Mug Makeover: Generating Realistic Ray-Traced Images from Rasterized Renders

CS231N: Thomas Yim, Daniel Lee, Caroline Cahilly

Background + Problem

Background:

- Ray tracing and rasterization are rendering methods that produce 2D images from 3D models
- Ray tracing looks much more realistic since it accurately simulates the effects of a large number of light rays hitting the material
- Rasterization is much faster than ray tracing but much less accurate due to polygon approximations
- Prior work in enhancing photorealism and low-exposure correction show promise in neural network's ability to learn lighting representations
- No work or dataset exists for rasterized to ray-traced image translation

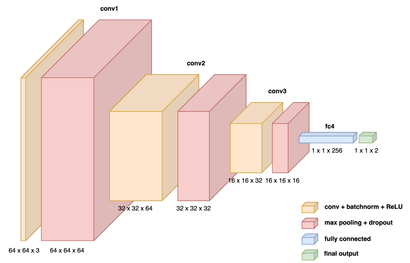
Subproblems:

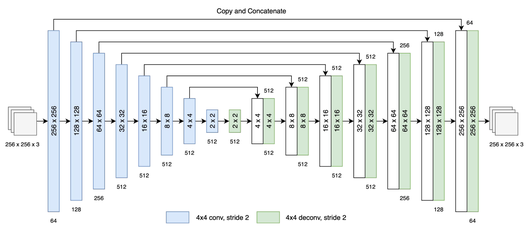
1) Train a convolutional neural network (CNN) to classify images as either ray-traced or rasterized
2) Generate ray-traced images based on rasterized images by training a pix2pix-based model, a type of conditional generative adversarial network (cGAN)

Dataset

- 117 mugs, train-val-test split of 82-18-17
- 100 viewpoints per mug, each with:
  - 1 rasterized image
  - 1 ray-traced image
  - 1 depth map
  - 1 normal map
- Self-created using Blender
  - Rasterized with Eevee engine
  - Ray-traced with Cycles engine
- Rendering times: 0.21s per Eevee, 1.68s per Cycles, 0.04s per depth map, 0.04s per normal map

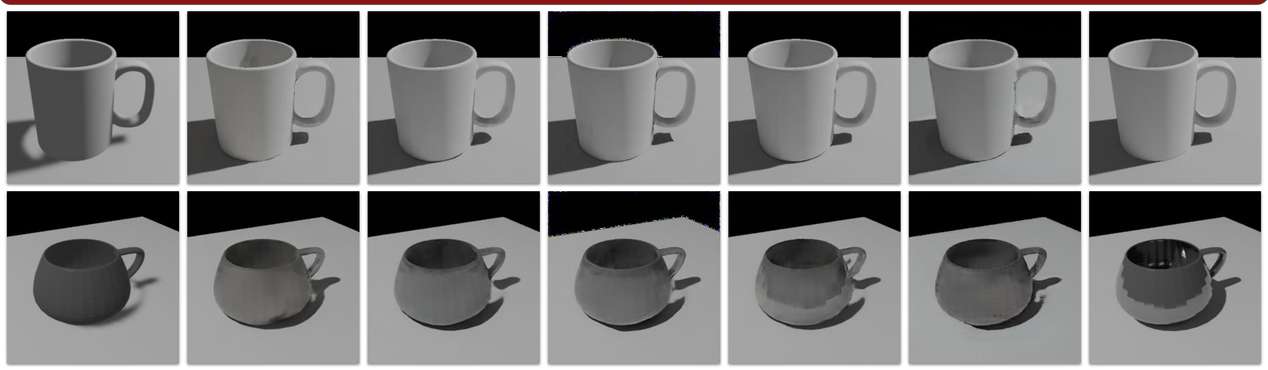
Models





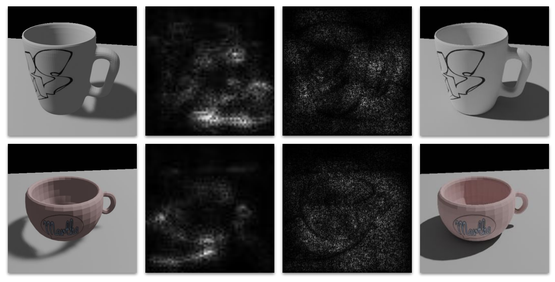
1) CNN Classifier: Classifies images as ray-traced or rasterized; 93.97% accuracy
2) pix2pix:
   a. Base Model: Vanilla pix2pix
   b. Depth Model: Input rasterized image stacked channel-wise with depth map
   c. Depth Model + Skip Connection: Additional skip connection for 3 channels of rasterized image only to output from UNet
   d. Depth + Normal Model: Input rasterized image stacked channel-wise with depth map and normal map
   e. Deeper Depth Model: Additional conv layers at each encoder and decoder layer for greater expressibility

Results



Rasterized | Base | Depth | Depth+Skip | Depth+Normal | Deeper Depth | GT Ray-traced

Model	Mean L1	CNN (%)	MSSIM
Ground Truth	0	100.00	1.000
Base	4.168	99.88	0.925
Depth	4.026	99.18	0.928
Depth+Skip	4.823	100.00	0.896
Depth+Normal	3.730	99.65	0.934
Deeper Depth	5.770	100.00	0.919



Rasterized | CNN Saliency | pix2pix Saliency | GT Ray-traced

Discussions

- Depth + Normal model performed the best, indicating that the model uses the extra channel for depth and three channels for normal mapping to better predict lighting
- It was noticeably better in adding reflections and filling-in/sharpening shadows
- Saliency map of the ray-traced/rasterized discriminator (our CNN) emphasizes shadows and the edges of mugs
- Saliency map of the ray-traced/generated discriminator (pix2pix) also focuses on shadows, and additionally on the handles where the lighting will shift the most from the initial rasterized representation
- Average inference time for the pix2pix model is 0.08 seconds
- The combined time to generate the rasterized rendering, depth map, and normal map and pass through cGAN is 0.37s
- This is over 4.5x speedup from the 1.68 seconds needed by Cycles to render a ray-traced version
- No model generated perfect output compared to the ground truth ray-traced version

Conclusion

- The model needs depth and normal maps to accurately predict ray-traced images
- For some mugs, the generated ray-traced image is almost indistinguishable from the ground truth but takes >4.5x less time to create
- For others, the outputs were imperfect, meaning that there are further improvements necessary

Future Work

- Give the model more information (lighting position, materials) to improve reflections
- Add more convolutional layers after extra skip connection
- Change depth differently at different model levels
- Train pix2pix or use DenseDepth to produce depth and normal maps automatically
- The best MSSIM performance on the training set was 0.937, barely better than our best performing model on the test set, indicating a need for a deeper more complex model