

ReconFusion: 3D Reconstruction with Diffusion Priors

摘要

神经辐射场 (NeRF) 等 3D 重建方法在渲染复杂场景的逼真新视图方面表现出色。然而，恢复高质量的 NeRF 模型通常需要数十到数百张输入图像，导致捕获过程十分耗时。为了解决这一问题，本论文提出了一种新方法——ReconFusion，该方法仅利用少数图像即可重建现实世界的场景。该方案结合了扩散先验进行新视图合成，并在合成与多视图数据集上进行训练，进而实现了在 NeRF 的 3D 重建框架中对新相机姿势的有效规范化，超越了传统输入图像集的限制。通过这种方式，该方法能够在保留观察区域外观一致性的同时，在欠约束区域内融合真实的几何和纹理信息。

关键词：神经辐射场；扩散先验；少数图像；重建

1 引言

现有的三维重建技术可以将图像转化为三维模型，从而实现不同视角下的逼真渲染。然而，传统方法如 NeRF [12] 从稀疏视角图像中重建的场景常常出现伪影问题，尤其是在视角捕获不足的区域。这意味着为了得到高质量的重建结果，就必须进行大量的图像采集，且每个区域需要从多个角度反复拍摄。甚至在一些简单的物体上也需要几十到上百张图像才能确保重建结果清晰无误。这对实际应用提出了巨大的挑战，特别是在仅依赖少量图像时，重建效果往往不理想。

为了减少对密集图像采集的依赖，许多方法通过低级启发式正则化来进行深度重建 [5]、可见性分析 [10]、外观恢复 [13] 或图像空间频率 [20] 的处理。然而，即使是最有效的方法，在与密集采集对比时，仍然表现出显著的效果下降，特别是在处理新视角时。近年来，扩散模型在图像生成 [8] 领域取得了显著成功，部分研究尝试将其应用于新视角合成任务 [6]。然而，这些模型通常无法生成一致的三维形状，因此无法作为通用的三维重建先验。

针对这一问题，该论文提出了一种结合二维图像先验的三维重建方法，利用扩散模型为新视角合成任务提供先验，并将其与传统 NeRF 重建流程结合。通过在多视角数据集上微调扩散模型，使其能够推断场景在未观测视角下的外观，从而为三维重建提供了有力的约束。

与现有方法相比，该方法在多个数据集上显著提高了重建质量。在稀疏视角观察场景中，它提供了一个强大的几何和外观重建先验；而在较为密集的采集条件下，它则能够减少常见的“雾霾”或“浮物”伪影，同时保留了采集充分区域的外观细节。该方法不仅在解决稀疏视角下的重建问题方面具有重要意义，也为利用更少的采集图像实现高质量的三维重建提供了可行的解决方案。

2 相关工作

2.1 稀疏视角 NeRF

许多研究集中在少视角条件下的优化方法，主要通过正则化几何信息来改进 NeRF 的表现。DS-NeRF [5] 使用稀疏深度输出作为监督信号来指导模型优化。DDP-NeRF [15] 则通过 CNN 模型从稀疏数据中获得稠密深度监督，进一步提升精度。SimpleNeRF [17] 通过减少位置编码频率和去除视角依赖部分来正则化几何和外观。FreeNeRF [20] 展示了简单地正则化位置编码频率范围，可以有效提升少视角情况下的重建质量。尽管这些方法在少视角场景中有显著改进，但当视角非常稀疏时，尤其是在大规模场景中，它们的效果仍然有限。

2.2 基于视图合成的回归模型

与 NeRF 针对每个场景进行优化不同，一些方法采用回归模型来进行视角合成。这些模型通过训练前馈神经网络，使用来自多个场景的多视角数据来学习泛化的视角合成能力 [4, 7]。例如，许多方法采用平面扫描体积的方式将输入图像提升到 3D 表示中，然后利用神经网络预测新的视角图像。这些方法在输入视角附近表现良好，但对于未观察到的视角，其推断能力往往较弱，尤其是在分布变得多模态时，重建效果会受到严重影响。

2.3 基于视图合成的生成模型

为了生成未知新视角的图像，生成模型（如生成对抗网络 GAN 和扩散模型）近年来得到了广泛应用。早期的研究主要依赖生成对抗网络 (GANs) 来合成新的视角图像 [1, 2]。然而，这些方法通常仅在已知视角数据的周围表现良好，而对于未知视角的合成，仍然存在较大的挑战。随着扩散模型在图像生成中的成功应用，越来越多的研究开始将扩散模型应用于视角合成任务。3DiM [19] 提出了训练一个基于扩散模型的姿势条件图像生成模型，用于合成 ShapeNet 数据上的新视角。GeNVS [3] 和 SparseFusion [18] 则将其应用于真实世界的多视角数据集，并结合 3D 几何先验，进一步提升了生成效果。然而，这些模型通常是针对特定类别的，并不具备泛化到任意场景的能力。Zero-1-to-3 [11] 通过对一个大型预训练的扩散模型进行微调，使其能在 Objaverse 数据集上进行 zero-shot 泛化并获得很好的效果，但它仅支持单一物体图像，并且要求背景干净，这使得其在复杂场景下的应用受到限制。为了克服这一局限，ZeroNVS [16] 对 Zero-1-to-3 进行了进一步微调，能够在一般场景中进行单图像重建。然而，这些方法大多仅限于单视图输入和特定场景类别。

2.4 提升 2D 扩散模型用于 3D 生成

鉴于现有 3D 数据集的稀缺，越来越多的研究尝试将 2D 扩散模型应用于 3D 资产的生成。这些方法通常通过文本提示或输入图像来生成 3D 内容。例如，DreamFusion [14] 提出了分数蒸馏采样 (SDS) 方法，在该方法中，2D 扩散模型作为一个评估员来监督三维模型的优化过程。SparseFusion [21] 则采用多步采样的方法，从当前渲染的噪声编码中生成图像，并将其作为三维重建的目标。这些方法的目标是通过强化 2D 扩散模型的条件生成能力，以便生成更为真实的 3D 结构。

3 本文方法

3.1 本文方法概述

本文提出的 ReconFusion 方法结合了基于扩散模型的新视角合成技术和 3D 重建流程，以有效地从稀疏视角的输入图像生成高质量的 3D 场景重建。具体的 pipeline 如图 1 所示：

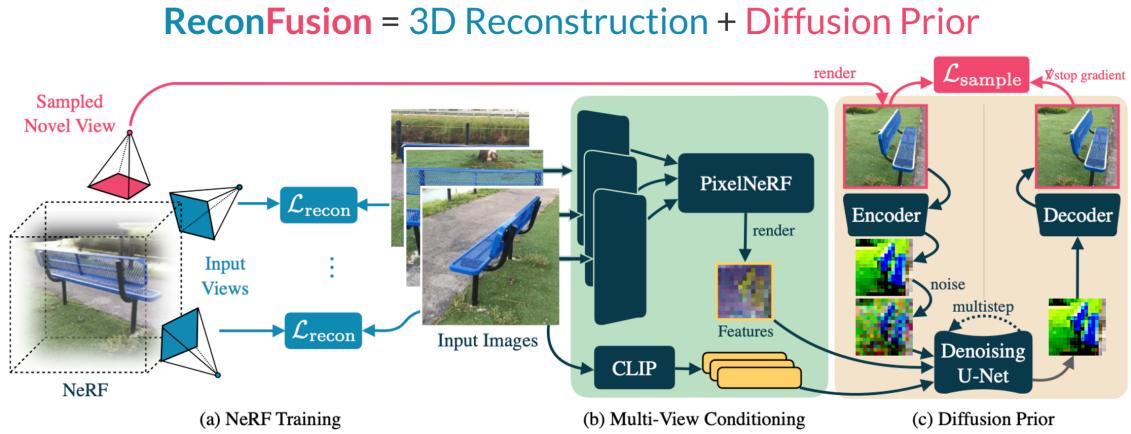


图 1. ReconFusion 主要框架示意图

ReconFusion 的核心思想主要分为两个部分：第一部分通过扩散模型生成合理的新视角图像，作为先验知识指导 3D 重建。扩散模型能够根据少量的已知视角图像和相机姿态，生成新的视角图像，从而学习场景的外观和几何特征。第二部分通过利用扩散模型的生成结果来优化 NeRF 的重建过程，通过将扩散模型的先验信息引入 NeRF 优化中，帮助改善几何一致性，并减少视角稀缺时产生的伪影和几何误差。

通过这种方法，ReconFusion 能够在少量视角输入的情况下，生成细节丰富且几何一致的高质量 3D 重建，极大地提高了 NeRF 在低视角场景中的表现。

3.2 基于新视角合成的扩散模型

本文提出的基于新视角合成的扩散模型框架旨在通过学习输入图像和相机姿态的条件分布来生成给定新视角下的图像。该方法利用了潜在扩散模型 (Latent Diffusion Models, LDM)，结合了变分自编码器 (VAE) 和去噪 U-Net 网络，在潜在空间内进行扩散，最终生成具有高质量和一致性的视角图像。这一方法的关键目标是通过生成逼真且符合几何约束的新视角图像，来为 3D 重建过程提供有效的先验信息。

扩散过程发生在潜在空间中，通过逐步引入噪声并使用去噪 U-Net 网络将噪声潜变量恢复为清晰的潜变量。在这一过程中，去噪 U-Net 网络的作用是将含有噪声的潜变量逐步恢复成清晰的潜变量，从而生成与输入条件一致的图像。生成的潜在变量通过 VAE 解码器进一步解码，最终恢复为新视角下的图像。

为了使扩散模型能够根据输入图像和相机姿态生成新视角图像，本文对传统的 U-Net 架构进行了关键性的改进。其搭建的预训练扩散模型的主要框架如图 2 所示：

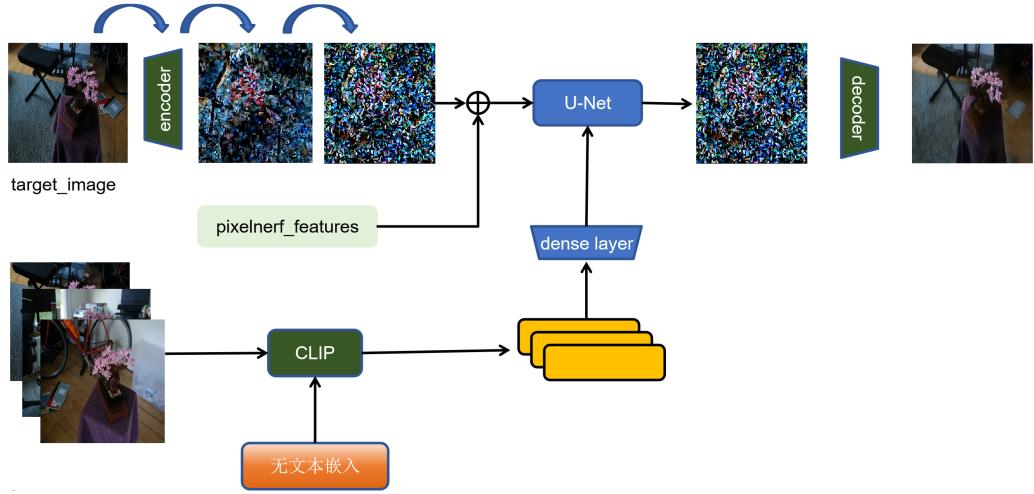


图 2. 扩散模型主要框架示意图

可以看到，扩散模型的条件输入不再局限于简单的图像特征，而是通过两种方式增强了条件信息的表达。一是为了传递输入图像的高层次语义信息，利用 CLIP (Contrastive Language-Image Pretraining) 模型对每一张输入图像进行嵌入 (embedding)。CLIP 的嵌入向量被作为条件信息传递给 U-Net 网络，以增强模型对图像语义的理解，从而能够在生成新视角图像时保持语义一致性。二则为了更好地结合几何信息，本文进一步利用 PixelNeRF 模型渲染目标视角的特征图。这一特征图的作用是提供相对于目标视角的几何信息，帮助扩散模型更准确地理解输入图像与目标视角之间的几何关系。通过这种方式，扩散模型不仅能够捕捉图像的语义信息，还能够有效结合几何信息，提高生成新视角图像的准确性。

其中选择 PixelNerf 作为几何信息提取模块的主要原因就是它本身也是对稀疏视图进行新视角合成工作，只是适用于 tiny data，并且对输入视角以及新视角的依赖较强。但不妨碍他可以从中提取几何特征指导整个扩散过程。PixelNerf 的主要框架如图 3 所示：

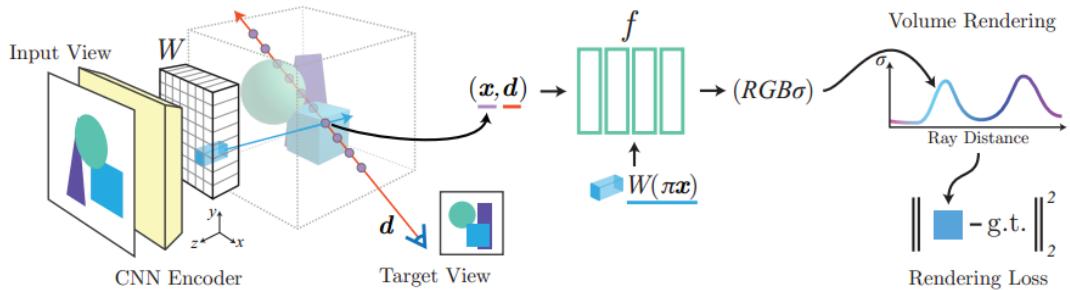


图 3. PixelNerf 主要框架示意图

语义和几何条件信息通过跨注意力机制 (cross-attention) 被传递到 U-Net 的不同层级，确保模型能够同时学习到图像的语义特征和几何结构，从而在生成新视角图像时更好地结合语义和几何信息。

3.3 基于扩散先验的三维重建

为了利用扩散模型进行高质量的三维重建，本文提出了一种增强的 3D 重建方法，基于 Zip-NeRF 的框架，并结合了扩散模型的先验信息。传统的 NeRF 方法通过优化 3D 场景的

参数，使渲染图像尽可能与给定视角的输入图像相匹配，但这种方法通常会出现几何不一致，尤其是在少视角或缺乏完整几何信息时，容易产生不真实的几何结构和浮动现象。

为了改进这一问题，本文通过将扩散模型生成的图像作为正则化先验信息，与传统 NeRF 重建过程相结合，提出了一种新的 3D 重建策略。在每次优化迭代中，扩散模型被用来生成一个目标视角的图像，随后将该图像与 NeRF 渲染的图像进行对比，从而计算差异并引入扩散损失。这一过程有效地帮助 NeRF 在缺少足够视角的情况下，生成更加稳定且几何一致的 3D 模型，避免了由于少量视角引发的几何不准确问题。

扩散模型通过学习输入图像及其相机姿态的条件分布，能够为新视角合成提供有力的指导。尽管扩散模型能够生成高质量的图像，但这些生成的图像在几何上不一定完全符合场景的真实结构。为此，本文引入扩散先验作为正则化项，确保 NeRF 生成的 3D 模型不仅在视觉上与真实图像一致，也能够在几何上保持一致性。

3.4 损失函数定义

本文的损失函数由两部分组成：重建损失和扩散损失。这两部分损失共同作用，优化 NeRF 模型，保证生成的 3D 重建既具有几何一致性，又具备高质量的纹理细节。下面将详细介绍这两种损失函数及其作用。

3.4.1 重建损失

重建损失的目标是使得 NeRF 渲染图像与实际观察到的图像之间的差异最小化。具体来说，对于每个输入的观察图像，NeRF 会根据相应的相机姿态渲染出一个对应的图像。重建损失通过计算渲染图像和实际图像之间的图像相似度来引导优化过程，从而使得 NeRF 的 3D 模型更好地匹配输入图像。

重建损失通常通过 L2 范数（均方误差）来度量图像之间的相似度：

$$L_{\text{Recon}}(\psi) = \mathbb{E}_{x_{\text{obs}}, \pi_{\text{obs}}} [\ell(x(\psi, \pi_{\text{obs}}), x_{\text{obs}})]$$

其中， $x(\psi, \pi_{\text{obs}})$ 是通过 NeRF 模型渲染得到的图像， x_{obs} 是实际观察到的图像， ℓ 是图像相似度损失函数，通常为 L2 损失或其他鲁棒损失函数。这个损失的优化目标是减少渲染图像与观察图像之间的差异，从而使得生成的三维模型更加准确。

3.4.2 扩散损失

扩散损失的目标是确保 NeRF 生成的图像在未观察视角下也能与真实场景相匹配。为此，扩散损失利用了从扩散模型获得的先验信息，指导 NeRF 优化过程生成更符合真实场景的三维模型。

每次优化迭代中，系统会随机选择一个新视角，并通过扩散模型生成该视角下的目标图像。生成图像作为目标图像后，与 NeRF 渲染的图像进行比较，计算它们之间的差异。该差异作为扩散损失的一部分，用于调整 NeRF 的优化方向。

扩散损失函数包含了以下两部分内容：

- L1 损失：衡量渲染图像与扩散模型生成图像之间的像素级差异。

- 感知损失 (Perceptual Loss): 采用 LPIPS (Learned Perceptual Image Patch Similarity) 等感知损失函数，衡量图像的结构和语义一致性。

因此，扩散损失可以表达为以下公式：

$$L_{\text{sample}}(\psi) = \mathbb{E}_{\pi,t} [w(t) (\|x(\psi, \pi) - \hat{x}_\pi\|_1 + L_p(x, \hat{x}_\pi))]$$

其中， $x(\psi, \pi)$ 是通过 NeRF 渲染得到的图像， \hat{x}_π 是通过扩散模型生成的目标图像， $L_p(x, \hat{x}_\pi)$ 是感知损失 (如 LPIPS)。此外， $w(t)$ 是一个噪声水平依赖的加权函数，用来平衡扩散过程中不同噪声级别的影响，确保优化过程中在不同噪声水平下的平衡。

4 复现细节

4.1 与已有开源代码对比

本文所复现的算法并没有公开源代码，然而，我们结合了该框架中涉及的多个开源项目，包括 Zip-NeRF、PixelNeRF、CLIP 图像文本编码器以及 Stable-Diffusion。这些开源技术为本项目提供了重要的技术支持，并且按照 ReconFusion 的实现细节复现了其在此基础上进行的创新与改进，最终基本实现了基于扩散先验的高质量 3D 重建方法。

4.1.1 Zip-Nerf

在三维重建方面，我们参考了 GitHub 上 SuLvXiangXin 提供的 zipnerf-pytorch 实现版本。Zip-NeRF 是一个基于 NeRF 的 3D 重建框架，能够从少量视角图像中重建出较为准确的 3D 场景。我们借用了 Zip-NeRF 的核心结构。然而，为了提升在稀疏视图新视角生成时的几何一致性和纹理真实感，我们将 Zip-NeRF 与扩散模型结合，利用扩散先验信息对重建过程进行正则化，从而有效避免了传统方法中出现的几何不一致和纹理浮动等问题。

具体细节调整如下：Distortion Loss 权重为 0.01；网格参数应用归一化权重衰减，强度为 0.1；采用宽度为 32、深度为 1 的较小视角相关性网络；对单位球外“收缩”区域的密度进行了降权处理：使用 Zip-NeRF 生成的密度乘以 $|\det(\mathbf{JC}(\mathbf{x}))|$ （收缩函数引入的各向同性缩放因子）。

4.1.2 PixelNeRF

在新视角图像合成方面，我们参考了 GitHub 上 kunkun0w0 提供的 Clean-Torch-Nerf 中 PixelNerf 的实现版本。PixelNeRF 采用了通过相机姿态信息和输入图像合成新视角图像的策略，尤其注重几何信息的融入。我们在此基础上，训练了自己的数据集，即打乱所有训练集数据随机取四张图片，前三张作为 input，第四张图的 pose 作为输入进行训练。训练完成后将 PixelNeRF 渲染的目标新视角图像提取特征图作为扩散模型的条件输入，以便将图像的语义信息与几何信息结合。

通过这一修改，扩散模型能够生成更加符合几何一致性的新视角图像，并且提高了图像质量。整体训练框架图如图 4 所示。

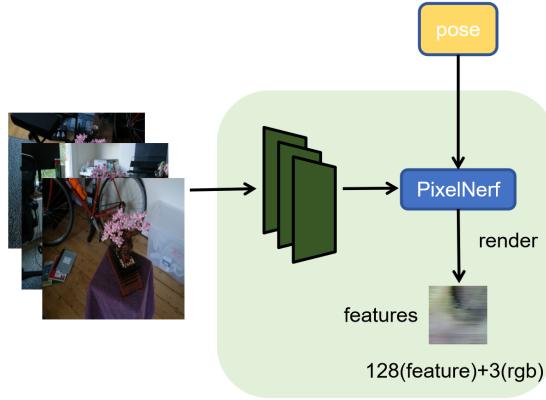


图 4. PixelNerf 训练框架示意图

4.1.3 CLIP 图像文本编码器

为了增强模型对输入图像的语义理解，我们使用了 CLIP 图像文本编码器。CLIP 能够将图像和文本映射到一个共享的嵌入空间，提供强大的语义表示。在本项目中，我们利用 CLIP 对输入图像进行嵌入，将其作为额外的条件信息传递给扩散模型。这种嵌入帮助扩散模型捕捉到图像的高层次语义信息，使得生成的新视角图像在视觉内容上更加一致和准确。

具体来说，在利用 CLIP 编码时我们不仅加入了三张输入图像的图像编码特征，同时也引入了无文本嵌入（即空串）的文本特征。通过增加一个密集层调整输入到 U-Net 网络的维度，并联合训练。其中密集层的权重设置采用了文本编码特征的均值，以让其开始训练时更偏向于无文本嵌入，使其训练后期更加关注输入的图像信息。

4.1.4 Stable-Diffusion

为了实现新视角图像的合成，我们基于 Github 上 CompVis 提供的 Stable-Diffusion 的开源代码进行修改。Stable-Diffusion 是一个高效的扩散模型，广泛应用于图像生成任务。在本项目中，我们将其从文本到图像生成的模型转变为基于输入图像和新颖相机姿态下的渲染图像来生成新视角真实图像的模型。通过修改 U-Net 架构，加入了 CLIP 嵌入和 PixelNeRF 生成的特征图，使得模型能够同时处理图像的语义和几何信息。

具体来说，PixelNerf 特征图与加噪到潜空间的渲染图拼接，并联合 CLIP 特征输入到 U-Net 训练。由于拼接后 U-Net 整个网络框架的维度发生了变化无法实现原论文中的微调目的。因此，我们参考了 Adapter 这篇文章 [9]，在 U-Net 前增加一个适配器，联合整个 U-Net 进行训练即可达到文章中微调的效果。这样能够使其更好地生成符合 3D 重建需求的图像。

4.2 实验环境搭建

4.2.1 硬件平台

为了加速模型的训练和推理过程，我们使用了以下硬件平台：

- **GPU**: 为了加速模型的训练和推理过程，我们使用了 NVIDIA RTX 4090 24GB 显卡。该显卡具备高效的深度学习计算能力，特别适合处理大规模的神经网络训练任务。

- **CPU**: 我们使用了 *Intel Core i9-14900K* 处理器，提供足够的计算能力来支持并行任务和数据预处理。

4.2.2 软件环境

本项目的软件环境主要参考了以下三个开源代码库：`zipnerf-pytorch`、`Clean-Torch-NeRFs` 和 `stable-diffusion`。具体的环境依赖文件基于这三个项目的 `environment.yaml` 进行了整合，确保了兼容性和高效性。以下是构建本项目部分所需的依赖项和环境配置，具体配置参考 Github：

- **PyTorch**: 用于深度学习模型的训练和推理，支持 GPU 加速。
- **CUDA**: 针对 NVIDIA GPU 的计算加速，确保了高效的模型训练。
- **Transformers**: 用于加载和使用预训练的语言模型，特别是 CLIP。
- **Diffusers**: 提供用于扩散模型的训练和推理的工具。
- **NumPy** 和 **SciPy**: 常用于数值计算和科学计算。
- **Pillow**: 用于处理图像输入输出。
- **OpenCV**: 用于图像处理和计算机视觉任务。
- **Matplotlib** 和 **TensorBoard**: 用于模型的可视化和监控。

4.3 创新点

本研究在现有的三维重建和扩散模型框架的基础上，提出了以下几个创新点：

- **结合扩散模型与 Zip-NeRF 进行新视角的三维重建**: 本研究首次将扩散模型作为先验信息引入到三维重建过程中，结合 Zip-NeRF 框架优化少量视角图像的三维重建。扩散模型通过生成新的视角图像来为传统的 NeRF 方法提供有效的正则化，有助于避免几何不一致和纹理浮动等问题，从而提高了三维重建的质量，特别是在少视角情况下。
- **扩散模型的改进**: 本研究对扩散模型进行了结构性改进，旨在更好地支持新视角的三维重建。我们结合了 CLIP 图像语义嵌入和 PixelNeRF 的几何特征渲染，将其作为条件信息传递给 U-Net 网络，从而增强了生成图像时对语义和几何的双重理解。通过这种方式，扩散模型不仅能够生成高质量的新视角图像，还能够有效地捕捉和保留输入图像的几何一致性和语义特征，为三维重建过程提供更强的先验信息，尤其在少视角输入情况下，显著提升了重建质量和准确性。
- **创新的扩散损失函数设计**: 本研究提出了一种新的扩散损失函数，结合了 L1 损失、感知损失 (LPIPS) 和噪声水平依赖加权函数，进一步提升了新视角生成的图像质量。该损失函数不仅有效地引导了图像生成的过程，还在优化过程中保持了新视角的几何一致性和纹理细节，克服了传统方法中由于缺乏先验知识而产生的几何和纹理不一致问题。

- **高效的多视角重建与优化方法：**为了更好地利用少量输入视角进行高效的三维重建，我们在优化过程中采用了新的图像合成方法，通过扩散模型和 NeRF 模型的联合优化，使得即使在有限视角下，仍然能够恢复出较为精确的三维结构，并且在生成新视角时避免了“浮动”效应和不准确的几何结构。

5 实验结果分析

本次复现工作主要集中在以下几个模块。

首先是 PixelNerf 训练模块，使其能够提取新视角下渲染图像的几何特征。具体的训练结果如图 5 所示。尽管该结果来源于 tiny data，并且该方法并不适用于真实场景数据集，因此效果上无法与真实场景进行直接对比。然而，由于我们仅利用 PixelNeRF 生成的几何特征来指导扩散模型的优化过程，整体上其作用主要体现在帮助生成更符合几何一致性的新视角图像，而非直接影响图像细节或纹理。

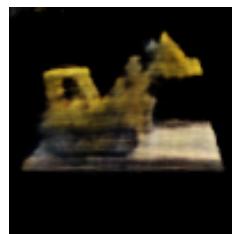


图 5. pixelnerf 训练结果图

接下来是 Diffusion VAE 的训练，以达到输入图像与重建图像相同。具体结果如图 6 所示。随着 VAE 的收敛，重建图像与输入图像逐渐一致。

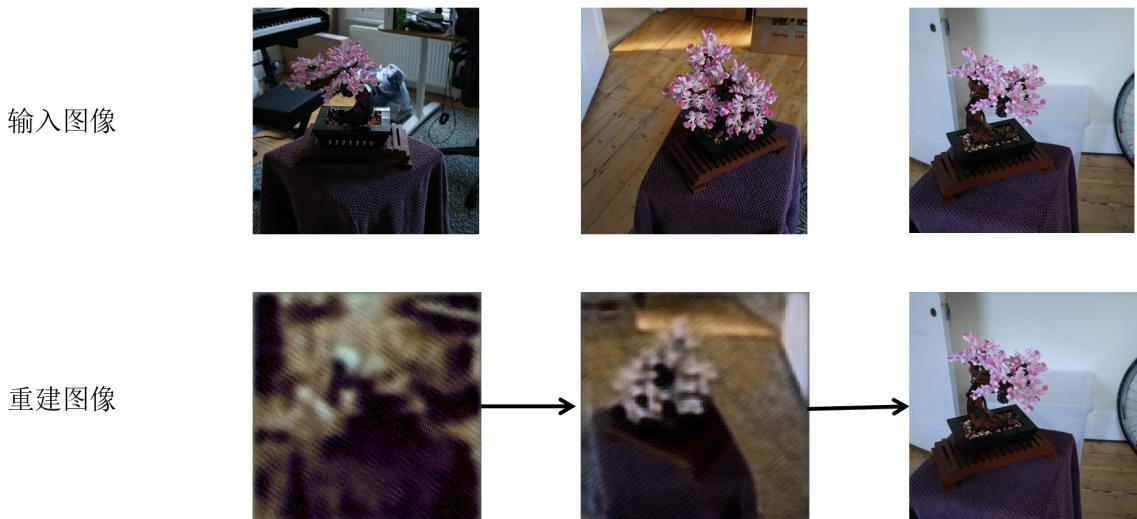


图 6. VAE 训练结果图

随后我们训练了集成 PixelNerf 和 CLIP 特征的扩散模型。为了评估模型在不同加噪步数下的表现，我们计算了生成图像与真实图像间三个常见的图像质量评估指标：PSNR、SSIM 和 LPIPS。实验结果如表 1 所示，表中列出了加噪步数为 200 步、300 步和 500 步时对应的定量评价指标。

表 1. 不同加噪步数下 PSNR、SSIM 和 LPIPS 的比较

Steps	PSNR	SSIM	LPIPS
200 steps	19.53	0.431	0.261
300 steps	18.54	0.403	0.315
500 steps	16.73	0.343	0.470

从表中可以看出，随着加噪步数的增加，模型的 PSNR 值有所下降，SSIM 和 LPIPS 指标也表现出一定程度的恶化。这里的 PSNR 相对较低是因为扩散模型的训练目前还不充分，仍未达到较好的收敛。但是从 LPIPS 可以看出生成图像与真实图像的感知相似度较好，即人类对图像内容的主观感知是相似的。

正如图 7 所展示的训练结果一样。尽管生成的图像与目标图像仍保持一定的相似性，但在较高噪声步数时图片细节表现有所下降，特别是桌子下方的书本部分。这也表明模型尚未充分学到所有细节信息，训练尚未完全收敛。若想获得更好的效果，仍需进行更长时间的训练和调参。



图 7. Diffusion 训练结果图

最后是将以上模型统一整合到主干 Zip-Nerf 中。由于 Nerf 运行速度太慢，同时每一轮训练都要调用 Diffusion 模型，导致训练达到收敛消耗的计算资源过大。不过从训练了一百五十轮渲染出来的结果来看，整体渲染的颜色过度得比较平滑，并逐渐有深度的信息传递出来。后面在计算资源充足情况下可以自行训练达到收敛查看效果。

6 总结与展望

本研究提出了一种结合扩散模型与 Zip-NeRF 的新颖三维重建方法，旨在通过利用扩散先验信息，提升少量视角图像下的三维重建质量。本文通过引入扩散模型生成新视角图像并将其作为辅助输入，结合 Zip-NeRF 框架，在少视角条件下生成了更具几何一致性和纹理真实感的高质量三维场景。

尽管本文提出的方法在少量视角输入下能够得到较好的三维重建效果，但仍然存在一些挑战和改进空间。首先，Zip-NeRF 联合 Diffusion 的训练过程计算资源消耗巨大，且训练时间较长，进一步优化和加速模型训练仍然是一个重要问题。未来的研究可以考虑替换现有的

三维重建方法，如使用基于 3D 高斯的重建方法，来提高训练效率和减少计算资源消耗。此外也可以应用于 3D 生成任务，即结合 2D 先验信息来引导 3D 高斯球生成，以此来生成一个具有一致性的三维场景。未来也可以通过将这种方法与现有框架结合，探索如何更有效地生成和优化三维场景。

参考文献

- [1] Eric Chan, Connor Z. Lin, Matthew Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, S. Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16102–16112, 2021.
- [2] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5795–5805, 2020.
- [3] Eric Chan, Koki Nagano, Matthew Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4194–4206, 2023.
- [4] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14104–14113, 2021.
- [5] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.
- [6] Jiatao Gu, Alex Trevithick, Kai-En Lin, Josh Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, 2023.
- [7] Philipp Henzler, Jeremy Reizenstein, Patrick Labatut, Roman Shapovalov, Tobias Ritschel, Andrea Vedaldi, and David Novotný. Unsupervised learning of 3d object categories from videos in the wild. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4698–4707, 2021.
- [8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

- [9] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. *ArXiv*, abs/1902.00751, 2019.
- [10] Minseop Kwak, Jiuhn Song, and Seungryong Kim. Geconerf: Few-shot neural radiance fields via geometric consistency. *arXiv preprint arXiv:2301.10941*, 2023.
- [11] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9264–9275, 2023.
- [12] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.
- [13] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [14] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ArXiv*, abs/2209.14988, 2022.
- [15] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12882–12891, 2021.
- [16] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Fei-Fei Li, Deqing Sun, and Jiajun Wu. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *ArXiv*, abs/2310.17994, 2023.
- [17] Nagabhushan Somraj, Adithyan Karanayil, and Rajiv Soundararajan. Simplenerf: Regularizing sparse input neural radiance fields with simpler solutions. *SIGGRAPH Asia 2023 Conference Papers*, 2023.
- [18] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9031–9042, 2023.
- [19] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *ArXiv*, abs/2210.04628, 2022.

- [20] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [21] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12588–12597, 2022.