Multi-scale Residual Pyramid Attention Network for Monocular Depth Estimation

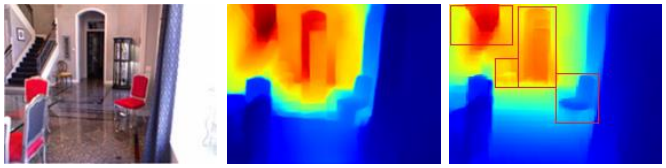
Jing Liu[1], Xiaona Zhang[1], Zhaoxin Li[2] and Tianlu Mao[2]

1. College of Computer and Cyber Security Hebei Normal University Shijiazhuang, Hebei, China
2. Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences Beijing, China

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Introduction

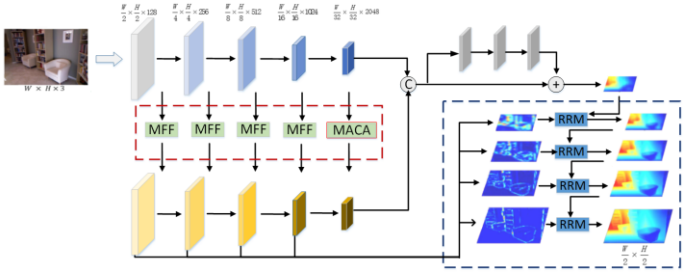
1. We propose a multi-scale attention context aggregation (MACA) module, which can adaptively aggregate spatial and scale context information by learning the similarity between pixels.
2. We propose an improved residual refinement module (RRM) that captures deeper semantic information and more details information to refine the scene structure.
3. Our method achieved competitive performance in object boundaries and local details.



(a) (b) (c)

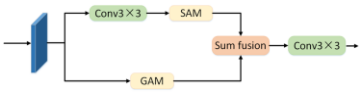
The visual comparison of estimated depth maps. (a) RGB input, (b) the result by the state-of-the-art method [1], and (c) the result by our method. Our method achieve better results for the local details (e.g., stairs) and boundaries of objects (e.g., chairs).

Method

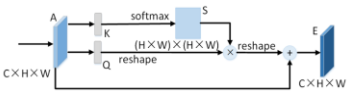


Overview

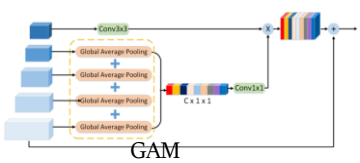
- We proposed a multi-scale attention context aggregation (MACA) module and an improved residual refinement module (RRM).
- The MACA module consists of spatial attention module (SAM) and global attention module (GAM), which adaptively learns the similarities between pixels to aggregation the spatial and scale context information.
- The RRM module can capture more details information to further refine the scene structure.



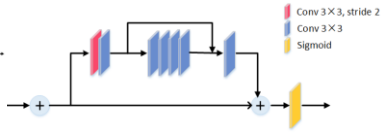
MACA



SAM



GAM



RPM

Loss Function

To train our residual pyramid network, we compute the difference between the predicted depth map D^i and the ground-truth G^i at each scale. For each scale, It consists of three terms, l_{depth} considering the pixel-wise difference between D^i and the ground truth G^i, l_{grad} penalizing errors around edges, and l_{normal} further improving fine details.

Combing all the L scales, our loss function for the entire network is

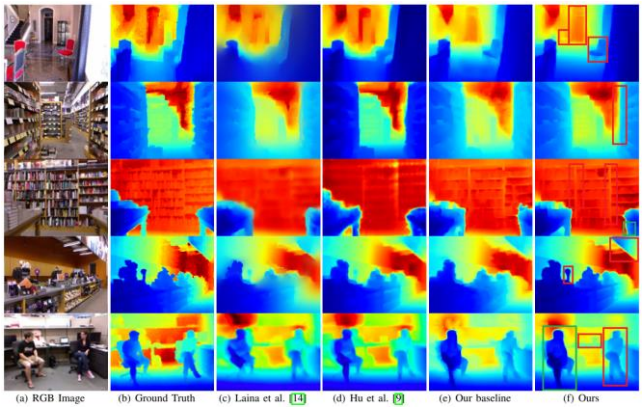
$$Loss = \sum_{i=1}^{L} (l_{depth}^{i} + l_{grad}^{i} + l_{normal}^{i})$$

Experiments

Implementation

- initialize the encoder module by pre-trained model on ImageNet, use SENet as the backbone.
- Adam optimizer with initial learning rate 10^{-4}, reduce 10% every 5 epoch. $\beta 1 = 0.9, \beta 2 = 0.999$ and weight decay as 10^{-4}
- Network was trained for 20 epochs with a batch size of 4.

Results (NYU Depth V2)



(a) RGB Image (b) Ground Truth (c) Laina et al. [14] (d) Hu et al. [9] (e) Our baseline (f) Ours

The red regions are good results, and the green regions are misestimation.

TABLE I
COMPARISONS WITH STATE-OF-THE-ART DEPTH ESTIMATION APPROACHES ON NYU DEPTH V2 DATASET.

Method	lower is better			higher is better		
	Abs Rel	RMS	Log10	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Eigen et al. [12]	0.215	0.907	–	0.611	0.887	0.971
Laina et al. [14]	0.127	0.573	0.055	0.811	0.953	0.988
Xu et al. [26]	0.125	0.593	0.057	0.806	0.952	0.986
Chen et al. [20]	0.138	0.496	–	0.826	0.964	0.990
Fu et al. [16]	0.115	0.509	0.051	0.828	0.965	0.992
Jiao et al. [17]	**0.098**	**0.329**	**0.040**	**0.917**	**0.983**	**0.996**
Hu et al. [9]	0.115	0.530	0.050	0.866	0.975	0.993
Ding et al. [19]	0.101	0.519	0.044	0.847	0.967	0.992
Our Baseline	0.123	0.596	0.056	0.838	0.968	0.992
Our Baseline + MACA	0.121	0.537	0.051	0.854	0.973	0.993
Ours: Baseline + MACA + RRM	0.113	0.525	0.049	0.872	0.974	0.993

Conclusion

- We proposed an multi-scale residual pyramid attention network (MRPAN) for monocular depth estimation.
- The experiment results show that our method achieves competitive performance in comparsion with the state-of-the-art methods, especially the boundaries and local details of image in complex scenes.
- Our future work will focus on improving numerical image accuracy, and plan to extend our method to other dense labeling tasks, such as semantic segmentation.

Reference

1. J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Jan 2019, pp. 1043–1051.
2. Due to space limitations, the references in the ***Experiments*** are not listed here. Please refer to our paper for details.

Acknowledgment

This work is in part supported by the National Natural Science Foundation of China (61532002, 61702482, 61802109), the Major Program of National Natural Science Foundation of China (91938301), the National Defense Equipment Advance Research Shared Technology Program of China (41402050301-170441402065), and the Sichuan Science and Technology Major Project on New Generation Artificial Intelligence (2018GZDZX0034). the Natural Science Foundation of Hebei Province (F2020205006), the Top Youth Talents of Science and Technology Research Project in Hebei Province (BJ2020059), and the Science Foundation of Hebei Normal University (L2018K02). Open Foundation of Beijing Key Laboratory of Mobile Computing and Pervasive Device.