

# DEPARTMENT OF COMPUTER SCIENCE

# Motivation

- Visual-Inertial Odometry (VIO) enables ubiquitous mobility for mobile robots by providing accurate pose information.
- Recent deep learning approaches for VIO have proven successful.
- Previous work rarely focus on incorporating robust fusion strategies for dealing with **imperfect input** sensory data.
- **Real issues** include camera occlusion or operation in low-light conditions, measurement noises, temporal or spatial misalignment between two sensors.
- The learning-based methods are **not explicitly modelling** the sources of degradation in real-world usages.
- Naively using all features before fusion will lead to **unreliable** state estimation.

# Contributions

- A generic framework to learn **feature selections** from two modalities, enabling robust and accurate ego-motion estimation
- Our selective sensor fusion masks can be visualised and interpreted
- A new and complete systematic research on the accuracy and robustness of deep sensor fusion in presence of corrupted data



- The hard and soft fusion masks under different conditions
- Left: normal data; middle and right: corrupted data

# **Selective Sensor Fusion for Neural Visual-Inertial Odometry**

Changhao Chen<sup>1</sup>, Stefano Rosa<sup>1</sup>, Yishu Miao<sup>2</sup>, Chris Xiaoxuan Lu<sup>1</sup>, Wei Wu<sup>3</sup>, Andrew Markham<sup>1</sup>, Niki Trigoni<sup>1</sup> <sup>1</sup>Department of Computer Science, University of Oxford <sup>2</sup>MO Intelligence <sup>3</sup>Tencent

# **Neural Visual-Inertial Odometry Framework**



Visual Encoder: extracts visual features from a set of two consecutive monocular images. Inertial Encoder: extracts inertial features from a sequence of inertial measurements. Feature Fusion: combines the features from two modalities Temporal Modelling: employs LSTMs to model temporal dependencies. Pose Regression: maps the latent space to pose transformation.

# **Selective Sensor Fusion - Deep Features Selection**

### **Soft Fusion (Deterministic):**

- Re-weights each feature by conditioning on both the visual and inertial channels

$$\mathbf{s}_{V} = \operatorname{Sigmoid}_{V}([\mathbf{a}_{V}; \mathbf{a}_{I}])$$
$$\mathbf{s}_{I} = \operatorname{Sigmoid}_{I}([\mathbf{a}_{V}; \mathbf{a}_{I}])$$

#### Hard Fusion (Stochastic):

- Generates a binary mask that either propagates the feature or blocks it
- Gumbel-Softmax resampling to infer the stochastic layer

$$\mathbf{s}_V \sim p(\mathbf{s}_V | \mathbf{a}_V, \mathbf{a}_I) = \text{Bernoulli}(\alpha_V)$$
  
 $\mathbf{s}_I \sim p(\mathbf{s}_I | \mathbf{a}_V, \mathbf{a}_I) = \text{Bernoulli}(\alpha_I).$ 

### **Feature Fusion**:

• Features are element-wise multiplied with their corresponding soft or hard masks

$$g(\mathbf{a}_V,\mathbf{a}_I) = [\mathbf{a}_V \odot \mathbf{s}_V; \mathbf{a}_I \odot \mathbf{s}_I].$$

### Intuition

• The fusion masks can be viewed as similar to the gain and covariance matrix in classical filtering methods

The trajectories on the Sequence 05 of KITTI dataset are from the ground truth (GT), neural vision-only model (VO), neural visual inertial models with direct (VIO), soft (Soft), and hard fusion (Hard). Left: Seq 05 with vision degradation (10% occlusion, 10% blur, and 10% missing data); **Right**: Seq 05 with all degradation (5% for each).





# Evaluation





- Email: changhao.chen@cs.ox.ac.uk