Depth Estimation from Single Image Using Convolutional Neural Networks

Xiaobai Ma, Zhenglin Geng, Zhi Bie

Introduction

- Depth estimation using monocular images is important in many practical scenarios such as images on social media and real estate listings.
- Depth estimation using monocular images is challenging because no image correspondences are available are the problem is inherently ambiguous.
- Many depth hints such as perspective and object sized can be exploited from monocular images and CNN is an ideal tool to utilize these information.

Problem Statement

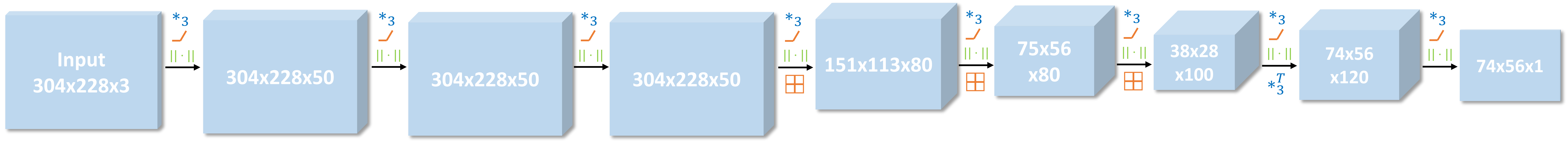
- Input: single RGB image
- Output: depth image
- Evaluation: both in quantitative metric comparison and qualitative visualization

Dataset

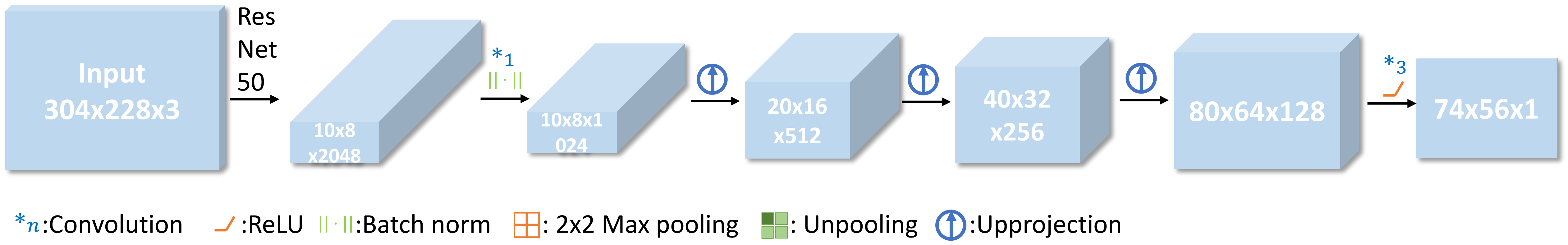
NYU Depth Dataset V2 [1]:
Both RGB and depth images of this dataset are acquired by Microsoft Kinect. We processed 266585 image and depth map pairs from 464 indoor scenes, while the results on this poster are obtained by training on a subset of the living room scene with 298 images.

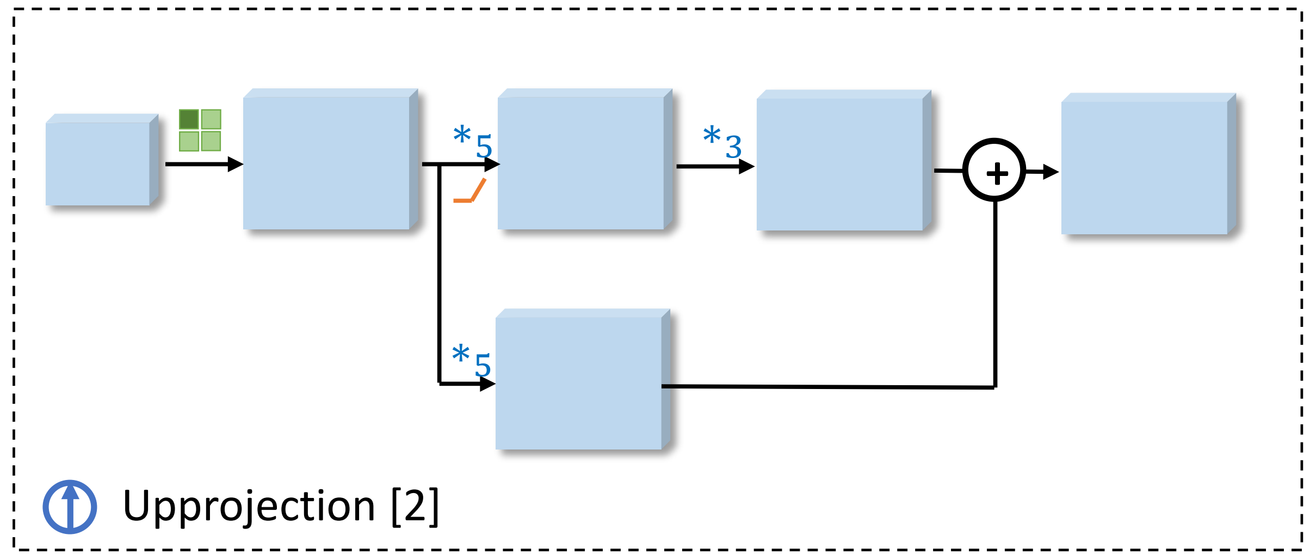
Method

- Fully convolution network:

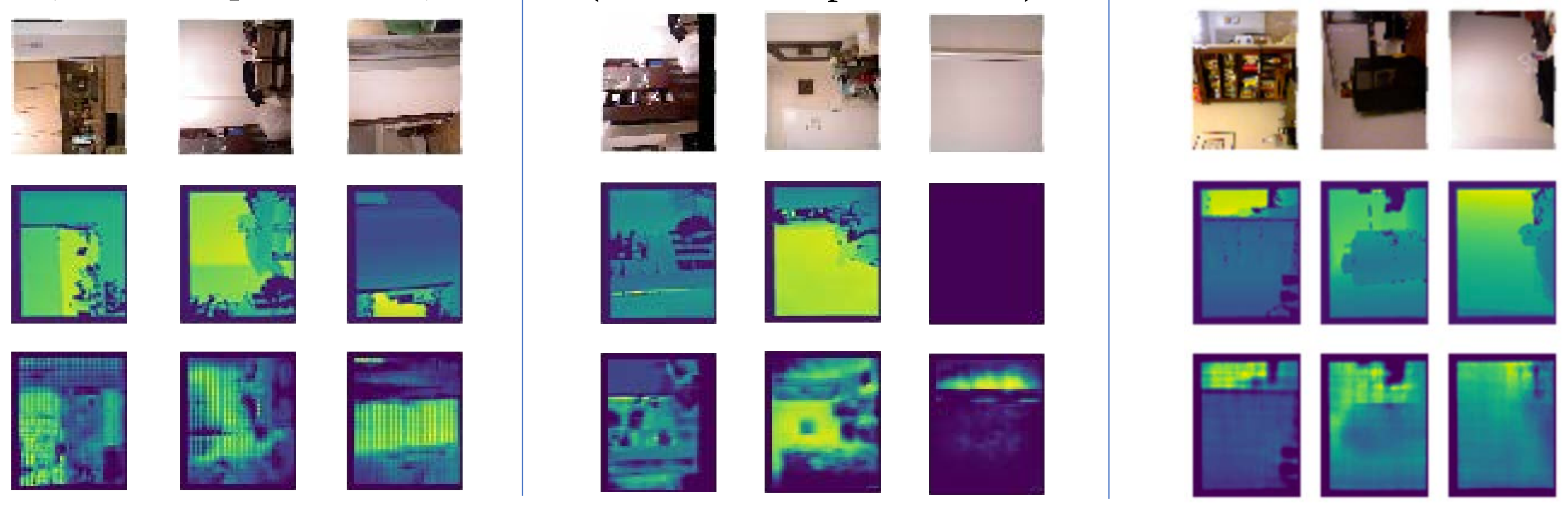


- Convolution + Residual network:





Experimental Evaluations



Loss measurement: scale-invariant mean squared error

$$L(y, y^{*}) = \frac{1}{n}\sum_i d_i^2 - \frac{\lambda}{n^2}\left(\sum_i d_i\right)^2, d_i = \log y_i - \log y_i^{*}, \lambda \in [0,1].$$

Evaluate on validation set:

	Threshold < 1.25	Abs relative difference	RMSE
ConvNet 1	0.924	3.84	44.2
ConvNet 2	0.975	3.66	43.7
Conv+ResNet	0.014	0.717	50.43

Conclusions

In this project, we explored the performance of single image based depth estimation with three network architectures. We see that ConvNets extracts spatial information and gives generally good depth estimation. When combined with features extracted from ResNet, the absolute relative difference is significantly lower. This network yields the best visual effect qualitatively, but also takes longer to train.

References

[1] Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. Computer Vision–ECCV 2012, 746-760.

[2] Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016, October). Deeper depth prediction with fully convolutional residual networks. In 3D Vision (3DV), 2016 Fourth International Conference on (pp. 239-248). IEEE.