Factoring Shape, Pose, and Layout from the 2D Image of a 3D Scene (Supplementary Material)

Shubham Tulsiani, Saurabh Gupta, David Fouhey, Alexei A. Efros, Jitendra Malik
University of California, Berkeley
{shubhtuls, sgupta, dfouhey, efros, malik}@eecs.berkeley.edu

A1. Network Details

All the convolution layers in the networks below, except when stated otherwise, have a kernel of size 3. padding of 1, and a stride of 1 (except otherwise stated). Similarly, all the 2D or 3D up-convolution layers have a kernel of size 4, padding 2, and a stride of 2 *i.e*. they double the spatial resolution of their input. All weights, except the ones explicitly mentioned as being initialized using a pretrained ResNet-18, are initialized randomly.

Layout Module. We use a skip-connected network similar to [1]. The input to this network is an image of of height 128, and width 256 pixels. It outputs an inverse disparity map of half the input resolution. We use 6 convolutional blocks, each with 2 convolution layers with the second one with a stride of 2. Then, we use 5 upconvolutional layers, each followed by a convolution, to produce the output. After each upconvolution, we also append the features from the decoder via skip layers.

Coarse Image Encoder. The coarse image encoder comprises of the first 4 blocks of a ResNet-18 model (initialized using a pretrained CNN), followed by 2 fully connected layers of 300 units each.

Fine Image Encoder. The fine image encoder comprises of the first 3 blocks of a ResNet-18 model (initialized using a pretrained CNN), followed by an RoI-pooling layer which crops features corresponding to the input box, followed by 2 fully connected layers of 300 units each.

Bounding Box Encoder. We use 3 fully-connected layers, with 50 units each to encode the input normalized bounding box location.

Per Bounding Box Features. After concatenating the features from the coarse image, cropped features from the fina image, and the features from the bounding box location, we use 2 fully-connected layers, with 300 units each, to compute the features used for the final prediction.

Shape Decoder. Our shape decoder starts from a bottleneck of size 20, and then, using 5 un-convolutional layers, each followed by a convolution, decodes to the voxel grid of spatial dimension 32.

A2. Additional Visualizations

We provide additional visualizations for 100 random validation images, comparing our predicted factored representation to depth or voxel based alternates, in Figure 1 and Figure 2.

References

[1] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 1

Figure 1: A visualization the various inferred representations from a single input image for 50 random validation images. For each input image shown on the left, we show the various inferred representations from two views each: a) camera view (left), and b) a novel view (right). The representations shown, from left to right, are a) Voxels, b) Depth, c) Factored (ours).

Figure 2: A visualization the various inferred representations from a single input image for 50 random validation images. For each input image shown on the left, we show the various inferred representations from two views each: a) camera view (left), and b) a novel view (right). The representations shown, from left to right, are a) Voxels, b) Depth, c) Factored (ours).