Digging Into Self-Supervised Monocular Depth Estimation

Shashank Kapoor, Syed Rizwan Ali Quadri, Shubham Patwar
School of computing and Electrical Electronics
Indian Institute of Technology Mandi,
Kamand 175-005
Email: S22022@students.iitmandi.ac.in
Email: T22113@students.iitmandi.ac.in
Email: T22108@students.iitmandi.ac.in

1. Problem Statement

Monocular Depth Estimation is the task of estimating the depth value (distance relative to the camera) of each pixel given a single (monocular) RGB image. Generally State-of-the-art methods usually fall into one of two categories: designing a complex network that is powerful enough to directly regress the depth map, or splitting the input into bins or windows to reduce computational complexity. The most popular benchmark is the KITTI datasets. Models are typically evaluated using RMSE or absolute relative error.

Per-pixel ground-truth depth data is challenging to acquire at scale. To overcome this limitation, self-supervised learning has emerged as a promising alternative for training models to perform monocular depth estimation. In this, we propose a set of improvements, which together result in both quantitatively and qualitatively improved depth maps compared to competing self-supervised methods.

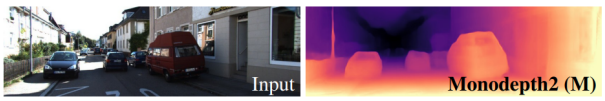

Figure 1. Depth from a single image.

Research on self-supervised monocular training usually explores increasingly complex architectures, loss functions, and image formation models, all of which have recently helped to close the gap with fully-supervised methods. We show that a surprisingly simple model, and associated design choices, lead to superior predictions. In particular, we propose:

(i) a minimum reprojection loss, designed to robustly handle occlusions

(ii) a full-resolution multi-scale sampling method that reduces visual artifacts

(iii) an auto-masking loss to ignore training pixels that violate camera motion assumptions.

We demonstrate the effectiveness of each component in isolation, and show high quality, state-of-the-art results on the KITTI benchmark.The goal is to find a dense depth map that minimizes the photometric error.

2. Literature Review

The paper discusses the challenge of inferring depth from a single image and proposes a self-supervised approach using monocular video or stereo pairs. The authors propose three innovations to improve monocular depth estimation:

- A novel appearance matching loss to address occluded pixels
- A simple auto-masking approach to ignore pixels with no relative camera motion
- A multi-scale appearance matching loss to reduce depth artifacts.

The proposed approach yields state-of-the-art results on the KITTI dataset and simplifies existing top-performing models.

The problem of estimating depth from a single image is difficult because there are multiple possible depths that an image can project to. Learning-based methods have been successful in addressing this problem by using predictive models that learn the relationship between color images and depth. Some of these methods require ground truth depth during training, which can be challenging to obtain in real-world settings. As a result, there is a growing body of work that exploits weakly supervised training data, such as known object sizes or synthetic depth data.

Recently, conventional structure-from-motion (SfM) pipelines have been used to generate sparse training signals for camera pose and depth. This approach has been shown to improve depth predictions when combined with traditional stereo algorithms.

3. Proposed Methods

The author discusses various methods for training depth estimation models in the absence of ground truth depth information. Self-supervised learning using stereo pairs or

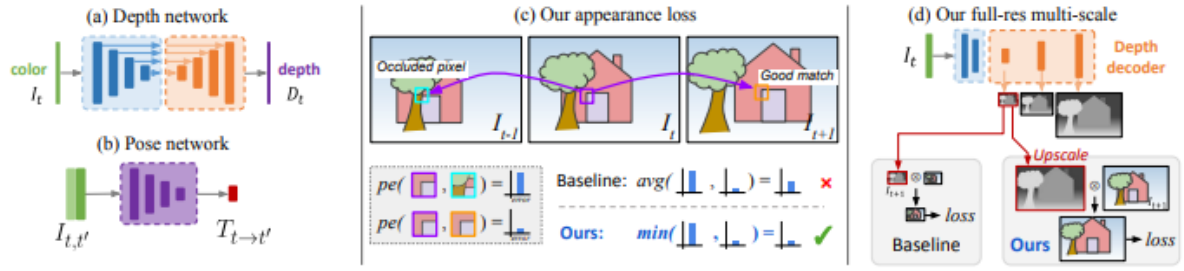

Figure 2. **Overview. (a) Depth network**: We use a standard, fully convolutional, U-Net to predict depth. **(b) Pose network**: Pose between a pair of frames is predicted with a separate pose network. **(c) Per-pixel minimum reprojection**: When correspondences are good, the reprojection loss should be low. However, occlusions and disocclusions result in pixels from the current time step not appearing in both the previous and next frames. The baseline average loss forces the network to match occluded pixels, whereas our minimum reprojection loss only matches each pixel to the view in which it is visible, leading to sharper results. (d) Full-resolution multi-scale: We upsample depth predictions at intermediate layers and compute all losses at the input resolution, reducing texture-copy artifacts.

monocular videos is a popular alternative, where the model is trained to minimize image reconstruction error by predicting depth or pixel disparities between frames.Appearance-based losses and careful resolution choices can improve performance, and recent approaches include geometry-based matching, instance segmentation, and explicit occlusion modeling. The article also proposes a novel minimum reprojection loss that improves the sharpness of depth predictions by only matching each pixel to the view in which it is visible.The training problem is formulated as the minimization of a photometric reprojection error, which measures the difference between the predicted image and the target image. The goal is to find a dense depth map that minimizes the photometric error.

The auto-masking method can also filter out pixels belonging to the object, but we found that this is a rare occurrence in practice. The auto-masking method can be seen as a form of data augmentation that helps the network learn to handle stationary scenes and moving objects.

Adaptive Depth Range Scaling, the predicted depths can vary between different scenes, making it difficult to compare results across datasets or even across different scenes within the same dataset. To address this issue, we propose a simple method to adaptively scale the predicted depth maps based on the median depth of each frame. Specifically, we compute the median depth of each frame and use it to scale the predicted depth map of that frame. This ensures that all frames have a similar scale, while still allowing for scene-specific depth ranges. We found that this simple method improves the accuracy of the predicted depth maps.

They also propose a multi-scale approach to depth estimation, where low-resolution depth maps are upsampled and the photometric error is computed at the high input resolution.

4. Why did you choose that particular research paper?

Genuinely, what i feel is that if i can go through this may be i contribute someway i.e Paper, updated version, to the computer vision research community as It presents a novel approach to depth estimation and pose estimation from monocular video and stereo images, which are important tasks in computer vision with many applications i.e inexpensively complement LIDAR sensor (SLAM). The paper also proposes several improvements to existing methods and achieves state-of-the-art results on benchmark dataset, which makes it a significant contribution to the field.

5. Future Plan. (Timeline for the rest of the semester):

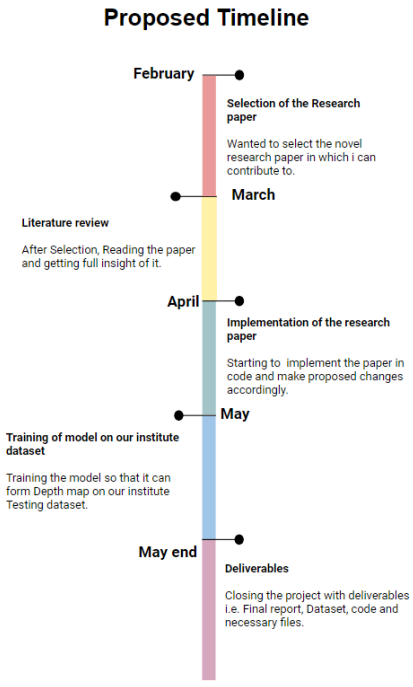

Figure 3. Proposed Timeline of the Project.

References

[1] Clement Godard et. al. *Digging Into self-Supervised Monocular Dept Estimation*, CVPR, August 2019