A new method for monocular depth estimation

Zhaofeng Niu
Nara Institute of Science and Technology

1. Introduction

Depth information is very important for various applications [1], such as 3D reconstruction, autonomous driving and augmented reality. Nowadays, there are several ways to obtain the depth information: 1) using depth camera, which measures the depth by structure light [2], etc.; 2) using stereo images or video streams, which provides some spatial and temporal information to estimate the depth map [3, 4]; 3) using monocular images [5], for which the input is a single RGB image without temporal information. Among these three ways, using the depth camera is the simplest way but it is expensive and usually has lots of limitation, such as short sensing range and low resolution, while using single RGB images is the most convenient way with lowest cost, but it is very challenging due to the limited information from the input RGB image.

My research is depth estimation on single RGB images, which is the third way mentioned above. There are lots of existing methods aiming to achieve the high accurate depth estimation with deep learning method, which is proved to have better performance than the traditional methods [6]. For example, Liu *et al*. [7] propose a convolutional neural network (CNN) model with a conditional random field (CRF) loss, which is used to minimize the log-likelihood between neighboring superpixels generated by the model, while Cao *et al*. [8] design a fully connected CRF to do the post-processing for refining the output. In addition, since depth is a kind of geometric information, combining other geometric information such as surface normal [9] and semantic information [10], may help to improve the accuracy of depth estimation. Zhang *et al*. [11] propose a multi-task network, which can predict the surface normal, semantic segmentation and depth map simultaneously. However, the performance can be improved by better utilizing the multi-scale information of the input images, which is proved to be one of the keys for generating high-accuracy depth estimation.

Inspired by BTS method [12], a new monocular depth estimation method is proposed in this work, *i.e.*, HMA-Depth method [13] (which means a Hierarchical Multi-scale Attention method for Depth estimation). In BTS method, a multi-scale method is proposed for depth estimation. For each scale of the image features, a local planar guidance (LPG) module is introduced to guide the features back to the original resolution of the input image. However, it uses a convolutional layer to combine the results from each scale, which does not fully utilize the advantages of the scaled information. Instead of using LPG module, HMA-Depth method adopts a hierarchical multi-scale attention method to do the depth estimation. Similar to BTS method, HMA-Depth also tries to obtain the various advantages among different resolution scales. Specifically, higher resolution can show more details of the image while lower resolution has relatively clear object contour information. HMA-Depth extracts the features of multiple scales and adopts an attention mechanism to generate the important area in the estimated depth maps. Different from BTS method, HMA-Depth generates the depth map for each scale, as well as the attention map, which can show the area that is important for estimation. The experiment results prove that HMA-Depth outperforms BTS method and other state-of-art methods.

2. Method

To obtain more precise local information from the image while keeping a good understanding of the global context, HMA-Depth method is proposed, which adopts a multi-scale attention frame to do depth estimation. The network architecture is shown in Fig.1. Firstly, a backbone network extracts the features into different scales, *i.e.*, *H/8*, *H/4*, *H/2* and *H* (*H/s* indicates the scales of *H/s* × *W/s* for short, where $s \in \{1, 2, 4, 8\}$, and *H* and *W* represent the height and weight of the input image, respectively). Specifically, an atrous spatial pyramid pooling (ASPP) module, skip connections and bilinear interpolation are used in the up-sampling process. The ASPP module, using convolutional kernels with different dilation rates ($r \in \{3, 6, 12, 18, 24\}$), is used to improve the feature quality.

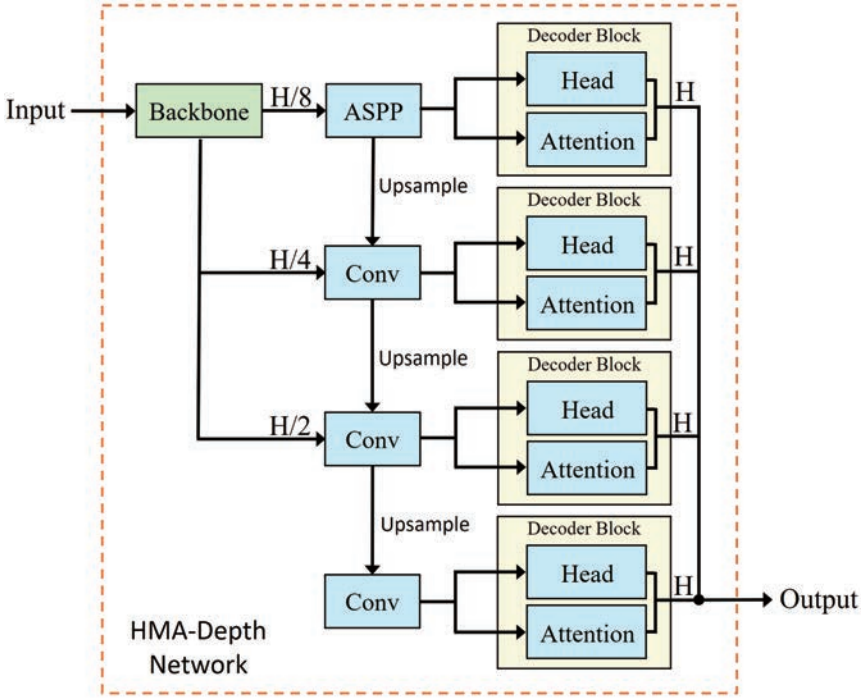

Fig.1 Network architecture

For each scale of resolution, a decoder block, which involves an attention module and a depth module, is used to generate the estimation results for the corresponding scale. The depth module estimates the depth map for each scale while the attention module can extract the preferred regions for depth map. To make each scale focus on independent parts of the input image, the calculation of attention masks can be explained as follows:

$$M_{H/8} = A_{H/8} \quad (1)$$

$$M_{H/4} = A_{H/4}(1 - A_{H/8}) \quad (2)$$

$$M_{H/2} = A_{H/2}(1 - A_{H/8})(1 - A_{H/4}) \quad (3)$$

$$M_{H} = (1 - A_{H/8})(1 - A_{H/4})(1 - A_{H/2}) \quad (4)$$

$A_{H/8}$, $A_{H/4}$ and $A_{H/2}$ indicate the attention maps for the scales of *H/8*, *H/4* and *H/2*, respectively, then $M_{H/8}$, $M_{H/4}$,$M_{H/2}$, and M_H indicate the corresponding weighted masks for each scale. These calculations can make each attention module focus on independent parts of the input, and the sum of masks equals to 1.

The scaled depth maps and attention maps are illustrated in Fig.2. It can be seen that the global information such as the object contour is estimated better in the lower resolution while higher resolution module predicts fine details. And the sum of four masks should be a whole white mask.

The final output can be represented as follows:

$$D_{final} = \sum_{s \in S} M_s \cdot D_s^m \quad (5)$$

where D^m is the scaled depth map generated from the depth module.

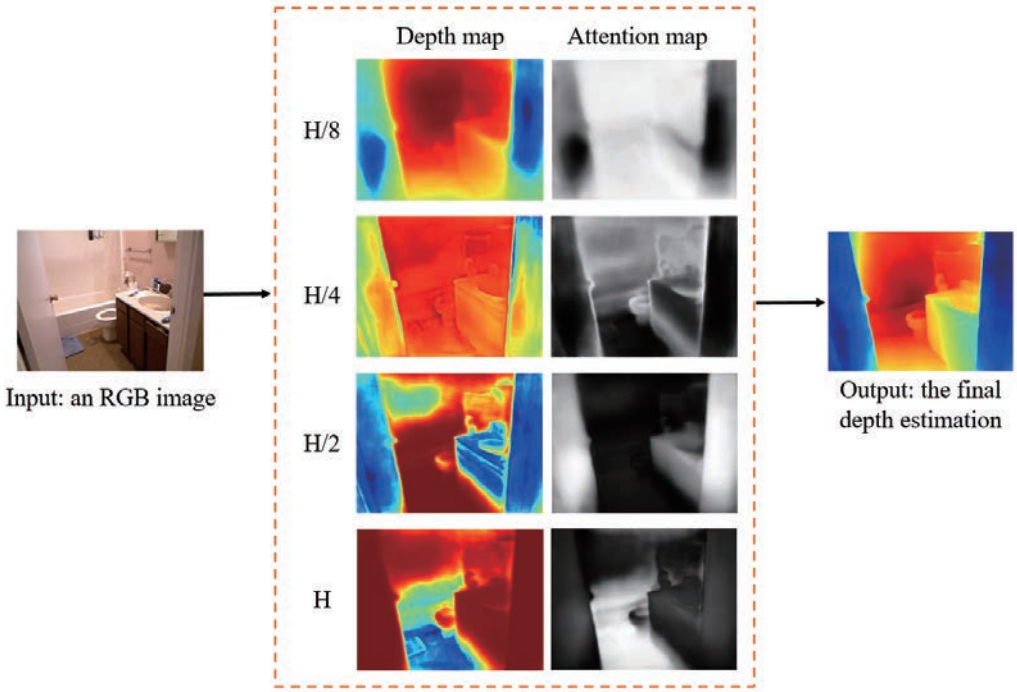

Fig.2 Depth and attention maps generated at different scales

As for the loss function, the scale-invariant error proposed by Eigen *et al.* [6] is adopted to calculate the error between the predicted y and the ground truth y^*, and the formula is shown as follows:

$$\text{Loss} = \frac{1}{n}\sum_i g_i^2 - \frac{\lambda}{n^2}(\sum_i g_i)^2 \quad (6)$$

in which $g_i = \log y_i - \log y_i^*$ and $\lambda \in [0,1]$; n represents the number of pixels that have valid depth values. Similar to BTS method [12], $\lambda = 0.85$ is set in the experiments.

3. Experimental results

In the experiment, PyTorch [14] is used to implement the whole network. The number of the epoch is set as 50 and the batch size is 16. Multiple networks, such as ResNet 50, ResNeXt 50, are used as the backbone network, extracting the dense feature. A part of the experiments is conducted with Reedbush-H server. To verify the effectiveness of the proposed method, one commonly-used dataset, *i.e.*, KITTI dataset [15], is adopted to conduct the experiments.

KITTI dataset is obtained by an autonomous driving platform, which is equipped with a laser scanner, a GPS localization system and a stereo camera rig. To compare with other methods, the commonly used Eigen split [6] is adopted in the experiments, involving 23488 images from 32 scenes for training and 697

images from 29 scenes for testing. Tab.I shows the comparison result on KITTI dataset.

Tab.I. Quantitative results on KITTI dataset

Methods	**δ1↑**	**δ2↑**	**δ3↑**	**AbsRel↓**	**RMSE↓**	**RMSElog↓**
Make3D [16]	0.601	0.820	0.926	0.280	8.734	0.361
Eigen *et al*. [6]	0.702	0.898	0.967	0.203	6.307	0.282
Liu *et al*. [17]	0.680	0.898	0.967	0.201	6.471	0.273
Kuznietso *et al*. [18]	0.862	0.960	0.986	0.113	4.621	0.189
Yin *et al*. [19]	0.938	0.990	**0.998**	0.072	3.258	0.117
DORN[5]	0.932	0.984	0.994	0.072	**2.727**	0.120
BTS-ResNet 50 [12]	0.950	0.991	**0.998**	0.062	2.878	0.101
BTS-DenseNet 161 [12]	0.952	0.992	**0.998**	0.062	2.871	0.094
HMA-Depth-ResNet 50	0.953	0.992	**0.998**	0.062	2.870	0.096
HMA-Depth-ResNeXt 50	0.951	0.992	**0.998**	0.062	2.867	0.094
HMA-Depth-DenseNet 121	0.952	0.991	**0.998**	0.063	2.874	0.096
HMA-Depth-DenseNet 161	**0.955**	**0.993**	**0.998**	**0.060**	2.850	**0.092**

* ↑ indicates that the performance is better when the value is greater; ↓ indicates the performance is better when the value is lower. The bold value represents the best value for each metric.

According to the table above, HMA-Depth method outperforms other methods for all the metric, except the root mean square error (RMSE) metric, which is a little larger than DORN method. In addition, Fig.3 shows some qualitative results to prove the visualization performance. It compares BTS method and HMD-Depth method by depth estimation results of two RGB images from KITTI dataset. It can be seen that HMD-Depth method can generate more accurate boundary of the objects and more smooth surface.

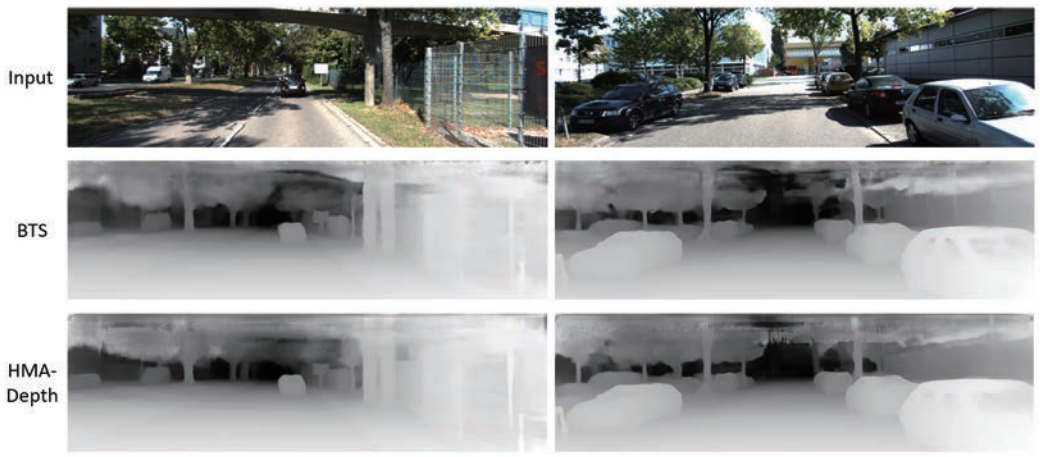

Fig.3 Visualization results of the KITTI dataset

4. Conclusion

In this work, a new method, name HMA-Depth method, is proposed for depth estimation of single RGB images. It utilizes the advantages of different scales of the image and adopts an attention mechanism to extract the important areas of each scale. Specifically, HMA-Depth makes each scale focus on independent area of the images, aiming to amplify the advantages of each scale. The quantitative results and qualitative results both prove that HMA-Depth method outperforms other methods on KITTI dataset. To obtain more solid comparison, however, another common-used dataset, NYU V2 dataset [20] as an example, need to be used. Besides, some ablation studies are also necessary to prove the effectiveness of 4-scale frame. The ablation study may include 3-scale, 5-scale frame or 4-scale frame without attention mechanism.

Therefore, the experiments on another datasets and ablation experiments will be conducted for the next step. To improve the performance further, combining depth estimation task with semantic segmentation results, which can help smooth the object surface, may also be considered in the future work.

Reference

[1] Silvan Weder, Johannes Schonberger, Marc Pollefeys, and Martin R Oswald, “RoutedFusion: Learning real-time depth map fusion,” in IEEE CVPR, 2020, pp.4887–4897.

[2] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al., “KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera,” in Annual ACM Symposium on User Interface Software and Technology, 2011, pp. 559–568.

[3] Jamie Watson, Oisin Mac Aodha, Daniyar Turmukham-betov, Gabriel J Brostow, and Michael Firman, “Learning stereo from single images,” in ECCV, 2020, pp. 722–740.

[4] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf, “Consistent video depth estimation,” arXiv preprint arXiv:2004.15021, 2020.

[5] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao, “Deep ordinal regression network for monocular depth estimation,” in IEEE CVPR, 2018, pp. 2002–2011.

[6] David Eigen, Christian Puhrsch, and Rob Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in Advances in Neural Information Processing Systems, 2014, pp. 2366–2374.

[7] Fayao Liu, Chunhua Shen, and Guosheng Lin, “Deep convolutional neural fields for depth estimation from a single image,” in IEEE CVPR, 2015, pp. 5162–5170.

[8] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen, “Estimating depth from monocular images as classification using deep fully convolutional residual networks,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 11, pp. 3174–3182,2017.

[9] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hen-gel, and Mingyi He, “Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs,” in IEEE CVPR, 2015, pp. 1119–1127.

[10] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille, “Towards unified depth and semantic prediction from a single image,” in IEEE CVPR, 2015, pp. 2800–2809.

[11] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, NicuSebe, and Jian Yang, “Pattern-affinitive propagation across depth, surface normal and semantic segmentation,” in IEEE CVPR, 2019, pp. 4106–

4115.
[12] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh, “From big to small: Multi-scale local planar guidance for monocular depth estimation,” arXiv preprint arXiv:1907.10326, 2019.
[13] Zhaofeng Niu, Yuichiro Fujimoto, Masayuki Kanbara, and Hirokazu Kato. "HMA-Depth: A New Monocular Depth Estimation Model Using Hierarchical Multi-Scale Attention," IEEE 17th International Conference on Machine Vision and Applications (MVA), pp. 1-5, 2021.
[14] Adam Paszke, Sam Gross, Francisco Massa, AdamLerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., “Pytorch: An imperative style, high-performance deep learning library,” in Advances in Neural Information Processing Systems, 2019, pp. 8026–8037.
[15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, “Vision meets robotics: The KITTI dataset,” International Journal of Robotics Research (IJRR), 2013.
[16] Ashutosh Saxena, Min Sun, and Andrew Y Ng, “Make3D: Learning 3D scene structure from a single still image,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 5, pp. 824–840, 2008.
[17] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid, “Learning depth from single monocular images using deep convolutional neural fields,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no.10, pp. 2024–2039, 2015.
[18] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe, “Semi-supervised deep learning for monocular depth map prediction,” in IEEE CVPR, 2017, pp. 6647–6655.
[19] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan, “Enforcing geometric constraints of virtual normal for depth prediction,” in IEEE ICCV, 2019, pp. 5684–5693.
[20] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, “Indoor segmentation and support inference from RGBD images,” in ECCV, 2012, pp. 746–760.