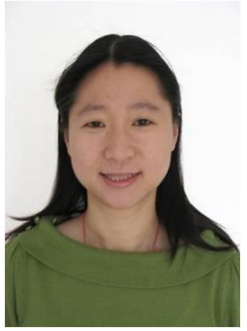
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Multi-scale Residual Pyramid Attention Network for Monocular Depth Estimation



Jing Liu[1] Xiaona Zhang[1] Zhaoxin Li[2] 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

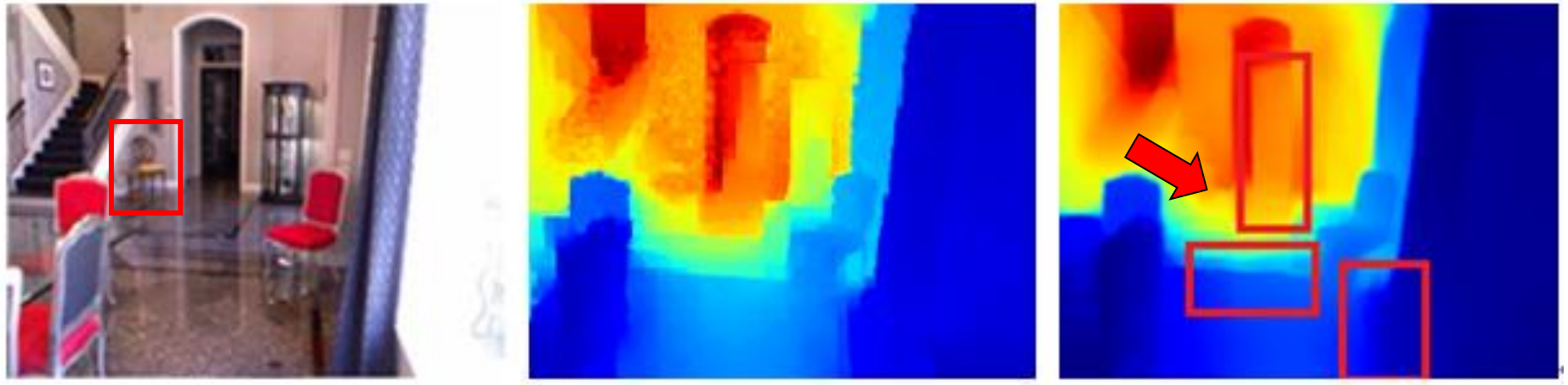
河北师范大学 HEBEI NORMAL UNIVERSITY

中国科学院计算技术研究所 INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Contents

Introduction

- Existing methods fail to consider complex textures and geometries in scenes.



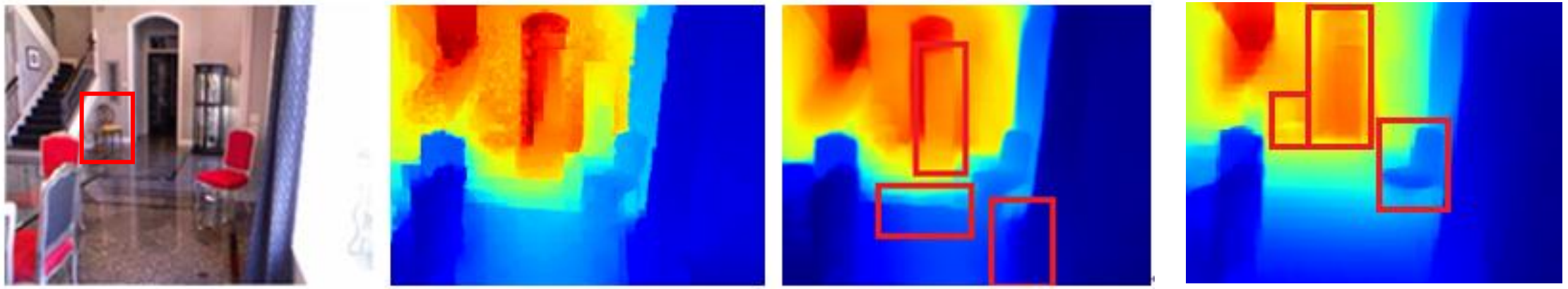
RGB GT Hu et al.[9]

Problems:

loss of local details, distorted object boundaries, and blurry reconstruction.

Introduction

- In this work, we proposed an end-to-end multi-scale residual pyramid attention network.

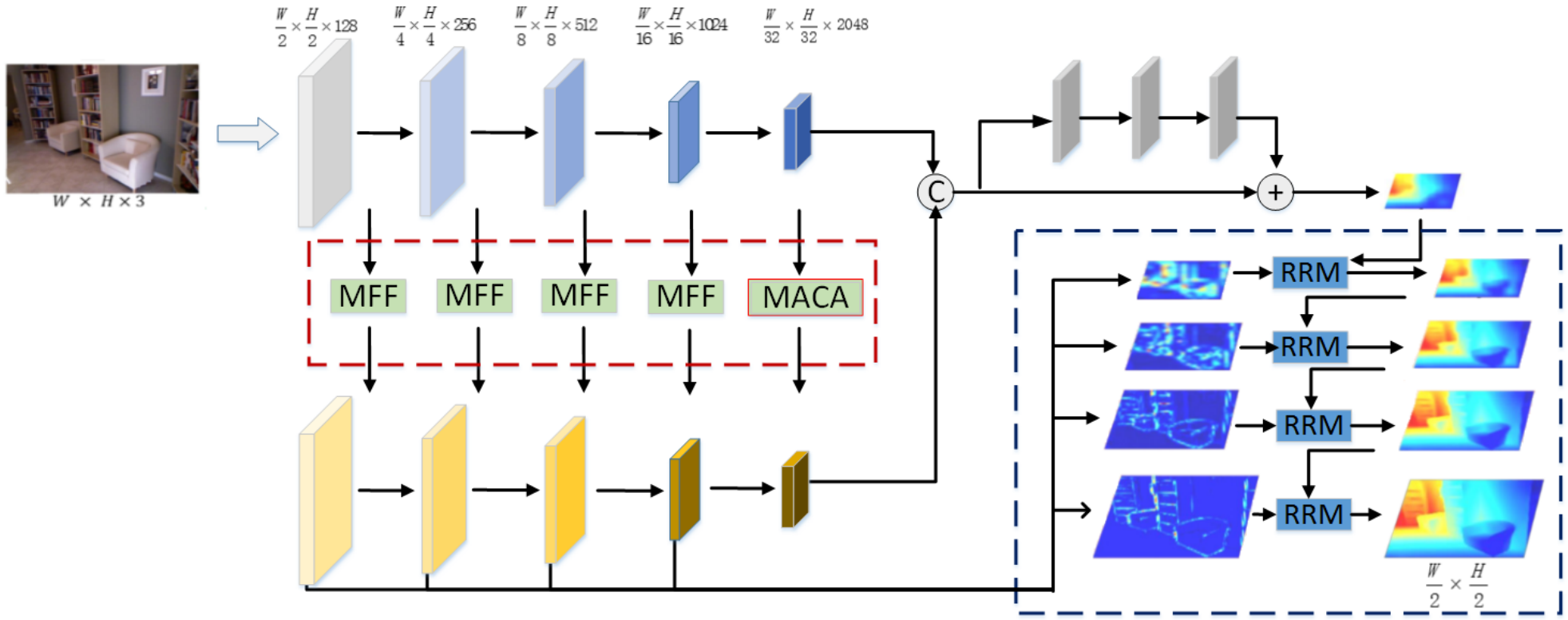


RGB GT Hu et al.[9] Ours

Our method achieved competitive performance in object boundaries and local details.

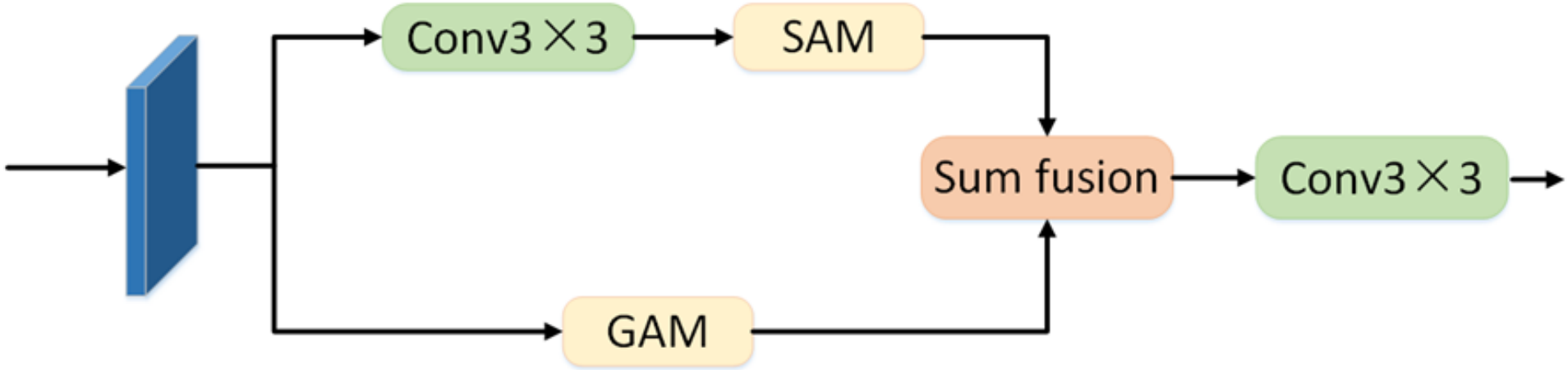
Method

- The proposed network architecture:



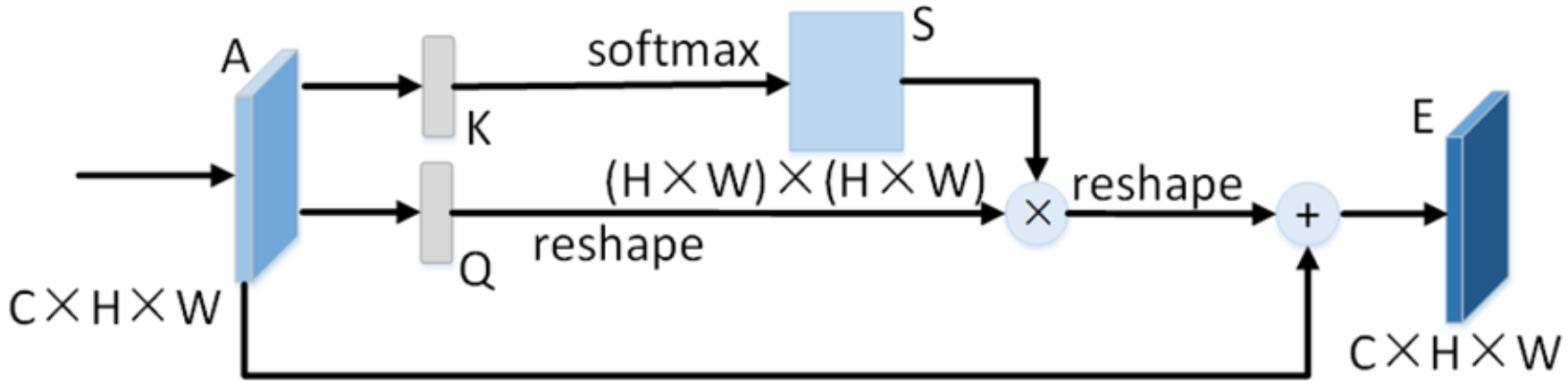
Method

- Multi-scale attention context aggregation (MACA)
  - We proposed MACA, the module consists of SAM and GAM, which adaptively learns the similarities between pixels to aggregation the spatial and scale context information.



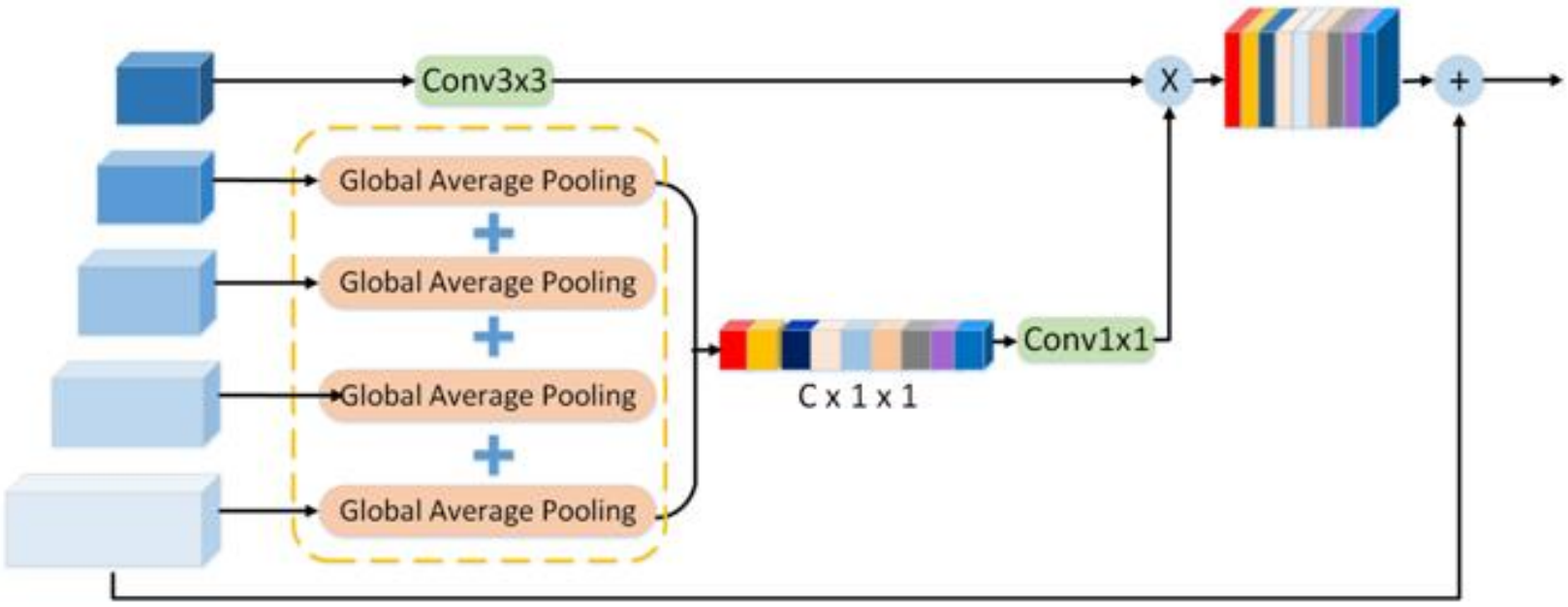
Method

- Spatial attention module (SAM)



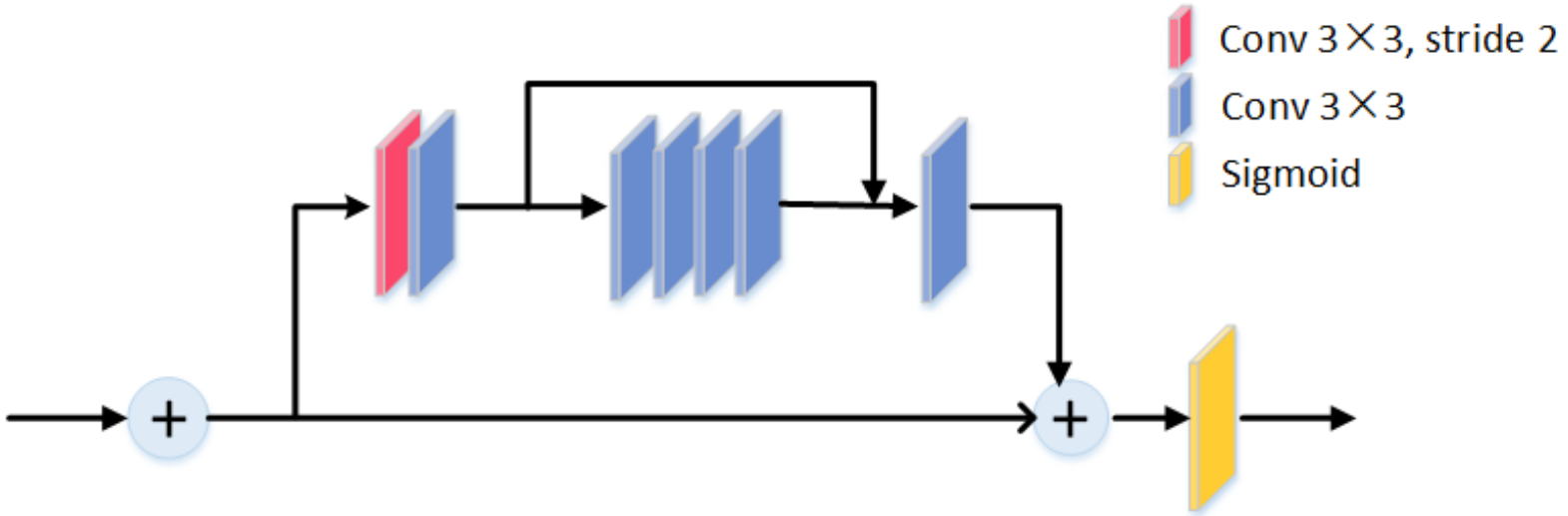
Method

- Global attention module (GAM)



Method

- Residual refinement module (RRM)
  - Improved RRM, the module can capture more details information to further refine the scene structure.



Method

- Computed the difference between the predicted depth map D^i and the ground-truth G^i at each scale. Combing all the L scales, our loss function for the entire network is:

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

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.

l_{normal} further improving fine details.

Experiment Results

- Implementation Details

  - 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.

Experiment Results

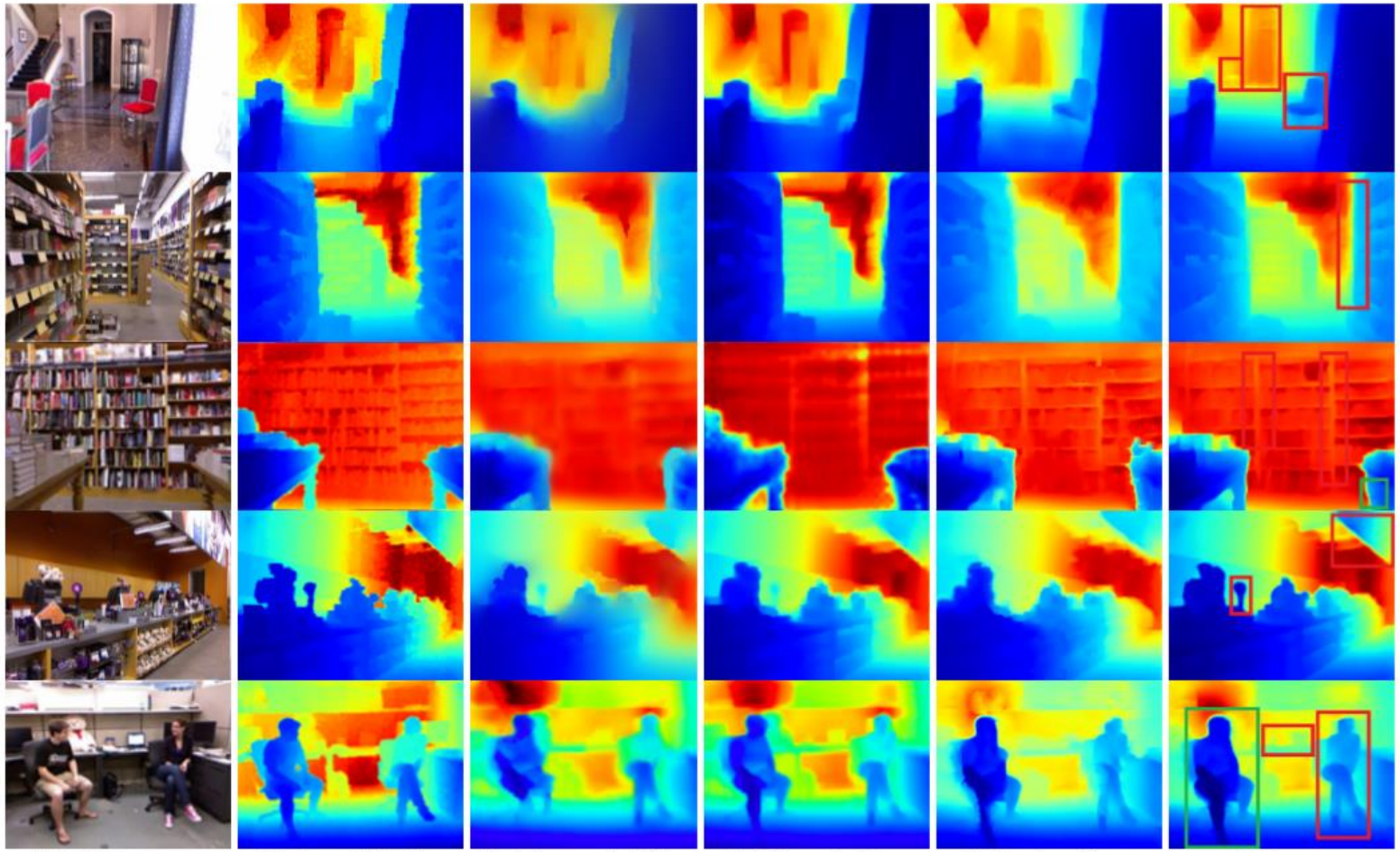
- Quantitative Evaluation(NYU Depth V2)

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

Experiment Results

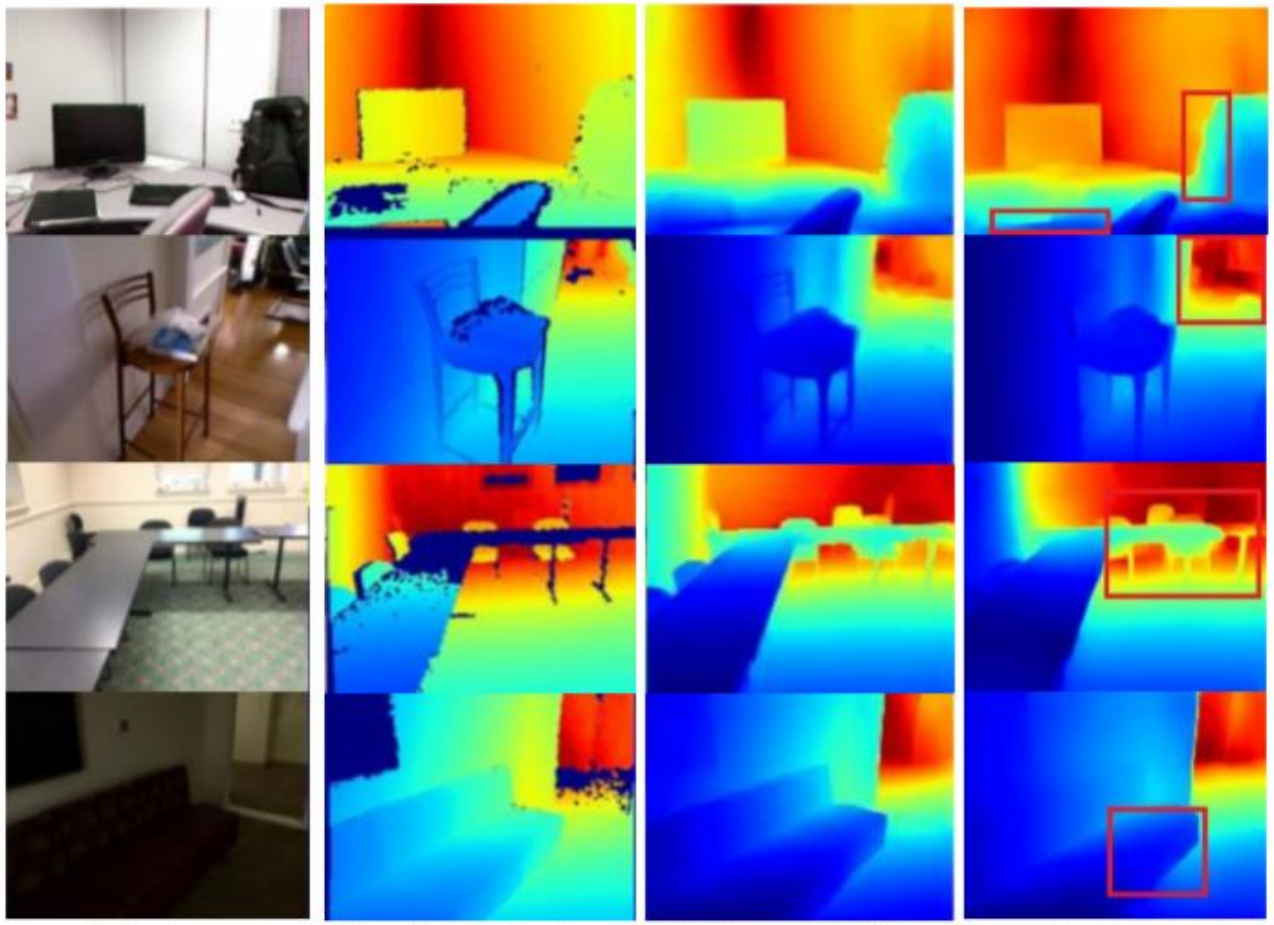
- Qualitative Evaluation(NYU Depth V2)



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

Experiment Results

- Qualitative Evaluation(SUN-RGBD)



(a) RGB Image (b) Ground Truth (c) Chen et al. [21] (d) Ours

Conclusion

- We proposed an MRPAN for monocular depth estimation.
- Achieves competitive performance in comparsion with the state-of-the-art methods, especially the boundaries and local details of image in complex scenes.

Thanks !
Q&A