SinLayout: Learning Generative 3D Scene Layouts from a Single Image

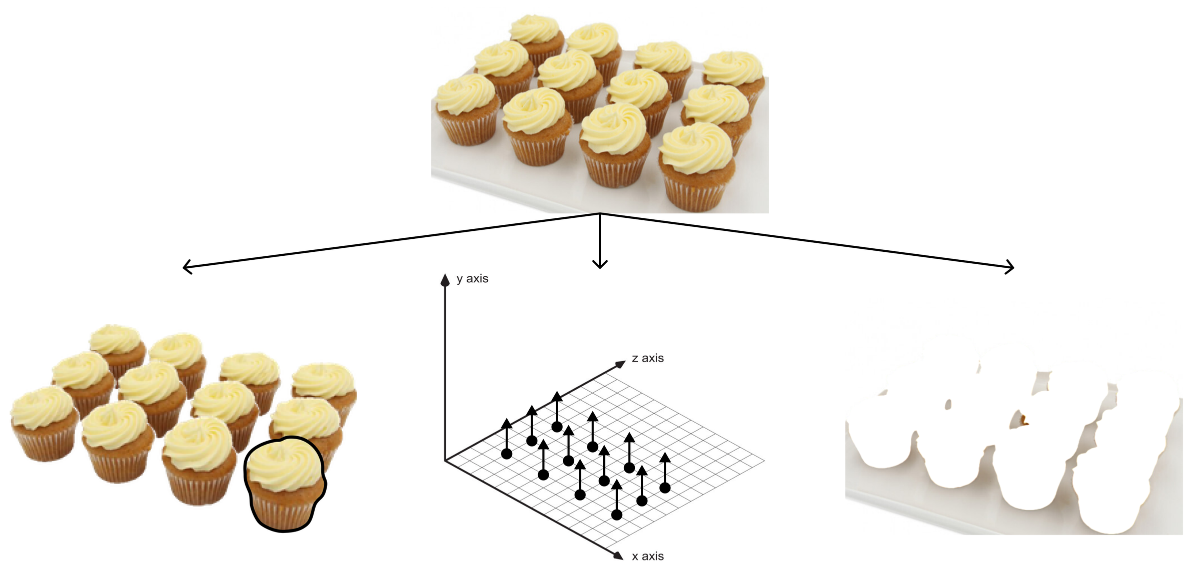
Linan (Frank) Zhao, Zeqing Yuan, Yunzhi Zhang, Shangzhe Wu, Jiajun Wu

Department of Computer Science, Stanford University

Stanford Computer Science

Research Overview

- Scene = Objects + **Spatial Arrangement** + Background.
- While recent works pushed the boundary for modelling 3D objects, learning the 3D layout is underexplored.
- Allows us to generate **plausible and diverse** 3D scenes.



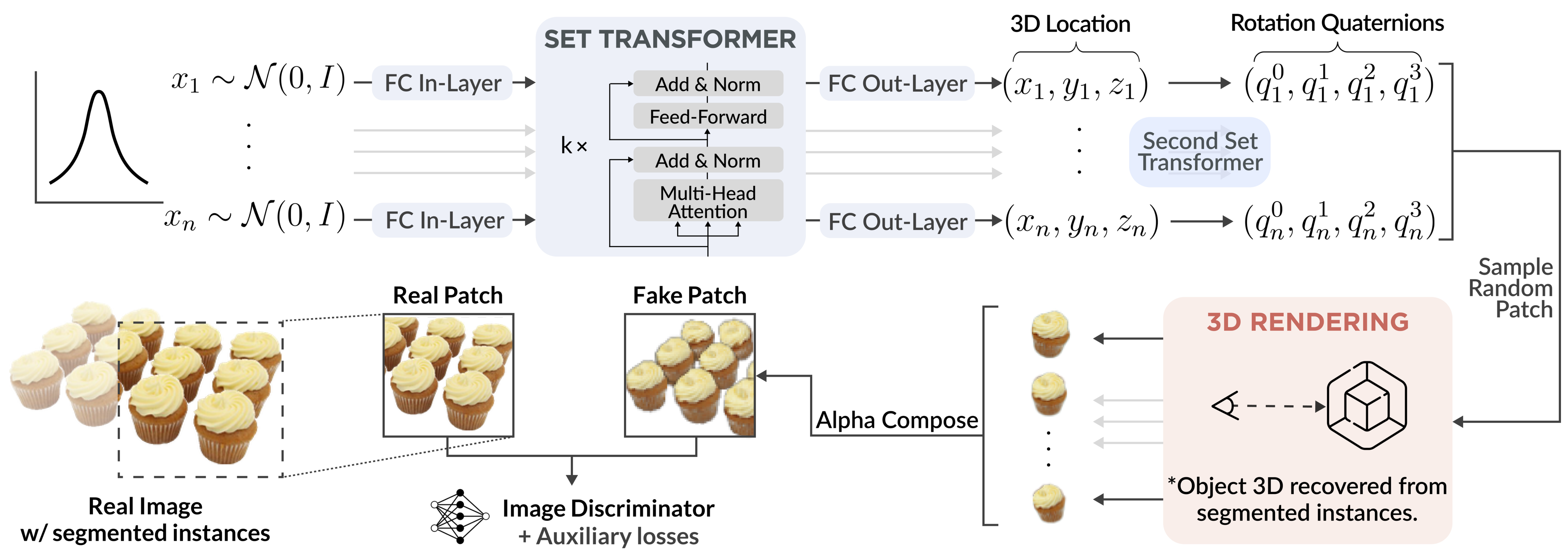
- **Problem Setting:**
  -- **Input:** a **single image** (segmented) of multiple objects from the same category.
  -- **Goal:** learn the objects' **3D layout distribution** (location + rotation) and generate similar scenes.
- **Main Approach:** Train a permutation-equivariant generator with a patch-based discriminator.
- **Main Contributions:**
  -- Propose the task of learning explicit 3D scene layout from a single image of multiple similar instances.
  -- Build a generative pipeline that fulfills this task.
  -- Demonstrate effectiveness on Internet images with interpolation and extrapolation results.

Datasets & Metrics

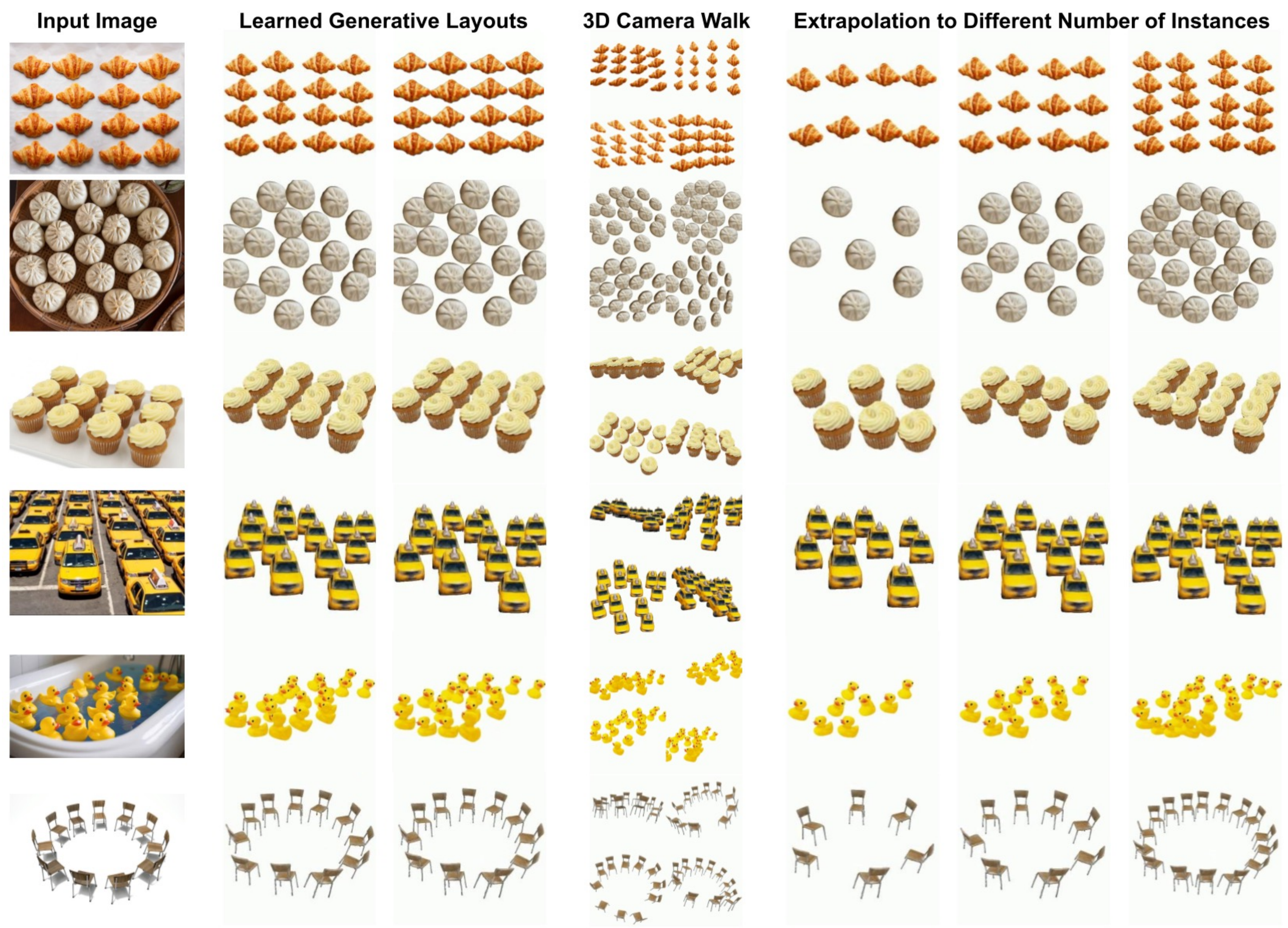
- Dataset contains single-view **Internet images**.
- Segment the foreground objects (e.g. with SAM).

- Rely on **qualitative** judgements since we have no ground-truth 3D poses of objects.
- Implement camera walks, latent walks, and extrapolation to different number of instances.

Main Pipeline



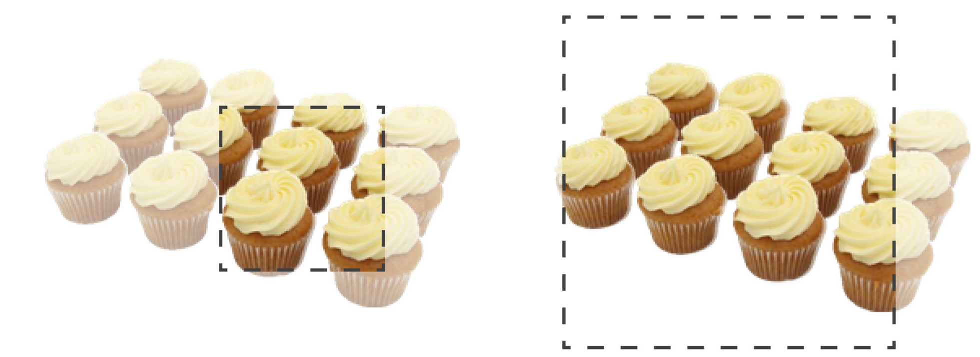
Results



Technical Details

- **Wonder3D** to obtain 3D object geometry from segmented instances. Mesh + layout = scene images.

- **2 Discriminators** with WGAN objective.



40% Crop   80% Crop

- **Augmentation:**
  Pose prediction + Random color background + Noise.
- **Auxiliary Losses:**

2D Location Loss   Overlapping Loss   Off-scene Loss

Analysis

- Learn **diverse** range of layouts from a single image.
- Novel views + Interpolation + Extrapolation abilities.
- **Disentanglement** between geometry and layout.

Layout transfer:

- Difficult to **quantitatively** evaluate the results.
- Single-view supervision → other views can have slight irregularities.
- Extrapolation to different number of instances seems to be interpolating among object instances.
- Training is unstable & fails for more **complex scenes.**
- Extend the pipeline to real images of multiple **different objects** (e.g. indoor scenes).