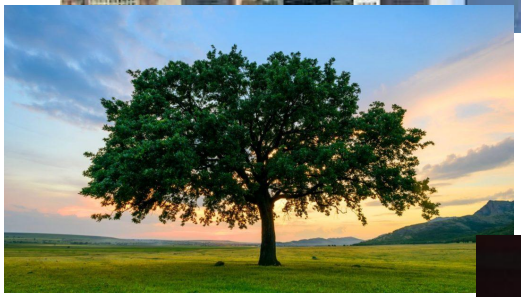
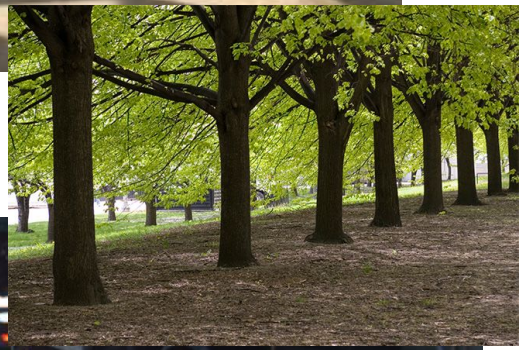
Geo-Supervised Visual Depth Prediction

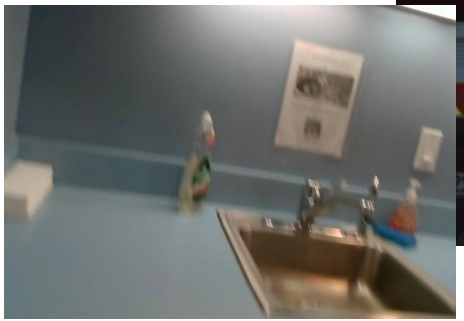
Xiaohan Fei, **Alex Wong**, and Stefano Soatto

UCLAVISIONLAB

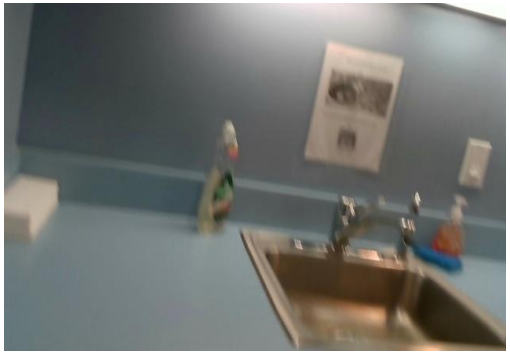
Gravity Biases Shape

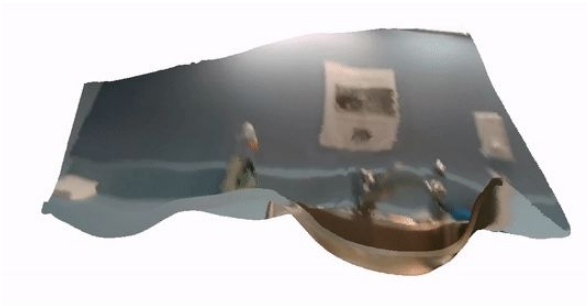


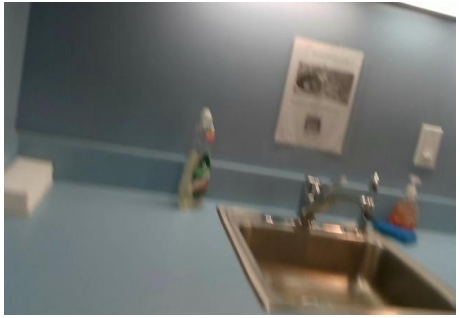


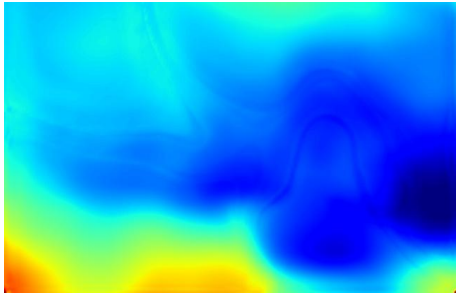


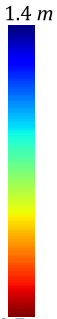
Generic Regularization



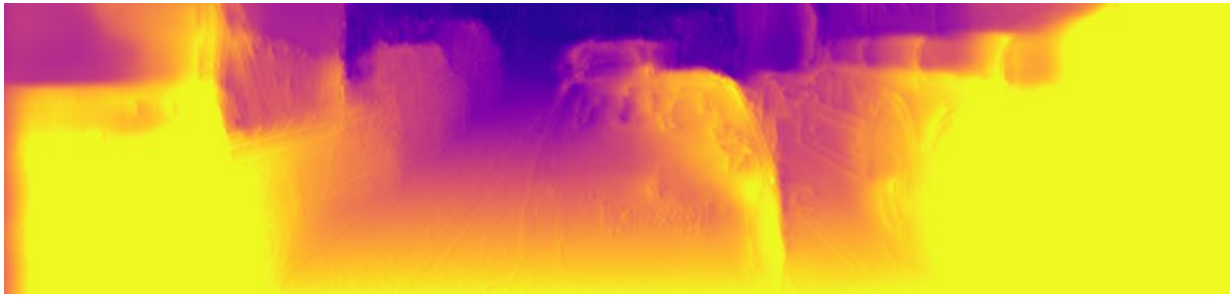




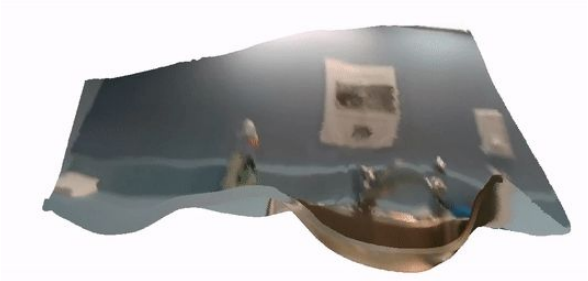




Yin *et al.*



Godard *et al.*

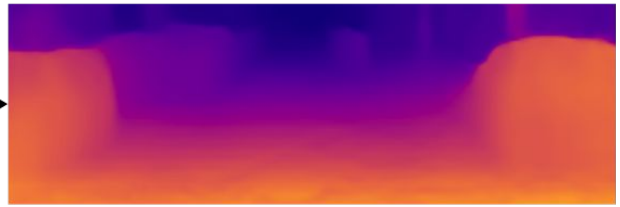


[1] Yin *et al.,* GeoNet: Unsupervised learning of depth, optical flow and camera pose. CVPR, 2018.
[2] Godard *et al.*, Unsupervised Monocular Depth Estimation with Left-Right Consistency. CVPR, 2017.

Single Image Depth Prediction

I_0

Depth Predictor



z

Single image depth prediction is an ill-posed problem

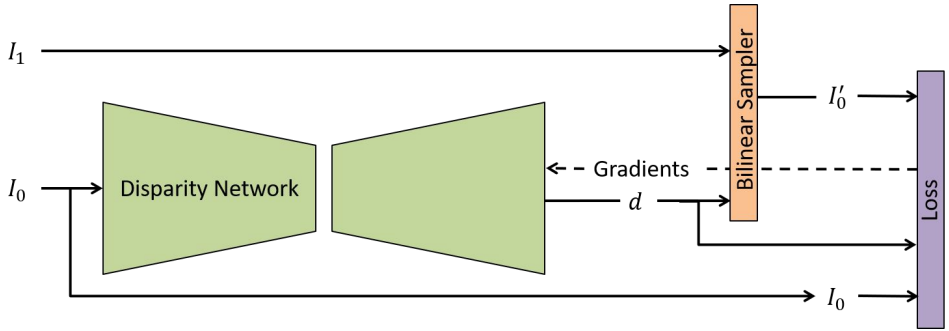
We **must** rely on a prior -- one that exploits the bias induced by gravity

Not All Objects are Biased by

y

Typical Architecture of Depth Prediction Systems

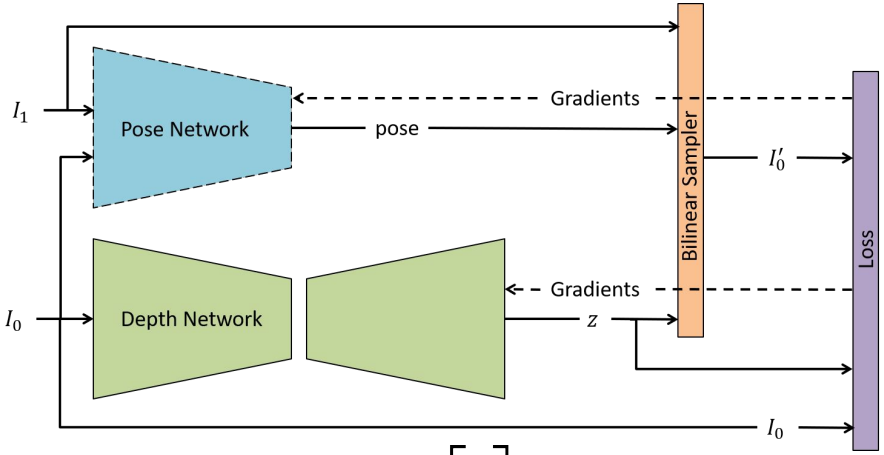
Stereo Training Pipeline [2]



$$I_0'(x, y) = I_1(x + d, y)$$

$$L = \sum_{x,y \in \Omega} |I_0(x, y) - I_0'(x, y)| + |\nabla d(x, y)|$$

Monocular Training Pipeline [3]



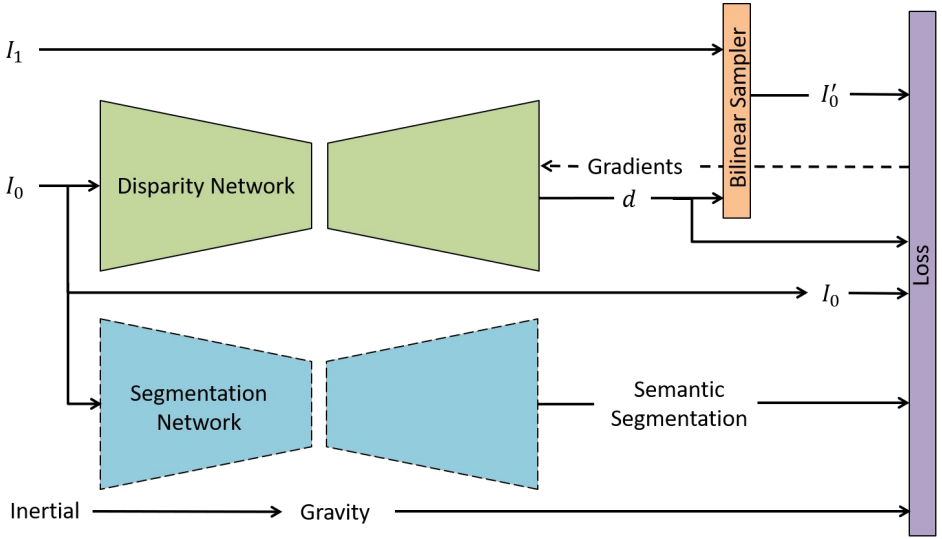
$$I_0'(x, y) = I_1(\pi g K^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} z)$$

$$L = \sum_{x,y \in \Omega} |I_0(x, y) - I_0'(x, y)| + |\nabla z(x, y)|$$

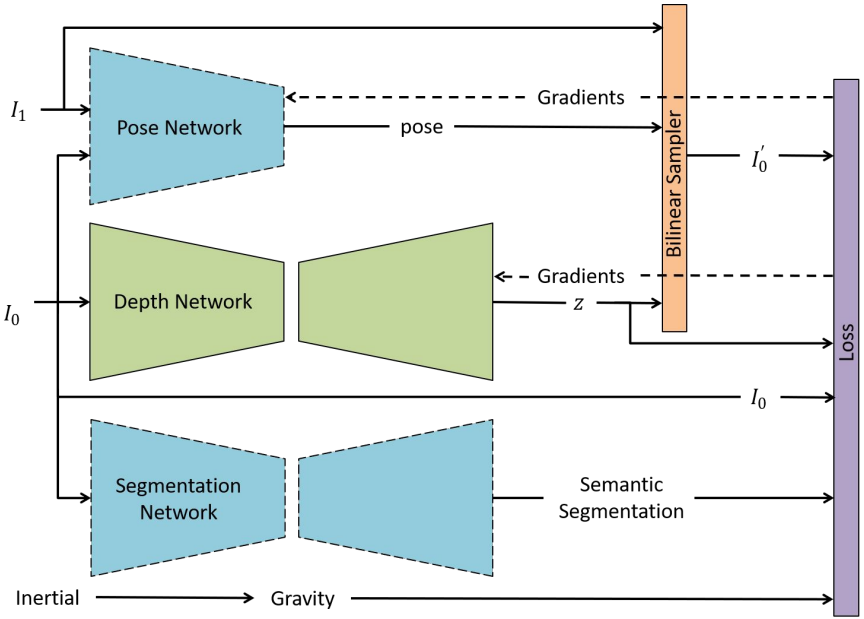
[2] Godard *et al.*, Unsupervised Monocular Depth Estimation with Left-Right Consistency. CVPR, 2017.
[3] Zhou *et al.*, Unsupervised learning of depth and ego-motion from vision. CVPR, 2017.

Our Proposed System

Stereo Training Pipeline



Monocular Training Pipeline



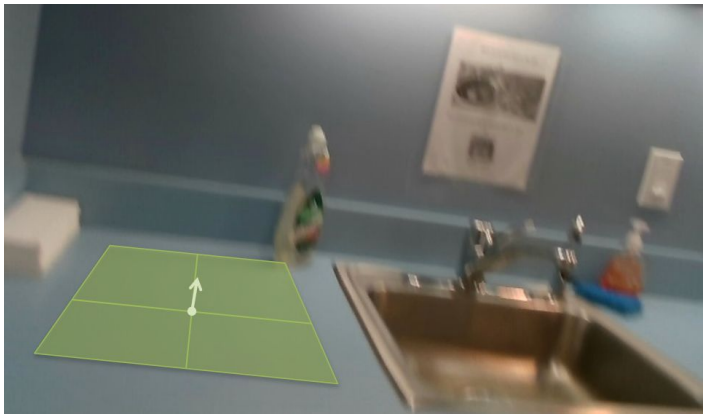
Incorporating a Semantically Induced Geometric Loss (SIGL):

$$L^* = L + \underbrace{L_{HP} + L_{VP}}_{\text{SIGL}}$$

SIGL: Horizontal Plane Loss

$$L_{HP}(\Omega_{HP}) = \frac{1}{|\Omega_{HP}|}\|(\mathbf{I} - \frac{1}{|\Omega_{HP}|}\mathbf{1}\mathbf{1}^\top)\bar{\mathbf{X}}\gamma\|_2^2$$

$$= \frac{1}{|\Omega_{HP}|}\sum_{i=1}^{|\Omega_{HP}|}((\mathbf{X}_i - \mu)^\top\gamma)^2$$



$\Omega_{HP} \subset \mathbb{R}^2$ is a subset of the image plane whose associated semantic classes have **horizontal surfaces**

$\bar{\mathbf{X}}$ is the collection of 3D points which project to Ω_{HP}

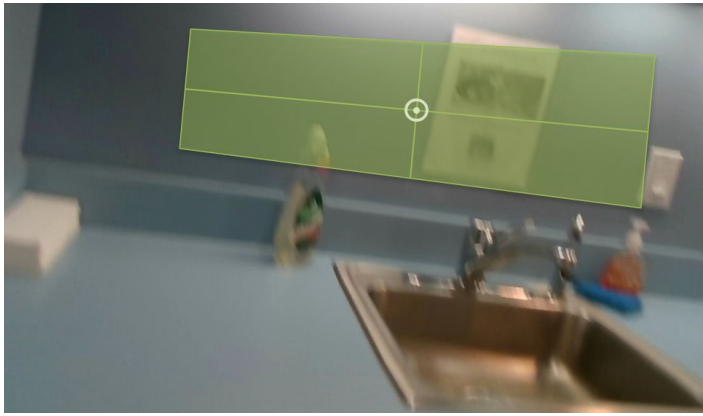
γ denotes gravity

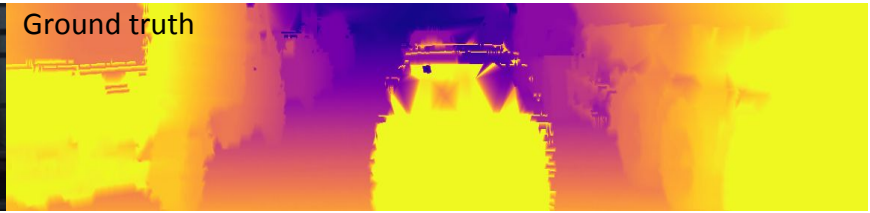
SIGL: Vertical Plane Loss

$$L_{VP}(\Omega_{VP}) = \min_{\substack{N \in \mathcal{N}(\gamma) \\ \|N\|=1}} \frac{1}{|\Omega_{VP}|} \|(\mathbf{I} - \frac{1}{|\Omega_{VP}|}\mathbf{1}\mathbf{1}^\top)\bar{\mathbf{X}}N)\|_2^2$$

$\Omega_{VP} \subset \mathbb{R}^2$ is a subset of the image plane whose associated semantic classes have **vertical surfaces**

$\mathcal{N}(\gamma)$ is the null space of γ

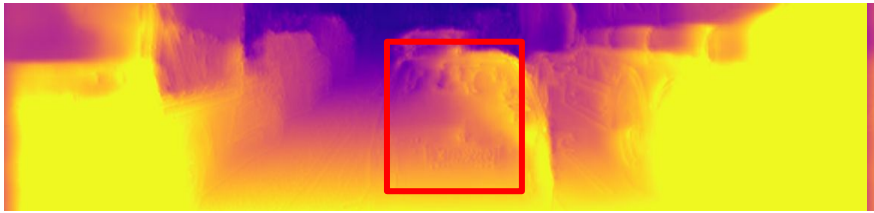


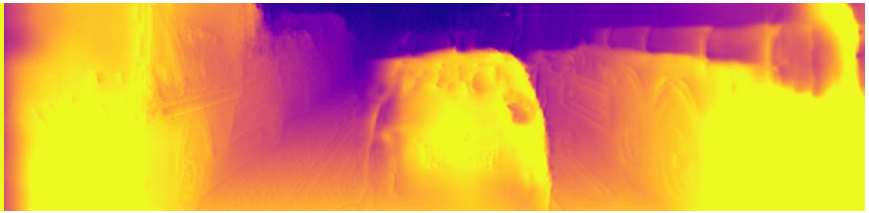


[2] w/out SIGL

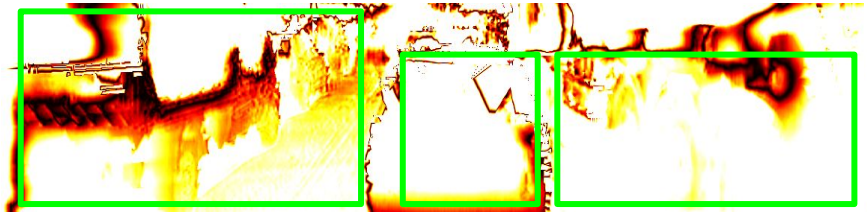
[2] with SIGL

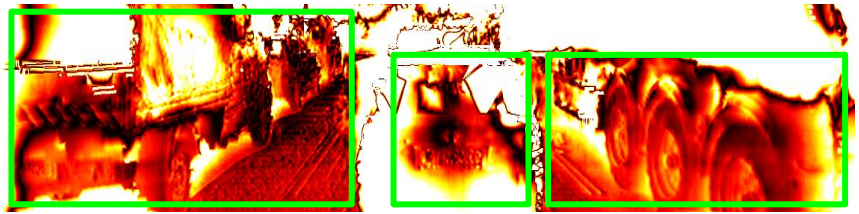
Disparity





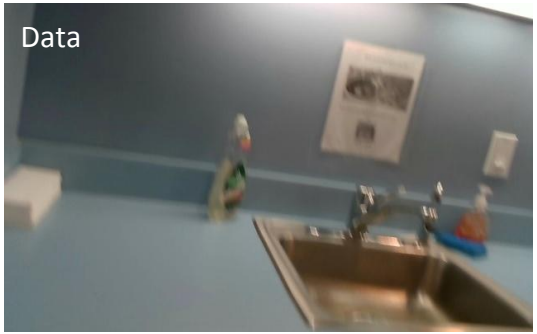
Error

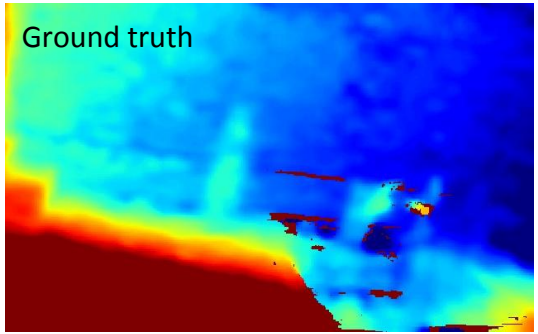




[Geiger *et al.*, Vision meets robotics: The KITTI dataset. IJRR, 2013.]

[2] Godard *et al.*, Unsupervised Monocular Depth Estimation with Left-Right Consistency. CVPR, 2017.

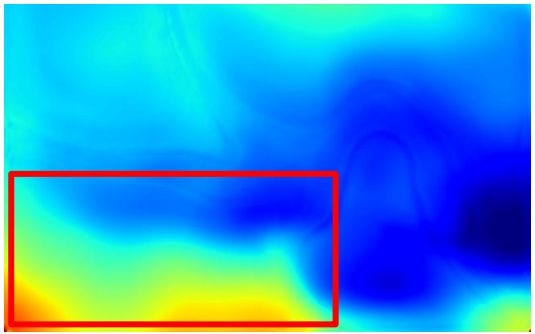


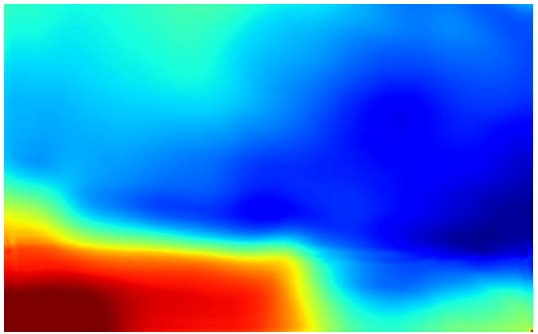


[1] w/out SIGL

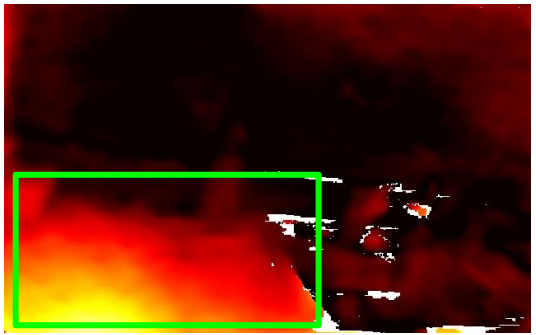
[1] with SIGL

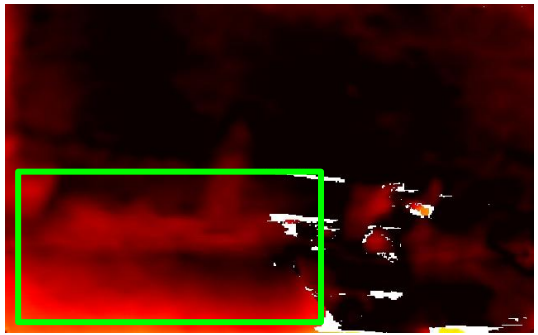
Depth





Error





Our Visual-Inertial and Depth Dataset

[1] Yin *et al.,* GeoNet: Unsupervised learning of depth, optical flow and camera pose. CVPR, 2018.

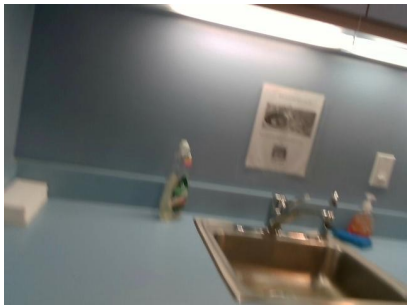
Method	Data	Error metric				Accuracy ($\delta <$)		
		AbsRel	SqRel	RMSE	RMSElog	1.25	1.25^2	1.25^3
Stereo Training								
Monodepth *VGG* [2]	K	0.148	1.344	5.937	0.247	0.803	0.922	0.964
+SIGL	K	**0.139**	**1.211**	**5.702**	**0.239**	**0.816**	**0.928**	**0.966**
Stereo-Temporal [4]	K	0.144	1.391	5.869	0.241	0.803	0.928	0.969
+SIGL	K	**0.137**	**1.061**	**5.692**	**0.239**	**0.805**	0.928	0.969
Monodepth *VGG* [2]	CS+K	0.124	1.076	5.311	0.219	0.847	0.942	0.973
+SIGL	CS+K	**0.114**	**0.885**	**4.877**	**0.203**	**0.858**	**0.950**	**0.978**
Monodepth *ResNet* [2]	CS+K	0.114	0.898	4.935	0.206	0.861	0.949	0.976
+SIGL	CS+K	**0.112**	**0.836**	**4.892**	**0.204**	**0.862**	**0.950**	**0.977**
Monocular Training								
GeoNet *ResNet* [1]	K	0.155	1.296	5.857	0.233	0.793	0.931	0.973
+SIGL	K	**0.142**	**1.124**	**5.611**	**0.223**	**0.813**	**0.938**	**0.975**
DDVO [5]	K	0.151	1.257	5.583	0.228	**0.810**	0.936	0.974
+SIGL	K	**0.146**	**1.068**	**5.538**	**0.224**	0.809	**0.938**	**0.975**
GeoNet *ResNet* [1]	CS+K	0.153	1.328	5.737	0.232	0.802	0.934	0.972
+SIGL	CS+K	**0.147**	**1.076**	**5.468**	**0.222**	**0.806**	**0.938**	**0.976**
DDVO [5]	CS+K	0.148	1.187	5.496	0.226	0.812	0.938	0.975
+SIGL	CS+K	**0.142**	**1.094**	**5.409**	**0.219**	**0.821**	**0.941**	**0.976**

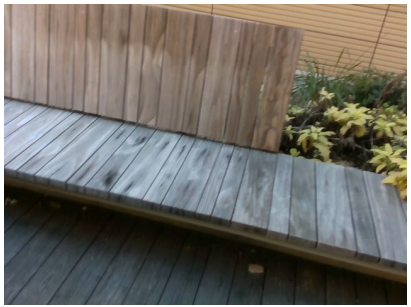
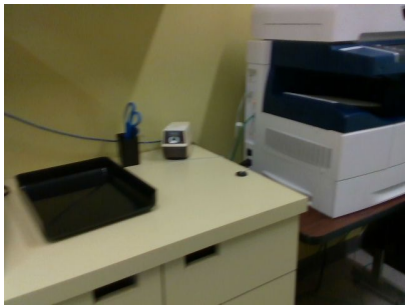
CS = Cityscapes
K = KITTI

[1] Yin *et al.,* GeoNet: Unsupervised learning of depth, optical flow and camera pose. CVPR, 2018.
[2] Godard *et al.*, Unsupervised Monocular Depth Estimation with Left-Right Consistency. CVPR, 2017.
[4] Zhan *et al.,* Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. CVPR, 2017.
[5] Wang *et al.,* Learning depth from monocular video using direct methods. CVPR, 2018.

Method	Data	Error metric				Accuracy (δ <)		
		AbsRel	SqRel	RMSE	RMSElog	1.25	1.25^2	1.25^3
Stereo Training								
Monodepth *VGG* [2]	K	0.148	1.344	5.937	0.247	0.803	0.922	0.964
+SIGL	K	**0.139**	**1.211**	**5.702**	**0.239**	**0.816**	**0.928**	**0.966**
Stereo-Temporal [4]	K	0.144	1.391	5.869	0.241	0.803	0.928	0.969
+SIGL	K	**0.137**	**1.061**	**5.692**	**0.239**	**0.805**	0.928	0.969
Monodepth *VGG* [2]	CS+K	0.124	1.076	5.311	0.219	0.847	0.942	0.973
+SIGL	CS+K	**0.114**	**0.885**	**4.877**	**0.203**	**0.858**	**0.950**	**0.978**
Monodepth *ResNet* [2]	CS+K	0.114	0.898	4.935	0.206	0.861	0.949	0.976
+SIGL	CS+K	**0.112**	**0.836**	**4.892**	**0.204**	**0.862**	**0.950**	**0.977**
Monocular Training								
GeoNet *ResNet* [1]	K	0.155	1.296	5.857	0.233	0.793	0.931	0.973
+SIGL	K	**0.142**	**1.124**	**5.611**	**0.223**	**0.813**	**0.938**	**0.975**
DDVO [5]	K	0.151	1.257	5.583	0.228	**0.810**	0.936	0.974
+SIGL	K	**0.146**	**1.068**	**5.538**	**0.224**	0.809	**0.938**	**0.975**
GeoNet *ResNet* [1]	CS+K	0.153	1.328	5.737	0.232	0.802	0.934	0.972
+SIGL	CS+K	**0.147**	**1.076**	**5.468**	**0.222**	**0.806**	**0.938**	**0.976**
DDVO [5]	CS+K	0.148	1.187	5.496	0.226	0.812	0.938	0.975
+SIGL	CS+K	**0.142**	**1.094**	**5.409**	**0.219**	**0.821**	**0.941**	**0.976**

CS = Cityscapes
K = KITTI



RGB-D 640x480 @ 30 Hz IMU @ 400 Hz

Limitations and Further Extensions

Exploit inertials at test time

Automatic selection of classes that are gravity-biased

Dynamic objects

Limitations and Further Extensions

Exploit inertials at test time

Automatic selection of classes that are gravity-biased

Dynamic objects

Limitations and Further Extensions

Exploit inertials at test time

Automatic selection of classes that are gravity-biased

Dynamic objects

Project site: shorturl.at/cgiAS