Monocular Depth Estimation for Robotics Applications

Keivalya Bhartendu Pandya pandya.kei
Gaurav Kothamachu Harish kothamachuharish.g
Rachel Lim lim.rac

Abstract

Depth estimation is a critical task in robotics for navigation, obstacle avoidance, and object interaction. This project focuses on developing a deep learning model for monocular depth estimation, where depth is predicted from a single RGB image without requiring stereo cameras or LiDAR. Contemporary state-of-the-art methods typically fall into two distinct categories: designing complex networks capable of direct depth map regression, or partitioning the input into bins or windows to reduce computational complexity. This project adopts the latter approach. The primary objective is to implement and conduct comparative analyses of CNN-based and Transformer-based architectures for depth estimation, optimizing for both accuracy and computational efficiency, with particular emphasis on edge device deployment. The proposed methodology includes training and evaluating models on established benchmark datasets, specifically NYU Depth v2 and KITTI Depth. Performance evaluation will be conducted using standardized metrics, including Root Mean Square Error (RMSE) and Absolute Relative Error. Furthermore, this research will investigate lightweight model optimizations suitable for real-time robotic applications.

1 Literature survey

Eigen et al. [2014] introduces a CNN-based approach for monocular depth estimation using multiscale feature extraction. Godard et al. [2017] proposes a self-supervised learning method using stereo image pairs without explicit depth labels. Hu et al. [2019] uses an ordinal regression approach for higher-resolution depth maps with better object boundary preservation. Ranftl et al. [2021] explores Vision Transformers for depth estimation, improving generalization across different scenes.

2 Datasets

NYU Depth Dataset V2, from Nathan Silberman and Fergus [2012], includes video sequences from indoor scenes with depth information from the Microsoft Kinect. KITTI includes stereo camera and LiDAR data from their autonomous driving platform Geiger et al. [2013]. Song et al. [2015] introduces SUN RGB-D, with Asus Xtion, Kinect v1, and Kinect v2 data.

This challenge has already been addressed by various research groups employing diverse methodologies. While some approaches prioritize accuracy optimization, others focus on reducing computational complexity. As previously stated in the *Abstract*, the primary objective of this research is to develop a computationally efficient model while subsequently optimizing its accuracy performance.

Submitted to Prof. Huaizu Jiang as the CS-7150: Deep Learning Course Project Proposal.

3 Plan of activities

The project is structured into four distinct phases. Phase 1 encompasses dataset selection, preprocessing, and the establishment of the deep learning framework and environment. Phase 2, the model development phase, focuses on implementing a CNN-based baseline model, conducting experiments with transformer-based architectures, and optimizing these models for efficient real-time inference. Phase 3 comprises model training and evaluation, wherein the developed models are trained on selected datasets and evaluated using standardized metrics (Root Mean Square Error, Absolute Relative Error, and log-scale errors), followed by comparative analysis against established depth-estimation benchmarks. Phase 4 consists of comprehensive ablation studies and final model selection, culminating in detailed documentation of the methodology, technical challenges, experimental results, and analytical discussions, accompanied by a formal presentation of findings.

3.1 Work Distribution

Phase	Keivalya P.	Rachel L.	Gaurav H.
Phase 1	Framework setup	Dataset preprocessing	Data pipeline
Phase 2	Optimization Techniques	Transformer architecture	CNN implementation
Phase 3	Metrics implementation	Training pipeline	Benchmark comparison
Phase 4	Results documentation	Ablation studies	Final presentation

Table 1: Work Distribution

References

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2366–2374, Cambridge, MA, USA, 2014. MIT Press.

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

Clement Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

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

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, October 2021.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015. doi: 10.1109/CVPR.2015.7298655.