Research Statement

Zelin Gao (jamesgzl@zju.edu.cn / jameskuma.github.io)
Master in Control Science and Engineering, Zhejiang University

0. Abstract

My research experience spans computer vision, with specific research fields of **Simultaneous Localization and Mapping (SLAM)**, **Neural Radiance Fields (NeRF)**, and **Generative Models**. After defending my thesis on localization and representation in unknown environments via NeRF, my research projects now focus on object/scene generation by distilling the prior information from diffusion models into implicit representations.

1. Introduction

My ultimate goal is to build a human-like artificial intelligence (AI) system that can reconstruct the world with its eyes and imagine marvelously like a human brain. This AI system could serve as the fundamental component of some downstream tasks (e.g., robotics perception, autonomous driving, and content generation). Therefore, my research projects mainly focus on the following two parts.

- **Object / Scene Generation**: Machines and humans differ in their capacity for imagination, and what an exciting difference it is! Recent achievement in distilling the prior of diffusion model into generation have gifted machines the ability to imagine. Just like human-beings, these methods can generate 2D images / 3D objects from one conditional image or even a text prompt!
- **Implicit Representation**: Our world is built upon a large coordinate system! Recent implicit representation methods such as NeRF, SIREN, and Gaussian Splatting can inherently represent anything as the continuous implicit function, rather than using classical Point Clouds. These groundbreaking researches can revolutionize the way to reconstruct our world.

2. Specific Research Summary

2.1 Object / Scene Generation

(1) D⁴-Dreamer: Text-to-Non-Rigid Scene Generation. The revolutionary achievements in the field of text-to-everything generation, including not only text-to-image, text-to-video, and text-to-3D generation, is a key milestone in the field of computer vision. In this paper, we propose D⁴-Dreamer, a text-to-non-rigid scene generation method. As shown in Fig. 1, the spatial and temporal K-Planes are used to represent 3D constant and temporal motion, respectively. Given the text prompt, the multi-view sampling images rendered from neural representations are optimized by scene content guidance $\mathcal{L}_{SDS-MV}$, while scene motion guidance $\mathcal{L}_{VSD-T}$ is proposed to optimize the multi-timestep sampling images and generate reasonable motion. Moreover, the text-to-video diffusion model can be lora-finetuned by $\mathcal{L}_{LoRA-T}$ to improve its performance of video generation and even non-rigid scene generation over the text prompt without using any additional training dataset.

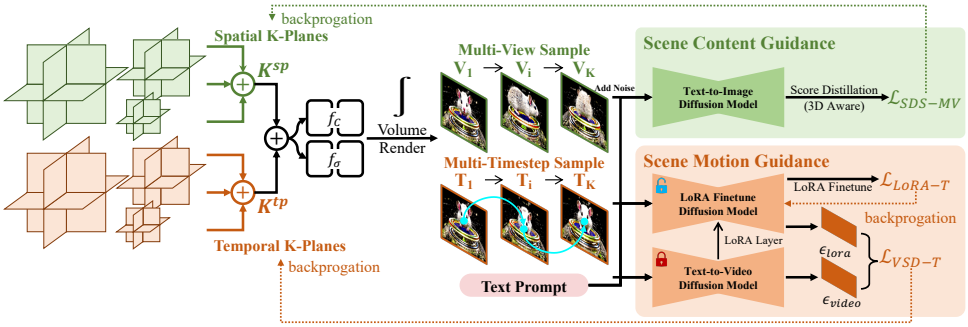


Figure. 1 The Overview of D⁴-Dreamer.

(2) SAP3D: The More You See in 2D, The More You Perceive in 3D. Humans can infer 3D structure from 2D images of an object based on past experience and improve their 3D understanding as they see more images. Inspired by this behavior, we introduce SAP3D, a system for 3D reconstruction and novel view synthesis from an arbitrary number of unposed images. Given a few unposed images of an object, we adapt a pre-trained view-conditioned diffusion model together with the camera poses of the images via test-time fine-tuning. The adapted diffusion model and the obtained camera poses are then utilized as instance-specific priors for 3D reconstruction and novel view synthesis.

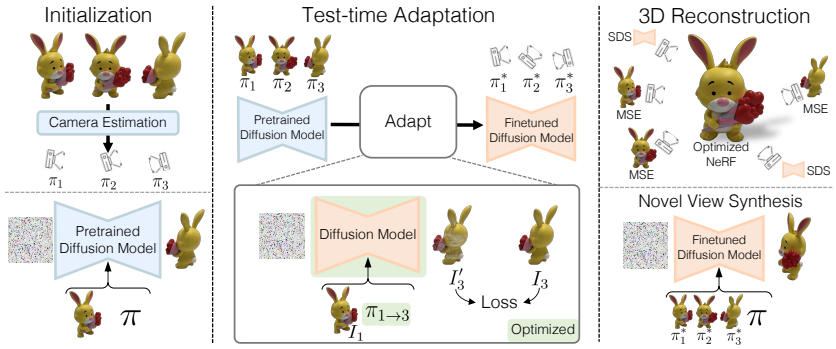


Figure. 2 The Overview of SAP3D.

As shown in Fig. 3, as the number of input images increases, the performance of our approach improves, bridging the gap between optimization-based prior-less 3D reconstruction methods and single-image-to-3D diffusion-based methods. We can thus demonstrate the great performance of our system on real images as well as standard synthetic benchmarks.

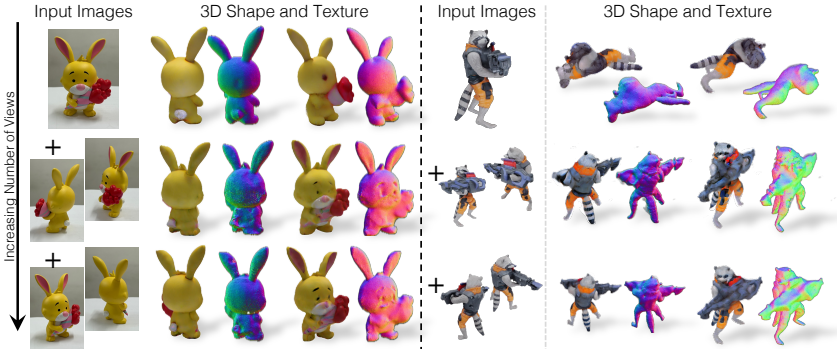


Figure. 3 3D Reconstruction from One or More Views

2.2 Implicit Representation Learning

(1) APE-BARF: Adaptive Positional Encoding for Bundle-Adjusting NeRF. In this paper, we address the issue of training neural radiance fields from unknown camera parameters (intrinsic and extrinsic) - the joint problem of reconstructing the 3D scene, registering the camera poses, and updating the camera intrinsic. We propose adaptive positional encoding (APE) for bundle-adjusting neural radiance fields that can simultaneously optimize implicit network parameters and camera parameters.

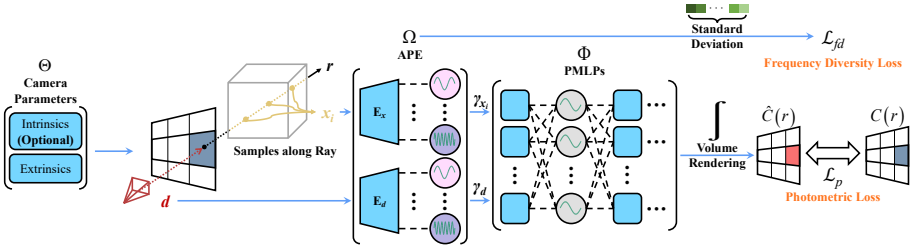
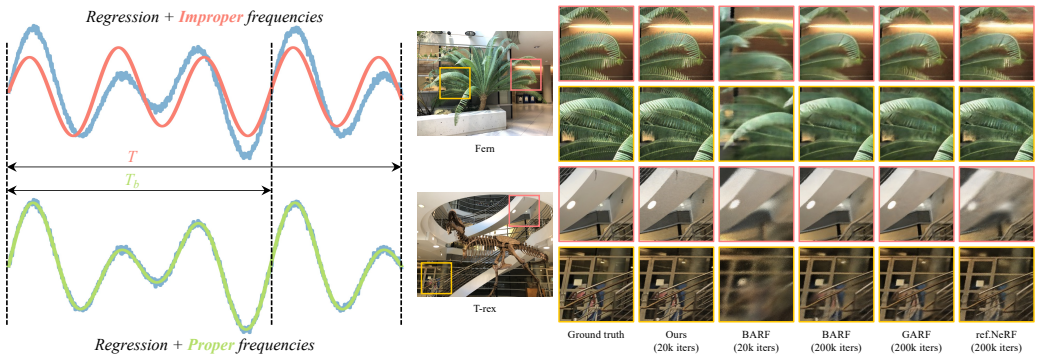


Figure. 3 The Overview of APE-BARF.

As shown in Fig. 4(a), the proposed APE is motivated by Fourier series regression, where improper frequencies in harmonic term degrade performance while adaptive frequencies can further improve the regression accuracy. Fig. 4(b) presents that our method 20k iterations, unknown camera poses) can synthesize more fine details as compared to other methods including BARF, GARF, and ref.NeRF (200k iterations, known camera poses).



(a) Fourier Series Regression. (b) Joint Optimization for Scene Representation.

Figure. 4 Qualitative Results of APE.

(2) HG³-NeRF: Hierarchical Guided NeRF for Sparse View Inputs. In this paper, we exploit the geometric, semantic, and photometric guidance to represent the neural radiance fields from sparse view inputs. We propose hierarchical geometric guidance (HGG) to sample volume points with the depth prior and hierarchical semantic guidance (HSG) to supervise semantic consistency of the complex real-world scenarios using CLIP.

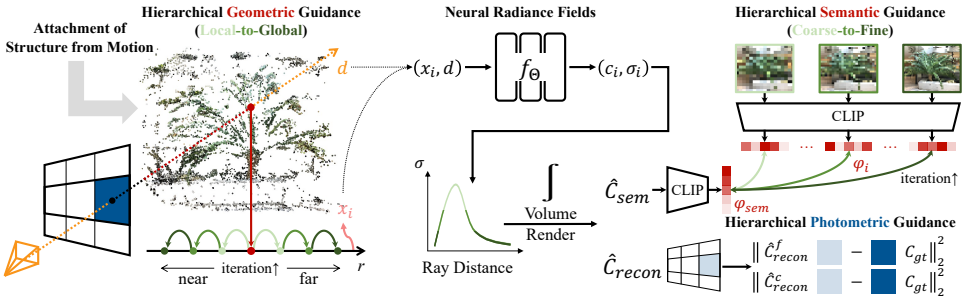
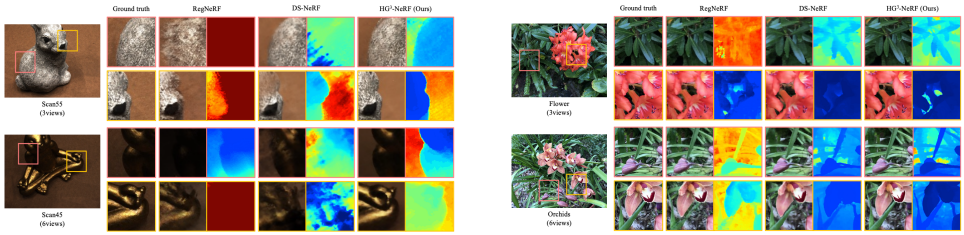


Figure. 5 The Overview of HG³-NeRF.

As shown in Fig. 6, we evaluate the effectiveness of the proposed HG³-NeRF in comparison to state-of-the-art baselines on various standard benchmarks, showing great performance in high-fidelity images synthesis and accurate depth estimation from sparse view inputs.



(a) Comparison on DTU. (b) Comparison on LLFF.

Figure. 6 Qualitative Comparison on Different Benchmarks.