# MONODEPTH2: DIGGING INTO SELF-SUPERVISED MONOCULAR DEPTH ESTIMATION

PAPER ANALYSIS

**Ege Berke Balseven**
Department of Computer Engineering
Hacettepe University
Beytepe, Ankara, Turkey
b21590776@cs.hacettepe.edu.tr

August 1, 2019

***Keywords*** monocular-depth · self-supervised · stereo · pose network

## 1 Introduction

In this paper [1], the authors propose a set of improvements that are improved in both quantity and quality compared to other self-supervised methods. Research on self-supervised monocular training often explores increasingly complex structures, loss functions, and image formation models, all of which help to compensate for the lack of a fully supervised approach.They show simple model and related design choices that can lead to better predictions. In particular, they propose a minimum re-projection loss designed to robustly handle occlusion, a full resolution multi-scale sampling method, reduce visual artifacts, and an automatic masking loss, ignored A training pixel that violates camera motion assumptions. The effectiveness of each component was demonstrated separately and high quality, most advanced results were shown on the KITTI benchmark. Of course, some failure cases are also shown in the paper.

### 1.1 Features

- Per pixel min reprojection loss: at each pixel, instead of averaging the reprojection loss, use the min of loss in all the images. This improves the sharpness of occlusion boundaries.
- Auto-masking stationary pixels. This filters out pixels which do not change appearance from frame to the next. This per pixel mask is calculated in forwarding pass.
- This criterion indicates a static camera, or static object, or a low texture region.
- Scale back to original scale then do photometric loss calculation. This helps removing holes in large low-texture region.

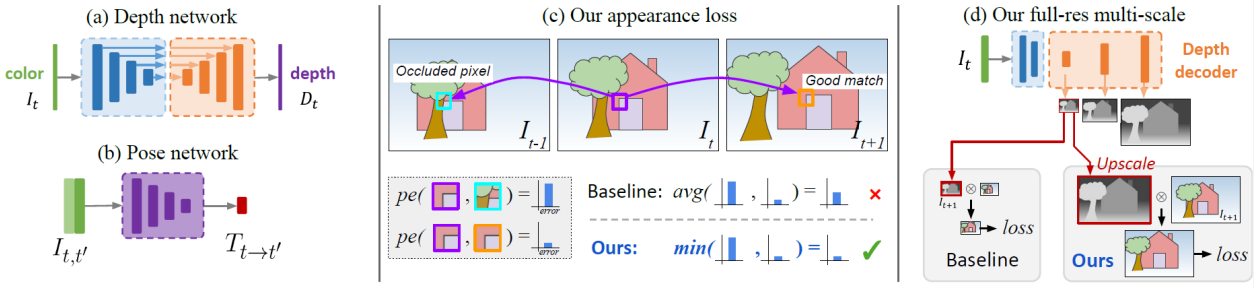


Figure 1: Monodepth2 Overview

The figure 1 is the core network structure of monodepth2. The core points in the paper can be summarized into three points:

- Deep network: A deep network is mentioned in the paper and the depth is obtained from the color map.
- Appearance feature loss: well-processed the prediction of occluded pixel values.
- Multi-scale synchronization for loss: loss training for groudtruth on multiple different scales.

### 1.2 Technical details

ResNet18 encoder used for both depth and pose networks. Implemented in PyTorch, trained for 20 epochs using Adam optimizer, with a batch size of 12 and an input/output resolution of 640x192 unless otherwise specified. Learning rate of $1 \times 10^{-4}$ for the first 15 epochs which is then dropped to $1 \times 10^{-5}$ for the remainder. This was chosen using a dedicated validation set of 10% of the data. Used L1 edge preserving loss.

### 1.3 Discussion

I think LSD-SLAM[2] method is better than this method for high-altitude aerial vehicles. LSD-SLAM is designed for tracking and mapping environment information, but maybe it will be better for low altitude vehicles. I think it can be used in drones. In addition Monocular2 so better in depth estimation. If we only consider the depth instead of tracking and mapping Monocular2 will be better choice. Therefore the depth performance of LSD-SLAM needs to be improved.

## References

[1] Clément Godard and Oisin Mac Aodha and Michael Firman and Gabriel J. Brostow, Digging into Self-Supervised Monocular Depth Prediction, *arXiv:1806.01260*, 2018.

[2] Jakob Engel and Thomas Schops and Daniel Cremers, LSD-SLAM: Large-Scale Direct Monocular SLAM, In ECCV, 2014.