Predicting 3D structure by latent posterior sampling

Azmi Haider, Dan Rosenbaum

Department of Computer Science, University of Haifa

Goal:

A probabilistic 3D perception model that can predict a 3D scene given various types and levels of observations.

Contributions:

1. Reconstruct 3D scenes by sampling the posterior of a compressed 3D latent representation.
2. Efficient two-stage training:
   a. auto-decode compressed representations of 3D scenes using a **conditional NeRF**.
   b. train a **diffusion model** as a prior over the representations.
3. Considering the full posterior leads to **better reconstruction** and **uncertainty prediction** in various tasks: reconstruction from ➢ sparse views ➢ sparse depth data ➢ noisy images ➢ sparse pixels.

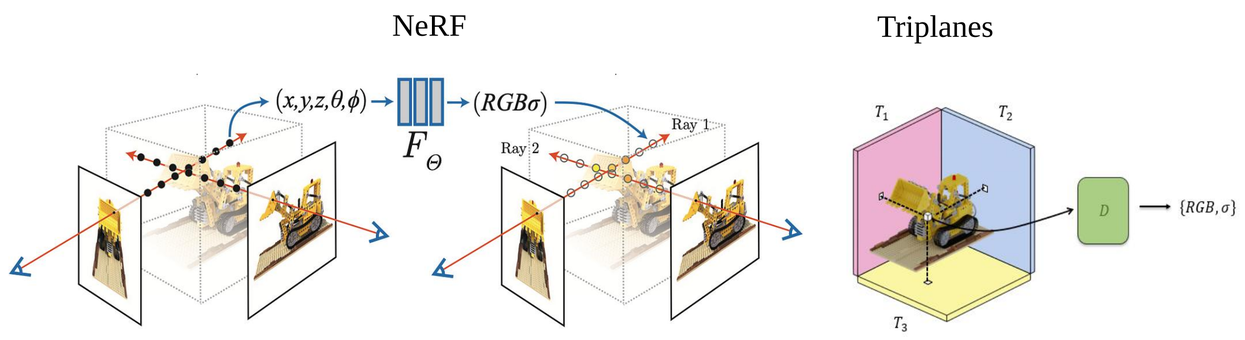
Background:

NeRF is an **MLP** trained to reconstruct images in a 3D scene.

- ➢ Requires many images of the scene to train.
- ➢ MLP is overfitted to one scene only.

Conditional NeRF – one shared MLP for all scenes.

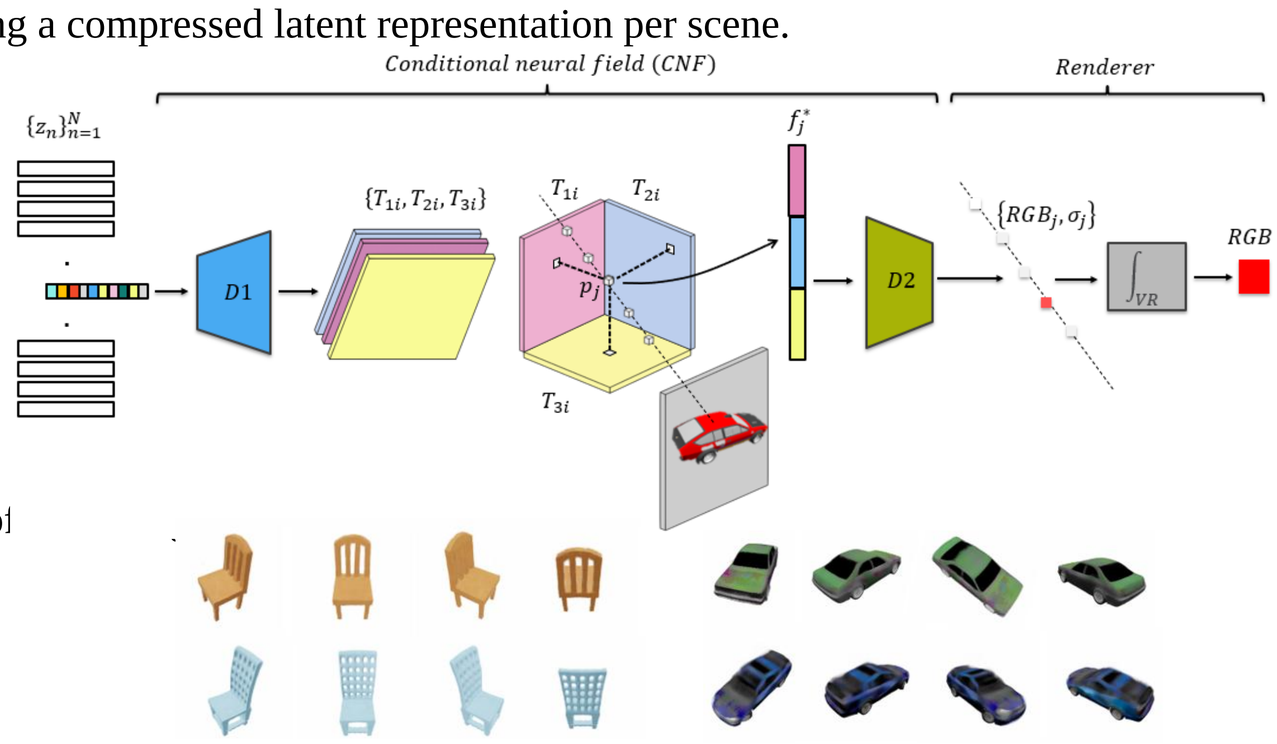
- ➢ Conditioned on a **triplane representation** of each scene.



Training the model:

1. Reconstruction model of latent representations

- ➢ Auto-decoding: Train a shared network while optimizing a compressed latent representation per scene.
- ➢ Latent decoder $D1$
  A shared network that maps latents to triplanes
- ➢ Triplane decoder $D2$
  a shared network that maps the triplane 3D interpolations to $\{RGB, \sigma\}$.



2. Train a generative model of the latents

- ➢ Diffusion model (DDPM): A learned iterative process of
- ➢ Learns a **prior** on the latents: $P(Z)$.
- ➢ Unconditional sampling from the prior: $Z_i \sim P(Z)$
  (views rendered via the trained reconstruction model)

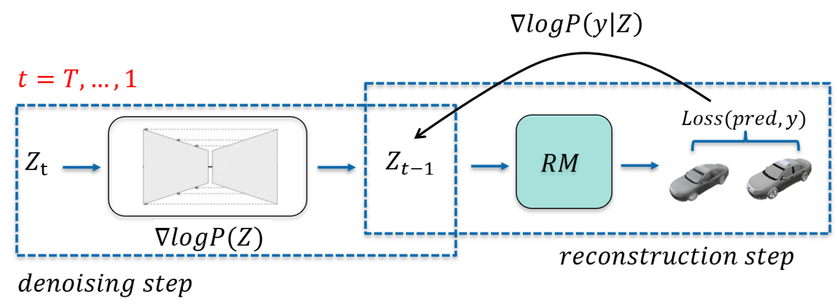
Inference – conditional generation:

Inference - conditional generation

- ➢ Guide the denoising process in the diffusion model
- ➢ Posterior sampling (given an observation y)
  (y: RGB, depth, full images or sparse pixels)

$$\nabla logP(Z|y) = \nabla logP(Z) + \nabla logP(y|Z)$$

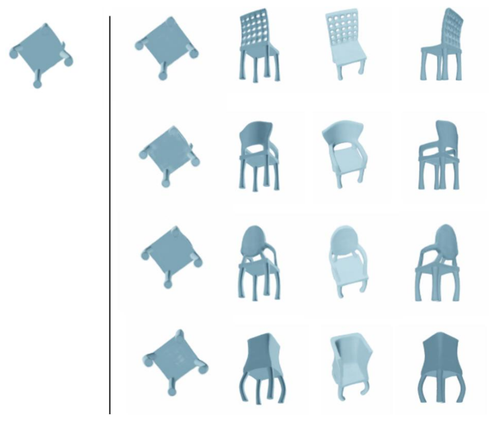
Prior: plausibility Likelihood: consistency



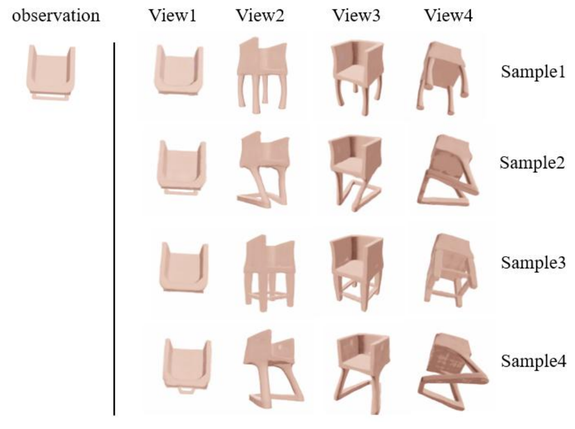
- ➢ Depending on the level of information given, different samples are generated (rows in the figures below).

likelihood is dominant

A tradeoff between prior and likelihood



Prior is dominant



- ➢ Different levels of information used as guidance:

Depth image

Half of an RGB image

Sparse pixels from RGB/depth image

Noised images