Semantic-MD: Infusing Monocular Depth with Semantic signals

3D Vision Project Proposal
Supervised by: Zuria Bauer, Mihai Dusmanu
March 10, 2023

GROUP MEMBERS

Alice Mazzoleni | Rolando Grave de Peralta Gonzalez | Kush Prasad

I. DESCRIPTION OF THE PROJECT

Monocular depth prediction is an important field in computer vision with applications in scene understanding, robot localization, augmented reality and other fields and areas. The objective of our project is to improve upon state-of-the-art monocular depth prediction by infusing semantic signals as part of the input. We will be using the semantic classes output from a deep learning model such as Mask R-CNN [3]. We will use the semantic classes output as an additional channel input and as an input to the loss function. We will be exploring different loss functions during training which make use of the semantic input. As a final step we will explore making modifications to an existing state-of-the-art network such as Binsformer [4], DwinFormer [6] or URCDC [7]. We will finally evaluate our model using the KITTI [2] and NYU Depth V2 [5] datasets.

II. WORK PACKAGES AND TIMELINE

The work packages will consist of the following steps.

1) As a first step we will explore the models from the following state-of-the-art performing papers on depth estimation [4], [6], [7]. We will choose one of the architectures as our base architecture depending on the general model complexity and the hardware required for training. If none of the models is feasible due to our hardware constraints we will use a less sophisticated architecture.
   We expect this to be completed by March 20, 2023.
2) As a second step we will train the model on KITTI and NYU Depth-V2, if the trained models are not already available. We will evaluate and note the performance of the model on the mentioned datasets. We need to pay attention to the format of the KITTI and NYU Depth-V2 data, as the depth data may not be aligned to the RGB image and there might be missing values. We will pretty much follow the procedure outlined in paper [1] to overcome those challenges.
   NOTE: For the purpose of training a model we will be using ETH's Euler cluster, with a 12GB Nvidia 2080Ti graphics card. Depending on the load we will be able to adjust the amount of cores and ram needed to load the datasets and train the model. We will be using PyTorch to code our models.
   We expect to do this step by April 3, 2023.
3) As we will need segmentation masks for all data, we will evaluate the Mask R-CNN network [3] on KITTI and NYU Depth-V2 if a trained model is already available. If the trained model is not available, we will train it. If the performance of the segmentation network is not up to the mark, we will consider other alternatives. We will use the output of Mask R-CNN as our segmentation masks for further steps.
   We expect to do this by April 3, 2023.
4) As a next step we will experiment with the following ways to infuse semantic signals into our model. We might also combine these ideas.
   - To begin with, we will train our network using loss-functions used in state-of-the-art models for depth prediction. This step should have been mostly done in step 2.
   - We will use the segmentation mask as an additional input channel to our base network.
   - We will incorporate the output from segmentation model in our training loss function. We will explore different loss functions which make use of the segmentation output. Neighbouring pixels are expected

to have similar depth values unless they belong to different objects. We will thus come up with a loss function that incorporates this by making use of the segmentation mask input by penalizing neighbouring pixels of similar depth but belonging to different object classes. Although two distinct cars belong to the same class they can have completely different depth values. We will thus form connected components of the segmentation mask and then provide them as an input to our model/algorithm.
- Also we would like to make sure that the overall error is scale-invariant as outlined in [1].

We expect to do this by April 24, 2023.

5) After completing the above-mentioned steps, we will explore making modifications to the backbone network and also explore pre-training methods to improve the depth prediction performance.
We expect to do this till May 8, 2023, depending upon how much time we have.
6) We will finally evaluate the model on KITTI, NYU Depth V2 datasets.
We will do this by May 19, 2023.

III. Outcomes and Demonstration

The expected final product will be a python script that will be able to take an input RGB image and output depth value for all pixels. The estimated depth value can be converted to grayscale to give qualitative visual results of the estimated depth. To compare the estimated depth against the ground truth we will use the scale-invariant error introduced in [1] as well as RMSE. We will present an ablation study of our proposed modifications to the back-bone model. The model will be compared to the results of our initial back-bone model. The consumer hardware we will be using has the following specs; CPU is an AMD 9 5900x, for the ram we have up to 32 GO Ram 3600mhz, and a RTX 3080 10Go graphics card.

References

[1] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *CoRR*, abs/1406.2283, 2014.

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[4] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation, 2022.

[5] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[6] Md Awsafur Rahman and Shaikh Anowarul Fattah. Dwinformer: Dual window transformers for end-to-end monocular depth estimation, 2023.

[7] Shuwei Shao, Zhongcai Pei, Weihai Chen, Ran Li, Zhong Liu, and Zhengguo Li. Urcdc-depth: Uncertainty rectified cross-distillation with cutflip for monocular depth estimation, 2023.