A traditional computer vision task, depth estimation (DI) involves predicting depth based on one or two-dimensional images. In this task, RGB pictures are taken as input and a depth map (representing the distance between the object and the camera viewpoint) will be generated by the algorithm [1]. The applications in scene understanding and 3D modeling, robotics, and autonomous driving, have made DI a hot research topic.

In the past, traditional methods (e.g., focus/defocus) were used for DI tasks [1]. These algorithms were then replaced by some automated algorithms such as SIFT and SURF with the advent of machine learning. In recent years, deep learning-based algorithms have been largely used to accomplish DI tasks[1]. DI can be classified into different subgroups. There are two main approaches to estimate depth in a scene depending on sensors used for data acquisition: active and passive. Active approaches use LIDAR sensors and RGB-D cameras to calculate Depth Information (DI). While in passive approaches, stereo and monocular images are used as two primary sources for achieving DI. Stereo DE extracts DI from stereo images by pixel matching among the pixels and rectifying both images. Monocular DE, on the other hand, does not require batches of rectified images. It needs a sequence of images from a specific camera. In recent years demand for MDE has increased since it requires less equipment than stereo DE.[1]

Monocular depth estimation is classified as an ill-posed and vague problem in computer vision [2]. The potential of deep learning in image classification and solving classical ill-posed problems advanced developments in MDE based on deep learning techniques [1].

Having a sufficient amount of data is a vital part of training a neural network. Numerous datasets have been prepared for indoor and outdoor DI. As an example, the KITTI dataset [3] is an outdoor road scene with a fixed camera position. An additional dataset is NYU Depth [4], which contains 464 indoor scenes. It accepts monocular video sequences of scenes, as well as the depth data from an RGB-D camera, unlike the KITTI dataset, which uses LIDAR to collect ground truth. Other famous datasets also can be utilized in the training section. We will analyze, contrast, and train our networks using these datasets as part of our research.

Throughout this project, we aim to take online courses in order to get hands-on experience on Deep Learning programming and implementation (October 30th), analyze different datasets appropriate for DI tasks and choose between them (October 10th), and obtain excellent accuracy by reconstructing results from previously published publications in the field of monocular depth estimation using convolutional neural networks (CNNs) (until November 10th). Achieving these short-term goals will pave the way for us to address still-remaining challenges of the field. Training deep NNs requires a large set of labeled data [2], which is often not available and costly to prepare. Leveraging unsupervised methods to train such networks with unlabeled data is our field of interest, that we hope we would be able to reach excellent accuracy by re-simulating related papers (and even proposing novel ideas) (by Dec 1st)

References:

[1] M. et al., "Monocular Depth Estimation Using Deep Learning: A Review," Sensors, 2022.

[2] A. Bhoi, "Monocular Depth Estimation: A Survey," Arxiv.

[3] Geiger, A. et al. Are we ready for autonomous driving? the Kitti vision benchmark suite. 2012 IEEE ICVPR

[4] Couprie et al. Indoor semantic segmentation using depth information. arXiv 2013, arXiv:1301.3572.