3D-Aware Scene Manipulation via Inverse Graphics

Shunyu Yao*[1], Tzu-Ming Harry Hsu*[2], Jun-Yan Zhu[2], Jiajun Wu[2], Antonio Torralba[2], William T. Freeman[23], and Joshua B. Tenenbaum[2]

[1]Tsinghua University, [2]MIT CSAIL, [3]Google Research

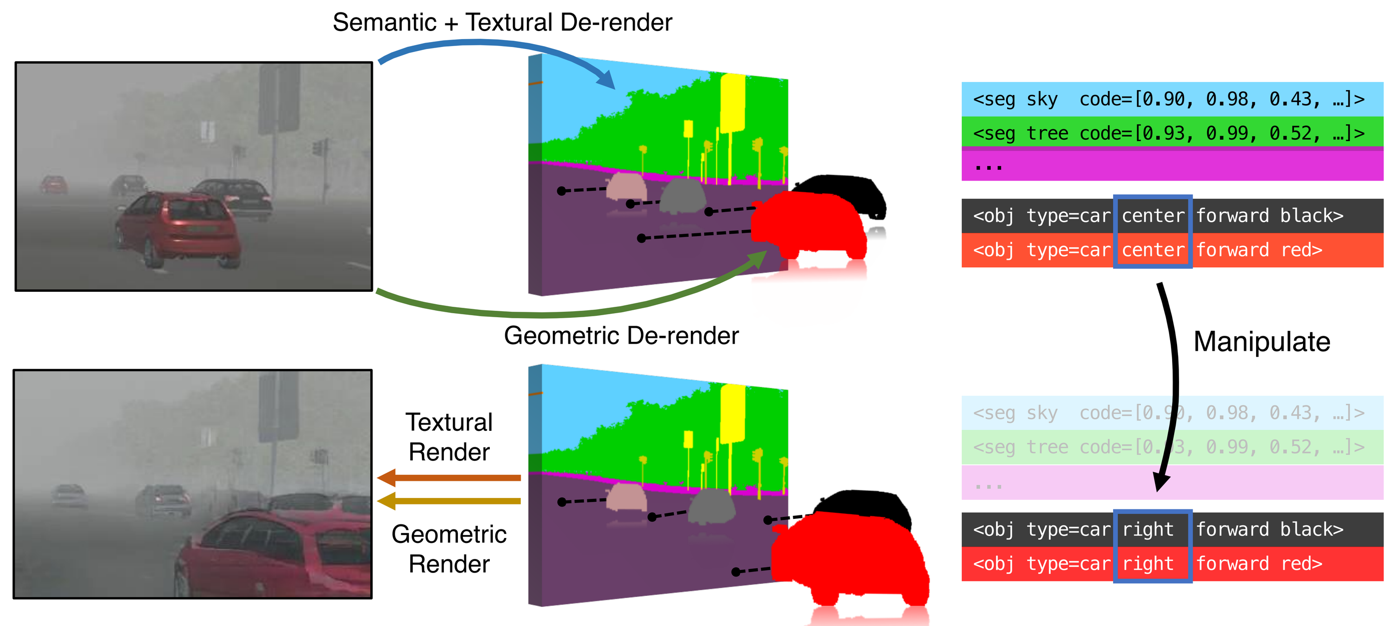
Motivation & Contributions

- Humans are good at ***perceiving*** and ***simulating*** the world with 3D structure in mind
- Previous deep generative models are often limited to a single object, hard to interpret, and missing the 3D structure

3D-SDN (ours) vs. 2D Method

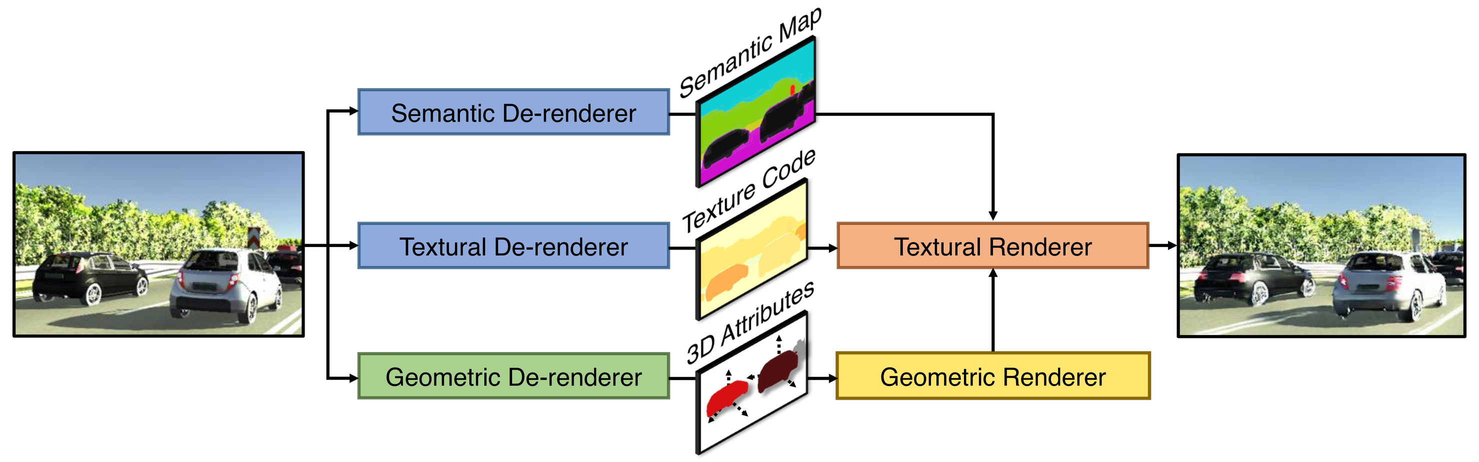
- **3D-SDNs** learn and incorporate
  (1) Scene semantic labels
  (2) Texture encodings for objects and the background
  (3) 3D geometry and pose for objects

Scene Manipulation via 3D-SDN

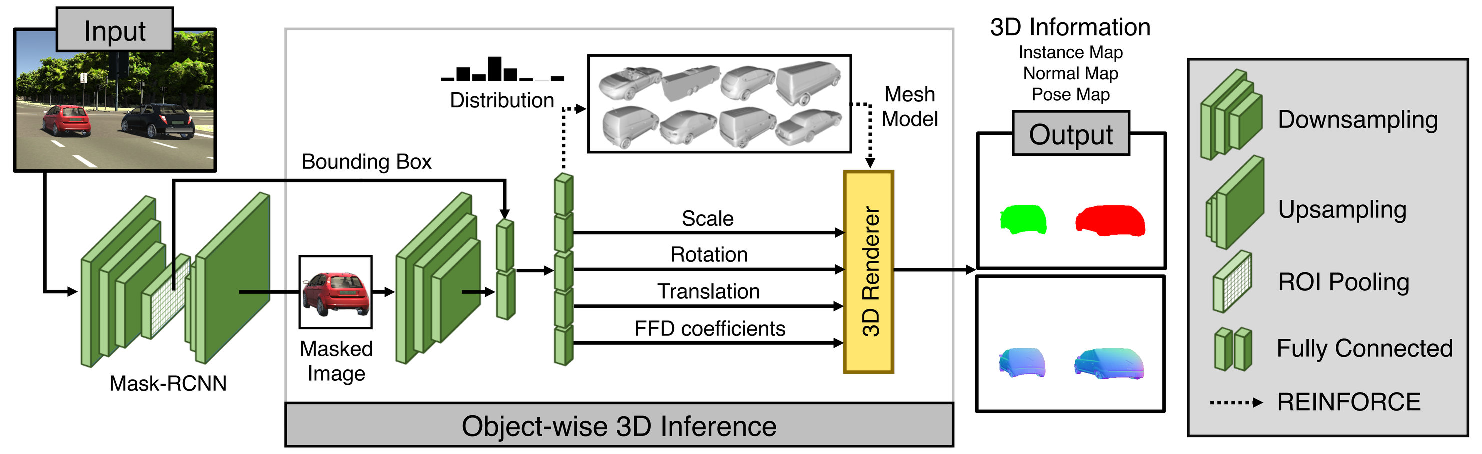


3D Scene De-rendering Networks (3D-SDN)

Overview



Geometric De-renderer & Renderer



- **Mask-RCNN** generates object proposals
- **3D De-renderer** infers object attributes and free form deformation (FFD) coefficients, and selects a mesh model

$$\mathcal{L}_{\text{pred}} = \underbrace{\|\log \tilde{\mathbf{s}} - \log \mathbf{s}\|_2^2}_{\text{scale}} + \underbrace{\left(1 - (\tilde{\mathbf{q}} \cdot \mathbf{q})^2\right)}_{\text{rotation}} + \underbrace{\|\tilde{\mathbf{e}} - \mathbf{e}\|_2^2}_{\text{2D offset}} + \underbrace{(\log \tilde{\tau} - \log \tau)^2}_{\text{depth}}$$

predicted

- **3D Mesh Renderer** renders silhouettes, and a **normal map** (byproduct: **object edge map** and **object pose map**)

$$\mathcal{L}_{\text{reproj}} = \left\| \tilde{\mathbf{S}} - \mathbf{S} \right\|$$

- **REINFORCE** + regular gradient train on the loss

$$\mathcal{L}_{\text{pred}} + \lambda_{\text{reproj}} \mathcal{L}_{\text{reproj}}$$

Semantic De-renderer

- DRN generates a **semantic label map** for the background

Textural De-renderer & Renderer

- **Image Translator Network** takes a **texture code map**, **semantic label map**, **normal map**, **object edge map** and **object pose map** as $\mathbf{L}$, generating an image $\tilde{\mathbf{I}}$ [pix2pixHD]
- **GAN Loss** accounts for photorealism

$$\tilde{\mathbf{I}} = \underbrace{G}_{\text{generator}}(\mathbf{L}, \overbrace{E}^{\text{encoder}}(\mathbf{L}, \mathbf{I}))$$

$$\mathcal{L}_{\text{GAN}}(G, D, E) = \mathbb{E}_{\mathbf{L},\mathbf{I}}\left[\log\left(\overbrace{D}^{\text{discriminator}}(\mathbf{L}, \mathbf{I})\right) + \log\left(1 - D\left(\mathbf{L}, \tilde{\mathbf{I}}\right)\right)\right]$$

- **Feature Matching Loss** stabilizes training

$$\mathcal{L}_{\text{FM}}(G, D, E) = \mathbb{E}_{\mathbf{L},\mathbf{I}}\left[\sum_{i=1}^{T_F} \frac{1}{N_i} \left\| \overbrace{F}^{\text{featurizer}}{}^{(i)}(\mathbf{I}) - F^{(i)}\left(\tilde{\mathbf{I}}\right) \right\|_1 + \sum_{i=1}^{T_D} \frac{1}{M_i} \left\| D^{(i)}(\mathbf{I}) - D^{(i)}\left(\tilde{\mathbf{I}}\right) \right\|_1\right]$$

- **Reconstruction Loss** encourages reconstruction

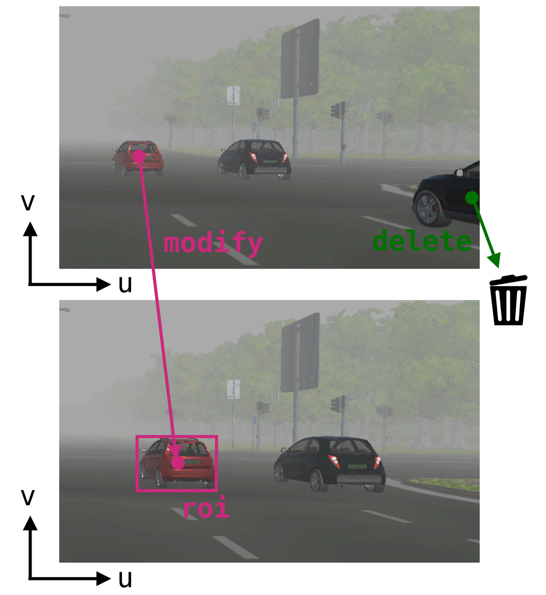
$$\mathcal{L}_{\text{Recon}}(G, E) = \mathbb{E}_{\mathbf{L},\mathbf{I}}\left[\left\| \mathbf{I} - \tilde{\mathbf{I}} \right\|_1\right]$$

- **Minimax Game** trains on the loss

$$G^*, E^* = \arg\min_{G,E} \left(\max_D \left(\mathcal{L}_{\text{GAN}}(G, D, E) \right) + \lambda_{\text{FM}} \mathcal{L}_{\text{FM}}(G, D, E) + \lambda_{\text{Recon}} \mathcal{L}_{\text{Recon}}(G, E) \right)$$

Virtual KITTI Editing Benchmark

- 92 pairs of images picked from **Virtual KITTI dataset**
- Each pair contains operations in `.json` format




```
{
  …
  "operations": [
    {
      "type": "modify",
      "from": {"u": "750.9", "v": "213.9"},
      "to": {
        "u": "804.4", "v": "227.1",
        "roi": [194, 756, 269, 865]
      },
      "zoom": "1.338",
      "ry": "0.007"
    },
    …
    {
      "type": "delete",
      "from": {"u": "1328.5", "v": "271.3"},
    }
  ]
}
```


vkitti_editing_benchmark.json

Results

Virtual KITTI Editing Benchmark

- **2D/2D+**: only **texture code map** and **semantic label map**; naïve translation and scaling (+out-of-plane rotation)

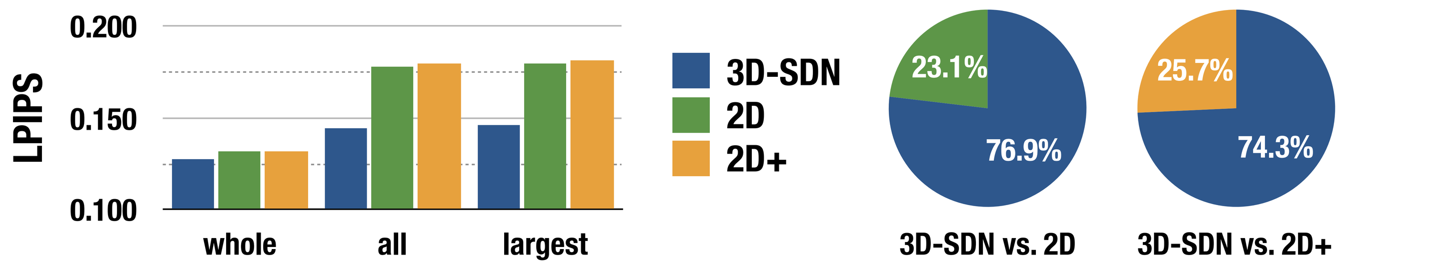


Image Editing Examples

References | [pix2pixHD] Wang et al. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs, In CVPR, 2018.