Multi-Modal Depth Estimation with CNN

Yuxiao Chen yuxiaoc@stanford.edu

Qingxi Meng qingxi@stanford.edu

Motivation

Depth estimation has become critical for autonomous driving to ensure safety maneuvers [1]. Previous work [2] shows that monocular images have already achieved promising results for depth estimation. However, in general, the monocular depth estimation is an ill-posed problem because it lacks reliable stereoscopic relationships.

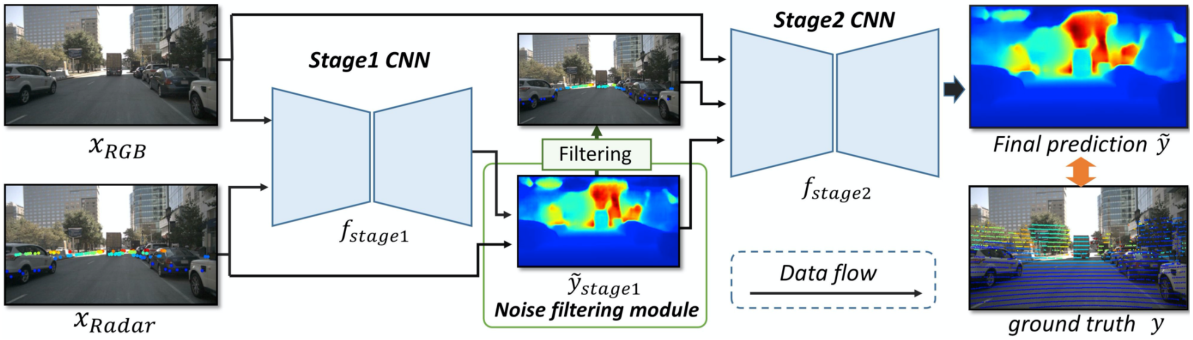
Although partial stereoscopic information could be inferred by monocular cues such as texture, occlusion and non-lambertian surfaces, it is much more computationally efficient to directly extract it from other signals [3]. In addition to monocular images, today more and more self-driving cars are equipped with radar and lidar sensors. One big advantage of radar and lidar signals is that they could perform well in harsh environments, which gives coarse but much more accurate measurements. The purpose of this project is to integrate the monocular images with radar signals to yield better resolution for the depth estimation.

Technical Overview

I plan to use Nuscenes dataset, which is one of the first datasets which features Camera, Radar, and LiDAR recordings in diverse scenes and weather conditions. I plan to use the RGB band, and implement a band pass filter to get a sparse radar band, to feed into a CNN model for generating the depth estimation. For the band pass filter, my initial plan is to use an existing package with well-tuned parameters. I could also try to treat the band pass filter as part of training parameters and dynamically learn them along the training. The band pass filter could be trained in either time domain or frequency domain. More advanced transforms like discrete cosine transform (DCT) could be also explored.

One potential challenge is that the sparse radar signals alone may not be adequate to help extract the stereoscopic information. Therefore, I plan to also use the measurements from lidar or use data augmentation techniques to have more densely sampled radar signals. Another potential challenge is to find a model to combine the inputs from different modes correctly. One idea is to train embedding layers for the image input and the radar signal input respectively, and then concatenate the embeddings into one input to the CNN model.

For baseline, I plan to use the RGB band only to train on traditional CNN. Then I will add the sparse radar band to retrain the CNN with a multimodal input. Then I will try what this work [4] proposes, which is to add a noise filter after the first CNN and then add another CNN model that intakes a multimodal input of RGB, radar, and the denoised depth prediction. For the noise filtering model, instead of what the work originally proposes, which is to use an adaptive threshold, I plan to try either bilateral filter or non-local means. The parameters for those filters could be pre-set as heuristics or trained dynamically during the training.



I will divide the dataset into a training set, validation set and test set. Since I will have a huge amount of data, I will use 80% of the original dataset as a training set, 10% of the original dataset as a validation set, and 10% of the original dataset as a testing set.

To evaluate the result, I will compare the predicted depth map to the ground truth depth map. I will use several different metrics to compare the quality of the predicted depth map. One metric I will use is to use the signal-to-noise ratio (SNR) to measure the quality of the depth map. I also plan to use root mean square error (RMS) in both the linear domain and log domain. Furthermore, I will also calculate both the square relative difference and absolute relative difference between the predicted depth map and the ground true depth map.

Milestones

2/20 Branch from https://github.com/brade31919/radar_depth repo, setup dataset
2/23 Apply band pass filter to extract a sparse radar band
2/26 Train Baseline (RGB band with CNN)
2/29 Train with multimodal input
3/1 Noise filtering on the trained depth map (try bilateral or non-local means)
3/5 Train second CNN with denoised depth map
3/8 Analyze result with ground truth depth map

Reference

[1] Godard, C., Mac Aodha, O., & Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 270-279).

[2] Godard, C., Mac Aodha, O., Firman, M., & Brostow, G. J. (2019, Oktober). Digging Into Self-Supervised Monocular Depth Estimation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

[3] Bhoi, A. (2019). Monocular depth estimation: A survey. arXiv preprint arXiv:1901.09402.

[4] Lin, J.T., Dai, D., & Van Gool, L. (2020). Depth Estimation from Monocular Images and Sparse Radar Data. *In International Conference on Intelligent Robots and Systems (IROS)*.