

使用 DDIM 反演的新型视图合成

Sehajdeep Singh¹, A V Subramanyam¹

¹Indraprastha Institute of Information Technology, Delhi
{ sehajs, subramanyam } @iiitd.ac.in

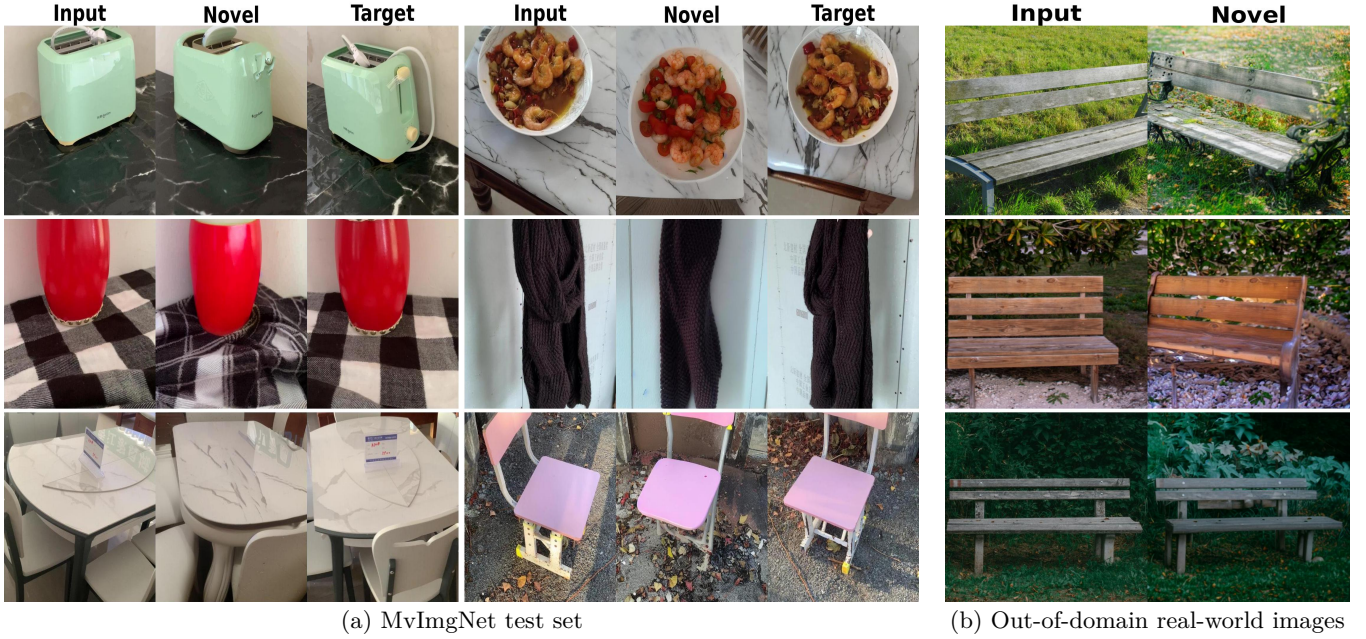


Figure 1: (a) High-resolution (512×512) novel-view synthesis on the MvImgNet test set from a single input image and camera parameters, (b) Zero-shot synthesis on out-of-domain images downloaded from Unsplash.

Abstract

从单张输入图像合成新的视图是一项具有挑战性的任务。它需要推测场景的三维结构，同时推断被遮挡区域的细节，并在不同视点之间保持几何一致性。许多现有方法必须使用多个视图微调大型扩散模型或从头开始训练一个扩散模型，这代价极高。此外，它们还面临着模糊重建和泛化能力差的问题。这个差距提供了探索一种显式轻量级视图转换框架的机会，该框架可以直接利用预训练扩散模型的高保真生成能力，同时从新视图重建场景。给定单个输入图像的 DDIM 反转潜在表示，我们采用一个摄像机姿态条件化翻译 U-Net，即 TUNet，来预测对应于目标视图的反转潜在表示。然而，使用预测的潜在表示采样的图像可能导致模糊的重建。为此，我们提出了一种新的融合策略，利用了 DDIM 反演中观察到的内在噪声相关结构。这种融合策略有助于保留纹理和细粒度细节。为了合成新视图，我们使用融合后的潜在表示作为 DDIM 采样的初始条件，利用预训练扩散模型的生成先验。在 MvImgNet 上的大量实验表明，我们的方法优于现有方法。代码已开放。

引言

新颖视图合成是计算机视觉和图形学中的一个基本任务。像 NeRFs (Mildenhall et al. 2021) 和 3DGS (Kerbl et al. 2023) 这样的杰出作品在 3D 场景理解中被广泛应用。有几项工作在这些基础作品上进行了改进。然而，它们对场景级优化的依赖以及对密集视图集合的需求限制了可用性。扩散模型 (Rombach et al. 2022; Podell et al. 2023) 已经在新颖视图合成任务 (Elata et al. 2025; Tang et al. 2025) 上获得了显著的关注。一个经典的方法是将这些模型在 3D 数据集上进行微调，同时加入一个模块将 3D 几何编码到架构中 (Tang et al. 2023; Liu et al. 2024; Long et al. 2024; Huang et al. 2024; Bourigault and Bourigault 2024; Gao et al. 2024)。然而，由于生成过程无法完全控制，生成的输出在多视图重建中缺乏一致性，导致图像质量不足，并且常常在长距离视点重建中产生模糊结果。

DDIM (Song, Meng, and Ermon 2020) 提出了一个确定性逆运算“DDIM 逆运算”，该逆运算通过顺序地

向图像添加噪声以获得一个有噪声的潜在变量。使用 DDIM 采样可以将这个有噪声的潜在变量追溯到原始图像。这个潜在变量封装了对均值和方差有贡献的信号和噪声，这改变了每次逆运算时噪声潜在变量的分布。先前的工作如 (Garibi et al. 2024; Mokady et al. 2023) 尝试优化或配置这个噪声表示，以更好地与给定任务对齐。(Staniszewski, Kuciński, and Deja 2025) 详细研究了逆运算噪声，并声称 DDIM 逆运算潜在空间不易操控，这使得直接与这个噪声潜在变量进行插值对于新视角合成和编辑等任务显得困难。

本文提出了一种从给定输入图像和相机参数生成新视图的方法。我们的流程完全在 DDIM 反转的潜在空间中进行。我们首先学习使用称为 TUNet 的翻译 U-Net 将输入视图潜在映射到目标潜在。此映射仅近似目标视图的粗粒度版本。这是因为扩散模型表现出光谱偏差，偏向低频分量 (Choi et al. 2022)。为了引入高频分量，我们引入了一种新颖的噪声潜在融合策略。值得注意的是，我们使用预训练的扩散模型，只训练一个轻量级的潜在空间翻译网络 TUNet 来进行视图转换。我们在不同设置下进行了广泛的实验，并表明我们的工作可以扩展到未见过的类别以及从网络获得的域外图像。图 1 展示了一些示例结果。我们声称以下是关键贡献：

- 我们提出了一种将输入的 DDIM 反演潜变量转换为目标潜变量的方法。可以通过预训练扩散模型的 VAE 解码器解码目标潜变量，以获得目标的新视图。
- 翻译后的潜在编码可能仅仅导致一个粗糙的图像，该图像保留了目标图像的大致结构。为了注入高频细节，我们提出了一种新颖的融合策略。TUNet 的粗略输出与通过我们的融合策略得到的高方差噪声相融合。融合后的潜在编码可以用于初始化 DDIM 采样，从而重建一个具有一致几何和生动细致细节的高质量新视图。
- 在我们的实验中，我们展示了该方法在 LPIPS、PSNR、SSIM 和 FID 方面取得了优异的结果。

相关工作

神经辐射场：神经场方法，例如神经辐射场 (NeRF) (Mildenhall et al. 2021)，使用可学习的函数将三维空间坐标和观察方向映射到体积密度和颜色。这些模型通过在学习到的场景表示中进行光线行进实现体积渲染来合成新视图。NeRF 已经展示出，当基于密集输入视图集进行训练时，可以渲染出高质量的新视图。

虽然最近的扩展，如 PixelNeRF (Yu et al. 2021)、IBRNet (Wang et al. 2021)、MultiDiff (Müller et al. 2024) 及其他 (Henzler et al. 2021; Liu et al. 2022; Wu et al. 2023)，旨在从较少的输入视角进行视图合成，但它们通常在缺失或遮挡内容的区域表现不佳。因为这些模型在没有显式不确定性建模的情况下进行确定性预测，生成的输出往往在有歧义的地方进行平均，从而导致在未观察到的区域出现模糊和较不合理的重建。

Gaussian Splatting：三维高斯散射 (3DGS) (Kerbl et al. 2023; Peng et al. 2024; Zheng et al. 2025) 使用一组各向异性的三维高斯来表示场景。高斯散射方法是确定性的，并且在很大程度上依赖于准确的多视图几何或密集采样的相机位置 (Li et al. 2024)。在稀疏视图或单视图情况下应用时，它们常常无法在未见区域生成合理的内容，因为它们缺乏生成性先验。

相比之下，我们的工作针对的是仅给定单个输入图像和一个目标相机姿态的新视图合成。这种设定涵盖了近距离和远距离的视点变化。在这种情况下，即使在如 (Szymanowicz et al. 2024; Tewari et al. 2023) 中那样加入生成性指导，方法如 NeRF (Mildenhall et al. 2021) 和 3DGS (Kerbl et al. 2023) 也难以从单个图像有效地进行外推。

Transformers：NViST (Jang and Agapito 2024) 采用了一种基于 Transformer 的编码器-解码器架构 (Vaswani et al. 2017; Dosovitskiy et al. 2021)，从单张图像预测辐射场，从而通过类似 NeRF 的体渲染实现新的视角合成。然而，由于激进的下采样（缩小了 12 倍），NViST 遭遇了细节损失，并且在合成远距离视角时遇到了困难（当目标帧与输入帧相距在 30 帧的视窗中超过 15 帧时）。

Diffusion models：扩散模型可以用于生成输入视图中未观察区域的合理内容。接下来，我们将其识别为微调预训练扩散模型，或从头开始训练扩散模型的那些方法。

Pretrained Diffusion Models：MVDiffusion (Tang et al. 2023)、Zero123++ (Shi et al. 2023)、SyncDreamer (Liu et al. 2024)、Wonder3D (Long et al. 2024)、EpiDiff (Huang et al. 2024)、BoostDream (Yu et al. 2024)、MVDiff (Bourigault and Bourigault 2024)、CAT3D (Gao et al. 2024)、Magic-Boost (Yang et al. 2024)、Cycle3D (Tang et al. 2025)，都使用了预训练或微调的扩散模型。MVDiffusion 通过引入一种称为对应注意力 (CAA) 的交叉分支注意力机制，修改了稳定扩散架构，以建模视图间依赖关系。SyncDreamer 从所有目标噪声视图中构建一个视槽特征体积，并使用深度注意力层将其注入到预训练的去噪 U-net 中。EpiDiff (Huang et al. 2024) 在 U-Net 的中间和解码阶段中集成了一个由极线约束指导的注意力模块，使模型能够跨视图捕捉广义极线几何。

大多数现有的方法要么对扩散模型进行微调，要么将与目标视角对应的空间特征注入到基本模型的去噪 U-Net 中。这些特征通常是从体积投影或深度估计中得出的。然而，遵循这种范式的模型往往在场景级别的重建上表现欠佳，并且通常是在特定物体的数据集上训练的，这可能限制它们对复杂场景的泛化能力。

相比之下，我们的方法不会修改或注入任何学习到的特征到扩散模型的 U-Net 中。相反，我们向 TUNet 提供外部条件输入，以获得与目标视图相对应的潜变量。

扩散模型的潜在文本空间也已被用于探索视图合成 Burgess, Wang, and Yeung-Levy (2024)。然而，它通常局限于小规模数据集，不能直接应用于场景级的新视图合成任务。

从头训练扩散模型：一些最新的工作从头训练扩散模型用于新视图合成，包括 Tseng et al. (2023)、Photometric-NVS (Yu et al. 2023a)、DiffDreamer (Cai et al. 2023)、GIBR (Anciukevičius et al. 2024) 和 (Henderson et al. 2024)。Photometric-NVS (Yu et al. 2023a) 引入了一种双流潜在在扩散架构，该架构独立处理源视图和噪声目标视图，同时通过姿态条件的交叉注意机制交换信息。GIBR (Anciukevičius et al. 2024) 使用 IB-planes 对 3D 场景进行建模，并在像素空间直接训练扩散过程，从而能够学习多视图观察和相机姿态的联合分布。

端到端训练整个扩散模型在计算上是昂贵的，并且

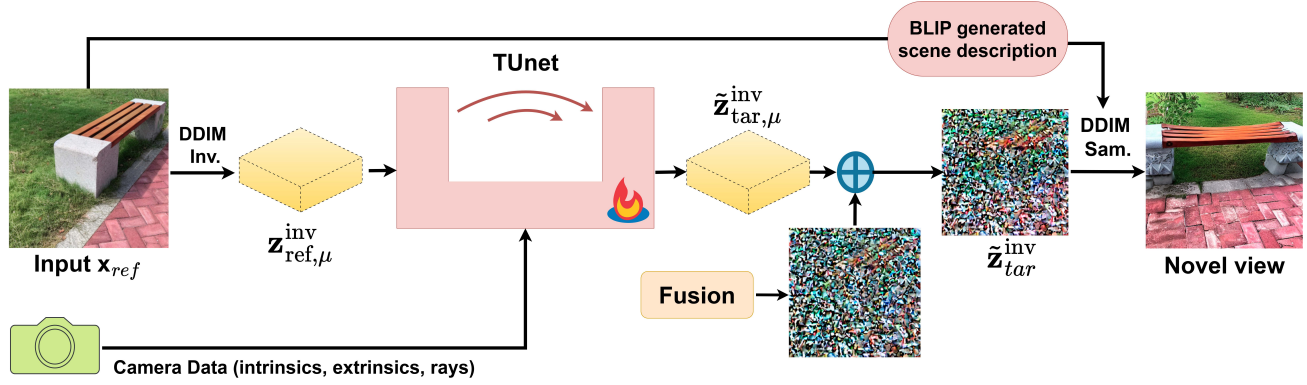


Figure 2: 概述：给定一个参考图像 $\mathbf{x}_{\text{ref}}$ ，我们首先应用 DDIM 反演到 $t = 600$ 以获得均值潜在变量 $\mathbf{z}_{\text{ref}, \mu}^{\text{inv}}$ 。这一过程，连同相机内参/外参、类别嵌入以及光线信息，一同输入到我们的翻译网络 TUNet。TUNet 预测目标视角的均值潜在变量 $\tilde{\mathbf{z}}_{\text{tar}, \mu}^{\text{inv}}$ ，我们通过一种融合策略将其与相应的噪声组件结合，形成初始的 DDIM 潜在变量 $\tilde{\mathbf{z}}_{\text{tar}}^{\text{inv}}$ 。最后，这个潜在变量通过一个预训练的扩散模型进行采样，以合成新的视角图像。

需要大规模的数据集来实现高分辨率和照片真实感的重建。相比之下，我们的方法在固定的时间步操作于 DDIM 反演的潜在空间，这对应于一个弱但信息丰富的信号。这使我们能够使用轻量级的 U-Net 进行从给定潜在到目标潜在的翻译。在潜在空间中操作显著简化了视图翻译任务，因为模型处理的是紧凑的、语义丰富的表示，而不是原始像素。我们的融合策略提供了关于高频场景细节的必要信息。最终的新视图是使用预训练的扩散管道合成的，该管道解码预测的潜在。

给定一个参考图像和目标视角的相机参数，我们的工作旨在解决新视图合成任务。受 DDIM 反演的确定性行为启发，我们完全在 DDIM 反演的潜空间中进行视图合成。训练一个专用的翻译网络 TUNet，以将源潜映射到与新视角对应的目标潜。为了引入高频场景细节，我们提出了一种融合策略。然后将生成的潜通过一个预训练的扩散模型，以生成最终高保真度的新视图。我们的方法在图 2 中进行了说明。

扩散的谱行为

在 (Kingma and Gao 2023; Falck et al. 2025) 中，作者研究了扩散的谱行为。前向扩散 (Ho, Jain, and Abbeel 2020) 过程由以下公式给出：

其中 $\mathbf{x}_0$ 是干净的潜在， $\mathbf{x}_t$ 是与时间步长 t 对应的潜变量， $\bar{\alpha}_t$ 是缩放因子。在正向扩散过程中，高频分量（代表精细细节）比低频分量更快且更早退化。这一特性在扩散的逆过程也一致。如 Choi et al. (2022) 所示，扩散模型固有地偏向于低频分量，这意味着必须更加重视对高频细节的建模。此外，噪声分量通常被观察到偏离标准多变量高斯分布 (Staniszewski, Kuciński, and Deja 2025)。在反演的后期迭代中，噪声涵盖图像的高频信息且具有较高的方差。信号的方差随着反演时间的增加而减小，而预测噪声的方差随着反演时间的增加而增大。有效的 DDIM 反演 (Song, Meng, and Ermon 2020) 迭代为：

$$\mathbf{x}_{t+1} = \underbrace{\left(\mathbf{x}_t - \sqrt{1 - \alpha_t} \epsilon_{\theta}(\mathbf{x}_t, t) \right) \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}}_{\text{signal / mean, } \mathbf{z}_{\mu, t+1}^{\text{inv}}} + \underbrace{\sqrt{1 - \alpha_{t+1}} \epsilon_{\theta}(\mathbf{x}_t, t)}_{\text{noise / variance, } \mathbf{z}_{\sigma, t+1}^{\text{inv}}} \quad (1)$$

其中 $\mathbf{x}_{t+1}$ 是在时间步 $t+1$ 的噪声潜变量。 $\mathbf{x}_t$ 是时间步 t 的潜变量，其噪声较少于 $\mathbf{x}_{t+1}$ 。 α_t 和 α_{t+1} 是控制时间步 t 和 $t+1$ 的噪声调度的参数，源自噪声方差调度。 $\epsilon_{\theta}(\mathbf{x}_t, t)$ 是在时间步 t 预测的噪声，通过逆过程中的扩散 U-Net 估计。

我们并不是逆推到 $t = T$ ，因为此时潜在变量类似于白噪声，反向轨迹变得不稳定，而是我们在中间的时间步 $t^* < T$ 停止。在 t^* ，DDIM 潜在变量仍然保留了足够的低频结构，以支持通过我们的 TUNet 进行直接视图转换。

方程 (1) 中的信号/均值是我们进行视图转换的粗略图像表示。此外，方程 (1) 中的噪声/方差编码了在去噪过程中 (Staniszewski, Kuciński, and Deja 2025) 恢复的特定于图像的特征。对于新视图合成任务，这个噪声/方差可以用来在经过视图转换的潜在变量中引入高频细节，然后可以将其输入到 DDIM 采样中。基于上述讨论，我们为新视图合成任务形式化了两件事：

- 扩散模型的频谱偏差可以被利用，在低频空间中通过我们的翻译网络 TUNet 进行视图变换。
- 为了补偿高频细节，我们利用 DDIM 逆变中方程 (1) 的噪声/方差项来制定融合策略。

DDIM-反转潜变量

设 $\mathbf{z}_t^{\text{inv}}$ 是 DDIM 反转的潜在变量。如果我们在 $t = T$ 使用这个潜在变量，我们可能会看到 DDIM 采样的图像偏离输入图像，特别是在较少的 DDIM 步骤中执行时 (Bao et al. 2025; Zeng, Suganuma, and Okatani 2025; Feng et al. 2024)。因此，我们固定 $t = 600$ ，并在 30 个 DDIM 步骤中得到我们的 DDIM 反转噪声初始潜在变量。此外，方程 (1) 中的信号/均值项是扩散网络对在 t 通过根据扩散得分模型 ϵ_{θ} 去噪 $\mathbf{z}_t^{\text{inv}}$ 所获得的干净潜在变量的估计。这个信号/均值项是我们输入 TUNet 进

行视图转换的内容。我们在图 3 中可视化了与信号/均值项对应的重建图像。我们观察到重建图像主要由输入图像的低频成分组成。因此，为了强加高频信息，我们利用方程 (1) 中的噪声/方差项 $z_{\sigma,t+1}^{\text{inv}}$ ，在 t 层级重新注入预测的噪声。我们使用一个预训练的潜在扩散模型 (LDM) (Rombach et al. 2022) 作为我们的生成先验，来计算 DDIM 反转的潜在变量。我们首先描述 TUNet 模型，其次是融合策略。由于我们将时间步长 t 固定为 600，在表示信号/均值和噪声/方差项时，我们省略下标 t ：从现在起，方程 (1) 中的 $z_{\mu,t+1}^{\text{inv}}$ 和 $z_{\sigma,t+1}^{\text{inv}}$ 。



Figure 3: 左图：DDIM 反转潜在变量在 $t=600$ 时的平均值。使用 VAE 解码潜在变量以进行可视化。右图：原始 512×512 图像。

TUNET 架构

TUNet 是一个受 U-Net (Ronneberger, Fischer, and Brox 2015) 启发的编码器-解码器架构，设计用于预测目标视图的 DDIM 反转潜变量的平均表示。TUNet 在输入或参考视图与目标视图之间引入了交叉注意力，能够在视点之间进行有效的特征转移。该架构在多个阶段基于摄像机参数和类别嵌入进行调整，以保持几何一致性和语义完整性。

输入与条件设置 首先通过 VAE 编码器将输入图像 x_{ref} 映射到潜在空间，得到 z_{ref} 。然后我们在这个潜在空间表示 z_{ref} 上执行 DDIM 反演，以获得作为 TUNet 输入的均值项 $z_{\text{ref},\mu}^{\text{inv}}$ 。以下信息在不同阶段用作 TUNet 的条件：

- 摄像机嵌入 $C = (K, R, t)$ ：相机的内参 K 和外参 (R, t) 的矢量化形式通过一个可学习的线性层来生成嵌入向量 $e_C \in \mathbb{R}^{d_C}$ 。
- 类别嵌入：一个可学习的类别嵌入，对应于场景类别，映射到 $e_c \in \mathbb{R}^{d_c}$ 。

这些嵌入与时间嵌入 $\gamma(t) \in \mathbb{R}^{d_t}$ 进行拼接，并将合并向量 $[\gamma(t) \oplus e_C \oplus e_c] \in \mathbb{R}^{d_t+d_C+d_c}$ 传递通过一个可学习的线性投影，以使其与时间嵌入空间对齐。投影后的向量在空间上进行广播，并在每个降采样、中间和升采样块增加到特征图 f 上：

$$f' = f + \text{Proj}_{\text{combined}}[\gamma(t) \oplus e_C \oplus e_c],$$

，其中 $\text{Proj}_{\text{combined}}$ 是一个学习的线性层，将拼接嵌入映射到 $\mathbb{R}^{d_t}$ 。这使得能够联合考虑时间、摄像机视角和场景类别。

编码器（下行模块） 编码器由一系列残差下采样块组成，这些块在降低空间分辨率的同时扩大特征深度。每个块的条件基于摄像机和输入或参考视图 ($C_{\text{ref}}, c_{\text{ref}}$) 的类别嵌入。这些嵌入在连接和投影后被添加：

$$f^{(i)} = \text{Down}_i(f^{(i-1)} + \text{Proj}_{\text{combined}}[\gamma(t) \oplus e_{C_{\text{ref}}} \oplus e_{c_{\text{ref}}}]),$$

其中 i 表示 TUNet 中的块深度。

瓶颈和解码器（中间 + 向上模块） 瓶颈模块既依赖于输入或参考视图和目标视图的相机嵌入 ($C_{\text{ref}}, C_{\text{tar}}$)，也依赖于类别嵌入，这使得模型能够在潜在层面上捕捉视点转换。上采样阶段仅依赖于目标视图的相机和类别嵌入 ($C_{\text{tar}}, c_{\text{tar}}$)，指导表示向所需的目标视图靠拢：

$$f^{\text{mid}} = \text{Mid}(f^{\text{enc}} + \text{Proj}_{\text{combined}}[\gamma(t) \oplus e_{C_{\text{ref}}} \oplus e_{C_{\text{tar}}} \oplus e_{c_{\text{tar}}}]),$$

$$f^{(i)} = \text{Up}_i(f^{(i-1)} + \text{Proj}_{\text{combined}}[\gamma(t) \oplus e_{C_{\text{tar}}} \oplus e_{c_{\text{tar}}}]).$$

交叉注意力模块 在中间块和上升块中集成了交叉注意力机制，使得信息可以通过光线信息和潜特征对齐从参考视图流向目标视图。设 r_{ref} 表示参考视图的光线嵌入， r_{tar} 表示目标视图的光线嵌入。我们使用 NeRF 中的标准光线参数化 (Mildenhall et al. 2021) 来计算相机姿态编码的光线起点和方向，以获得 r_{ref} 和 r_{tar} 。设 $z_{\text{ref},\mu}^{\text{inv}}$ 为参考图像的 DDIM 反向潜均值， f_{tar} 为交叉注意力块中的中间目标特征图。

注意力机制使用以下公式：

$$Q = W_Q[r_{\text{tar}} \| f_{\text{tar}}], K = W_K[r_{\text{ref}} \| z_{\text{ref},\mu}^{\text{inv}}], V = W_V z_{\text{ref},\mu}^{\text{inv}}$$

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \quad (2)$$

然后，将注意力的输出加回到目标特征中：

$$f'_{\text{tar}} = f_{\text{tar}} + \text{Attn}(Q, K, V).$$

TUNet 的输出是一个潜在的 $\tilde{z}_{\text{tar},\mu}^{\text{inv}}$ ，代表合成视图的 DDIM 反演均值项，对应于目标相机。使用 $\tilde{z}_{\text{tar},\mu}^{\text{inv}}$ ，我们接下来解释融合策略。

融合策略

为了从预测的 DDIM 反演均值潜在变量 $\tilde{z}_{\text{tar},\mu}^{\text{inv}}$ 合成语义丰富的目标视图潜在变量，我们引入了两种融合策略，这些策略将此均值潜在变量与从输入视图潜在变量派生的噪声成分结合。这些策略重新注入学习的噪声方差，即高频细节，进入粗潜在变量。我们利用了这样一个事实，即输入视图的 DDIM 反演潜在变量中的噪声/方差项方程 (1) 包含场景级别的属性和特征 (Staniszewski, Kuciński, and Deja 2025)，当与 TUNet 的预测融合时，可以用于从新的视角合成场景。

策略 A：通过 σ 组件进行方差融合 在这个策略中，我们从输入视图的 DDIM 反演潜变量中明确提取方差（或噪声）成分，记作 $z_{\text{ref},\sigma}^{\text{inv}}$ 。我们在 z_{ref} 上执行 DDIM 反演，并从方程 (1) 中提取等效的噪声/方差项 $z_{\text{ref},\sigma}^{\text{inv}}$ 。最终的潜变量计算如下：

$$z_{\text{noisy}} = \tilde{z}_{\text{tar},\mu}^{\text{inv}} + z_{\text{ref},\sigma}^{\text{inv}}. \quad (3)$$

融合后的潜变量 z_{noisy} 然后被传递到 Stable Diffusion U-Net 中以计算噪声预测：

$$\epsilon_{\theta} = \text{U-Net}(z_{\text{noisy}}, t). \quad (4)$$

DDIM 采样的初始潜变量如下获得：

$$\tilde{z}_{\text{tar}}^{\text{inv}} = \tilde{z}_{\text{tar},\mu}^{\text{inv}} + \sqrt{1 + \alpha_{t+1}} \epsilon_{\theta}. \quad (5)$$

策略 B: 从参考反演直接添加噪声 在此，我们直接使用输入视图 $z_{\text{ref}}^{\text{inv}}$ 的完整 DDIM 反演潜在向量中的噪声组件，而不是单独提取其方差。用于 DDIM 采样的初始潜在向量 $\tilde{z}_{\text{tar}}^{\text{inv}}$ 计算如下：

$$\tilde{z}_{\text{tar}}^{\text{inv}} = \tilde{z}_{\text{tar},\mu}^{\text{inv}} + \sqrt{1 + \alpha_{t+1}} z_{\text{ref}}^{\text{inv}}. \quad (6)$$

我们使用从方程 (5) 和方程 (6) 中生成的 $\tilde{z}_{\text{tar}}^{\text{inv}}$ 样本，并选择最佳生成结果，其中我们选择 LPIPS 分数较低的结果。

训练目标

我们的训练目标是对齐预测值和真实值的 DDIM 反演潜在均值。我们通过最小化预测目标潜在均值 $\tilde{z}_{\text{tar},\mu}^{\text{inv}}$ 与目标视图的真实 DDIM 反演潜在均值 $z_{\text{tar},\mu}^{\text{inv}}$ 之间的均方误差 (MSE) 损失来实现这一点：

$$\mathcal{L}_{\text{MSE}} = \|\tilde{z}_{\text{tar},\mu}^{\text{inv}} - z_{\text{tar},\mu}^{\text{inv}}\|_2^2. \quad (7)$$

实验

数据集：MvImgNet (Yu et al. 2023b) 包含 238 个类别的 650 万帧现实场景。我们使用 MvImgNet 的两个子集：(i) 三个场景类别：沙发、椅子和桌子。训练、验证和测试分别使用 90-5-5 (Anciukevicius et al. 2024) 的分割，由场景标识符的词典排序决定。(ii) 我们使用 167 个类别中的 85 万帧，对于每个类别，我们在测试集中保留 99 个场景中的 1 个，以评估我们的结果并与其他方法进行比较。

预处理：我们将图像的较短边缩放到 512，并调整其他维度以保持纵横比，然后对中心区域进行裁剪为 512×512 。这些 512×512 RGB 图像随后通过 VAE 编码器和 DDIM 反转管道，以获得反转后的潜变量 z^{inv} ，并提取它们的均值和方差组件。我们从 $t = 0$ 到 $t = 600$ 以 30 步进行 DDIM 反转。在反转后的潜在空间 z^{inv} 中，数据的维度是 $4 \times 64 \times 64$ 。

实现细节：TUnet 约有 148M 的参数。类别和相机嵌入的维度均为 64，交叉注意力维度为 768，注意力头维度为 64。我们使用潜在扩散骨干 (Rombach et al. 2022)。训练时，我们随机将每个场景的第 1-10 帧与第 15-25 帧配对。我们有效地使用每个场景的 20 帧进行训练。我们的评估采用相同的帧配对策略。我们在我们的子集上训练了两个模型 (i) 3 类和 (ii) 167 类。我们的 167 类模型在单台 40GB A400 上以批量大小为 32 和学习率为 $1e-5$ 的条件下训练了 450 个 epoch，总计 17 个 GPU 天，同时使用循环调度器衰减学习率。在推理期间，我们通过 30 次 DDIM 采样步骤生成最终结果，初始潜在变量为方程式 (5) 或方程式 (6)，表示在 $t = 600$ 时的噪声潜在变量。我们将我们的 3 类模型与 GIBR (Anciukevicius et al. 2024) 进行比较，将 167 类模型与 NViST (Jang and Agapito 2024) 进行比较。对于 GIBR，我们拥有相同的训练/测试划分，并直接报

Method	LPIPS ↓	PSNR ↑	SSIM ↑
GIBR	0.510	17.61	0.554
Ours	0.490	15.71	0.523

Table 1: 对 3 个类别的比较——椅子、沙发、桌子。分辨率为 256×256 。(注意：由于代码不可用，GIBR 的准确设置无法复现。)

Method	LPIPS ↓	PSNR ↑	SSIM ↑	FID ↓
NViST	0.448	14.31	0.566	91.63
Ours	0.409	16.16	0.578	65.50

Table 2: 167 个类别的比较。分辨率是 90×90

告其论文中的结果。对于 NViST，我们使用其预训练模型测试精确输入/目标帧对。所有测试帧均来自训练中使用的类别内未见过的场景。为了与 GIBR 进行比较，我们将 512×512 结果调整大小至 256×256 。为了与 NViST 直接比较，我们将结果调整大小至 90×90 。我们报告调整大小结果的 LPIPS 和 FID 分数以进行比较。我们报告原始 512×512 分辨率结果的 PSNR 和 SSIM。

定量比较

3 类模型：表 1 展示了在分辨率为 256×256 的 3 类上与 GIBR 的比较。测试中使用了来自 168 个未见过场景的输入和目标配对。我们在 LPIPS 方面优于 GIBR。GIBR 在 RGB 空间中训练整个扩散过程，并在训练和体积渲染时使用多个视图生成最终图像。因此，GIBR 在 PSNR 和 SSIM 方面表现更好。然而，在像素空间中训练扩散是非常昂贵的。另一方面，我们只在使用潜在变量的情况下训练了 148M 参数的 TUnet。

167 类模型：与分辨率为 90×90 的 NViST 的比较见表 2。在测试中使用了来自 360 个未见场景的输入和目标对。在这里，我们看到我们的方法在 LPIPS、PSNR、SSIM 和 FID 方面表现更好。

我们在图 4 中展示了合成结果。在水壶的例子中，我们可以看到未观察到的区域被高保真合成。同样，在碗的例子中（第三行，中间列），阴影也被忠实地合成。

定性比较

我们在图 5 中将我们的结果与 NViST (Jang and Agapito 2024) 进行了比较。显然，我们的方法可以高保真地合成目标，并且能够生成接近和远距离目标视图的结果，而 NViST 无法做到这一点。

未见类别：我们在图 6 中展示了 6 个未见类别的定性结果。我们涵盖了室外和室内场景，并在测试集中包含了大物体和小物体类别。即使在未见类别中，我们也能够为多样化的场景生成高分辨率的重建。

域外数据：为了评估超越 MvImgNet 的零样本泛化能力，我们通过下载来自 Unsplash 的自然场景自由许可照片，组建了一个域外测试集。我们展示了结果并与 Zero123++ (Shi et al. 2023) 进行了对比，见图 7。虽然这两种方法都成功生成了合理的新视角，但我们的方法产生了更真实的表面纹理，并保持了自然场景特征。



Figure 4: 我们的 167 类训练模型的定性结果。

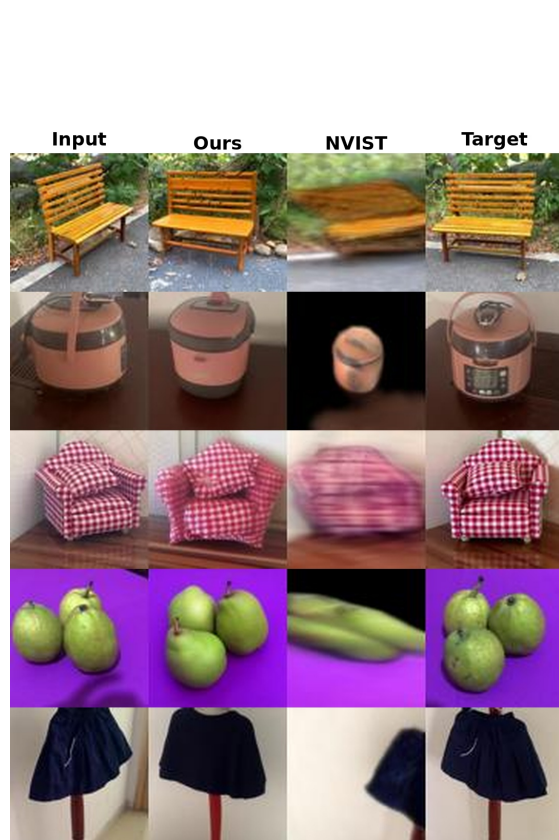


Figure 5: 我们将我们的结果调整为 90×90 ，以展示与 NViST 在来自 5 个类别的未见测试场景上的比较。



Figure 6: MVImgNet 中 6 个未见类别的结果

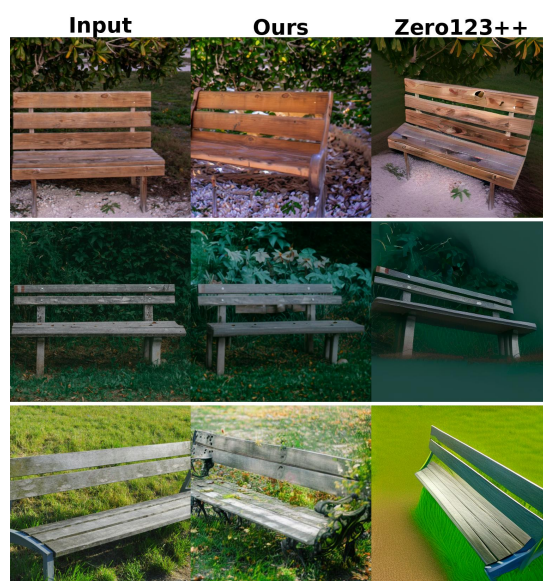


Figure 7: 领域外图像

消融

烧蚀结果在表 ?? 中展示。我们评估了以下设置。在第一个设置 拼接 中,我们将类别和相机嵌入与输入连接,但没有将它们注入每个 ResNet 块。在这里,我们看到性能有显著下降。第二个设置 无交叉注意力 中,我们移除了所有交叉注意力层。所有三个指标的结果都有所下降。

在这项工作中,我们提出了一种新方法,使用 TUNet 和融合策略来合成高质量的新视角。我们的方法使用单张输入图像和相机参数来合成新视角。相较于之前需要训练一个庞大扩散模型的工作,我们的方法训练了一个轻量级的翻译网络以在潜在空间中获取视角转换。为了丰富预测的潜在变量的高频场景细节,我们提出了一种新颖的融合策略。我们的实验在各种设置下显示出强大的性能。

References

- Anciukevičius, T.; Manhardt, F.; Tombari, F.; and Henderson, P. 2024. Denoising diffusion via image-based rendering. In ICLR.
- Bao, Y.; Liu, H.; Gao, X.; Fu, H.; and Kang, G. 2025. FreeInv: Free Lunch for Improving DDIM Inversion. arXiv preprint arXiv:2503.23035.
- Bourigault, E.; and Bourigault, P. 2024. Mvdif: Scalable and flexible multi-view diffusion for 3d object reconstruction from single-view. In CVPR, 7579–7586.
- Burgess, J.; Wang, K.-C.; and Yeung-Levy, S. 2024. Viewpoint textual inversion: Discovering scene representations and 3d view control in 2d diffusion models. In ECCV, 416–435.
- Cai, S.; Chan, E. R.; Peng, S.; Shahbazi, M.; Obukhov, A.; Van Gool, L.; and Wetzstein, G. 2023. Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models. In ICCV, 2139–2150.
- Choi, J.; Lee, J.; Shin, C.; Kim, S.; Kim, H.; and Yoon, S. 2022. Perception prioritized training of diffusion models. In CVPR, 11472–11481.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR.
- Elata, N.; Kawar, B.; Ostrovsky-Berman, Y.; Farber, M.; and Sokolovsky, R. 2025. Novel view synthesis with pixel-space diffusion models. In CVPR, 26756–26766.
- Falck, F.; Pandeva, T.; Zahirnia, K.; Lawrence, R.; Turner, R.; Meeds, E.; Zazo, J.; and Karmalkar, S. 2025. A Fourier Space Perspective on Diffusion Models. arXiv preprint arXiv:2505.11278.
- Feng, Y.; Gao, S.; Bao, Y.; Wang, X.; Han, S.; Zhang, J.; Zhang, B.; and Yao, A. 2024. Wave: Warping ddim inversion features for zero-shot text-to-video editing. In ECCV, 38–55. Springer.
- Gao, R.; Holynski, A.; Henzler, P.; Brussee, A.; Martin-Brualla, R.; Srinivasan, P.; Barron, J. T.; and Poole, B. 2024. Cat3d: Create anything in 3d with multi-view diffusion models. NeurIPS.
- Garibi, D.; Patashnik, O.; Voynov, A.; Averbuch-Elor, H.; and Cohen-Or, D. 2024. Renoise: Real image inversion through iterative noising. In ECCV, 395–413.
- Henderson, P.; de Almeida, M.; Ivanova, D.; et al. 2024. Sampling 3d gaussian scenes in seconds with latent diffusion models. arXiv preprint arXiv:2406.13099.
- Henzler, P.; Reizenstein, J.; Labatut, P.; Shapovalov, R.; Ritschel, T.; Vedaldi, A.; and Novotny, D. 2021. Unsupervised learning of 3d object categories from videos in the wild. In CVPR, 4700–4709.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33: 6840–6851.
- Huang, Z.; Wen, H.; Dong, J.; Wang, Y.; Li, Y.; Chen, X.; Cao, Y.-P.; Liang, D.; Qiao, Y.; Dai, B.; et al. 2024. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In CVPR, 9784–9794.
- Jang, W.; and Agapito, L. 2024. Nvst: In the wild new view synthesis from a single image with transformers. In CVPR, 10181–10193.
- Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Trans. Graph., 42(4): 139–1.
- Kingma, D.; and Gao, R. 2023. Understanding diffusion objectives as the elbo with simple data augmentation. NeurIPS, 36.
- Li, J.; Zhang, J.; Bai, X.; Zheng, J.; Ning, X.; Zhou, J.; and Gu, L. 2024. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In CVPR, 20775–20785.
- Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2024. Syncdreamer: Generating multiview-consistent images from a single-view image. ICLR.
- Liu, Y.; Peng, S.; Liu, L.; Wang, Q.; Wang, P.; Theobalt, C.; Zhou, X.; and Wang, W. 2022. Neural rays for occlusion-aware image-based rendering. In CVPR, 7824–7833.
- Long, X.; Guo, Y.-C.; Lin, C.; Liu, Y.; Dou, Z.; Liu, L.; Ma, Y.; Zhang, S.-H.; Habermann, M.; Theobalt, C.; et al. 2024. Wonder3d: Single image to 3d using cross-domain diffusion. In CVPR, 9970–9980.
- Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1): 99–106.
- Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In CVPR, 6038–6047.
- Müller, N.; Schwarz, K.; Rössl, B.; Porzi, L.; Bulo, S. R.; Nießner, M.; and Kotschieder, P. 2024. Multidiff: Consistent novel view synthesis from a single image. In CVPR, 10258–10268.

- Peng, R.; Xu, W.; Tang, L.; Jiao, J.; Wang, R.; et al. 2024. Structure consistent gaussian splatting with matching prior for few-shot novel view synthesis. *NeurIPS*, 37: 97328–97352.
- Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*, 10684–10695.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241.
- Shi, R.; Chen, H.; Zhang, Z.; Liu, M.; Xu, C.; Wei, X.; Chen, L.; Zeng, C.; and Su, H. 2023. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*.
- Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *ICLR*.
- Staniszewski, Ł.; Kuciński, Ł.; and Deja, K. 2025. There and Back Again: On the relation between Noise and Image Inversions in Diffusion Models. *ICLR*.
- Szymanowicz, S.; Insafutdinov, E.; Zheng, C.; Campbell, D.; Henriques, J. F.; Rupprecht, C.; and Vedaldi, A. 2024. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. *arXiv preprint arXiv:2406.04343*.
- Tang, S.; Zhang, F.; Chen, J.; Wang, P.; and Furukawa, Y. 2023. MVDiffusion: Enabling Holistic Multi-view Image Generation with Correspondence-Aware Diffusion. In *NeurIPS*.
- Tang, Z.; Zhang, J.; Cheng, X.; Yu, W.; Feng, C.; Pang, Y.; Lin, B.; and Yuan, L. 2025. Cycle3d: High-quality and consistent image-to-3d generation via generation-reconstruction cycle. In *AAAI*, volume 39, 7320–7328.
- Tewari, A.; Yin, T.; Cazenavette, G.; Rezhikov, S.; Tenenbaum, J.; Durand, F.; Freeman, B.; and Sitzmann, V. 2023. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *NeurIPS*, 36: 12349–12362.
- Tseng, H.-Y.; Li, Q.; Kim, C.; Alsisan, S.; Huang, J.-B.; and Kopf, J. 2023. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, 16773–16783.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.
- Wang, Q.; Wang, Z.; Genova, K.; Srinivasan, P. P.; Zhou, H.; Barron, J. T.; Martin-Brualla, R.; Snavely, N.; and Funkhouser, T. 2021. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 4690–4699.
- Wu, C.-Y.; Johnson, J.; Malik, J.; Feichtenhofer, C.; and Gkioxari, G. 2023. Multiview compressive coding for 3D reconstruction. In *CVPR*, 9065–9075.
- Yang, F.; Zhang, J.; Shi, Y.; Chen, B.; Zhang, C.; Zhang, H.; Yang, X.; Li, X.; Feng, J.; and Lin, G. 2024. Magic-Boost: Boost 3D Generation with Multi-View Conditioned Diffusion. *arXiv preprint arXiv:2404.06429*.
- Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 4578–4587.
- Yu, J. J.; Forghani, F.; Derpanis, K. G.; and Brubaker, M. A. 2023a. Long-term photometric consistent novel view synthesis with diffusion models. In *ICCV*, 7094–7104.
- Yu, X.; Xu, M.; Zhang, Y.; Liu, H.; Ye, C.; Wu, Y.; Yan, Z.; Zhu, C.; Xiong, Z.; Liang, T.; et al. 2023b. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, 9150–9161.
- Yu, Y.; Zhu, S.; Qin, H.; and Li, H. 2024. BoostDream: Efficient Refining for High-Quality Text-to-3D Generation from Multi-View Diffusion. In *IJCAI*, 5407–5415.
- Zeng, Y.; Suganuma, M.; and Okatani, T. 2025. Inverting the generation process of denoising diffusion implicit models: Empirical evaluation and a novel method. In *WACV*, 4516–4524. *IEEE*.
- Zheng, Y.; Jiang, Z.; He, S.; Sun, Y.; Dong, J.; Zhang, H.; and Du, Y. 2025. NexusGS: Sparse View Synthesis with Epipolar Depth Priors in 3D Gaussian Splatting. In *CVPR*, 26800–26809.

由于我们使用 DDIM 反演噪声，允许我们在最终采样过程中通过文本提示来扩展编辑我们的管道。对于我们的新视图重建，我们使用 BLIP 为输入图像 x_{ref} 生成标题。我们编辑这些提示，以便在我们重建的新的视图图中添加所需的场景。图 ?? 展示了我们的编辑机制的结果。

在图 ?? 中，我们使用单一输入图像生成多个帧。我们以多个目标相机参数查询相同的输入图像，并根据不同的目标视图重建新的视图。即使是远距离的视点，所提出的方法也可以获得良好的合成结果。

在图 ?? 中，我们展示了设置消融实验的定性结果。可以观察到，在“无交叉注意力”设置中，视图变换几何受损。在“Concat”设置中，物体在合成新视图时失去了细微的细节。完整模型显示了最佳的性能。

在表 ?? 中，我们报告了在 512×512 的原始合成分辨率上的结果。

我们在图 ?? 和图 ?? 中分别展示了使用我们的 3 类和 167 类模型的其他结果。