Towards Unsupervised Learning of Generative Models for 3D Controllable Image Synthesis

Yiyi Liao, Katja Schwarz, Lars Mescheder, Andreas Geiger
Max Planck Institute for Intelligent Systems, Tübingen
University of Tübingen, Amazon, Tübingen

**Task:** 3D Controllable Image Synthesis

- 3D controllability is essential in many applications, e.g., gaming, simulation, virtual reality and data augmentation
- 3D controllable properties: 3D pose, shape, appearance of multiple objects and camera viewpoint
- Is it possible to learn the simulation pipeline including 3D content creation from raw 2D image observations?

**Idea:** Learning the image generation process jointly in 3D and 2D space

**Method**

- **3D Representations:**
  - Foreground objects $o_f$: Primitive type: Point clouds, cuboids, spheres
  - Scene background $o_b$:
    - Spherical environment map

- **Loss Functions:**
  - Adversarial Loss:
    $$ L_{adv} (\theta, \psi, c) = \mathbb{E}_{p_{data}} [ f(d_o (g_n (\mathbf{z}, c), \psi))] + \mathbb{E}_{p_{U(I, c)}} [ f(-d_o (\mathbf{I}, c))] $$
  - Compactness Loss:
    $$ L_{compact} (\theta) = \mathbb{E}_{p_{data}} \left[ \sum_{i=1}^N \max \left\{ \tau : \| (A_i^* \odot (X_i - \mathbf{X})) \| \right\} \right] $$
  - Geometric Consistency Loss:
    $$ L_{geometric} (\theta) = \mathbb{E}_{p_{data}} \left[ \sum_{i=1}^N \| (A_i^* \odot (D_i^* - D_i)) \|_2 \right] + \mathbb{E}_{p_{data}} \left[ \sum_{i=1}^N \| (A_i^* \odot (D_i^* - D_i)) \|_2 \right] $$

**Generative Models**

- Classical Rendering Pipeline
- 3D Generator
- 2D Generator
- Differentiable Projection
- Render

**3D Generative Model**

- 3D controllable: Expensive and inefficient to design 3D models
- Not 3D controllable: Efficient and can be learned from only 2D images

**Our Approach**

- 3D controllable: Efficient and can be learned from only 2D images
- Not 3D controllable: Efficient and can be learned from only 2D images

**Quantitative Results**

<table>
<thead>
<tr>
<th>3D Representations</th>
<th>FID FID FID FID MVC’</th>
<th>X A I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point cloud</td>
<td>38 43 44 66 Good</td>
<td></td>
</tr>
<tr>
<td>Cuboid</td>
<td>38 45 45 60 Good</td>
<td></td>
</tr>
<tr>
<td>Sphere</td>
<td>33 45 45 53 Good</td>
<td></td>
</tr>
<tr>
<td>Deformable primitive w/o $g$</td>
<td>60 71 74 69 Good</td>
<td></td>
</tr>
<tr>
<td>Single primitive</td>
<td>30 38 44 – Poor</td>
<td></td>
</tr>
</tbody>
</table>

**Comparison to Baselines**

<table>
<thead>
<tr>
<th></th>
<th>Car Dataset</th>
<th>Indoor Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FID FID FID FID AX</td>
<td></td>
</tr>
<tr>
<td>Vanilla GAN</td>
<td>43</td>
<td>89</td>
</tr>
<tr>
<td>Lapa2d</td>
<td>43 56</td>
<td>84</td>
</tr>
<tr>
<td>2D Baseline</td>
<td>80 79</td>
<td>107</td>
</tr>
<tr>
<td>Ours (w/o o)</td>
<td>65 71 75</td>
<td>120 120 120</td>
</tr>
<tr>
<td>Ours</td>
<td>44 54 66</td>
<td>88 90 100</td>
</tr>
</tbody>
</table>

**Qualitative Results**

- Car Dataset
- Indoor Dataset
- Failure Cases