Problem Definition

- Data: tuples of (image, pose, intrinsics) of several scenes.
- Use novel view synthesis to investigate whether model understands 3D appearance & geometry.
- Define neural scene representation $\Phi$ and differentiable neural renderer.
- Goal: Jointly solve for neural scene representation and renderer, supervised only with posed images!

Scene Representation Networks (SRNs) are a continuous neural scene representation. Along with a neural renderer, they model both 3D scene geometry and appearance, enforce 3D structure in a multi-view consistent manner, and naturally generalize shape and appearance across scenes.

Method

Representing scenes as functions

- Represent scene as function $\Phi$ that maps 3D world coordinates $x$ to feature descriptor of scene properties $\nu$:
  $$\Phi: \mathbb{R}^3 \rightarrow \mathbb{R}^n, \ x \mapsto \Phi(x) = \nu$$
  - $\nu$ can encode visual information, but also higher-level information (e.g. distance to closest scene surface)
  - Represent $\Phi$ as Multi-Layer Perceptron.
  - Phi represents geometry implicitly! How to render?

Neural Rendering via Differentiable Ray Marching

- Neural ray-marcher finds scene-camera-ray intersections.
- Begin at near plane. For every ray, sample $\Phi$ at current position, use feature vector $\nu$ to predict step length.
- When surface found, predict zero step size.
- Convert feature vector $\nu$ at last step to color with MLP.

Generalizing across scenes

- $i - th$ scene can be represented by weights of MLP $\phi_i$.
- Manifold assumption: weights $\phi_i$ live on low-dimensional manifold.
- Predict weights $\phi_i$ from scene latent vector $z_i$
  $$\Psi: \mathbb{R}^k \rightarrow \mathbb{R}^l, \ z_i \mapsto \Psi(z_i) = \phi_i$$
  - $\Psi(z_i)$ is parameterized as MLP (Hypernetwork), learned end-to-end with renderer, raymarcher.
  - $z_i$ are optimized jointly with network weights (Auto-decoder)

Abstract

Classic scene representations (voxel grids, points clouds, meshes) are discrete, limiting achievable spatial resolution and only sparsely sample smooth surfaces.

Neural scene representations have so far not exploited 3D structure.

---

Results

Appearance & geometry from 50 images

Few-shot reconstruction

Latent space interpolation

Non-rigid deformation

---

Project Page & Contact

vsitzmann.github.io
sitzmann@cs.stanford.edu