Exploration for Lightweight Monocular Depth Estimation

Wenxuan Zhang
Wzhang121@hawk.iit.edu
September 29, 2023

Abstract

To be completed..

1 Background

Monocular Depth Estimation (MDE) 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. 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. 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 selfsupervision. For instance, in extremely noisy environments, self-supervised signals might be too weak or ambiguous.

2 A brief survey

Lite-Mono [1] 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.

The Lite-Mono architecture proposed in the document makes the following contributions:

- Lightweight Architecture: Lite-Mono is a new lightweight architecture for self-supervised monocular depth estimation. It achieves a good balance between model size and computational complexity, making it efficient for real-time applications.
- Superior Accuracy: Lite-Mono outperforms competitive larger models on the KITTI dataset, achieving state-of-the-art results with the least trainable parameters. It also demonstrates good generalization ability on the Make3D dataset.
- Effective Feature Extraction: Lite-Mono incorporates Consecutive Dilated Convolutions (CDC) modules and Local-Global Features Interaction (LGFI) modules to extract rich hierarchical features. This allows the model to perceive different scales of objects and handle challenging scenarios like moving objects close to the camera.
- Trade-off between Complexity and Inference Speed: Lite-Mono achieves a good trade-off between model complexity and inference speed. It has been tested on both NVIDIA TITAN Xp and Jetson Xavier platforms, demonstrating its efficiency for real-time applications.

3 How the proposed work is different

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

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

4 Preliminary plan (milestones)

- By Oct. 7, finish the investigation in literature.
- By Oct. 14, come up with ideas and finish reproducing the baseline.
- By Oct. 21, finish the core experiments as well as the experiment partition of the final paper.
- By Nov. 1, finish the intermediate draft.
- BY Nov. 7, finish the supplementary experiments.
- BY Nov. 31, finish the final script.

References

[1] 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.