Part I : Paper Analysis

Digging Into Self-Supervised Monocular Depth Estimation

Abstract

This research is dedicated to the self-supervised learning to perform monocular depth estimation. The researcher has improved depth maps compared to competing self-supervised methods. The research is focused simple method, and associated design choices. The researcher also offers a minumum reprojection loss, full-resolution multi-scale sampling method and an auto-masking loss to ignore training pixels. Researcher shows that on KITTI dataset.

Introduction

High quality depth-from color is important for lidar sensors used in self-driving cars and enable new single-photo applications. Solving for depth can be used large unlabeled image datesets. Collecting large and varied training datasets need ground truth depth for supervised learning. As a an another way last recent self-supervised approaches have shown that it is instead possible to train monocular depth estimation. There are two main points(streo pairs, monocular video) of view on depth estimation. First point of monocular video is better aternative to stereo-based supervision but it inroduces its own set of challenges. Using stereo data for training makes the camera-pose estimation but can cause issues related to blocking and texture-copy produce.

This research is aimed at improving monocular depth estimation when training with monocular video, stereo pairs, or both. The research will also offer three architectural and loss innovations. First one original apperance matching loss to address the problem when using monocular supervision. Second one that new and simple automasking way to ignore pixels where no relative camera motion is observed in monocular training. Third one is multi-scale matching loss that performans at the input resolution. These three combinations make monocular and stereo self-supervised depth estimation on KITTI dataset.

Literature Review

Input→Single color image
Output→ Predict the depth of each pixel

Supervised Depth Estimation
Supervised requiring ground-truth depth during training. This is challenging to gain in varied real-world settings. As a conclusion there is developing work to achieve weakly supervised training data.

Self-supervised Depth Estimation
Without ground truth depth, one alternative is to train depth estimation models using image resconstruction as the organizational signal.

Research Design
Researcher use standard fully convolutional, U-Net(encoder-decoder) to predict depth. As encoder they use ResNet18 which contains 11M parameters. They make data augmentations. Such as color augmentations and horizantal flip. Their models implemented in PyTorch trained for 20 epochs using Adam optimizer with batch size 12 and an input resolution 640x192. They set validation %10. Learning rate of 10^{-4} for 15 epochs.

Conclusion
Researcher claimed that they achieved state-of-the art for depth predictions. They propesed three improvements on depth predictions. There are a minimum reprojection loss, an auto-masking loss to ignore confusing and a full-resolution multi-scale sampling method.

Usage of Self-Supervised Monocular Depth Estimation A High-Altitude Aerial Vehicle

Self-Supervised Learning has been using since the late 1990s. Autonomously driving car is a major practical use. These method largely give up the need for manually labeled data as they are designed to work in unseen environments. In most recent studies on self-supervised learning for terrain classification, the ground truth is always used during operation. The monocular information is used a path-planning task, requiring a cost function for either exploring unknown potential obstacles. Since checking whether a potential obstacle is traversable is costly, the robot(drone) learns to classify the terrain ahead with vision.

In a flying robot first uses optical flow to select a landing site that is flat and free of obstacles. In order this to work, the robot has to move sufficiently with respect to the objects on the landing site. While flying, the robot uses self-supervised learning to learn a regression function that appearance-based values coming from the optical flow process. The learned function extends the capabilities of the robot, as after learning it is also able to select landing sites without moving.

References

[1] https://journals.sagepub.com/doi/pdf/10.1177/1756829318756355