体与主体在语义和物理可供性方面的匹配程度，搜索适合这些接触点的物体；然后依据人体运动以及前序步骤推断出的物体，将其他未与人接触的物体布置到场景中。

我们使用 PROXD [Hassan等人 2019]和 GIMO [Zheng等人 2022] 数据集开展实验。在接触估计方面，ContactFormer 的表现优于以往的单帧接触预测方法 [Hassanet al. 2021b]。在场景合成方面，与基线方法相比，我们提出的系统生成的场景更加逼真、符合物理规律且更加多样，并在各项评估指标和人工评估中表现更优。

我们的贡献主要有三点。首先，我们提出 SUMMON，这是一种仅根据人体运动轨迹合成语义合理、符合物理规律且多样化场景的系统。其次，作为 SUMMON 的组成部分，我们提出接触预测模块 ContactFormer，它通过建模语义标签的时间一致性，表现优于现有方法。第三，我们证明，SUMMON 合成的场景在定性和定量评估中均持续优于现有方法。

2 相关工作

场景可供性学习。从人-场景交互中学习可供性近年来受到了广泛关注 [Chen 等人 2019；Chuang 等人 2018；Delaitre 等人 2012； Fouhey 等人 2012； Gupta 等人 2011； Wang 等人 2019a；Zhu 等人 2014]。在相关文献中，研究人员探讨如何将人体骨架置于场景中。例如，Wang 等人 [2017] 提出从情景喜剧视频中学习可供性，以便在静态图像中定位人体骨架。Li 等人 [2019a] 提出了一种三维姿势生成模型，用于预测场景中合理的人体姿势。随着更优人体表示方法的发展，已有一些方法尝试将三维人体置于场景中 [Zhang 等，2020a]。最近，POSA [Hassanet al. 2021b] 学习了一个模型，通过为 SMPL-X 人体模型的顶点添加接触概率和语义标签，将人体姿势置于三维场景网格中。Blinn 等人 [2021]提出了一种拟合与舒适度损失，用于训练一个具有可供性感知能力的模型，以生成适合人体姿态的椅子。一些研究也尝试收集或生成涉及人-场景交互的数据。例如，VirtualHome [Puig 等人 2018] 提供了一个模拟的三维环境，类人智能体可以在其中与三维物体交互。BEHAVE [Bhatnagar 等人 2022]提供了一个数据集，其中包含使用 SMPL 参数化的真实全身人体在三维场景中通过接触点与物体交互并操控物体的数据。我们的工作在可供性学习研究的基础上更进一步：我们首先学习理解可供性，然后利用它们合成场景，以用于其他相关任务。

人体运动合成。 动作合成是计算机图形学和视觉领域的一个长期研究问题 [Brand and Hertzmann 2000; Holden et al. 2016; Kovar and Gleicher 2003; Park et al. 2002; Spallone 2015]。Xu等人 [2020] 提出了一种层级式方法，通过记忆库检索短时程参考片段来生成长时程运动。Harvey等人 [2020] 提出借助额外的嵌入来稳健地预测运动。近年来，许多研究也将环境纳入考量 [Hassan等人 2021a; Rempe et al. 2021; Wang et al. 2021a]。例如，Wang等人 [2021a] 将以场景网格为条件的长时程人体运动合成与可供性优化相结合，从而生成逼真的人体轨迹。SAMP [Hassan等人 2021a] 学习跨同一物体类别不同实例的泛化交互。我们的工作旨在解决逆问题：给定人体运动轨迹，生成合理的场景。

场景合成。 我们的工作也与合成合理的三维场景和房间布局密切相关 [Li 等人，2019b；Luo 等人，2020年；Purkait 等人，2020年；Ritchie 等人，2019年；Wang 等人，2019b、2021b；Zhang 等人，2020b；Zhou 等人， 2019]。例如，ATISS [Paschali-dou 等人 2021] 学习一种用于家具摆放的自回归生成模型。它可用于生成合理的新颖房间布局、根据现有物体补全场景，以及在给定空间约束的情况下建议可能的摆放位置。另一项工作 Pose2Room [Nie 等人 2022], 根据三维人体姿态轨迹预测物体的边界框。MOVER [Yi 等人 2022]利用单目 RGB 视频中的三维人体预测作为约束，重建三维物体。与这些先前方法不同，我们的模型不仅能生成布局，还能仅根据人体轨迹生成可负担得起的物体。

3 方法

我们的目标是仅根据人体运动序列预测一组家具物体及其物理合理的三维构型。我们首先在第 3.1 节介绍人体和接触表示。SUMMON 为人体的每个顶点生成时间一致的接触语义估计，以检索合适的物体（第 3.2 节）。随后，我们根据接触位置和物理合理性优化物体摆放（第 3.3 节）。图2展示了我们的方法示意图。

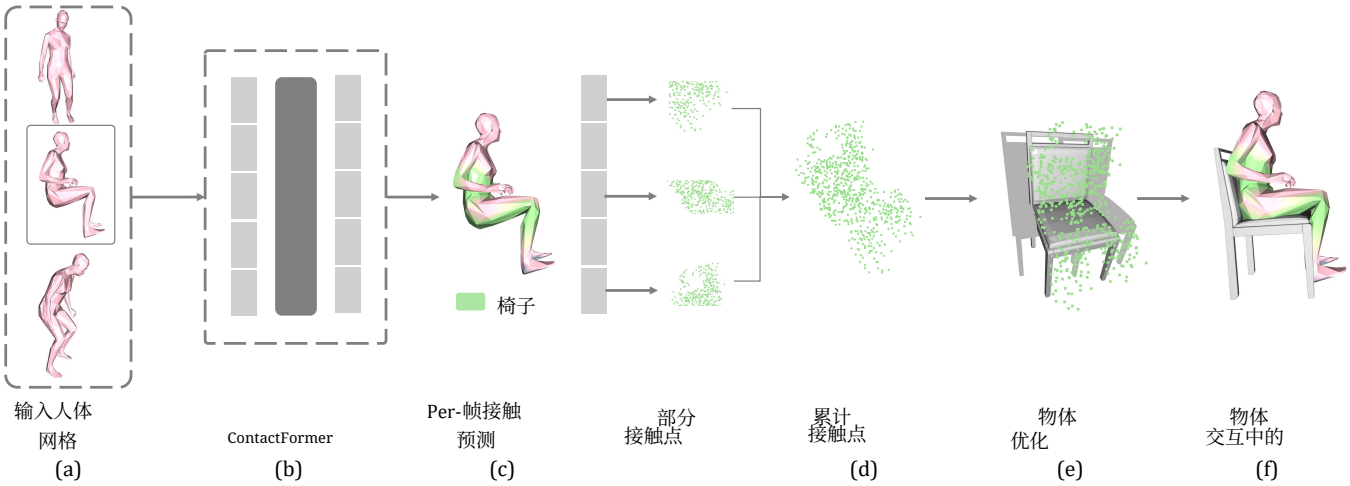

图2：SUMMON 概览：(a) 与场景交互的人体网格输入序列，(b) 预测逐帧接触标签的 Contact- Former，(c) 逐帧接触预测，(d) 估计的接触点，(e) 合成物体，以及 (f) 交互中的物体。

3.1 人体与接触表示

我们使用经修改的 SMPL-X [Pavlakos等人 2019] 提出的模型来表示人体姿态。具体而言，我们将人体参数化为 $M(\theta,\ \beta)$：$\mathbb{R}^{|\theta|\times|\beta|} \rightarrow \mathbb{R}^{3N}$，其中 θ表示姿态参数，β表示学习得到的形状空间中的系数，N表示 SMPL-X人体网格中的顶点数量。为提高计算效率，我们遵循 Hassan等人 [2021b]的先前工作，将顶点数从 10，475 下采样至 655 个点。

我们使用逐顶点特征表示接触信息。对于每个顶点 $v_b \in V_b$，其中 V_b 表示人体的所有顶点，我们使用独热向量 f表示该顶点的接触语义标签。每个向量 f的长度为 $|f| = C + 1$，其中 C 是物体类别的数量。我们额外引入一个“无效”类别，用于表示没有接触的顶点。我们使用 F表示人体姿态中所有顶点的接触语义标签。

3.2 人体-场景接触预测

我们的数据集由一系列成对的顶点和接触语义标签组成 $\{(V_b^1,\ F^1),\ (V_b^2,\ F^2)\ldots\ldots,\ (V_b^n,\ F^n)\}$，其中 V_b^i表示人体顶点（图2(a)），F^i 表示第 i 帧的接触语义标签，且 n 表示可变的序列长度。我们首先训练一个条件变分自编码器（cVAE），以学习一个以顶点位置为条件的接触语义标签概率模型。然后，我们在 cVAE 上方部署 Transformer 层，以提高时间一致性。我们将该框架称为 ContactFormer。整体网络架构如图3所示。

接触语义预测。我们首先训练一个模型，为每个单独的姿态预测接触语义标签。给定一组人体顶点和接触语义标签（$V_b,\ F$），我们首先融合这两个部分：I_e = 拼接($V_b,\ F$)。我们将 I_e输入图神经网络（GNN）编码器 G_{Enc}，得到潜在高斯空间及其均值 H_μ和标准差 H_σ。然后，我们从潜在高斯空间中采样一个潜在向量 z，并将其与每个顶点的位置拼接：I_d = 拼接($V_b,\ z$)。我们将 I_d 输入

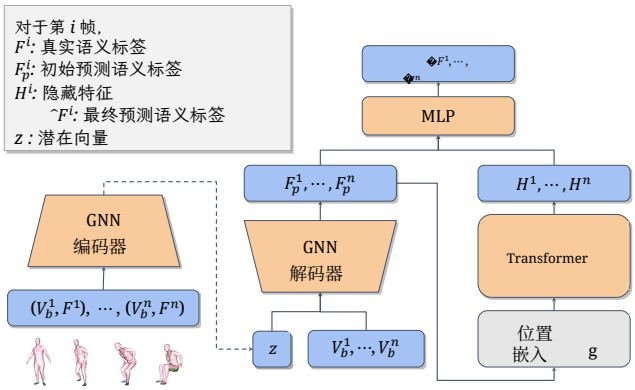

图3：ContactFormer的架构。我们首先使用基于GNN的变分自编码器对接触点进行编码。然后应用Transformer来改善时序信息融合。我们还在GNN解码器的输出中加入了正弦位置编码。

GNN解码器 G_{Dec} 以预测重建后的接触语义标签 F_p。请注意，编码器和解码器中的图神经网络均采用与Hassan等人 [2021b]相同的结构。第 k 层顶点 x 的每个顶点特征 h_x^k 都会更新为

$$h_x^k = \text{Linear}(\text{Concat}(\{h_{x'}^{k-1} : x' \in N(x)\})), \quad (1)$$

其中，$N(x)$表示在螺旋排序序列中，x的 m近邻顶点，正如 Gong等人 [2019]所提出的那样。

ContactFormer：我们训练了一个 Transformer，从姿态序列中提取时序信息，以提高预测的一致性，如图3所示。具体而言，给定从第1帧到 n帧的一组姿态和接触语义标签 $\{(V_b^1,\ F^1)\ldots\ldots(V_b^n,\ F^n)\}$，我们首先使用前一个模型，针对每一帧独立地重建接触语义标签 F_p^i i。然后，我们将每个 F_p^i 嵌入到隐特征空间中，并用正弦位置

嵌入对其进行增强，然后将其输入 Transformer 模块。Transformer 模块的输出是由 n 个向量组成的序列 $\{H_1$、……、$H_n\}$。对于每一帧 i，我们将 H_i 与初始预测 F_p^i 拼接，并使用多层感知机（MLP）得到最终预测 $\hat{F}^i$。最终预测如图2(c)所示。

训练： 我们通过以下损失函数优化模型参数：

$$\mathcal{L} = \mathcal{L}_{rec} + \alpha \cdot \mathcal{L}_{KL}, \tag{2}$$

其中，L_{rec} 是任意帧 i的真实语义标签 F^i 与模型预测$\hat{F}^i$ 之间的类别交叉熵（CCE）损失之和：

$$\mathcal{L}_{rec} = \sum_i \mathrm{CCE}(F^i, \hat{F}^i), \tag{3}$$

其中，L_{KL} 是潜在高斯空间与正态分布 N 之间的 Kullback-Leibler 散度损失：

$$\mathcal{L}_{KL} = KL(Q(z|F, V_b)||\mathcal{N}). \tag{4}$$

这里，我们使用 Q 表示 cVAE 中结合了重参数化技巧采样过程的编码器网络。受 Higgins等人 [2016],的启发，我们还将 L_{KL} 乘以权重 α，以控制重建准确率与多样性之间的平衡。

3.3 场景合成

接触物体恢复。 根据 ContactFormer 从每帧预测得到的累积接触点（图2(d)），我们进一步通过局部物体类别多数投票来减少空间预测噪声，如图4所示。随后，将每个预测物体类别的顶点聚类为可能的接触实例V_c，并使用该类别中所有物体边的最短长度作为聚类的 ϵ。实际操作中，我们对接触顶点进行下采样，以保证后续计算的可处理性。

然后，我们通过最小化以下损失来优化物体点云 V_o 的姿态：

$$\mathcal{L}(V_c, V_o) = \mathcal{L}_{contact} + \mathcal{L}_{pen}. \tag{5}$$

接触损失 $L_{contact}$ 定义为

$$\mathcal{L}_{contact} = \lambda_{contact} \frac{1}{|V_c|} \sum_{v_c \in V_c} \min_{v_o \in V_o} ||v_c - v_o||_2^2, \tag{6}$$

其中， $\lambda_{contact}$ 是一个可调超参数。该损失促使物体与预测的人体顶点接触。穿透损失 L^{pen} 定义为：

$$\mathcal{L}_{pen} = \lambda_{pen} \sum_{d_c^i < t} {d_c^i}^2, \tag{7}$$

其中， d_c^i 是物体与人体序列之间的有符号距离， t 是穿透距离阈值。该损失可防止物体穿透人体序列。

直观而言，这些损失促使物体与人体网格接触，但不会穿透人体网格。图2(e)展示了优化后的物体摆放示意图。为了改进计算

为提高计算效率，我们选择根据运动序列合并后的人体网格计算人体符号距离场。为了使不同物体的损失尺度保持一致，我们根据物体的大小选择采样点的数量。

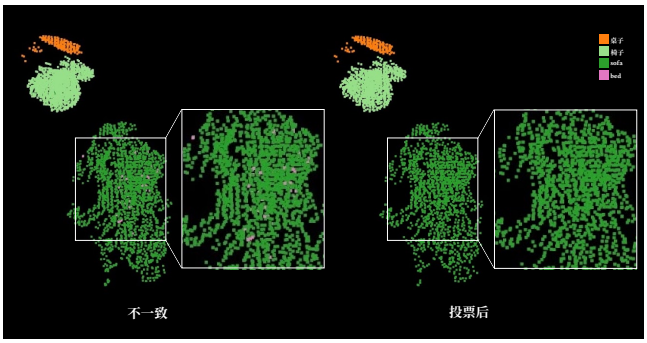

图4：局部多数投票示意图。从放大框中可以看到，原始接触点中存在多个不一致的点。粉色点表示语义标签"床"，绿色点表示标签"沙发"。我们通过加入多数投票来缓解这一问题。

受约束的场景补全。 为了获得完整场景，我们还将预测非接触物体，将其作为一项场景补全任务，并受3D人体轨迹和现有接触物体的约束。地板被划分为网格；如果脚部顶点或物体顶点与某个网格单元距离较近，则将该单元标记为已占用。考虑到房间中的家具类别构成一个序列，我们在3D-FRONT数据集 [Fu et al. 2021a] 上训练一个自回归Transformer模型。该模型以现有物体的类别为输入，并返回下一个潜在物体类别的概率分布。我们从该分布中采样一个类别，并将该类别的一个物体随机放置在未占用的地板网格上。为防止采样得到的物体穿入人体序列，我们进一步使用 L^{pen} 优化物体的平移和旋转（参见公式7）。

4 实验设置

本节介绍整个 SUMMON 框架所用的数据集及其实现细节。

4.1 数据集

我们使用 PROXD [Hassan等人 2019] 数据集训练 ContactFormer。PROXD 使用 RGB-D 相机捕捉了 20 名人类受试者与 12 个场景进行交互的过程。我们使用 SMPL-X 格式表示人体姿势，以重建人体网格。PROXD 中的姿态序列使用 SMPLify [Bogo 等人 2016] 估计，包含许多抖动。我们应用 LEMO [Zhang 等人2021], 学习到的时间运动平滑先验，以生成平滑的人体运动作为训练数据。我们的逐顶点接触语义标签真实标注通过场景有符号距离函数生成，并使用 PROX-E [Zhang 等，2020a], 提供的接触语义标签；PROX-E 通过人工标注具有预定义物体类别的场景网格扩展了 PROXD。我们将人场景接触定义为人体顶点与场景之间的有符号距离小于 0。05。

我们从 3D-FUTURE [Fu 等，2021b]中挑选物体，放置到场景中。3D-FUTURE 是一个包含按类别划分的家具三维模型及其原始尺寸的数据集。我们使用 3D-FUTURE 的一个精选子集，以缩短候选搜索时间。为了简化接触估计，并将预测的物体类别限制在我们的物体数据集所提供的类别范围内

我们将 PROX-E 数据集中的接触物体类别从 42 类缩减至 8 类。

我们使用 GIMO 数据集 [Zheng 等人 2022] 作为另一个测试数据集，以评估所提出方法在分布外数据上的泛化能力。

4.2 实现

对于编码器和解码器中的图神经网络，我们将隐藏层数设为 3。图神经网络中每个隐藏顶点特征的维度为 64。在图神经网络编码器中，我们在每个隐藏层之后将人体顶点下采样 4 倍。我们采用与先前工作 [Vaswani 等人 2017]类似的 Transformer 层架构。训练细节和超参数选择见补充材料。

为了比较不同架构提取时间信息的能力，我们还实现了将 MLP 和长短期记忆网络 [Greff 等人 2016] 模块用作 GNN 解码器顶层最终模块的模型。对于使用 MLP 模块的模型，我们在 GNN解码器的输出上沿顶点维度应用最大池化层。然后将结果输入 MLP 模块，以获得整个序列的嵌入。随后，通过线性投影将序列嵌入与 GNN解码器的输出融合，得到最终预测。对于使用长短期记忆网络模块的模型，我们将 GNN解码器的输出线性投影到更高维的嵌入空间，并将其输入双向长短期记忆网络层，以提取每帧的特征。然后将帧特征与 GNN解码器的输出拼接，得到最终语义标签。

接触物体恢复。 为了减少接触语义估计中的噪声，我们在带有 $\epsilon = 0.1$ 和 $minPts = 10$的点云簇中使用多数语义投票。在用于物体实例拟合的点云聚类中，考虑到不同类别的尺寸不同，我们对不同类别使用不同的 ϵ 值。

为了将物体以合适的高度放置到场景中，我们首先对所有与地板接触的人体顶点进行聚类。然后，取所有簇中位数的最小值作为估计的地面高度。接下来，我们平移物体，使其最低顶点落在地板上。

为避免陷入局部极小值，我们沿地面平面对平移进行网格搜索，并围绕向上轴对旋转进行网格搜索，以此为初始变换提供良好的起点。然后，我们对相同的
变换参数进行优化，并以网格搜索的结果为基础。在这两种情况下，我们使用不同的 $\lambda_{contact}$、λ_{pen} 和 t，以适应不同物体类别的属性。我们将损失最低的变换作为优化结果。

为实现场景多样性，我们考虑类间多样性和类内多样性。类间多样性是指人体运动可能与不同类别的物体发生交互。例如，坐下这一动作可以在椅子、床或沙发上完成。为实现这一点，我们首先根据 ContactFormer 预测的接触概率分布，对每个顶点的接触语义进行采样。在接触物体恢复的局部聚类过程中，我们将局部聚类簇中的类别标签视为概率分布，并对聚类接触类别进行采样。类内多样性是指人体运动可能与同一物体类别的不同实例发生交互。为

表 1：接触预测结果。我们使用重建准确率和一致性分数作为评估指标。我们的 ContactFormer 明显优于基线方法。

模型	重建准确率 ↑	一致性分数 ↑
MLP预测器	0.9082	0.8922
LSTM预测器	0.9087	0.9209
POSA	0.9106	0.8816
ContactFormer（我们的方法）	**0.9120**	**0.9518**

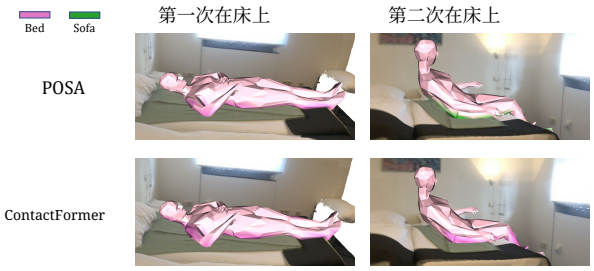

图 5：POSA 和我们的 ContactFormer 的接触预测结果可视化。左：人物躺在床上时，POSA 和 ContactFormer 的接触预测结果。右：人物在四处走动后再次躺在床上时，POSA 和 ContactFormer 的接触预测结果。人物第二次躺在床上时，ContactFormer 的一致性更好。

实现这一点，我们对该物体类别的所有实例进行网格搜索和优化。

5 评估

本节介绍接触预测和场景合成的评估指标、基线方法及结果。我们建议读者观看补充材料中的视频。

5.1 接触语义预测

基线方法。 我们与三种基线方法进行比较，包括 POSA [Hassan 等人 2021b],、一种使用基于多层感知机（MLP）的预测器的架构变体，以及一种融合时间信息并使用双向长短期记忆网络的变体 [Greff 等人 2016]。

评估指标。 我们使用两项评估指标来评估接触语义预测：重建准确率和一致性分数。重建准确率是通过计算每个顶点的预测标签与真实标签之间的平均正确率得出的。一致性分数的设计基于以下直觉：如果我们累积每一帧的预测接触点，那么彼此接近的接触点应具有一致的接触语义标签。因此，该损失的计算方式如下：给定一个姿态序列及累积的接触点，对于每个点，我们将其预测的接触标签与其邻近点的接触标签进行比较，以判断该预测是否与邻近点中占多数的接触标签一致（即一致性分数较高）。

结果。 表 1 展示了 PROXD 验证集上所有方法的重建准确率和一致性分数。我们发现，ContactFormer 在重建精度方面表现出色

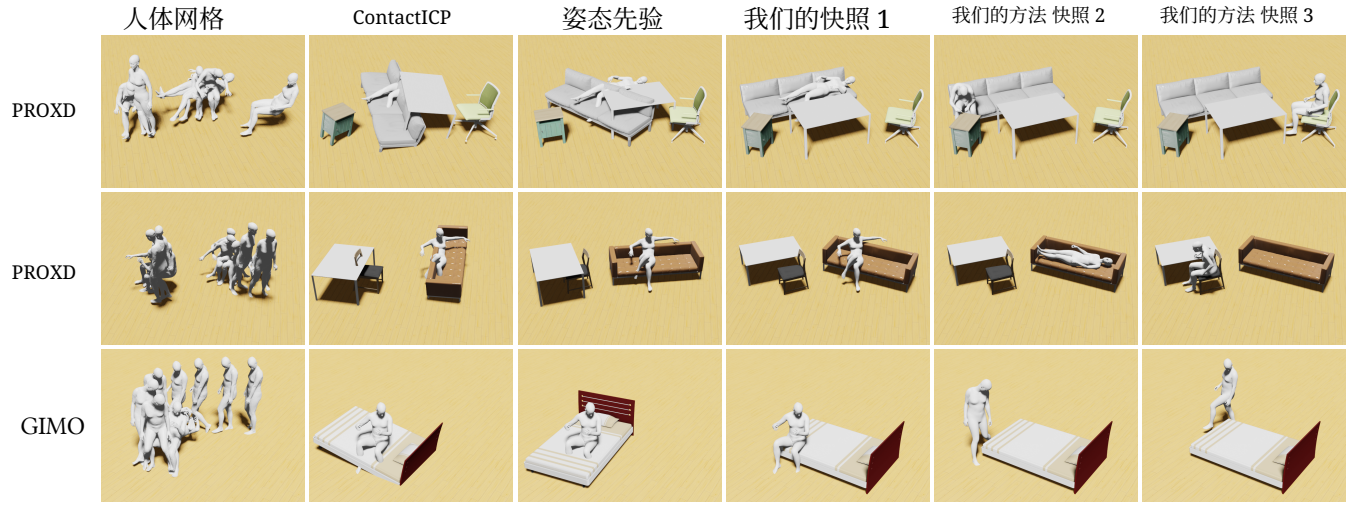

图6：基于人体轨迹生成的物体的可视化结果。人体轨迹来自 PROXD 数据集和未见过的 GIMO 数据集。第一列展示轨迹。第二列展示 ContactICP基线的结果。第三列展示姿态先验基线的结果。第四至第六列是 SUMMON 生成结果的快照。

，并且在一致性得分方面显著优于所有基线方法。这证明了基于Transformer的架构在预测时间上保持一致且准确的接触标签方面具有优越性。

图 5 展示了 ContactFormer 和 POSA 输出的接触标签。我们注意到，当人体第二次尝试躺在床上时，ContactFormer 会预测出一致的标签；而 POSA 由于缺乏时间信息，会预测出不同的标签。

用户研究。 我们开展了一项用户研究，以评估接触语义标签预测的质量，并比较 ContactFormer 与 POSA。对于验证数据集中的每个姿态序列，我们都会渲染一段视频，展示人体运动、预测的接触语义标签以及真实场景。我们通过以不同颜色渲染人体顶点周围的小块区域来展示预测的接触语义标签，颜色取决于相应的标签。每段视频的摄像机角度都经过设置，以便清晰呈现人体运动和语义标签。对于每个姿态序列，我们向人类受试者提出以下问题："哪段视频的接触标签预测看起来更合理？"在 22 名用户中，78.12% 的用户选择了 ContactFormer 而非 POSA，认为 ContactFormer 提供的结果更加合理且令人信服。这一结果与表 1 中的定量结果相呼应。

5.2 接触物体恢复

基线方法。 由于我们研究的问题是全新的，且没有现成的基线方法，因此我们自行设计了两种合理的基线方法：基于点对点 ICP [的接触信息引导的点云配准（ContactICP） 1992] ，以及基于髋部朝向并采用姿态先验的方法（Pose Pri- ors）。相关方法的详细信息见补充材料。

评估指标。 我们采用 Zhang 等人提出的无碰撞分数 [2020a]。该分数用于估算人体网格与场景物体之间的碰撞比例。由于所有方法，包括

表2：在平滑后的 PROXD 和未见过的 GIMO 数据集上进行接触物体恢复时的无碰撞分数。对于每个序列，分数取所有可能生成场景的平均值。分数越高越好。

方法	PROXD	GIMO
ContactICP	0.654	0.820
姿态先验	0.703	0.798
SUMMON（不进行优化）	0.815	0.937
SUMMON（我们的方法）	0.851	0.951

SUMMON，都会先将物体对齐到接触点的质心，因此接触约束自然得到满足。

结果。 对于每个序列，我们计算场景中所有物体的无碰撞得分的平均值。表 2 比较了经过平滑处理的 PROXD 数据集 [Zhang 等人 2021], 上的平均无碰撞得分；该数据集用于训练。此外，我们还在未见过的 GIMO 数据集 [Zheng 等人 2022], 上进行了比较，该数据集还提供了与场景交互的人体的 SMPL-X 参数。

图6展示了我们的方法与基线方法之间的对比。我们发现，SUMMON 能够合成物理上合理、语义上合乎情理的物体。Contac-tICP 通常会出现严重穿透，因为用于配准的接触点可能较为稀疏。虽然 Pose Priors 得到的物体位置和方向看起来可能正确，但它往往未能考虑物理约束。

图7展示了 SUMMON 根据同一人体运动轨迹生成的多种可能场景。我们发现，SUMMON 能够泛化到类内（例如外观各异的椅子）和类间（例如从沙发到床）的情况。补充材料中提供了更多示例。

人体用户研究。 我们遵循第 5.1 节中的相同流程。我们不展示接触预测，而是向用户展示

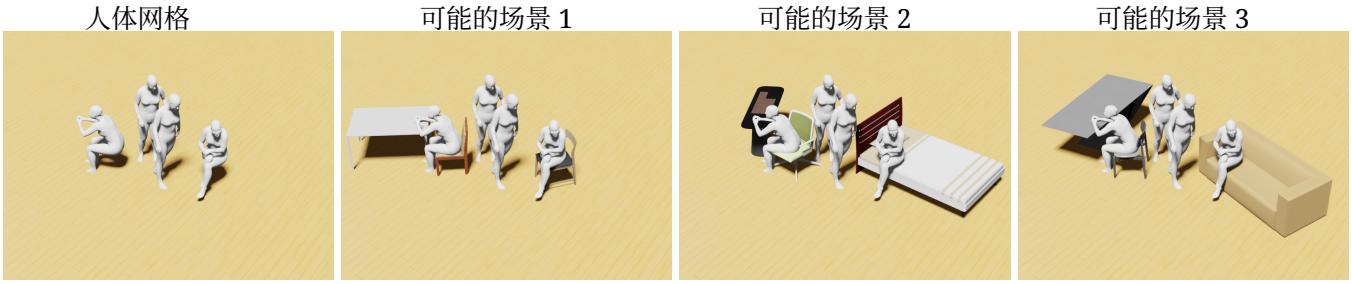

图7：基于相同的人体轨迹，SUMMON 生成的可能替代物体放置位置的可视化示例。在此示例中，接触物体可以是椅子、沙发或床，只要不违反物理约束即可。SUMMON 还可以在同一类别中生成不同的实例（例如，椅子）。

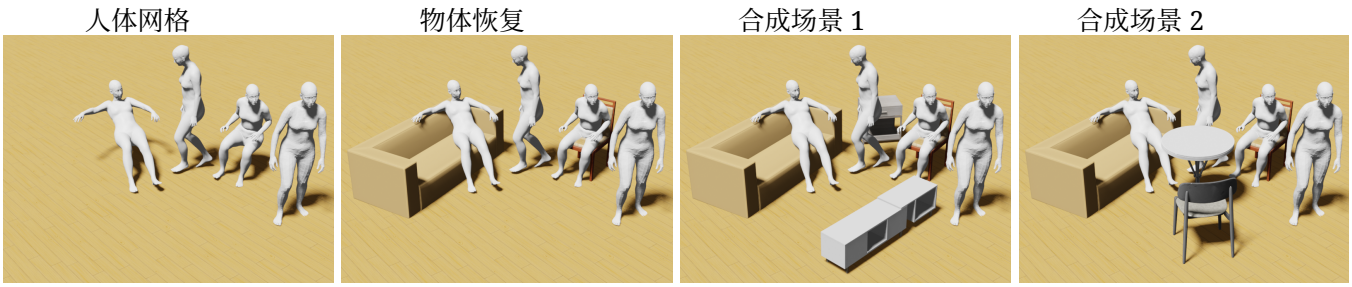

图 8：场景补全的可视化示例。基于所有接触物体和人体运动轨迹，SUMMON 现在会生成人体网格未接触的物体。虽然这些物体没有发生接触，但它们使场景更加完整，并为未来合成的人体运动序列与其他物体进行交互创造了可能。

场景中的动画人体运动序列和预测得到的物体，并请他们选择最合理、最逼真的摆放方式。根据统计结果，我们发现，74.5% 的用户选择了 SUMMON，而非 ContactICP 和姿态先验。我们发现，姿态先验的用户选择率为 23.5%，表明它在某些情况下能够产生合理的结果。

消融实验。 我们还对优化目标进行了消融实验。表 3 显示，穿透损失和接触损失对 SUMMON 都很重要。直观而言，穿透损失有助于避免物体发生碰撞，而接触损失有助于使物体保持在人体附近。我们同时使用无碰撞分数 和接触分数。接触分数 的计算方式为：场景中与人体轨迹接触的物体所占的比例（Zhang 等人 [2020a]）。

5.3 场景补全

为了生成完整的场景，我们按照 Paschalidou 等人 [2021] 在第 3.3 节中的做法，训练了另一个物体生成模型。该模型会输出一组可能的物体，这些物体不会与人体网格接触或发生穿透。利用该模型，我们生成了包含接触物体和非接触物体的更完整场景。可视化结果如图 8 所示。补全后的场景中增加了电视柜或茶几等物体。虽然这些物体与人体网格没有接触，但它们使场景在语义上更加真实。

表 3：损失项的消融研究。消融了穿透损失和接触损失。我们使用无碰撞分数和接触分数作为评估指标。

方法	无碰撞分数 ↑	接触分数 ↑
SUMMON	0.894	1
不使用穿透损失	0.656	1
不含接触损失	0.995	0.194

6 结论

我们提出了**场**景合成：基于人**类运**动**合**成（**SUMMON**），这是一个能够根据人类交互序列生成多物体场景的框架。SUMMON 利用人体接触估计和场景先验，生成能够真实支撑交互并符合语义上下文的场景。SUMMON 的灵活性还使其能够从单个运动序列合成多种不同场景。我们希望这也能为生成低成本、多样化的人-场景交互数据集带来启发。未来，我们有兴趣探索以下方向。由于 PROXD 未考虑软体交互，一个潜在方向是考虑床和沙发等物体的软体变形。我们的方法假设合成的场景是静止的，因此未来的工作可以考虑在人-场景交互过程中移动和重新布置家具。由于 PROXD 将书籍、杯子或电视遥控器等较小的交互物体都归为同一类别，我们方法的一个潜在扩展方向是纳入与更多特定小物体的交互。

致谢

这项工作部分得到了斯坦福以人为本人工智能研究院（HAI）、丰田研究院（TRI）、Innopeak、Meta、博世和三星的支持。

参考文献

P.J. Besl 和 Neil D. McKay。1992年。一种三维形状配准方法。*IEEE*模式分析与机器智能汇刊14, 2 (1992年), 239–256。Bharat Lal Bhatnagar、Xianghui Xie、Ilya A. Petrov、Cristian Sminchisescu、Chris- tian Theobalt 和 Gerard Pons-Moll。2022年。BEHAVE：用于跟踪人体与物体交互的数据集和方法。载于计算机视觉与模式识别会议（*CVPR*）。15935–15946。Bryce Blinn、Alexander Ding、Daniel Ritchie、R Kenny Jones、Srinath Sridhar 和 Manolis Savva。2021。学习身体感知三维形状生成模型。*arXiv*预印本 *arXiv:2112.07022* (2021)。Federica Bogo、Angjoo Kanazawa、Christoph Lassner、Peter Gehler、Javier Romero 和 Michael J. Black。2016年。善用 SMPL：从单幅图像自动估计三维人体姿态和形状。载于欧洲计算机视觉会议（*ECCV*）。561–578。Matthew Brand 和 Aaron Hertzmann。2000。风格机。载于计算机图形学与交互技术会议。183–192。Yixin Chen、Siyuan Huang、Tao Yuan、Siyuan Qi、Yixin Zhu 和 Song-Chun Zhu。2019年。整体++ 场景理解：结合人-物交互和物理常识的单视图三维整体场景解析与人体姿态估计。载于国际计算机视觉会议（*ICCV*）。8648–8657。Ching-Yao Chuang、Jiaman Li、Antonio Torralba 和 Sanja Fidler。2018年。学会恰当行动：从图像预测并解释可供性。载于计算机视觉与模式识别会议（*CVPR*）。975–983。Vincent Delaitre、David F Fouhey、Ivan Laptev、Josef Sivic、Abhinav Gupta 和 Alexei A Efros。2012。通过长期观察人来理解场景语义。载于欧洲计算机视觉会议（*ECCV*）。284–298。David F Fouhey、Vincent Delaitre、Abhinav Gupta、Alexei A Efros、Ivan Laptev 和 Josef Sivic。2012。观察人：人体动作作为单视图几何的线索。载于欧洲计算机视觉会议（*ECCV*）。732–745。Huan Fu、Bowen Cai、Lin Gao、Ling-Xiao Zhang、Jiaming Wang、Cao Li、Qixun Zeng、Chengyue Sun、Rongfei Jia、Binqiang Zhao 等。2021a。3D-FRONT：具有布局和语义的三维带家具房间。载于国际计算机视觉会议（*ICCV*）。10933–10942。Huan Fu、Rongfei Jia、Lin Gao、Mingming Gong、Binqiang Zhao、Steve Maybank 和 Dacheng Tao。2021b。3D-FUTURE：具有纹理的三维家具形状。国际计算机视觉期刊（*IJCV*）129, 12 (2021), 3313–3337。Shunwang Gong、Lei Chen、Michael Bronstein 和 Stefanos Zafeiriou。2019。Spiralnet++：一种快速且高效的网格卷积算子。载于国际计算机视觉会议研讨会（*ICCVW*）。4141–4148。Klaus Greff、Rupesh K Srivastava、Jan Koutník、Bas R Steunebrink 和 Jürgen Schmid- huber。2016年。长短期记忆网络：搜索空间奥德赛。*IEEE*神经网络与学习系统汇刊28, 10 (2016), 2222–2232。Abhinav Gupta、Scott Satkin、Alexei A Efros 和 Martial Hebert。2011。从三维场景几何到人体工作空间。载于计算机视觉与模式识别会议（*CVPR*）。1961–1968。Félix G Harvey、Mike Yurick、Derek Nowrouzezahrai 和 Christopher Pal。2020年。鲁棒的动作补间。*ACM*图形学汇刊（*TOG*）39, 4 (2020), 60–1。Mohamed Hassan、Duygu Ceylan、Ruben Villegas、Jun Saito、Jimei Yang、Yi Zhou 和 Michael Black。2021a。随机场景感知动作预测。载于国际计算机视觉会议（*ICCV*）。11354–11364。Mohamed Hassan、Vasileios Choutas、Dimitrios Tzionas 和 Michael J Black。2019年。利用三维场景约束消解三维人体姿态歧义。载于国际计算机视觉会议（*ICCV*）。2282–2292。Mohamed Hassan、Partha Ghosh、Joachim Tesch、Dimitrios Tzionas 和 Michael J. Black。2021b。通过学习人-场景交互来构建三维场景。载于计算机视觉与模式识别会议（*CVPR*）。14708–14718。Irina Higgins、Loic Matthey、Arka Pal、Christopher Burgess、Xavier Glorot、Matthew Botvinick、Shakir Mohamed 和 Alexander Lerchner。2016年。β变分自编码器：通过受约束的变分框架学习基本视觉概念。载于国际学习表征会议（*ICLR*）。Daniel Holden、Jun Saito 和 Taku Komura。2016年。用于角色动作合成与编辑的深度学习框架。*ACM*图形学汇刊（*TOG*）35, 4 (2016), 1–11。Lucas Kovar 和 Michael Gleicher。2003年。借助配准曲线实现灵活的自动动作混合。载于计算机动画研讨会（*SCA*）。214–224。Manyi Li、Akshay Gadi Patil、Kai Xu、Siddhartha Chaudhuri、Owais Khan、Ariel Shamir、Changhe Tu、Baoquan Chen、Daniel Cohen-Or 和 Hao Zhang。2019b。Grains：用于室内场景的生成式递归自编码器。*ACM*图形学汇刊（*TOG*）38, 2 (2019), 1–16。Xueting Li、Sifei Liu、Kihwan Kim、Xiaolong Wang、Ming-Hsuan Yang 和 Jan Kautz。2019a。《将人置于场景中：在三维室内环境中学习可供性》。收录于计算机视觉与模式识别会议（*CVPR*）。12360–12368。Andrew Luo、Zhoutong Zhang、Jiajun Wu 和 Joshua B Tenenbaum。2020。《场景布局的端到端优化》。收录于计算机视觉与模式识别会议（*CVPR*）。3754–3763。Naureen Mahmood、Nima Ghorbani、Nikolaus F Troje、Gerard Pons-Moll 和 Michael J Black。2019。AMASS：以表面形状表示的动作捕捉档案。收录于国际计算机视觉会议（*ICCV*）。5442–5451。Yinyu Nie、Angela Dai、Xiaoguang Han 和 Matthias Nießner。2022。Pose2Room：从人类活动中理解三维场景。收录于欧洲计算机视觉会议（*ECCV*）。Sang Il Park、Hyun Joon Shin 和 Sung Yong Shin。2002。基于动作混合的在线运动生成。收录于计算机动画研讨会（*SCA*）。105–111。Despoina Paschalidou、Amlan Kar、Maria Shugrina、Karsten Kreis、Andreas Geiger 和 Sanja Fidler。2021。ATISS：用于室内场景合成的自回归变换器。收录于神经信息处理系统进展（*NeurIPS*）。12013–12026。Georgios Pavlakos、Vasileios Choutas、Nima Ghorbani、Timo Bolkart、Ahmed A. A. Osman、Dimitrios Tzionas 和 Michael J. Black。2019。富表现力的人体捕捉：从单幅图像中捕捉三维手部、面部和身体。收录于计算机视觉与模式识别会议（*CVPR*）。10975–10985。Xavier Puig、Kevin Ra、Marko Boben、Jiaman Li、Tingwu Wang、Sanja Fidler 和 Antonio Torralba。2018。Virtualhome：通过程序模拟家庭活动。收录于计算机视觉与模式识别会议（*CVPR*）。8494–8502。Pulak Purkait、Christopher Zach 和 Ian Reid。2020。Sg-vae：用于生成新室内场景的场景文法变分自编码器。收录于欧洲计算机视觉会议（*ECCV*）。155–171。Davis Rempe、Tolga Birdal、Aaron Hertzmann、Jimei Yang、Srinath Sridhar 和 Leonidas J Guibas。2021。Humor：用于鲁棒姿态估计的三维人体运动模型。收录于国际计算机视觉会议（*ICCV*）。11488–11499。Daniel Ritchie、Kai Wang 和 Yu-an Lin。2019。通过深度卷积生成模型实现快速灵活的室内场景合成。收录于计算机视觉与模式识别会议（*CVPR*）。6182–6190。Manolis Savva、Angel X Chang、Pat Hanrahan、Matthew Fisher 和 Matthias Nießner。2016。Pigraphs：从观测中学习交互快照。*ACM*图形学汇刊（*TOG*）35，4（2016），1–12。Roberta Spallone。2015。已拆除建筑杰作的数字重建、三维建模与动画：Mollino 都灵赛马场案例研究。收录于建筑测绘、建模与表达新兴数字工具研究手册（2015），476–509。Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017。注意力就是你所需要的一切。收录于神经信息处理系统进展（*NIPS*）。5998–6008。Jiashun Wang、Huazhe Xu、Jingwei Xu、Sifei Liu 和 Xiaolong Wang。2021a。合成长期三维人体运动及其在三维场景中的交互。收录于计算机视觉与模式识别会议（*CVPR*）。9401–9411。Kai Wang、Yu-An Lin、Ben Weissmann、Manolis Savva、Angel X Chang 和 Daniel Ritchie。2019b。Planit：借助关系图和空间先验网络规划并实例化室内场景。*ACM*图形学汇刊（*TOG*）38，4（2019），1–15。Xiaolong Wang、Rohit Girdhar 和 Abhinav Gupta。2017。追剧：从情景喜剧中扩展可供性学习。收录于计算机视觉与模式识别会议（*CVPR*）。2596–2605。Xinpeng Wang、Chandan Yeshwanth 和 Matthias Nießner。2021b。Sceneformer：利用变换器生成室内场景。收录于国际三维视觉会议（*3DV*）。106–115。Zhe Wang、Liyan Chen、Shaurya Rathore、Daeyun Shin 和 Charless Fowlkes。2019a。几何姿态可供性：结合场景约束的三维人体姿态。*arXiv*预印本 *arXiv:1905.07718* （2019）。Jingwei Xu、Huazhe Xu、Bingbing Ni、Xiaokang Yang、Xiaolong Wang 和 Trevor Darrell。2020。用于动作合成的分层风格网络。收录于欧洲计算机视觉会议（*ECCV*）。178–194。Hongwei Yi、Chun-Hao P Huang、Dimitrios Tzionas、Muhammed Kocabas、Mohamed Hassan、Siyu Tang、Justus Thies 和 Michael J Black。2022。用于视觉环境重建的感知人体的物体摆放。收录于计算机视觉与模式识别会议（*CVPR*）。3959–3970。Siwei Zhang、Yan Zhang、Federica Bogo、Marc Pollefeys 和 Siyu Tang。2021。学习用于三维场景中四维人体捕捉的运动先验。收录于国际计算机视觉会议（*ICCV*）。11343–11353。Song-Hai Zhang、Shao-Kui Zhang、Wei-Yu Xie、Cheng-Yang Luo 和 Hong-Bo Fu。2020b。通过离散且精确的布局模式提取快速合成三维室内场景。*arXiv*预印本 *arXiv:2002.00328* （2020）。Yan Zhang、Mohamed Hassan、Heiko Neumann、Michael J Black 和 Siyu Tang。2020a。生成人员缺席场景中的三维人物。收录于计算机视觉与模式识别会议（*CVPR*）。6194–6204。Yang Zheng、Yanchao Yang、Kaichun Mo、Jiaman Li、Tao Yu、Yebin Liu、Karen Liu 和 Leonidas Guibas。2022。GIMO：上下文中的注视信息引导的人体运动预测。收录于欧洲计算机视觉会议（*ECCV*）。

Yang Zhou、Zachary While 和 Evangelos Kalogerakis。2019。Scenegraphnet：用于三维室内场景增强的神经消息传递。收录于国际计算机视觉会议（*ICCV*）。7384–7392。Yuke Zhu、Alireza Fathi 和 李飞飞。2014。基于知识库表示推理物体可供性。收录于欧洲计算机视觉会议（*ECCV*）。408–424。