Replacing LIDAR: 3D Object Classification using Single Image Depth Estimation

Benjamin Göing, Lars Jebe

Introduction

3D object classification is an important task that appears across many industries, such as autonomous driving or real-time surveillance or agriculture. The most advanced 3D object classification systems rely heavily on RGBD-data, usually captured with a camera and a LIDAR. Recent advances in depth estimation from single images allow estimation of an RGBD image from a single RGB image. In this project we want to explore how we can use an estimate of depth, rather than actively captured geometry, to support 3D object classification.

Related Work

Many papers discuss 3D object detection and localization and evaluate their performance on the KITTY dataset. [1][2][3] use the associated LIDAR data and obtain an object detection AP in between 55% and 80%. Others use the RGB image only [5][6], and obtain an object detection AP of <15%. This large gap is due to the missing depth estimation. The problem of estimating depth from a single image has been explored by various groups [6][7][8]. Their networks often work very well when trained on a homogenous dataset, such as KITTI or NYUv2. Many research groups have tackled this problem, so that a comprehensive overview over CNN-based state-of-the-art methods was necessary [9].

Approach

We want to use the model from [1] (Frustrum PointNet 3D) for object classification and from [6] (FCRN) for single image depth estimation and investigate the effects of replacing the LIDAR-depth data with the estimated data. We further want explore how changes to the depth channel (e.g. adding noise / blurring etc.) impacts the performance of [1]. This might lead to insights about which aspects of the depth maps are important for object classification and where accuracy can potentially be compromised. Most existing code is in TensorFlow, so we are most likely going to use that.

Milestones

1. Setup: Get Frustrum PointNet 3D to run and create predictions on the KITTY dataset, make use of existing code and pretrained weights.
2. Setup2: Get [6] to run and to predict depth on the KITTY dataset, also making use of existing code and pretrained weights.
3. Add different perturbations to the depth channel of the input to the Frustrum PointNet 3D, observing the effects on the detection AP.
4. Use predictions from FCRN [6] instead of the LIDAR depth for the Frustrum PointNet 3D and observe effects.

5. (Bonus, if time allows) Think about training [1] and [6] end-to-end, finetuning
6. (Bonus, if time allows) Using an appropriate prior, enhance the depth maps resulting from [6] in a post-processing optimization step (They look a bit blurry, so there might be room for improvement)
7. (Bonus, if time allows) Use insights about which aspects of depth maps are important and re-train / finetune [6] based on those insights. Observe results on detection AP.

References

[1] Qi, C. R., Liu, W., Wu, C., Su, H., & Guibas, L. J. (2018). Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 918-927).

[2] Zhou, Y., & Tuzel, O. (2018). Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4490-4499).

[3] Chen, X., Ma, H., Wan, J., Li, B., & Xia, T. (2017, July). Multi-view 3d object detection network for autonomous driving. In *IEEE CVPR* (Vol. 1, No. 2, p. 3).

[4] Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., & Urtasun, R. (2016). Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2147-2156).

[5] Tung, F., & Little, J. J. (2017, May). MF3D: Model-free 3D semantic scene parsing. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on* (pp. 4596-4603). IEEE.

[6] Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016, October). Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on* (pp. 239-248). IEEE.

[7] Li, J., Klein, R., & Yao, A. (2017, October). A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy* (pp. 22-29).

[8] Liu, F., Shen, C., & Lin, G. (2015). Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5162-5170).

[9] Koch, T., Liebel, L., Fraundorfer, F., & Körner, M. (2018). Evaluation of CNN-based single-image depth estimation methods. *arXiv preprint arXiv:1805.01328*.