浙江大學

硕 士 研 究 生 读 书 报 告

题目 ViewExtrapolator：基于视频扩散模型的视角外推技术读书报告

作者姓名 刘其

作者学号 22451058

指导教师 李启雷

学科专业 软件工程

所在学院 软件学院

提交日期 2024 年 12 月

摘要

由于辐射场方法的发展，新颖的视图合成领域取得了重大进展。然而，大多数辐射场技术在新视图插值方面比新视图外推要好得多，其中合成的新视图远远超出了观察到的训练视图。作者设计了 ViewExtrapolator，这是一种新颖的视图合成方法，它利用稳定视频扩散 (SVD) 的生成先验来实现逼真的新颖视图外推。通过重新设计 SVD 去噪过程，ViewExtrapolator 细化了由辐射场渲染的容易出现伪影的视图，大大增强了合成新视图的清晰度和真实感。ViewExtrapolator 是一种通用的新型视图外推器，可以处理不同类型的 3D 渲染，例如当只有单个视图或单目视频可用时从点云渲染的视图。此外，ViewExtrapolator 不需要对 SVD 进行微调，从而使其数据效率和计算效率都高。大量的实验证明了 ViewExtrapolator 在新颖的视图外推方面的优越性。

关键词：扩散模型，3D 生成

Abstract

The field of novel view synthesis has made significant strides thanks to the development of radiance field methods. However, most radiance field techniques are far better at novel view interpolation than novel view extrapolation where the synthesis novel views are far beyond the observed training views. We design ViewExtrapolator, a novel view synthesis approach that leverages the generative priors of Stable Video Diffusion (SVD) for realistic novel view extrapolation. By redesigning the SVD denoising process, ViewExtrapolator refines the artifact-prone views rendered by radiance fields, greatly enhancing the clarity and realism of the synthesized novel views. ViewExtrapolator is a generic novel view extrapolator that can work with different types of 3D rendering such as views rendered from point clouds when only a single view or monocular video is available. Additionally, ViewExtrapolator requires no fine-tuning of SVD, making it both dataefficient and computation-efficient. Extensive experiments demonstrate the superiority of ViewExtrapolator in novel view extrapolation.

Keywords： diffusion models, 3D generation

一、引言

随着计算机视觉和图形学领域的不断进步，特别是深度学习技术的崛起，3D 重建和视角合成已经成为计算机视觉中的重要研究课题。在这一领域中，传统的 3D 渲染方法，如光场（Light Fields）和辐射场（Radiance Fields）技术，已经取得了显著的进展。这些方法通过重建场景的高维信息，提供了精确的视角合成能力，并广泛应用于虚拟现实、增强现实、游戏开发以及其他沉浸式体验中。然而，尽管这些技术在视角合成（view interpolation）方面已经取得了令人满意的结果，但在视角外推（view extrapolation）任务中仍然存在显著挑战。

视角外推是指在给定有限的训练视角数据后，生成新的视角图像。传统的 3D 重建方法往往依赖于充分的视角数据，无法有效地从现有数据中推断出未见过的视角。因此，如何从有限的训练视角中有效地外推新视角，一直是该领域的难点。本文提出的 **ViewExtrapolator** 方法，通过引入最新的扩散模型（Diffusion Models），特别是基于视频扩散模型（SVD），在视角外推任务中取得了显著的效果。

本文将详细介绍 **ViewExtrapolator** 方法的背景、相关工作、模型原理和实验结果，并结合扩散模型的最新进展，探讨其在视角外推任务中的创新性贡献。

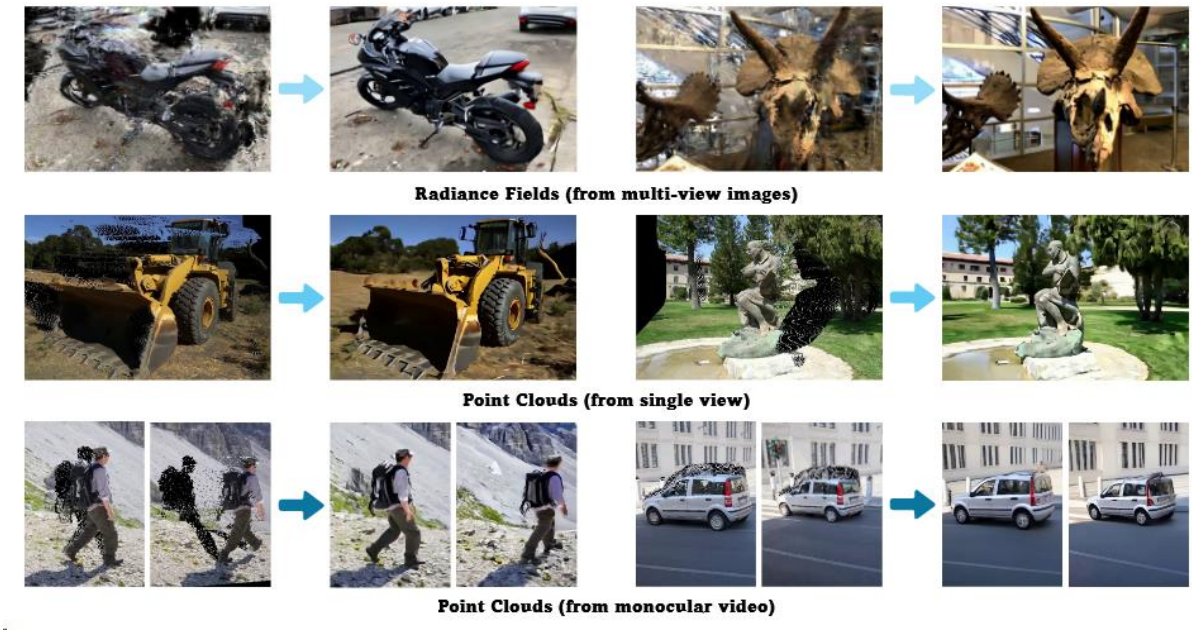

图 1 视角外扩效果

二、相关工作

视角合成和 3D 重建技术近年来取得了显著进展。早期的工作主要集中在基于几何模型的渲染技术，如**光场（Light Fields）**、辐射场（Radiance Fields）等。这些方法依赖于从多个视角采集场景数据，并通过体积渲染技术合成新的视角

图像。光场技术通过捕获场景的光照信息和方向信息，在不同视角下生成逼真的图像。Radiance Fields（NeRF）方法则通过将场景中的光照、密度信息映射到三维空间中的点，并通过体积渲染生成高质量的视角图像。然而，这些方法通常需要大量的训练数据和计算资源，并且在视角外推任务上仍然存在局限性。

随着生成对抗网络（GANs）和变分自编码器（VAEs）的提出，生成模型逐渐成为图像生成领域的重要工具。这些生成模型能够通过学习数据的潜在分布来生成新的样本，广泛应用于图像生成、图像修复、超分辨率等任务。特别是扩散模型（Diffusion Models）的提出，为图像生成任务带来了新的突破。扩散模型通过逐步去噪的过程来生成图像，这一过程在理论上具有较强的稳定性和生成质量。

在视角外推领域，尽管已有一些研究尝试解决这一问题，如 **ExtraNeRF** 和 **RapNeRF** 等方法，但它们依然受到训练数据有限性的制约，难以有效地生成新的视角。因此，如何在有限视角下进行高质量的视角外推，仍然是该领域的挑战。

扩散模型（Diffusion Models）作为一种新兴的生成模型，近年在图像生成、视频生成和 3D 重建领域展现了强大的潜力。尤其是 **DDPM（Denoising Diffusion Probabilistic Models）**，它通过逐步去噪的方式生成图像，表现出了比传统生成模型更高的质量和稳定性。

三、扩散模型（Diffusion Models）及 DDPM 的原理

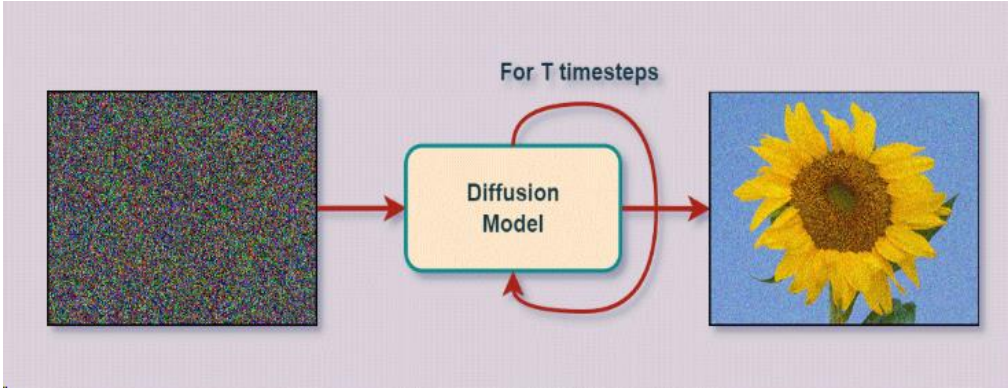

图 2 扩散模型生成清晰图片过程

扩散模型的核心思想来源于**热力学中的扩散过程**，这一过程描述了系统中粒子的逐步扩散，从无序到有序的转变。扩散模型将图像生成过程建模为一个逐步去噪的过程，通过从噪声图像开始，逐步恢复到清晰图像。在训练过程中，模型学习如何通过噪声的逐步去除生成目标图像。

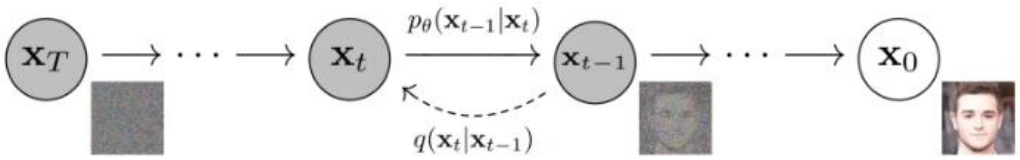

图 3 DDPM 流程图

3.1 基本框架

扩散模型的基本框架可以分为两个阶段：**正向过程**（Forward Process）和**反向过程**（Reverse Process）。

1. **正向过程**：在正向过程中，模型将真实数据逐步添加噪声，直到数据完全变为噪声。通过引入噪声，正向过程可以看作是数据从清晰图像到纯噪声图像的转变。这个过程通过一个马尔可夫链逐步实现，每一步都会对图像加入噪声。

2. **反向过程**：反向过程是扩散模型的关键，目标是从噪声图像恢复到原始图像。模型通过学习一个反向去噪过程，将噪声图像逐步去除噪声，还原成清晰的图像。反向过程的目标是最小化一个损失函数，该函数度量了模型生成图像和真实图像之间的差异。

3.2 生成模型：DDPM

DDPM（Denoising Diffusion Probabilistic Models）是扩散模型的一种实现方式，它通过学习正向过程中的噪声逐步去除的逆过程来生成高质量的图像。具体而言，DDPM 的生成过程如下：

1. 在正向过程中，从原始图像开始，逐步加入噪声，直到生成一个完全是噪声的图像。

2. 在反向过程中，模型学习如何逐步去噪，从纯噪声图像逐步恢复出原始图像。

DDPM 通过最大化数据似然函数来训练模型，使用变分推断方法优化去噪网络。通过对比正向和反向过程，模型能够有效地学习到如何去除噪声，并生成清晰的图像。

3.3 模型架构

DDPM 的网络架构通常包括一个去噪网络（Denoising Network），该网络通过接收噪声图像和时间步信息来预测原始图像的清晰版本。为了有效地进行去噪，DDPM 通常使用卷积神经网络（CNN）或者变换器（Transformers）等架构。这些网络能够捕捉图像的空间和时间信息，在去噪过程中逐步恢复图像的细节。

四、ViewExtrapolator 方法

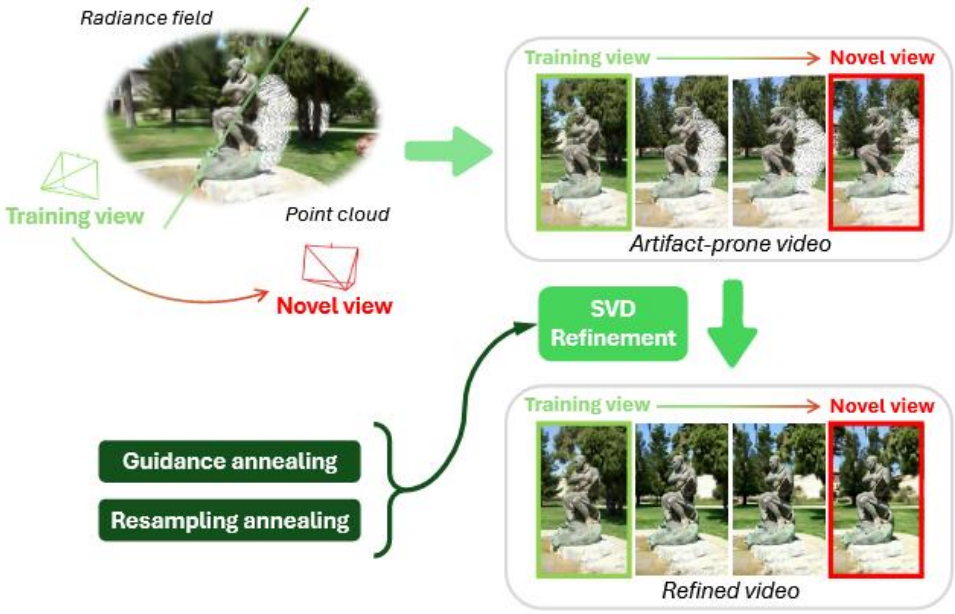

图 4 ViewExtrapolator 方法流程

本文提出的 **ViewExtrapolator** 方法（见图 4）结合了扩散模型的生成能力，特别是视频扩散模型（SVD），来解决视角外推问题。ViewExtrapolator 方法的核心创新在于，它不仅使用了基于视角的生成模型，还借助了扩散模型的强大先验知识，通过去噪过程有效地生成新的视角。

4.1 视频扩散模型（SVD）

Stable Video Diffusion（SVD）是本文方法的基础，它是扩散模型在视频生成任务中的应用。与图像扩散模型不同，SVD 需要处理视频中的时间维度。视频扩散模型通过对视频序列中的每一帧逐步去噪，生成连续、自然的视频流。在生成过程中，SVD 不仅关注每一帧的图像质量，还考虑到视频中帧与帧之间的连贯性。SVD 通过将视频帧建模为一系列状态，能够在生成过程中捕捉到时间信息，从而生成更为自然的视频。

4.2 引导退火与重采样退火

为了克服传统视角外推方法中出现的伪影，**ViewExtrapolator** 引入了两种新的退火策略：

4.2.1 引导退火（Guidance Annealing）

引导退火是一种用于扩散模型生成过程中的策略，它通过逐步减少噪声的影响，能够在去噪过程中提高模型生成的精度，特别是在外推视角的生成上起到关键作用。该方法不仅帮助模型在生成过程中更好地捕捉外推图像的真实细

节，还能有效避免生成图像中出现伪影或不自然的边缘，使得外推视角显得更加自然和流畅。

在扩散模型的训练中，噪声是模型生成图像的关键因素之一。正向扩散过程通过逐步将噪声加入到原始图像中，直到图像完全变成噪声，而反向去噪过程则是逐步去除噪声，从而恢复图像。在这个过程中，噪声对图像的影响不仅会改变其内容，还可能导致图像的细节丧失，尤其是对于视角外推任务，过多的噪声会导致图像中出现伪影或不一致的元素。

引导退火通过在训练和生成的每个步骤中逐渐降低噪声的影响，从而使模型在生成过程中能够更精确地控制图像的细节恢复。例如，在初期阶段，扩散模型会引入更多的噪声，以便探索广阔的潜在空间，并生成多样化的候选图像；而在后期阶段，噪声的强度逐渐减小，从而使得模型可以更加专注于恢复图像的细节，确保外推视角的图像内容与前后帧之间的连贯性。这种逐渐降低噪声的策略帮助去除了早期阶段的粗糙生成，减少了图像中的伪影，增强了模型对于外推图像的精确控制。

在外推任务中，视角外推需要对图像进行多次推测和反向调整。引导退火能够有效防止在这一过程中产生无法修复的伪影或细节缺失。通过精确控制噪声的强度和引导过程，模型可以逐步在空间和时间的维度上恢复细节，生成连贯且自然的外推视角。这一策略在动态视频生成和三维场景重建中尤其重要，因为它能够保证生成的视频不仅具有高质量的单帧内容，而且帧与帧之间保持自然的过渡。

Algorithm 1: Video refinement with guidance annealing and resampling annealing.

Input: artifact-prone video $\tilde{\mathbf{x}}$, opacity mask $\mathbf{m}$

$x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$
for $t = T, \ldots, 1$ **do**
  if $t > T - T^{\text{guide}}$ **then**
    for $r = 1, \ldots, R$ **do**
      $\hat{\mathbf{x}}_0 = \text{Predict}(\mathbf{x}_t)$
      if $r \leq R^{\text{guide}}$ **then**
        $\hat{\mathbf{x}}_0^{\text{dir}} = \tilde{\mathbf{x}} \odot \mathbf{m} + \hat{\mathbf{x}}_0 \odot (1 - \mathbf{m})$
      else
        $\hat{\mathbf{x}}_0^{\text{dir}} = \hat{\mathbf{x}}_0$
      $\mathbf{x}_{t-1} = \text{Denoise}(\mathbf{x}_t, \hat{\mathbf{x}}_0^{\text{dir}})$
      if $r < R$ **then**
        $\mathbf{x}_t \sim \mathcal{N}(\hat{\mathbf{x}}_0^{\text{dir}}, \sigma_t)$
  else
    $\hat{\mathbf{x}}_0 = \text{Predict}(\mathbf{x}_t)$
    $\mathbf{x}_{t-1} = \text{Denoise}(\mathbf{x}_t, \hat{\mathbf{x}}_0)$
return $\mathbf{x}_0$

算法 1 整体算法图

4.2.2 重采样退火（Resampling Annealing）

重采样退火是一种优化扩散模型生成质量的策略，它通过调整重采样过程中噪声分布的方式，使得生成的图像更加自然、一致，并且具备更高的视觉连贯性。这一策略的核心目标是在图像的生成过程中，尤其是在视角外推时，改善图像的质量和细节，以使得外推的图像不仅符合空间内容的要求，还能够有效地保持图像之间的时序一致性和过渡平滑性。

扩散模型的生成过程通常包括**重采样**步骤。在每次去噪的过程中，模型会对噪声进行重新采样，并通过一系列的计算步骤逐步去除噪声，恢复图像的原始结构。重采样的过程至关重要，因为它决定了生成图像的多样性和质量。在传统的扩散模型中，噪声的分布通常是固定的，这可能导致在生成过程中出现一些不自然的特征，尤其是在复杂的动态场景中，可能会导致视频帧之间的不一致性。

重采样退火通过动态调整噪声分布，在重采样的过程中引入变化，使得生成过程更加灵活，从而提升图像的质量。具体而言，重采样退火会在初期阶段采用较大的噪声尺度，以增加生成的多样性；而在后期阶段，逐步减小噪声的尺度，从而细化图像的细节并保证生成内容的一致性。通过这种方式，重采样退火不仅能够提高单帧图像的质量，还能增强图像之间的过渡平滑性，确保视频帧之间的连续性和自然流畅感。

在外推任务中，重采样退火尤为重要。因为外推图像通常需要从已有的图像内容中进行推测和延伸，而这一过程可能导致图像的细节丧失或不一致。通过引入重采样退火，模型可以在外推过程中灵活调整噪声分布，从而在保持图像连贯性的同时，优化图像的质量和自然性。此外，重采样退火还能够提高模型在生成多样性方面的能力，允许生成更为丰富的图像内容，使得外推视角更具真实性和细腻感。

与引导退火类似，**重采样退火**也通过调整噪声的尺度，使得模型在生成过程中能够更精细地控制噪声的影响，尤其是在时序一致性和空间细节之间找到平衡。尤其在视频生成中，重采样退火可以有效地减少生成过程中的不一致性，避免帧间产生不必要的跳跃或断裂。

4.3 训练与实验

在训练阶段，ViewExtrapolator 通过大量的视角数据进行学习。每个视频序列都通过扩散模型进行建模，并且通过反向去噪的方式恢复出新视角图像。在实验中，ViewExtrapolator 表现出了较传统方法显著的优势，尤其在复杂场景和长时间序列的生成上，能够生成更加真实且流畅的视频。

五、实验与结果分析

5.1 数据集与实验设置

为了验证 ViewExtrapolator 的有效性，本文使用了多个公开的视角外推数据集，包括 **Tanks and Temples** 和 **Multi-View Image Dataset**。在这些数据集上，模型训练通过对多个视角进行建模，并使用扩散模型生成新的视角。

5.2 结果分析

Methods	SSIM ↑	PSNR ↑	LPIPS ↓
3DGS	0.416	14.46	0.429
DRGS	0.406	14.68	0.457
ViewExtrapolator (video)	0.427	14.83	0.379
ViewExtrapolator (3DGS)	0.460	15.46	0.378
ViewExtrapolator w/o GA	0.442	15.14	0.448
ViewExtrapolator w/o RA	0.456	15.33	0.382

表 1 实现效果对比图

实验结果显示（见表 1），ViewExtrapolator 在多个任务中表现出了卓越的性能。与传统的视角外推方法相比，ViewExtrapolator 能够生成更为细腻的图像，并且在视频的连贯性方面取得了显著提升。通过扩散模型的引导退火和重采样退火策略，ViewExtrapolator 能够有效地避免传统方法中常见的伪影和不连续现象。

六、结论与未来工作

本文提出的ViewExtrapolator方法通过结合视频扩散模型（SVD），为视角外推任务提供了一种新的解决方案。实验结果表明，该方法能够有效地生成高质量的视角外推图像，并克服了传统方法中的一些局限性。未来的工作将进一步优化模型的生成效率，并探讨如何将该方法扩展到更多的实际应用场景中，例如虚拟现实和增强现实中。

扩散模型的成功应用不仅推动了视角外推任务的进展，也为图像生成领域带来了新的思路。随着生成模型和计算机视觉技术的不断发展，基于扩散模型的视角外推方法无疑将在 3D 重建、虚拟现实等领域发挥越来越重要的作用。