

ViewCrafter: 驯服视频扩散模型以实现高保真新视角合成

ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis

Wangbo Yu*, Jinbo Xing*, Li Yuan*, Wenbo Hu†, Xiaoyu Li, Zhipeng Huang,
Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian†, *Fellow, IEEE*

摘要—神经三维重建近来虽已取得长足进展，但它对稠密多视角采集的依赖限制了应用范围。本文提出 **ViewCrafter**，一种借助视频扩散模型先验、从单张或稀疏图像合成通用场景高保真新视角的方法。该方法一方面发挥视频扩散模型强大的生成能力，另一方面利用基于点的表示所提供的粗略三维线索，从而在精确控制相机位姿的前提下生成高质量视频帧。为进一步扩大新视角的生成范围，我们量身设计了一套迭代式视角合成策略，并辅以相机轨迹规划算法，逐步扩展三维线索与新视角覆盖的区域。借助 **ViewCrafter**，我们能够支撑多种应用：例如利用重建出的三维点与生成的新视角高效优化 3D 高斯泼溅 (3D-GS) 表示，实现实时渲染的沉浸式体验；又如面向更富想象力内容创作的场景级文本到 3D 生成。在多样数据集上的大量实验表明，本方法在合成高保真且一致的新视角上具有很强的泛化能力与优越性能。项目主页与代码见 <https://drexubery.github.io/ViewCrafter/>。

Index Terms—新视角合成，视频扩散模型，三维场景生成。

1 引言

视角合成在计算机视觉与图形学中扮演着关键角色，可为新游戏、混合现实和视觉特效创造沉浸式体验。尽管 NeRF [1]、3D-GS [2] 等三维神经重建技术已取得显著成功，但它们依赖稠密的多视角观测，在仅有少量视角可用的场景中适用性受限。

实践中更理想的问题设定，是从稀疏观测乃至单张图像出发，合成通用场景的新视角。这一任务相当困难，因为它要求模型全面理解三维世界，包括结构、外观、语义以及遮挡关系。早期研究 [3], [4], [5], [6], [7], [8], [9], [10] 着眼于训练基于回归的模型，从稀疏或单张输入合成新视角。然而受限于表示能力，这类方法大多针对特定类别，只能处理室内场景等有限领域。近来强大的扩散模型不断进步，使从单一视角进行零样本新视角合成 [11], [12], [13] 成为可能。但这些方法要么只能处理物体级图像，要么因依赖高层位姿提示来引导视

角合成而无法精确控制相机位姿。也有一些工作 [14], [15] 尝试用基于深度的图像变形加扩散式修补，从单张图像合成新视角。然而受制于逐视角修补的机制，它们在遮挡区域往往生成前后不一致的内容。

本文聚焦于从单张或稀疏图像出发、对通用场景进行高保真新视角合成，同时保持对相机位姿的精确控制以及所生成新视角的一致性。为此，我们探索将视频扩散模型的生成能力与点云表示所提供的显式三维信息结合起来。一方面，在网络规模视频数据上训练的视频扩散模型 [16], [17], [18] 形成了对世界较为合理的理解，能够从单张图像或文本提示生成合理的视频内容。但它们缺乏场景底层的三维信息，难以实现精确的相机视角控制。另一方面，近来的稠密立体方法 [19], [20] 让从单张或稀疏图像快速重建点云成为可能。点云表示提供了宝贵的粗略三维场景信息，可支持自由视角渲染中的精确位姿控制。但受限于其表示能力不足，加之极稀疏参考图像所能提供的三维线索有限，点云会出现遮挡、缺失区域和几何畸变，妨碍其在新视角合成中的应用。有鉴于此，我们提出把视频扩散模型的生成能力与基于点的表示所提供的粗略三维先验融合起来，以实现通用场景更高保真的新视角合成。

我们的方法 **ViewCrafter** 通过一个点云条件视频扩散模型来完成新视角合成：以从单张或稀疏图像重建的点云所渲染的对应帧为条件，沿新视角轨迹生成高保真、一致的视

- *: Equal contribution. †: Corresponding authors. Wangbo Yu, Li Yuan, Zhipeng Huang, Yonghong Tian are with Peking University and Peng Cheng Laboratory. Jinbo Xing is with The Chinese University of Hong Kong. Wenbo Hu, Xiaoyu Li are with Tencent AI Lab. Ying Shan is with ARC Lab, Tencent PCG and Tencent AI Lab. Xiangjun Gao is with Hong Kong University of Science and Technology. Tien-Tsin Wong is with Monash University.
- 本文为 AI 辅助翻译版本，仅供学习交流；引用请以英文原文为准：arXiv:2409.02048。

频。借助点云提供的显式三维信息与视频扩散模型的生成能力，本方法能够精确控制 6 自由度相机位姿，并生成高保真、一致的新视角。此外，由于内存和计算开销会不可接受地激增，视频扩散模型在生成视频时面临困难。为应对这一挑战，我们提出一套迭代式视角合成策略，并配以内容自适应的相机轨迹规划算法，逐步扩展重建出的点云以及新视角覆盖的区域。从输入图像得到的初始点云出发，我们先用相机轨迹规划算法，基于当前点云预测一条内容自适应的相机位姿序列，以有效揭示遮挡区域。随后依照预测的位姿序列渲染点云，并以渲染结果为条件用 ViewCrafter 合成新视角；接着再用合成出的新视角更新点云，扩展全局点云表示。反复执行上述步骤，最终即可得到覆盖大视角范围的高保真新视角以及一份扩展后的点云。

除新视角合成外，我们还探索了本方法的若干应用。例如，基于构建好的点云和合成的新视角，我们可以在数分钟内高效优化一个 3D-GS 表示，从而实现面向沉浸式体验的实时渲染。此外，本方法还展现出适配场景级文本到 3D 生成的潜力，可催生更富想象力的艺术创作。

我们在 Tanks-and-Temples [21]、RealEstate10K [7] 和 CO3D [22] 等多个数据集上，对本方法在零样本新视角合成与稀疏视角 3D-GS 重建两方面进行了充分评测。在零样本新视角合成上，本方法在图像质量与位姿精度两类指标上均优于基线，表明其在合成高保真新视角、实现精确位姿控制上更胜一筹。在 3D-GS 重建上，本方法始终超越以往最先进方法，进一步印证了它在稀疏视角场景重建中的有效性。

本文的贡献可归纳如下：

- 我们提出 **ViewCrafter**，一个专为从单张或稀疏图像合成通用场景高保真新视角序列、并能精确控制相机位姿而设计的新视角合成框架。
- 我们提出一套迭代式视角合成策略，并结合内容自适应的相机轨迹规划算法，逐步扩展新视角覆盖区域与重建点云，从而实现长距离、大范围的新视角合成。
- 本方法在多个具有挑战性的数据集上，无论是合成新视角的质量还是相机位姿控制的精度都取得了优越表现。它还能支撑新视角合成之外的多种应用，例如从我们的结果高效优化 3D-GS 表示以实现面向沉浸式体验的实时渲染，以及面向更富想象力艺术创作的场景级文本到 3D 生成。

2 相关工作

2.1 基于回归的新视角合成

基于回归的方法旨在训练一个前馈模型，从稀疏或单张图像输入生成新视角。这通常借助基于 CNN/Transformer [23] 的

架构，为输入图像建立三维表示来实现。例如，一些工作 [24], [25] 把这一思路用于人脸等特定模态，通过生成三平面表示来做新视角合成。LRM [26] 将该策略推广到通用物体；[7], [8], [9] 等方法采用多平面表示；PixelNeRF [10] 则以 NeRF [1] 作为三维表示来合成新视角。受 3D-GS [2] 成功的启发，PixelSplat [27]、MVSplat [28] 等近期方法探索训练基于回归的模型来输出三维高斯表示，以获得实时渲染能力。此外，[3], [4], [5], [6] 等方法将单目深度估计与图像修补模块统一到一个框架中来合成新视角。然而，这些方法局限于特定类别的领域（如物体和室内场景），且因模型表示能力有限而易产生伪影。相比之下，本方法能够合成通用场景的高保真新视角。

2.2 基于扩散模型的新视角合成

扩散模型 [29], [30], [31] 的迅猛发展在合成高质量图像上展现出卓越能力，也显示出可用于从单张或稀疏输入合成新视角的潜力。一些基于优化的方法 [32], [33], [34] 直接在文本到图像 (T2I) 扩散模型 [31] 的监督下训练三维表示，但它们需要针对具体场景进行优化，损害了泛化能力。为解决这一问题，GenVS [35] 在大规模多视角数据集 [22] 上训练一个以三维特征为条件的扩散模型，提出了一个可泛化的新视角合成框架。类似地，Zero-1-to-3 [11] 在合成数据集 [36], [37] 上训练以相机位姿为条件的扩散模型，能够从更多样的输入合成新视角。然而这些模型要么针对特定类别 [35], [38]，要么只能处理背景简单的玩具级物体。近来，ZeroNVS [12] 在同时包含合成数据 [36] 与真实数据 [7], [39], [40] 的混合数据集上训练，提升了 Zero-1-to-3 [11] 的生成能力，实现了从单张输入图像对通用场景的零样本新视角合成。不过，由于它把相机位姿条件当作高层文本嵌入处理，仍难以合成一致的新视角，也缺乏精确的位姿控制。Reconfusion [41] 提出一个以 PixelNeRF [10] 特征为条件的扩散模型，在新视角合成中实现了相对精确的位姿控制；但由于无法建模采样视角之间的关联，它未能合成一致的新视角，而且需要多张图像作为输入，无法处理单张图像。还有一些工作 [14], [15], [42] 利用基于深度的图像变形来合成新视角，并用预训练的 T2I 扩散模型 [43] 来修饰变形后的图像；然而这些方法生成的新视角在修补区域常出现伪影和不真实的内容，限制了其适用性。

2.3 条件式视频扩散模型

随着从多种输入模态生成内容的生成模型快速演进，增强用户对生成过程的控制受到了广泛关注。ControlNet [44]、T2I-adapter [45] 和 GLIGEN [46] 率先为 T2I 生成引入了条件信号。类似策略也被用于视频生成，可实现 RGB 图像 [17], [18], [47]、深度 [48], [49]、轨迹 [50], [51] 和语义图 [52] 等条件控制。然而，相机运动控制受到的关注相对较少。AnimateDiff

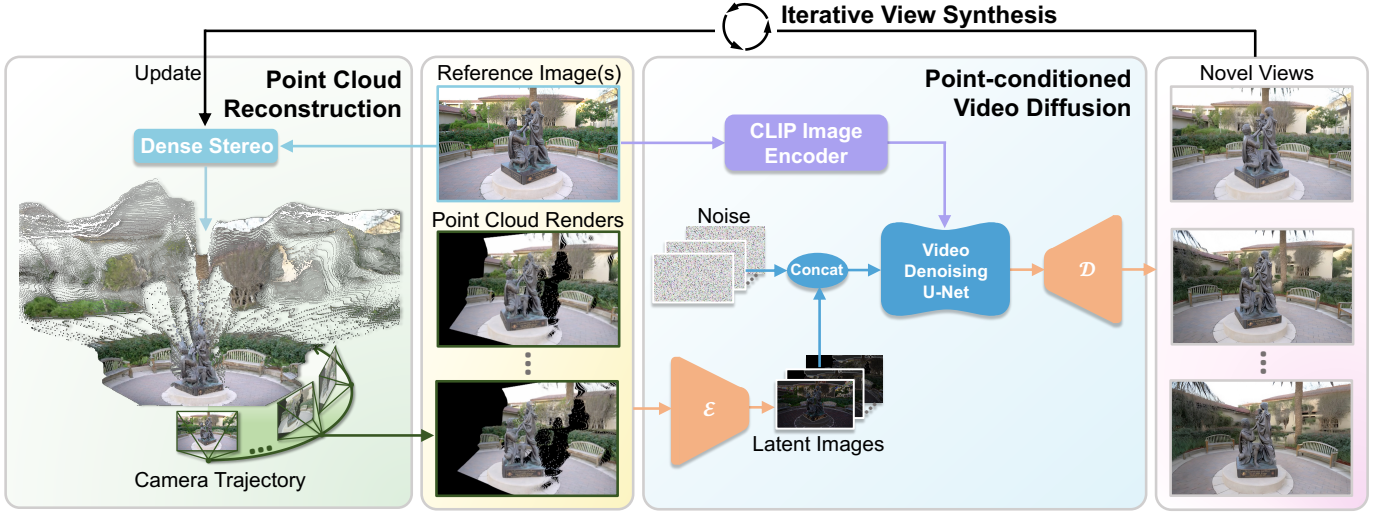


图 1. **ViewCrafter 方法总览**。给定单张参考图像或稀疏图像集，我们首先用稠密立体模型构建其点云表示，从而能够精确移动相机进行自由视角渲染。随后，为解决点云渲染结果中出现的大面积缺失区域、几何畸变和点云伪影，我们训练一个点云条件视频扩散模型作为增强渲染器，基于粗略的点云渲染生成高保真、一致的新视角。为实现长距离新视角合成，我们采用迭代式视角合成策略，反复移动相机、生成新视角并更新点云，从而实现更完整的点云重建，并惠及 3D-GS 优化等下游任务。

[53] 和 SVD [17] 研究类别条件的视频生成，将相机运动分组，并利用 LoRA [54] 模块生成特定的相机运动。MotionCtrl [13] 以相机外参矩阵作为条件信号，改进了控制能力；但它依赖一维数值，虽对简单轨迹有效，在复杂的真实场景中却难以精确控制。MultiDiff [55] 利用基于深度的图像变形生成变形图像，并以变形图像为视频扩散模型的条件来提供显式三维先验；然而它在特定类别的数据集 [7], [56] 上训练视频扩散模型，因而缺乏处理通用场景的泛化能力。近来，CamCo [57] 和 CameraCtrl [58] 在视频扩散模型中引入普吕克坐标 [59] 来控制相机运动；但由于从数值化的相机参数到视频之间的映射十分复杂，这些方法仍无法精确控制相机运动。本文提出利用显式的点云表示来实现视频生成中的精确相机控制，从而满足我们对一致且精确的新视角合成的需求。

3 方法

下文中，我们先在第 3.1 节简要介绍视频扩散模型，随后在第 3.2 节说明点云重建流程，并在第 3.3 节阐述点云条件视频扩散模型。接着，我们在第 3.4 节详述迭代式视角合成与相机轨迹规划策略，并在第 3.5 节展示如何将本方法用于高效的 3D-GS 优化与文本到 3D 生成。

3.1 预备知识：视频扩散模型

扩散模型 [30] 主要由两个部分组成：前向过程 q 和反向过程 p_θ 。前向过程从干净数据 $x_0 \sim q_0(x_0)$ 出发，逐步向 x_0 注入噪声，在不同时间步形成含噪状态，数学上表示为

$x_t = \alpha_t x_0 + \sigma_t \epsilon$ ，其中 $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 。超参数 α_t 与 σ_t 满足约束 $\alpha_t^2 + \sigma_t^2 = 1$ 。反向过程 p_θ 则借助噪声预测器 ϵ_θ 从含噪数据中去除噪声，其优化目标为：

$$\min_{\theta} = \mathbb{E}_{t \sim \mathcal{U}(0,1), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\epsilon_\theta(x_t, t) - \epsilon\|_2^2]. \quad (1)$$

在基于扩散的视频生成中，人们常采用潜在扩散模型 (Latent Diffusion Model, LDM) [60] 以降低计算开销。在 LDM 中，视频数据 $x \in \mathbb{R}^{L \times 3 \times H \times W}$ 通过预训练的 VAE 编码器逐帧编码到潜在空间，记作 $z = \mathcal{E}(x)$ ， $z \in \mathbb{R}^{L \times C \times h \times w}$ 。随后，前向过程 q 与反向过程 p_θ 均在潜在空间中进行。最终生成的视频通过 VAE 解码器得到，即 $\hat{x} = \mathcal{D}(z)$ 。本文中，我们的模型基于开源的图像到视频 (I2V) 扩散模型 DynamiCrafter [18] 构建，它能够从单张输入图像生成动态视频，这与我们从稀疏或单张输入合成新视角的目标天然契合。

3.2 从单张或稀疏图像重建点云

为在新视角合成框架中实现精确的位姿控制，我们首先从参考图像构建点云表示。具体而言，我们采用稠密立体模型 (e.g. DUS3R [19]) 来同时重建点云并估计相机参数。它以一对 RGB 图像 $\mathbf{I}^0, \mathbf{I}^1 \in \mathbb{R}^{H \times W \times 3}$ 为输入，生成对应的点图 $\mathbf{O}^{0,0}, \mathbf{O}^{1,0} \in \mathbb{R}^{H \times W \times 3}$ ，以及各自的置信度图 $\mathbf{D}^{0,0}, \mathbf{D}^{1,0} \in \mathbb{R}^{H \times W}$ ，后者反映点图精度的置信程度。 $\mathbf{O}^{0,0}, \mathbf{O}^{1,0}$ 的下标表示它们都表达在锚定视角 $\mathbf{I}^0$ 的同一相机坐标系中。为恢复相机内参，这里假设主点位于中心且像素

为正方形，于是只剩焦距 f_0^* 未知，可用 Weiszfeld 算法 [61] 经过几步优化求解：

$$f_0^* = \arg \min_{f_0} \sum_{u=0}^W \sum_{v=0}^H \mathbf{D}_{u,v}^{0,0} \left\| (u', v') - f_0 \frac{(\mathbf{O}_{u,v,0}^{0,0}, \mathbf{O}_{u,v,1}^{0,0})}{\mathbf{O}_{u,v,2}^{0,0}} \right\|, \quad (2)$$

其中 $u' = u - \frac{W}{2}$, $v' = v - \frac{H}{2}$ 。当仅有单张输入图像时，我们把该图像复制一份构成成对输入，再估计其点图与相机内参。当输入图像多于两张时，它还能通过几步优化迭代实现全局点图对齐。

将点图与对应的 RGB 图像结合，即可得到彩色点云，为场景提供粗略的三维信息。然而，点云表示能力有限，加之稀疏或单张输入所提供的三维线索不足，重建出的点云会出现明显的缺失区域、遮挡和伪影，导致渲染结果质量低下。因此，我们提出引入视频扩散模型，在这种不完美的点云基础上实现高保真新视角合成。

3.3 用视频扩散模型渲染高保真新视角

如图 1 所示，以单张参考图像 $\mathbf{I}^{\text{ref}}$ 为例，我们先通过稠密立体模型 [19] 得到其点云、相机内参和相机位姿 $\mathbf{C}^{\text{ref}}$ 。随后，我们可以让相机沿一条包含 $\mathbf{C}^{\text{ref}}$ 的相机位姿序列 $\mathbf{C} = \{\mathbf{C}^0, \dots, \mathbf{C}^{L-1}\}$ 移动，渲染点云并得到一系列渲染结果，记作 $\mathbf{P} = \{\mathbf{P}^0, \dots, \mathbf{P}^{L-1}\}$ 。点云渲染虽能准确体现视角间的关系，却饱受大量遮挡、缺失区域和视觉保真度下降之扰。为提升新视角渲染质量，我们的目标是学习一个条件分布 $\mathbf{x} \sim p(\mathbf{x} | \mathbf{I}^{\text{ref}}, \mathbf{P})$ ，使其能够根据点云渲染 $\mathbf{P}$ 和参考图像 $\mathbf{I}^{\text{ref}}$ 生成高质量的新视角 $\mathbf{x} = \{\mathbf{x}^0, \dots, \mathbf{x}^{L-1}\}$ 。鉴于视频扩散模型 [16], [17], [18] 在合成高质量、一致视频上的成效，我们通过训练一个以点云渲染和参考图像为条件的视频扩散模型来学习该条件分布。于是，新视角合成过程可自然地建模为点云条件视频扩散模型的反向过程，表示为 $\mathbf{x} \sim p_\theta(\mathbf{x} | \mathbf{I}^{\text{ref}}, \mathbf{P})$ ，其中 θ 表示模型参数。

点云条件视频扩散模型的架构如图 1 所示。它沿用 LDM [60] 架构，主要包括：用于图像压缩的一对 VAE 编码器 $\mathcal{E}$ 与解码器 $\mathcal{D}$ ；一个视频去噪 U-Net，其空间层之后接时序层，用于时序感知的噪声估计；以及一个用于理解参考图像的 CLIP [62] 图像编码器。我们把点云渲染作为条件信号加入视频去噪 U-Net：先用 $\mathcal{E}$ 对其编码，再将得到的潜在图像与噪声沿通道维拼接。

为训练模型，我们构造成对训练数据，既包含点云渲染 $\mathbf{P} = \{\mathbf{P}^0, \dots, \mathbf{P}^{L-1}\}$ ，也包含对应的真值参考图像 $\mathbf{I} = \{\mathbf{I}^0, \dots, \mathbf{I}^{L-1}\}$ 。我们强制点云渲染至少经过一个真值视角，即在 L 帧中的某个随机位置至少包含一张真值参考图像。这有助于模型更好地学会把参考图像中的细节迁移到点云渲染上，并使模型能够灵活处理任意数量的参考图像。遵循

LDM [60] 的做法，我们冻结 VAE 编码器 $\mathcal{E}$ 与解码器 $\mathcal{D}$ 的参数，在潜在空间中开展训练。具体而言，我们将训练数据对 $\mathbf{I} = \{\mathbf{I}^0, \dots, \mathbf{I}^{L-1}\}$ 与 $\mathbf{P} = \{\mathbf{P}^0, \dots, \mathbf{P}^{L-1}\}$ 编码到潜在空间，得到真值潜在表示 $\mathbf{z} = \{\mathbf{z}^0, \dots, \mathbf{z}^{L-1}\}$ 以及将与采样噪声沿通道拼接的条件信号 $\hat{\mathbf{z}} = \{\hat{\mathbf{z}}^0, \dots, \hat{\mathbf{z}}^{L-1}\}$ 。随后，视频去噪 U-Net 由扩散损失优化：

$$\min_{\theta} = \mathbb{E}_{t \sim \mathcal{U}(0,1), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\epsilon_\theta(\mathbf{z}_t, t, \hat{\mathbf{z}}, \mathbf{I}^{\text{ref}}) - \epsilon\|_2^2], \quad (3)$$

其中 $\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \epsilon_0$ 。

在推理过程中，我们渲染出一列点云渲染 $\mathbf{P} = \{\mathbf{P}^0, \dots, \mathbf{P}^{L-1}\}$ ，并把参考视角的渲染结果替换为对应的参考图像。随后将它们编码到潜在空间，得到潜在图像 $\hat{\mathbf{z}} = \{\hat{\mathbf{z}}^0, \dots, \hat{\mathbf{z}}^{L-1}\}$ ，采样噪声 $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ，再沿通道拼接构成含噪潜在表示。除潜在空间条件外，我们还把参考图像送入 CLIP [62] 图像编码器，通过交叉注意力调制 U-Net 特征，以获得更好的三维理解。借助训练好的 U-Net，含噪潜在表示被迭代去噪为干净潜在表示，再由 VAE 解码器 $\mathcal{D}$ 解码为高质量的新视角 $\mathbf{x} = \{\mathbf{x}^0, \dots, \mathbf{x}^{L-1}\}$ 。

3.4 迭代式视角合成与相机轨迹规划

现有视频扩散模型在生成帧数众多的长视频时面临困难。推理时随着视频长度增加，视频稳定性下降、计算开销上升。因此，训练好的 ViewCrafter 模型在生成更长视频以覆盖更大视角范围时也会遇到困难。为应对这一挑战，我们采用迭代式视角合成方法。具体来说，给定由参考图像建立的初始点云，我们把相机从某个参考视角移动到一个目标相机位姿，以揭示当前点云的遮挡和缺失区域；随后用 ViewCrafter 生成高保真新视角，并将生成的视角反投影回去以补全点云。通过反复移动相机、生成新视角、更新点云，我们最终能得到视角范围更大的新视角，以及更完整的场景点云表示。

在迭代式视角合成过程中，相机轨迹的设计对合成结果影响显著。[14], [63] 等方法采用预定义的相机轨迹进行场景生成，忽视了不同场景中多样的几何关系，导致明显的遮挡。为在迭代式视角合成中有效揭示遮挡、促成更完整的场景生成，我们设计了一种基于下一最佳视角（Next-Best-View, NBV）[64], [65], [66] 的相机轨迹规划算法，能够自适应地生成契合各类场景的相机轨迹。该相机轨迹规划算法如算法 1 所示。从输入参考图像 $\mathcal{I}_{\text{ref}}$ 出发，我们用稠密立体模型 [19] 构建初始点云 $\mathcal{P}_{\text{ref}}$ 。参照 [64], [66], [67], [68]，我们选取一个相机位姿均匀分布、朝向前方的四分之一球面作为搜索空间，记作 $\mathcal{S}$ ，并把它中心放在点云世界坐标系的原点，半径设为参考图像中心像素的深度。相机轨迹从某个参考相机位姿 $\mathbf{C}_{\text{ref}}$ 初始化。为预测下一个位姿，我们在当前相机位姿 $\mathbf{C}_{\text{curr}} = \mathbf{C}_{\text{ref}}$ 周围的搜索空间中均匀采样 K 个候选相机位姿

算法 1 相机轨迹规划

输入： 参考图像 $\mathcal{I}_{\text{ref}}$, 稠密立体模型 $\mathcal{D}(\cdot)$, 点云条件视频扩散模型 $\mathcal{V}(\cdot)$, 初始点云 $\mathcal{P}_{\text{ref}}$, 搜索空间 $\mathcal{S}$, 初始位姿 $\mathcal{C}_{\text{ref}}$, 最大预测位姿数 N , 候选位姿数 K , 效用函数 $\mathcal{F}(\cdot)$

```

1: 初始化 当前点云  $\mathcal{P}_{\text{curr}} \leftarrow \mathcal{P}_{\text{ref}}$ , 当前相机位姿  $\mathcal{C}_{\text{curr}} \leftarrow \mathcal{C}_{\text{ref}}$ ,
    $\text{step} \leftarrow 0$ 
2: while  $\text{step} \leq N$  do
3:   从当前位姿  $\mathcal{C}_{\text{curr}}$  周围的搜索空间  $\mathcal{S}$  中均匀采样  $K$  个
     候选位姿  $\mathcal{C}_{\text{can}} = \{\mathcal{C}_{\text{can}}^1, \dots, \mathcal{C}_{\text{can}}^K\}$ , 初始化候选掩码集合
      $\mathcal{M}_{\text{can}} = \{\}$ 
4:   for  $\mathcal{C}$  in  $\{\mathcal{C}_{\text{can}}^1, \dots, \mathcal{C}_{\text{can}}^K\}$  do
5:      $\mathcal{M}_{\mathcal{C}} = \text{Render}(\mathcal{P}_{\text{curr}}, \mathcal{C})$ 
6:      $\mathcal{M}_{\text{can}}.\text{append}(\mathcal{M}_{\mathcal{C}})$ 
7:   end for
8:    $\mathcal{C}_{\text{nbv}} = \arg \max_{\mathcal{C} \in \mathcal{C}_{\text{can}}} \mathcal{F}(\mathcal{C})$ 
9:    $\mathcal{I}_{\text{nbv}} = \mathcal{V}(\text{interpolate}(\mathcal{C}_{\text{curr}}, \mathcal{C}_{\text{nbv}}), \mathcal{P}_{\text{curr}})$ 
10:   $\mathcal{P}_{\text{curr}} \leftarrow \mathcal{D}(\mathcal{I}_{\text{nbv}}, \mathcal{P}_{\text{curr}})$ 
11:   $\mathcal{C}_{\text{curr}} \leftarrow \mathcal{C}_{\text{nbv}}$ 
12:   $\text{step} \leftarrow \text{step} + 1$ 
13: end while
14: 返回

```

$\mathcal{C}_{\text{can}} = \{\mathcal{C}_{\text{can}}^1, \dots, \mathcal{C}_{\text{can}}^K\}$, 然后从当前点云 $\mathcal{P}_{\text{curr}}$ 渲染出一组候选掩码 $\mathcal{M}_{\text{can}}$ (其中 1 表示遮挡与缺失区域, 0 表示已填充区域)。接着, 我们建立一个效用函数 [64] $\mathcal{F}(\cdot)$ 来确定下一步的最优相机位姿, 定义为:

$$\mathcal{F}(\mathcal{C}) = \begin{cases} \frac{\text{sum}(\mathcal{M}_{\mathcal{C}})}{W \times H}, \frac{\text{sum}(\mathcal{M}_{\mathcal{C}})}{W \times H} < \Theta \\ 1 - \frac{\text{sum}(\mathcal{M}_{\mathcal{C}})}{W \times H}, \frac{\text{sum}(\mathcal{M}_{\mathcal{C}})}{W \times H} > \Theta, \end{cases} \quad (4)$$

其中 $\mathcal{C} \in \mathcal{C}_{\text{can}}$, $\mathcal{M}_{\mathcal{C}} \in \mathcal{M}_{\text{can}}$, 且 $\text{sum}(\mathcal{M}_{\mathcal{C}}) = \sum_{u=0}^W \sum_{v=0}^H \mathcal{M}_{\mathcal{C}}(u, v)$ 。该效用函数有助于挑选一个合适的相机位姿: 既能揭示足够面积的遮挡与缺失区域, 又能避免那些揭示出的空洞过大、明显偏离阈值 Θ 的位姿 (这类位姿可能影响 ViewCrafter 的生成能力)。一旦预测出下一最佳相机位姿 $\mathcal{C}_{\text{nbv}}$, 我们便在 $\mathcal{C}_{\text{curr}}$ 与 $\mathcal{C}_{\text{nbv}}$ 之间插值出一条相机路径, 并沿该路径运行 ViewCrafter 生成一系列高保真新视角。随后, 我们将生成的新视角 $\mathcal{I}_{\text{nbv}}$ 反投影并对齐到当前点云 $\mathcal{P}_{\text{curr}}$ 上, 把 $\mathcal{C}_{\text{nbv}}$ 指定为新的 $\mathcal{C}_{\text{curr}}$, 再重复上述过程, 直到预测的位姿数达到预设上限 N 。通过反复预测相机位姿、合成新视角、更新点云, 我们最终能得到更完整的场景点云表示。

3.5 应用

ViewCrafter 能够从单张或稀疏输入有效生成精确、一致、高保真的新视角。然而, 由于多步去噪过程较慢, 它在提供沉

浸式体验上仍有困难。为实现实时渲染, 我们进一步深入研究如何从 ViewCrafter 的结果优化出一个 3D-GS [2] 表示。为此, 一种直接的做法是在初始构建的点云上并发多次运行 ViewCrafter 生成多组新视角, 再据此优化 3D-GS。但由于初始点云并不完整, 不同次生成的视角序列在遮挡区域会引入不一致, 这种做法会导致次优的优化结果。

如图 2 所示, 为促成更一致的 3D-GS 优化, 我们利用前述迭代式视角合成策略, 用 ViewCrafter 迭代地补全初始点云并合成新视角。这样既提供了一致的新视角作为训练数据, 也为 3D-GS [2] 提供了强有力的几何初始化。训练时, 每个三维高斯的中心由补全后的稠密点云初始化, 其各项属性则在合成新视角的监督下优化。我们简化了 3D-GS 优化流程, 弃用了稠密化、分裂和不透明度重置等技巧 [69], 并将总优化时长压缩到 2000 次迭代, 比原始 3D-GS 训练快得多。

除对真实世界图像合成新视角外, 我们还探索将 ViewCrafter 与富有创意的文本到图像扩散模型结合, 用于文本到 3D 生成。其做法是: 先用文本到图像扩散模型根据给定的文本提示生成一张参考图像, 再用 ViewCrafter 进行新视角合成与三维重建。

4 实验

本节中, 我们先在第 4.1 节说明实现细节, 随后在第 4.2 节比较零样本新视角合成、在第 4.3 节比较场景重建。接着, 我们在第 4.4 节开展消融研究, 评估各项设计选择与训练策略。最后, 我们在第 4.5 节展示文本到 3D 生成的结果。

4.1 实现细节

我们采用渐进式训练策略。第一阶段, 我们在 320×512 分辨率下训练 ViewCrafter 模型, 帧长设为 25。整个视频去噪 U-Net 训练 50,000 次迭代, 学习率为 5×10^{-5} , 小批量大小为 16。第二阶段, 为适配高分辨率, 我们在 576×1024 分辨率下微调视频去噪 U-Net 的空间层 (即 2D 卷积层与空间注意力层), 迭代 5,000 次, 学习率为 1×10^{-5} , 有效小批量大小为 16。我们的模型在由 DL3DV [70] 和 RealEstate10K [7] 组成的混合数据集上训练。我们将视频数据切分为视频片段, 每段含 25 帧。为生成条件信号 (即点云渲染), 我们用 DUStr3R [19] 处理视频片段, 得到片段的相机轨迹以及每一视频帧全局对齐后的点云。然后, 我们随机选取若干视频帧所构建的点云, 用 Pytorch3D [71] 沿估计出的相机轨迹渲染。总计生成 632,152 对视频对作为训练数据。推理时, 我们采用带无分类器引导 [72] 的 DDIM 采样器 [30]。

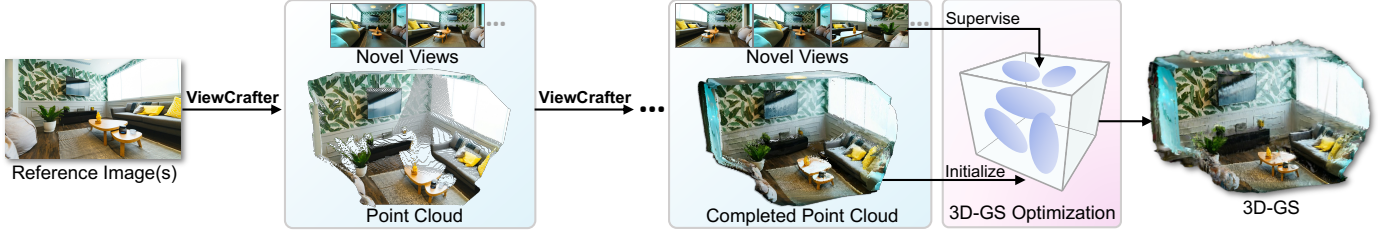


图 2. 3D-GS 优化的应用。为促成更一致的 3D-GS 优化，我们利用迭代式视角合成策略，用 ViewCrafter 逐步补全初始点云并合成新视角。随后，我们用补全后的稠密点云初始化 3D-GS，并用合成的新视角监督 3D-GS 的训练。

表 1

Tanks-and-Temples [21]、RealEstate10K [7] 和 CO3D [39] 数据集上零样本新视角合成的定量对比。由于 ZeroNVS [12] 和 LucidDreamer [14] 只能处理正方形图像，我们在计算定量指标时，把本方法和 MotionCtrl [13] 生成的新视角裁剪至与它们对齐。

数据集	简单集						困难集					
	LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓	$R_{\text{dist}} \downarrow$	$T_{\text{dist}} \downarrow$	LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓	$R_{\text{dist}} \downarrow$	$T_{\text{dist}} \downarrow$
Tanks-and-Temples												
LucidDreamer [14]	0.413	14.53	0.362	42.32	6.137	5.695	0.558	11.69	0.267	200.8	8.998	9.305
ZeroNVS [12]	0.482	14.71	0.380	74.60	8.810	6.348	0.569	12.05	0.309	131.0	8.860	8.557
MotionCtrl [13]	0.400	15.34	0.427	70.3	7.299	8.039	0.473	13.29	0.384	196.8	9.801	9.112
ViewCrafter (本文)	0.194	21.26	0.655	27.18	0.471	1.009	0.283	18.07	0.563	38.92	1.109	0.910
RealEstate10K												
LucidDreamer [14]	0.315	16.35	0.579	56.77	5.821	10.02	0.400	14.13	0.511	71.43	7.990	10.85
ZeroNVS [12]	0.364	16.50	0.577	96.18	6.370	9.817	0.431	14.24	0.535	105.8	8.562	10.31
MotionCtrl [13]	0.341	16.31	0.604	89.90	4.236	9.091	0.386	16.29	0.587	70.02	8.084	9.295
ViewCrafter (本文)	0.145	21.81	0.796	33.09	0.380	2.888	0.178	22.04	0.798	24.89	1.098	2.867
CO3D												
LucidDreamer [14]	0.429	15.11	0.451	78.87	12.90	6.665	0.517	12.69	0.374	157.8	16.43	8.301
ZeroNVS [12]	0.467	15.15	0.463	93.84	15.44	8.872	0.524	13.31	0.426	143.2	15.02	10.22
MotionCtrl [13]	0.393	16.87	0.529	69.18	16.87	5.131	0.443	15.46	0.502	112.7	18.81	5.575
ViewCrafter (本文)	0.243	21.38	0.687	24.63	2.175	1.033	0.324	18.96	0.641	36.96	2.849	1.480

4.2 零样本新视角合成对比

数据集与评价指标。在研究中，我们采用三个不同规模的真实世界数据集作为零样本新视角合成的评测基准，包括 CO3D [39]、RealEstate10K [7] 和 Tanks-and-Temples [21] 数据集。对于以物体为中心的 CO3D [39]，我们在 10 个场景上评测。RealEstate10K [7] 由室内场景视频片段构成，我们从其测试集中选取 10 个场景评测。对于包含大规模室外与室内场景的 Tanks-and-Temples [21]，我们使用全部 9 个场景。对所有基准，我们都从原始拍摄视频中抽取帧，并对原视频施加不同的采样率，构造两类测试集。简单 (easy) 测试集用较小的抽帧步长生成，相机运动缓慢、视角范围有限；困难 (hard) 测试集则用较大的采样步长生成，相机运动快速、视角范围大。

我们采用 PSNR、SSIM [73]、LPIPS [74] 和 FID [75] 作

为评估图像质量的指标。其中，PSNR 是用于比较图像相似度的传统指标；SSIM [73] 和 LPIPS [74] 则用于评估生成图像与真值图像之间的结构相似度和感知相似度，这两个指标专门设计得与人类感知判断更为吻合。参照 [3], [4], [5]，我们进一步把 FID 纳入评估流程来衡量合成视角的质量；在评估含有大量缺失与遮挡区域的困难测试集时，FID 尤为有效。此外，为评估生成新视角的位姿精度，我们估计生成新视角的相机位姿并与真值相机位姿比较。遵循 [58]，我们把估计位姿的相机坐标变换为相对于第一帧，并用最远帧对平移尺度归一化。随后，我们对每条生成的新视角序列，计算其与真值旋转矩阵之间的旋转距离 (R_{dist})，表示为：

$$R_{\text{dist}} = \sum_{i=1}^n \arccos\left(\frac{\text{tr}(\mathbf{R}_{\text{gen}}^i \mathbf{R}_{\text{gt}}^{iT}) - 1}{2}\right), \quad (5)$$

其中 $\mathbf{R}_{\text{gt}}^i$ 与 $\mathbf{R}_{\text{gen}}^i$ 分别表示真值旋转矩阵和生成旋转矩阵，我



图 3. Tanks-and-Temples [21]、RealEstate10K [7] 和 CO3D [39] 数据集上零样本新视角合成的定性对比。参考图像位于最左列，真值新视角位于最右列。

们将所有帧的距离求和作为最终结果。我们还计算平移距离 (T_{dist}), 表示为:

$$T_{\text{dist}} = \sum_{i=1}^n \|\mathbf{T}_{\text{gt}}^i - \mathbf{T}_{\text{gen}}^i\|_2. \quad (6)$$

值得注意的是, 由于 COLMAP [76] 对不一致的特征较为敏感, 容易无法从生成的新视角中提取位姿, 我们改用 DUST3R [19] 进行更鲁棒的位姿估计。

对比基线。作为一个基于扩散、可泛化的新视角合成框架, 我们将本方法与三个基于扩散的基线进行比较: ZeroNVS [12]、MotionCtrl [13] 和 LucidDreamer [14]。具体来说, ZeroNVS

[12] 从 Zero-1-to-3 [11] 微调而来, 能够以参考图像和相对相机位姿为条件生成新视角; 相机位姿被处理为 CLIP [62] 文本嵌入, 并通过交叉注意力注入扩散 U-Net。MotionCtrl [13] 是从 SVD [17] 微调而来的相机条件视频扩散模型, 能够根据条件参考图像和相对相机位姿序列生成一致的新视角; 这些位姿序列同样被处理为高层嵌入, 通过交叉注意力注入视频扩散 U-Net。LucidDreamer [14] 利用基于深度的图像变形来合成新视角, 并采用预训练的基于扩散的修补模型 [43] 来修补新视角中的缺失区域。在零样本新视角合成对比中, 由于这些基线只能进行单视角新视角合成, 我们对所有基线和本

方法都使用单张参考图像作为输入。

定性对比。定性结果如图 3 所示，其中参考图像位于最左列，真值新视角位于最右列。LucidDreamer [14] 的结果出现严重伪影，因为它用基于深度的图像变形来生成新视角，在处理相机内参未知的野外图像时尤其容易出问题，导致新视角不准确；此外，它用一个现成的修补模型 [43] 来修饰变形结果，往往在原始内容与修补内容之间引入不一致。ZeroNVS [12] 生成的新视角质量和精度也相对较低，主要原因是它通过文本嵌入把相机位姿条件引入扩散模型，无法对新视角的生成实现精确控制，从而得到次优结果。类似地，MotionCtrl [13] 虽能生成保真度更好的新视角，却难以生成与给定相机条件精确对齐的新视角，这是因为它同样采用高层相机嵌入来控制相机位姿，使新视角合成的精度下降。相比之下，本方法融合了显式点云先验与视频扩散模型，结果表明本方法在位姿控制精度和生成新视角的整体质量上都更胜一筹。

定量对比。定量对比结果见表 1。由于 ZeroNVS [12] 和 LucidDreamer [14] 只能处理正方形图像，我们在计算定量指标时，把本方法和 MotionCtrl [13] 生成的新视角裁剪至与 ZeroNVS、LucidDreamer 对齐。在图像质量方面，可以看到本方法在所有指标上都稳定优于基线：更高的 PSNR 和 SSIM 表明本方法保持了更好的图像质量以及与真值的相似度；更低的 LPIPS 进一步说明本方法生成的图像在感知上更准确；而 FID 大幅改善则表明本方法更有效地捕捉了数据的底层分布。在位姿精度方面，更小的 R_{dist} 和 T_{dist} 证明了本模型设计的有效性，使新视角合成中的位姿控制更加精确。

4.3 场景重建对比

数据集与评价指标。在场景重建对比中，我们使用 Tanks-and-Temples 数据集 [21] 中的 6 个场景进行评测。我们构建了一个具有挑战性的稀疏视角基准，每个场景仅含 2 张真值训练图像，并用 12 个视角进行评测。我们采用 PSNR、SSIM [73] 和 LPIPS [74] 作为图像质量评估指标。

对比基线。我们将本方法与三种基于 3D-GS 表示的稀疏视角重建方法进行比较：DNGaussian [77]、FSGS [78] 和 InstantSplat [79]。具体而言，DNGaussian [12] 和 FSGS [78] 采用 COLMAP [76] 生成的点云作为初始化，并同时利用图像监督和深度正则化进行稀疏视角重建。InstantSplat [79] 则探索采用 DUST3R [19] 生成的点云作为初始化，从而能够从稀疏图像高效训练 3D-GS。

定性对比。定性对比结果如图 4 所示。可以看到，DNGaussian [77] 的结果出现明显伪影。类似地，FSGS [78] 的结果在偏离真值训练图像的新视角下也出现伪影。InstantSplat [79] 虽用 DUST3R [19] 做点云初始化、能较好保留真值训练图像的细节，但由于省略了稠密化过程 [2]，无法恢复遮挡区域，导致

表 2

Tanks-and-Temples [21] 上场景重建的定量对比。我们对每个场景使用 2 张真值训练图像，并采用 12 个视角进行评测。

方法	LPIPS ↓	PSNR ↑	SSIM ↑
DNGaussian [77]	0.331	15.47	0.541
FSGS [78]	0.364	17.53	0.558
InstantSplat [79]	0.275	18.61	0.614
ViewCrafter (本文)	0.245	21.50	0.692

表 3

位姿条件策略的消融研究。除条件信号外，Plücker 模型的架构和训练策略与 ViewCrafter 完全一致。

方法	LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓	R_{dist} ↓	T_{dist} ↓
Plücker 模型	0.370	17.51	0.546	49.33	2.688	2.570
本文方法	0.270	20.25	0.649	38.17	0.552	0.983

新视角下出现严重空洞。相比之下，本方法借助视频扩散模型的先验，仅凭 2 张真值训练图像即可生成高保真新视角。

定量对比。定量对比结果见表 2。可以看到，本方法在所有指标上都稳定优于对比基线，进一步印证了本方法在稀疏视角场景重建中的有效性。

4.4 消融研究

位姿条件策略的讨论。在本方法中，我们利用点云渲染作为视频扩散模型的显式条件，从而在新视角合成中实现高度精确的位姿控制。一些同期工作 [57], [80] 采用普吕克坐标 [81] 作为位姿条件来进行可控位姿的视频生成。为比较我们基于点云的位姿条件策略与基于普吕克坐标的位姿条件策略在位姿精度上的差异，我们训练了一个以普吕克坐标为条件的视频扩散模型（简称 Plücker 模型），它接受普吕克坐标作为条件来合成新视角。普吕克坐标描述逐像素的运动：对给定的一帧 RGB 图像及其相机位姿，其普吕克坐标与该 RGB 帧具有相同的空间尺寸，并在每个像素位置包含 6 个通道。给定一条位姿序列，我们把对应的普吕克坐标缩放到与潜在空间尺寸对齐，并沿通道维与噪声拼接。除位姿条件策略外，我们让 Plücker 模型的其余架构与 ViewCrafter 完全一致，并按第 4.1 节所述的训练细节，在 320×512 分辨率下训练该模型。我们在 ViewCrafter (320×512 分辨率) 与 Plücker 模型之间进行零样本新视角合成对比。定性定量结果分别见图 6 和表 3，表明 ViewCrafter 采用的基于点云的位姿条件策略在新视角合成中实现了更精确的位姿控制。我们还观察到，Plücker 模型容易忽略相机的高频运动。图 7 给出一个例子，其中我们比较真值相机位姿与从生成新视角估计出的位姿之间的对齐程度。结果显示，从 ViewCrafter 生成的新视角估计



图 4. Tanks-and-Temples [21] 数据集上场景重建的定性对比。我们用每个场景的 2 张真值训练图像进行训练，并渲染新视角与真值新视角（最右列）比较。

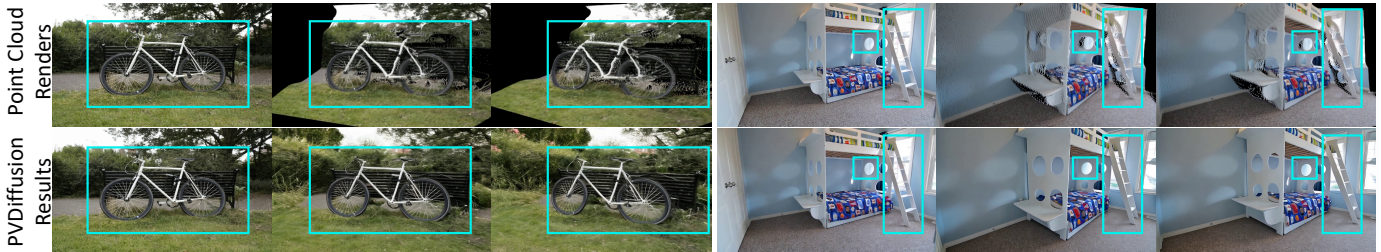


图 5. 点云条件的鲁棒性。上行展示点云渲染结果，下行展示 ViewCrafter 相应生成的新视角（放大观看效果最佳）。

表 4

不同训练范式的定量对比。我们分析同时训练视频去噪 U-Net 的空间层与时序层的有效性，以及渐进式训练策略和推理更多帧带来的收益。

训练范式	LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓
仅训练空间层	0.301	18.82	0.595	42.30
直接在 576×1024 上训练	0.314	18.55	0.582	41.01
16 帧模型	0.289	19.07	0.610	38.43
本文方法	0.280	19.52	0.615	37.77

出的位姿与真值位姿对齐得更紧密，进一步证明了我们基于点云的位姿条件策略的有效性。

点云条件的鲁棒性。 ViewCrafter 以点云渲染结果作为条件，从而实现高度精确的位姿控制。然而，这些渲染可能含有伪影和几何畸变。图 5 给出一个例子：第一行显示，作为条件

的点云渲染出现遮挡和缺失区域，以及前景边界处的几何畸变；第二行展示 ViewCrafter 相应生成的新视角，表明它能够填补空洞并纠正不准确的几何。这说明 ViewCrafter 已形成对三维世界的全面理解，使其能够从不完美的条件信息生成高质量新视角，并对点云条件表现出鲁棒性。

训练范式的消融。本消融研究考察所采用训练范式的有效性。为评估应训练视频去噪 U-Net 中哪个模块的选择，我们比较仅训练空间层与同时训练空间层和时序层（本文方法）的新视角合成质量，结果见表 4 第一行。为评估渐进式训练的重要性，我们比较直接在 576×1024 分辨率下训练模型与采用渐进式训练策略训练模型（本文方法），结果见表 4 第二行。此外，我们观察到 ViewCrafter 的推理长度会影响新视角合成质量：在相同的视角变化范围内，推理更多帧能提升生成帧的时序一致性。为在计算开销与合成质量之间取得平衡，我们训练了两个模型：推理 16 帧的基础模型与推理 25 帧的更



图 6. 位姿条件策略的定性消融。除条件信号外，我们让 Plücker 模型的架构和训练策略与 ViewCrafter 完全一致。

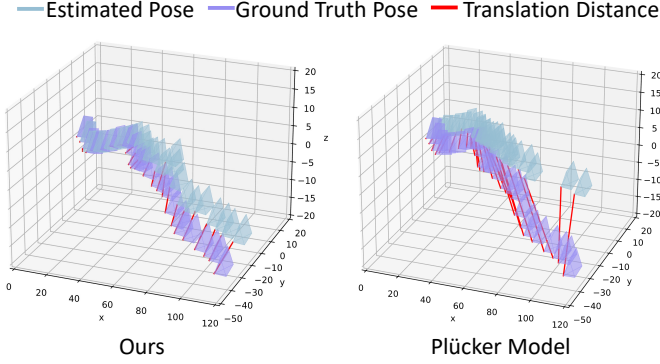


图 7. 位姿精度可视化。我们比较真值相机位姿与从 ViewCrafter、Plücker 模型生成的新视角所估计位姿之间的对齐程度。

强模型（本文方法）。两个模型的比较见表 4 第三行。上述结果展示了所实现训练范式的有效性。

相机轨迹规划的消融。本消融研究评估我们提出的相机轨迹规划算法在揭示遮挡和补全点云方面的有效性。图 8 给出一个例子，其中我们比较两种做法生成的点云：一种用预定义的圆形相机轨迹（与 [14] 类似）进行迭代式视角合成，另一种用我们的相机轨迹规划算法。给定一张参考图像和一个初始点云，我们采用一个以初始点云世界坐标系原点为中心的四分之一球面搜索空间，半径设为参考图像中心像素的深度。我们先探索搜索空间的左半区，相机轨迹规划参数设为 $N = 3$ 、 $K = 5$ 、 $\Theta = 0.6$ 。相应地，圆形相机轨迹设置为：从参考位姿出发，向搜索空间的左方向均匀移动相机三次，每次移动 20 度。随后，我们用相同参数探索搜索空间的右半区。最终生成的点云如图 8(a) 和图 8(b) 所示。在图 8(a) 中可以看到，用预定义圆形相机轨迹重建出的点云未能有效补全遮挡区域；相比之下，图 8(b) 展示了用我们的相机轨迹规划算法进行的重建。更完整的重建结果表明，该算法能够有效揭示场景的遮挡区域，提升整体场景重建质量。

4.5 文本到 3D 生成

除对真实世界图像合成新视角外，我们还探索将本框架与富有创意的文本到图像（T2I）扩散模型结合，用于文本到 3D 生成。其做法是：给定一段文本提示，我们先用 T2I 模型生

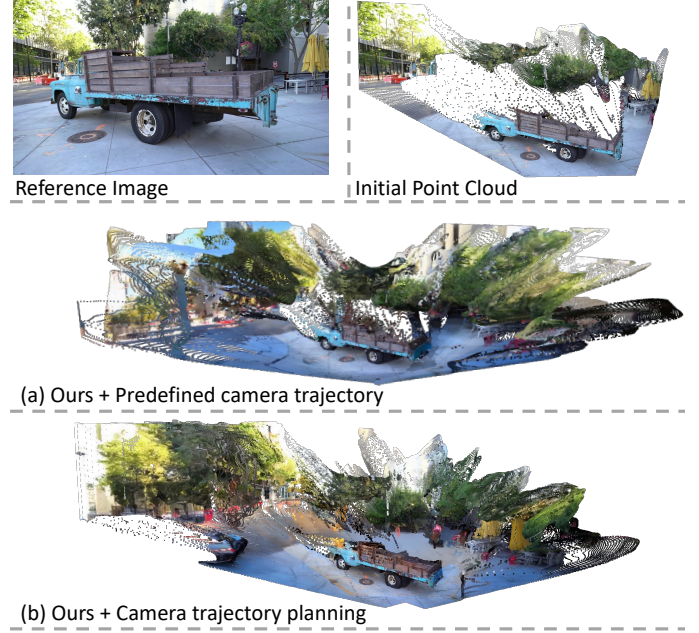


图 8. 相机轨迹规划的消融研究。用预定义相机轨迹重建的点云无法有效补全遮挡区域；相比之下，通过我们的相机轨迹规划算法生成的点云能更有效地揭示场景的遮挡区域，提升点云的整体重建质量。

成对应图像，再用 ViewCrafter 合成一致的新视角。结果如图 9 所示。

5 结论与局限

本文提出 ViewCrafter，一个结合视频扩散模型与点云先验、用于高保真且精确的新视角合成框架。本方法克服了现有方法的局限：既能泛化到各类场景，又能适配单张和稀疏图像两种输入，同时在新视角质量上保持一致性与精确性。此外，我们提出一种迭代式视角合成方法和一套自适应相机轨迹规划流程，支撑长距离新视角合成，并为各类场景自动生成相机轨迹。除新视角合成外，我们还探索了高效优化 3D-GS 表示以实现实时、高帧率的新视角渲染，以及将本框架适配到文本到 3D 生成。

局限。尽管有诸多优势，本方法仍存在若干局限。其一，在三维线索有限的情况下合成视角范围极大的新视角时，本方法

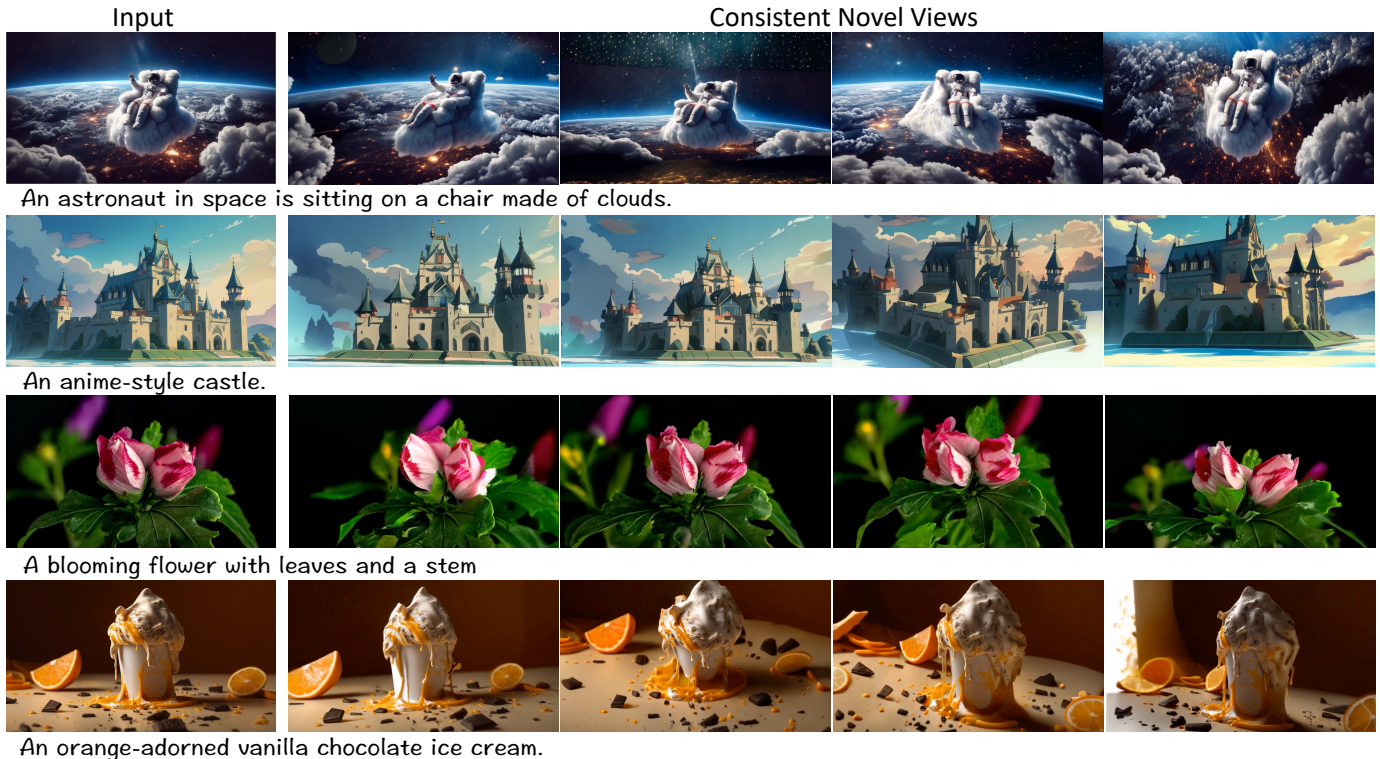


图 9. 文本到 3D 生成结果。最左列展示输入文本提示以及 T2I 模型生成的对应图像，其余各列展示 ViewCrafter 基于该生成图像所产生的一致新视角。

可能遇到困难，例如仅凭一张后视图生成前视图。其二，我们以点云作为显式先验，并验证了本方法对低质量点云的鲁棒性；但在条件点云严重不准确的场景中，困难可能依然存在。其三，作为一个视频扩散模型，本方法在推理过程中需要多步去噪，计算开销相对较高。

参考文献

- [1] 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,” in *ECCV*, 2020.
- [2] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3d gaussian splatting for real-time radiance field rendering,” *ACM TOG*, 2023.
- [3] O. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson, “Synsin: End-to-end view synthesis from a single image,” in *CVPR*, 2020.
- [4] R. Rombach, P. Esser, and B. Ommer, “Geometry-free view synthesis: Transformers and no 3d priors,” in *ICCV*, 2021.
- [5] C. Rockwell, D. F. Fouhey, and J. Johnson, “Pixelsynth: Generating a 3d-consistent experience from a single image,” in *ICCV*, 2021.
- [6] B. Park, H. Go, and C. Kim, “Bridging implicit and explicit geometric transformation for single-image view synthesis,” *IEEE TPAMI*, 2024.
- [7] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, “Stereo magnification: Learning view synthesis using multiplane images,” *ACM TOG*, 2018.
- [8] Y. Han, R. Wang, and J. Yang, “Single-view view synthesis in the wild with learned adaptive multiplane images,” in *SIGGRAPH Conference*, 2022.
- [9] R. Tucker and N. Snavely, “Single-view view synthesis with multiplane images,” in *CVPR*, 2020.
- [10] 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*, 2021, pp. 4578–4587.
- [11] 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 *ICCV*, 2023.
- [12] K. Sargent, Z. Li, T. Shah, C. Herrmann, H.-X. Yu, Y. Zhang, E. R. Chan, D. Lagun, L. Fei-Fei, D. Sun, and J. Wu, “ZeroNVS: Zero-shot 360-degree view synthesis from a single real image,” in *CVPR*, 2024.
- [13] Z. Wang, Z. Yuan, X. Wang, T. Chen, M. Xia, P. Luo, and Y. Shan, “Motionctrl: A unified and flexible motion controller for video generation,” in *SIGGRAPH Conference*, 2024.
- [14] J. Chung, S. Lee, H. Nam, J. Lee, and K. M. Lee, “Luciddreamer: Domain-free generation of 3d gaussian splatting scenes,” *arXiv preprint arXiv:2311.13384*, 2023.
- [15] J. Shriram, A. Trevithick, L. Liu, and R. Ramamoorthi, “Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion,” *arXiv preprint arXiv:2404.07199*, 2024.
- [16] P.-Y. Lab and T. A. etc., “Open-sora-plan,” <https://github.com/PKU-YuanGroup/Open-Sora-Plan>, 2024. [Online]. Available: <https://doi.org/10.5281/zenodo.10948109>

- [17] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, “Stable video diffusion: Scaling latent video diffusion models to large datasets,” *arXiv preprint arXiv:2311.15127*, 2023.
- [18] J. Xing, M. Xia, Y. Zhang, H. Chen, X. Wang, T.-T. Wong, and Y. Shan, “Dynamicrafter: Animating open-domain images with video diffusion priors,” *arXiv preprint arXiv:2310.12190*, 2023.
- [19] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, “Dust3r: Geometric 3d vision made easy,” in *CVPR*, 2024.
- [20] V. Leroy, Y. Cabon, and J. Revaud, “Grounding image matching in 3d with mast3r,” *arXiv:2406.09756*, 2024.
- [21] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, “Tanks and temples: Benchmarking large-scale scene reconstruction,” *ACM TOG*, 2017.
- [22] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordon, P. Labatut, and D. Novotny, “Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction,” in *ICCV*, 2021.
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017.
- [24] A. Trevisan, M. Chan, M. Stengel, E. Chan, C. Liu, Z. Yu, S. Khamis, M. Chandraker, R. Ramamoorthi, and K. Nagano, “Real-time radiance fields for single-image portrait view synthesis,” *ACM TOG*, 2023.
- [25] W. Yu, Y. Fan, Y. Zhang, X. Wang, F. Yin, Y. Bai, Y.-P. Cao, Y. Shan, Y. Wu, Z. Sun *et al.*, “Nofa: Nerf-based one-shot facial avatar reconstruction,” in *SIGGRAPH Conference*, 2023.
- [26] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan, “Lrm: Large reconstruction model for single image to 3d,” in *ICLR*, 2024.
- [27] D. Charatan, S. L. Li, A. Tagliasacchi, and V. Sitzmann, “pixel-splat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction,” in *CVPR*, 2024.
- [28] Y. Chen, H. Xu, C. Zheng, B. Zhuang, M. Pollefeys, A. Geiger, T.-J. Cham, and J. Cai, “Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images,” in *ECCV*, 2024.
- [29] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *NeurIPS*, 2020.
- [30] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in *ICLR*, 2021.
- [31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *CVPR*, 2022.
- [32] W. Yu, L. Yuan, Y.-P. Cao, X. Gao, X. Li, L. Quan, Y. Shan, and Y. Tian, “Hifi-123: Towards high-fidelity one image to 3d content generation,” in *ECCV*, 2024.
- [33] J. Tang, T. Wang, B. Zhang, T. Zhang, R. Yi, L. Ma, and D. Chen, “Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior,” in *ICCV*, 2023.
- [34] J. Sun, B. Zhang, R. Shao, L. Wang, W. Liu, Z. Xie, and Y. Liu, “Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior,” in *ICLR*, 2024.
- [35] E. R. Chan, K. Nagano, M. A. Chan, A. W. Bergman, J. J. Park, A. Levy, M. Aittala, S. De Mello, T. Karras, and G. Wetzstein, “Generative novel view synthesis with 3d-aware diffusion models,” in *ICCV*, 2023.
- [36] 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 *CVPR*, 2023.
- [37] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, “Shapenet: An information-rich 3d model repository,” *arXiv preprint arXiv:1512.03012*, 2015.
- [38] D. Watson, W. Chan, R. M. Brualdi, J. Ho, A. Tagliasacchi, and M. Norouzi, “Novel view synthesis with diffusion models,” in *ICLR*, 2023.
- [39] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordon, P. Labatut, and D. Novotny, “Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction,” in *ICCV*, 2021.
- [40] X. Yu, M. Xu, Y. Zhang, H. Liu, C. Ye, Y. Wu, Z. Yan, C. Zhu, Z. Xiong, T. Liang *et al.*, “Mvimgnet: A large-scale dataset of multi-view images,” in *CVPR*, 2023.
- [41] R. Wu, B. Mildenhall, P. Henzler, K. Park, R. Gao, D. Watson, P. P. Srinivasan, D. Verbin, J. T. Barron, B. Poole *et al.*, “Reconfusion: 3d reconstruction with diffusion priors,” in *CVPR*, 2024.
- [42] J. Zhang, X. Li, Z. Wan, C. Wang, and J. Liao, “Text2nerf: Text-driven 3d scene generation with neural radiance fields,” *IEEE TVCG*, 2024.
- [43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *CVPR*, 2022.
- [44] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *ICCV*, 2023.
- [45] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, “T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models,” in *AAAI*, 2024.
- [46] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, “Gligen: Open-set grounded text-to-image generation,” in *CVPR*, 2023.
- [47] J. Xing, H. Liu, M. Xia, Y. Zhang, X. Wang, Y. Shan, and T.-T. Wong, “Toonrafter: Generative cartoon interpolation,” *arXiv preprint arXiv:2405.17933*, 2024.
- [48] J. Xing, M. Xia, Y. Liu, Y. Zhang, Y. Zhang, Y. He, H. Liu, H. Chen, X. Cun, X. Wang *et al.*, “Make-your-video: Customized video generation using textual and structural guidance,” *IEEE TVCG*, 2024.
- [49] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, “Structure and content-guided video synthesis with diffusion models,” in *ICCV*, 2023.
- [50] S. Yin, C. Wu, J. Liang, J. Shi, H. Li, G. Ming, and N. Duan, “Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory,” *arXiv preprint arXiv:2308.08089*, 2023.
- [51] M. Niu, X. Cun, X. Wang, Y. Zhang, Y. Shan, and Y. Zheng, “Mofa-video: Controllable image animation via generative motion field adaptations in frozen image-to-video diffusion model,” *arXiv preprint arXiv:2405.20222*, 2024.
- [52] E. Peruzzo, V. Goel, D. Xu, X. Xu, Y. Jiang, Z. Wang, H. Shi, and N. Sebe, “Vase: Object-centric appearance and shape manipulation of real videos,” *arXiv preprint arXiv:2401.02473*, 2024.
- [53] Y. Guo, C. Yang, A. Rao, Y. Wang, Y. Qiao, D. Lin,

- and B. Dai, “Animatediff: Animate your personalized text-to-image diffusion models without specific tuning,” *arXiv preprint arXiv:2307.04725*, 2023.
- [54] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” in *ICLR*, 2022.
- [55] N. Müller, K. Schwarz, B. Rössl, L. Porzi, S. R. Bulò, M. Nießner, and P. Kotschieder, “Multidiff: Consistent novel view synthesis from a single image,” in *CVPR*, 2024.
- [56] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “ScanNet: Richly-annotated 3d reconstructions of indoor scenes,” in *CVPR*, 2017.
- [57] D. Xu, W. Nie, C. Liu, S. Liu, J. Kautz, Z. Wang, and A. Vahdat, “Camco: Camera-controllable 3d-consistent image-to-video generation,” *arXiv preprint arXiv:2406.02509*, 2024.
- [58] H. He, Y. Xu, Y. Guo, G. Wetzstein, B. Dai, H. Li, and C. Yang, “Cameractrl: Enabling camera control for text-to-video generation,” *arXiv preprint arXiv:2404.02101*, 2024.
- [59] V. Sitzmann, S. Rezchikov, B. Freeman, J. Tenenbaum, and F. Durand, “Light field networks: Neural scene representations with single-evaluation rendering,” in *NeurIPS*, 2021.
- [60] G. Metzger, E. Richardson, O. Patashnik, R. Giryes, and D. Cohen-Or, “Latent-nerf for shape-guided generation of 3d shapes and textures,” in *CVPR*, 2023.
- [61] F. Plastria, “The weiszfeld algorithm: proof, amendments and extensions, ha eiselt and v. marianov (eds.) foundations of location analysis, international series in operations research and management science,” 2011.
- [62] 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 *ICML*, 2021.
- [63] R. Gao, A. Holynski, P. Henzler, A. Brussee, R. Martin-Brualla, P. Srinivasan, J. T. Barron, and B. Poole, “Cat3d: Create anything in 3d with multi-view diffusion models,” *arXiv preprint arXiv:2405.10314*, 2024.
- [64] R. Zeng, Y. Wen, W. Zhao, and Y.-J. Liu, “View planning in robot active vision: A survey of systems, algorithms, and applications,” *Computational Visual Media*, vol. 6, 2020.
- [65] H. Dhami, V. D. Sharma, and P. Tokekar, “Pred-nbv: Prediction-guided next-best-view planning for 3d object reconstruction,” in *IROS*, 2023.
- [66] L. Jin, X. Chen, J. Rückin, and M. Popović, “Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering,” in *IROS*, 2023.
- [67] S. Isler, R. Sabzevari, J. Delmerico, and D. Scaramuzza, “An information gain formulation for active volumetric 3d reconstruction,” in *ICRA*, 2016.
- [68] D. Peralta, J. Casimiro, A. M. Nilles, J. A. Aguilar, R. Atienza, and R. Cajote, “Next-best view policy for 3d reconstruction,” in *ECCVW*, 2020.
- [69] M.-L. Shih, S.-Y. Su, J. Kopf, and J.-B. Huang, “3d photography using context-aware layered depth inpainting,” in *CVPR*, 2020.
- [70] L. Ling, Y. Sheng, Z. Tu, W. Zhao, C. Xin, K. Wan, L. Yu, Q. Guo, Z. Yu, Y. Lu *et al.*, “DL3dv-10k: A large-scale scene dataset for deep learning-based 3d vision,” in *CVPR*, 2024.
- [71] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari, “Accelerating 3d deep learning with pytorch3d,” *arXiv:2007.08501*, 2020.
- [72] J. Ho and T. Salimans, “Classifier-free diffusion guidance,” *arXiv preprint arXiv:2207.12598*, 2022.
- [73] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.
- [74] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *CVPR*, 2018.
- [75] 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,” in *NeurIPS*, 2017.
- [76] J. L. Schönberger and J.-M. Frahm, “Structure-from-motion revisited,” in *CVPR*, 2016.
- [77] J. Li, J. Zhang, X. Bai, J. Zheng, X. Ning, J. Zhou, and L. Gu, “Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization,” in *CVPR*, 2024.
- [78] Z. Zhu, Z. Fan, Y. Jiang, and Z. Wang, “Fsgs: Real-time few-shot view synthesis using gaussian splatting,” in *ECCV*, 2024.
- [79] Z. Fan, W. Cong, K. Wen, K. Wang, J. Zhang, X. Ding, D. Xu, B. Ivanovic, M. Pavone, G. Pavlakos *et al.*, “Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds,” *arXiv:2403.20309*, 2024.
- [80] S. Bahmani, I. Skorokhodov, A. Siarohin, W. Menapace, G. Qian, M. Vasilkovsky, H.-Y. Lee, C. Wang, J. Zou, A. Tagliasacchi *et al.*, “Vd3d: Taming large video diffusion transformers for 3d camera control,” *arXiv preprint arXiv:2407.12781*, 2024.
- [81] Y.-B. Jia, “Plücker coordinates for lines in the space,” *Problem Solver Techniques for Applied Computer Science, Com-S-477/577 Course Handout*, 2020.