Geo-Supervised Visual Depth Prediction

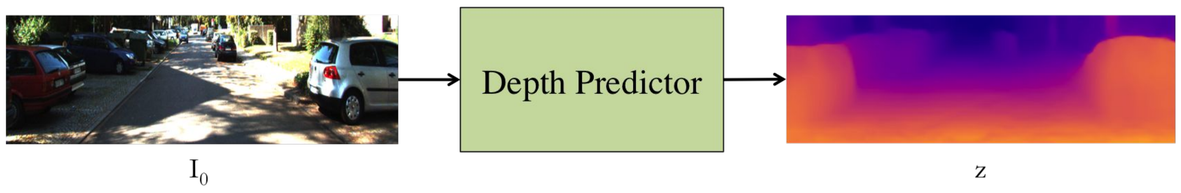
Xiaohan Fei <feixh@cs.ucla.edu>
Alex Wong <alexw@cs.ucla.edu>
Stefano Soatto <soatto@cs.ucla.edu>
UCLAVISIONLAB

ICRA 2019, Montreal

Motivation

- Agents can benefit from understanding the shape of objects.
- The shape of objects surrounding us is biased by gravity.
- Gravity provides a persistent global orientation reference and can be easily inferred from inertial sensors.
- Many devices today have both visual and inertial sensors.

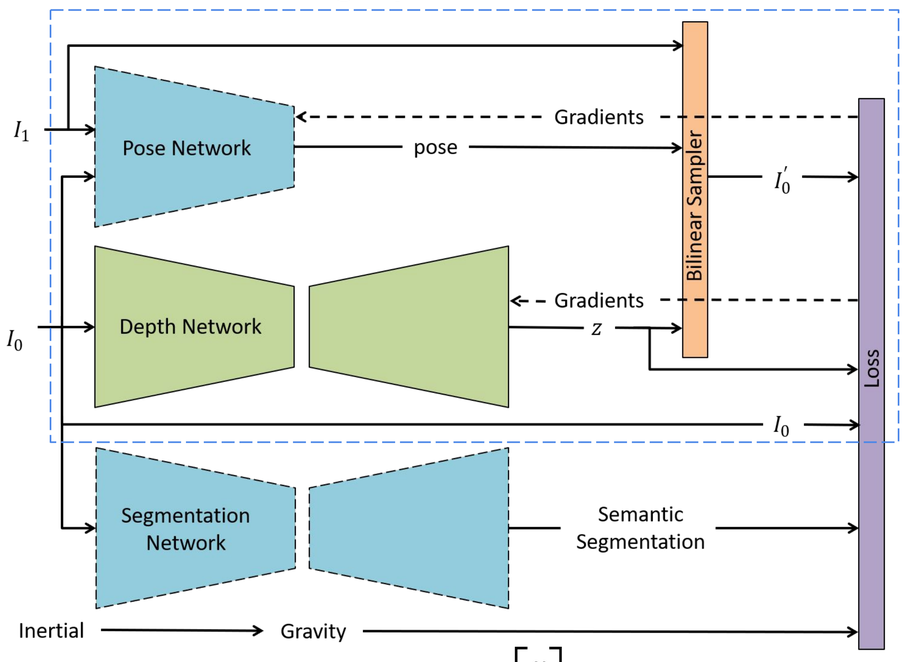
Goal: to leverage gravity to impose priors on shapes to improve visual depth prediction.



- Inferring shape from images is ill-posed so prior/regularization is needed.
- Not every object is biased by gravity so imposing the bias induced by gravity has to be done selectively, in a way that is object dependent.

Proposed System

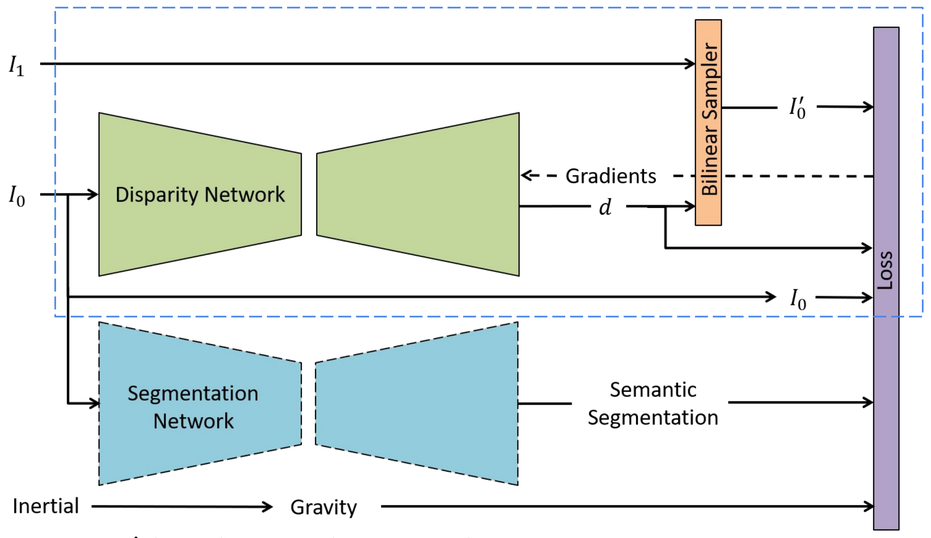
Monocular Training Pipeline



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

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

Stereo Training Pipeline



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

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

Legends

- Typical architecture of depth prediction systems
- Generic regularizers, e.g., piecewise smoothness
- Modules only needed in training

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

SIGL = Semantically Informed Geometric Loss

Typical architecture

- Encoder-Decoder structure with skip-connections.
- Learn to predict depth by minimizing a view synthesis loss (data term) and some generic regularizers, such as piecewise smoothness.
- Two popular training diagrams:
  - Monocular videos
  - Stereo pairs
- At inference time, only one RGB image is needed.

Proposed system

- At training time, apply *category-specific* shape priors *selectively*, which requires:
  1) a semantic segmentation module to provide per pixel class labels
  2) a Visual-Inertial Odometry (VIO) system to provide reliable gravity estimation
- At inference time, still only one RGB image is needed.

Semantically Informed Geometric Loss (SIGL)

Vertical Plane Loss

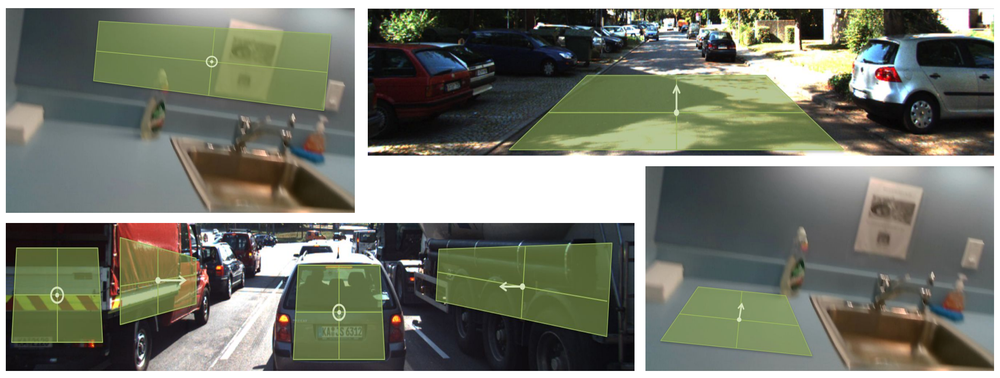
$$L_{VP}(\Omega_{VP}) = \min_{\substack{N \in \mathcal{N}(\gamma) \\ \|N\|=1}} \frac{1}{|\Omega_{VP}|} \|(\mathbf{I} - \frac{1}{|\Omega_{VP}|}\mathbf{1}\mathbf{1}^\top)\bar{\mathbf{X}}N)\|_2^2$$

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

$\bar{\mathbf{X}} \in \mathbb{R}^{|\Omega_{VP}| \times 3}$: collection of 3D points which project to Ω_{VP}

$\gamma \in \mathbb{R}^3$: gravity

$\mathcal{N}(\gamma)$: Null space of γ



Horizontal Plane Loss

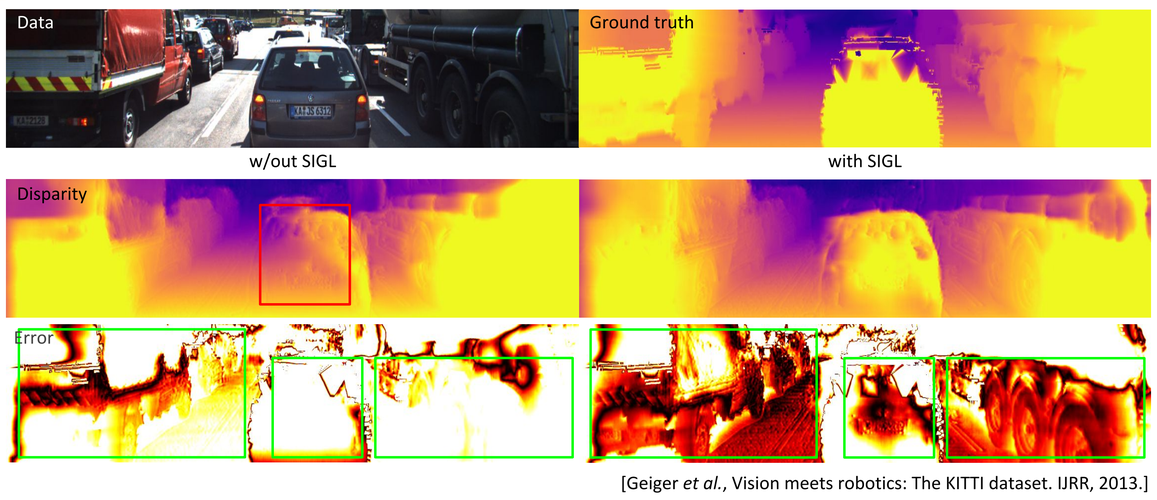
$$L_{HP}(\Omega_{HP}) = \frac{1}{|\Omega_{HP}|} \|(\mathbf{I} - \frac{1}{|\Omega_{HP}|}\mathbf{1}\mathbf{1}^\top)\bar{\mathbf{X}}\gamma\|_2^2$$

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

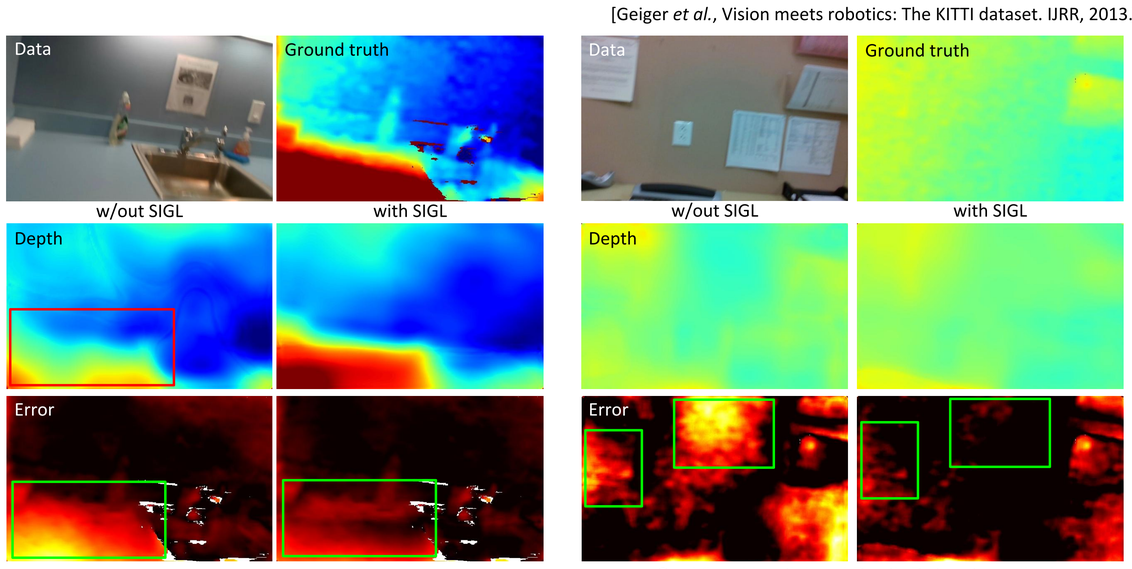
$\Omega_{HP} \subset \mathbb{R}^2$: subset of the image plane whose associated semantic classes have **horizontal surfaces**

$\bar{\mathbf{X}} \in \mathbb{R}^{|\Omega_{HP}| \times 3}$: collection of 3D points which project to Ω_{HP}

Results



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



Our Visual Inertial and Depth Dataset, available online

Method	Data	Error metric				Accuracy (δ <)		
		AbsRel	SqRel	RMSE	RMSElog	1.25	1.25^2	1.25^3
Stereo Training								
Monodepth *VGG* [1]	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 [2]	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* [1]	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* [1]	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* [3]	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 [4]	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* [3]	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 [4]	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**

K=KITTI, CS=Cityscapes

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

Our Visual Inertial and Depth Dataset

- **Extended version of the VISMA dataset used in paper**
- **RGB-D of VGA size @ 30 Hz**
- **Acce & Gyro @ 400 Hz**
- **Non-trivial 6 DoF motion**
- **> 40,000 frames, suitable for learning-based visual-inertial sensor fusion**

Code & Data

SCAN ME

Acknowledgement

This work was supported by ONR N00014-17-1-2072 and ARO W911NF-17-1-0304.