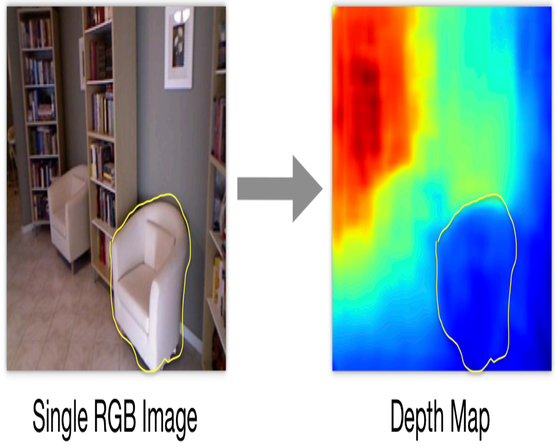
Monocular Depth Refinement with Segmentation Priors and Ordinal Constraints: Towards a Deep Learning Approach

Researcher: Gabriel Birman, Advisor: Jia Deng

Background

- 2D image to 3D object mobile app would require depth estimation
- Existing Depth Estimation using Convolutional Neural Networks (CNNs) do not make use of user input → performance is bad
- Techniques that do make use of user input are relatively complicated → not robust, complicated
- CNNs can effectively learn complicated structure in images by leveraging large datasets



Example of depth estimation with single RGB image. The yellow outline indicates that we want to consider the local depth estimation around the armchair.

Research Question

Is it possible to train a CNN to accurately refine the predicted depth of an existing base network by adding a segmentation channel and user inputted ordinal depth constraint channel?

Methods and Materials

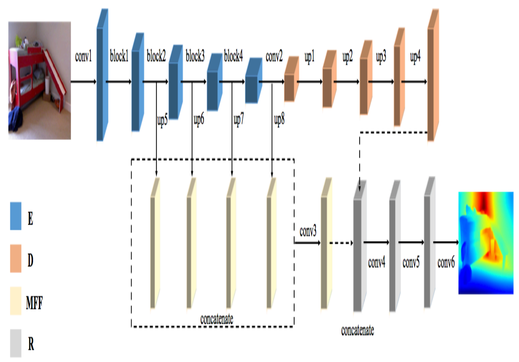
- State-of-the-art Base Network: Hu et. al
- NYUV2 Dataset (14k mask/image pairs)
- Simulate User Clicks:

d_i: predicted map at pixel index i
g_i: ground truth map at pixel index i
P_i: constraint matrix at pixel index i

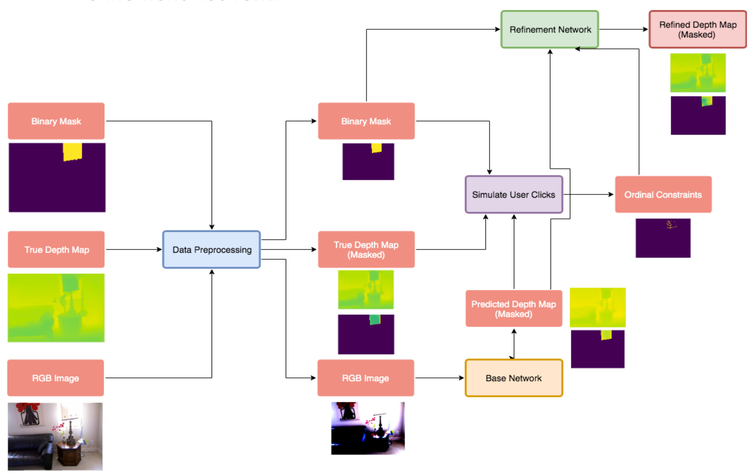
If $\max\left(\frac{d_i}{g_i}, \frac{g_i}{d_i}\right) > \frac{1}{4}$:

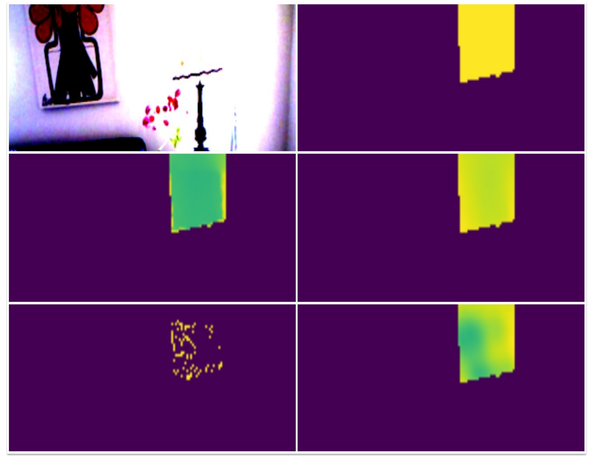
$$P_i = \begin{cases} \operatorname{sgn}\left(\frac{d_i}{g_i} - 1\right) \text{ with prob. } p \\ 0 \text{ otherwise} \end{cases}$$

- CNN trained for 1 epoch (Adam optimizer, learning rate=1E-5) using pre-initialized weights that represented the identity transform



Base network architecture, comprising an encoder, decoder, multi-feature fusion, and refinement module. A slightly modified architecture is used for our depth refinement network.





Pre-processed image (top-left) with lampshade binary mask (top-right). Comparison of ground truth depth (middle-left) with base network output (middle-right). Adding simulated user clicks (bottom-left) leads to refined output (bottom-right).

p	RMS	MAE	REL	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Base Network Output	0.612	0.371	0.133	0.831	0.964	0.991
0	0.650	0.383	0.140	0.819	0.961	0.991
0.05	0.576	0.373	0.148	0.816	0.979	0.995
0.1	0.536	0.354	0.119	0.870	0.983	0.998
0.25	**0.443**	**0.276**	**0.100**	**0.935**	0.993	0.998
0.5	0.511	0.349	0.122	0.845	**0.994**	**0.999**

Comparing the results of the base network with and without our refinement network for various p values, representing the probability of adding a user click. Bolded results show best performing metrics. All metrics are masked.

Results

- Base output improved 27.6% with respect to masked RMS error for p=0.25 (avg. 4 clicks per image)
- *p=0* performing worse than base network indicates that this was not simply due to overfitting the network
- Successive applications of refinement network does *not* improve final depth map (empty constraints, overcompensation)

Conclusion

Conclusion/Discussion:

- The refinement network did a decent job of refining the base depth map, but lacks the capability for iterative refinement, and produces worse performance than state-of-the-art interactive depth estimation
- This has shown is possible to use ordinal constraints to provide guidance for depth refinement without incorporating complicated optimization frameworks

Future Directions:

- Try to implement an iterative process (possibly using an RNN)
- Test on non-NYUV2 datasets
- Finetune hyperparameters/architecture
- Stratify ordinal constraints to provide more guidance

Acknowledgments

Special thanks to my advisor, Jia Deng, TA Hei Law, and the Princeton University COS department.

References

- Hu, J., Ozay, M., Zhang, Y. & Okatani, T. Revisiting Single Image Depth Estimation: Toward Higher Resolution Maps with Accurate Object Boundaries. *CoRR* **abs/1803.08673**, (2018).
- Fergus, D. E. a. P. a., Depth Map Prediction from a Single Image using a Multi-Scale DeepNetwork. *CoRR* **abs/1406.2283**, (2014).
- Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from RGBD images. In *Computer Vision, ECCV 2012 - 12th European Conference on Computer Vision, Proceedings* (PART 5 ed., Vol. 7576 LNCS, pp. 746-760).