Scene Understanding for Robots using RGB-Depth Information

Aman Raj
amanrajdce@gmail.com
Senior Undergraduate
Delhi Technological University

Carnegie Mellon University

THE ROBOTICS INSTITUTE

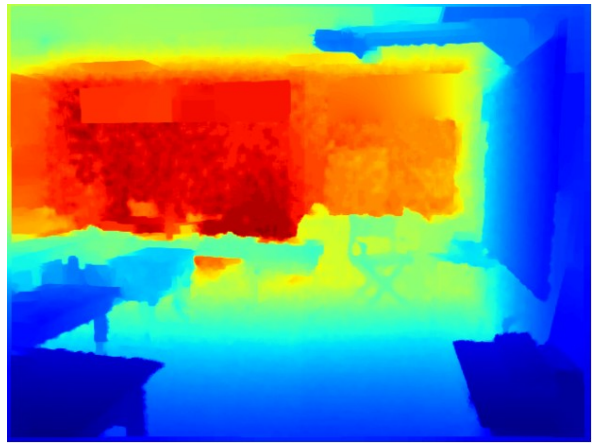
Why do we need it?

- Robots can understand about the rich features of surrounding
- Can help to achieve automation in decision making
- Path Planning and Navigation
- Better assist humans

How can we accomplish it?

- There are many approaches to this problem
- My work focuses on developing deep convolutional net and utilizing RGB-Depth data from camera feed.

RGB



Depth

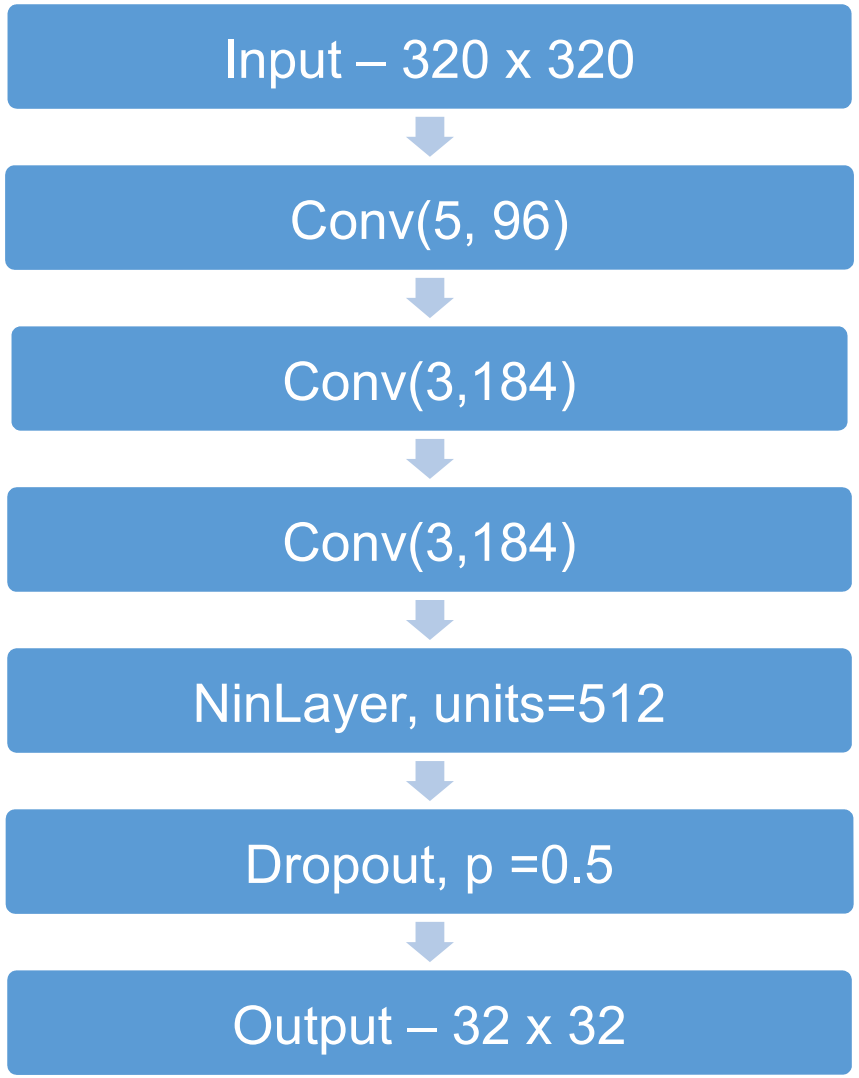
NYU dataset

Labels

Does depth information really help or RGB is enough?

Four layered ConvNet trained:

- RGB
- RGB+Depth



Conv3-184, means a convolutional layer with filter-size(3,3) and filter units = 184

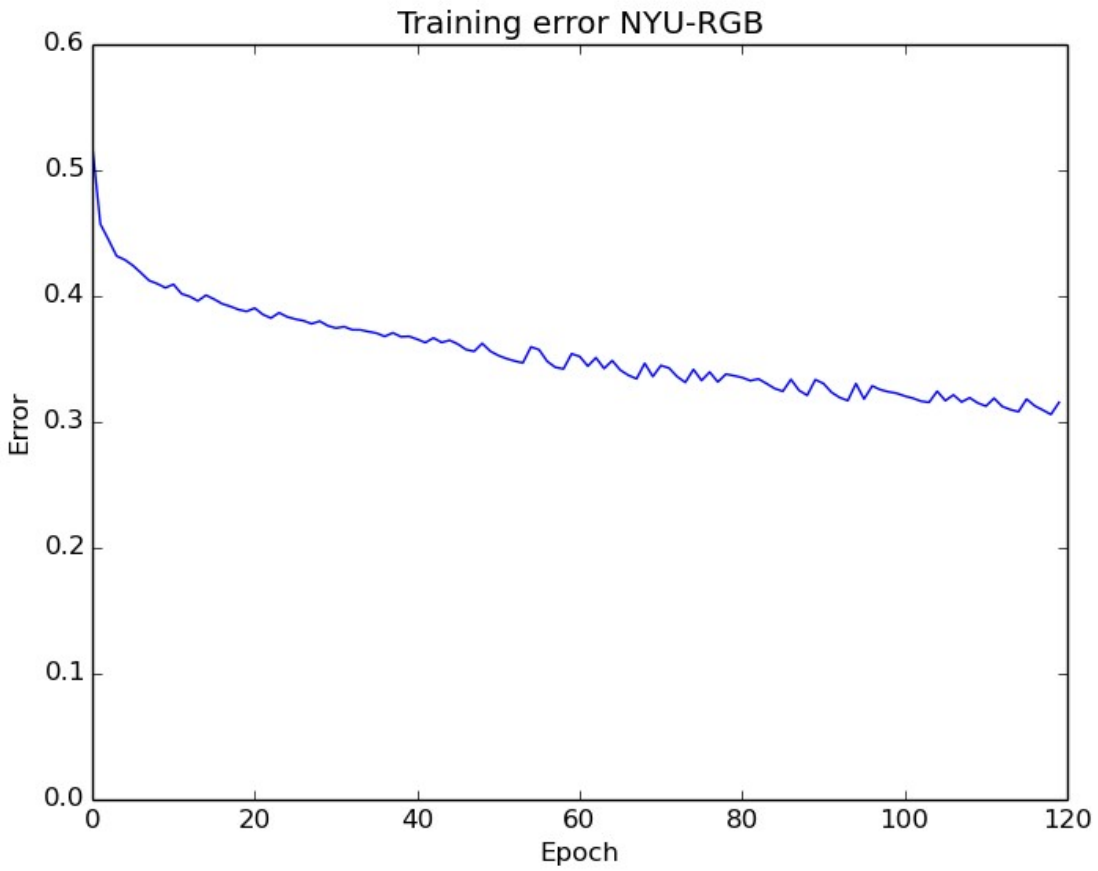
Training - Loss

Loss saturation >1.0

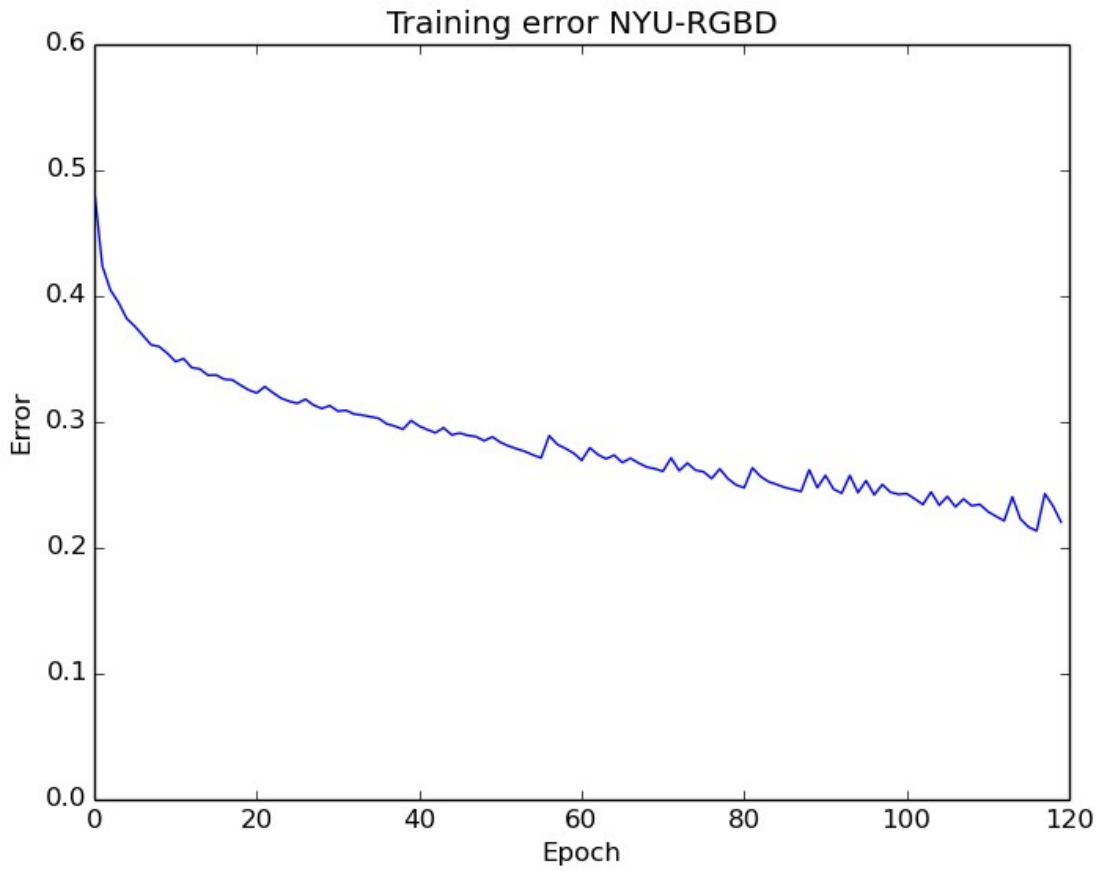
Loss saturation <1.0

Losses are calculated by standard SGD(Stochastic Gradient Descent)

Training - Error



Error saturation >30%

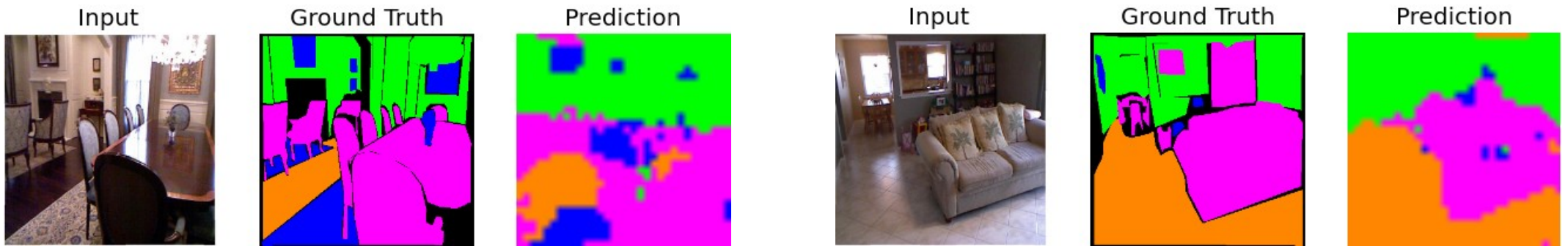


Error saturation <30%

Testing results

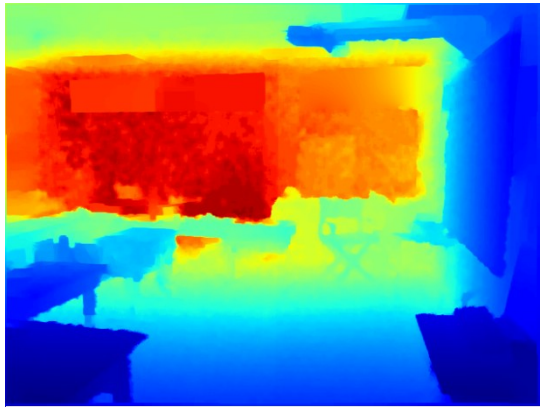
	Avg. Pixel Accuracy	Avg. Pixel Error
RGB	62.29	37.71
RGB-D	67.02	32.98

RGBD – Some predictions on Test set



Is there better way to represent depth?

- Encoding of depth information like RGB
  - H – Horizontal Disparity
  - H – Height above the ground
  - A – Angle Normal makes with gravity



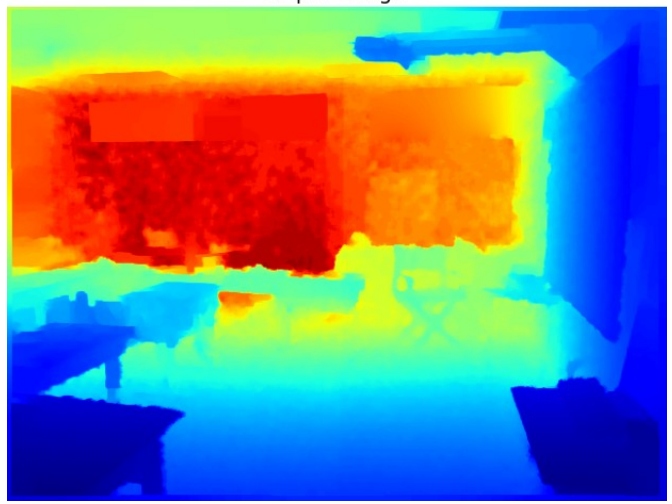
=

?

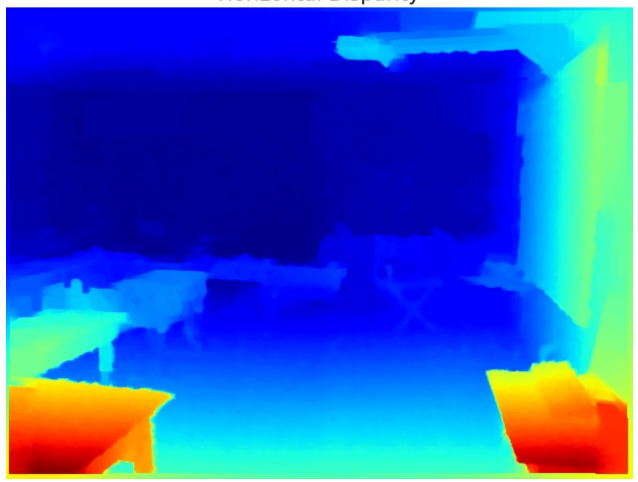
- Experimentation proved use of HHA encoding helps

HHA Encoding of Depth Information

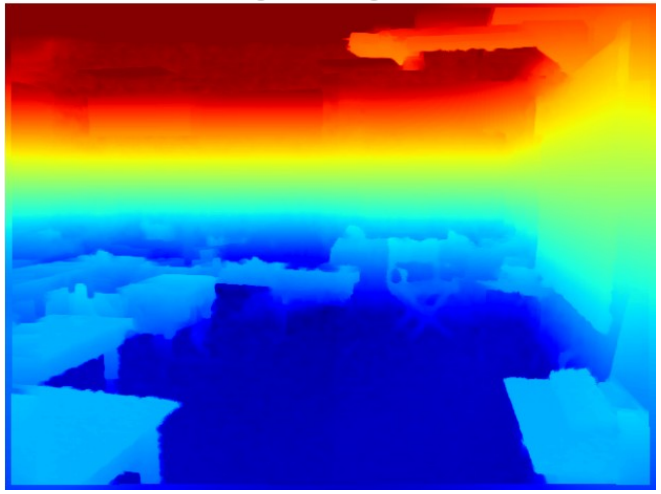
Depth Image



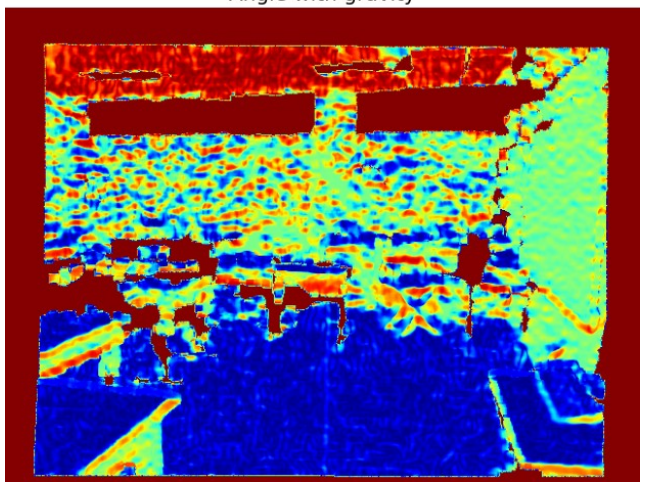
Horizontal Disparity



Height above ground



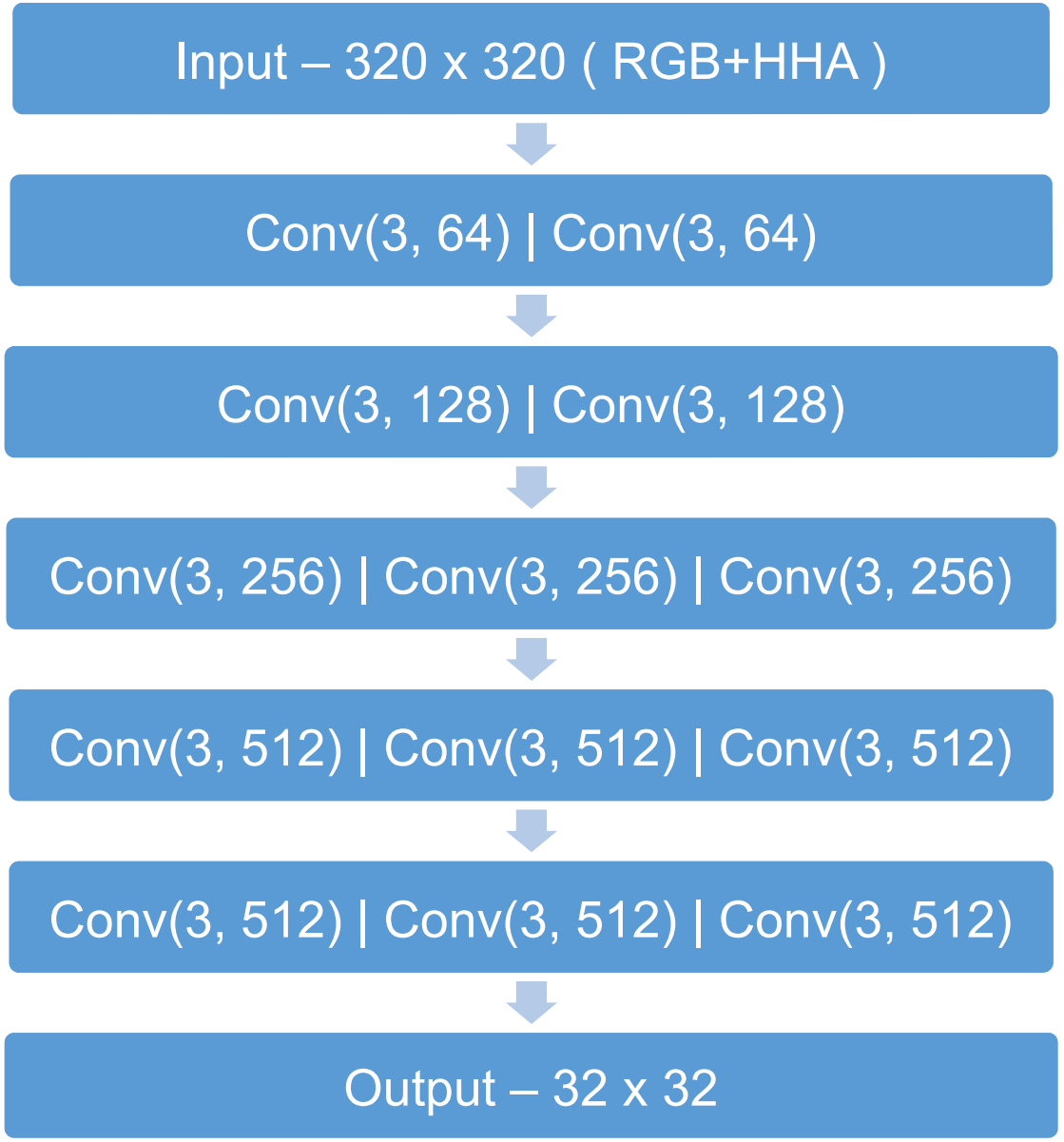
Angle with gravity



S. Gupta, 2014

VGG16-FC ConvNet from Imagenet-2014

- Every stack of convolution layer is followed by a pooling layer(not shown).
- Conv3-64, means a convolutional layer with filter-size(3,3) and filter units = 64
- Fully connected layers from original VGG16 aren't used.

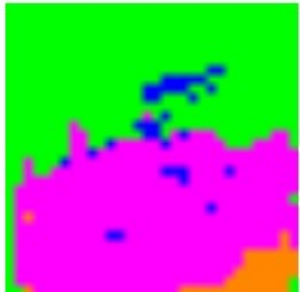


VGG16-FC Results

Input

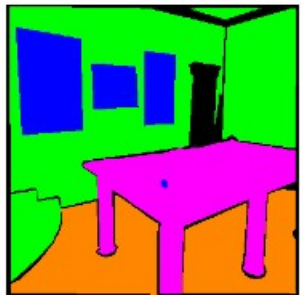
Ground Truth

Prediction



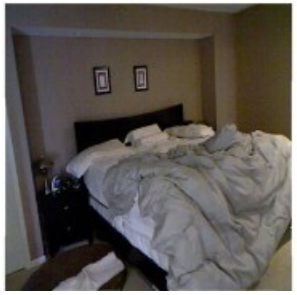
Input

Ground Truth

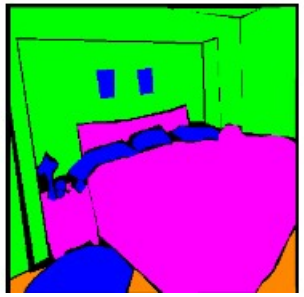


Prediction

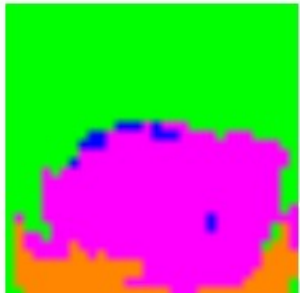
Input



Ground Truth



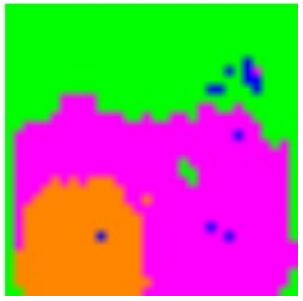
Prediction



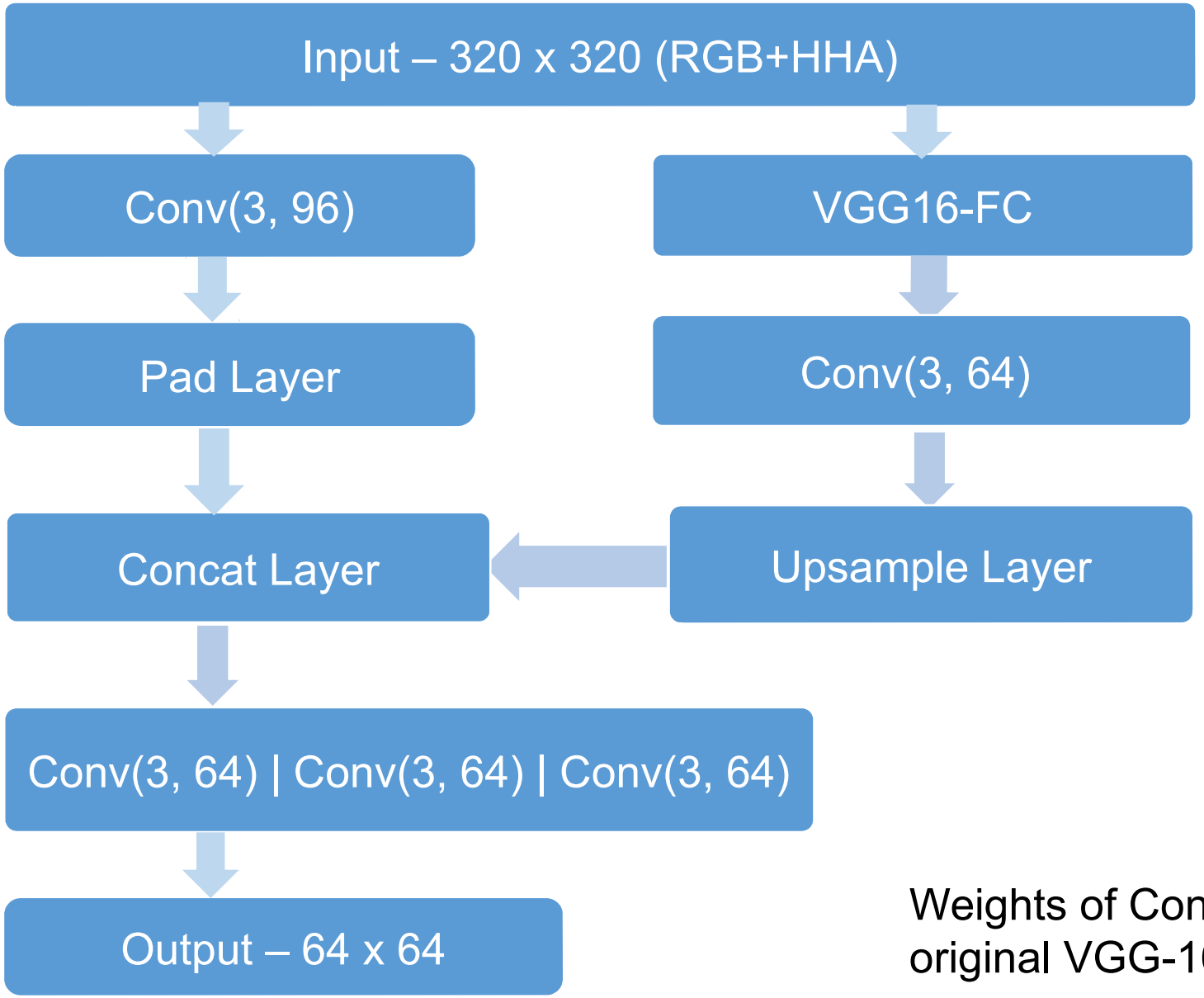
Input

Ground Truth

Prediction

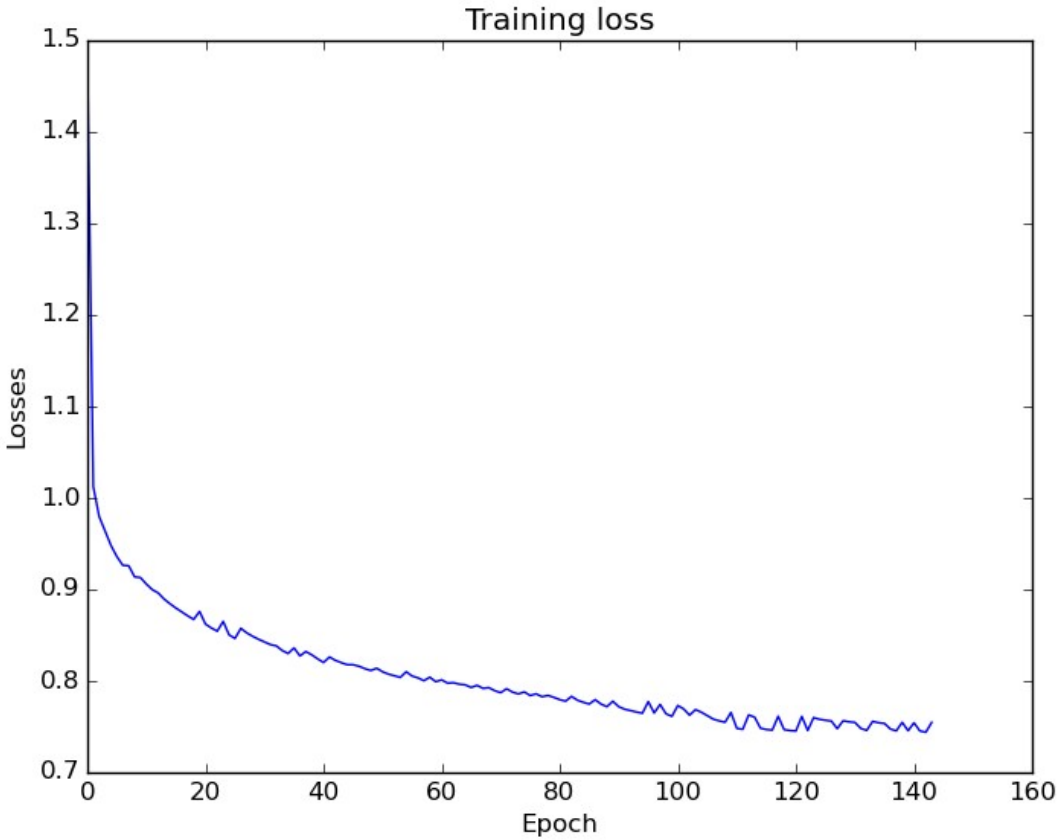


Modified VGG16-FC (VGG-M)

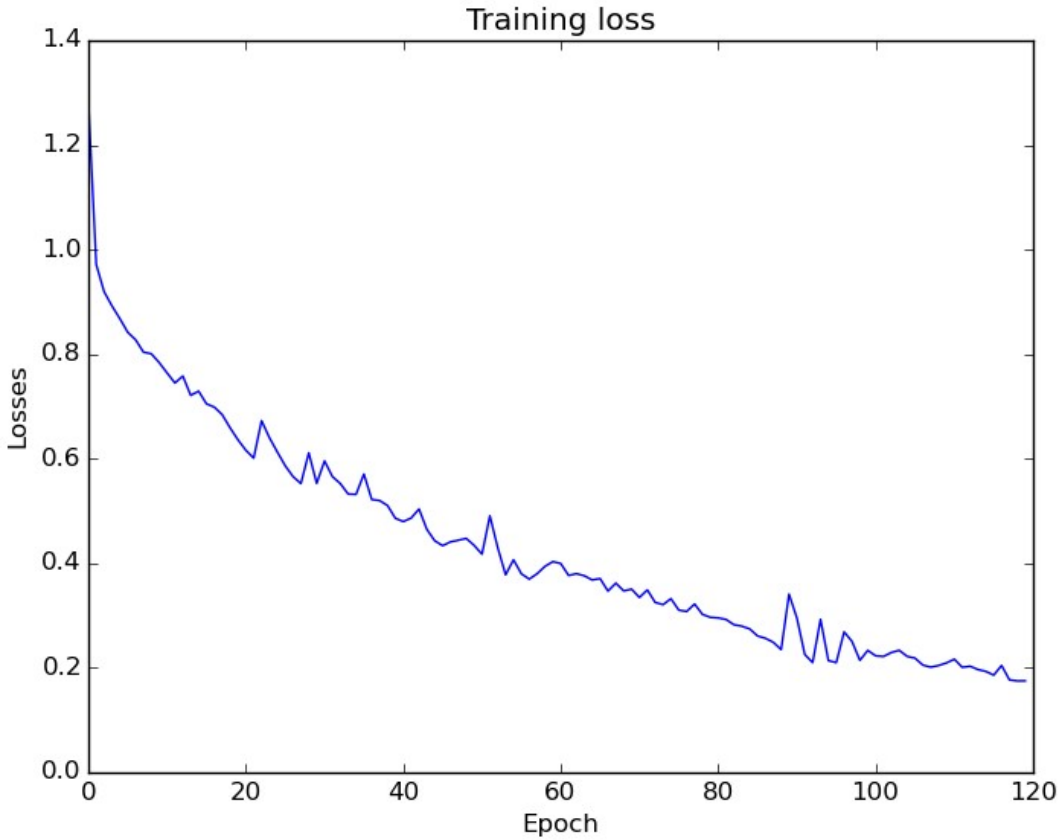


Weights of Conv Layers initialized with original VGG-16 Imagenet weights

How good is VGG-M ?

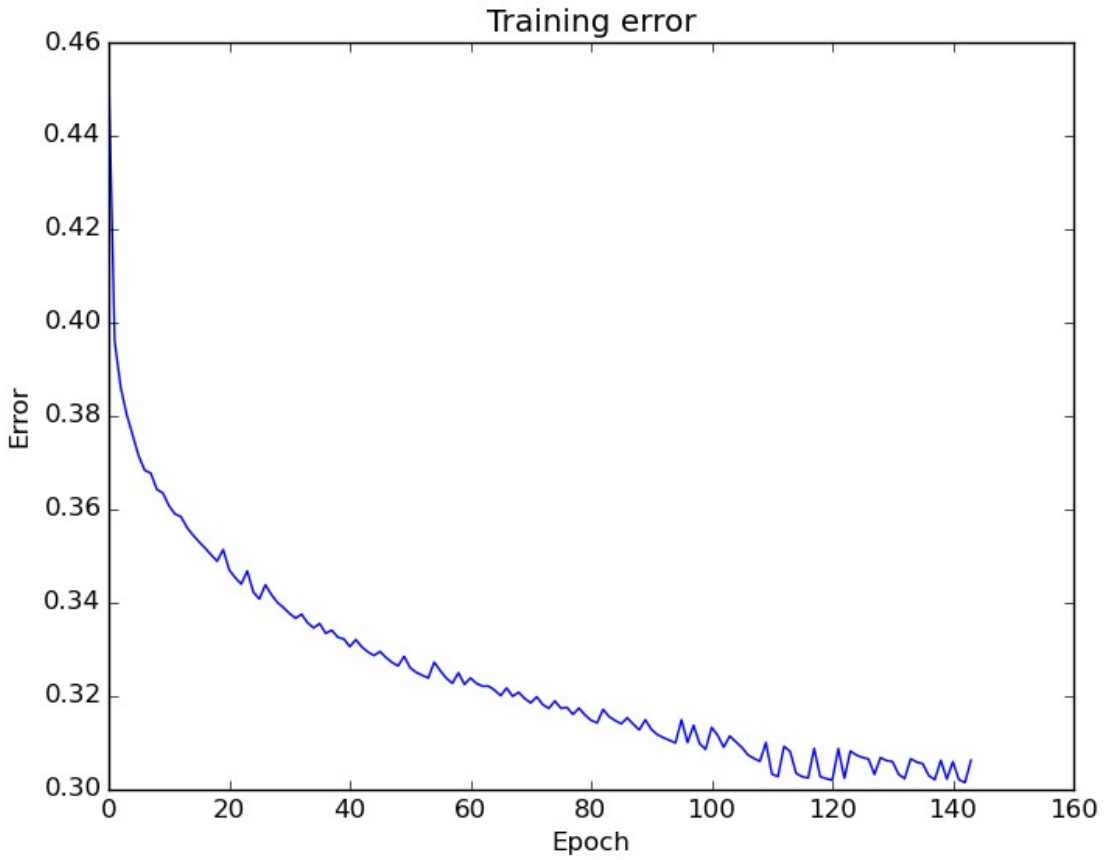


VGG16-FC
Loss Saturates > 0.7

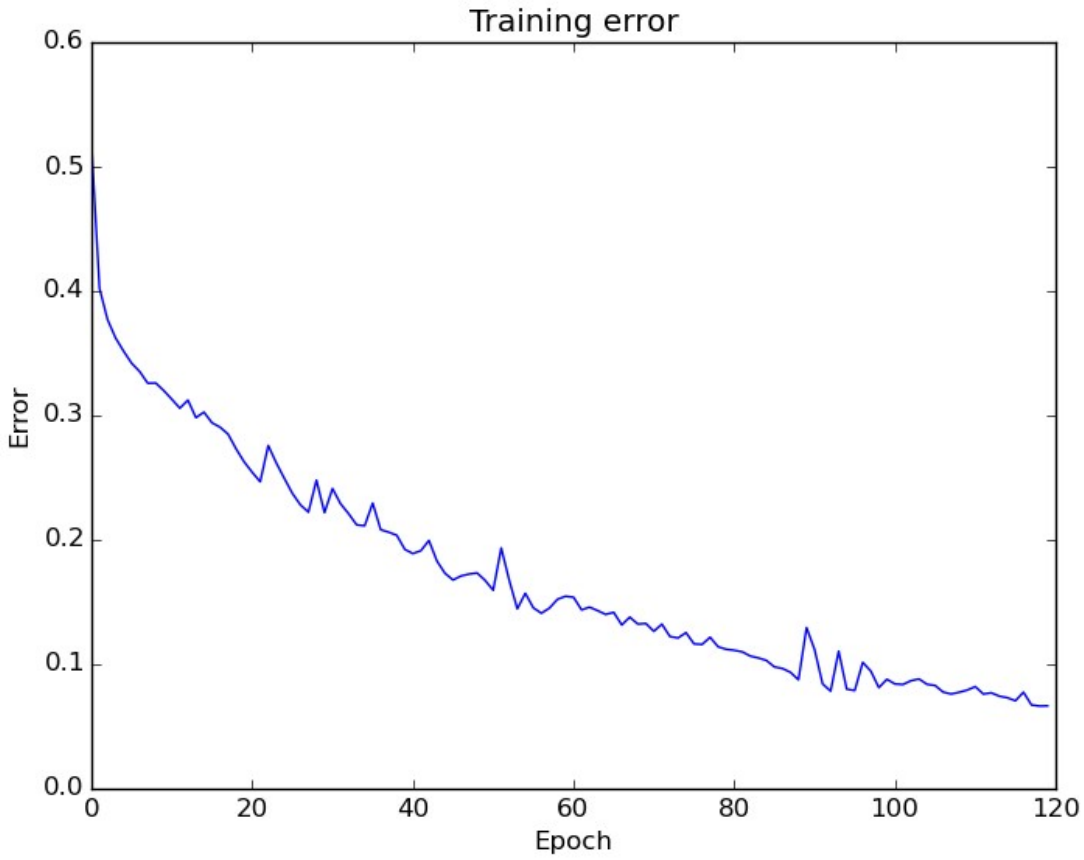


VGG-M
Loss Saturates < 0.2

Training Error



VGG16-FC
Error Saturates > 30%

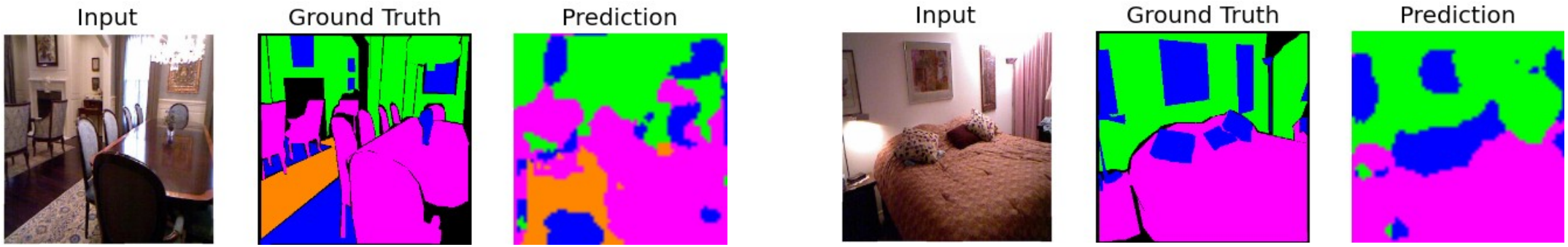


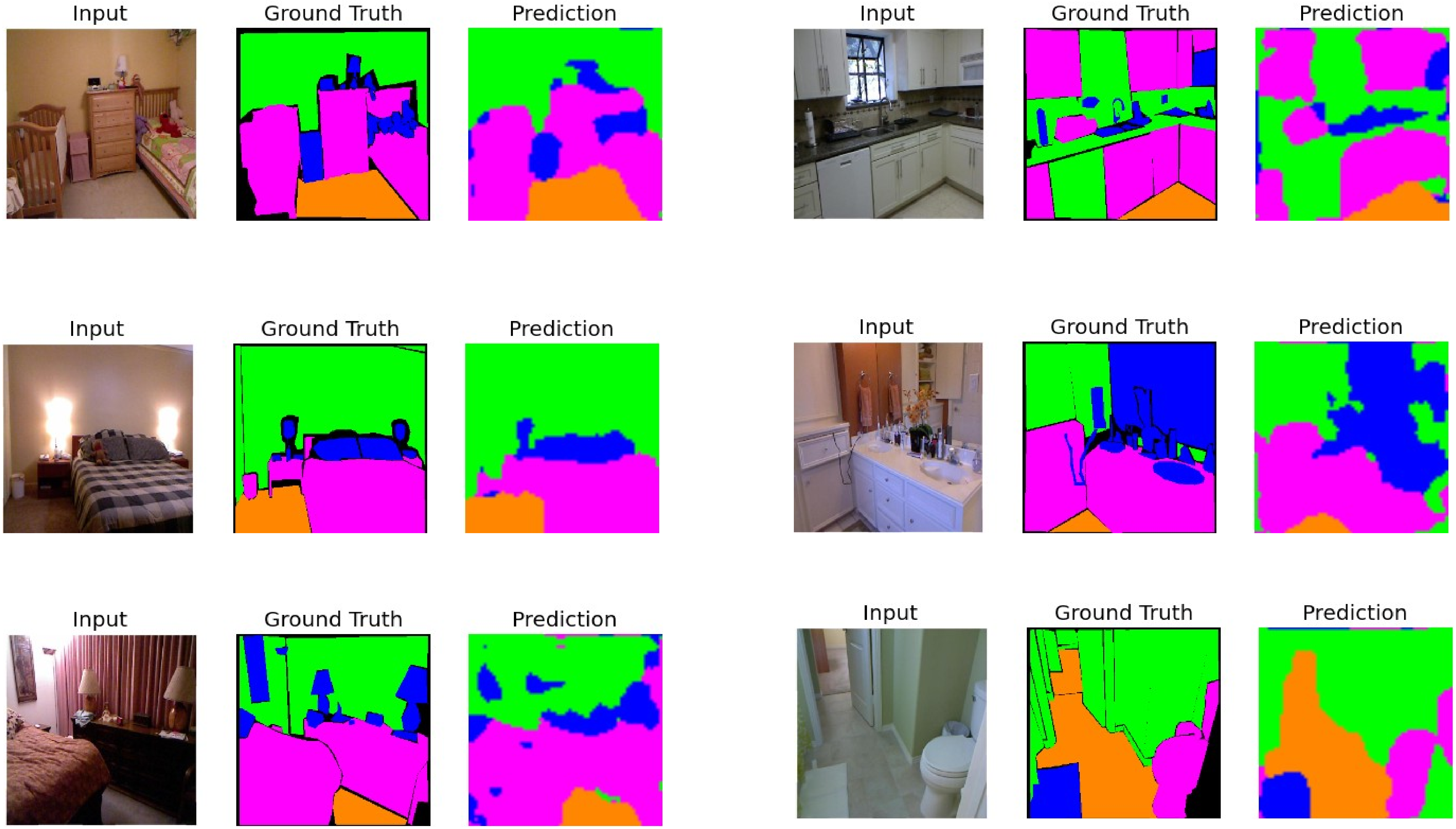
VGG-M
Error Saturates < 10%

Some Results

	Avg. Pixel Accuracy	Avg. Pixel Error
VGG16-FC	66.07	33.93
VGG-M	71.45	29.55

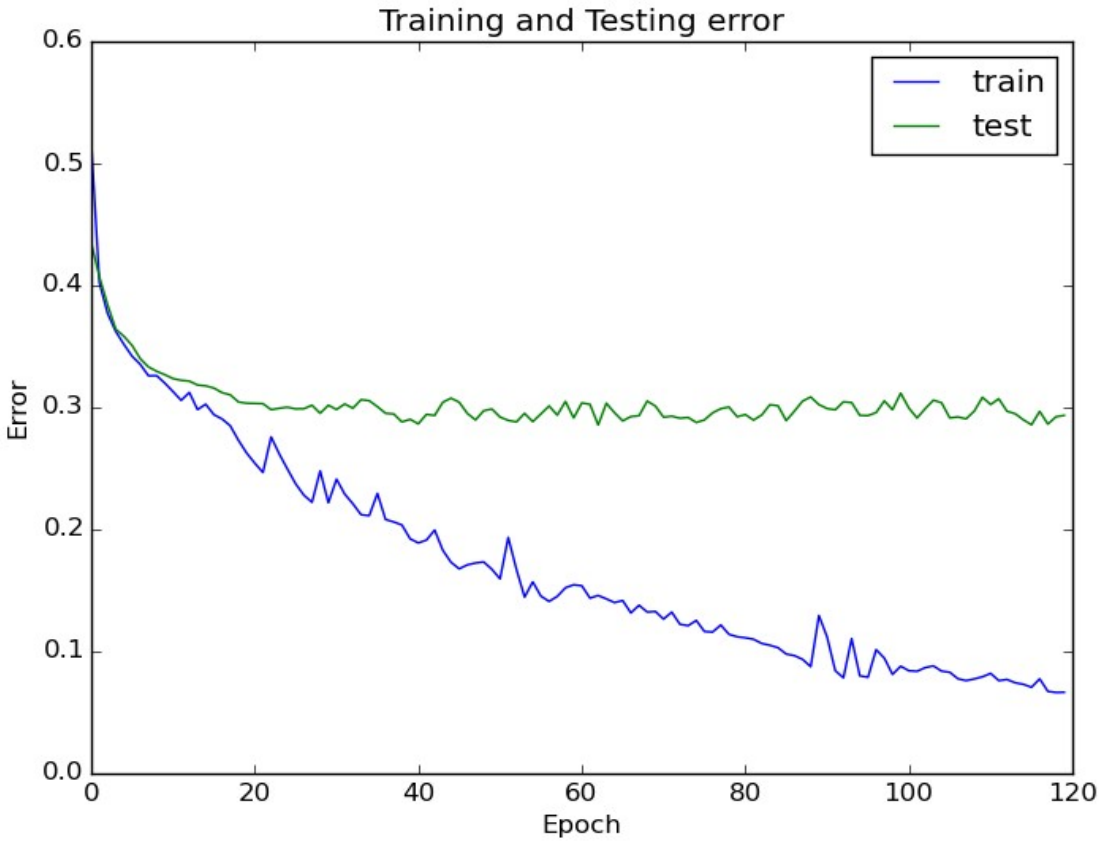
VGG-M Results




Input
Ground Truth
Prediction
Input
Ground Truth
Prediction
Input
Ground Truth
Prediction
Input
Ground Truth
Prediction
Input
Ground Truth
Prediction
Input
Ground Truth
Prediction

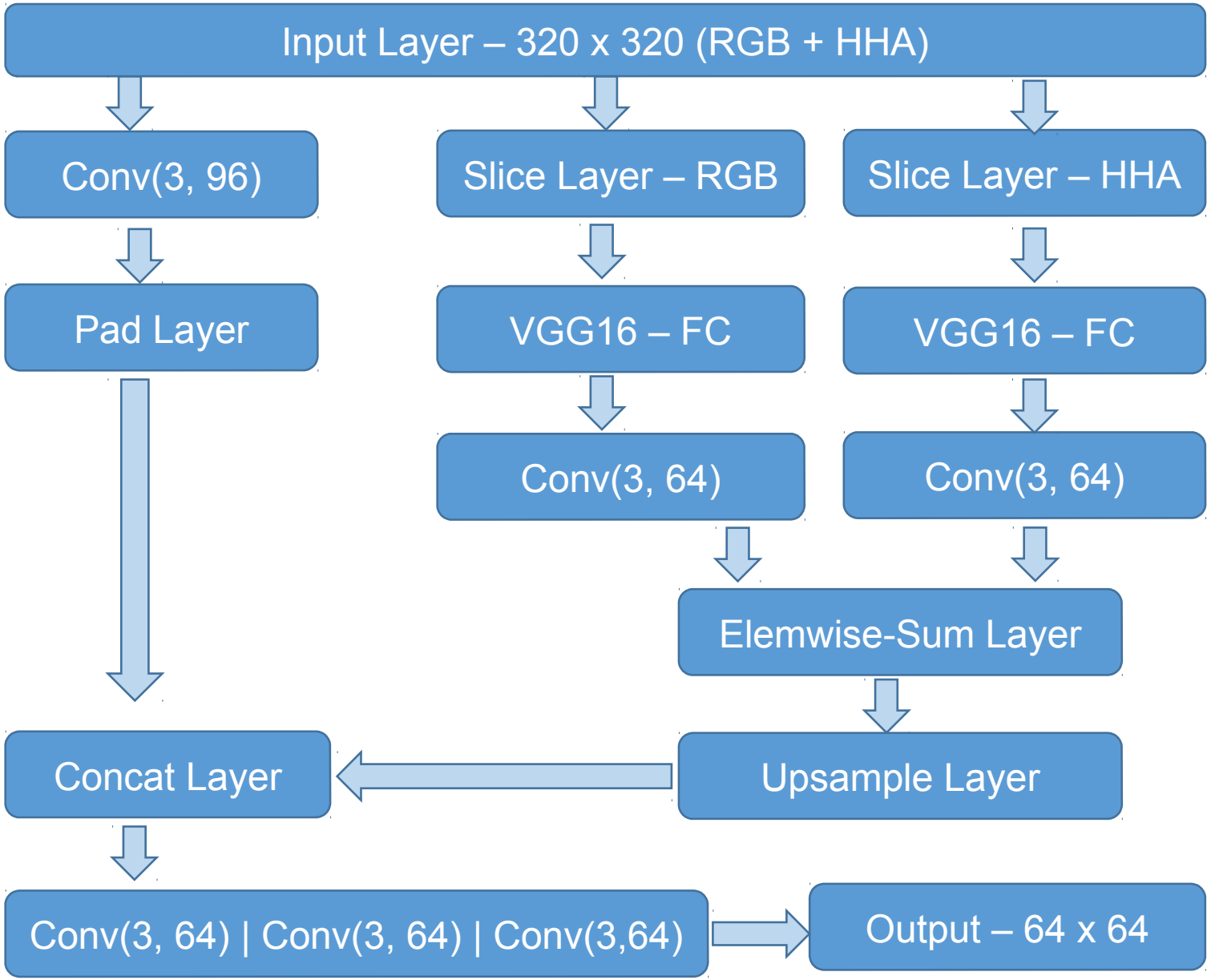
Further Improvement on VGG-M

- Testing error saturates after 30th epoch although the training error keeps going down.
- Possible explanation network isn't learning anything more, it's improving it's confidence score of prediction on training set.



Next Step?

- Train separately on RGB and HHA, combine the coarse output of both network.
- Concatenate the above output with low level prediction
- Refine the predictions with another stack of layers.
- Generate losses using SGD.


Input Layer – 320 x 320 (RGB + HHA)
Conv(3, 96)
Slice Layer – RGB
Slice Layer – HHA
Pad Layer
VGG16 – FC
VGG16 – FC
Conv(3, 64)
Conv(3, 64)
Elemwise-Sum Layer
Concat Layer
Upsample Layer
Conv(3, 64) | Conv(3, 64) | Conv(3,64)
Output – 64 x 64

Conclusion

- For scene understanding depth information with RGB data can greatly help, improving perception of robots.
- Very deep convolutional neural nets modelled for image classification problems can be used and further improved for semantic labelling.
- HHA encoding of depth information brings rich features.