Unspervised Monocular Depth Estimation with Left-Right Consistency

Abstract

Despite the absense of ground truth depth data, it performs single image depth estimation, genearing disparity images by training the network with an image reconstruction loss. To overcome the poor result of the reconstruction loss, the author proposes a novel training loss that enforces consistency between the disparities.

1. Introduction

- Most existing techniques relyes on the assumption that multiple observations of the scene of interest are available.
- The FC model does not require any depth data.
- It trains the depth as an intermediate.
- It learns to predict the pixel-level correspondence between pairs of rectified stereo images.
- 35 msec to predict a dense depth map
- The contribution would be
  1) A network architecture that performs end-to-end unsupervised monocular depth estimation preserving left-right depth consistency
  2) An evaluation of several training losses and models
  3) new dataset

2. Related Work

Many approaches are typically only applicable with more than one input image for the scene of interest.

Learning-Based Stereo

Models like DispNet requires expensive ground truth data at training time. Also, those approaches use synthetic data for training.

Supervised Single Image Depth Estimation

Still needs high quality, pixel aligned, ground truth depth at training time. Monodepth network perform single depth image estimation with an added binocular color image instead of massive ground truth depth.

Unsupervised Depth Estimation

- It does not require ground truth depth at training time.
- Some of them (i.e. DeepStereo) are not suitable because it requires several nearby posed images at test time.
- The Deep3D network requires heavy memory usage, which means not scalable with bigger resolutions.
- Some models are not fully differentiable.
- The monodepth model are fully differentiable with in-model left-right consistency check.

3. Method

3.1 Depth Estimation as Image Reconstruction

Given Image I at test time, predicting the per-pixel scene depth,

$$\hat{d} = f(I)$$

- Given a calibrated pair of binocular cameras, we can learn function in order to reconstruct on image from other and 3D shape of the scene being imaged.

At training time, $\tilde{I}^r = I^l(d^r), \tilde{I}^l = I^r(d^l)$, where d corresponds to the image disparity - a scalar value per pixel that the model will learn to predict.

After that, given the base line distance b between the cameras and the focal length f, trivailly recover the depth,

$$\hat{d} = bf/d$$

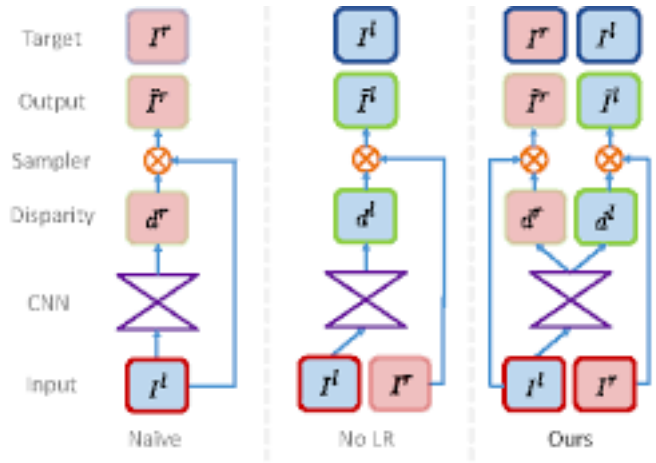
3.2 Depth Estimation Network

At a high level, our network estimates depth by inferring the disparities that warp the left image to match the right one. We can simultaneously infer both disparities, using only the left input image, and obtain better depths by enforcing them to be consistent with each other.

3.3 Training Loss

For each scale s, the total loss $C = \sum_{s=1}^{4} C_s$

$$C_s = \alpha_{ap}(C_{ap}^l + C_{ap}^r) + \alpha_{ds}(C_{ds}^l + C_{ds}^r) + \alpha_{lr}(C_{lr}^l + C_{lr}^r)$$



gure 3. Sampling strategies for backward mapping. With naï

Figure 1: enter image description here

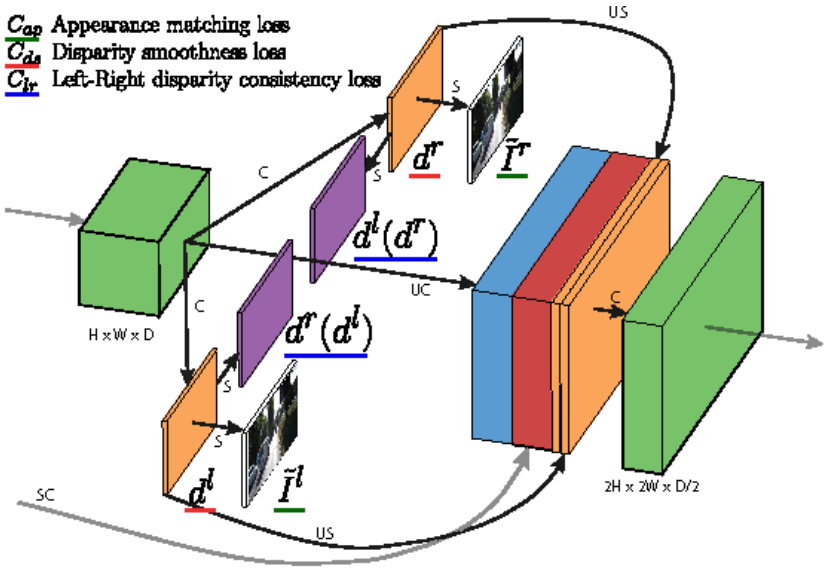


Figure 2. Our loss module outputs left and right disparity maps d^l

Figure 2: enter image description here

Appearance Matching Loss

$$C_{ap}^{l} = \frac{1}{N}\sum_{i,j} \alpha \frac{1 - \mathrm{SSIM}(\mathrm{I}_{\mathrm{ij}}^{\mathrm{l}}, \tilde{\mathrm{I}}_{\mathrm{ij}}^{\mathrm{l}})}{2} + (1-\alpha)\left|\left|I_{ij}^{l} - \tilde{I_{ij}^{l}}\right|\right|$$

Disparity Smoothness Loss

$$C_{ds}^{l} = \frac{1}{N}\sum_{i,j} \left|\partial_x d_{ij}^{l}\right| e^{-||\partial_x I_{ij}^{l}||} + \left|\partial_y d_{ij}^{l}\right| e^{-||\partial_y I_{ij}^{l}||}$$

Left-Right Disparity Consistency Loss

$$C_{lr}^{l} = \frac{1}{N}\sum_{i,j} \left|d_{ij}^{l} - d_{ij+d_{ij}^{l}}^{r}\right|$$

4. Results

It does better at resolving small objects such as the pedestrains and poles.

4.6 Limitations

- There are still some artifacts visible at occlusion boundaries due to the pixels in the occlusion region not being visible in both images.
- It requires rectified and temporally aligned stereo paris during trainig, meaning we cannot use existing single-view datasets for training purposes
- It relies on the image reconstruction term, namly that specular and transparent surfaces will produce inconsistent depths. – This could be improved with more sophisticated similarity measures.