CSCI.731 - Project Proposal

Monocular Depth-Estimation for Autonomous Driving using Neural Networks

Proposed by: Bardh Rushiti
Mentored by: Dr. Thomas Kinsman

Background

Monocular depth estimation is the process of estimating the depth of every pixel in a 2-dimensional image captured by a camera. It is a new and important research topic in computer vision, with numerous applications in autonomous driving, robotics, and augmented reality. In recent years there has been significant development in this field of depth estimation and several approaches have been proposed.

One of the earliest approaches to depth estimation was based on stereo matching, where the depth is estimated by finding the corresponding pixel of interest in a stereo pair of images, and given the distance between the cameras is known, the distance between the object and the camera can be calculated. Nonetheless, this approach requires two cameras which have limited practicality in real-world applications. To overcome this problem, researchers developed monocular depth estimation methods which rely on the predictive capabilities of machine learning algorithms.

Current research is mainly focused on utilizing deep learning approaches to achieve depth estimation, with lower computational resources and energy consumption. These approaches can be categorized into supervised, semi-supervised, and unsupervised. Supervised learning methods follow the $\mathbf{x} \rightarrow \mathbf{y}$ format, where the method uses a single image **x** and it's corresponding depth information **y** for training. However, for this approach requires a large amount of data to be able to accurately generalize on the depth estimation problem. To overcome this obstacle, a number of semi-suppervised approaches have been suggested [1, ,2, 3]. By nature, this kind of approach requires less amount of labeled data and a larger amount of unlabelled data for training. However, the limitation of this methods is that the networks is not able to overcome their biases and require more information about camera focal length and sensor data. Last but not least, self-supervised approaches need a small amount of unlabeled data to train the networks for depth estimation, and make us of input modalities to estimate depth [4, 5, 6]. However, with this approach the corresponding models perform well in a limited set of scenarios, with similar features as training set.

For the following of this Background & Literature Review I will be focusing on top 5 papers ranked (starting with the highest number of implementations) by [7] for the Monocular Depth Estimation, which denotes the most implemented papers on GitHub for the topic.

[8] present a deep learning approach to monocular depth estimation that achieves state-of-the-art results on benchmark datasets. The authors propose a transfer learning strategy where a pre-trained network is

fine-tuned on a small dataset of stereo pairs to learn depth from a single input image. The proposed method is able to generate high-quality depth maps with sharp edges and fine details, while being computationally efficient and requiring little memory.

[9] present an unsupervised learning approach for monocular depth estimation that uses left-right consistency to enforce geometric constraints between the estimated depth and image geometry. The authors propose a deep learning architecture that takes a single image as input and produces a dense depth map, without requiring any ground truth depth information for training. The key idea of their method is to use left-right consistency to ensure that the depth estimates are consistent across pairs of stereo images, even when no ground truth depth information is available.

[10] propose a transformer-based approach for dense prediction tasks, such as semantic segmentation and monocular depth estimation. The authors demonstrate that the transformer architecture can effectively capture global context and long-range dependencies in image data, outperforming previous state-of-the-art methods on benchmark datasets for dense prediction tasks. The authors also provide insights into the performance trade-offs of different transformer architectures and demonstrate the effectiveness of a hybrid architecture that combines transformers with convolutional neural networks.

[11] addresses the challenge of training monocular depth estimation models that can generalize to new environments and datasets. The authors propose a method for combining multiple datasets during training, enabling the model to learn from a larger and more diverse set of environments. The resulting model can then be fine-tuned on a new dataset without requiring any labeled depth data, a process known as zero-shot transfer. The authors show that their approach outperforms previous state-of-the-art methods on benchmark datasets for monocular depth estimation, demonstrating the importance of robustness and generalization in this task.

[12] proposes a self-supervised approach to monocular depth estimation that does not require any manual annotation of depth information in the training data. The method involves training a neural network to predict the depth of an image from a single viewpoint, based on the assumption that the image contains sufficient information for depth estimation. The authors show that their approach outperforms previous self-supervised methods on benchmark datasets for monocular depth estimation, achieving state-of-the-art results.

Datasets

The depth estimation problem benefits from certain datasets that provide images and depth maps from multiple viewpoints. The subsequent section focuses on well-known datasets that are used to analyze various scenes. The ground truth depth images for these datasets are usually captured using consumer-level sensors such as the Kinect and Velodyne laser scanner. A summary of the datasets can be found in Table 1.

Table 1 lists the datasets commonly used for monocular depth estimation, including:

- NYU-v2: This dataset contains 1449 densely labelled RGB images with corresponding depth images, captured from three different cities. It is primarily used for indoor scene depth estimation, segmentation, and classification and comprises 407K frames of 464 scenes.
- Make3D: This dataset contains 400 training and 134 testing outdoor images, as well as indoor and synthetic scenes, which are used for depth estimation. It presents a more complex set of features.
- KITTI: This dataset comprises 394 road scenes with RGB stereo sets and corresponding ground truth depth maps. It has several versions and is divided into RD, CD, SD, ES, and ID. The Velodyne laser scanner is used to capture high-quality ground truth images. It is commonly used for various tasks, including 3D object detection and depth estimation.
- Pandora: This dataset contains 250K full-resolution RGB and corresponding depth images with annotations. It is primarily used for head center localization, head pose estimation, and shoulder pose estimation.
- SceneFlow: One of the first large-scale synthetic datasets introduced, SceneFlow includes 39K stereo images with corresponding disparity, depth, optical flow, and segmentation masks.

References

[1] Kuznietsov, Y.; Stückler, J.; Leibe, B. Semi-Supervised Deep Learning for Monocular Depth Map Prediction. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2215–2223.

[2] Chen, Y.; Zhao, H.; Hu, Z. Attention-based context aggregation network for monocular depth estimation. arXiv Prepr. **2019**, arXiv:1901.10137.

[3] Alhashim, I.; Wonka, P. High quality monocular depth estimation via transfer learning. arXiv Prepr. **2018**, arXiv:1812.11941.

[4] Godard, C.; Aodha, O.M.; Brostow, G.J. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6602–6611.

[5] Godard, C.; Mac Aodha, O.; Firman, M.; Brostow, G.J. Digging into self-supervised monocular depth estimation. In Proceedings of the 2019 IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 3828–3838.

[6] Yin, W.; Liu, Y.; Shen, C.; Yan, Y. Enforcing geometric constraints of virtual normal for depth prediction. In Proceedings of the 2019 IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 5684–5693.

[7] Papers with Code. Monocular Depth Estimation. Retrieved February 21, 2023, from https://paperswithcode.com/task/monocular-depth-estimation#papers-list.

[8] Alhashim, I., & Wonka, P. (2018). High Quality Monocular Depth Estimation via Transfer Learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018-June, 567–576.

[9] Godard, C., Mac Aodha, O., & Brostow, G. J. (2017). Unsupervised Monocular Depth Estimation with Left-Right Consistency. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6602-6611.

[10] Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision Transformers for Dense Prediction. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 11322-11331.

[11] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2019). Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2624-2634.

[12] Godard, C., Mac Aodha, O., Firman, M., & Brostow, G. J. (2019). Digging Into Self-Supervised Monocular Depth Estimation. Proceedings of the IEEE International Conference on Computer Vision, 3828-3838.