Hierarchical Deep Stereo Matching on High-resolution Images

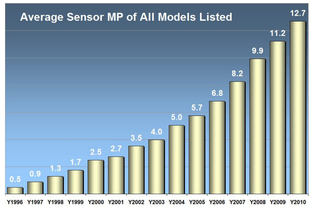
Gengshan Yang[1], Joshua Manela[2], Michael Happold[2], Deva Ramanan[1,2]

Carnegie Mellon University[1], Argo AI[2]

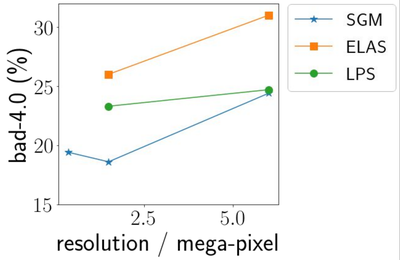
Data and Pytorch code available:

Introduction

Motivation



"Progression of Pixel Resolution in Digital Cameras." Indranil's World, 18 Sept. 2014, indranilsinharoy.com/2014/02/09/pixel-resolution-in-digital-cameras/



*On Middlebury benchmark.

LiDAR vs Stereo

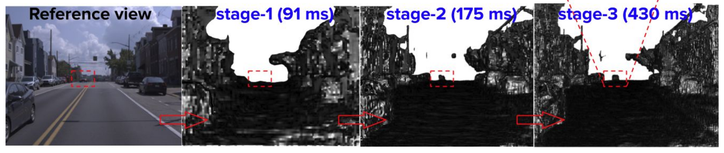
Why high-res?

$$\epsilon_Z = \frac{\epsilon_d}{bf} Z^2$$

ϵ_Z : depth error
ϵ_d : disparity error
b : baseline
f : focal length

Results

·SOTA on Middlebury while significantly faster than prior arts.
·**Anytime** on-demand reports of disparity from coarse-to-fine.



Challenges

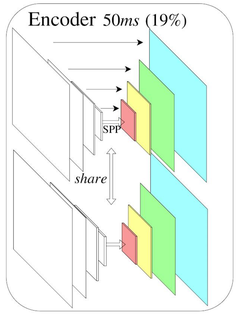
·Memory and running time overhead
·Insufficient high-resolution training data
·Vertical disparities

Approach

Hierarchical Stereo Matching (HSM) network

Reference & Target Image

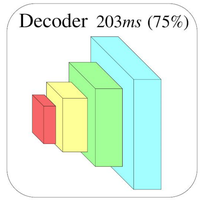
$H \times W \times 3$



Pyramid Features

$\frac{H \times W}{\{8, 16, 32, 64\}} \times C_k$

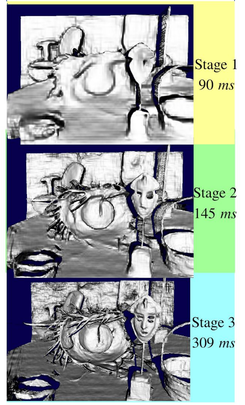
- 3D residual blocks
- Volumetric pyramid pooling
- Feature volume refinement
- Soft-argmin regression



Pyramid Cost Volumes

$\frac{H \times W \times D_k}{\{8, 16, 32, 64\}}$

Loss: Multi-scale smoothed-L1



High-resolution datasets

·For training, we collected High-res Virtual Stereo (HR-VS) dataset using Carla simulator.
·For testing, we collected High-res Real Stereo (HR-RS) dataset while driving in the urban scenes.

Name	Res*	Size	Scene	Real
Sintel	0.45	1064	In/Outdoor	N
ETH3D	0.46	27	Campus	Y
KITTI-15	0.47	200	Driving	Y
Sceneflow	0.52	30k	All	N
Middlebury	6.00	23	Indoor	Y
HR-VS (Ours)	5.07	780	Driving	N
HR-RS (Ours)	4.65	120	Driving	Y

*: in Megapixel

reference image | target image | ground-truth

reference image | target image | stage 3 prediction

Stereo data augmentation

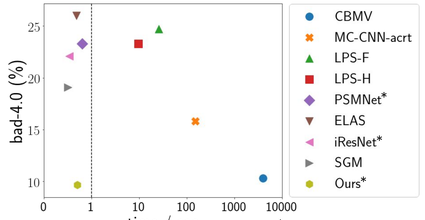
- y-disparity augmentation: randomly apply rotation and translation to the target image;
- Asymmetric color augmentation: randomly apply color transformations to the target image;
- Asymmetric occlusion: randomly occlude a region of the target image.

Experiments

Setup

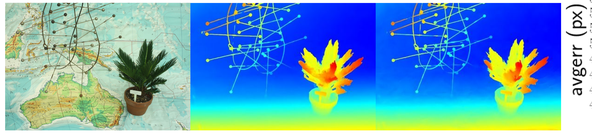
·We jointly train on 4 publicly available datasets, including Middlebury-14, KITTI-15, ETH3D and SceneFlow, together with proposed HR-VS.

Middlebury

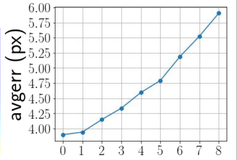


*: "deep" stereo methods

Resilience to rectification error

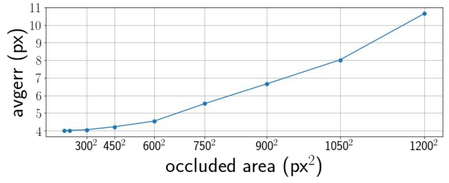


reference image | perfect rectification | imperfect rectification



y-translation in pixels

Resilience to occlusion



Diagnostics

Method	bad-2.0 (%)
full-method	**26.88**
feature fusion → cost fusion	31.03
w/o HR-VS	30.24
w/o y-disp. augmentation	28.98
w/o multi scale loss	28.83
w/o occlusion augmentation	28.20
w/o VPP	27.73