Ray-ONet: Efficient 3D Reconstruction From A Single RGB Image (Supplementary Material)

Wenjing Bian
wenjing@robots.ox.ac.uk
Zirui Wang
ryan@robots.ox.ac.uk
Kejie Li
kejie.li@eng.ox.ac.uk
Victor Adrian Prisacariu
victor@robots.ox.ac.uk

Active Vision Lab
University of Oxford
Oxford, UK

A Implementation Details

The following sections include a detailed description of our model architecture, data preprocessing steps and evaluation procedure.

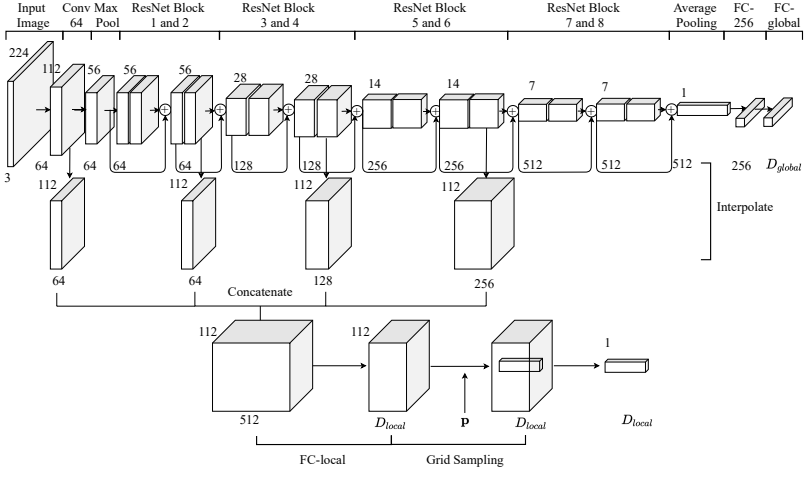

Figure 1: **Model Architecture for Encoder $\mathcal{H}$.** With a single image as input, the global latent code $\mathbf{z}$ is generated from a ResNet-18[5]. The local image feature is extracted from the position of 2D location $\mathbf{p}$ on the concatenated feature maps from 4 different encoder stages. Symbol $\oplus$ represents additive operation.

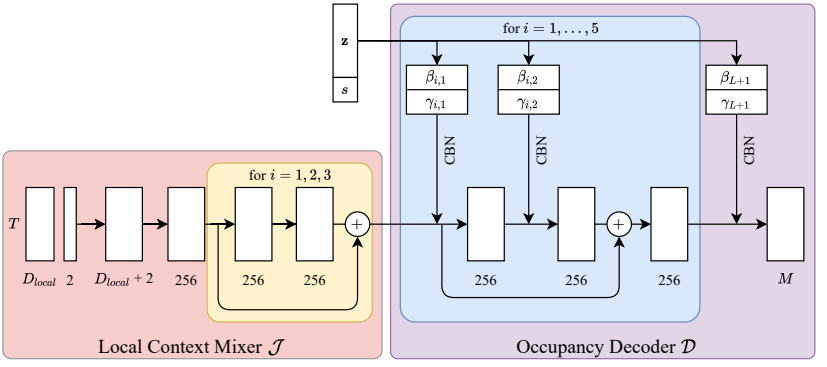

Figure 2: **Model Architecture for Local Context Mixer $\mathcal{J}$ and Occupancy Decoder $\mathcal{D}$.** Using the local feature and a 2D location **p** as input, we first generate a fused local context feature through Local Context Mixer $\mathcal{J}$. The Occupancy Decoder $\mathcal{D}$ uses Conditional Batch-Normalization (CBN) to condition the fused local feature on global latent code **z** and the scale calibration factor s, and predicts occupancy probabilities for M points along the ray.

Encoder $\mathcal{H}$. The encoder is built on a ResNet-18 architecture[5] with an additional upsampling step to generate feature maps. The encoder is initialised with pretrained weights on ImageNet dataset[4] except the last fully connected layer. The global feature **z** is obtained from a fully connected layer with D_{global} dimensions. The outputs from the 2^{nd}, 4^{th} and 6^{th} ResNet Blocks, are upsampled to 112×112 using bilinear interpolation and concatenated, togenether with the the 112×112 feature maps output from the 'Conv64' layer, to form 512 dimensional feature maps. The dimension is changed to D_{local} with a fully connected layer. The local feature is then extracted from the corresponding position of 2D point **p** on the image. In practice, we choose $D_{global} = D_{local} = 256$.

Local Context Mixer $\mathcal{J}$. As shown in Fig. 2, the Local Context Mixer $\mathcal{J}$ takes a batch of T local features with D_{local} dimensions and the corresponding 2D points as input. The local features and points are first concatenated and projected to 256 dimensions with a fully-connected layer. It then passes through 3 residual MLPs with ReLU activation before each fully-connected layer. The output local feature has T batches with 256 dimensions.

Occupancy Decoder $\mathcal{D}$. The Occupancy Decoder $\mathcal{D}$ follows the architecture of occupancy network[6], with different inputs and output dimensions. The inputs are the global feature $(\mathbf{z}, s)$ and the local feature output from $\mathcal{J}$. The local feature first passes through 5 pre-activation ResNet-blocks. Each ResNet-block consist of 2 sub-blocks, where each sub-block applies Conditional Batch-Normalization (CBN) to the local feature followed by a ReLU activation function and a fully-connected layer. The output from the ResNet-block is added to the input local feature. After the 5 ResNet-blocks, the output passes through a last CBN layer and ReLU activation and a final fully-connected layer which produces the M dimensional output, representing occupancy probability estimations for M points along the ray.

A.1 Data Preprocessing

We use the image renderings and train/test split of ShapeNet[1] as in 3D-R2N2[3]. Following ONet[6], the training set is subdivided into a training set and a validation set.

In order to generate ground truth occupancies for rays from each view, we first create

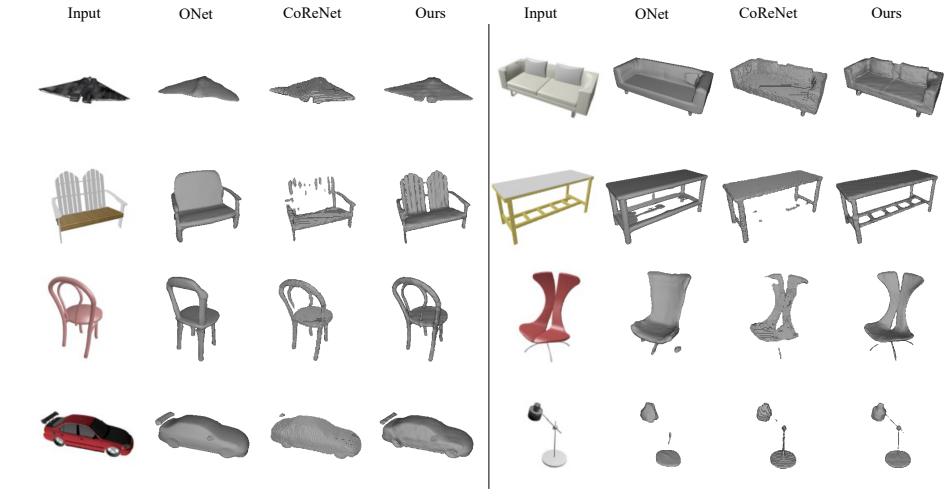

Figure 3: **Additional Qualitative Reconstruction Results on ShapeNet.**

watertight meshes using the code provided by [8]. With camera parameters for 3D-R2N2 image renderings, we place the camera at corresponding location of each view, and generate 5000 random rays passing through the image. We sample equally spaced points on each ray between defined distances d_{min} and d_{max}. In practice, we choose $d_{min} = 0.63$ and $d_{max} = 2.16$, which guarantee all meshes are within this range.

A.2 Evaluation

To make a fair comparison with previous approaches, we use normalised ground truth points and point clouds produced by ONet[6] for evaluation. With the scale calibration factor, our predicted mesh has the same scale as the raw unnormalised ShapeNet mesh. In order to make the predicted mesh in a consistent scale as the normalised ground truth, we use the scale factor between the raw ShapeNet mesh and the normalised ShapeNet mesh to scale our prediction before evaluation. The threshold parameter for converting occupancy probabilities into binary occupancy values is set to 0.2 during inference.

B Additional Experimental Results

B.1 Additional Qualitative Results on ShapeNet

Additional qualitative results on ShapeNet are shown in Fig. 3 and Fig. 4, where Fig. 3 is the standard reconstruction task results on categories seen during training, and Fig. 4 is the test results after trained on 3 different categories.

B.2 Qualitative Results on Online Products and Pix3D Datasets

With the model trained on 13 categories of synthesis ShapeNet[1] images, we made further tests on 2 additional datasets with real world images to validate the generalisation ability of the model.

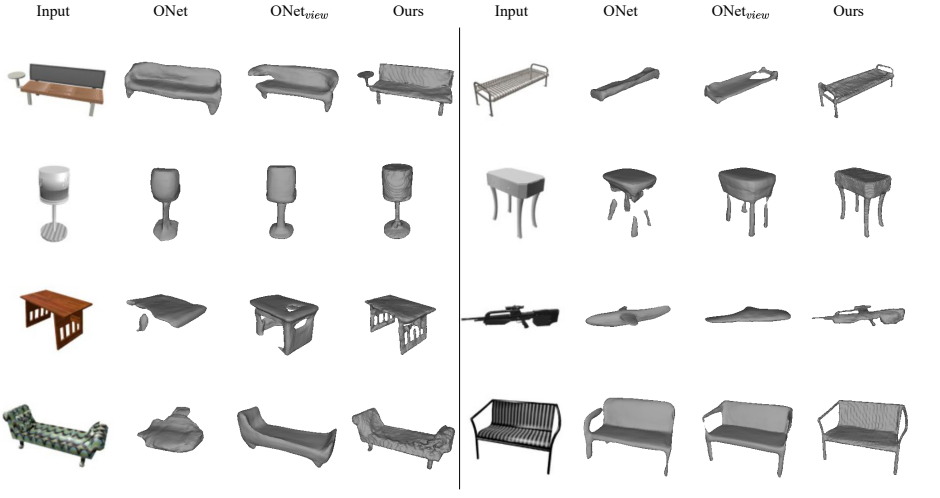

Figure 4: **Additional Qualitative Reconstruction Results on Unseen Categories.**

Online Products Dataset. We use the chair category in Online Products dataset[7]. As the training data has white background, We first feed the image into DeepLabV3[2] to generate a segmentation mask and changed the color of the image outside the mask to white. As there is no camera parameters available, the reconstruction shown in Fig. 5 is of correct proportion only.

Pix3D Dataset. Similarly, we test our model on chair category of Pix3D dataset[9], with the ground truth segmentation mask provided. Some results are shown in Fig. 6.

B.3 Limitations

As shown in Fig. 7, for certain objects in unseen categories, the 3D object reconstructions are much more accurate in the image view than other views, as our model is able to predict shape for visible parts from image features but lacks of shape priors for invisible parts on those unseen categories.

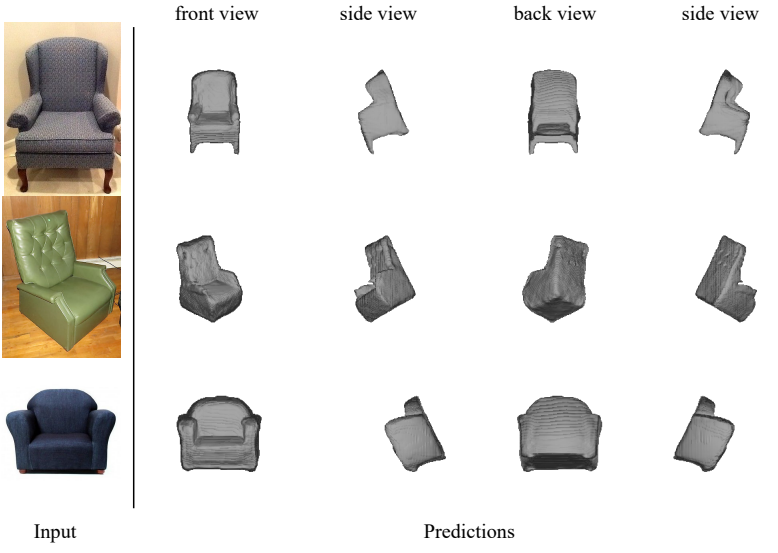

Figure 5: **Qualitative Reconstruction Results on Online Products dataset.**

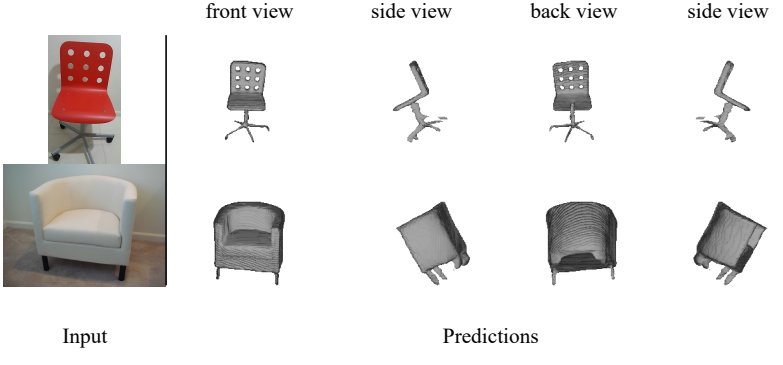

Figure 6: **Qualitative Reconstruction Results on Pix3D dataset.**

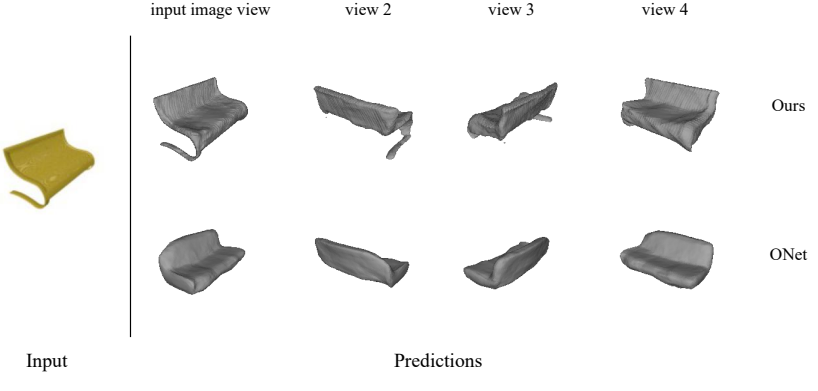

Figure 7: **A Failure Case for Novel Categories.**

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012, 2015.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI, 2017.

[3] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In ECCV, 2016.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[6] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In CVPR, 2019.

[7] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In CVPR, 2016.

[8] David Stutz and Andreas Geiger. Learning 3d shape completion from laser scan data with weak supervision. In CVPR, 2018.

[9] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In CVPR, 2018.