An overview of monocular depth estimation with applicability in intelligent transportation

Hasan Hotait
DTU Electro
Denmark Technical University
Kongens Lyngby, Denmark
s203211@dtu.dk

Alexandru Forrai
Research and Technology Development
Siemens Industry Software Netherlands B.V.
Helmond, The Netherlands
alexandru.forrai@siemens.com

***Abstract*—The paper deals with monocular depth estimation and performance assessment considering three publicly available algorithms: MiDAS, DepthAnything and ZoeDepth. In automated driving systems depth estimation is an essential part of the sensing and perception subsystem and usually relies on different sensors (camera, radar, LiDAR) and sensor fusion.**

Monocular depth estimation using a single camera image, mathematically is an ill-conditioned problem. However, recently several papers reported significant progress using sophisticated neural network-based architectures and intriguing learning approaches. Therefore, the paper compares these algorithms on a fair basis showing their estimation accuracy considering relative and metric depth. Furthermore, results are presented related to performance gains in case of metric depth estimation by fine tuning the network on a given data set.

The paper presents the achieved accuracy in terms of absolute relative error at pixel and object level, considering different monocular depth estimation methods. After fair comparison, the final judgment if such algorithms could be applied in automated driving systems is left to the reader.

***Index Terms*—monocular depth estimation; sensing and perception; automated driving systems; relative and metric depth; absolute relative error**

I. Introduction

This paper overviews three state-of-the-art (SOTA) monocular depth estimation (MDE) algorithms and compares their performances on a fair basis. Existing work either focused on relative depth estimation (disregarding the metric scale), especially on generalization performance or metric depth estimation, achieving a certain performance level on a specific data set.

Monocular depth estimation is inherently an ill-posed mathematical problem, but due to its cost-effective approach (using a single camera) could have numerous applications such as generative AI, 3D reconstruction and autonomous driving [1].

Learning-based approaches that aimed to estimate metric depth have used supervised training on homogeneous datasets with representative environments (e.g., focusing on indoor or outdoor scenes) to encourage the supervised network to learn an appropriate metric scale. However, this results in overfitting to narrow depth ranges and degrades generalization across environments [1].

As an alternative, relative depth estimation aims to regress pixel-wise depth predictions that are accurate relative to each other, has better generalization capabilities but carry no metric meaning.

MiDaS is trained on labeled datasets in disparity space—defined as inverse depth up to an arbitrary scale and shift—such that the disparity d is normalized to the range $[0, 1]$. This formulation enables the construction of a large meta-dataset by combining multiple existing datasets and incorporating 3D movies as an additional data source, where disparity labels are readily available.

DepthAnything on the other hand,for the first time, pays attention to large-scale unlabeled data with their teacher-student model approach [2]. To make positive use of such data, The student model is challenged with a more difficult optimization target when learning the pseudo labels. The student model is enforced to seek extra visual knowledge and learn robust representations under various strong perturbations to better handle unseen images [2].

Unlike the previously mentioned models which disregard the metric scale, ZoeDepth combines the two worlds (relative and metric depth) leading to a model with good generalization performance while maintaining metric scale.

ZoeDepth flagship model, ZoeD-M12-NK follows a two-stage approach. In the first-stage a relative MDE model is pre-trained smilar to [1] [3] on 12 datasets. While in the second stage, metric depth estimation heads are added to the encoder-decoder architecture and fine-tuned on two datasets using metric depth [4].

We also highlight that DepthAnything authors also extend their relative MDE model, following the two-stage approach in ZoeDepth to also provide a metric MDE model where the MiDaS encoder is simply replaced with their improved Depth Anything encoder, leaving other components unchanged [2].

Models are evaluated on six datasets: DIW [5], ETH3D [6], Sintel [7], KITTI [8], NYU Depth v2 [9] and TUM [10].

 HORIZON JU Innovation Actions - 101139048 - ENVELOPE - HORIZON-JU-SNS-2023. This project has received funding from the European Union's Smart Networks and Services Joint Undertaking (SNS JU) under grant agreement No. 101139048.