Geo-Supervised Visual Depth Prediction

Xiaohan Fei
<feixh@cs.ucla.edu>

Alex Wong
<alexw@cs.ucla.edu>

Stefano Soatto
<soatto@cs.ucla.edu>

UCLA VISION LAB

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.

**Proposed System**

**Monocular Training Pipeline**

- Pose Network
- Depth Network
- Inertial Network
- Semantic Segmentation

**Stereo Training Pipeline**

- Disparity Network
- Segmentation Network
- Inertial Network
- Gravity

**Proposed System**

- At inference 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

**Results**

- Vertical Plane Loss
  
  \[ L_{VP}(\Omega_{VP}) = \frac{1}{N_{VP}} \sum_{x,y \in \Omega_{VP}} \left[ |I'(x,y) - I_{d}(x,y)| + \|\nabla I(x,y)\| \right] \]

- Horizontal Plane Loss
  
  \[ L_{HP}(\Omega_{HP}) = \frac{1}{N_{HP}} \sum_{x,y \in \Omega_{HP}} \left[ (x_1 - p_1)^2 \right] \]

**Semantically Informed Geometric Loss (SIGL)**

**Our Visual Inertial and Depth Dataset**

- Extended version of the VSDA dataset used in paper
- RGB-D of VSDA size @ 30 Hz
- Pose & Gyro @ 400 Hz
- Non-ideal e/d ball motion
- 11,000 frames, suitable for learning-based visual-inertial sensor fusion

**Acknowledgments**

This work was supported by ONR N00014-17-1-2037 and ARO W911NF-17-1-0436.