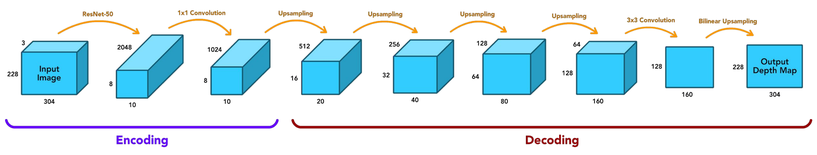
DeepFrame

Jing Lin
Rebecca Agustin
Magnus Johnson

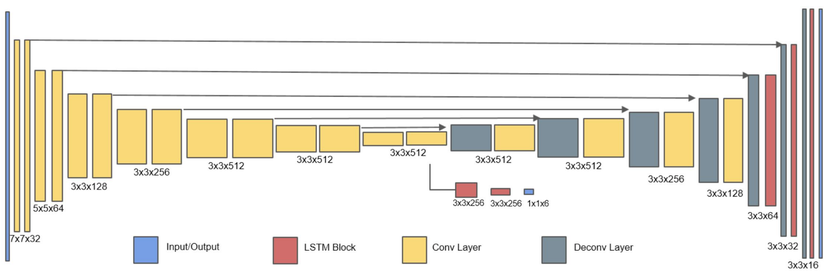
Background

Depth Estimation (predicting depth from 2D images) is a difficult problem with wide range of applications ranging from self-driving cars to robotics. In the field of computer vision, this challenge is being explored through a variety of different techniques.

In one such method, depth estimation can be accomplished through application of convolutional neural networks, such as *DeMoN* below [Benjamin et al. 2017]:



However, more complex approaches, such as the *DenseSLAMNet* presented in [Wang et al. 2018], explore methods for predicting depth from sequential 2D images by utilizing temporal visual information. Towards this end, the network architecture below utilizes both convolutional and LSTM layers:

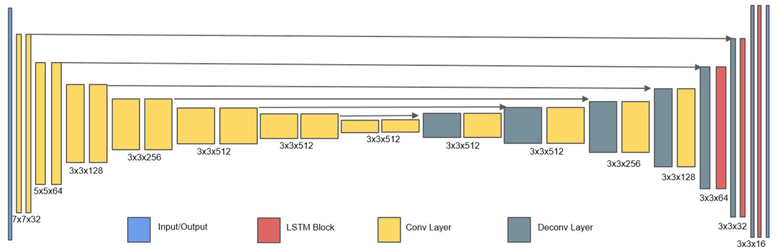


This architecture also includes outputs for predicting camera angles near the most compressed latent layer in the network. This is implemented so that the network can produce more accurate depth predictions based on camera orientation in a multiview context.

Approach

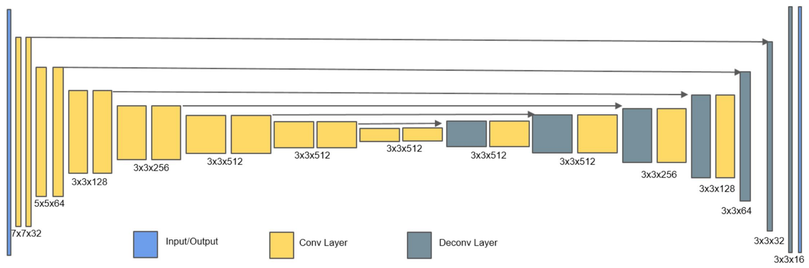
We believe the method proposed by [Wang et al. 2018] can be made more concise without sacrificing functionality.

In our method, we used the following variant of the *DenseSLAMNet* architecture:



...which notably lacks extra outputs and LSTM layers for producing camera angles.

We also test variants of the *CNN-Single* and *CNN-Stack* architectures proposed by [Wang et al. 2018] to compare the results between each. Here is the architecture we used for *CNN-Single*:



Likewise, the architecture we used for *CNN-Stack* is identical to that used for *CNN-Single* except each input is a stacked vector of sequential video frames rather than a single frame.

References

[1] F. Mal and S. Karaman, "Sparse-to-dense: Depth Prediction From Sparse Depth Samples and a Single Image," in IEEE International Conference on Robotics and Automation (ICRA), 2018.

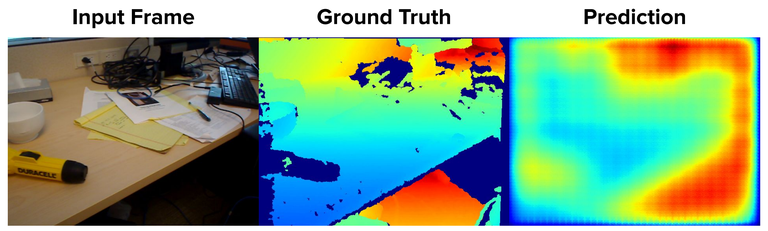
[2] R. Wang, J.-M. Frahm, and S. M. Pizer, "Recurrent Neural Network for Learning Dense Depth and Ego-Motion from Video," arXiv preprint arXiv:1805.06558, 2018.

Results

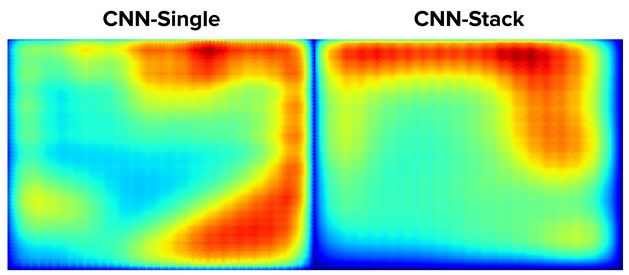
We compare our specialized RCNN architecture with non-recurrent alternatives.

We recreated variants of the *CNN-Single* and *CNN-Stack* architectures from [Wang et al. 2018], and tested their performance on the _ dataset.

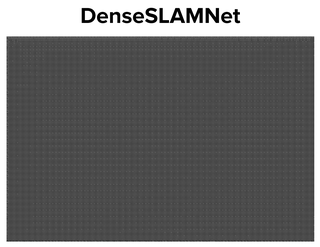
Here are some results from our *CNN-Single* architecture:



Comparing our *CNN-Single* prediction with the results from the *CNN-Stack* architecture, we get the same decrease in accuracy as described in [Wang et al. 2018]



Lastly, we tested the performance of our specialized RCNN *DenseSLAMNet* network. Unfortunately, we did not have enough time & resources to train this model to produce interesting results:



Clearly, given the same hyperparameters and training time, CNN-Single appears to outperform CNN-Stack.