专题综述

CAST：从 RGB 图像重建组件对齐的 3D 场景

姚凯欣[*1,2] 张龙文[*1,2] 严新豪[1,2] 曾焱[1,2] 张启煊[†1,2] 杨卫[3] 许岚[‡1] 顾家远[‡1] 虞晶怡[‡1]
1. 上海科技大学 2. 影眸科技 3. 华中科技大学

本文是上海科技大学、影眸科技和华中科技大学团队合作研究的成果，发表在SIGGRAPH 2025并获得了最佳论文奖。从单个RGB图像中恢复高质量的3D场景是计算机图形学中的一项挑战性任务。现有方法通常受限于特定领域或低质量的物体生成。为了解决这些问题，论文提出了CAST（**C**omponent-**A**ligned 3D **S**cene reconsTruction from a Single RGB Image），一种用于3D场景重建的新方法（图1）。CAST首先从输入图像中提取物体级别的2D分割和相对深度信息，然后使用基于GPT的模型分析物体间的空间关系。这使得场景中物体如何相互关联能够被理解，从而确保更整体一致的重建。CAST接着采用一个具备遮挡感知的大规模3D生成模型，独立生成每个物体的完整几何形状，使用Masked Auto Encoder（MAE）和点云条件来减轻遮挡和部分物体信息的影响，确保与源图像的几何形状和纹理准确对齐。为了使每个物体与场景对齐，姿态对齐生成模型计算所需的变换，使得单个生成的物体网格（mesh）能够准确放置并融入到场景的点云中。最后，CAST采用一个基于物理的校正机制，利用细粒度关系图生成约束图，指导物体姿态的优化，确保物理一致性和空间一致性。通过利用有符号距离场（SDF），该方法有效解决了遮挡、物体穿透和浮动物体等问题，确保生成的场景准确反映真实的物理交互。实验结果表明：CAST显著提升了单图像3D场景重建的质量，在场景理解和重建任务中提供了增强的真实感和准确性。CAST具有实际应用价值，在虚拟内容创作中，例如沉浸式游戏环境和电影制作，可以将现实世界无缝集成到虚拟景观中。此外，CAST还可以用于机器人领域，实现高效的真实到模拟工作流，并为机器人系统提供真实、可扩展的模拟环境。

一、导言

人类生活在清晰的关系网络中——家庭、朋友、同事——这些网络指导我们的决策和行为。这些联系塑造了我们的世界并赋予其结构。类似地，空间中的物体也在其自身网络中发挥作用[42]，但较少被注意到。它们并非孤立存在；其放置、设计和材料源于物理限制、功能作用和人类设计意图，并影响我们移动、互动和感知空

图 1　CAST 从单张图像中将多样化的 3D 场景生动呈现，展现出物体间丰富的物理和空间相互作用关系。

间的方式。例如，椅子靠在桌子上以获得支撑，杯子放在碟子上，台灯的光线与周围表面相互作用，投射出塑造整个场景的阴影。识别这些关系对于准确的场景解析、建模以及最近的 3D 生成至关重要，确保虚拟环境感觉真实和与现实世界一致。

在从文本或图像提示生成单个物体方面取得了显著进展。神经渲染方法[62,76]优化了隐式表示，而原生 3D 生成器[80,92]则通过端到端学习直接创建 3D 形状和纹理。虽然这些方法在单个物体方面显示出前景，但将它们应用于通过组合物体来生成整个场景时，面临显著的挑战。一个关键挑战是准确的姿态估计。现有方法通常假设物体是视图对齐的，这在真实世界场景中很少见。物体可能以不同的方向出现，受设计、物理或部分遮挡的限制。然而，大多数现有方法优先考虑几何保真度而非姿态对齐，使这一关键方面未得到充分探索。

一个更根本的问题源于物体间空间关系的缺乏。即使姿态相对准确，生成的场景也常常出现物理上不合理的瑕疵：物体相互穿透、漂浮或未能进行必要的接触。这些错误源于缺乏自然地将物体绑定在一起的空间和物理约束，就像人类关系构建我们的社会一样。虽然最近的一些方法[46,93]隐式编码空间关系，使用编码器-解码器架构，但它们仍局限于室内场景等特定领域。其他场景级生成器[21]将物体放置在全局坐标系中，但忽略了它们的相对姿态和依赖关系，进一步损害了真实性和下游应用的可用性，如编辑、动画和模拟。

为此，我们提出了 CAST，一种用于从单个 RGB 图像生成与图像对齐的组件的 3D 场景重建方法。CAST 为单个物体生成高质量的 3D 网格，并生成它们的相似变换（旋转、平移、缩放），确保与参考图像对齐并强制执行物理上合理的相互依赖关系。CAST 首先通过使用 2D 基础模型（例如 Florence-2[81]、GroundingDINO[49]、SAM[64]、Grounded-SAM[65]）处理非结构化 RGB 图像，以开放词汇方式识别、定位和分割物体。现有的单目深度估计器[75]提供部分 3D 点云和物体间空间关系的初步估计，包括相对变换和尺度。

CAST 的第一个核心组件是 3D 实例生成器，它包含两个模块：一个具备遮挡感知的物体生成模块和一个姿态对齐生成模块。物体生成模块采用基于潜在扩散的生成模型，根据分割出的部分图像（和点云）生成高保真物体网格。该模块包含一个遮挡感知 2D 图像编码器，能够推断被遮挡区域，确保稳健地从图像条件中提取特征。为了提高对真实世界点云条件的鲁棒性，我们在训练过程中模拟了带遮挡区域的部分点云，使模型能够有效处理遮挡。姿态对齐模块采用一个生成模型，生成一个变换后的部分点云，与潜在空间中隐式表示的完整几何形状对齐。相似变换是通过生成的变换点云和从相机估计的部分点云推导出来的。与直接姿态回归方法[35,39]不同，我们的方法通过生成来估计变换，捕捉了姿态对齐的多模态性质。

CAST 的第二个核心组件解决了物体间空间关系问题。尽管像素层面进行了对齐，但如果缺乏对物理约束的显式建模，仍可能出现物理上的不合理，如穿透或漂浮。CAST 引入了一个基于物理的校正过程，以确保空间和物理一致性。GPT-4v[1]被用于识别源图像的常识性物理关系，然后利用这些约束来优化物体姿态。这个过程确保重建的场景表现出真实的物理依赖关系，使其适用于模拟、编辑和渲染等应用。

CAST 可以从各种图像生成真实的 3D 场景，无论这些图像是来源于室内外环境、真实世界拍摄还是 AI 生成。与之前的方法[18,46]不同，CAST 通过精心设计的管线，支持开放词汇重建，甚至支持具有挑战性的室外图像。定量上，CAST 在室内数据集 3D-Front[22]上超越了基线方法，在物体级和场景级几何质量方面表现出色。另外，通过视觉语言模型和用户研究进行验证，它在各种图像上表现出优越的感知和物理真实性。

仅凭一张图像，CAST 就能真实地重建场景，包括细致的几何形状、生动的物体纹理，以及更重要的，物体之间的空间和物理相互依赖关系。这一能力使虚拟创作大众化：一个房间或室外空间的单一快照变成了一个实例化的 3D 环境，包含姿态精确的物体，交互自然，并考虑了遮挡。游戏开发者可以将真实世界的设置集成到沉浸式景观中，电影制作人可以毫不费力地生成复杂的虚拟场景——释放创意潜力。除了娱乐之外，CAST 还为更智能的机器人铺平了道路。通过使机器人研究人

员能够从真实世界演示数据中构建数字副本，它能够促进真实到模拟的流程[44,72]，从而实现更高效、可扩展的物理模拟工作流。

二、相关工作

将真实世界场景转换为数字领域，增强了我们理解、重建和与我们周围 3D 世界互动的能力。这种做法在动画、电影、游戏、建筑和制造等行业中得到广泛应用。它使得沉浸式电影体验、历史文物的数字保存以及交互式游戏环境的开发成为可能。例如，詹姆斯·卡梅隆在《阿凡达》（2009）中采用了开创性的 3D 扫描技术，将潘多拉郁郁葱葱、真实的生境带入生活。同样，在游戏行业中，《巫师 3：狂猎》融合了受波兰真实世界地点启发的逼真地形和建筑，将真实的文化和自然元素与富有想象力的开放世界探索相结合。

摄影测量是一种广泛使用的方法，可以高细节地捕捉物理世界并将其转换为数字形式[3,11,27,36,57,58]，但它需要数十到数百张来自多个视角的图像，这既耗时、资源密集，又难以扩展。相比之下，基于单图像的方法更高效、可扩展，只需要一张图像即可轻松从在线存储库获取，无需昂贵的扫描设备或多视图设置。

2.1 单图像场景重建

从单个图像进行场景级重建面临物体多样性、遮挡以及保持空间关系的挑战。一个例子是单目深度估计，即从单个图像推断深度，进一步得到深度点云[7,61,75,85,87]。虽然这提供了有价值的信息，但它在处理遮挡和场景隐藏部分时会遇到困难。为了解决这个问题，新视图合成方法使用辐射场[71,88]和 3D 高斯[67,68]等表示来学习 3D 数据集中的遮挡先验[10,17,25,66]。尽管取得了进展，单目重建方法仍然难以提供详细而精确的场景表示。

有些方法侧重于直接回归场景中的几何形状及其语义标签[12,15,16,26]。这些方法通常依赖于带有物体真值标注的场景数据集，例如 Matterport3D[22]和 3DFront[22]。这些数据集规模通常较小且仅限于室内房间环境。然而，这些方法的前馈性质导致生成的几何形状通常缺乏足够的细节和质量。

为了更好地将真实世界场景转化为数字，其他方法转向基于检索的方法[18,23,28,38,41]，通过在场景中搜索并替换相似物体来提高场景质量。这些方法结合了 GPT-4[1]、SAM[37,65]和深度先验等先进工具来分解场景。虽然这些方法通过集成真实世界物体提高了场景的真实感，但它们受限于所依赖数据集的丰富性和范围。对于超出数据集领域限制的场景，基于检索的方法要么产生错误结果，要么无法找到合适的替换，显著降低了重建场景的质量。

2.2 重建即生成

随着该领域的不断进步，从各种开放词汇图像或文本提示创建高质量 3D 数字资产的能力显著提高。这一进步促使范式转变，将单视图重建问题演变为生成式 3D 合成框架。这种范式转变允许生成 3D 资产而不受限于固定数据集，从而实现更灵活和可扩展的场景重建。

目前大部分 3D 资产生成研究都集中在从 2D 图像生成模型中提取 3D 几何形状[62,70,76]。最近的发展通过纳入多视图图像进行监督[47,48,51,52,74,78]，通常在大型物体数据集（如 Objaverse[20]）上进行训练以增强生成过程中的视图一致性。一些方法根据输入图像直接回归单个物体的形状和外观[32,69]。虽然这些方法取得了令人满意的视觉结果，但它们经常无法再现精细的几何细节。为了提高 3D 几何质量，越来越多工作已完全脱离 2D 监督，转而直接在 3D 资产上进行训练[19,20]。这些方法通过先进的处理技术[80,92,94]生成高质量的物体级几何形状。然而，这些方法侧重于孤立的物体，未能解决场景级挑战，如建模空间层次结构、物体间关系和环境光照。由于建模物体关系、光照和材料的表示复杂性高，场景生成仍处于不发达状态。尽管取得进展，现有方法仍难以生成完全实现、可编辑的 3D 场景。现有范式要么使用视频扩散模型[8,30,31]生成可导航的 2D 投影[9, 89]，要么依赖扩散先验通过 3D 高斯飞溅[24,45,79]进行体素场景近似。虽然这些方法产生吸睛的视觉效果，但它们与传统生产管线不兼容，缺乏可编辑网格、UV 映射和可分解的 PBR 材料。

一种更可行的方法是将场景分解为模块化组件——物体、背景和环境，并将它们生成和重新组合成可编

辑的场景图，以实现更大的灵活性和精度。例如，Gen3DSR[21]使用 DreamGaussian[70]进行开放词汇重建。然而，它在处理遮挡、姿态估计和编辑单个物体方面存在困难，同时依赖 2D 模型导致几何细节差和低保真表示。另一项最近的工作，Midi[33]，学习场景中物体间的空间关系，但需要基于真值 3D 网格和标注的数据集进行训练。这种对特定数据集的依赖限制了其可扩展性和对任意场景的泛化能力。

我们的方法与经典的分析-合成方法[91]共享概念基础，因为两者都旨在通过生成对观察到的图像的解释来推断 3D 结构。然而，分析-合成依赖于迭代渲染和像素级优化，而我们的方法利用预训练的生成模型和学习的先验直接合成可信的 3D 场景，通常绕过显式渲染和优化循环，从而提高了可扩展性、效率和对开放世界场景的适配性。

在此基础上，我们提出了一种新颖的场景重建管线，独立生成每个物体并将其对齐到整体的场景中。与现有方法不同，我们的方法保留了准确的几何形状、纹理和一致的空间关系，从而产生了更真实、可靠和可编辑的重建，提高了质量和灵活性。

2.3 具备物理感知的 3D 建模

生成物理上合理的 3D 资产对于确保动画、游戏和机器人等应用中的真实感和功能性至关重要。虽然最近的 3D 生成模型在创建视觉上真实的物体方面表现出色，但它们通常无法达到物理合理性。为了解决这一限制，物理感知 3D 生成模型被开发出来，将物理原理集成到生成过程中。有些方法使用软体模拟来动画化 3D 高斯[82, 96]，或者通过基于物理的惩罚来指导生成关节物体[50]。而另一些方法则通过刚体模拟[13,55,56]或 FEM[29,83]确保自支撑结构。这些方法利用离线或在线物理模拟来检查生成形状的物理有效性，进而指导生成。然而，这些方法通常局限于单个物体，忽略了场景中多个物体之间的相互影响。

将物理约束纳入场景合成更具挑战性，因为涉及更复杂的物体间接触等关系。Yang 等人[86]将物体碰撞、房间布局和物体可达性等约束集成到其场景级生成管线中。然而，它仅限于室内场景合成，并依赖封闭词汇数据库执行形状检索。Ni 等人[59]解决了多视图神经重建中物理不合理性问题。它利用可微分渲染和物理模拟来学习隐式表示。然而，它需要多视图图像作为输入，专注于单个物体，并且主要只解决稳定性问题。相比之下，我们的方法在开放词汇设置下运行，只需要一张输入图像。此外，它考虑了更复杂的物体间关系，特别是支撑和接触，使其更通用，适用于不同的场景。

三、方法概述

从单个图像进行场景级重建是计算机图形学中的一个基本挑战，在动画、虚拟现实和互动游戏中有广泛应用。与专注于孤立物体的物体级重建不同，场景级重建强调多个实体在真实（或风格化）物理下的排列和关系。通过捕捉每个物体的结构、空间关系和上下文线索，这种整体方法能够实现更沉浸的体验、引人入胜的叙事和高效的工作流程。尽管之前的方法探索了使用固定 3D 模板的前馈管线或基于检索的方法[18, 46]，但这些方法往往难以捕捉细微的场景语义和复杂的物体关系。针对这些局限，我们提出了一种以生成驱动的场景重建方法，强调物体关系，从单个未标注的 RGB 图像构建高保真、上下文一致的 3D 环境，无论是来源于真实世界摄影还是合成数据（见图 2）。

我们方法的关键在于对场景上下文信息进行全面的物体关系分析。首先，我们进行物体分割以识别和定位图像中的组成物体。然后，我们获取初步的几何信息（点云），并探索物体间的语义和空间关系。这些预处理结果，作为上下文，为我们后续的物体级生成管线提供了信息，确保每个重建的物体不仅保持其几何保真度，而且在场景中位置正确。最后，我们合成一个考虑物理合理性的整体 3D 环境，实现结构合理的布局和场景元素间真实的交互。

我们的研究聚焦两个主要目标：探索生成模型如何有效捕捉复杂的物体间关系，以从单个图像生成逼真、场景级重建；以及识别整合几何线索和上下文信息的策略，以最大化 3D 重建的准确性和合理性。通过本文研究，我们证明生成方法提供了比传统前馈和基于检索的技术更灵活和稳健的替代方案。CAST 允许对物体级细节和全局场景构成进行细粒度控制，从而简化了动画、

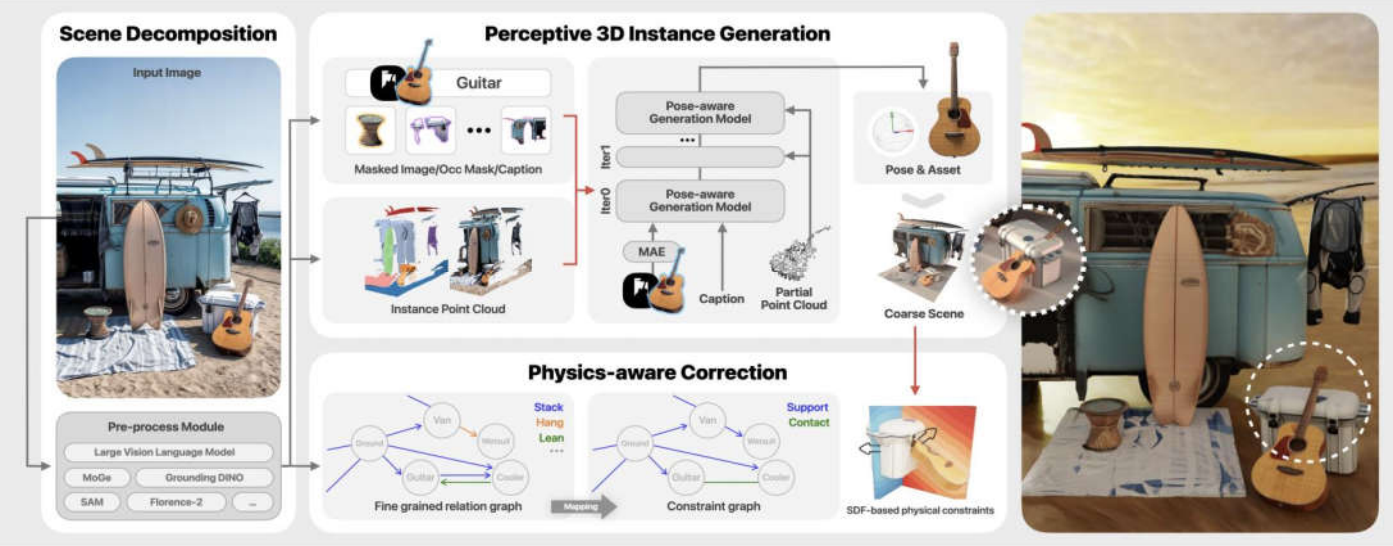

图 2　CAST 管线的概述。首先通过场景分析提取 RGB 输入图像的关键信息，然后通过具备姿态感知的生成过程创建初始 3D 模型。最后基于物理约束的优化促进真实的交互和空间关系，从而生成高质量的网格化 3D 场景。

游戏开发和其他需要 3D 模型的领域的内容创建管线。本文突出了以生成为中心的框架的优势，并为 3D 场景重建的未来发展奠定了基础。它还强调了上下文驱动方法在弥合 2D 图像与沉浸式、交互式虚拟环境之间差距方面日益增长的重要性。

预处理 为了辅助从单个图像进行全面的场景重建，我们首先进行全方位的语义提取，为后续处理提供坚实基础。具体而言，我们使用 Florence-2[81]识别物体，生成它们的描述，并用边界框定位每个物体。然后，我们利用 GPT-4v[1]过滤掉虚假检测，并分离出有意义的组成物体，从而实现不受预定义类别限制的开放词汇物体识别。接下来，我们使用 GroundedSAM-v2[65]为每个识别出的物体$\{o_i\}$生成精细的分割掩模$\{M_i\}$，从而获得精确的物体边界和相应的遮挡掩模，这在物体生成阶段起着关键的辅助作用。除了语义线索，我们还通过提取场景级点云来整合几何信息。我们使用 MoGe[75]生成像素对齐的点云$\{q_i\}$，用于每个物体$\{o_i\}$，　$i \in \{1, \ldots, N\}$，以及场景坐标系中的全局相机参数。这些额外的几何数据随后与每个物体的分割掩模匹配，为最终的 3D 场景重建提供了可靠的结构参考。

四、感知3D实例生成

在从单个 RGB 图像重建高保真 3D 场景的尝试中，一种简单粗暴的方法是使用单图像深度估计或扩散先验等技术直接生成整个场景网格。然而，由于真实世界场景的复杂和交织性质，这种方法在处理遮挡、渲染不可见组件以及准确表示物体关系方面固有地存在困难。因此，如图 3 所示，我们的方法不是直接生成整个场景网格，而是专注于单个物体生成，然后通过精确的对齐来排列物体。这种策略具有以下几个优点：1. 专注于单个物体可确保更高的几何保真度，并允许进行详细建模，从而产生更准确和视觉吸引力的场景组件。2. 在规范化空间中操作可确保生成的资产符合标准化方向和比例，与艺术家定义的坐标系统无缝集成，并促进数字内容创作工具之间的一致性。3. 模块化方法支持各种应用，如编辑、渲染和模拟，从而实现对物体进行独立操作，以获得更大的灵活性和效率。通过将场景重建分解为物体级生成和对齐，我们的方法提高了资产质量和可管理性，同时增强了 3D 场景的整体一致性和功能性。这种方法解决了几何精度和高效后处理等挑战，推动了单图像 3D 场景生成的发展。

场景中的物体级生成同样面临重大挑战，主要原因是场景中物体部分被遮挡以及传感器覆盖范围有限。此外，现有生成方法往往无法协调多个物体，导致场景不一致和不真实。为了克服这些限制，我们提出了一种具备遮挡感知的 3D 物体生成框架，将部分观察结果与全面的场景理解相结合。具体而言，给定图像及其点云，该框架生成一个高质量的 3D 资产，不仅与输入图像相

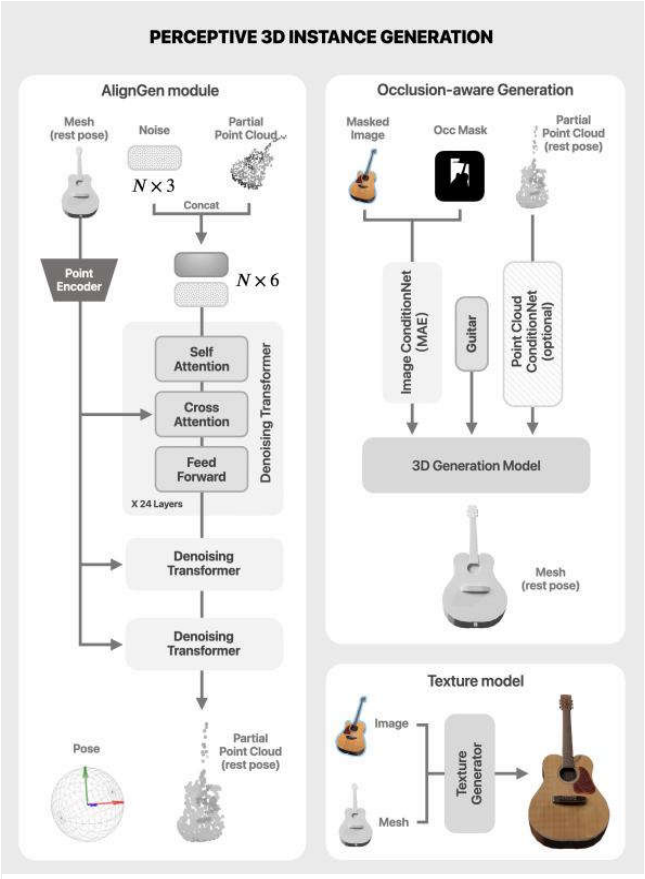

图 3　对齐生成模型（第 4.2 节）、遮挡感知物体生成模型（第 4.1 节）以及纹理生成模型的示意图

似，而且还与规范空间中对应的部分点云准确对齐。此外，我们计算一个变换矩阵，将生成的物体从其规范空间映射回原始场景空间，确保场景内的空间一致性。

我们物体生成过程的一个关键是利用大型生成模型，从部分图像和点云观测中生成整体且高保真度的物体网格。为此，我们首先遵循最先进的原生 3D 生成模型[80,92,94]对大型 3D 生成模型进行预训练。该模型以文本和图像输入为条件。

本文参照现有的生成框架，基于 3DShape2VecSet 表示[92,94]，通过几何变分自编码器（VAE）进行几何生成。这个 VAE 框架将均匀采样的表面点云编码为无序潜在编码，并将这些潜在表示解码为有符号距离场（SDFs）。VAE 编码器 $\mathcal{E}$ 和解码器 $\mathcal{D}$ 定义如下：

$$\boldsymbol{Z} = \mathcal{E}(\boldsymbol{X}), \quad \mathcal{D}(\boldsymbol{Z}, \boldsymbol{p}) = \mathrm{SDF}(\boldsymbol{p}), \tag{1}$$

其中 $\boldsymbol{X}$ 表示几何体的表面采样点云，$\boldsymbol{Z}$ 为对应潜在编码，$\mathrm{SDF}(\boldsymbol{p})$表示查询点 $\boldsymbol{p}$ 处的 SDF 值（用于通过 Marching Cubes 方法提取网格）。为了将图像信息有效融入几何生成过程，我们采用 DINOv2[60]作为图像编码器（遵循 Xiang 等人[80]与 Zhang 等人[92,94]的方法论）。几何潜在扩散模型（LDM）形式化表示为：

$$\epsilon_{\mathrm{obj}}(Z_t; t, c) \to Z, \tag{2}$$

其中 ϵ 代表扩散模型（Transformer 架构），Z_t为时间步 t 的带噪几何潜在编码，c 表示 DINOv2 编码的图像特征。我们遵循先前研究[92,94]，在 Objaverse[20]上预训练基础模型。训练后的生成模型ϵ能仅根据图像特征生成精细的三维几何。

4.1 具有遮挡感知的 3D 物体生成

直接使用基于 3D 生成模型面临着相当大的挑战，因为现实世界场景通常存在输入图像中的部分遮挡，这会严重降低生成物体几何形状的质量和准确性。为了解决这个问题，我们利用 DINOv2 的 Masked Auto Encoder（MAE）能力。具体而言，在推理过程中，我们提供一个遮挡掩模$\boldsymbol{M}$ 和输入图像 $\boldsymbol{I}$，使编码器能够通过推断被遮挡区域的潜在特征来处理缺失像素。这形式化为：

$$\boldsymbol{c_m} = \mathcal{E}_{\mathbf{DINOv2}}(\boldsymbol{I} \odot \boldsymbol{M}) \tag{3}$$

其中，$\boldsymbol{M}$ 是一个二进制掩模，指示哪些令牌应该被遮蔽并替换为[mask]令牌。在预训练阶段，**DINOv2** 在随机设置的掩模下进行训练，使其能够根据可见区域稳健地推断缺失部分。因此，在推理过程中，即使物体图像的部分被遮挡，编码器也能有效重建必要的特征，确保生成模型保持高质量和准确性。这种图像条件和遮挡处理的集成对我们的管线至关重要，因为它确保生成的 3D 物体在视觉上与输入图像一致，并在真实地反映集合结构。

规范点云条件 尽管物体生成模型能从输入物体图像产生视觉合理的网格，但由于编码图像条件 $\boldsymbol{c}$ 的高层特性及缺乏像素级监督，生成像素对齐的几何仍具挑战。我们通过额外基于变换到规范坐标系中观测到的部分点云作为生成条件来解决该问题。这种双重条件化确保生成几何不仅与输入图像视觉对齐，更准确反映其内在尺度、形状与深度。在条件化训练期间，我们通过从多

视角渲染每个三维资产来模拟真实局部扫描或估计深度图，从而获取对应 RGB 图像、相机参数与真值深度图。这些 RGB 图像随后通过先进深度估计技术（包括MoGe[75]与 Metric3D[87]）处理，生成估计深度图并投影为部分点云。为保障尺度一致性，我们根据有效深度值的中位数与中位数绝对偏差对 MoGe 与 Metric3D 的估计深度图进行缩放与平移，使其与真值深度图对齐。最终点云被归一化至规范$[-1,1]^3$空间，确保粗略的物体对齐所需的空间表征一致性。

为增强模型鲁棒性及跨现实场景的泛化能力，我们采用数据增强策略：在真值部分点云 $\boldsymbol{p}_{\text{gt}}$（从真值深度图投影以模拟精确深度）与含噪声估计部分点云 $\boldsymbol{p}_{\text{est}}$（从估计深度图投影并对齐以模拟RGB估计噪声深度）间进行插值。数学表示为 $\boldsymbol{p}_{\text{disturb}} = \alpha \cdot \boldsymbol{p}_{\text{gt}} + (1-\alpha) \cdot \boldsymbol{p}_{\text{est}}$，其中$\alpha \in [0,1]$为训练期间均匀采样的权重因子。我们的物体生成器，命名为ObjectGen，在部分点云条件下的形式化表示为：

$$\epsilon(Z_t; t, c, \boldsymbol{p}_{\text{disturb}}) \to Z, \tag{4}$$

其条件化适配方案类似Zhang等人[92,94]的注意力机制。此外，为模拟真实遮挡与数据缺失，我们在不同相机视角的深度图中随机掩码基本图元（如圆形与矩形），从而产生含遮挡与不完整区域的局部点云，进一步提升模型处理不完美输入的能力。本方法的关键设计是保持训练数据集中部分点云与几何的对齐：不同于对增强点云施加随机缩放、平移或旋转的方法，我们对齐的部分点云确保生成模型能更有效契合输入点云。这种对齐约束着模型去紧密遵循物体的实际形状与尺度，从而实现更精确、一致的三维重建。通过对这些良好对齐的部分点云进行条件化，模型在整体尺寸与局部几何细节上均实现卓越对齐。

4.2 生成对齐

每个生成的3D物体都在标准化体积内，并假定一个规范姿态，该姿态可能与图像和场景空间点云不完全对齐。这是因为图像条件使用了例如DINOv2的高级特征，以实现更好的泛化。确保每个物体都被正确变换和缩放以与它在场景中的样子对齐，对于场景组合至关重要。尽管可以采用传统对齐方法，如迭代最近点法（ICP）[2,6]，但它们通常无法考虑语义上下文，导致常出现未对齐和低准确性（见图9）。相反，我们引入了一个对齐生成模型，以场景空间部分点云 $\boldsymbol{q} \in \mathbb{R}^{N\times 3}$ 和规范空间几何潜在编码 $\boldsymbol{Z}$ 为条件。正式地，我们定义我们的对齐生成器AlignGen如下：

$$\epsilon_{\text{align}}(\boldsymbol{p}_t; t, \boldsymbol{q}, Z) \to \boldsymbol{p}, \tag{5}$$

其中 ϵ_{align} 是一个点云扩散变换器，$\boldsymbol{p} \in \mathbb{R}^{N\times 3}$ 是场景空间部分点云转换到规范空间后的版本，与生成的物体网格对齐。$\boldsymbol{Z}$ 是由物体生成模型生成的与 $\boldsymbol{p}$ 对应的物体几何潜在表示。$\boldsymbol{p_t}$ 是时间步 t 处 $\boldsymbol{p}$ 的噪声版本。本质上，生成模型将场景空间部分点云 $\boldsymbol{q}$ 映射到规范的$[-1,1]^3$空间中的 $\boldsymbol{p}$，并与生成的物体网格对齐。我们随后可以使用Umeyama算法[73]从 $\boldsymbol{q}$ 和 $\boldsymbol{p}$ 中恢复相似变换（即缩放、旋转和平移），因为它们是逐点对应的。这一最终步骤在数值上比直接预测变换参数更稳定。

实际上，我们对输入点云 $\boldsymbol{q}$ 和几何潜在 $\boldsymbol{Z}$ 采用不同的条件策略。对于 $\boldsymbol{q}$，我们沿着特征通道维度将输入点云与扩散样本 $\boldsymbol{p_t}$ 拼接起来，使变换器架构能够学习噪声规范帧部分点云与世界空间部分点云之间的显式对应关系。对于几何潜在 $\boldsymbol{Z}$，我们应用交叉注意力机制将其注入点扩散模型（Transformer架构）。这种方法确保模型有效整合空间和几何关系。此外，由于对称性和重复的几何形状，对于给定的 $\boldsymbol{q}$ 和 $\boldsymbol{Z}$ 可能存在多个有效的 $\boldsymbol{p}$。我们的扩散模型通过采样多个噪声实现并聚合结果变换来解决这个问题，以选择置信度最高的表示。

4.3 迭代生成过程

回想一下，在我们的设计中，物体点云最初无法用于物体生成，因为它以场景空间表示，而我们的物体生成模型需要规范空间点云进行条件化。仅仅依赖图像线索进行物体生成通常无法产生像素对齐的几何形状。幸运的是，我们的设计通过联合迭代过程，实现了物体生成和对齐模块的无缝集成。这种集成确保了每个生成的3D物体不仅在视觉上与输入图像一致，而且在场景中准确地定位和缩放。这个迭代的工作流可以概括为以下三个关键步骤（k 表示迭代次数）：

步骤1：物体生成。对于带掩码的物体图像，物体生

成模块（第4.1节）基于DINOv2提取的图像特征 c 与规范坐标系中的对齐点云 $\boldsymbol{p}^{(k)}$，合成几何潜在编码 $\boldsymbol{z}^{(k)}$。我们初始化 $\boldsymbol{p}^{(0)}$为场景空间点云 $\boldsymbol{q}$，并设置点云条件化比例因子 $\beta^{(k)}$ 随迭代进程从0逐步增至1，使部分点云的影响随时间逐步增强。形式化表述为：

$$z^{(k)} = \mathrm{ObjectGen}\left(c, \boldsymbol{p}^{(k)} \otimes \beta^{(k)}\right). \quad (6)$$

步骤2：对齐。随后，生成式对齐模块（第4.2节）接收新生成的几何潜在编码 $\boldsymbol{z}^{(k)}$ 与场景坐标系中的部分点云 $\boldsymbol{q}$，预测变换后的规范空间部分点云 $\boldsymbol{p}^{(k+1)}$：

$$\boldsymbol{p}^{(k+1)} = \mathrm{AlignGen}\left(\boldsymbol{q}, \boldsymbol{z}^{(k)}\right). \quad (7)$$

此变换后的点云 $\boldsymbol{p}^{(k+1)}$作为改进的对齐参考用于下一次迭代。通过生成式变换模型，确保缩放、旋转和平移调整既精确又符合语义理解。

步骤3：细化。利用更新后的部分点云 $\boldsymbol{p}^{(k+1)}$，系统可估算新的相似变换以优化生成几何在场景中的对齐。更新后的点云随后反馈至物体生成模块进行下一次迭代，实现几何精度与空间定位的渐进式提升。

迭代循环在几何生成与变换估计间交替进行，持续直至满足收敛标准。当变换参数变化低于预设阈值或达到最大迭代次数时即实现收敛。最终获得既视觉精确又与输入数据几何对齐的高保真三维物体。通过将物体生成模块与对齐生成模块紧密集成于迭代框架内，我们的方法有效平衡了美学保真度与几何精确度。该联合生成过程综合利用视觉与深度信息，确保每个三维资产均具备高质量特性与精确定位能力。因此，该流程为构建物理正确且视觉一致的三维场景奠定了坚实基础，助力编辑、渲染与动画等下游应用。

确定物体几何后，我们采用最先进的纹理生成模块创建逼真表面细节。遵循成熟纹理合成流程[92,94]，我们分配UV贴图并训练生成网络将细致纹理绘制至三维网格。该模块能稳健处理各种增强条件下的图像，确保最终纹理即使存在遮挡或有限可见性仍与输入外观匹配。

五、基于物理的校正

第4节详述的管线独立生成每个3D物体实例，并根据单个输入图像估计其相似变换（缩放、旋转和平移）。虽然我们提出的模块实现了高精度，但生成的场景有时并不物理合理。例如，如图4所示，一个物体（例如吉他）可能与另一个物体（例如冷藏箱）相交，或者一个物体（例如冲浪板）可能在没有任何支撑（例如来自货车）的情况下不自然地漂浮。

为了解决这些问题，我们引入了一个基于物理的校正过程，该过程优化了物体的旋转和平移，确保场景符合常识性的物理约束。校正过程受物理模拟（第5.1节）启发，并公式化为基于从图像中提取的场景图（第5.3节）的优化问题（第5.2节）。

5.1 刚体模拟简要介绍

我们介绍物理（刚体）模拟的基本原理，这些原理启发了我们的问题建模，并使框架更易于应用于游戏和机器人等下游应用。更详细的介绍请参阅 Bender[5]。

在刚体模拟中，世界被建模为常微分方程（ODE）过程。在每个模拟步骤中，首先应用牛顿-欧拉（微分）方程，它们描述了刚体在无接触情况下的动态运动。进行碰撞检测以找到刚体之间的接触点，这些点是确定接触力的必要条件。对于接触处理和碰撞解决，通常有几个条件：无穿透约束以防止物体重叠，摩擦模型确保接触力在其摩擦锥内，以及互补约束以实现变量之间特定的析取关系。求解器用于解决包含方程和不等式的系统，随后更新每个刚体的速度和位置。

增强物理合理性的直接方法是利用现成的刚体模拟器来处理场景，从之前管线估计的初始状态开始，并在模拟后获得静止状态。然而，这种方法存在几个挑战。

（1）部分场景：由于2D基础模型的限制，某些物体可能缺失，因此无法重建。在完全物理规则下模拟部分场景可能导致次优结果。

（2）不完美几何形状：虽然我们的3D生成模型生成高质量几何形状，但仍可能出现微小缺陷。刚体模拟器通常需要对物体进行凸分解[53,54,77]，这会引入额外的复杂性和超参数。过度细粒度的分解可能导致非平坦、复杂表面，导致物体在模拟过程中掉落或意外移动。相反，粗粒度分解可能由于视觉和碰撞几何体之间的差异而导致视觉上的浮动物体。

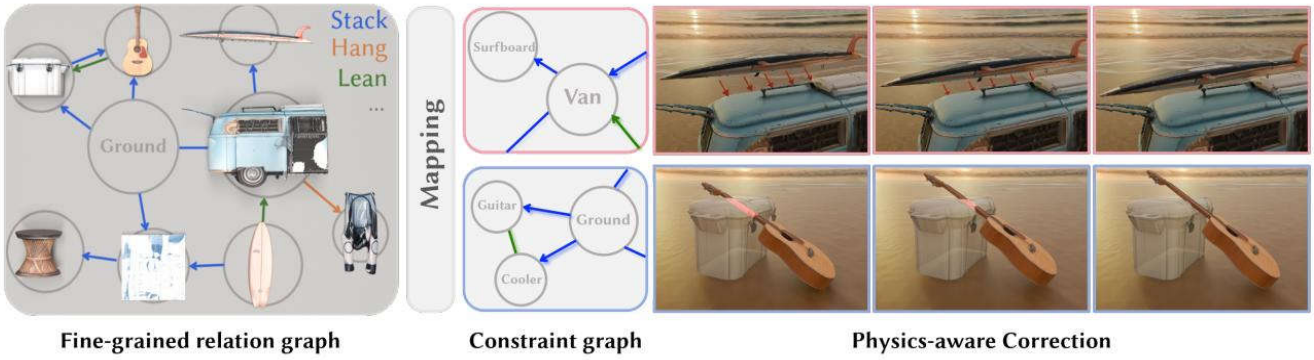

图 4　通过来自细粒度的关系图映射得到的约束图实现基于物理的校正。
右上方：浮动的冲浪板支撑在货车上。右下方：穿透的吉他和冷藏箱分离。

(3) 初始穿透：尽管姿态估计精度高，但在初始状态下可能存在显著的物体间穿透。这些穿透会给标准刚体求解器带来不稳定性，在某些情况下，如果求解器非定制，甚至会导致无解的情况。

因此，我们提出了一种定制和简化的“物理模拟”，以优化物体姿态，确保场景符合从单个图像中推导出的常识性物理原理。请注意，我们的方法不模拟完整的动力学。例如，一个物体可能无法在其当前姿态下长时间保持稳定。然而，我们认为，我们优化后的结果可以作为后续物理模拟的可靠初始化。

5.2 问题公式化和物理约束

我们将物理感知校正过程构建为一个优化问题，旨在最小化针对物体间相互关系约束的总成本：

$$\min_{T=\{T_1,T_2,\ldots,T_N\}} \sum_{i,j} C\left(T_i, T_j; \boldsymbol{o}_i, \boldsymbol{o}_j\right) \tag{8}$$

其中N为物体数量，T_i表示第 i个物体 $\boldsymbol{o}_i$ 的刚体变换（旋转与平移）。C 为成本函数，表征 $\boldsymbol{o}_i$ 与 $\boldsymbol{o}_j$ 间的关系。需注意成本函数随关系类型而变化。

受物理仿真启发，我们将关系分为两类：接触（contact）与支撑（support）。这些关系通过视觉语言模型（VLM）辅助识别，详见第5.3节。

（1） **接触关系**。描述两个物体 $\boldsymbol{o}_i$ 与 $\boldsymbol{o}_j$ 是否处于接触状态。令 $D_i(p)$ 表示物体 $\boldsymbol{o}_i$ 在点p处的符号距离函数（SDF），用于定义约束条件。$D_i(p) = D_j(p) = 0$表明 p 是 $\boldsymbol{o}_i$与 $\boldsymbol{o}_j$ 的接触点。当$D_i(p) = 0$ (即 p 为 $\boldsymbol{o}_i$表面点)时，$D_j(p) < 0$表示物体间穿透，而$D_j(p) > 0$则表示物体分离。因此成本函数定义为：

$$C\left(T_i, T_j; o_i \to o_j\right) = -\frac{\sum_{p\in\partial o_j} D_i\left(p\left(T_j\right)\right) \mathbb{I}\left(D_i\left(p\left(T_j\right)\right) < 0\right)}{\sum_{p\in\partial o_j} \mathbb{I}\left(D_i\left(p\left(T_j\right)\right) < 0\right)}$$

$$+ max\left(min_{(p\in\partial o_j)} D_i\left(p\left(T_j\right)\right), 0\right)$$

$$C\left(T_i, T_j\right) = C\left(T_i, T_j; o_i \to o_j\right) + C\left(T_i, T_j; o_j \to o_i\right)$$
$$if\ o_i\ and\ o_j\ are\ in\ contact \tag{9}$$

其中 $\partial \boldsymbol{o}_i$表示 $\boldsymbol{o}_i$的表面，$\mathbb{I}$ 为指示函数。该约束确保物体间无穿透且至少存在一个接触点。注意 $p \in \partial o_i$是 T_i 的函数。此处定义的接触约束是双向的，即同时适用于两个物体。

（2） **支撑关系**。作为单向约束，是接触关系的特例。若 $\boldsymbol{o}_i$ 支撑 $\boldsymbol{o}_j$，意味着需要优化 $\boldsymbol{o}_j$ 的位姿 T_j，而假设 $\boldsymbol{o}_i$ 保持静止。此情形通常出现在物体垂直堆叠时。该情况下的成本函数与接触关系类似，但仅涉及单向计算：

$$C\left(T_i, T_j\right) = \left|\min_{p\in\partial o_j} D_i\left(p\left(T_j\right)\right)\right|, \quad \text{if } \boldsymbol{o}_i \text{ supports } \boldsymbol{o}_j \tag{10}$$

此外，对于地面或墙壁等平坦支撑表面，我们正则化接触区域附近的SDF值，以确保物体与这些表面紧密接触。这种正则化针对物体部分重建的场景，例如图4中只有两个轮子的货车。

$$C\left(T_i, T_j\right) = \frac{\sum_{p\in\partial o_j} D_i\left(p\left(T_j\right)\right) \cdot \mathbb{I}\left(0 < D_i(p) < \sigma\right)}{\sum_{p\in\partial o_j} \mathbb{I}\left(0 < D_i(p) < \sigma\right)} \tag{11}$$

其中 $\mathbb{I}$ 为指示函数，σ 是判定点是否足够接近表面阈值。

5.3 场景关系图

物体间关系的物理线索在图像中直观存在。我们利用视觉语言模型（像 GPT-4v[1]）强大的常识推理能力[14,43,63]来识别第 5.2 节中定义的成对物理约束。给定图

像，我们采用 Set of Mark[84](SoM) 技术，通过视觉提示 GPT-4v 描述物体间关系，并随后从回答中提取场景关系图。为了解决 VLM 固有的采样不确定性，我们采用集成策略，结合多次试验的结果。如果关系在超过一半的样本中出现，则将其定义为正确，以生成相对鲁棒的推断图。更具体地说，我们多次应用随机着色和数值排序的 Set-of-Mark 方法，为进一步基于 GPT 的问答任务提供更可靠和一致的输出。

我们不是直接要求 GPT-4v 识别支撑和接触关系，而是首先提供更细粒度的物理关系，例如堆叠（物体 2 支撑物体 1）、倚靠（物体 1 倚靠物体 2）和悬挂（物体 2 从上方支撑物体 1）。我们指示 GPT-4v 分析 Set-of-Mark 方法中的编号物体，并输出所有基于接触的关系，涵盖六种类型：堆叠、倚靠、悬挂、夹紧、包含和边缘/点接触。提示词指定只有接触物体才具有关系，并且对于模糊情况默认为堆叠。

然后，我们将这些详细关系映射到预定义的支撑和接触类别，以进行进一步优化。具体来说，如果在两个之间存在相互指向的边，则该边被归类为接触；否则，归类为支撑。通过向 GPT-4v 提示这些细微的关系，有助于消除二元关系分类中潜在的歧义，并促进 GPT-4v 更准确的推理。由此生成的图示例如图 4 所示。

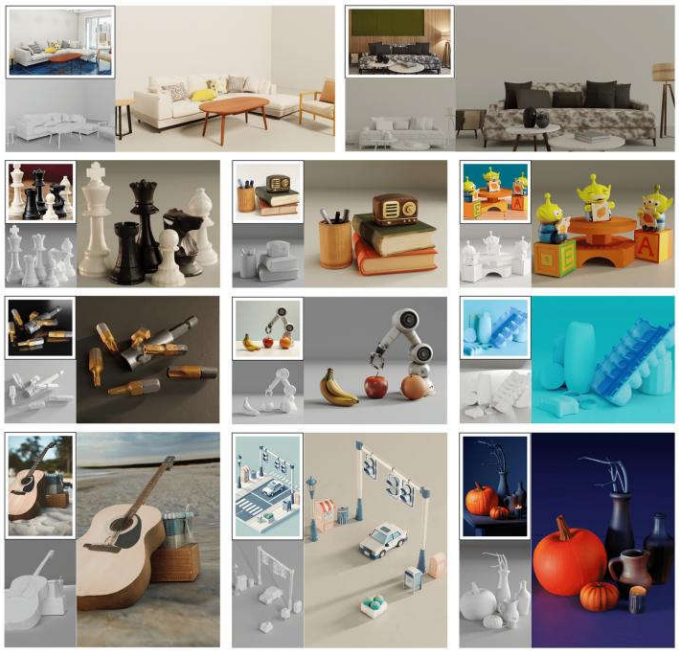

图 5　CAST 将开放词汇场景重新构想为沉浸式数字环境，捕捉每个独特场景的丰富性，将真实世界的生动多样性带入虚拟领域。对于每个场景，图像显示如下：左上角是输入图像，中上角是渲染的几何形状，右侧是带有真实纹理的渲染图像。

映射的场景约束图是一个有向图，其中节点表示物体实例，边表示物体间的物理关系。接触关系由双向边表示，而支撑关系由有向边表示。该图作为定义式(8)中使用的成本函数的基础。

5.4 基于物理感知关系图的优化

给定推断关系图定义的物理约束，我们可以实例化式(8)中描述的成本函数。该图允许我们减少需要优化的成对约束的数量，与全局物理模拟不同。

在实现方面，我们从每个物体的静止姿态表面均匀采样固定数量的点。然后根据当前物体的姿态参数对这些点进行变换，并用于查询相对于另一个物体（及其姿态）的 SDF 值。SDF 计算由 Open3D 处理，PyTorch 用于自动微分损失函数。

六、结果

图 5 展示了我们的方法从单视图输入生成的一系列 3D 场景，涵盖了各种开放词汇场景，包括详细的室内环境、物体特写以及 AI 生成的图像。这些示例突出了我们方法的多功能性和鲁棒性，展示了高保真几何、真实纹理和令人信服的场景组合。

6.1 实现细节

ObjectGen（第 4.1 节）模型的预训练遵循 3DShape2VecSet[92]和 CLAY[94]中概述的方法，其中我们利用变分自编码器（VAE）和潜在扩散模型（LDM）来生成 3D 物体几何形状。VAE 和 LDM 模块均采用 24 层 Transformer 实现，总参数量为 1.5 亿。模型在 Objaverse[20]数据集上进行训练，该数据集包含约 50 万个经过过滤的 3D 资产。部分点云条件遵循 CLAY 的适应框架的类似方法。我们将规范空间部分点云编码为位置嵌入，特征维度为 512，并使用交叉注意力机制将其注入 LDM Transformer。对于每个 3D 资产，我们渲染 32 个视图，并使用 MoGe[75]和 Metric3D[87]预计算深度图。这些深度图在训练期间被反投影为点云，并应

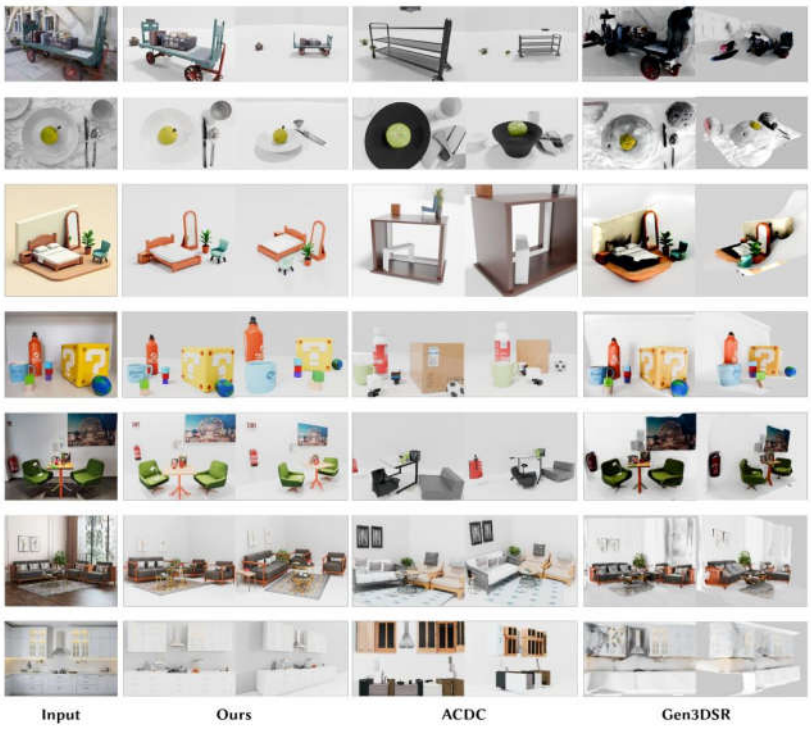

图 6　CAST 与最先进的单图像场景重建方法的定性比较。从左到右：输入图像、CAST、ACDC 和 Gen3DSR。从上到下：随机开放词汇数据集（第 1-3 行）、Gen3DSR 输入（第 4-5 行）、ACDC 输入（第 6-7 行）。

用随机掩模以模拟遮挡。我们使用最远点采样（FPS）从点云中采样 2048 个点，作为 LDM 的条件输入。为了增强模型的鲁棒性，我们随机插值真值和预测的部分点云，使得系统能够处理不同质量的数据。条件模块在 20 万个清洗过的 Objaverse 数据上训练 3000 个 epochs，使用 64 块 Nvidia A800 GPU，耗时约一周。采用 AdamW 优化器，学习率为 1e-5。对于单个物体的推理，物体生成约需要 7 秒，纹理生成大约需要 10 秒，均在一块 NVIDIA A6000 GPU 上运行。

AlignGen（第 4.2 节）模块负责生成姿态对齐，使用一个 24 层 Transformer，特征维度为 512，总共有 1.5 亿个参数。在训练期间，我们从事先计算好的深度图反投影得到的点云中随机采样规范空间中的部分点云，并对此点云应用随机变换。变换后的点云以及 ObjectGen 生成的几何潜在编码z被用作条件输入。2048 个点通过 FPS 从部分点云中被采样，以确保 Transformer 的固定输入数量。该模块在相同的 20 万个物体的数据集上训练 1500 个 epochs，使用 64 块 Nvidia A800 GPU，大约需要两天时间。采用 AdamW 优化器，学习率为 1e-5。在推理过程中，AlignGen 模块为单个物体生成姿态大约需要 1 秒。

6.2 比较

定性比较　我们首先在开放词汇场景中评估我们的方法 CAST 与最先进的单图像场景重建技术。我们还包括 ACDC 和 Gen3DSR 使用的图像，以进一步展示不同方法的场景重建结果。图 6 展示了三种方法的性能——(1) 基于检索的方法 ACDC[18]，(2) 基于生成的方法 Gen3DSR[21]，以及(3) 我们提出的 CAST——包括参考视图和新视图。我们的结果突出了 CAST 在各种设置中准确重建场景的优越能力，包括室内和室外环境、特写视角以及 AI 生成图像。

如图 6 所示，CAST 通过创新，在 ACDC 和 Gen3DSR 中脱颖而出。ACDC 受限于室内场景，并依赖大型数据集进行物体检索，通常生成与场景中物体相似而非完全相同的物体，而 CAST 支持开放词汇。这使得 CAST 能够准确重建各种复杂环境中的物体。ACDC 使用简单的边界框作为代理，而 CAST 将基于图像的物理先验与网格优化相结合，以有效处理复杂场景。与 Gen3DSR 相比，CAST 通过 Masked Autoencoder 直接进行 3D 生成，消除了容易出错的 2D 修补步骤。这带来了更平滑的网格，显著优于 Gen3DSR 在单物体生

成质量方面的表现，尤其是在具有挑战性的场景中。此外，Gen3DSR 缺乏模拟常常导致物体穿透或浮动等问题，使得场景仅从输入视角看起来一致，并降低了新视图渲染质量。相比之下，CAST 确保了跨视角的场景一致性。CAST 在各种条件下展示了稳健的场景重建，突显了其对广泛的真实世界和生成场景的适用性。

为了评估生成场景的视觉保真度和语义准确性，我们采用了两种互补的评估方法，包括 CLIP 分数 [95] 和 GPT-4 推理。我们计算渲染场景与输入图像之间的 CLIP 分数，以衡量整体重建质量和视觉相似性。为了最小化无关影响，我们在计算分数之前从渲染图像和参考图像中移除了背景。我们还利用 GPT-4 对生成场景进行排名，基于各种语义方面，包括物体排列、物理关系和场景真实感。这种语义反馈有助于识别像素级分数可能不明显的对齐或上下文错误。

除了上述指标，我们还进行了一项用户研究，重点关注视觉质量（VQ）和物理合理性（PP）两个关键方面。我们随机选择配对的参考、新颖和目标视图，要求参与者选择哪种方法的输出与输入图像在相似度和整体美学方面最匹配。为了减少视觉相似性引入的潜在偏差，参与者在单独的会话中仅查看渲染结果——没有原始输入图像——并根据物理约束和常识（例如，是否有浮动物体或不可能的接触）判断哪个场景更真实。

如表 1 所示，CAST 在所有四个评估指标中均优于 ACDC 和 Gen3DSR，证实了其在生成视觉一致和物理合理场景方面的有效性。

定量比较 尽管 CAST 旨在处理开放词汇场景，但许多此类场景缺乏网格真值，这使得直接定量比较变得困难。为了解决这个问题，我们在 3DFront 数据集[22] 进行了额外评估。该数据集提供了真值网格以及对应的渲染图像，从而能够更精确地评估物体级和场景级重建。我们将 CAST 与 InstPIFu[46]、ACDC[18]和 Gen3DSR[21] 进行比较。我们计算物体级的 Chamfer 距离和 F-Score，以及场景级的 IoU、Chamfer 距离和 F-Score，以评估单个物体几何的保真度及其空间布局的准确性。为了确保公平性，我们用真值掩模替换了其他方法中的分割模块，使得任何差异纯粹源于重建能力而非物体分割。

Method	CLIP↑	GPT-4↓	VQ↑	PP↑
ACDC	69.77	2.7	5.58%	22.86%
Gen3DSR	79.84	2.175	6.35%	5.72%
CAST	**85.77**	**1.125**	**88.07%**	**71.42%**

表 1　将场景重建方法在 CLIP 分数、GPT-4 排名、视觉质量（VQ）和物理合理性（PP）四个指标上的定量比较

如表 2 所示，CAST 不仅实现了更高的物体级生成质量，而且在场景布局精度方面也超越了现有方法。即使在室内数据集的限制下，我们的方法也表现出稳健的性能，优于比较的基线。

6.3 评估

为了阐明 CAST 中关键组件的个体贡献，我们进行了一系列消融研究。这些实验系统地移除或修改了特定组件，以评估它们对整体性能的影响。消融研究侧重于几个关键设计选择：遮挡感知物体生成、点云条件、姿态对齐生成和基于物理的校正过程。

遮挡感知生成消融 遮挡是复杂场景中的一个重大挑战。为了评估 Masked Autoencoder（MAE）在处理遮挡方面的有效性，我们进行了一项消融研究，比较了有无 MAE 组件的生成结果。如图 7 所示，结果突出了遮挡感知模块的重要性。没有 MAE，部分遮挡区域生成的物体表现出显著退化。例如，飞船显示为破碎不完整，而杯子被描绘为破碎带有缺失部分。相比之下，当应用 MAE 条件时，模型成功推断并填充了被遮挡区域，从而产生更准确和视觉一致的生成，与输入图像更好地

Method	CD-S↓	FS-S↑	CD-O↓	FS-O↑	IoU-B↑
Vanila	0.079	53.38	0.069	52.83	0.515
+MAE	0.064	53.79	0.066	54.32	0.548
+ PCD	0.056	53.91	0.060	54.60	0.582
+ iter.	**0.052**	**56.18**	**0.057**	**56.50**	**0.603**

表 2　3D-Front 室内数据集上场景重建性能的定量比较。我们根据形状精度计算 Chamfer 距离（CD）、物体级重建质量计算 F-Score (FS) 以及场景级重叠计算 Intersection over Union（IoU）评估不同方法。

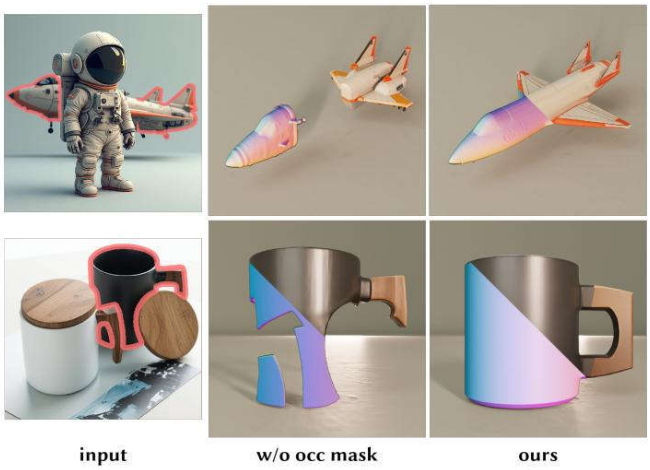

图 7　我们评估了有无遮挡感知生成模块的生成性能。物体渲染图和法线渲染图突显了该模块在确保生成物体的完整性和高质量方面的重要性。

对齐。这表明遮挡感知模块在确保准确重建被遮挡物体、提高最终 3D 场景的完整性和真实感方面具有关键作用。

部分点云条件消融　我们进行了一项消融研究，以研究规范空间部分点云条件在生成过程中的作用。尽管直接从输入图像生成可以产生视觉上合理的结果，但在缺乏像素级对齐的情况下，模型难以保持正确的物体数量和尺度，导致生成不令人满意。为了更有效地展示点

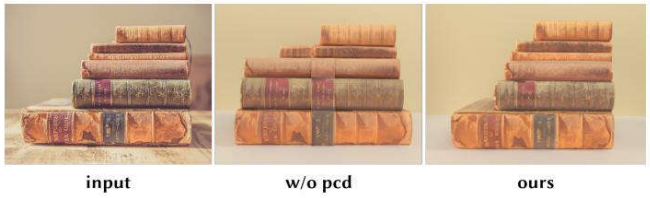

图 8　一叠不同长度和宽度的书籍。没有点云条件，模型直接生成一个复杂的单一物体。这展示了点云条件增强了尺度、维度和局部细节的保留。

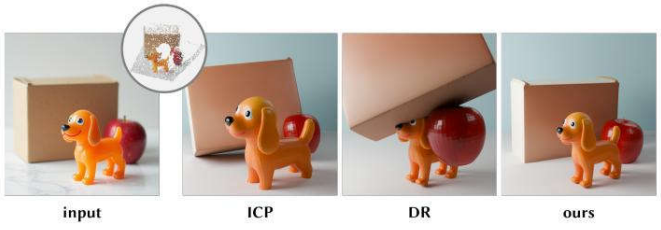

图 9　姿态估计方法的比较。我们的姿态对齐模块相比迭代最近点（ICP）和可微分渲染（DR）显示出优越的对齐精度。

云条件在生成单个实例中的重要性，我们选择直接生成一个更复杂的实例结构：一叠六本长度和宽度各异的书籍。如图 8 所示，当生成过程仅依赖输入图像时，在没有点云条件的情况下，结果频繁出现生成的物体数量和维度不准确。相比之下，点云条件引入了稳健的几何先验，显著提高了生成场景的精度。这种增强确保了具有复杂形状和不同维度的物体能够更准确地重建，与输入图像中描绘的真实世界对应物更加相似。这表明几何先验在通过保留真实尺度和形状来增强 3D 场景生成保真度方面具有关键作用。

姿态对齐生成方法的有效性　为了评估姿态对齐模块的有效性，我们将其与常用的姿态估计算法进行了比较，如迭代最近点（ICP）[2,6]和可微分渲染[40]。生成的网格被提供给不同的姿态估计算法，以使其与参考 RGB 图像及其对应的深度预测对齐。对于 ICP 方法，我们从生成的网格中均匀采样点云，并通过其边界框对采样点云和估计点云进行归一化，避免尺度差异。我们使用了 Open3D[97]中的 ICP 实现来对齐这两个归一化点云。对于可微分渲染，我们优化了旋转和平移参数，使变换后的物体网格的渲染图像与参考 RGB 图像对齐。如图 9 所示，我们的方法在对齐精度方面优于 ICP 和可微分渲染。ICP 在处理点云中的异常值、未知物体尺度以及对称或重复几何形状时，通常难以进行准确的姿态估计，可能导致局部最小值。另一方面，可微分渲染受到 RGB 输入中遮挡的显著影响，干扰了物体姿态的优化，并阻碍了与输入图像的精确对齐。我们的结果表明，我们的姿态对齐模块优于传统的 ICP 和可微分渲染方法，证明了其在准确估计生成网格的物体姿态和改进与输入图像对齐方面的鲁棒性。

物理一致性校正的效果　在 CAST 中，物理约束对于实现真实的物体交互和维持场景内的空间一致性至关重要。虽然我们解决了遮挡和不完整视图等常见挑战，但浮动物体、穿透和未对齐的空间关系等问题仍然存在。如图 10 所示，在没有关系约束的情况下生成的场景可能在物理上不一致；当应用完整的物理模拟时，物体会遵守物理定律，但它们的相对位置和整体排列可能与预期场景显著不同（例如，洋葱可能会从表面掉落，打乱

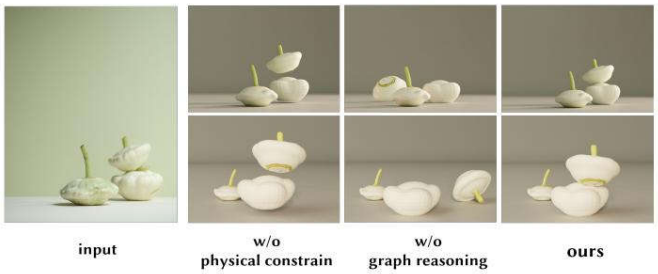

图 10　有无关系图约束的场景重建比较。通过整合关系图约束，我们的方法确保了物理合理性和与预期场景的准确对齐，保持了正确的空间关系。

原始构图）。通过整合关系图约束，我们的方法确保物体不仅符合物理可行性，而且与预期场景布局对齐，同时保留了物理合理性和所需的空间关系。

不同模块的定量消融研究 为了定量评估每个模块的贡献，我们进行了一项全面的消融研究。如表 3 所示，我们评估了移除或修改关键组件对最终场景质量的影响。结果表明，每个组件都对我们方法的整体性能做出了显著贡献。定量分析进一步突出了每个模块在实现高质量、物理一致和逼真的场景重建方面的重要性。

应用 如图 11 所示，CAST 将单个图像转换为一个实例化的 3D 场景，从而支持广泛的应用。这种重建详细环境的能力通过确保真实的物体交互，为基于物理的动画提供了动力。它还支持机器人领域的真实到模拟（real-to-sim）工作流，允许从真实世界数据集进行准确的场景复制。在游戏开发中，CAST 促进了沉浸式环境的创建，使得如实重建的场景可以无缝集成到基于虚幻引擎的交互式世界中。

Method	CD-S↓	FS-S↑	CD-O↓	FS-O↑	IoU-B↑
Vanilla	0.079	53.38	0.069	52.83	0.515
+ MAE	0.064	53.79	0.066	54.32	0.548
+ PCD	0.056	53.91	0.060	54.60	0.582
+ iter.	**0.052**	**56.18**	**0.057**	**56.50**	**0.603**

表 3　MAE 模块、点云条件（PCD）和迭代细化策略（iter.）的定量消融研究。为简洁起见，每行仅显示新增关键组件。

七、结论

在本文中，我们介绍了 CAST，一种新颖的单图像 3D 场景重建方法，它结合了几何保真度、像素级对齐和物理约束。通过整合场景分解、感知 3D 实例生成框架和物理校正技术，CAST 解决了姿态未对齐、物体相互依赖和部分遮挡等关键挑战。这种结构化的管线生成了视觉准确且物理一致的 3D 场景，超越了传统以物体为中心方法的局限性。我们通过广泛的实验和用户研究验证了 CAST，证明其在视觉质量和物理合理性方面显著优于现有最先进方法。我们预计 CAST 将为 3D 生成、场景重建和沉浸式内容创建的未来发展奠定坚实基础。

局限性和未来工作 CAST 中场景生成的质量严重依赖于底层的物体生成模型。目前，该模型仍缺乏足够的细节和精度，这一局限性导致生成的物体存在显著不一致，影响它们在场景中的对齐和空间关系。

此外，当前网格表示在处理纺织品、玻璃或织物等材料时仍存在困难，常常显得不自然，并且无法准确描绘透明材料，如图 12 所示。虽然已加入额外模块以增强物体鲁棒性和相似性，但仍需要更先进和鲁棒生成模型。更详细和准确的物体生成器可以显著提高整体场景质量并增强其现实世界适用性。

图 11　CAST 实现了逼真的基于物理的动画、沉浸式游戏环境和高效的真实到模拟过渡，推动各领域创新。

图 12　在某些场景中，透明玻璃、纺织品和织物难以表达，因为网格难以真实地表示它们。

当前方法的一个显著局限是缺乏光照估计和背景建模。没有真实的光照，物体与其周围环境之间的相互作用可能缺乏自然的阴影和光照效果，影响生成的环境的视觉真实感和沉浸感。为了增强视觉真实感，我们采用现成的全景 HDR 生成工具[34]，并结合 Blender 中预设的光照条件进行手动操作。CAST 未来可以受益于整合先进的光照估计和背景建模技术，从而显著丰富场景的上下文深度和视觉保真度。

在更复杂的场景中，当前方法的性能可能会略有下降。复杂的空间布局和密集的物体配置等挑战可能会在一定程度上影响场景重建的准确性。尽管 CAST 目前在重建单个场景方面表现出色，但一个有潜力的方向是利用其输出构建大型数据集，从而促进基于学习的场景生成或视频生成管线。扩展生成场景的多样性和真实感，可以进一步提高 3D 生成模型在电影制作、模拟和沉浸式媒体等领域的鲁棒性和适用性。

致谢

这项工作得到了国家重点研发计划（2022YFF0902301）、国家自然科学基金项目（61976138，61977047）、上海市科学技术委员会（2015F0203-000-06）和上海市教育委员会（2019-01-07-00-01-E00003）的支持。我们还要感谢上海人工智能前沿科学中心（ShangHAI）、教育部智能感知与人机协作重点实验室（上海科技大学）、上海科技大学计算机科学与通信学科平台以及上海科技大学高性能计算平台的支持。

责任编委　**魏秀参**

参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023).

[2] K Somani Arun, Thomas S Huang, and Steven D Blostein. 1987. Least-squares fitting of two 3-D point sets. IEEE Transactions on pattern analysis and machine intelligence 5 (1987), 698–700.

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo MartinBrualla, and Pratul P Srinivasan. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In Proceedings of the IEEE/CVF international conference on computer vision. 5855–5864.

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5470–5479.

[5] Jan Bender, Kenny Erleben, Jeff Trinkle, and Erwin Coumans. 2012. Interactive Simulation of Rigid Body Dynamics in Computer Graphics. In 33rd Annual Conference of the European Association for Computer Graphics, Eurographics 2012 - State of the Art Reports, Cagliari, Sardinia, Italy, May 13-18, 2012, Marie-Paule Cani and Fabio Ganovelli (Eds.). Eurographics Association, 95–134. https://doi.org/10.2312/CONF/ EG2012/STARS/095-134

[6] Paul J Best. 1992. A method for registration of 3-D shapes. IEEE Trans Pattern Anal Mach Vision 14 (1992), 239–256.

[7] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023).

[8] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023).

[9] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. 2024. Genie: Generative interactive environments. In Forty-first International Conference on Machine Learning.

[10] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015).

[11] Anpei Chen, Minye Wu, Yingliang Zhang, Nianyi Li, Jie Lu, Shenghua Gao, and Jingyi Yu. 2018. Deep surface light fields. Proceedings of the ACM on Computer Graphics and Interactive Techniques 1, 1 (2018), 1–17.

[12] Yixin Chen, Junfeng Ni, Nan Jiang, Yaowei Zhang, Yixin Zhu, and Siyuan Huang. 2024a. Single-view 3d scene reconstruction with high-fidelity shape and texture. In 2024 International Conference on 3D Vision (3DV). IEEE, 1456–1467.

[13] Yunuo Chen, Tianyi Xie, Zeshun Zong, Xuan Li, Feng Gao, Yin Yang, Ying Nian Wu, and Chenfanfu Jiang. 2024b. Atlas3D: Physically Constrained Self-Supporting Text-to-3D for Simulation and Fabrication. arXiv preprint arXiv:2405.18515 (2024).

[14] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. 2024. Navila: Legged robot visionlanguage-action model for navigation. arXiv preprint arXiv:2412.04453 (2024).

[15] Tao Chu, Pan Zhang, Qiong Liu, and Jiaqi Wang. 2023. Buol: A bottom-up framework with occupancy-aware lifting for panoptic 3d scene reconstruction from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4937–4946.

[16] Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. 2021. Panoptic 3d scene reconstruction from a single rgb image. Advances in Neural Information Processing Systems 34 (2021), 8282–8293.

[17] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In Proceedings of the IEEE conference on computer vision and pattern recognition. 5828–5839.

[18] Tianyuan Dai, Josiah Wong, Yunfan Jiang, Chen Wang, Cem Gokmen, Ruohan Zhang, Jiajun Wu, and Li Fei-Fei. 2024. Automated Creation of Digital Cousins for Robust Policy Learning. arXiv preprint arXiv:2410.07408 (2024).

[19] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. 2024. Objaverse-xl: A universe of 10m+ 3d objects. Advances in Neural Information Processing Systems 36 (2024).

[20] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 13142–13153.

[21] Andreea Dogaru, Mert Özer, and Bernhard Egger. 2024. Generalizable 3D Scene Reconstruction via Divide and Conquer from a Single View. arXiv preprint arXiv:2404.03421 (2024).

[22] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 2021. 3d-front: 3d furnished rooms with layouts and semantics. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 10933–10942.

[23] Daoyi Gao, Dávid Rozenberszki, Stefan Leutenegger, and Angela Dai. 2024b. Diffcad: Weakly-supervised probabilistic cad model retrieval and alignment from an rgb image. ACM Transactions on Graphics (TOG) 43, 4 (2024), 1–15.

[24] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo MartinBrualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. 2024a. Cat3d: Create anything in 3d with multi-view diffusion models. arXiv preprint arXiv:2405.10314 (2024).

[25] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. The International Journal of Robotics Research 32, 11 (2013), 1231–1237.

[26] Georgia Gkioxari, Nikhila Ravi, and Justin Johnson. 2022. Learning 3d object shape and layout without 3d supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1695–1704.

[27] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M. Seitz. 2007. Multi-View Stereo for Community Photo Collections. In 2007 IEEE 11th International Conference on Computer Vision. 1–8. https://doi.org/10.1109/ICCV.2007.4408933

[28] Can Gümeli, Angela Dai, and Matthias Nießner. 2022. Roca: Robust cad model retrieval and alignment from a single image. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 4022–4031.

[29] Minghao Guo, Bohan Wang, Pingchuan Ma, Tianyuan Zhang, Crystal Elaine Owens, Chuang Gan, Joshua B Tenenbaum, Kaiming He, and Wojciech Matusik. 2024. Physically Compatible 3D Object Modeling from a Single Image. arXiv preprint arXiv:2405.20510 (2024).

[30] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022a. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022).

[31] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022b. Video diffusion models. Advances in Neural Information Processing Systems 35 (2022), 8633–8646.

[32] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. 2023. Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023).

[33] Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. 2024. MIDI: Multi-Instance Diffusion for Single Image to 3D Scene Generation. arXiv preprint arXiv:2412.03558 (2024).

[34] Hyper3D. 2025. Omnicraft. https://hyper3d.ai/omnicraft/hdri

[35] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. 2017. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In Proceedings of the IEEE international conference on computer vision. 1521–1529.

[36] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics 42, 4 (July 2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[37] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 4015–4026.

[38] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. 2021. Patch2cad: Patchwise embedding learning for in-the-wild shape retrieval from a single image. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 12589– 12599.

[39] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. 2020. Cosypose: Consistent multi-view multi-object 6d pose estimation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. Springer, 574–591.

[40] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. 2020. Modular primitives for high-performance differentiable rendering. ACM Transactions on Graphics (TOG) 39 (2020), 1 – 14.

[41] Florian Langer, Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. 2022. SPARC: Sparse render-and-compare for CAD model alignment in a single RGB image. arXiv preprint arXiv:2210.01044 (2022).

[42] Bruno Latour. 2005. Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, Oxford, UK.

[43] Wenhao Li, Zhiyuan Yu, Qijin She, Zhinan Yu, Yuqing Lan, Chenyang Zhu, Ruizhen Hu, and Kai Xu. 2024b. LLM-enhanced Scene Graph Learning for Household Rearrangement. In SIGGRAPH Asia 2024 Conference Papers. 1–11.

[44] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. 2024a. Evaluating Real-World Robot Manipulation Policies in Simulation. arXiv preprint arXiv:2405.05941 (2024).

[45] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. 2024. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6517–6526.

[46] Haolin Liu, Yujian Zheng, Guanying Chen, Shuguang Cui, and Xiaoguang Han. 2022. Towards high-fidelity single-view holistic reconstruction of indoor scenes. In European Conference on Computer Vision. Springer, 429–446.

[47] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. 2024. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. Advances in Neural Information Processing Systems 36 (2024).

[48] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023c. Zero-1-to-3: Zero-shot one image to 3d object. In Proceedings of the IEEE/CVF International Conference on Computer Vision.

[49] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2025. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European Conference on Computer Vision. Springer, 38–55.

[50] Xueyi Liu, Bin Wang, He Wang, and Li Yi. 2023b. Few-Shot Physically-Aware Articulated Mesh Generation via Hierarchical Deformation. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 854–864.

[51] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2023a. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. In arXiv preprint arXiv:2309.03453.

[52] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2024. Wonder3d: Single image to 3d using cross-domain diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9970–9980.

[53] Khaled Mamou and Faouzi Ghorbel. 2009. A simple and efficient approach for 3D mesh approximate convex decomposition. In 2009 16th IEEE international conference on image processing (ICIP). IEEE, 3501–3504.

[54] Khaled Mamou, E Lengyel, and A Peters. 2016. Volumetric hierarchical approximate convex decomposition. Game engine gems 3 (2016), 141–158.

[55] Mariem Mezghanni, Théo Bodrito, Malika Boulkenafed, and Maks Ovsjanikov. 2022. Physical simulation layer for accurate 3d modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 13514–13523.

[56] Mariem Mezghanni, Malika Boulkenafed, Andre Lieutier, and Maks Ovsjanikov. 2021. Physically-aware generative network for 3d shape modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9330–9341.

[57] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In European conference on computer vision. Springer, 405–421.

[58] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. ACM Trans. Graph. 41, 4, Article 102 (July 2022), 15 pages. https://doi.org/10.1145/3528223.3530127

[59] Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Puhao Li, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. 2024. PhyRecon: Physically Plausible Neural Scene Reconstruction. Advances in Neural Information Processing Systems.

[60] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023).

[61] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. 2024. UniDepth: Universal Monocular Metric Depth Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10106–10116.

[62] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. In arXiv preprint arXiv:2209.14988.

[63] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. 2023. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In 7th Annual Conference on Robot Learning.

[64] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024).

[65] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024).

[66] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. 2018. Pix3d: Dataset and methods for single-image 3d shape modeling. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2974–2983.

[67] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F Henriques, Christian Rupprecht, and Andrea Vedaldi. 2024a. Flash3D: Feed-Forward Generalisable 3D Scene Reconstruction from a Single Image. arXiv preprint arXiv:2406.04343 (2024).

[68] Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. 2024b. Splatter image: Ultra-fast single-view 3d reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10208–10217.

[69] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. 2025. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In European Conference on Computer Vision. Springer, 1–18.

[70] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2023. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023).

[71] Fengrui Tian, Shaoyi Du, and Yueqi Duan. 2023. Mononerf: Learning a generalizable dynamic radiance field from monocular videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 17903–17913.

[72] Marcel Torne, Anthony Simeonov, Zechu Li, April Chan, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. 2024. Reconciling reality through simulation: A real-to-sim-toreal approach for robust manipulation. arXiv preprint arXiv:2403.03949 (2024).

[73] Shinji Umeyama. 1991. Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis & Machine Intelligence 13, 04 (1991), 376–380.

[74] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. 2025. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In European Conference on Computer Vision. Springer, 439–457.

[75] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. 2024b. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. arXiv preprint arXiv:2410.19115 (2024).

[76] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2024a. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in Neural Information Processing Systems 36 (2024).

[77] Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. 2022. Approximate convex decomposition for 3d meshes with collision-aware concavity and tree search. ACM Transactions on Graphics (TOG) 41, 4 (2022), 1–18.

[78] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. 2024a. Unique3D: High-Quality and Efficient 3D Mesh Generation from a Single Image. arXiv preprint arXiv:2405.20343 (2024).

[79] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. 2024b. Reconfusion: 3d reconstruction with diffusion priors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 21551–21561.

[80] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. 2024. Structured 3D Latents for Scalable and Versatile 3D Generation. arXiv preprint arXiv:2412.01506 (2024).

[81] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4818–4829.

[82] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. 2024. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4389–4398.

[83] Qingshan Xu, Jiao Liu, Melvin Wong, Caishun Chen, and Yew-Soon Ong. 2024. PrecisePhysics Driven Text-to-3D Generation. arXiv preprint arXiv:2403.12438 (2024).

[84] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. arXiv preprint arXiv:2310.11441 (2023).

[85] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024b. Depth anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10371–10381.

[86] Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. 2024a. Physcene: Physically interactable 3d scene synthesis for embodied ai. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16262–16272.

[87] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. 2023. Metric3d: Towards zero-shot metric 3d prediction from a single image. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 9043–9053.

[88] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelnerf: Neural radiance fields from one or few images. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 4578–4587.

[89] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. 2024. Wonderjourney: Going from anywhere to everywhere. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6658–6667.

[90] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. 2022. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. Advances in neural information processing systems 35 (2022), 25018– 25032.

[91] Alan Yuille and Daniel Kersten. 2006. Vision as Bayesian inference: analysis by synthesis? Trends in cognitive sciences 10, 7 (2006), 301–308.

[92] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 2023. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. ACM Transactions on Graphics (TOG) 42, 4 (2023), 1–16.

[93] Jiancheng Zhang, Haijin Zeng, Yongyong Chen, Dengxiu Yu, and Yin-Ping Zhao. 2024b. Improving Spectral Snapshot Reconstruction with Spectral-Spatial Rectification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 25817–25826.

[94] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. 2024a. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets. ACM Transactions on Graphics (TOG) 43, 4 (2024), 1–20.

[95] SUN Zhengwentai. 2023. clip-score: CLIP Score for PyTorch. https://github.com/taited/ clip-score. Version 0.1.1.

[96] Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. 2025. Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. In European Conference on Computer Vision. Springer, 407–423.

[97] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. 2018. Open3D: A Modern Library for 3D Data Processing. ArXiv abs/1801.09847 (2018).

姚凯欣

上海科技大学信息科学与技术学院 2023 级博士研究生，导师为虞晶怡教授和许岚教授，主要研究方向为场景生成、3D 生成、3D 高斯。
Email: yaokx2023@shanghaitech.edu.cn

张龙文

上海科技大学信息科学与技术学院 2022 级博士生，导师为虞晶怡教授和许岚教授，主要研究方向为 3D 生成。
Email: zhanglw2@shanghaitech.edu.cn

严新豪

上海科技大学信息学院 2023 级博士研究生,导师为许岚助理教授， 主要研究方向为多模态,3D 生成式模型。
Email: yanxh@shanghaitech.edu.cn

曾 焱

上海科技大学信息科学与技术学院 2024 级博士生，导师为虞晶怡教授，主要研究方向为视频生成。
Email: zengyan2024@shanghaitech.edu.cn

张启煊

上海科技大学研究生，同时担任数字人 AI 公司 DeemosTech 的首席技术官。专攻计算机图形学、计算摄影学和生成式人工智能领域。其研究成果屡获 SIGGRAPH、ICCV、CVPR 等顶尖学术会议录用，相关技术已应用于多部影视作品及游戏项目。
Email: zhangqx1@shanghaitech.edu.cn

杨 卫

华中科技大学副教授、博士生导师、湖北省百人、东湖学者。2017 年毕业于美国特拉华大学，获得计算机科学专业博士学位，师从 Jingyi Yu 教授，毕业后曾在 Google 的先进科技与计划部门和知名初创公司 DGene 工作，从事基于视觉的场景理解与虚拟现实的应用研究。2021 年进入华中科技大学计算机学院智能媒体计算与网络安全实验室从事教学科研工作。主要研究方向包括：三维视觉理解与重建、基于物理特性的视觉和图形学、先进感知与图像传感器等。在 TPAMI、TOG、IJCV、CVPR、 ICCV、ECCV 等期刊会议上发表高水平论文 20 余篇。现任中国图学学会动漫图学工程专业委员会委员，将担任 CVPR 23 的 Area Chair, AAAI 20 & 21, 以及 WACV 21, BMVC 18 的程序委员，TPAMI, TIP, TVCJ, CVPR, ICCV, NeurIPS, ECCV 等顶级会议期刊的审稿人。

Email: weiyangcs@hust.edu.cn

许 岚

上海科技大学信息科学与技术学院助理教授、研究员、博士生导师。许岚教授在浙江大学信息与电子工程学系获学士学位；在香港科技大学电子与计算机工程获博士学位。之后加入上海科技大学，任助理教授、研究员。他的研究方向包括计算机视觉、计算机图形学、机器学习、计算摄像学，目前的研究兴趣侧重于动静态三维重建、虚拟现实、数字孪生，终极目标是实现个人数字资产化和沉浸式全息立体通信。他已发表了多篇顶级期刊和会议文章，包括 CVPR、ECCV、ICCP、IROS、IEEE TRO、IEEE TVCG、IEEE TPAMI 等。主要研究内容包括人体动态捕捉、动静态三维重建与理解、数字孪生、虚拟现实、增强现实。

Email: xulan1@shanghaitech.edu.cn

顾家远

上海科技大学信息科学与技术学院助理教授、研究员、博士生导师。顾家远教授 2018 年于北京大学信息科学技术学院智能科学系获得本科学位，并于 2024 年在美国加州大学圣地亚哥分校计算机科学与工程学院获得博士学位。他曾在 Uber ATG、Waymo、Facebook AI、Qualcomm AI、Google DeepMind 等机构担任实习或学生研究员。他的研究方向包括具身智能与三维视觉，在计算机视觉、机器学习、机器人等国际顶会上发表 20 余篇工作。

Email: gujy1@shanghaitech.edu.cn

虞晶怡

上海科技大学副教务长，信息学院院长。在加入上海科技大学前，他任职美国特拉华大学计算机与信息科学系正教授。他于 2000 年获美国加州理工大学应用数学及计算机学士学位， 2003 年获美国麻省理工大学计算机与电子工程硕士学位，2005 年获美国麻省理工大学计算机与电子工程博士学位。他长期从事计算机视觉、计算成像、计算机图形学、生物信息学等领域的研究工作，已发表 120 多篇学术论文，其中超 70 篇发表于国际会议 CVPR/ICCV/ECCV 和期刊 TPAMI。他已获得美国发明专利 20 余项，并于 2009 和 2010 年分别获得美国国家科学基金的杰出青年奖和美国空军研究院的杰出青年奖。他是 IEEE TPAMI、IEEE TIP 和 Elsevier CVIU 的编委，担任 ICPR 2020、CVPR 2021、WACV 2021、ICCV 2027 的程序主席和 ICCV 2025 的大会主席。因为他在计算机视觉和计算成像上的贡献，当选 IEEE Fellow。

Email: yujingyi@shanghaitech.edu.cn