Fraunhofer IML

Master Thesis Exposé

A Monocular Vision Architecture for Free Volume Estimation in Shipping Containers

Ashwin Nedungadi

Academic Supervisors
Prof. Dr.-Ing. Alice Kirchheim
Christopher Rest, M.Sc.
Jonas Stenzel, M.Sc.

FLW Technische Universität Dortmund
Fraunhofer IML
Department of Robotics & Cognitive Systems

May 2024

Introduction

Accurately estimating point clouds from a single image using monocular depth estimation is a crucial challenge in computer vision with numerous downstream applications in robotics, autonomous navigation, augmented reality, and 3D reconstruction [6][7][8][10]. Recent advancements in deep learning [5] have enabled the rapid recovery of 3D information from 2D images, eliminating the need for expensive sensors. The aim of this research is to leverage learning-based metric monocular depth estimation for 3D reconstruction [2][3][4] and propose a novel monocular vision-based architecture designed explicitly for estimating the free volume of loaded shipping containers at logistical warehouses worldwide and validated at our industrial partner DB Schenker. The architecture aims to calculate a container's free volume using only RGB images captured from mobile devices [8], promoting a cost-effective and scalable approach to cargo optimisation and warehouse management.

Problem Statement

Optimising the loading of cargo trucks and shipping containers is a fundamental aspect of logistics with far-reaching consequences for sustainability and the economy [12][13]. Efficient loading directly impacts numerous factors [12] within the supply chain, helping to maximise overall efficiency and making it an area of significant research. Computer vision techniques offer low-cost and scalable advantages compared to measurements done manually or with the help of specialised instruments. Analysing images taken from a mobile device with a system that allows for precise volume measurement has the potential to improve loading configurations and generate insights into space utilisation.

This research addresses the challenge of improving container loading efficiency at logistics warehouses by achieving accurate metric 3D reconstruction using monocular depth estimation [5]. This is done by estimating the “utilised” or free volume of shipping containers after reconstructing a watertight mesh of the container [11]. Inherently, an ill-posed problem, 3D reconstruction using monocular depth estimation, has recently matured through the use of robust Transformer [9][16] and ConvNet [2][17] models in computer vision tasks. However, reconstructing 3D point clouds from a single viewpoint and accurately estimating their volume remains challenging for several reasons, such as the ambiguity of depth perception, scale ambiguity, occlusion, lighting and partial point clouds leading to lower accuracies in real-world environments [18].
We aim to leverage existing state-of-the-art metric monocular depth estimation models such as Metric3D [2], PixelFormer [3] and ZoeDepth [4] and fine-tune them on domain-specific data from a high-precision Navvis 3D scanner [15] and data obtained from an iOS device with LiDAR sensor [14]. Moreover, we investigate possible improvements to the architecture of these models and implement appropriate volume estimation methods to demonstrate the feasibility and advantages of using monocular depth estimation for cargo free-volume estimation with 3D reconstruction while utilising the latest deep learning techniques and domain-specific knowledge to enhance operational efficiency for our industrial partner DB Schenker.

State of the Art

Monocular Depth Estimation and 3D Reconstruction from images are active areas in computer vision and have been widely researched with a plethora of methods in existing literature [1][5]. While previous methods attempted to solve the reconstruction problem using feature-based triangulation from multiple images (SfM, MVS) and stereo-based methods that rely on epipolar geometry to estimate the depth and calculate points in 3D space [19][20][21], deep learning approaches have revolutionised this field by being able to identify, encode and abstract image features and learn non-linear 2D-3D relationships, achieving impressive depth prediction accuracies [2][3][5].

Several key factors categorise deep learning approaches to monocular depth estimation and 3D reconstruction. First, the **input image modality** can be single-view [1][22][23], utilising a single image for depth prediction, which is inherently an ill-posed “inverse” problem, or multi-view[24][25], combining information from multiple images of the same scene for improved accuracy using geometric priors. The **training strategy** can be supervised [5][2], requiring paired images with corresponding depth map ground truth, or self-supervised [26][27][28], learning depth from unlabeled images through tasks like image pair reconstruction, overcoming the need for ground truth depth [6][7]. The **predicted depth** can be relative [4], estimating the relative depth between pixels without a specific scale, or absolute [2], directly predicting the actual distance of each pixel from the camera [2][4]. Finally, the **output representation** of the final reconstruction can be surface-based, generating 3D surfaces through meshes, volumetric, representing the scene as a grid of voxels with density information, or implicit, utilising techniques like Neural Radiance Fields for a compact and flexible representation [29][30]. Understanding these categories and their advantages is crucial for selecting the most suitable deep learning approach for various applications in 3D reconstruction.

Research Objective

The primary objective of this research is to evaluate the feasibility of learning-based monocular depth estimation for the practical task of free volume estimation in logistics and to assess the associated advantages and disadvantages. The secondary objective is to optimise the overall system by incorporating architectural improvements based on current state-of-the-art literature. These objectives can be achieved by dividing the project into three main components:

1. 3D Reconstruction & Volume Estimation: Reconstruct a watertight mesh of the interior of shipping containers through monocular depth estimation and 3D reconstruction using images taken from a mobile device and estimate its free volume.
2. Data Collection & Fine-Tuning: Collect domain-specific data at DB Schenker using a Navvis 3D Scanner [15] and an iOS device equipped with a LiDAR sensor [14] to fine-tune existing monocular depth estimation models.
3. Evaluation & Benchmarking: Choose appropriate metrics and evaluate the models before and after fine-tuning. Benchmarking will measure the accuracy of the predicted depth maps and the final volume calculated from the 3D reconstructed container.

Methodology

The methodology will be structured into three primary components: *3D reconstruction and volume estimation*, *data collection and fine-tuning*, and *evaluation and benchmarking*. Initially, high-resolution 3D scans from DB Schenker will be utilized to generate ground truth depth maps and RGB images to benchmark various monocular depth estimation models for zero-shot performance. Additional data will be collected using an iOS device with LiDAR to capture the depth map as ground truth for container scenes at Fraunhofer IML. The collected data will be benchmarked using state-of-the-art monocular depth estimation models such as Metric3D [2], UniDepth [31], PixelFormer [3] and ZoeDepth [4]. The depth maps will then be projected to a point cloud to reconstruct a watertight 3D mesh of the container interior, enabling precise volume estimation through surface reconstruction algorithms.
Concurrently, a comprehensive dataset will be collected at DB Schenker warehouses using a Navvis 3D scanner [15] and a stereo camera to provide additional data for training the models on domain-specific data. This dataset will undergo preprocessing for alignment and normalisation, followed by augmentation to enhance model robustness.

The collected data will be used to fine-tune the monocular depth estimation models, leveraging transfer learning techniques for task-specific adaptation. Hyperparameter optimisation and architectural improvements will be considered to improve model performance and robustness. Evaluation of the models will be performed using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Intersection over Union (IoU) and F1 Score, to assess depth prediction accuracy and volumetric precision [32][33]. The fine-tuned models will be benchmarked against baseline and state-of-the-art approaches to evaluate real-time performance and scalability. This comprehensive methodology aims to demonstrate the feasibility and advantages of using monocular depth estimation for cargo free-volume estimation, enhancing operational efficiency in warehouse logistics using real-world validation at DB Schenker warehouses in cooperation with Fraunhofer IML.

Bibliography

[1] Wei Yin et al. Learning to Recover 3D Scene Shape from a Single Image. 2020. arXiv:2012.09365 [cs.CV].

[2] Wei Yin et al. Metric3D: Towards Zero-shot Metric 3D Prediction from A Single Image. 2023. arXiv: 2307.10984 [cs.CV].

[3] A. Agarwal and C. Arora, 'Attention Attention Everywhere: Monocular Depth Prediction With Skip Attention', in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 5861–5870.

[4] Shariq Farooq Bhat et al. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. 2023. arXiv: 2302.12288 [cs.CV].

[5] David Eigen, Christian Puhrsch, and Rob Fergus. Depth Map Pre-diction from a Single Image using a Multi-Scale Deep Network. 2014. arXiv: 1406.2283 [cs.CV].

[6] Armin Masoumian et al. "Monocular Depth Estimation Using Deep Learning: A Review". In: Sensors 22.14 (2022). issn: 1424-8220. Doi: 10.3390/s22145353.

[7] Alican Mertan, Damien Jade Duff, and Gozde Unal. "Single image depth estimation: An overview". In: Digital Signal Processing 123 (Apr.2022), p. 103441. issn: 1051-2004. Doi: 10.1016/j.dsp.2022.103441.

[8] Ashkan Ganj et al. Mobile AR Depth Estimation: Challenges & Prospects – Extended Version. 2023. arXiv: 2310.14437 [cs.CV].

[9] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. 2021. arXiv: 2103.13413 [cs.CV].

[10] X. Dong, M. A. Garratt, S. G. Anavatti, and H. A. Abbass, 'Towards Real-Time Monocular Depth Estimation for Robotics: A Survey', IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 10, pp. 16940–16961, 2022.

[11] Wen-Chung Chang et al. “Object volume estimation based on 3D point cloud”. In: 2017 International Automatic Control Conference (CACS) (2017), pp. 1–5. CorpusID:46826758.

[12] Zhao, X., Bennell, J.A., Bektaş, T. and Dowsland, K. (2016), A comparative review of 3D container loading algorithms. Intl. Trans. in Op. Res., 23: 287-320.

[13] J. Xue and K. K. Lai, “Effective Methods for a Container Packing Operation,” Mathematical and Computer Modelling 25, no. 2 (January 1, 1997): 75–84.

[14] “Modeling Spatial Uncertainty for the iPad Pro Depth Sensor | ISIF,” isif.org. https://isif.org/media/modeling-spatial-uncertainty-ipad-pro-depth-sensor accessed May 23, 2024).

[15] NavVis VLX, “the most accurate SLAM-based laser scanner”, www.navvis.com. https://www.navvis.com/ppc-campaign/overview (accessed May 23, 2024).

[16] A. Vaswani et al., ‘Attention Is All You Need’, arXiv [cs.CL]. 2023.

[17] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, ‘A ConvNet for the 2020s’, CoRR, vol. abs/2201.03545, 2022.

[18] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, ‘Deep Learning for 3D Point Clouds: A Survey’, *CoRR*, vol. abs/1912.12033, 2019.

[19] O. Ozyesil, V. Voroninski, R. Basri, and A. Singer, ‘A Survey of Structure from Motion’, arXiv [cs.CV]. 2017.

[20] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein and R. Szeliski, "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 2006, pp. 519-528, doi: 10.1109/CVPR.2006.19.

[21] Peter Sturm. A historical survey of geometric computer vision. CAIP - 14th International Conference on Computer Analysis of Images and Patterns, Aug 2011, Seville, Spain. Pp.1-8, ⟨10.1007/978-3-642-23672-3_1⟩. ⟨hal-00644982⟩

[22] S. Szymanowicz, C. Rupprecht, and A. Vedaldi, ‘Splatter Image: Ultra-Fast Single-View 3D Reconstruction’, arXiv [cs.CV]. 2024.

[23] X. Yang, G. Lin, and L. Zhou, ‘Single-View 3D Mesh Reconstruction for Seen and Unseen Categories’, IEEE Transactions on Image Processing, vol. 32, pp. 3746–3758, 2023.

[24] D. Wang *et al.*, ‘Multi-view 3D Reconstruction with Transformer’, *CoRR*, vol. abs/2103.12957, 2021.

[25] C. Wen, Y. Zhang, Z. Li, and Y. Fu, ‘Pixel2Mesh++: Multi-View 3D Mesh Generation via Deformation’, CoRR, vol. abs/1908.01491, 2019.

[26] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, ‘Digging Into Self-Supervised Monocular Depth Estimation’, arXiv [cs.CV]. 2019.

[27] Z. Liu, R. Li, S. Shao, X. Wu, and W. Chen, ‘Self-Supervised Monocular Depth Estimation With Self-Reference Distillation and Disparity Offset Refinement’, IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 12, pp. 7565–7577, Dec. 2023.

[28] R. Marsal, F. Chabot, A. Loesch, W. Grolleau, and H. Sahbi, ‘MonoProb: Self-Supervised Monocular Depth Estimation with Interpretable Uncertainty’, arXiv [cs.CV]. 2023.

[29] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao, ‘NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video’, CoRR, vol. abs/2104.00681, 2021.

[30] J. Huang, Z. Gojcic, M. Atzmon, O. Litany, S. Fidler, and F. Williams, ‘Neural Kernel Surface Reconstruction’, arXiv [cs.CV]. 2023.

[31] L. Piccinelli et al., ‘UniDepth: Universal Monocular Metric Depth Estimation’, arXiv [cs.CV]. 2024.

[32] J. Spencer et al., ‘The Third Monocular Depth Estimation Challenge’, arXiv [cs.CV]. 2024.

[33] N. Padkan, P. Trybala, R. Battisti, F. Remondino, and C. Bergeret, ‘EVALUATING MONOCULAR DEPTH ESTIMATION METHODS’, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLVIII-1/W3-2023, pp. 137–144, 2023.

Appendix

Preliminary Table of Contents

Proposed Timeline

Tasks	Duration
Data Acquisition and Preprocessing (Weeks 1-5) ● Collect data from DB Schenker warehouse. ● Collect a dataset of RGB images of shipping containers from various angles and lighting conditions. ● Capture LiDAR ground truth data for the containers using high-precision LiDAR. ● Pre-process the images by converting the Omnicam camera model into pinhole camera mode and cleaning, correcting and augmenting the data.	4 Weeks
Model Selection and Training (Weeks 5-8) ● Select appropriate monocular depth estimation models based on prior knowledge and benchmarks. ● Fine-tune the chosen models on the preprocessed container image and LiDAR data.	3 Weeks
Evaluation and Benchmarking (Weeks 8-11) ● Evaluate the performance of the fine-tuned models in three stages using metrics like RMSE, MAE, PSNR, SSIM, IoU, and F1 score. ● Analyse the results and identify areas for improvement.	3 Weeks
Model Refinement (Weeks 11-14) ● Explore architecture improvements (e.g., Spatial Transformer Networks) to enhance robustness. ● Retrain and re-evaluate the models with the improvements.	3 Weeks
System Integration and Testing (Weeks 14-16) ● Integrate the entire pipeline into the proposed architecture and test on unseen data obtained from DB Schenker. ● Benchmarking on the existing dataset.	2 Weeks
Optional Thesis Tasks & Buffer (Weeks 16-18)	2 Weeks
Writing the Thesis & Defense (Weeks 18-25)	7 Weeks