# Single RGB Image Depth and Certainty Estimation via Deep Network and Dropout

Yuanfang Wang, Julian Gao, Yinghao Xu
Graduate Department of Computer Science
Stanford University

Stanford | ENGINEERING
Computer Science

## Project Description

**Goal:**
- Estimate pixel depth from single RGB image
- Generate the certainty of depth estimation at the same time

**Areas of Impact:**
- Machine Learning — dropout usage in testing
- Computer vision — from 2D to 3D using deep neural network

## Bayesian DepthNet

As the approximate inference in Bayesian neural network, dropout can be used as a way of getting samples from the posterior distribution of models [1]. We performed probabilistic inference over the multi-scale depth estimation neural network model [2][5].

- Find the posterior distribution over the convolutional weights, **W**, given our observed training data **X** and depth **Y**. $p(\mathbf{W} \mid \mathbf{X}, \mathbf{Y})$
- Learn the distribution over weights, $q(\mathbf{W})$, by minimizing the Kullback-Leibler (KL) divergence between this approximating distribution and the full posterior; $KL(q(\mathbf{W}) \,||\, p(\mathbf{W} \mid \mathbf{X}, \mathbf{Y}))$
- Approximate variational distribution $q(\mathbf{W}i)$ for every $K \times K$ dimensional convolutional layer $i$, with units $j$ , is defined as: $\mathbf{b}_{ij} \sim$ Bernoulli($p_i$) for $j = 1, ..., K_i$, $\mathbf{W}_i = \mathbf{M}_i \text{diag}(b_i)$,

## Pipeline

### Input Data

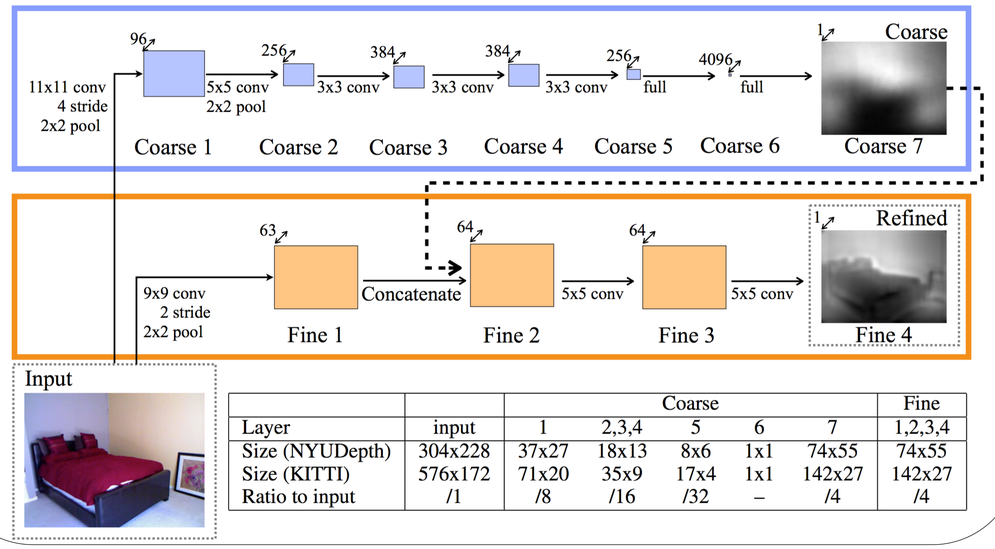
- KITTI dataset [3] (outdoor scenes)
- NYU-D2 dataset [4] (indoor scenes)

### Dropout Experiment

Probabilistic Variant
- **Bayesian Encoder**: dropout after each encoder unit ('Coarse 1' to 'Coarse 5')
- **Bayesian Center Encoder**: dropout after the last encoder unit ('Coarse 5')
- **Bayesian Center-2 Encoder**: dropout after the last 2 encoder unit ('Coarse 4' & 'Coarse 5')

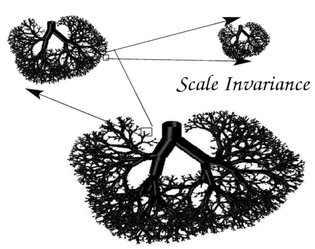
Dropout Rate
- **5%, 10%, 20%, 30%**

### Multi-Scale Deep Network [2]




| | | Coarse | | | | | Fine |
|---|---|---|---|---|---|---|---|
| Layer | input | 1 | 2,3,4 | 5 | 6 | 7 | 1,2,3,4 |
| Size (NYUDepth) | 304x228 | 37x27 | 18x13 | 8x6 | 1x1 | 74x55 | 74x55 |
| Size (KITTI) | 576x172 | 71x20 | 35x9 | 17x4 | 1x1 | 142x27 | 142x27 |
| Ratio to input | /1 | /8 | /16 | /32 | – | /4 | /4 |

## Scale-Invariant




- Finding the average scale of the scene accounts for the large fraction of total error [2]
- Use scale-invariant error to measure relationships between points in the scene
- First define alpha

$$D(y, y^*) = \frac{1}{2n}\sum_{i=1}^{n}(\log y_i - \log y_i^* + \alpha(y, y^*))^2$$

- Use following scale-invariant mean squared error in log space

$$D(y, y^*) = \frac{1}{2n^2}\sum_{i,j}\left((\log y_i - \log y_j) - (\log y_i^* - \log y_j^*)\right)^2$$

## Qualitative Results

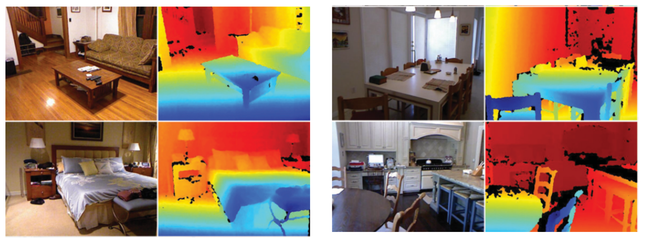
Model = "Bayesian Center Encoder"
Dropout Rate = 5%

- Up: input RGB image
- Medium: depth prediction
- Down: prediction certainty

## Quantitative Results

| Outdoor scenes depth prediction | | | | |
|---|---|---|---|---|
| Probabilistic Variants | 5% dropout | 10% dropout | 20% dropout | 30% dropout |
| Bayesian Encoder | 1.186 | 1.158 | 1.084 | 1.023 |
| Bayesian Center Encoder | 1.225 | 1.225 | 1.227 | 1.226 |
| Bayesian Center-2 Encoder | 1.225 | 1.230 | 1.229 | 1.233 |
| No dropout | 1.214 | | | |

## Next Steps



- **Run on NYUD2 dataset (indoor scenes)**
- **More analysis**

## References

[1] Gal, Yarin, et al. *"Bayesian convolutional neural networks with Bernoulli approximate variational inference."* ICLR 2016.
[2] Eigen, David, et al.*"Depth map prediction from a single image using a multi-scale deep network."* NIPS 2014.
[3] Geiger, Andreas, et al. *"Vision meets robotics: The KITTI dataset."* IJRR 2013
[4] Silberman, Nathan, et al. *"Indoor segmentation and support inference from RGBD images."* ECCV 2012.
[5] Kendall, Alex, et al. *"Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding."* arXiv preprint arXiv:1511.02680 (2015).