Physics-based Feature Extraction and Image Manipulation via Autoencoders

Abstract

- We experiment with the extraction of physics-based features by utilizing synthesized data as ground truth.
- We utilize these extracted features to perform image space manipulations.
- We model our network as a semisupervised adversarial autoencoder, and train our encoder to extract features corresponding to physical properties of the image-described scene.

Motivation

An existing issue with style transfer is that results are often asthetically pleasing yet not very realistic. We think it would be interesting to see if we could work on this problem by enforcing physical constraints on image generation via semi supervised methods.

Motivated by recent work on interesting applications of deep learning to image synthesis, we

explore a hybrid technique between completely data-based methods and physics-based generative models, by training a joint encoder-decoder network that performs

extraction of graphical appearance on the encoder end, and learns

an feature-based 2D render engine on the decoder end.

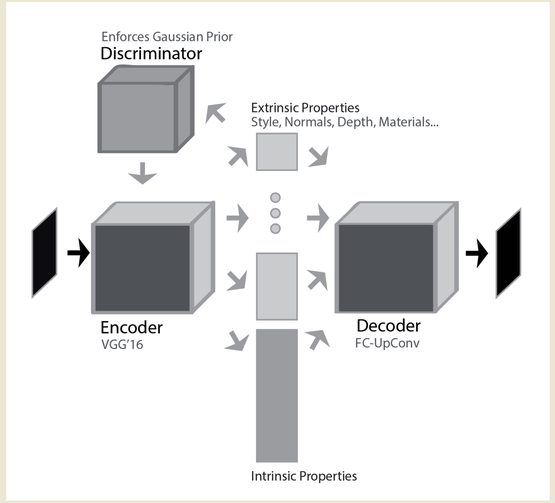
Our Approach

We use a similar framework to [3], where we train a encoder on some number of extrinsic features such as depth, surface normals, texture, and lighting, as well as some variational amount of hidden intrinsic parameters,

and train a decoder to act as a 2D image space renderer, which attempts to output the original image given the feature vector as generated by the encoder.

Based on the positive results reported by [1] and [9] for image generation from feature vector modifications, we use a pretrained VGG19 network as our first layer for the encoder network, and combine that with existing architecture from [3] for our initial results.

Architecture



The architecture can be split into three general components, the encoders, the decoder, and the discriminators.

We train separate encoders via semi-supervised methods, where first for the intrinsic feature vector, we append the ground truth features we obtained to the encoder output to pass to the decoder.

For each extrinsic feature encoder, we remove the ground truth feature corresponding to it and train an encoder for it, still appending the other ground truth features and the trained intrinsic vector. This is a variation on [3]'s methods.

For each feature encoder, we utilize an adversarial network discriminator during training time to enforce a gaussian prior on the encoder outputs.

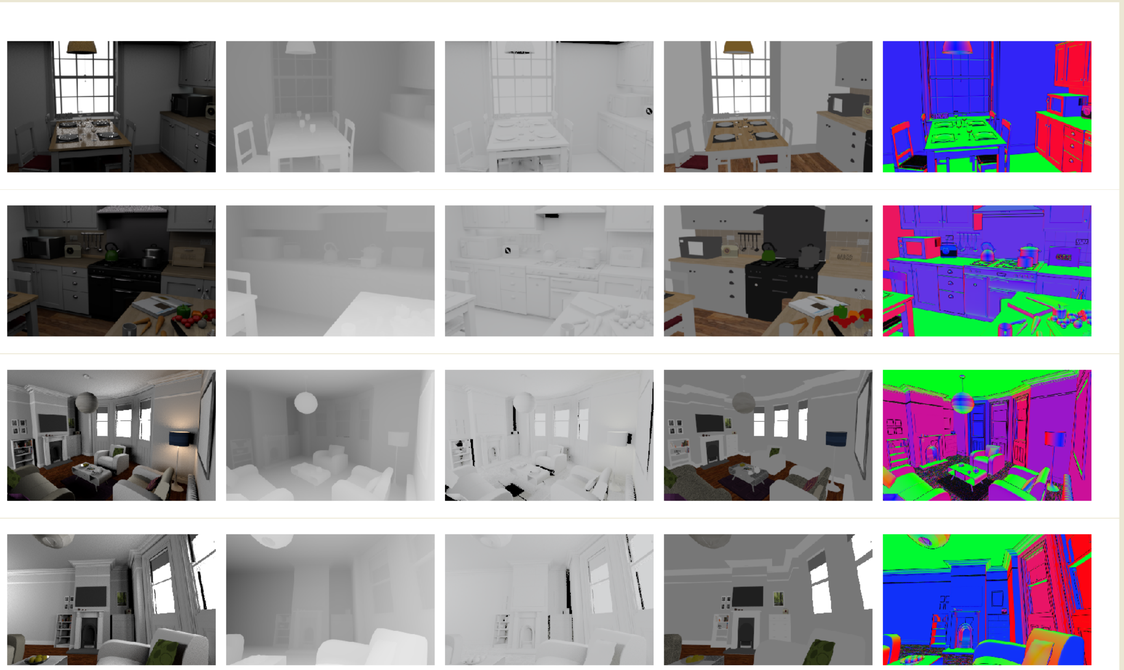
For our decoder, we swap out our first fully connected layer each time we replace the ground truth features with the feature encodings.

Data

The main challenge of this project is the data collection; as described above, we need ground truth that is not easily measurable in the real world. Few existing datasets go beyond RGBD, so we ended up having to spend a significant amount of time on data collection.

We synthesize our own data by heavily augmenting pbrt3, a state of the art research-oriented renderer, to output qualities such as material approximations, surface normals, depth, lighting etc. A visualization of two scenes from different angles are as shown below.

From left to right: original image, depth map, light intensity map, material approximation, and normal map.



Implementation

We used pytorch to implement our models, and augmented existing C++ code to synthesize our data.

Results

Still running due to complications with the dataset,

will paste a page on when poster gets printed...

Future Work

1. How does the number of intrinsics in the encoder output affect accuracy?
2. Can we effective perform image relighting or recoloring via the manipulation of the decoder's input?
3. What types of images are hard to work with? Can we capture complex phenomena such as reflection and subsurface scattering?

References

- D. Eigen and R. Fergus.
  Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture.
  In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.
- P. Khungurn, D. Schroeder, S. Zhao, K. Bala, and S. Marschner.
  Matching real fabrics with micro-appearance models.
  ACM Transactions on Graphics (TOG), 35(1):1, 2015.
- T. D. Kulkarni, W. Whitney, P. Kohli, and J. B. Tenenbaum.
  Deep convolutional inverse graphics network.
  CoRR, abs/1503.03167, 2015.
- M. M. Loper and M. J. Black.
  Opendr: An approximate differentiable renderer.
  In *European Conference on Computer Vision*, pages 154–169. Springer, 2014.
- A. Mahendran and A. Vedaldi.
  Understanding deep image representations by inverting them.
  In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5188–5196, 2015.
- A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey.
  Adversarial autoencoders.
  arXiv preprint arXiv:1511.05644, 2015.
- R. Ng, R. Ramamoorthi, and P. Hanrahan.
  All-frequency shadows using non-linear wavelet lighting approximation.
  In *ACM Transactions on Graphics (TOG)*, volume 22, pages 376–381. ACM, 2003.
- P. Ren, Y. Dong, S. Lin, X. Tong, and B. Guo.
  Image based relighting using neural networks.
  ACM Transactions on Graphics (TOG), 34(4):111, 2015.
- P. Upchurch, J. Gardner, K. Bala, R. Pless, N. Snavely, and K. Weinberger.
  Deep feature interpolation for image content changes.
  arXiv preprint arXiv:1611.05507, 2016.

Winnie Lin
CS 231N 2017: Final Project
Stanford University