Winter school project: Monocular Depth Estimation via Transfer Learning

Supervisor: Yongbo Chen*

June 23, 2021

Problem description: The monocular depth estimation is the task of estimating a dense depth map for each pixel using a single monocular RGB image. It can be widely used in different areas, including simultaneous localization and mapping(SLAM) [1], navigation [2], object detection [3] and semantic segmentation [4], etc. This problem is often described as an ill-posed and inherently ambiguous problem. The accurate depth estimation is a very challenging problem by considering the fact that most scenes have large texture and structural variations, object occlusions, and rich geometric detailing. Currently, with the rapid development of deep neural networks, deep learning has been widely used in this problem and achieved promising performance in accuracy. As an example, the following image Fig. 1 shows the inputs and outputs of this problem:

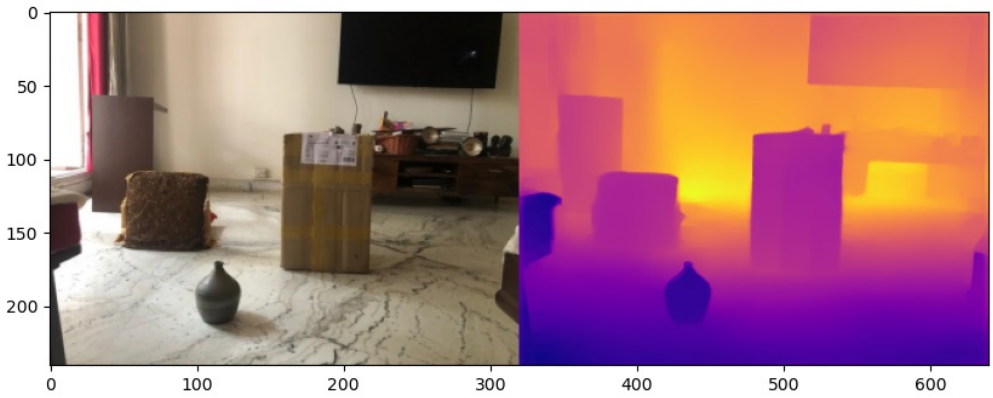

Figure 1: An example to show the monocular depth estimation

Purpose: This project is to introduce a deep learning-based method to solve monocular depth estimation. Based on this project, we would like to introduce the basic operations and solutions for this problem and further promote its application in the deformable SLAM.

Main reference: Paper: Alhashim I, Wonka P. High quality monocular depth estimation via transfer learning[J]. arXiv preprint arXiv:1812.11941, 2018. https://arxiv.org/abs/1812.11941

Code: Our deep learning tool is based on Tensorflow 2.0. It will use the trained well-known network Desenet-169. https://github.com/cyb1212/Monocular-Depth-Estimation-vis-Transfer-learning-Winter-school-.git

Dataset: The dataset is from NYU Depth Dataset V2, cited: Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from RGB-D images[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2012: 746-760. We offer a link directly used in our provided code: https://studentutsedu-my.sharepoint.com/:u:/g/personal/12514586_student_uts_edu_au/EU2J4152eDpMr7HXNvpvfSIB
wzfEl3CB6Q?e=OQ8OC9

Solution framework: The introduced framework is based on the transfer learning method and an encoder & decrder structure. Transfer learning is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. For our encoder, the input RGB image is encoded into a feature vector using the DenseNet-169 [5] network pre-trained on ImageNet [6]. This vector is then fed to a successive series of up-sampling layers, in order to construct the final depth map at half the input resolution. These upsampling layers and their associated skip-connections form our decoder. The decoder does not contain any Batch Normalization or other advanced layers recommended in recent state-of-the-art methods. The basic network architecture is shown in Fig. 2.

*Yongbo Chen is a postdoct research fellow surpervised by A.Prof Shoudong Huang at Faculty of Engineering and Information Technology, University of Technology, Sydney, Ultimo, NSW, 2007 Australia, e-mail: Yongbo.Chen@student.uts.edu.au, Shoudong.Huang@uts.edu.au.

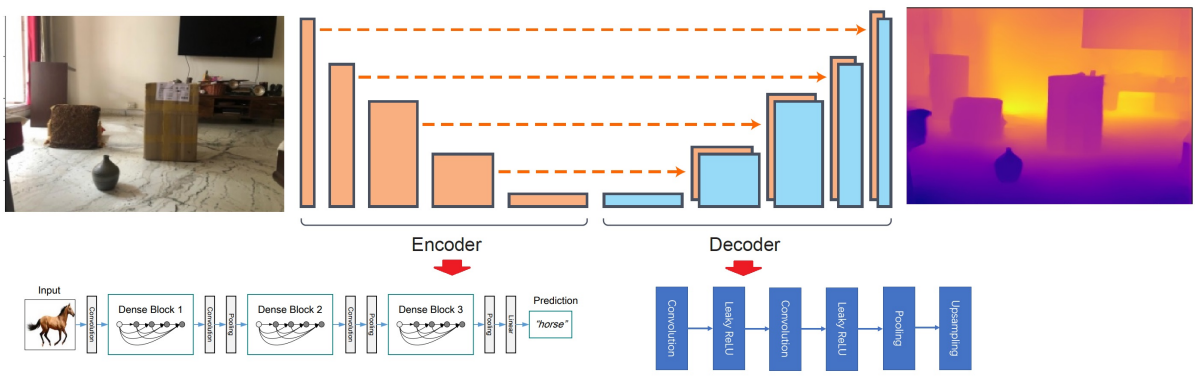

Figure 2: Overview of our network architecture

Project step: The participants will be asked to complete the following exploration steps based on the provided code, including:

- **Building Network** In our provided code, the encoder and decoder network has been deleted as a question for participants. You need to configure environment, revise the dataset folder, adjust parameters, and fill this gap. The used functions may include: "tensorflow.keras.applications.DenseNet169", "tf.keras.layers.Conv2D", "tf.keras.layers.UpSampling2D", "tf.keras.layers.Concatenate", "tf.keras.layers.LeakyReLU", and "tf.keras.Model". This step is the basic and most difficult step in this project.

- **Improving loss function** In our provided code, only the point-wise L1 loss defined on the depth values is offered. The participants are asked to add some other commonly used loss functions in this project and then combine them using a weighted method. The potential loss functions may include: (1) L1 loss defined over the image gradient of the depth image; (2)Structural Similarity; (3) Berhu Loss [7]; (4) the point-wise L2 loss defined on the depth values, and so on.

- **Adding data augmentation** The data provided in this project is a commonly used benchmark dataset. The image number is limited. In order to improve the network performance, the data augmentation is a good operation. The participants are asked to select several classical operations, including: flip (cv2.flip), rotation (cv2.rotate), scale (skimage.transform.rescale), crop (tf.random_crop), translation (tf.image.crop_to_bounding_box), gussian noise ($tf.random\_normal$, $tf.add$), and salt-and-pepper noise, in data augmentation for the image dataset.

- **Design new encoder and decorder network (optional)** Any new network following this encoder and decoder architecture with good performance is welcome. The well-known benchmark is https://paperswithcode.com/task/monocular-depth-estimation.

Final evluation: The marks for these four steps are respectively: Building Network (60), Improving loss function (20), Adding data augmentation (10), and Design new encoder and decoder network (10). The total marks are 100. Each part will be evaluated by combining the code and the final report.

References

[1] Hu G, Huang S, Zhao L, et al. A robust rgb-d slam algorithm[C]//2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2012: 1714-1719.

[2] de Queiroz Mendes R, Ribeiro E G, dos Santos Rosa N, et al. On deep learning techniques to boost monocular depth estimation for autonomous navigation[J]. Robotics and Autonomous Systems, 2021, 136: 103701.

[3] Chang J, Wetzstein G. Deep optics for monocular depth estimation and 3d object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 10193-10202.

[4] Zhang Z, Cui Z, Xu C, et al. Joint task-recursive learning for semantic segmentation and depth estimation[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 235-251.

[5] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2261–2269, 2017. 2, 3, 5, 11.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009. 3, 5

[7] Bhoi A. Monocular depth estimation: A survey[J]. arXiv preprint arXiv:1901.09402, 2019.