Deep attention-based classification network for robust depth prediction

Ruibo Li[1*], Ke Xian[1*], Chunhua Shen[2], Zhiguo Cao[1], Hao Lu[1], Lingxiao Hang[1]

[1]Huazhong University of Science and Technology, China
[2]The University of Adelaide, Australia
e-mail: {liruibo, kexian}@hust.edu.cn

Abstract. In this paper, we present our deep attention-based classification (DABC) network for robust single image depth prediction, in the context of the Robust Vision Challenge 2018 (ROB 2018)[1]. Unlike conventional depth prediction, our goal is to design a model that can perform well in both indoor and outdoor scenes with a single parameter set. However, robust depth prediction suffers from two challenging problems: a) How to extract more discriminative features for different scenes (compared to a single scene)? b) How to handle the large differences of depth ranges between indoor and outdoor datasets? To address these two problems, we first formulate depth prediction as a multi-class classification task and apply a softmax classifier to classify the depth label of each pixel. We then introduce a global pooling layer and a channel-wise attention mechanism to adaptively select the discriminative channels of features and to update the original features by assigning important channels with higher weights. Further, to reduce the influence of quantization errors, we employ a soft-weighted sum inference strategy for the final prediction.

Experimental results on both indoor and outdoor datasets demonstrate the effectiveness of our method. It is worth mentioning that we won the 2-nd place in single image depth prediction entry of ROB 2018, in conjunction with IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018.

[*] The first two authors contributed equally.

[1] http://www.robustvision.net/index.php

indoor depth range, and vice versa. This is why our model can adapt to the mixed data and avoid the mutual interference of different datasets.

Table 3. Comparison between classification and regression on the ScanNet validation dataset.

Method	sqRel	absRel	irmse	imae
Regression*	5.93	17.05	18.65	11.84
Regression	6.51	17.47	18.65	11.73
Classification*	**4.33**	12.84	16.05	8.88
Classification	4.34	**12.67**	**15.93**	**8.71**

Table 4. Comparison between classification and regression on the KITTI validation dataset.

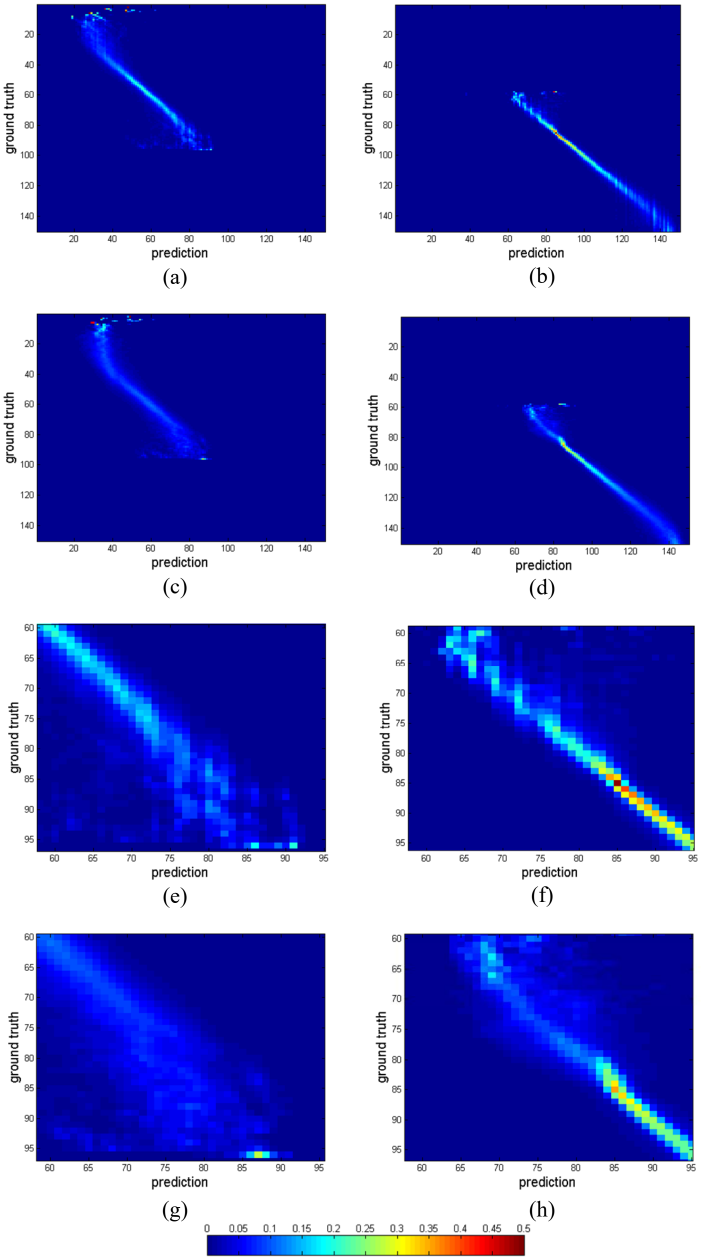
Method	SILog	sqRel	absRel	irmse
Regression*	14.55	2.26	9.33	11.45
Regression	14.49	2.28	9.44	11.63
Classification*	13.42	**1.92**	**7.86**	9.72
Classification	**13.34**	1.95	8.01	**9.58**

5.2 Effect of attention mechanism

In order to reveal the effectiveness of the channel-wise attention mechanism, we conduct an ablation study. Results are shown in Tables 5 and 6. Specially, each metric in Table 5 is multiplied by 100 for easy comparison. In these tables, “DABC w/o attention” represents a DABC model that ignores the attention vectors and directly sums the multi-scale features in each fusion block. Compared with “DABC w/o attention”, our DABC model makes a significant improvement on the ScanNet dataset and achieves comparable performance on the KITTI dataset.

Table 5. Evaluation of the attention mechanism on the ScanNet validation dataset.

Method	sqRel	absRel	irmse	imae
DABC w/o attention	4.93	13.50	16.37	9.073
DABC	**4.34**	**12.67**	**15.93**	**8.71**



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 5. Visualization of confusion matrices. (a) is the confusion matrix of our DABC model on the ScanNet. (b) is the confusion matrix of our DABC model on the KITTI. (c) is the confusion matrix of the regression network on the ScanNet. (d) is the confusion matrix of the regression network on the KITTI. (e)∼(h) show the details of the above four confusion matrices in the label range of 60 to 95, respectively.

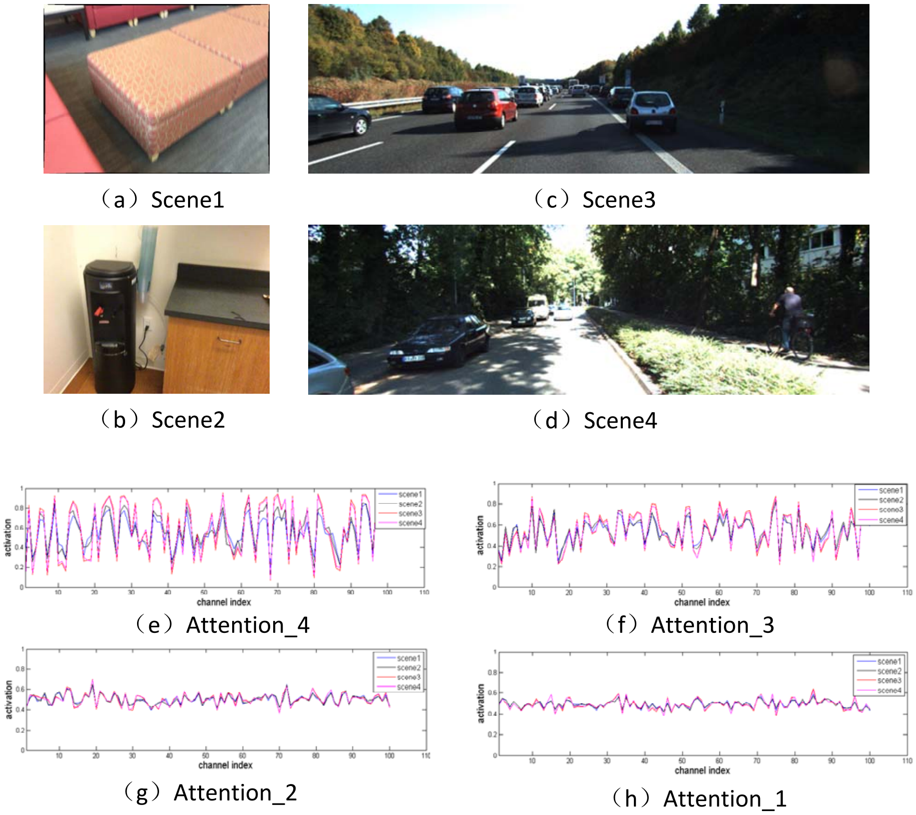


Fig. 6. Activations of different AFA modules. (a) and (b) are the color images from the ScanNet, (c) and (d) are from the KITTI. (e)∼(h) are the activations of AFA modules from the high-level to the low-level in sequence.

Table 6. Evaluation for attention mechanism on the KITTI validation dataset.

Method	SILog	sqRel	absRel	irmse
DABC w/o attention	**13.33**	2.00	**8.00**	**9.58**
DABC	13.34	**1.95**	8.01	**9.58**

Further, in order to visualize the activations, we choose four images from the KITTI and ScanNet datasets, and draw the activations of four AFA blocks, as shown in Figure 6. For clarity, only one hundred activations of each block are visualized. We make three observations about the attention mechanism in robust depth prediction. First, we find that the activations of four scenes are more different in the high-level block than in the low-level one, which suggests that the values of each channel in the high-level features are scene-specific. Second, as per Figure 6(e), we observe a significant difference between indoor and outdoor activations. It indicates that the attention mechanism can give different scenes with different activations based on the characteristics and the layout of the scene. Third, by comparing the activations of scene1 and scene2 in Figure 6(e), we can also observe a non-negligible difference of activations between the two indoor scenes, which suggests that the attention mechanism still has a strong discriminating ability when processing the scenes from the same dataset.

Therefore, we believe that the channel-wise attention mechanism plays a vital role in choosing discriminative features for diverse scenes and improving the performance in robust depth prediction.

6 Conclusion

In this paper, we study the task of robust depth prediction that requires a model suitable for both indoor and outdoor scenes with a single parameter set. Unlike conventional depth prediction tasks, robust depth prediction task needs the model to extract more discriminative features for diverse scenes and to tackle the large differences in depth ranges between indoor and outdoor scenes. To this end, we proposed a deep attention-based classification network to learn a universal RGB-to-depth mapping which is suitable for both indoor and outdoor scenes, where a channel-wise attention mechanism is employed to update the features according to the importance of each channel. Experimental results on both indoor and outdoor datasets demonstrate the effectiveness of our method. Specifically, we won the 2nd place in the single image depth prediction entry of ROB 2018, in conjunction with CVPR 2018. In the future, we plan to extend our method to other dense prediction tasks.

References

1. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems. (2014) 2366–2374
2. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE transactions on pattern analysis and machine intelligence **38** (2016) 2024–2039
3. Xian, K., Shen, C., Cao, Z., Lu, H., Xiao, Y., Li, R., Luo, Z.: Monocular relative depth perception with web stereo data supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 311–320

4. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European Conference on Computer Vision, Springer (2012) 746–760
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE. (2017)
6. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32** (2013) 1231–1237
7. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241
8. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. arXiv preprint arXiv:1611.06612 (2016)
9. Bo Li, Yuchao Dai, M.H.: Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference. (2018)
10. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE (2016) 239–248
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
12. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: Proceedings of CVPR. (2017)
13. Cao, Y., Wu, Z., Shen, C.: Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Transactions on Circuits and Systems for Video Technology (2017)
14. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: Advances in Neural Information Processing Systems. (2016) 730–738
15. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507 (2017)
16. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. arXiv preprint arXiv:1704.06904 (2017)
17. Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., Ricci, E.: Structured attention guided convolutional neural fields for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 3917–3925
18. Kong, S., Fowlkes, C.: Pixel-wise attentional gating for parsimonious pixel labeling. arXiv preprint arXiv:1805.01556 (2018)
19. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 3640–3649
20. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Learning a discriminative feature network for semantic segmentation. (2018)
21. Kim, Y., Denton, C., Hoang, L., Rush, A.M.: Structured attention networks. arXiv preprint arXiv:1702.00887 (2017)
22. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE (2017) 5987–5995

23. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: IEEE conference on computer vision and pattern recognition (CVPR). Volume 1. (2017) 3
24. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant cnns. In: International Conference on 3D Vision (3DV). (2017)
25. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. In: ACM Transactions on Graphics (ToG). Volume 23., ACM (2004) 689–694
26. Vedaldi, A., Lenc, K.: Matconvnet – convolutional neural networks for matlab. In: Proceeding of the ACM Int. Conf. on Multimedia. (2015)
27. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 1–9
28. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2002–2011