Text-Guided Camera Pose Optimization with 3D Gaussian Splatting

Jirong Li[1], Satoshi Ikehata[2,3], Shuhei Kurita[2], Ikuro Sato[1,3]

[1]Institute of Science Tokyo, [2]National Institute of Informatics, [3]Denso IT Laboratory, Inc.

Recognition, Control and Learning Algorithm Lab.

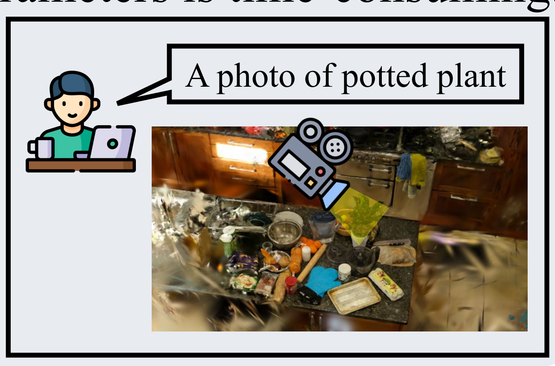
Introduction

Background

NeRF / 3D Gaussian Splatting (3DGS) has enabled us to render photos from any viewpoint. However, manually setting camera parameters is time-consuming.

Goal

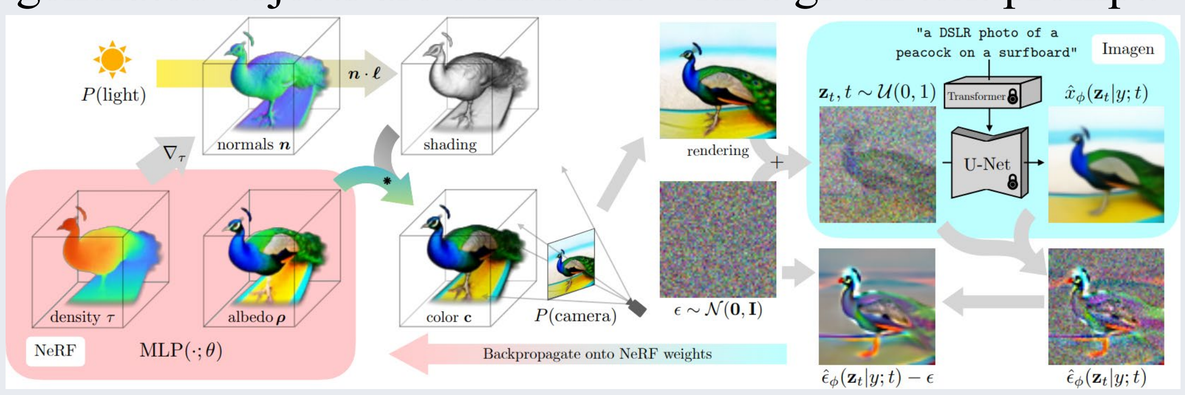
User can control camera by text prompt for a given NeRF / 3DGS.



Existing Method

DreamFusion [B. Poole+, ICLR 2023]

Text-to-3D task creates 3D assets from text so that generated objects are consistent with a given text prompt.



Score Distillation Sampling (SDS) can finetune the NeRF parameters θ so that rendered images are more consistent with a given text prompt.

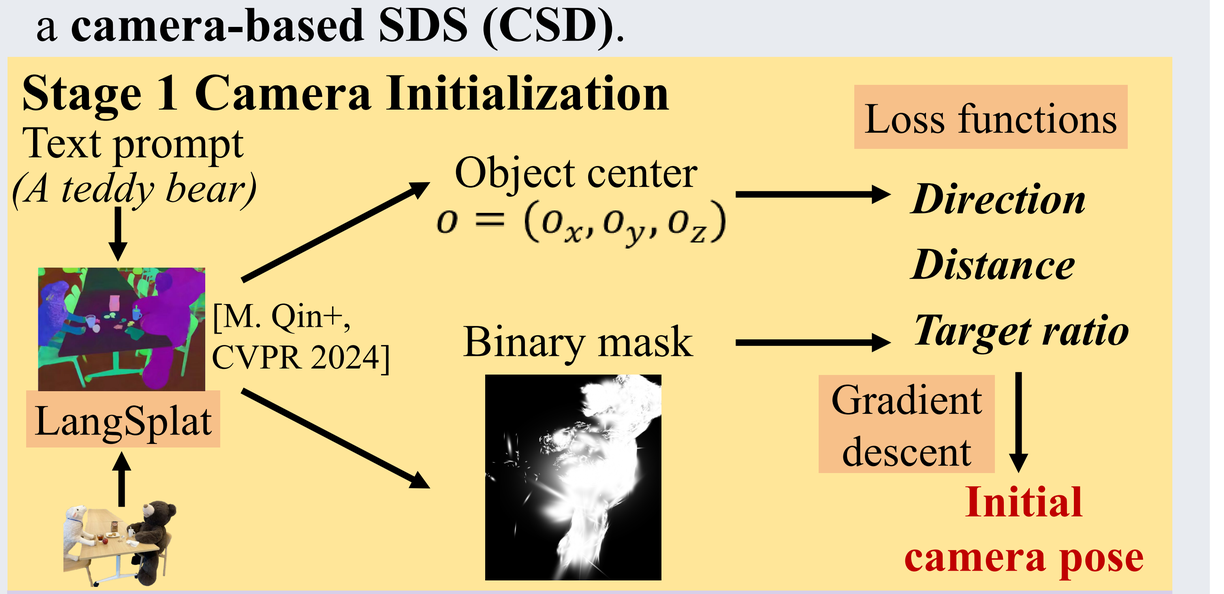
SDS Loss: $L_{SDS}(\phi, x)=\mathbb{E}_{t,\epsilon}\left[\omega(t)\left\|\epsilon_\phi(x_t, y, t)-\epsilon\right\|_2\right]$

Gradient: $\nabla_\theta L_{SDS}(\phi, x)=\mathbb{E}_{t,\epsilon}\left[\omega(t)\left(\epsilon_\phi(x_t, y, t)-\epsilon\right)\frac{\partial x}{\partial \theta}\right]$

ϵ_ϕ: diffusion model x: image y: text t: noise level ϵ: noise

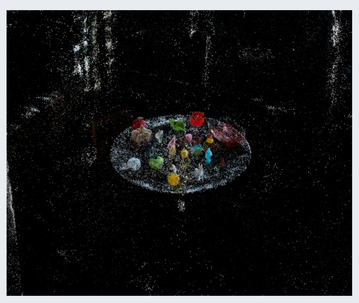
Proposed Method

Inspired by SDS, we backpropagate through diffusion model and 3DGS to optimize camera poses, proposing a **camera-based SDS (CSD)**.



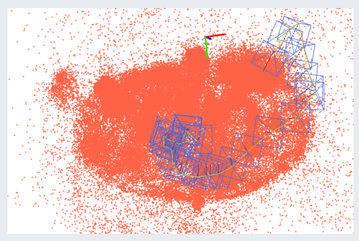
Experiment

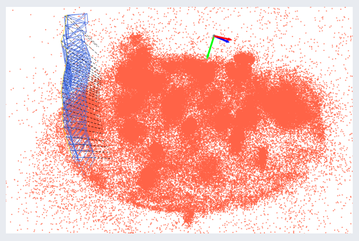
Scene
[J. Kerr+, ICCV 2023]
Figurines
（小さな造形）

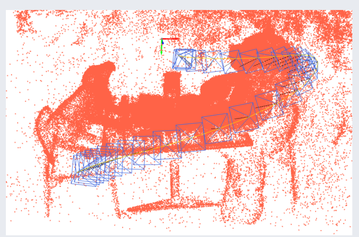


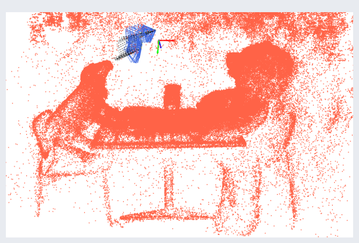
Teatime

Camera sequence









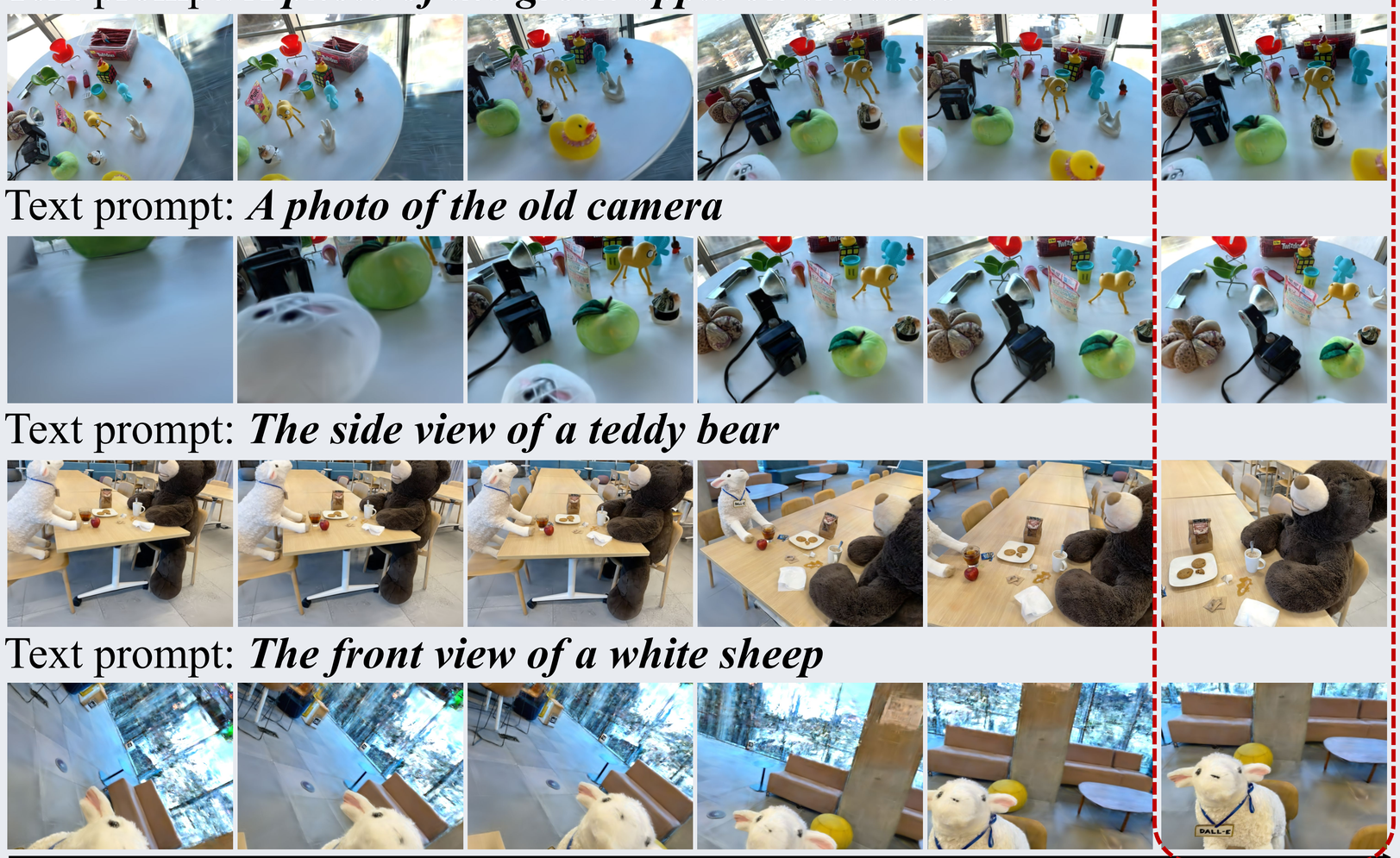
Text prompt: ***A photo of the green apple on the table***

Text prompt: ***A photo of the old camera***

Text prompt: ***The side view of a teddy bear***

Text prompt: ***The front view of a white sheep***

Final frame



This work is supported by DENSO IT LAB Recognition, Control, and Learning Algorithm Collaborative Research Chair (Science Tokyo).