Neural Scene De-rendering

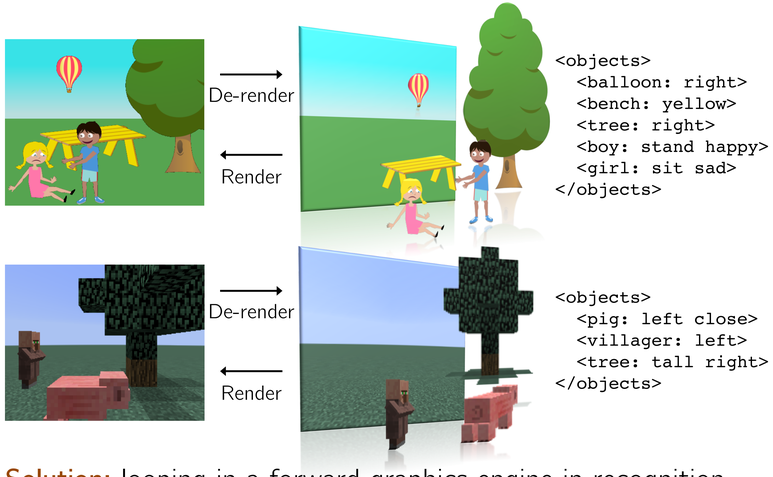
Jiajun Wu[1] Joshua B. Tenenbaum[1] Pushmeet Kohli[2,*]
1 Massachusetts Institute of Technology 2 DeepMind * Work done when the author was with Microsoft Research

Scene De-rendering

Goal: a compact, interpretable scene representation

Motivation

- An object-based, disentangled representation has wide applications
- Representations learned by current deep nets are hard to interpret

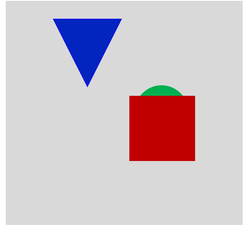


Solution: looping in a forward graphics engine in recognition

Advantages

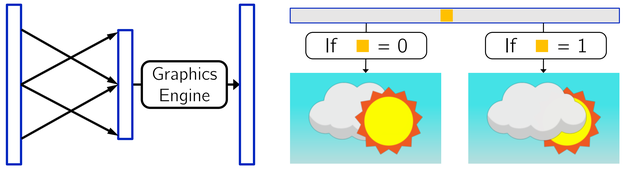
- Graphics engines bring in symbolic representation naturally
- Graphics engines generalize well to a variable number of objects
- The learned representation is rich, and has wide applications.

Scene XML



```
<object>
    <category>triangle</category>
    <size>1.5</size>
    <color>blue</color>
    <position>1.5,2,1</position>
    <yaw>0</yaw>
    …
</object>
<object>
…
```

Inference & Reconstruction

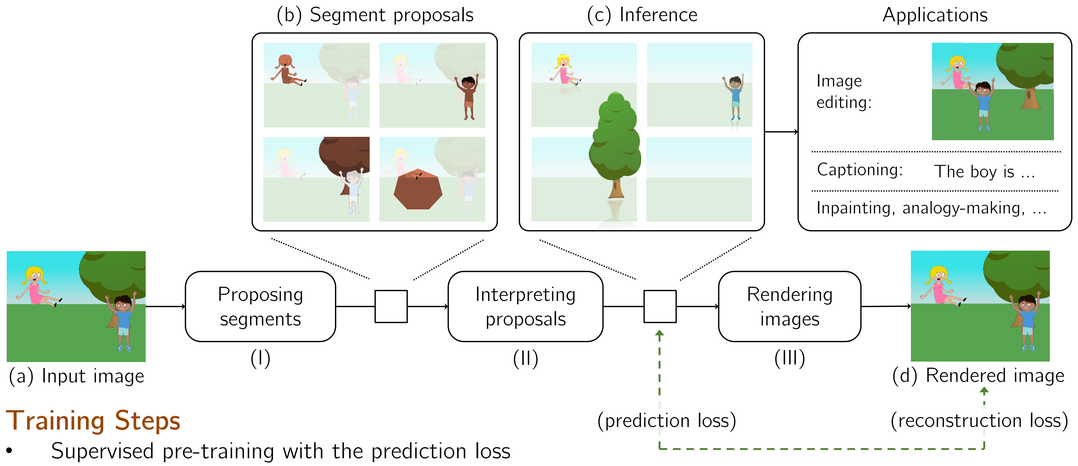


- Graphics engines as generalized decoders
- Visually distinctive images may have similar representations
- Solution: Optimizing in both spaces

Results

(a) Input image (b) CNN+LSTM (c) NSD (d) NSD (full)

Model

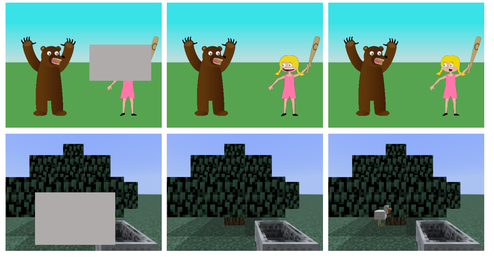


Training Steps

- Supervised pre-training with the prediction loss
- End-to-end fine-tuning with the reconstruction loss (with REINFORCE)

Applications

Inpainting

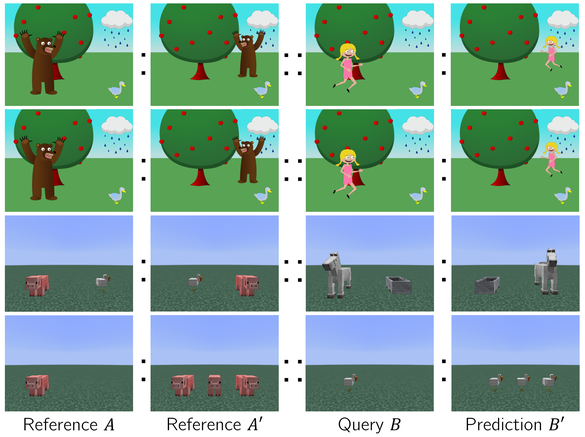


(a) Corrupted input (b) Reconstruction (c) Original image

Caption Retrieval

jenny and mike are having fun in the sandbox unaware of the storm that's coming their way

Analogy-Making



Reference A Reference A' Query B Prediction B'