HIGH QUALITY MONOCULAR DEPTH ESTIMATION VIA A MULTI-SCALE NETWORK AND A DETAIL-PRESERVING OBJECTIVE

Hualie Jiang[⋆], *Rui Huang*[⋆†*]

[⋆]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen,
Guangdong, 518172, P.R. China
[†]Shenzhen Institute of Artificial Intelligence and Robotics, Shenzhen, Guangdong, 518172, P.R. China

ABSTRACT

Monocular depth estimation is an important and challenging task in computer vision. Significant progress has been made recently due to deep convolutional neural networks. However, esitmating depth maps with high quality lacks sufficient attention. This paper proposes to recover detailed depth map by training a multi-scale network architecture with a detail-preserving loss function. Firstly, we construct our architecture inspired by the design of atrous spatial pyramid pooling for semantic segmentation. Secondly, we simplify the loss on depth map gradients for preserving details. Experiments on the NYU Depth V2 dataset show that our approach is effective and it achieves state-of-the-art performance, especially in the root mean squared error.

Index Terms— Monocular Depth Estimation, Multi-Scale Network, Detail-Preserving Loss

1. INTRODUCTION

Recovering depth information from images is an important problem in computer vision. In particular, **M**onocular **D**epth **E**stimation (MDE) [1] is an ill-posed and very challenging problem, because there are not enough geometrical constraints to recover depth from a single image. Thus, it is not well studied as many traditional techniques for acquiring depth from multiple images, for instance, stereovision [2].

With the fast development of deep learning in recent years, CNN-based methods have largely improved the performance of MDE [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. However, most of the works just count the depth error on pixels individually as loss, which causes that many detailed structures of the depth map cannot be well estimated. Following Hu *et al.* [15], we investigate the problem of estimating detailed depth map from a single image in this paper.

To address this problem, we propose to adopt a multi-scale network based on Atrous Spatial Pyramid Pooling (ASPP) [16, 17, 18]. Specifically, we use DeepLabv3+ [18] instead of DeepLabv3 [17], as DeepLabv3+ employs not only the ASPP module but also a simple decoder structure which increases the resolution of the prediction. In addition, we further modify the decoder structure to achieve higher resolution as a depth map contains more detailed structures than the semantic segmentation result. To train the network, we combine a robust loss on depth, BerHu [7] and a loss on depth gradients for preserving details inspired by [15]. We simplify the loss on depth normals by minimizing the $\mathcal{L}_1$ of the cross product instead of maximizing the cosine [15]. Our simplication can avoid division operation on gradient components without losing its effectiveness.

Experiments on the NYU Depth V2 dataset show that our modification of the network and our simplication of the normal loss works well and our approach can estimate high quality depth map and achieves state-of-the-art performance, especially in the root mean squared error. Our code is available on github[1].

2. RELATED WORK

Make3D [1] is a seminal work on MDE and Saxena *et al.* used a hierarchical, multi-scale Markov Random Field (MRF) on multi-scale image features to predict depth on the dataset that they collected by a 3D laser scanner. The performance has not been greatly improved until Eigen *et al.* [5] proposed the first CNN based solution. Their architecture consists of a coarse-scale network and a fine-scale network for producing coarse and fine prediction respectively. They later improved the architecture to a three-scale network [6]. Afterwards, Laina *et al.* [7] combined Fully Convolutional Networks (FCN) [19] with ResNet [20] to construct Fully Convolutional Residual Networks (FCRN) and used a robust loss, BerHu to increase the performance further.

Some works combined CNN and graphical models [4, 12, 13]. For example, Liu *et al.* [4] proposed to learn the unary and pairwise potentials as a Conditional Random Field (CRF) loss for CNN training without using geometric priors. Other

*Corresponding author: Rui Huang (email: ruihuang@cuhk.edu.cn). This work is partially supported by National Key R&D Program of China (Grant No. 2017YFB1303701), and Robotics Discipline Development Fund from Shenzhen Municipal Government (Grant No. 2016-1418).

[1] `https://github.com/HalleyJiang/HQMDE`

works tried to formulate MDE as a dense multi-label classification problem [8, 11, 14] by discretizing depth. Although DORN [14] boosts the performance by using the deep ordinal regression technique, the predicted depth map loses many details. Moreover, it is hard to extend classfication based approaches to high quality MDE as it is difficult introduce detail-preserving terms into the classificaiton loss.

Our work is mainly inspired by [9, 10, 15]. Carvalho *et al.* compared many regression losses, including $\mathcal{L}_1$, $\mathcal{L}_2$ and BerHu, which is actually a combination of $\mathcal{L}_1$ and $\mathcal{L}_2$. But these regression losses on depth are inferior in producing accurate and detailed predication to the losses on depth gradients proposed by Hu *et al.* [15]. However, we find that their loss on depth normals is complex so we simplify it in this paper. Michel *et al.* [10] examined many multi-scale network architectures in the task of monocular depth estimation and found that the ASPP achieved the best performance. Therefore, we perform MDE by the most advanced ASPP based architecture, DeepLabv3+ [18] in this paper.

3. METHODOLOGY

In this section, we illustrate the adopted multi-scale network architecture and the proposed detail-preserving loss function.

3.1. Multi-Scale Convolutional Network

Many kinds of multi-scale convolutional networks have been proposed for the semantic segmentation task, which is in fact a dense prediction problem similar to MDE. In [10], Michel *et al.* has examined many multi-scale network architectures in MDE and found that ASPP [16, 17] performed much better than other techniques, like skip connections in U-Nets [21] and feature pyramid outputs [22]. ASPP was proposed by Chen *et al.* along with DeepLab architectures [16, 17, 18], which are the state of the arts in the field of semantic segmentation. DeepLab architectures adopt atrous convolutions to enlarge the receptive field without reducing the size of features, and spatial pyramid pooling to effectively extract multi-scale features from deep features.

We adopt the most advanced DeepLabv3+ [18] as our baseline architecture in this paper. The DeepLabv3+ adds a simple decoder structure with atrous separable convolution upon DeepLabv3 [17] to boost segmenation performance further. In order to obtain finer depth map, we decode the features at $\frac{1}{4}\times$ resolution into 16 channel features and bilinearly upsmaple them to the input image size, following by a 1×1 convolution to produce 1 channel logits for the prediction. While DeepLabv3+ directly decodes the features at $\frac{1}{4}\times$ resolution into 1 channel logits and bilinearly upsmaple them to the input image size. We denote our modification as baseline+ in this paper. Figure 1 shows the network architecture used in this paper. We use ResNet-50 [20, 23] as backbone to implement the architecture without using atrous separable convolution in the decoder, as it requests much effort in hyperparameter tuning. Experiments indicate the adopted architecture greatly reduce the root mean squared error.

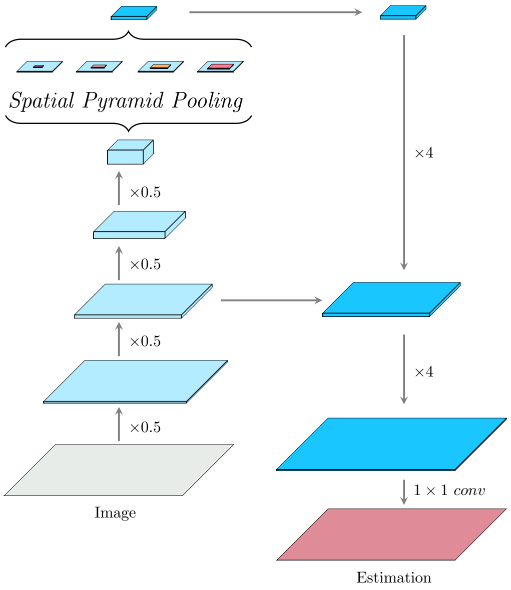

Fig. 1: **Illustration of The Multi-Scale Architecture Based on Atrous Spatial Pyramid Pooling.**

3.2. Detail-Preserving Objective Function

According to [15], in MDE, the loss on depth helps to produce a rough estimated depth map and upon that the loss on depth gradients helps to recover detailed structures of the depth map. For the loss on depth, we directly adopt the BerHu function by Laina *et al.* [7] as it is a robust loss function constructed with $\mathcal{L}_1$ and $\mathcal{L}_2$. Suppose that g_i, d_i and $e_i = |g_i - d_i|$ are the ground truth, the predicted depth and the absolute error, respectively, then the BerHu on error e_i is,

$$\mathcal{B}(e_i) = \begin{cases} |e_i| & |e_i| \le c \\ \frac{e_i^2+c^2}{2c} & |e_i| > c \end{cases}, \tag{1}$$

where the threshold c is set to $1/5$ of the maximum absolute error of the batch during training. So, the loss on depth is,

$$l_{depth} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{B}(e_i). \tag{2}$$

To note that our l_{depth} is different from that of Hu *et al.* [15], as it is accumulated from the logarithm transform of the absolute error. Define the surface normal of ground truth and estimated depth map as $n_i^g \equiv [-\nabla_x(g_i), -\nabla_y(g_i), 1]^\top$ and $n_i^d \equiv [-\nabla_x(d_i), -\nabla_y(d_i), 1]^\top$, respectively, then the loss on depth normals by [15] is,

$$l_{\text{normal}} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{\langle n_i^d, n_i^g \rangle}{\sqrt{\langle n_i^d, n_i^d \rangle}\sqrt{\langle n_i^g, n_i^g \rangle}}\right). \tag{3}$$

Table 1: **Quantitative Performance Comparison and Ablation Study on the NYU Depth V2 Test Set.** The best results are marked with bold-face and the runner-up results are marked with blue color.

Method	Error (lower is better)				Accuracy (higher is better)		
	Abs Rel	$\log_{10}$	RMSE	RMSE$_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Make3D [1]	0.349	-	1.214	0.409	0.447	0.745	0.897
Liu *et al.* [4]	0.213	0.087	0.759	-	0.650	0.906	0.976
Eigen *et al.* [5]	0.215	-	0.907	0.285	0.611	0.887	0.971
Eigen *et al.* [6]	0.158	-	0.641	0.214	0.769	0.950	0.988
FCRN [7]	0.127	0.055	0.573	0.195	0.811	0.953	0.988
Carvalho *et al.* [9]	0.135	0.059	0.600	0.199	0.819	0.957	0.987
Moukari *et al.* [10]	0.133	0.057	0.569	-	0.830	0.966	**0.993**
MS_CRF [12]	0.121	0.052	0.586	-	0.811	0.954	0.987
DORN [14]	**0.115**	**0.051**	0.509	-	0.828	0.965	0.992
Hu *et al.* [15] (ResNet-50)	0.126	0.054	0.555	-	**0.843**	**0.968**	0.991
Baseline w/ l_{depth}	0.129	0.056	0.487	0.170	0.834	0.962	0.990
Baseline+ w/ loss of Hu *et al.* [15]	0.129	0.056	0.488	0.169	0.828	0.961	0.991
Baseline+ w/ l_{depth}	0.127	0.056	0.486	0.169	0.834	0.963	0.991
Baseline+ w/ $l_{depth} + \lambda l_{grad\_mg}$	0.128	0.055	0.476	0.167	0.838	0.964	0.991
Baseline+ w/ full loss	0.127	0.054	**0.468**	**0.165**	0.841	0.966	**0.993**

l_{normal} is in fact designed to maximize the cosine value of the angle between two depth normals, but this loss contains division of gradient components.

To avoid the division, we adopt the $\mathcal{L}_1$ on the cross product of normals and we find that this loss can be divided into one error on the magnitude of the gradients and one error on the the direction of the gradients. The cross product of the two normals is,

$$n_i^d \times n_i^g = \begin{pmatrix} \nabla_y(g_i) - \nabla_y(d_i) \\ \nabla_x(d_i) - \nabla_x(g_i) \\ \nabla_y(g_i)\nabla_x(d_i) - \nabla_y(d_i)\nabla_x(g_i) \end{pmatrix}. \quad (4)$$

Accordingly, we use $\mathcal{L}_1$ to define the losses on the magnitude and the direction of the depth gradients as l_{grad_mg} and l_{grad_dr},

$$l_{grad\_mg} = \frac{1}{n}\sum_{i=1}^{n} |\nabla_y(g_i) - \nabla_y(d_i)| + |\nabla_x(d_i) - \nabla_x(g_i)|, \quad (5)$$

$$l_{grad\_dr} = \frac{1}{n}\sum_{i=1}^{n} |\nabla_y(g_i)\nabla_x(d_i) - \nabla_y(d_i)\nabla_x(g_i)|. \quad (6)$$

l_{grad_mg} and l_{grad_dr} are corresponding to l_{grad} and l_{normal} in [15] for preserving details. As the name indicates, l_{grad_mg} aims at constraining the gradient magnitudes to be the same and l_{grad_dr} reduces the difference on the gradient directions.

Finally, we define our whole detail-preserving objective function as,

$$L = l_{depth} + \lambda l_{grad\_mg} + \mu l_{grad\_dr}, \quad (7)$$

where $\lambda, \mu \in \mathbb{R}^+$ are weights to balance the objectives.

4. EXPERIMENTS

We demonstrate the effectiveness of our approach by experiments on the NYU Depth V2 [24] dataset in this section.

4.1. Implementation Details

We implement our approach by tensorflow on a Nvidia TITAN Xp GPU. We adopt ResNet-50 [20] as backbone and set the network input size as 289×385. Following data augmentation like [5], we train on $120k$ samples for 20 epochs with batchsize 8 and the weights λ and μ both as 1. The initial learning rate is 10^{-4} and decreases by 0.9 every $10k$ steps.

4.2. Evaluation Metrics

We adopt some conventional metrics in [7] for evaluation. Let d be the predicted depth and g the ground truth depth for total T pixels in an image. The metrics are: (1) Mean Absolute Relative Error (Abs Rel): $\frac{1}{T}\sum\frac{|d-g|}{g}$; (2) Mean $\log_{10}$ Error ($\log_{10}$): $\frac{1}{T}\sum|\log_{10} d - \log_{10} g|$; (3) Root Mean Squared Error (RMSE): $\sqrt{\frac{1}{T}\sum(d-g)^2}$; (4) Root Mean Squared Error in log space (RMSE$_{log}$): $\sqrt{\frac{1}{T}\sum(\log d - \log g)^2}$ and (5) the accuracy with threshold t, i.e. the percentage of such that $\delta = \max(\frac{d}{g}, \frac{d}{g}) < t$, where $t \in [1.25, 1.25^2, 1.25^3]$.

4.3. Experimental Results

Quantitative Performance. Table 1 shows the quantitative performance of some state of the art methods and variants of our approach. Obviously, our approach achieves the state of

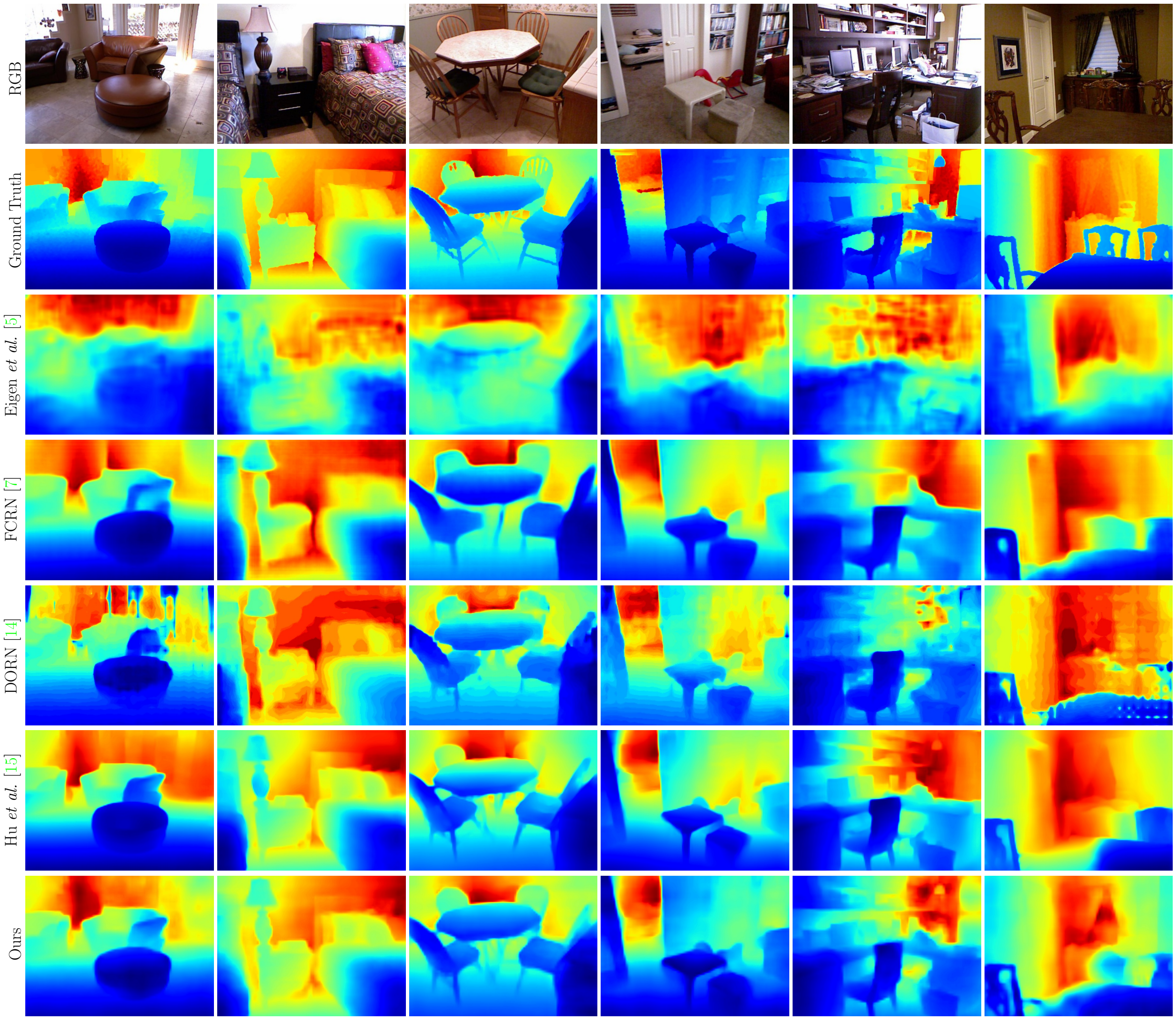

Fig. 2: **Qualitative Comparison on Some Examples of the NYU Depth V2 Test Set.**

the art on the accuracy metrics and makes a large improvement in RMSE. According to the performance of the variants of our approach, we find that our modification of DeepLabv3+ increases the performance on almost every metric, and the proposed l_{grad_mg} and l_{grad_dr} boosts the performance too, especially on RMSE. We also experiment on the full loss of Hu *et al.* [15] with our baseline+, and results indicate that our full loss performs markedly better than it.

Qualitative Comparison. Figure 2 demonstrates the qualitative results on some examples in NYU Depth V2 test set. Apparently, Eigen *et al.* [5] could not produce accurate prediction, while FCRN [7] over smoothes the estimation. Although DORN [14] outperforms our approach in Abs Rel and $\log_{10}$, its estimated depth map contains much more noise. Our approach outputs higher quality depth maps than other methods, and is slightly better than the method by Hu *et al.* [15]. For example, in the sample of the last column in Figure 2, our approach can preserve the inner structures of the left chair better and perceive the existence of the middle chair which is very difficult to be recognized due to the similar background.

5. CONCLUSIONS

We have presented our work on high quality monocular depth estimation. We solve this problem by constructing a multi-scale network inspired by DeepLabv3+ and proposing a simplified detail-preserving objective. Experiments show that we significantly decreased the RMSE with the adopted architecture, and our modification of it can slightly boost the performance too. In addition, our simplified detail-preserving loss function is effective and works better than the previous one with the network used in this paper. In conclusion, we have made a good progress in the target problem.

6. REFERENCES

[1] Ashutosh Saxena, Min Sun, and Andrew Y Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE TPAMI*, vol. 31, no. 5, pp. 824–840, 2009. 1, 3

[2] Heiko Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE TPAMI*, vol. 30, no. 2, pp. 328–341, 2008. 1

[3] Zhan Song and Ronald Chung, "Determining both surface position and orientation in structured-light-based sensing," *IEEE TPAMI*, vol. 32, no. 10, pp. 1770–1780, 2010.

[4] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian D Reid, "Learning depth from single monocular images using deep convolutional neural fields.," *IEEE TPAMI*, vol. 38, no. 10, pp. 2024–2039, 2016. 1, 3

[5] David Eigen, Christian Puhrsch, and Rob Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *NIPS*, 2014. 1, 3, 4

[6] David Eigen and Rob Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *CVPR*, 2015. 1, 3

[7] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab, "Deeper depth prediction with fully convolutional residual networks," in *3DV*, 2016. 1, 2, 3, 4

[8] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE TCSVT*, 2017. 1, 2

[9] Marcela Carvalho, Bertrand Le Saux, Pauline Trouvé-Peloux, Andrés Almansa, and Frédéric Champagnat, "On regression losses for deep depth estimation," in *ICIP*. IEEE, 2018. 1, 2, 3

[10] M. Moukari, S. Picard, L. Simoni, and F. Jurie, "Deep multi-scale architectures for monocular depth estimation," in *ICIP*. IEEE, 2018, pp. 2940–2944. 1, 2, 3

[11] Bo Li, Yuchao Dai, and Mingyi He, "Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference," *Pattern Recognition*, 2018. 1, 2

[12] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *CVPR*, 2017. 1, 3

[13] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *CVPR*, 2018. 1

[14] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao, "Deep ordinal regression network for monocular depth estimation," in *CVPR*, 2018. 1, 2, 3, 4

[15] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *WACV*, 2019. 1, 2, 3, 4

[16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE TPAMI*, vol. 40, no. 4, pp. 834–848, 2018. 1, 2

[17] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017. 1, 2

[18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018. 1, 2

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015. 1

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 1, 2, 3

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015. 2

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *ECCV*, 2016. 2

[24] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012. 3