# 3D Position Estimation Using Recursive Neural Networks

*Hanna Winter hannawii@stanford.edu*

## Problem Statement

Today, video feeds from phone or droid cameras are used for augmented reality or 3D reconstruction. Additionally, recent exploration using rendered image data for deep learning has shown we can perform viewpoint estimation [4] in single images. Using these observations, we create a dataset of rendered image sequences to emulate video frames and use a a VGG Net [3] attached to a Recursive Neural Net (RNN) to predict an object's 3D bounding box for a given image sequence. Extending this technique to predict feature points for point cloud creation would be the most useful application.

## Data

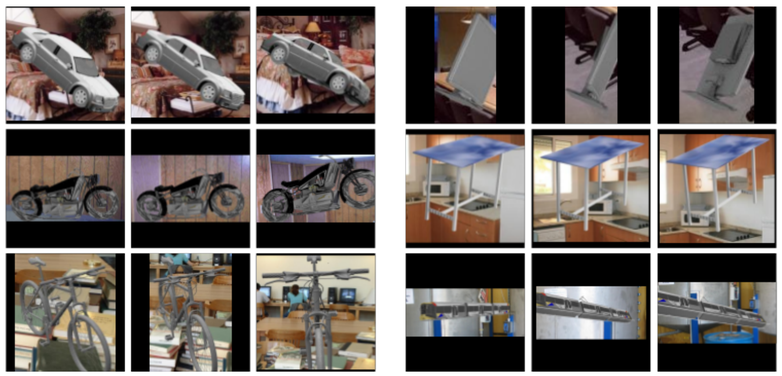
We created a dataset of synthetic images by rendering ShapeNet models [2] from a variety of different viewpoints. We adjusted the rendering pipeline from the Render For CNN paper [4] to meet our specific needs.

**Rendering Pipeline:**

1. ***Render:*** render each image sequence of a 3D model using Blender viewpoints calculated by semi-randomly interpolating the initial viewpoint
2. ***Crop:*** crop the images according to [4] except images within a given sequence are cropped the same with minimal randomization
3. ***Overlay:*** each rendered image sequence are given a randomly sampled from the SUN397 dataset
4. ***Preprocess:*** all images are resized and/or padded to be 224x224 and given a 4$^{th}$ channel of 0 or 1 indicating the 2D bounding box around the object

**Data Set:**

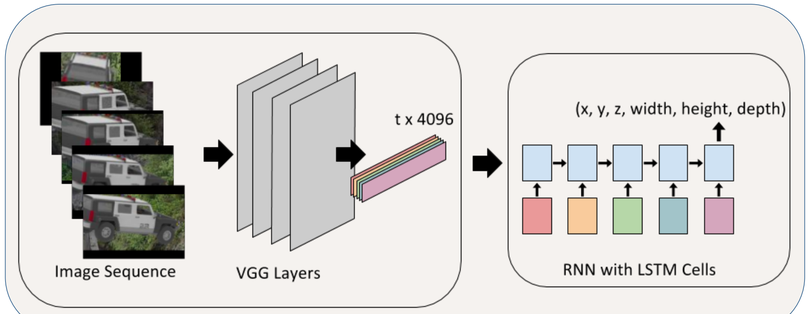
- ***Training examples:*** 50 image sequences of length 20 for each of the 12 classes from PASCAL 3D+ [5] for a total of 530 image sequences (10% removed for validation and testing)
- ***Labels:*** ground truth 3D bounding boxes, (x, y, z, width, height, depth)



Sample of rendered image sequences after resizing and padding. This displays 6 different classes from the 12 total categories of PASCAL 3D+.

## Implementation

We build a deep learning architecture with Tensorflow [1] using an out-of-box VGG19 network and replacing all but the first fully connected layer with an RNN. The image sequences are fed through the VGG as a large batch of images and reshaped back into a sequence before going through the RNN. The output of the network is the 3D bounding box of the object.




**Hyperparameters:** initial network parameters (no hypertuning)

- ***Image Sequence Length:*** 15
- ***Batch Size:*** 5
- ***Learning Rate:*** 0.1
- ***Optimizer:*** Momentum update
- ***RNN hidden size:*** 2048
- ***RNN layers:*** 1

## Future

Extending this method it to perform 3D feature point prediction would be the logical next step. Additionally, the model probably can be improved by training on larger datasets with longer image sequences. This can be achieved by training and testing on more computationally powerful machines. Finally, aspects of the dataset can be improved. The current cropping scheme is too randomized to mimic video data accurately and some of the 3D models do not render correctly in Blender.

## References

[1] ABADI, M., AGARWAL, A., BARHAM, P., BREVDO, E., CHEN, Z., CITRO, C., CORRADO, G. S., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., GOODFELLOW, I., HARP, A., IRVING, G., ISARD, M., JIA, Y., JOZEFOWICZ, R., KAISER, L., KUDLUR, M., LEVENBERG, J., MANE´, D., MONGA, R., MOORE, S., MURRAY, D., OLAH, C., SCHUSTER, M., SHLENS, J., STEINER, B., SUTSKEVER, I., TALWAR, K., TUCKER, P., VANHOUCKE, V., VASUDEVAN, V., VIE´GAS, F., VINYALS, O., WARDEN, P., WATTENBERG, M., WICKE, M., YU, Y., AND ZHENG, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[2] CHANG, A. X., FUNKHOUSER, T., GUIBAS, L., HANRA- HAN, P., HUANG, Q., LI, Z., SAVARESE, S., SAVVA, M., SONG, S., SU, H., XIAO, J., YI, L., AND YU, F. 2015. ShapeNet: An Information-Rich 3D Model Repository. Tech. Rep. arXiv: 1512.03012 [cs.GR], Stanford University — Prince- ton University — Toyota Technological Institute at Chicago.

[3] SIMONYAN, K., AND ZISSERMAN, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR abs/ 1409.1556*.

[4] SU, H., QI, C. R., LI, Y., AND GUIBAS, L. J. 2015. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *The IEEE International Conference on Computer Vision (ICCV)*.

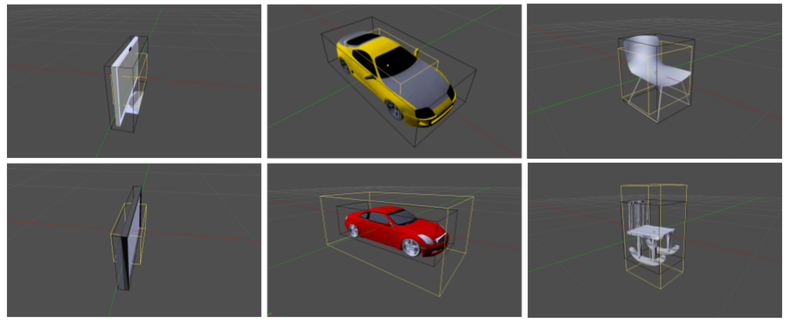
[5] XIANG, Y., MOTTAGHI, R., AND SAVARESE, S. 2014. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

## Results

Results of our VGG+RNN model on our dataset.

| | MSE | IOU | Accuracy |
|---|---|---|---|
| **Train** | 0.0408 | 0.362 | 0.53 |
| **Test** | 0.0483 | 0.275 | 0.41 |

**MSE:** the mean squared difference between the ground truth bounding box and our model's output
**IOU:** intersection volume over the union volume of the 3D bounding boxes
**Accuracy:** we say our model is correct if the IOU > 0.3

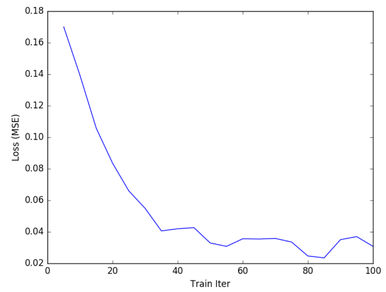


IOU > 0.2 IOU > 0.3 IOU > 0.5

Preliminary test results for our VGG19+RNN model. The ground truth bounding box is show in black around the 3D model. Our model's output is the yellow bounding box. These results are displayed in Blender.

## Discussion

The model was able to produce semi-reasonable results with minimal training of 100 training iterations and batch size of 5. The loss continually decreases over these training iterations showing that with more training and hypertuning, the model should be able to improve. Additionally, training on image sequences longer than 15 should produce even better results.




**Test data:** Currently testing is done on the rendered image sequences. It would produce a more accurate Other 3D datasets with real images are available for testing but do not have the 3D bounding box labels we require. Additionally, there does not seem to exist a 3D dataset with image sequences or video frame data.