Exploration for Lightweight Monocular Depth Estimation

Wenxuan Zhang

November 1, 2023

Abstract

To be completed..

1 Introduction

In recent years, the surge of interest in computer vision, augmented reality, and autonomous navigation systems has stimulated a growing demand for depth perception from images. Depth perception, the ability to judge the relative distances of objects from the observer, plays a pivotal role in many applications, from self-driving vehicles to robotics. While traditional methods relied on the use of expensive and bulky sensors like LiDAR or stereo camera systems to gauge depth information, the challenges posed by cost, power consumption, and form-factor constraints have driven researchers to seek lightweight and efficient alternatives.

Monocular Depth Estimation (MDE) emerges as a promising solution to this end. Unlike stereo vision systems, which infer depth based on parallax from two cameras, monocular depth estimation harnesses sophisticated algorithms to predict depth from a single image. This technique not only eliminates the need for additional hardware, making it a cost-effective approach, but also offers immense potential for applications where mounting multiple sensors may be impractical.

Monocular depth estimation refers to the task of predicting depth information from a single image. It's an active area of research in computer vision because while human beings (and many animals) can infer depth and three-dimensional structures from a single image due to our cognitive abilities and experience, computers usually need multiple viewpoints (like in stereo vision) to accurately infer depth. Hence, the inherent ambiguities in converting a 2D image to depth values without additional context makes it a challenging task in the 3D vision realm. Recent advances, especially in deep learning, have made significant strides in this area by training models on large datasets with depth annotations. Still, as previously mentioned, there are inherent challenges that make monocular depth estimation an active area of research in 3D computer vision. One of the challenge involved is the large amount of training data needed for the training process.

Self-supervised monocular depth estimation is an area of computer vision that deals with predicting depth from a single image, without requiring explicit ground truth depth data for training. Self-supervised learning is a paradigm in which the learning algorithm generates its supervision signal from the input data, often obviating the need for explicit external labels. While this approach has shown promising results in various domains, it also brings forth several challenges:

- Designing Effective Pseudo-Tasks: A significant part of self-supervised learning involves designing tasks (or pretext tasks) that force the model to learn useful features [5]. Crafting these tasks so they are neither too easy (which won't lead to meaningful feature learning) nor too hard (which may make convergence challenging) is non-trivial.

- Transfer Learning and Downstream Tasks: The ultimate goal of self-supervised learning is often to transfer the learned representations to a

downstream task [1]. Ensuring that the learned features are general enough to be useful for a broad range of tasks is challenging.

- Domain-Specific Challenges: In some domains or applications, the input data might not contain enough information for meaningful self-supervision. For instance, in extremely noisy environments, self-supervised signals might be too weak or ambiguous [6].

2 Related Work

Lite-Mono [9] discusses a self-supervised monocular depth estimation model with a hybrid CNN and Transformer architecture. The proposed model incorporates Consecutive Dilated Convolutions (CDC) modules to capture enhanced multi-scale local features and a Local-Global Features Interaction (LGFI) module to calculate the Multi-Head Self-Attention (MHSA) and encode global contexts into the features. The article also explores advanced architectures for depth estimation, such as ResNet, channel-wise attention modules, and feature modulation modules. The proposed model, called Lite-Mono, achieves good results in terms of model complexity, inference speed, and accuracy on the KITTI dataset. The generalization ability of the model is also validated on the Make3D dataset. Ablation studies are conducted to evaluate the importance of different design choices in the architecture.

[4] designs a lightweight monocular depth estimation on edge devices, where it develops a two-stage channel pruning method to, respectively, prune the encoder and decoder based on their characteristics. Extensive experiments show that their strategies are effective on different edge GPU devices, when input resolutions differ in outdoor or indoor scenes.

[3] introduces a lightweight, unsupervised neural network for monocular depth estimation using video sequences. Addressing the challenges of limited labeled data and resource-intensive models, the proposed network efficiently integrates features with fewer parameters. Tested on the KITTI dataset, it outperforms state-of-the-art models in speed and accuracy, particularly on low-end devices like the Raspberry Pi 3. This work paves the way for real-time depth prediction on cost-effective, embedded devices.

In the work by [7], a lightweight encoder-decoder model for monocular depth estimation is introduced, optimized for embedded systems. The distinct feature of this model is the Guided Upsampling Block (GUB), inspired by guided image filtering, which enhances depth map reconstruction. Employing multiple GUBs, the model surpasses competitors in accuracy on the NYU Depth V2 and KITTI datasets and achieves impressive frame rates on NVIDIA platforms.

In [7], the authors introduce a novel method for enhancing monocular depth estimation using 3D points as depth guidance. Distinct from traditional depth completion techniques, their approach excels with sparse, uneven point clouds, making it versatile to the 3D point source. Central to their success is a new multi-scale 3D point fusion network that's streamlined yet effective. Demonstrated on two depth estimation challenges, using structure-from-motion and LiDAR derived 3D points, the network competes with leading depth completion methods in accuracy, especially with minimal points, and stands out in compactness. It also surpasses several modern deep learning multi-view stereo and structure-from-motion techniques in both accuracy and efficiency.

Yet, existing depth estimation techniques, rooted in intricate neural networks, lack real-time efficiency on embedded platforms. Addressing this, the authors in [8] introduce an optimized encoder-decoder architecture, named FastDepth, enhanced with network pruning to curtail computational demands. Their approach, with a particular emphasis on a low-latency decoder, offers speeds vastly superior to existing methods while retaining comparable accuracy. Specifically, FastDepth boasts 178 fps on NVIDIA Jetson TX2 GPU and 27 fps on its CPU, consuming under 10 W. On the NYU Depth v2 dataset, it closely aligns with top-tier accuracy. The authors assert that their work represents the fastest, most efficient monocular depth estimation on an embedded system suitable for micro aerial vehicles.

3 Data

The KITTI dataset [2], a cornerstone in the realm of computer vision and autonomous driving research, was birthed from a collaboration between the Karlsruhe Institute of Technology and the Toyota Technological Institute at Chicago. This rich dataset offers a comprehensive suite of benchmarks aimed at various tasks pivotal for the progression of autonomous vehicles.

Data collection took place in the picturesque city of Karlsruhe, Germany. A station wagon equipped with cutting-edge sensors—including a Velodyne LiDAR and a mix of grayscale and color cameras—roamed through urban streets, highways, and rural sequences, capturing a diverse array of scenarios. This data forms the foundation upon which researchers worldwide have benchmarked their algorithms, especially those tailored for the challenges of autonomous driving.

The breadth of the KITTI dataset is what truly sets it apart. It serves as a testing ground for a plethora of tasks. From stereo vision, which assists in evaluating stereo and optical flow algorithms, to depth evaluation, which aids both monocular and stereo depth estimation methods—the KITTI dataset has it all. Moreover, its meticulously curated annotations for 3D object detection, 2D object tracking, road/lane detection, and even semantic segmentation underscore its comprehensive nature.

However, like all benchmarks, KITTI is not without its limitations. A significant challenge arises from its specific environmental bias. Given its focus on German streets and surroundings, there's a palpable geographical constraint. Models that are solely trained on the KITTI dataset might grapple with generalization when faced with diverse driving terrains from different corners of the globe.

In MDE realm, the KITTI dataset is also of great importance. Monocular Depth Estimation (MDE), the intricate task of discerning depth from a singular image, finds its proving ground in real-world datasets. Among these, the KITTI dataset stands out prominently. Originating from the streets of Karlsruhe, Germany, KITTI serves as a critical benchmark for a myriad of computer vision tasks, especially MDE.

What makes KITTI invaluable for MDE is its treasure trove of diverse driving scenarios. From bustling urban streets to tranquil rural settings and fast-paced highways, the dataset captures the multifaceted realities of driving. Researchers, using KITTI, are not merely testing their algorithms in sterile, synthetic environments but in dynamic, real-world situations. This realism, combined with the dataset's LiDAR-based depth maps, sets a rigorous standard for depth estimation models. With these depth maps acting as ground truth, the dataset challenges and validates the prowess of MDE algorithms.

Moreover, the multi-modal nature of KITTI's data, encompassing both visual imagery and LiDAR readings, brings forth another layer of depth (pun intended) to the testing environment. It allows a direct juxtaposition between predicted depth from monocular images and actual depth readings, enabling researchers to fine-tune and rectify their models with unparalleled precision. Additionally, the competitive spirit fostered by the KITTI leaderboard pushes researchers to continuously innovate, ensuring that the MDE field remains in perpetual motion.

Yet, while KITTI's virtues are many, it is not without its limitations. The dataset, deeply rooted in the specific landscapes and environments of Germany, could inadvertently introduce geographical biases. Models honed and perfected on KITTI might falter when faced with terrains or driving nuances alien to the dataset's confines.

To encapsulate, the KITTI dataset, with its robust and real-world data, has profoundly influenced the trajectory of Monocular Depth Estimation research. It has been both a challenge and a validator, a beacon guiding the evolution of MDE. However, as the field matures, there's a growing emphasis on diversifying datasets, ensuring that depth estimation models are as versatile as they are accurate.

4 Model

In this project, we explore the possibility to further improve the performance of efficient self-supervised monocular depth estimation, in the following aspects:

- Higher accuracy.

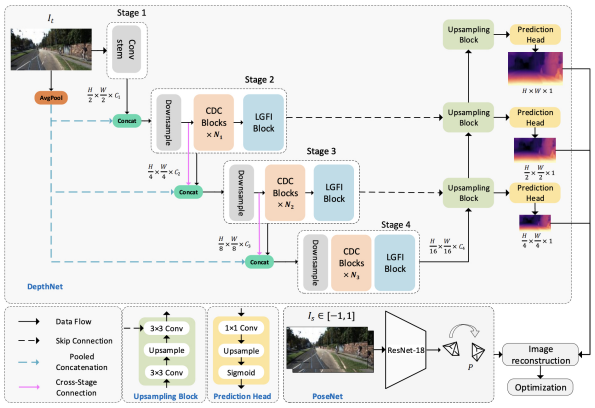

Figure 1: An illustration for our proposed network architecture.

- Lower time-lag and better real-time prediction.
- Less calculation and lower resource consumption.
- Less dependency on the training data.
- Better generalization ability.

An illustration for our proposed network architecture is depicted as in Figure 1.

5 Have Done So Far

- By Oct. 7, finish the investigation in literature.
- By Oct. 14, come up with ideas and finish reproducing the baseline.
- By Oct. 21, reproduce the core experiments as well as the experiment partition of the work.
- By Nov. 1, submit the intermediate draft.

6 To be Done

- By Nov. 7, finish all the experiments and submit supplementary experiments.
- By Nov. 31, finish the final script.

References

[1] Md Abul Bashar and Richi Nayak. Active learning for effectively fine-tuning transfer learning to downstream task. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(2):1–24, 2021. 2

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3

[3] Jun Liu, Qing Li, Rui Cao, Wenming Tang, and Guoping Qiu. Mininet: An extremely lightweight convolutional neural network for real-time unsupervised monocular depth estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:255–267, 2020. 2

[4] Siping Liu, Laurence Tianruo Yang, Xiaohan Tu, Renfa Li, and Cheng Xu. Lightweight monocular depth estimation on edge devices. *IEEE Internet of Things Journal*, 9(17):16168–16180, 2022. 2

[5] Elliot Meyerson and Risto Miikkulainen. Pseudo-task augmentation: From deep multitask learning to intratask sharing—and back. In *International Conference on Machine Learning*, pages 3511–3520. PMLR, 2018. 1

[6] Michael Moser, Michael Pfeiffer, and Josef Pichler. Domain-specific modeling in industrial automation: Challenges and experiences. In *Proceedings of the 1st International Workshop on Modern Software Engineering Methods for Industrial Automation*, pages 42–51, 2014. 2

[7] Michael Rudolph, Youssef Dawoud, Ronja Güldenring, Lazaros Nalpantidis, and Vasileios Belagiannis. Lightweight monocular depth estimation through guided decoding. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2344–2350. IEEE, 2022. 2

[8] Diana Wofk, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. Fastdepth: Fast monocular depth estimation on embedded systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6101–6108. IEEE, 2019. 2

[9] Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18537–18546, June 2023. 2