Computer Vision Project - Phase 1

21 november 2025

1 Paper Overview

For the first milestone, our approach was, generally speaking, to carefully read the paper Monocular Relative Depth Perception with Web Stereo Data Supervision and try to understand different components that we need to implement and to anticipate potential challenges we are going to face in the next milestones.

The paper proposes a method for estimating relative depth using a deep neural network built on top of a ResNet backbone. While the overall workflow is clearly presented, several technical components were not sufficiently detailed in the paper. These unclear parts, mainly the network architecture and the loss function, are discussed in the following subsections.

1.1 Network Architecture

The paper describes its model as using a ResNet encoder followed by residual and multi-scale fusion modules. However, the architectural description does not go into more narrow details on how these modules are implemented, and many details are missing. In particular, the paper does not specify:

- The exact configuration and depth of the residual blocks
- How features from different resolutions are fused
- The structure of the decoder or upsampling

The absence of these details makes it difficult to reconstruct the architecture precisely from the paper alone.

1.2 Loss Function

The loss function was one of the most challenging parts of the paper to interpret. Although the paper provides a mathematical expression, it relies on a ranking loss formulation that was not explained good enough. Also because of the unclear symbols and ambiguous indexing structure, it was challenging to follow how the loss evaluates whether a predicted depth ordering is correct. This lack of clarity made it difficult to fully grasp the intuition behind the loss and how to implement it correctly. These issues ultimately required external clarification from our professor to resolve.

1.3 Dataset Exploration

As part of this milestone, we downloaded the dataset referenced in the paper and examined its structure. During this exploration, one important observation was the high variability in the dimensions and aspect ratios of the images. The dataset contains a wide range of both horizontal and vertical images. This variability indicates that proper preprocessing will be necessary.

1.4 Related Work Review (MiDaS)

We also explored the MiDaS repository, which implements a ResNet-based architecture with multi-scale fusion for monocular depth estimation. Although it does not reference the paper directly, its design is highly similar and will be valuable when we begin implementing the model.

2 External Guidance and Communication

To resolve the unclear aspects of the paper, we used two strategies:

2.1 Meeting with the Professor

After our own investigation and attempts through large language models did not yield sufficient clarity, we arranged a meeting with our professor. In this meeting, we discussed the following.

- How is the ResNet backbone used in the architecture
- What is feature fusion, and how is it used in the paper
- The intuition of the loss function

This meeting significantly clarified the workflow of the model and provided the guidance needed to proceed with implementation in the next milestones.

2.2 Attempting to Contact the Authors

We also reached out to the author of the paper by email to request access to any additional available resources, such as presentation slides, talks, or code implementations. As of this report, we have not received a response. If further materials become available, they may support us in refining our implementation.

3 Next Steps

In this milestone, we developed a clear understanding of the overall workflow of the paper and identified the main challenges we will face during implementation. For the next milestone, we have considered implementing a baseline architecture for the second phase of the project. Since we now have access to the MiDaS repository, it might be clearer what to do. Nevertheless, many difficulties will remain, and **meetings with the professor will be highly necessary.**