Piecewise monocular depth estimation by plane fitting

Junhwa Hur
(junhwa.hur@visinf.tu-darmstadt.de)

Motivation

Monocular depth estimation is the task of estimating a dense depth map out of a single image, which is different from estimating a disparity map from a pair of stereo images. For the past few years, monocular depth estimation has been studied in depth and has demonstrated remarkable progress.

Figure 1: Monocular depth estimation

However, typical approaches demonstrate that the depth estimates near object boundaries are sometimes less accurate, **resulting in blurry artifacts**. These artifacts may be due to *i)* the decoder that directly regresses the depth value, which is typically known to result in blurry outputs, or *ii)* a lack of a regularization module that can make depth boundaries much sharper.

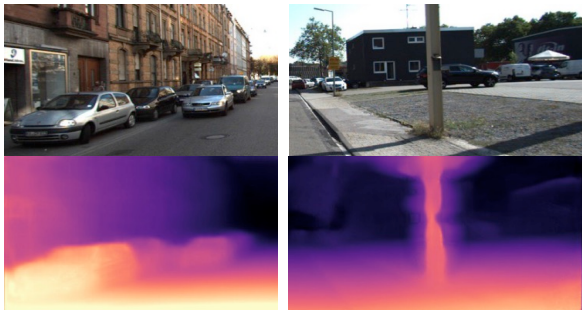

Figure 2. Inaccurate depth estimation near object boundaries

In this project, we aim to resolve this problem by exploiting **the explicit plane representation and fitting the plane model onto the scene**. Assuming that the scene is composed of multiple planes, i.e., superpixels (see Fig. 3), we aim to directly estimate **the plane coefficient** of each superpixel using Convolutional Neural Networks (CNNs).

Figure 3. Superpixel

$$ax + by + cz = 1$$

Then, with the known camera intrinsics, we can explicitly calculate the depth value of each pixel from its estimated plane coefficients. The benefit of estimating plane coefficients instead of directly estimating depth values is that they can effectively represent depth values of multiple pixels by estimating only 3 coefficients, and this explicit plane representation has the potential to reduce the blurry artifacts.

The input of the network is **a monocular image with its given superpixel map**, and the output of the network is **3 plane coefficients for each superpixel**. For easier training, the network can be trained in a **semi-supervised way** by using sparse LIDAR ground truth.

Goals

The main goal is to demonstrate that it is possible to directly estimate plane coefficients using CNNs, which are converted into a dense depth map.

Tasks

Students interested in this lab should work on the following steps:

Task 0. Understanding the literature and getting acquainted with a deep learning library (~20h)

- Read the reference papers and understand them clearly
- Choose one deep learning library and study how to use it

Task 2. Proceed with the project (~190h)

1) Prepare the dataset (e.g., KITTI raw dataset [1]) (~10h)
2) Choose one baseline (Monodepth [2] or Monodepth2 [3]) network to work on (~20h) and check whether it is working correctly
3) Choose an off-the-shelf superpixel generation algorithm and calculate the superpixel map of each image in the given dataset (~20h)
4) Train the network to estimate the plane coefficient of each superpixel and find a suitable training setup (~130h)
5) Evaluate the result on public benchmark datasets (~10h)

Task 3. Writing of a report (~60h)

References

[1] Geiger, Andreas, et al. "Vision meets robotics: The KITTI dataset." *The International Journal of Robotics Research* 32.11 (2013)

[2] Godard, Clément, Oisin Mac Aodha, and Gabriel J. Brostow. "Unsupervised monocular depth estimation with left-right consistency." CVPR 2017.

[2] Godard, C., Mac Aodha, O., Firman, M., & Brostow, G. "Digging into self-supervised monocular depth estimation." ICCV 2019