DeOcc-1-to-3: 通过自监督多视图扩散从单张图像实现三维去遮挡

Yansong Qu[1], Shaohui Dai[1], Xinyang Li[1], Yuze Wang[2],
You Shen[1], Liujuan Cao[1], Rongrong Ji[1]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University,
[2]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University

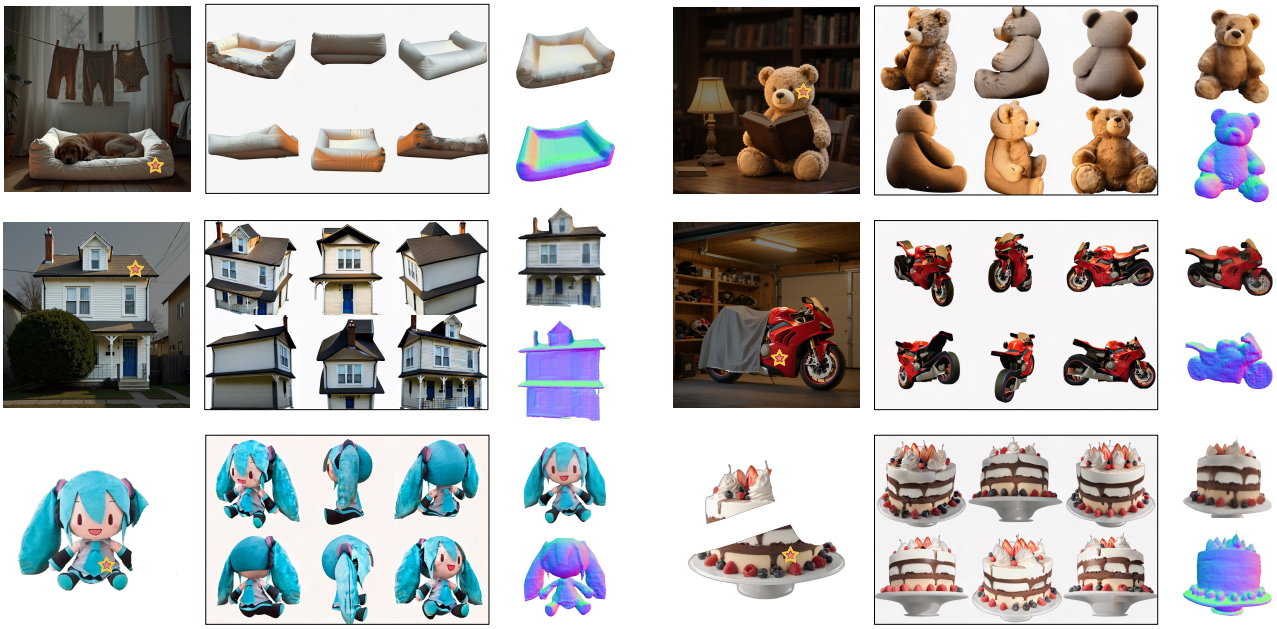


图 1. **去偶**-1 至 3 接收单个被遮挡的图像（左）作为输入，并合成结构一致的多视图去遮挡图像（中）。这些输出可以无缝集成到各种 3D 重建或生成框架中，以产生准确的网格和表面法线（右）。所提出的流水线展示了在不同物体类别和遮挡场景中的泛化能力。

Abstract

从单张图像重建 3D 物体是一个长期存在的挑战，特别是在现实世界的遮挡情况下。虽然最近基于扩散的视图合成模型可以从单一 RGB 图像生成一致的新视角，但它们通常假设输入是完全可见的，并且在对象部分被遮挡时会失败。这导致了不一致的视图和降级的 3D 重建质量。为了解决这一限制，我们提出了一种针对遮挡感知多视图生成的端到端框架。我们的方法直接从一张部分遮挡的图像合成六个结构上一致的新视角，从而在不需要预先修复或手动标注的情况下进行下游 3D 重建。我们使用 Pix2Gestalt 数据集构建了一个自监督训练流程，利用遮挡-未遮挡图像对和伪地面真实视图来教模型感知结构完成和视图一致

性。无需修改原始架构，我们将视图合成模型完全微调以联合学习完成和多视图生成。此外，我们引入了第一个针对遮挡感知重建的基准测试，涵盖多种遮挡级别、物体类别和掩码模式。该基准提供了评估未来方法在部分遮挡情况下的标准化协议。我们的代码可以在 https://github.com/Quyans/DeOcc123 获取。

1. 介绍

从单张图像重建完整的三维对象是计算机视觉中的一个长期挑战。基于扩散的多视角生成模型 [1–4] 的最新进展使得能够从单一 RGB 图像生成结构一致的新视角，为下游三维重建提供高质量输入。然而，这些模型通常假设输入的对象完全可见。实际上，图像常常因为杂乱、对象交互或有限视点而受到部分遮挡的影响，使这一假设变得不现实。遮挡对视图合成和三维建模都构成了重大挑战。当对象的部分不可见时，现有方法往往无法正确推断其几何结构和外观，导致新视角不一致以及重建结果破碎或不完整。

一个常见的解决方法是两阶段管道：首先应用 2D inpainting[5–10] 来完成被遮挡的区域，然后对已完成的图像应用视图合成或 3D 重建。然而，这种方法存在三个主要限制：(1) 2D inpainting 缺乏 3D 先验知识，无法保证多视角一致性；(2) 视图合成不了解生成内容中的不确定性，导致产生伪影；(3) 解耦设计阻止了联合优化，导致阶段间误差累积。

为了解决这些挑战，我们提出了一种端到端的、具备遮挡意识的多视图生成框架，该框架可以直接从单个被遮挡的图像预测出六个结构一致的新视图。无需改变原始视图合成架构，我们将模型完全微调以在一个统一的生成过程中共同学习遮挡补全和新视图合成。为消除手动标注的需求，我们引入了一种自监督训练策略。通过随机对 2D 输入应用遮挡，我们构建了配对的被遮挡与未被遮挡图像。对于每个未被遮挡的图像，预训练的多视图生成模型会生成六个视图伪真实值，当只有被遮挡的图像可用时，这些伪真实值监督网络。这使模型能够学习结构感知补全和视角一致合成。

此外，我们构建了一个全面的遮挡感知重建基准测试，涵盖了多个遮挡级别、对象类别和掩码模式。该基准测试为定量评估遮挡感知视图合成和 3D 重建方法的性能提供了一种标准化的评估协议。总结来说，我们的主要贡献如下：

- 我们引入了遮挡感知多视角生成任务，并提出了一种端到端框架，该框架可以直接从单个部分被遮挡的图像中合成结构一致的新视图。
- 我们开发了一种自监督训练范式，利用配对的被遮挡和未被遮挡图像，实现结构感知学习而无需人工标注。
- 我们在不修改其架构的情况下对多视角生成模型进行了微调，实现了强大的性能并与其下游的 3D 重建框架（如 InstantMesh[3]）无缝兼容。
- 我们建立了首个考虑遮挡的重建基准，涵盖多样化的遮挡模式和对象类别，以及标准化的评估指标。

2. 相关工作

2.1. 单图像到 3D 表示

早期的单图像三维重建方法主要集中在预测显式或隐式表示，如网格、点云或有符号距离场（SDFs）[11–13] 。然而，这些方法通常在处理复杂几何形状和遮挡时遇到困难。随着神经渲染技术的发展，NeRF [14–17] 通过可微体积渲染实现了高保真的三维重建，而 PixelNeRF [18] 将这一方法扩展到了单图像输入。为了提高效率，最近的方法如 3D 高斯喷绘（3DGS）[19–24] 采用显式的高斯原语进行实时渲染和编辑。尽管如此，直接从单一视角回归 NeRF 或 3DGS 表示仍然具有挑战性，这主要是由于几何歧义性和多视图一致性不足。

随着强大的图像条件扩散模型的兴起，DreamFusion [25] 及其后续作品 [26–28] 引入了得分蒸馏采样（SDS），该方法利用预训练的 2D 文本到图像扩散模型作为隐式能量函数来引导三维表示的优化。这一范式使文本驱动的三维生成无需明确的三维监督。虽然有效，但基于 SDS 的方法通常会遭受高计算成本、多面不一致和饱和外观的问题。另一条研究路线直接微调扩散模型以从单个输入合成结构上一致的多视角图像，如 Zero-1-to-3 [1]、Zero123++ [2] 和 MVDream [4] 所示。这些方法支持通过标准网格管道实现高效和高质量的三维重建。然而，它们假设输入完全可见，并且当目标对象部分被遮挡时会遇到困难。这突出显示了一个关键限制：现有模型缺乏关于遮挡结构进行推理并完成的能力，在实际应用中这种能力对于常见遮挡情

况至关重要。

2.2. 2D 无模完成

二维无模态补全旨在恢复图像中部分被遮挡对象的完整形状和外观 [10]。早期方法 [29, 30] 依赖于几何启发式方法——如欧拉螺旋线和贝塞尔曲线——基于预定义的遮挡顺序来外推被遮挡的边界。然而，这些方法仅限于简单形状，并且对于复杂的真实世界场景缺乏鲁棒性。后续的工作 [31–34] 采用合成数据集上的监督学习，但通常受限于特定的对象类别和遮挡模式。最近，随着生成模型的进步，一些方法 [5, 9, 35, 36] 使用强大的图像生成框架——如 Pix2Gestalt [5] 和 SynergyAmodal [10]——实现了有前景的零样本性能。

然而，直接将这些方法应用于 3D 重建存在若干限制。首先，2D 补全模型固有的多视角推理不足可能导致其扩展到 3D 时产生几何不一致的输出。其次，这些流水线将补全过程和 3D 重建阶段分开，阻碍了结构一致性的同时优化。为了克服这些限制，我们提出了一种端到端框架，该框架在生成过程中整合了 3D 意识，从而能够从单一遮挡图像中实现结构上一致的多视角补全。

2.3. 3D 去遮挡

早期的 3D 去遮挡尝试通常依赖于 2.5D 深度补全 [37, 38] 或基于模板的人体网格拟合 [39, 40]，这些方法通常需要 RGB-D 输入、身体先验知识或多帧观察。虽然在受控场景中有效，但这些方法难以推广到多样化的物体类别和复杂的遮挡模式。最近的方法探索了生成策略。CHROME [41] 采用姿态条件扩散模型合成被遮挡人物的多视角图像，然后通过高斯投射进行重建。OccFusion [42] 渲染粗糙的人体网格，并使用基于扩散的图像修复技术对其进行细化。Slice3D [43] 从遮挡视图预测横截面切片并将其组合成 3D 体积。

尽管取得了有希望的结果，这些方法往往依赖于密集的监督、任务特定的先验知识或定制的架构，限制了它们的可扩展性。相比之下，我们的方法直接从单个被遮挡的图像生成结构一致的新视角，并与现成的基于网格的重建管道无缝集成。

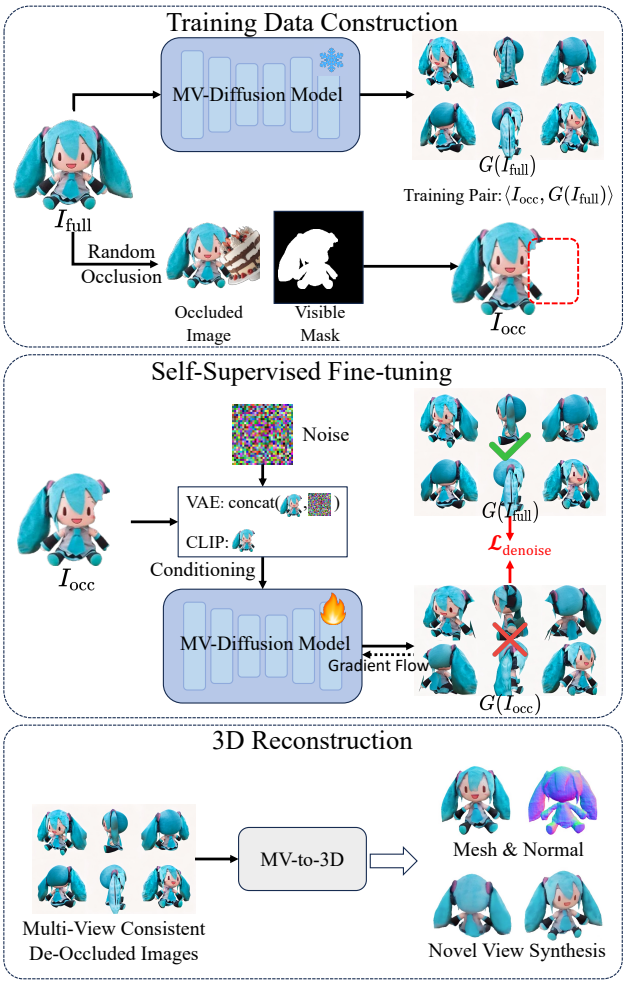


图 2. DeOcc-1-to-3 概述。**顶部**：通过向完整图像 I_{full} 应用随机遮挡生成被遮挡的图像 I_{occ}。一个冻结的多视角扩散模型产生六个视图伪地面真实值 $G(I_{\text{full}})$，形成训练对 $\langle I_{\text{occ}}, G(I_{\text{full}})\rangle$。**中间**：学生模型被完全微调以预测一致的新视图 $G(I_{\text{occ}})$，由去噪损失 $\mathcal{L}_{\text{denoise}}$ 监督。**底部**：预测的六视图图像被输入到下游重建模型中进行三维重建。

3. 方法

本节介绍了我们提出的遮挡感知多视图生成方法的整体框架，涵盖了训练数据构建（3.3）、自监督微调（3.4）以及与下游 3D 重建管道的集成（3.5）。无需修改原始视图合成架构，我们的方法通过全模型微调实现了结构完整性和一致性的新型视图生成。

3.1. 预备知识

我们基于 Zero123++ [2]，这是一个基于扩散的多视角生成模型，可以从单个 RGB 图像合成六个新的视图。输出格式为一个 3×2 拼贴图像，对应于由以下内容定义的六个预定义相机姿态：

- 高程角：$\{30°, -20°\}$
- 方位角：$\{30°, 90°, 150°, 210°, 270°, 330°\}$

给定输入图像 I_{input}，该模型使用基于速度的去噪目标进行训练：

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t \mid I_{\text{input}})\|^2\right], \quad (1)$$

其中 x_t 是噪声潜伏变量，ϵ_θ 表示模型预测的噪声。

Zero123++ 同时利用局部条件（通过缩放的参考注意力）和全局条件（通过 CLIP 图像嵌入），实现了合成视图之间的强空间一致性和语义一致性。这些特性使其成为我们的遮挡感知微调框架的强大且架构无关的基础。此外，由于 Zero123++ 是在显式的多视角监督下训练的，它固有地学习了多视角一致性，使其非常适合在遮挡情况下的视角一致生成任务。

3.2. 总体框架

先前的工作专注于从单张图像开始进行图像补全，然后生成三维模型，这通常会导致较差的三维重建完整性。这是因为此类二维去遮挡模型仅考虑二维图像补全，而没有考虑到三维结构完整性。受 MVDream [4] 和 Zero123++[2] 的启发，我们提出了一种基于多视角扩散模型的原生三维去遮挡框架。我们的方法使用伪多视图监督构建遮挡增强训练对，并微调视图合成模型以生成感知遮挡的新视图。然后将生成的图像传递到三维重建模块（例如 InstantMesh[3]）中，以恢复完整的三维对象。整个流程无需注释、类别无关且能稳健应对各种遮挡模式。具体来说，我们的目标是从一张被遮挡的图像中合成六张预定义的 RGB 视图，确保跨视图的结构和外观一致性，从而实现有效的下游三维重建。

如图 2 所示，提出的流水线包括三个阶段：(1) 使用伪多视图监督构建遮挡增强的训练对；(2) 微调一个多视图扩散模型以实现遮挡感知生成；(3) 将生成的视图输入到 3D 重建模块中以获得完整的 3D 模型。

3.3. 训练数据构建

真实世界三维数据的有限可用性也使得获得普遍一致的遮挡示例变得具有挑战性。虽然可以使用合成数据集 [44] 来生成此类数据，但它们通常会引入显著的领域差异，限制模型在现实世界场景中的泛化能力。相比之下，真实世界的二维数据丰富且多样。为了利用这一点，我们采用两阶段策略：首先，构建大规模二维遮挡数据集，然后使用预训练的多视图扩散模型生成用于监督的三维一致伪地面真值视图。

二维遮挡数据构建。我们基于 SA-1B 数据集构建感知遮挡的图像对 [45]，利用 Segment Anything Model (SAM) [45] 对前景物体进行分割。随机选择一个分割出的物体并将其覆盖到自然背景上，以合成遮挡场景。此过程产生：

- 包含完整前景对象 I_{raw} 的原始图像，
- 全前景对象掩码，M_{full}，
- 一个被遮挡的复合图像，I_{mix}，
- 和目标对象可见（未被遮挡）部分的掩码，M_{occ}。

然后，我们计算配对的前景图像如下：

$$I_{\text{full}} = I_{\text{raw}} \odot M_{\text{full}},$$
$$I_{\text{occ}} = I_{\text{mix}} \odot M_{\text{occ}},$$

其中 $\odot$ 表示逐元素相乘。

我们还应用过滤来移除前景对象本质上不完整的遮挡样本，防止模型学习生成不完整形状 [5]。此外，因为我们的 2D 数据集用于预测后续 3D 重建的伪真实值，我们将它整理以匹配所用多视角扩散模型的输入分布。特别是，我们排除了前景对象位于图像边界处的完整样本。最终的数据集包含 10 万份样本。

三维遮挡数据构建。我们首先将清晰图像 I_{full} 输入到预训练的多视角扩散模型 G（用作教师）中，生成具有高结构一致性的六视角伪真实图像 $G(I_{\text{full}})$。这使得可以构建用于三维去遮挡学习的配对训练样本 $\langle I_{\text{occ}}, G(I_{\text{full}})\rangle$。然后，我们使用 I_{occ} 作为条件输入来微调同一模型（学生），直接合成相应的新型视图。

为了增强鲁棒性并防止过度拟合到任何特定的遮挡模式，我们进一步从两个方面增强了训练集：(1) 我们对遮挡掩码应用膨胀和腐蚀操作以模拟更广泛的遮挡严重程度和形状，使模型能够学习来自重遮挡和轻遮挡场景；这些增强后的样本被包含在训练集中以提高鲁棒性；(2) 我们还将未被遮挡的前景对象子集

$\langle I_{\text{full}}, G(I_{\text{full}})\rangle$ 作为身份对纳入训练数据，以防模型在输入已经完整的情况下产生不必要的补全。

在过滤掉有缺陷或模糊的样本后，最终的 3D 去遮挡数据集包含大约 40K 个高质量训练对，涵盖了广泛的物体类别、遮挡水平和掩码模式。该流水线完全为自我监督，不需要诸如分割掩码或深度图等人工标签。

3.4. 模型结构与训练策略

如图 2 所示，我们采用了一个基于扩散的多视角生成主干网络 G（例如，Zero123++），并且没有修改其架构。该模型以单个被遮挡的 RGB 图像 I_{occ} 作为输入，并产生一个代表六个预定义新视图的拼接图像 3×2，表示为 $G(I_{\text{occ}})$。重要的是，在感知遮挡的生成过程中没有使用文本提示，遵循我们确保跨类别和遮挡类型的强大可用性和泛化能力的设计原则。

训练采用标准的去噪目标：

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t \mid I_{\text{occ}})\|^2\right],$$

其中，x_t 是伪真实值 $G(I_{\text{full}})$ 的噪声版本，而 ϵ_θ 是在时间步 t 处预测的噪声。

我们微调整个 U-Net，包括残差块和注意力模块。训练使用 AdamW 优化器，初始学习率为 2×10^{-5}，批量大小为 32，并且总共进行 150k 步。我们保留与原始模型相同的噪声调度以及局部（参考注意力）和全局（基于 CLIP 的图像编码器）条件以确保训练稳定性和生成一致性。

为进一步提高生成视图的鲁棒性和视觉保真度，我们在训练过程中维护模型权重 θ 的指数移动平均值：

$$\theta_{\text{EMA}} \leftarrow \beta \cdot \theta_{\text{EMA}} + (1 - \beta) \cdot \theta,$$

衰减率为 $\beta = 0.9999$。在推理时使用 EMA 权重以提升生成稳定性和质量。

3.5. 三维重建集成

所提出的感知遮挡的多视图生成框架合成了六视角 RGB 图像，这些图像不仅几何上一致而且结构完整。这些图像可以作为各种三维重建或生成方法 [3, 46, 47] 的通用输入，使能在不同的表示中恢复三维模型，例如 NeRF[14]、3D 高斯散射 [19] 和基于网格的表面。

在我们的实现中，我们采用 InstantMesh [3] 作为代表性的后端，因为其高效性和生成密封网格的能力。InstantMesh 从输入视图提取三平面特征，并执行稀疏视图几何推理以解码三维网格。值得注意的是，由于我们的模型保持与原始多视图扩散架构相同的输出格式，因此可以无缝集成到现有管道中而无需任何架构修改。实验结果表明，我们的遮挡感知生成器在遮挡场景中显著提升了三维重建性能，产生更完整、一致且平滑的表面。

3.6. Occ-LVIS 基准测试

为了系统地评估遮挡下的 3D 物体重建，我们基于 Objaverse-LVIS 数据集构建了一个新的基准测试 [48]。具体来说，我们利用了 LVIS 子集中高质量的资源，并在受控遮挡设置下将它们渲染成多视角图像序列。

每个对象首先从规范的正面视图渲染，然后使用随机选择的前景对象进行遮挡。然后我们生成：

- **六种规范视图** 的方位角为 {30°, 90°, 150°, 210°, 270°, 330°}，仰角交替为 {30°, -20°}；
- **四个随机视图** 用于评估对未见视角的泛化能力；
- **几何完整的网格地面真实值** ，用于评估三维重建的保真度。

为了在不同难度下进行分析，我们进一步根据遮挡比例（即目标对象可见区域的比例）将基准测试划分为五个遮挡级别：

表 1. 遮挡级别统计在 Occ-LVIS 标准中的情况。

级别	遮挡比例范围	比例（%）
L1	0% – 10%	12.3%
L2	10% – 20%	28.5%
L3	20% – 30%	31.6%
L4	30% – 40%	19.4%
L5	$\geq$ 40%	8.2%

基准中的每个对象都附带其遮挡级别标签，从而可以在不同的遮挡场景下进行细粒度的模型鲁棒性分析。

基准和评估协议旨在为遮挡感知的 3D 生成方法提供一个标准化和全面的平台。

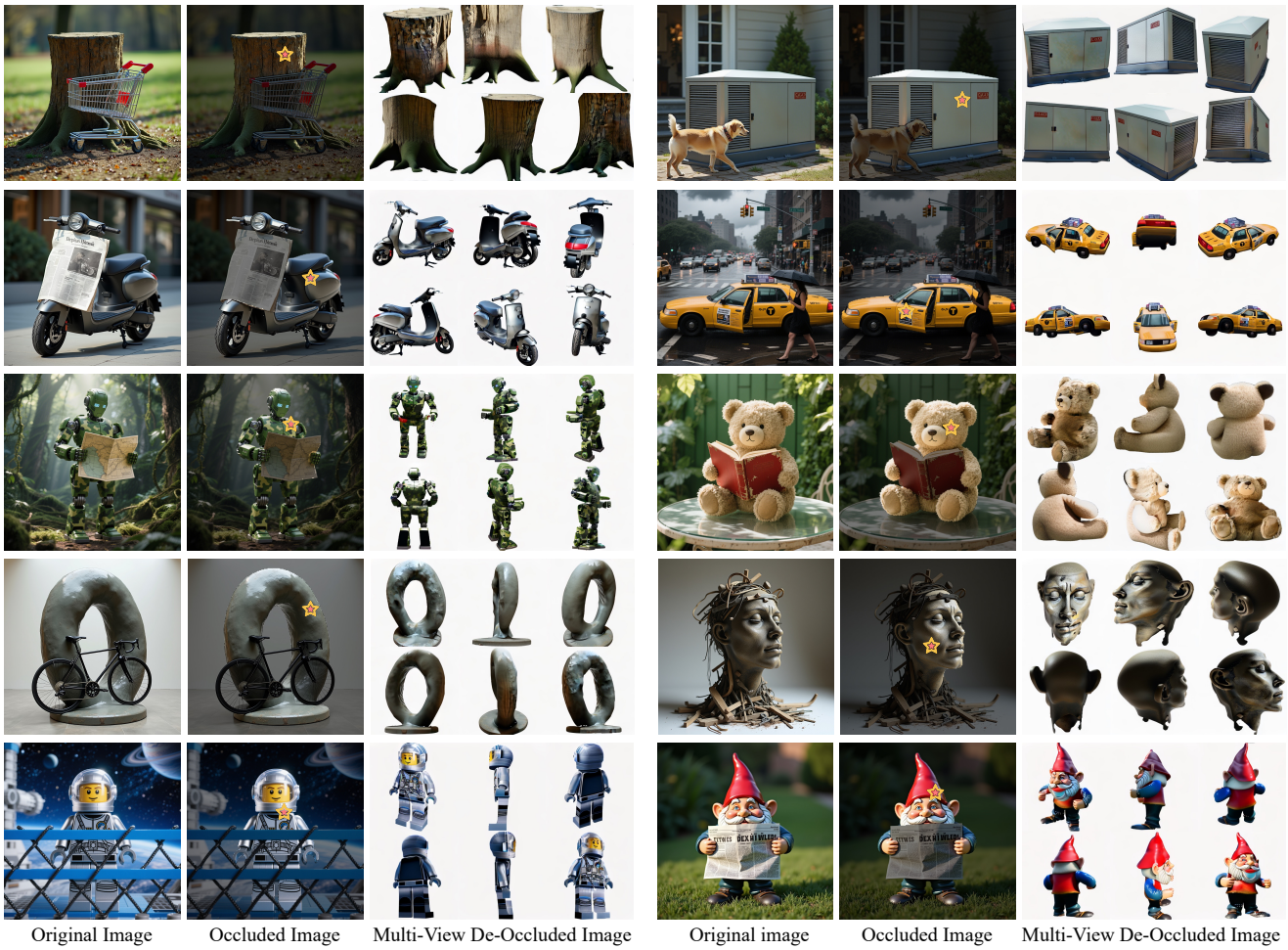


图 3. 定性去遮挡结果在多种对象上的表现。每个三联图显示（左）原始图像，（中）被遮挡的输入，和（右）我们的多视角去遮挡输出（六个视角）。我们的方法恢复了各种形状、材质和遮挡类型下的连贯几何结构和纹理。

4. 实验

4.1. 评估设置

数据集。我们在提出的 Occ-LVIS 基准（第 3.6 节）上进行了定量实验，并使用互联网图像和合成生成的图像进行定性评估。

基线。我们将我们的方法与以下基线进行比较：(1) 图像到三维流水线 (3DRecon)：使用 Vanilla Zero123++ [2]，随后是 InstantMesh [3]，不处理任何遮挡。(2) 二维去遮挡方法 + 图像到三维流水线 (P2G-3DRecon)：使用 Pix2Gestalt [5] 进行二维无模态补全，随后是 Zero123++ 和 InstantMesh。(3) 三维去遮挡流水线 (我们的方法)：我们提出的方法，该方法联合执行感知遮挡的视图合成和下游三维重建。

评估指标。我们从二维和三维两个角度评估生成结果的质量。

二维评估：我们使用以下方法评估生成的多视角图像的视觉保真度：

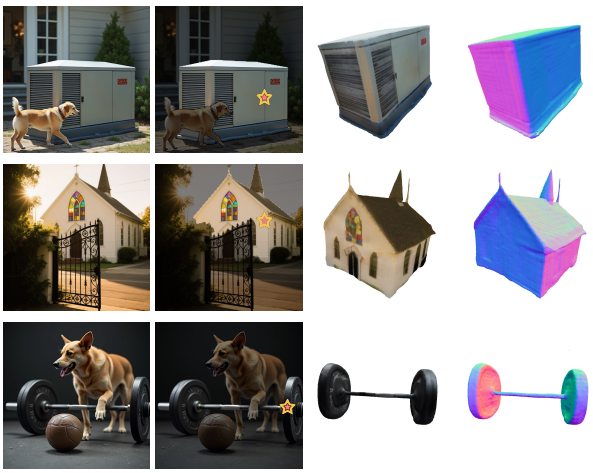
- Fréchet **初始距离**（FID）[49]：衡量真实图像和合成图像特征分布之间的距离。数值越低表示分布统计上的对齐程度越高。
- **核初始距离**（KID）[50]：度量真实图像特征与生成图像特征之间的平方最大均值差异，采用多项式核函数。
- CLIP **分数** [51]：通过计算预训练的 CLIP 模型的余弦相似度来评估真实图像和生成图像之间的语义一致性。

三维评估：为了评估几何重建的质量，我们将生

图 4. 定性比较在 Occ-LVIS 基准上的去遮挡结果。我们对比了 3DRecon [2, 3], P2G-3DRecon [2, 3, 5] 和我们的方法。3DRecon 无法恢复被遮挡的内容，而 P2G-3DRecon 则遭受纹理退化和偶尔的失败。我们的方法产生了一致且高质量的去遮挡视图，实现了高质量的 3D 去遮挡。

成的网格和真实值在原点为中心的单位球内进行对齐，并保持一致的坐标系统。我们报告：

- **切角距离**（CD）：计算在预测网格和参考网格表面均匀采样的点之间的平均双向距离。它反映了表面级别的重建保真度，数值越低表示对应关系越好。
- **F 值**：在固定距离阈值下，定义为精度和召回率的调和平均值，该指标评估了重建几何形状的完整性和准确性。在我们的评估中，我们报告了 F 分数。
- **体积交并比**（V-IoU）：测量预测形状与真实形状之间的体积重叠。



(a) Natural Occlusion

(b) Manual Occlusion

图 5. 定性 3D 重建结果使用预测的多视图去遮挡图像在（a）自然和（b）手动遮挡下。对于每个对象，我们显示（左）原始图像，（第二）被遮挡的输入，（第三）重建的网格，以及（右）对应表面法线。尽管存在真实世界和合成遮挡，我们的方法成功恢复了完整的几何形状和准确的法线，证明我们的模型实现了多视图一致的遮挡补全。

4.2. 定量结果

我们在 Occ-LVIS 基准上报告了二维图像质量指标和三维几何保真度指标，以评估在遮挡输入设置下我们方法的有效性。

表 2. 在 Occ-LVIS 基准上的二维比较。

Method	CLIP ↑	FID ↓	KID ↓
3DRecon	0.7430	41.3836	0.01361
P2G-3DRecon	0.7833	30.1892	0.0043
Ours	0.7892	29.0836	0.0035

表 3. 在 Occ-LVIS 基准上的 3D 比较。

Method	CD ↓	F-Score ↑	V-IoU ↑
3DRecon	0.0125	0.3721	0.1565
P2G-3DRecon	0.0106	0.4470	0.3232
Ours	0.0086	0.4835	0.3445

二维结果。如表 2 所示，我们的方法在所有指标中均表现最佳。具体来说，它获得了最低的 FID 和 KID 分数，这表明了更好的感知质量和与真实图像分布的一致性。此外，我们的方法还获得了最高的 CLIP 相似度得分，证明了生成视图与输入之间的语义一致性更强。相比于基线管道（三维重建）和两阶段完成管道（P2G-3D 重建），我们的遮挡感知生成框架始终能够提高多视角质量和连贯性。

3D **结果**。我们还使用 Chamfer 距离 (CD)、F-Score 和 V-IoU 来评估重建的 3D 网格的保真度，如表 3 所示。我们的方法实现了最低的 CD，表明表面重建准确，并且 F-Score 和 V-IoU 最高，这反映了更好的几何完整性和体积一致性。这些结果证实了我们多视图生成在跨视图时保留结构完整性并为下游 3D 重建提供高质量监督。

4.3. 定性结果

如图 3 所示，我们展示了定性结果，证明了我们的遮挡感知多视图生成框架在广泛的对象类别和遮挡场景中的有效性。实验是在未见过的测试样本上进行的，包括合成遮挡和现实世界的遮挡照片。值得注意的是，一些案例涉及超过对象区域 40%的严重遮挡以及具有挑战性的视觉结构——例如，第三列显示了一个“抱着书的泰迪熊”，第四列表示一个“用柴火堆叠的艺术装置”。尽管存在这些复杂性，我们的方法成功地在所有

六个视图中恢复了连贯的对象几何和外观。合成输出表现出强烈的多视角一致性，最小化了伪影，并合理地完成了遮挡区域。这些结果突显了我们微调模型的鲁棒性和泛化能力，即使在高度遮挡和分布外条件下也是如此。

定性比较。如图 4 所示，我们展示了在提出的 OccLVIS 基准测试上的定性比较结果。三维重建基线由于缺乏遮挡补全机制，无法恢复缺失内容，导致损坏或不完整的 3D 重建。虽然 P2G-3D 重建流程可以补全大部分遮挡区域，但它受到跨阶段错误累积的影响。这导致明显的纹理退化（例如，第一行中扭曲的汽车油漆和第三行中混乱的蛋糕颜色），以及偶尔的失败案例（例如，第六行中缺失的水瓶内容）。相比之下，我们的方法始终生成高质量、结构一致的去遮挡视图，涵盖各种遮挡类型和物体类别，从而实现更忠实和完整的 3D 重建。

为了进一步验证我们的去遮挡图像生成的质量和多视角一致性，我们利用 InstantMesh [3] 从预测的六视图输出中重构 3D 几何。我们还可视化了所得网格的表面法线以进行定性检查。

如图 5 所示，我们的方法能够在各种场景中实现高质量的 3D 重建，生成几何一致的多视图补全和逼真的物体表面。这些结果证实了我们生成的视图不仅能够合理地完成被遮挡的内容，还能作为下游基于网格的 3D 建模的可靠输入。

5. 结论

我们提出了 DeOcc-1-to-3，这是一种从单一遮挡图像进行多视角三维去遮挡的自监督范式。通过微调一个多视图扩散模型，我们的方法能够同时完成缺失结构并合成几何一致的新视图。它不需要架构修改，可以无缝集成到现有的三维重建流程中，并且很好地泛化到了各种遮挡场景。为了支持未来的研究，我们还引入了一个基准测试，用于标准化评估具有遮挡意识的三维建模。

参考文献

[1] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 9298–9309, 2023.

[2] R. Shi, H. Chen, Z. Zhang, M. Liu, C. Xu, X. Wei, L. Chen, C. Zeng, and H. Su, "Zero123++: a single image to consistent multi-view diffusion base model," arXiv preprint arXiv:2310.15110, 2023.

[3] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan, "Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models," arXiv preprint arXiv:2404.07191, 2024.

[4] Y. Shi, P. Wang, J. Ye, M. Long, K. Li, and X. Yang, "Mvdream: Multi-view diffusion for 3d generation," arXiv preprint arXiv:2308.16512, 2023.

[5] E. Ozguroglu, R. Liu, D. Surís, D. Chen, A. Dave, P. Tokmakov, and C. Vondrick, "pix2gestalt: Amodal segmentation by synthesizing wholes," in 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3931–3940, IEEE Computer Society, 2024.

[6] K. Xu, L. Zhang, and J. Shi, "Amodal completion via progressive mixed context diffusion," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9099–9109, 2024.

[7] A. Dogaru, M. Özer, and B. Egger, "Generalizable 3d scene reconstruction via divide and conquer from a single view," arXiv preprint arXiv:2404.03421, 2024.

[8] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin, and C. C. Loy, "Self-supervised scene de-occlusion," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3784–3792, 2020.

[9] G. Zhan, C. Zheng, W. Xie, and A. Zisserman, "Amodal ground truth and completion in the wild," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 28003–28013, 2024.

[10] X. Li, C. Yi, J. Lai, M. Lin, Y. Qu, S. Zhang, and L. Cao, "Synergyamodal: Deocclude anything with text control," arXiv preprint arXiv:2504.19506, 2025.

[11] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction," in Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11-14, 2016, proceedings, part VIII 14, pp. 628–644, Springer, 2016.

[12] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges,

and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in 2011 10th IEEE international symposium on mixed and augmented reality, pp. 127–136, Ieee, 2011.

[13] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 165–174, 2019.

[14] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," Communications of the ACM, vol. 65, no. 1, pp. 99–106, 2021.

[15] Y. Wang, J. Wang, Y. Qu, and Y. Qi, "Rip-nerf: Learning rotation-invariant point-based neural radiance field for fine-grained editing and compositing," in Proceedings of the 2023 ACM international conference on multimedia retrieval, pp. 125–134, 2023.

[16] Y. Qu, Y. Wang, and Y. Qi, "Sg-nerf: Semantic-guided point-based neural radiance fields," in 2023 IEEE International Conference on Multimedia and Expo (ICME), pp. 570–575, IEEE, 2023.

[17] C. Huang, X. Li, S. Zhang, L. Cao, and R. Ji, "Nerf-dets: Enhancing multi-view 3d object detection with sampling-adaptive network of continuous nerf-based representation," arXiv e-prints, pp. arXiv–2404, 2024.

[18] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelnerf: Neural radiance fields from one or few images," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4578–4587, 2021.

[19] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering.," ACM Trans. Graph., vol. 42, no. 4, pp. 139–1, 2023.

[20] Y. Shen, Z. Zhang, X. Li, Y. Qu, Y. Lin, S. Zhang, and L. Cao, "111," 2025.

[21] Y. Shen, Z. Zhang, X. Li, Y. Qu, Y. Lin, S. Zhang, and L. Cao, "Evolving high-quality rendering and reconstruction in a unified framework with contribution-adaptive regularization," arXiv preprint arXiv:2503.00881, 2025.

[22] Y. Wang, J. Wang, R. Gao, Y. Qu, W. Duan, S. Yang, and Y. Qi, "Look at the sky: Sky-aware efficient 3d gaussian splatting in the wild," IEEE Transactions on Visualization and Computer Graphics, 2025.

[23] Y. Guo, J. Hu, Y. Qu, and L. Cao, "Wildseg3d: Segment any 3d objects in the wild from 2d images," arXiv preprint arXiv:2503.08407, 2025.

[24] S. Dai, Y. Qu, Z. Li, X. Li, S. Zhang, and L. Cao, "Training-free hierarchical scene understanding for gaussian splatting with superpoint graphs," arXiv preprint arXiv:2504.13153, 2025.

[25] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," arXiv preprint arXiv:2209.14988, 2022.

[26] X. Li, Z. Lai, L. Xu, Y. Qu, L. Cao, S. Zhang, B. Dai, and R. Ji, "Director3d: Real-world camera trajectory and 3d scene generation from text," Advances in Neural Information Processing Systems, vol. 37, pp. 75125–75151, 2024.

[27] Y. Qu, D. Chen, X. Li, X. Li, S. Zhang, L. Cao, and R. Ji, "Drag your gaussian: Effective drag-based editing with score distillation for 3d gaussian splatting," arXiv preprint arXiv:2501.18672, 2025.

[28] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu, "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," Advances in Neural Information Processing Systems, vol. 36, pp. 8406–8441, 2023.

[29] B. B. Kimia, I. Frankel, and A.-M. Popescu, "Euler spiral for shape completion," International journal of computer vision, vol. 54, no. 1, pp. 159–182, 2003.

[30] N. Silberman, L. Shapira, R. Gal, and P. Kohli, "A contour completion model for augmenting surface reconstructions," in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13, pp. 488–503, Springer, 2014.

[31] N. Zhang, N. Liu, J. Han, K. Wan, and L. Shao, "Face de-occlusion with deep cascade guidance learning," IEEE Transactions on Multimedia, vol. 25, pp. 3217–3229, 2022.

[32] X. Yan, F. Wang, W. Liu, Y. Yu, S. He, and J. Pan, "Visualizing the invisible: Occluded vehicle segmentation and recovery," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7618–7627, 2019.

[33] Q. Zhou, S. Wang, Y. Wang, Z. Huang, and X. Wang, "Human de-occlusion: Invisible perception and recovery for humans," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3691–3701, 2021.

[34] D. P. Papadopoulos, Y. Tamaazousti, F. Ofli, I. Weber, and A. Torralba, "How to make a pizza: Learning a compositional layer-based gan model," in proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8002–8011, 2019.

[35] Z.-P. Duan, J. Zhang, S. Liu, Z. Lin, C.-L. Guo, D. Zou, J. Ren, and C. Li, "A diffusion-based framework for occluded object movement," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 2816–2824, 2025.

[36] J. J. Lee, B. Benes, and R. A. Yeh, "Tuning-free amodal segmentation via the occlusion-free bias of inpainting models," arXiv preprint arXiv:2503.18947, 2025.

[37] Y. Zhang and T. Funkhouser, "Deep depth completion of a single rgb-d image," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 175–185, 2018.

[38] F. Ma, G. Cavalheiro, and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 1–8, IEEE, 2018.

[39] T. Alldieck, M. A. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3d people models," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8387–8397, 2018.

[40] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10975–10985, 2019.

[41] A. Dutta, M. Zheng, Z. Gao, B. Planche, A. Choudhuri, T. Chen, A. K. Roy-Chowdhury, and Z. Wu, "Chrome: Clothed human reconstruction with occlusion-resilience and multiview-consistency from a single image," arXiv preprint arXiv:2503.15671, 2025.

[42] A. Sun, T. Xiang, S. Delp, L. Fei-Fei, and E. Adeli, "Occfusion: Rendering occluded humans with generative diffusion priors," arXiv preprint arXiv:2407.00316, 2024.

[43] Y. Wang, W. Lira, W. Wang, A. Mahdavi-Amiri, and H. Zhang, "Slice3d: Multi-slice occlusion-revealing single view 3d reconstruction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9881–9891, 2024.

[44] Y.-T. Hu, H.-S. Chen, K. Hui, J.-B. Huang, and A. G. Schwing, "Sail-vos: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3105–3115, 2019.

[45] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., "Segment anything," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 4015–4026, 2023.

[46] J. Xiang, Z. Lv, S. Xu, Y. Deng, R. Wang, B. Zhang, D. Chen, X. Tong, and J. Yang, "Structured 3d latents for scalable and versatile 3d generation," in Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 21469–21480, 2025.

[47] Z. Zhao, Z. Lai, Q. Lin, Y. Zhao, H. Liu, S. Yang, Y. Feng, M. Yang, S. Zhang, X. Yang, et al., "Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation," arXiv preprint arXiv:2501.12202, 2025.

[48] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13142–13153, 2023.

[49] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," Advances in neural information processing systems, vol. 30, 2017.

[50] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," arXiv preprint arXiv:1801.01401, 2018.

[51] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning, pp. 8748–8763, PmLR, 2021.