

Visual SLAM: An Online Learning Approach

Hongbin Zha

Key Laboratory of Machine Perception (MOE)

Peking University

zha@cis.pku.edu.cn

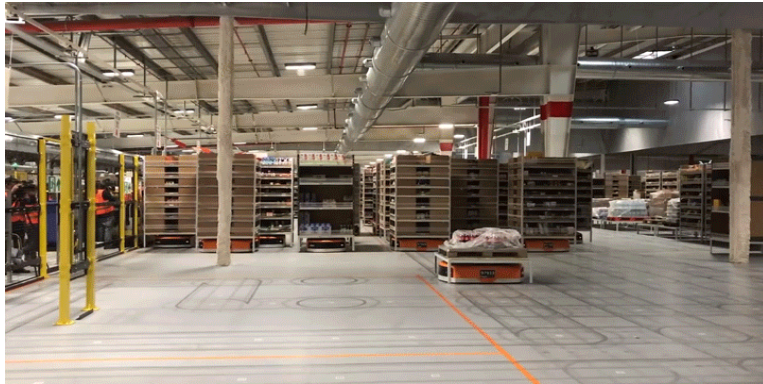
Outline

- ◆ **Introduction to Online Learning for SLAM**
- ◆ **Related Research Topics:**
 - ◆ **Sequential Adversarial Learning for Self-Supervised Deep Visual Odometry**
- ◆ **Conclusions**

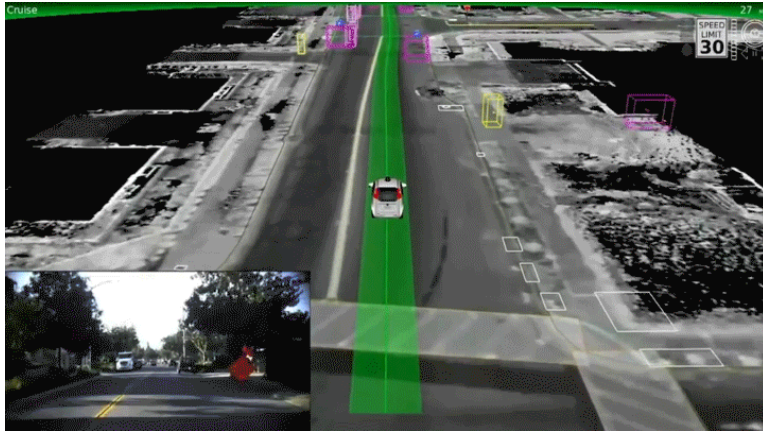
Localization and Mapping

◆ Localization

Odometry of sensors or robot systems



Amazon warehouse robot



Google autonomous driving car

◆ Mapping

3D reconstruction of environments



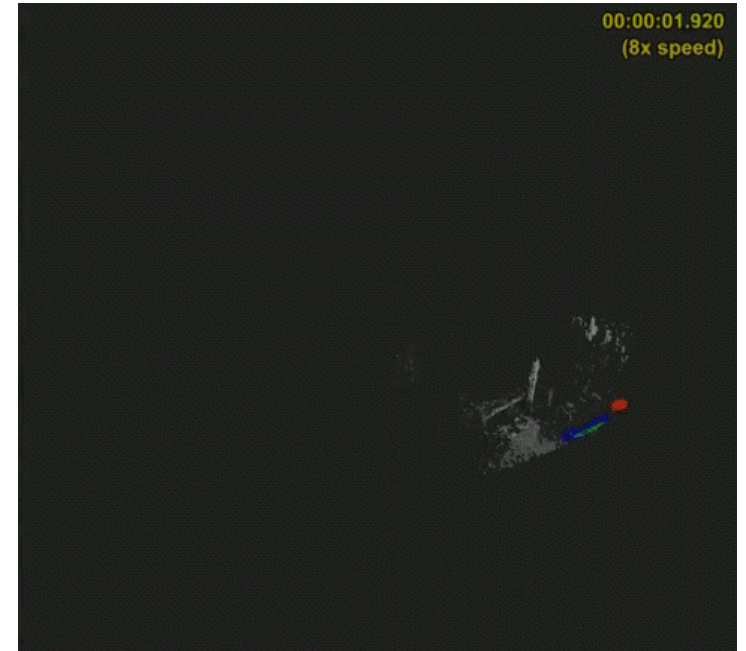
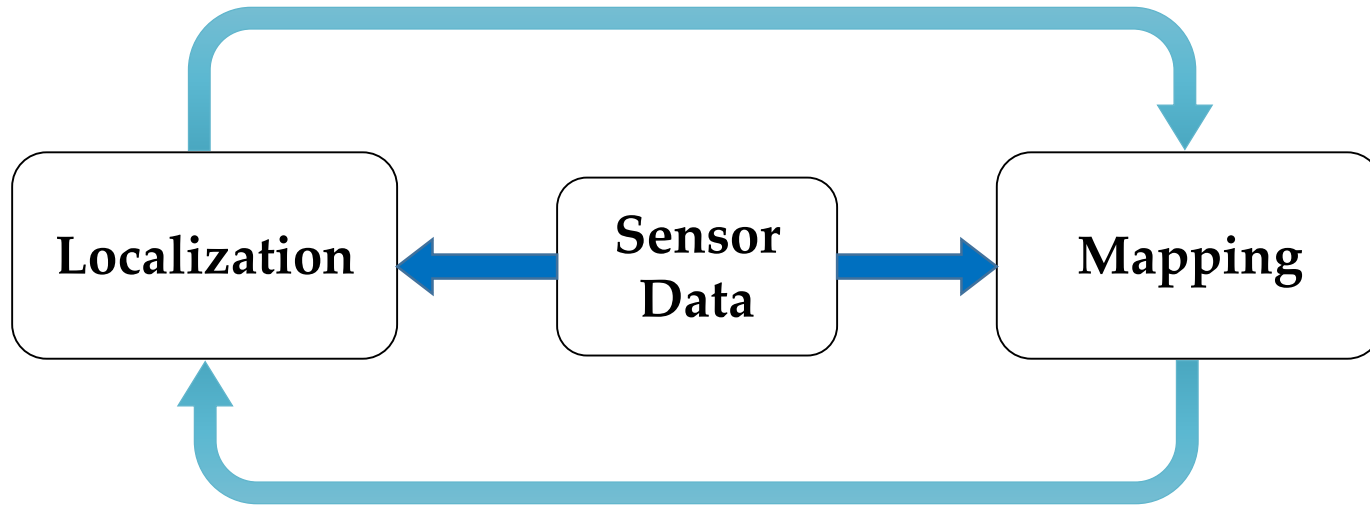
Google earth



Microsoft Holoportation

SLAM: Simultaneous Localization And Mapping

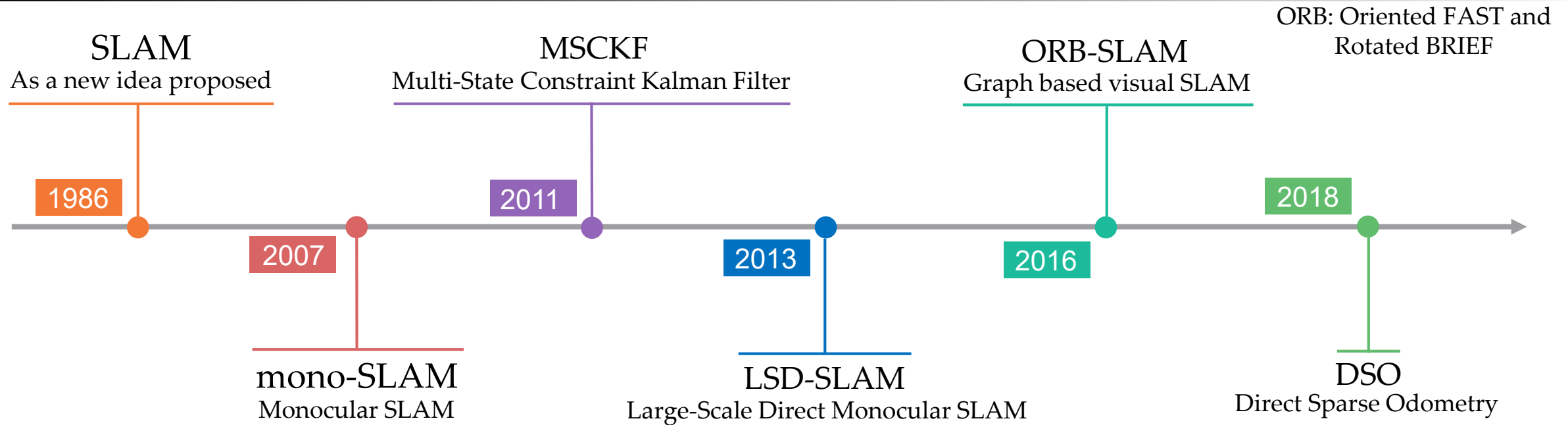
- ◆ Tight coupling of localization and mapping



LSD-SLAM

A fundamental function of dynamic vision systems as for human

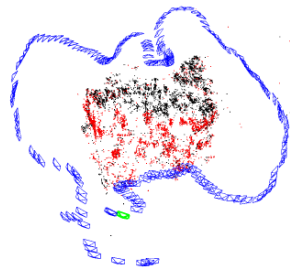
History of SLAM Research



- ◆ Make good use of explicit geometrical relationships between consecutive frames: multi-view geometry
- ◆ Enhance performance by fusing different kinds of sensors
- ◆ Work well in limited environments for specific tasks

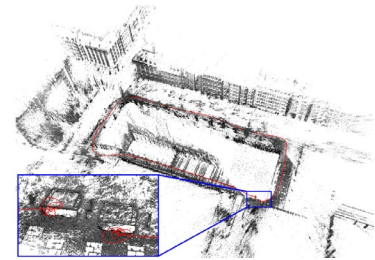
Critical Problem

- ◆ Error accumulation; high computational costs
- ◆ Lack of Systematic Formulation
 - Require delicate hand-crafted design and ad hoc strategies
 - Complex optimizations for different situations with lots of constraints
 - Poor generalization ability to different situations



ORB-SLAM

- Poor performance on texture-less scenes



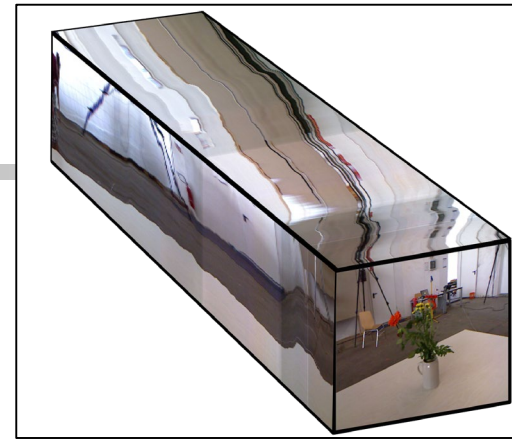
DSO

- Requires photometric calibration
- Not robust to illumination changes

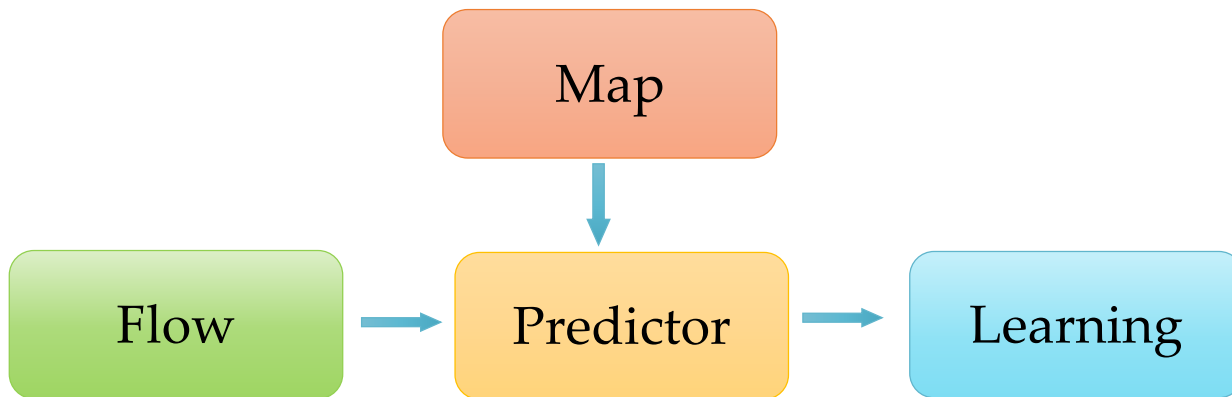
Key Concept: What is Flow?

Sensor Data Flow:

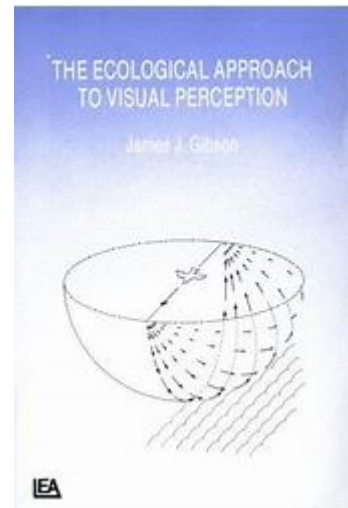
- Continuous motion patterns of time varying sequential data
- Explicit representation of temporal consistency of input data
- Comply to regular patterns according to laws of physics
- **Make the unpredictable, predictable**



Data Flow



James Gibson: The Ecological Approach to Visual Perception, 1986

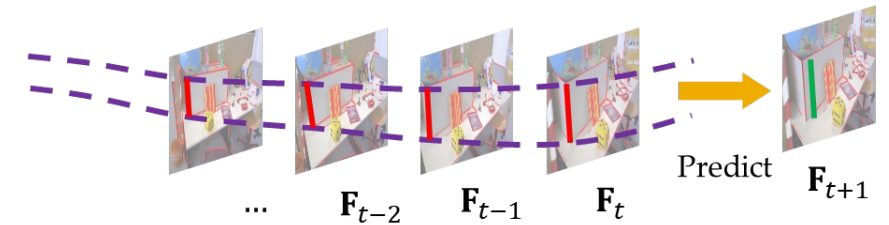


EA

Predictor and Map

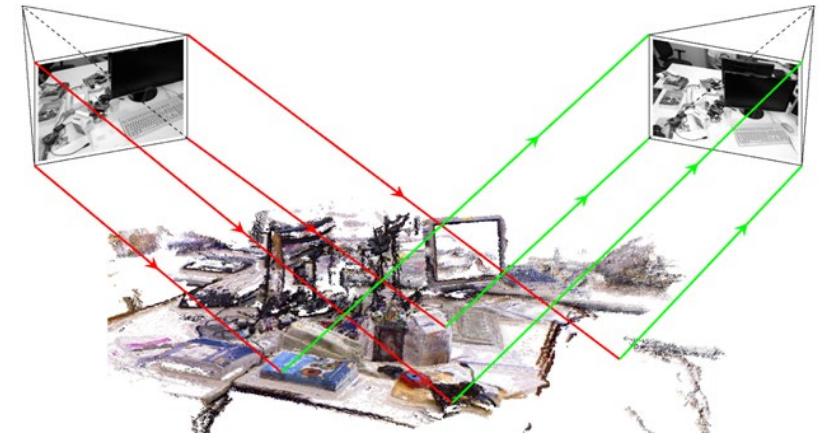
◆ Predictor: the engine of efficient SLAM

- Recurrent state inference: a generative model
- Infer the next state from history
- Provide guidance to online incremental processing



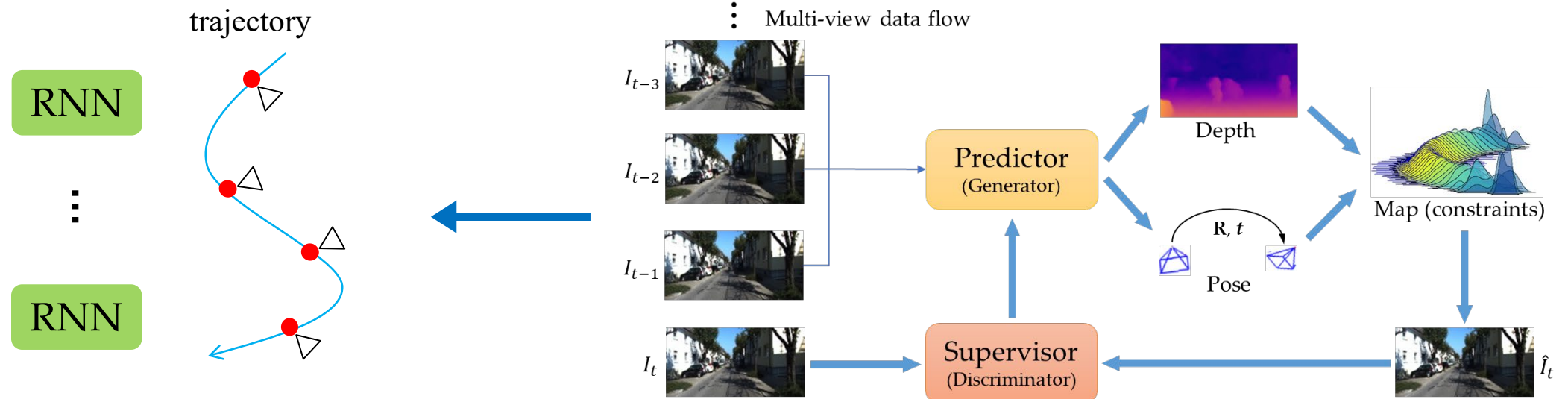
◆ Map: a global, invariant representation

- Implicit/explicit representation of world
- Provide global constraints as regularization
- Supervisory information for prediction



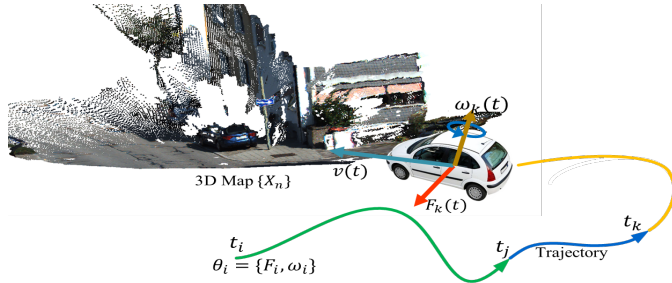
Online Learning: A Systematic Solution

- Learning camera pose in a supervised manner
 - Modelling the data flow using RNN
-
- Self-supervised, online learning
 - Generative Adversarial Networks (GAN)

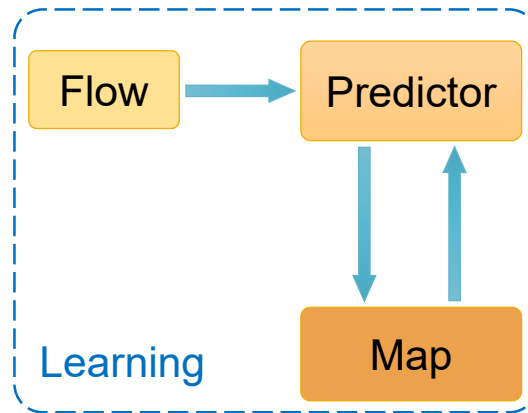
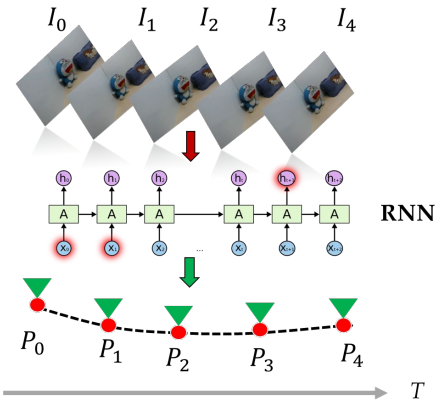


Our Related Research

◆ Dynamics Model

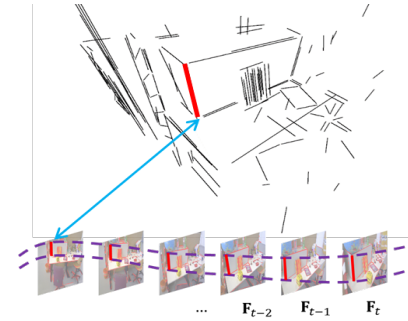


◆ RNN Learning

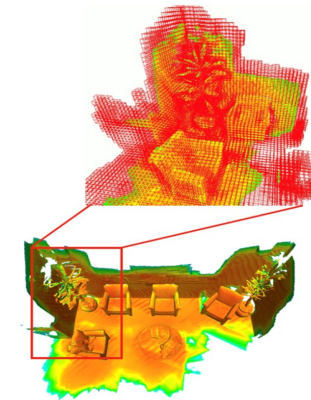


Goal: **Unsupervised online learning** for SLAM

◆ Line Flow



◆ Probabilistic Map Representation



IEEE T-PAMI (accepted), IEEE T-RO (accepted)

ICCV'21, CVPR'21 (2篇), CVPR'20(oral), ICCV'19 (2篇), CVPR'19(oral), ECCV'18, ICRA'18, ICRA'17, 3DV'20, PRL(2019), ACCV'18 (2篇), ICPR'18 (Track Best Paper), IROS'17, BMVC'16

Outline

- ◆ Introduction to Online Learning for SLAM
- ◆ **Related Research Topics:**
 - ◆ **Sequential Adversarial Learning for Self-Supervised Deep Visual Odometry**
- ◆ Conclusions

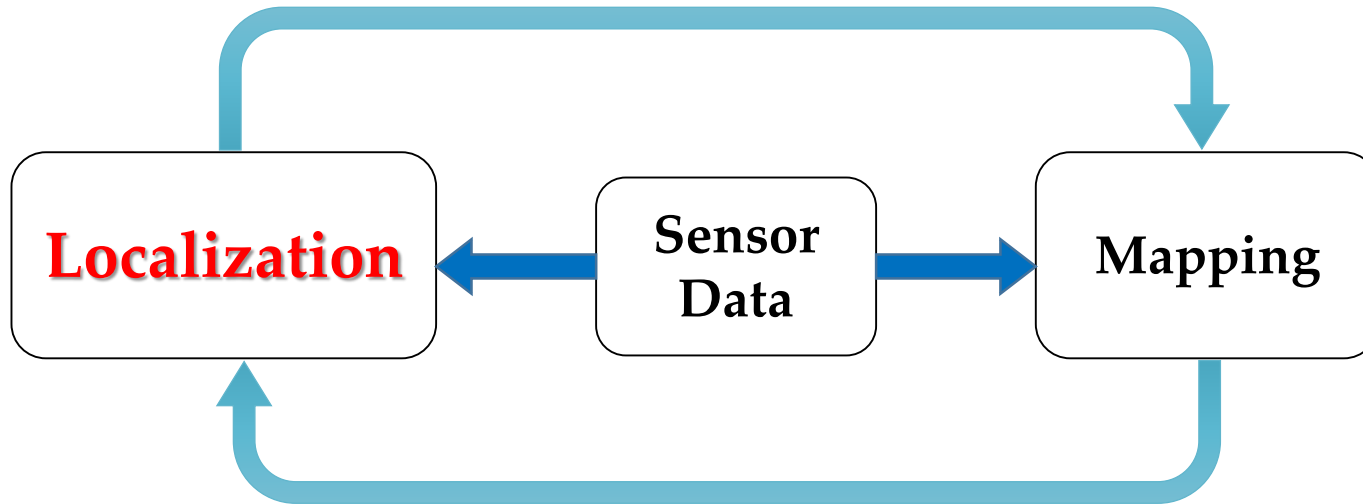
Sequential Adversarial Learning for Self-Supervised Deep Visual Odometry

Shunkai Li, Fei Xue, Xin Wang, Zike Yan, Hongbin Zha

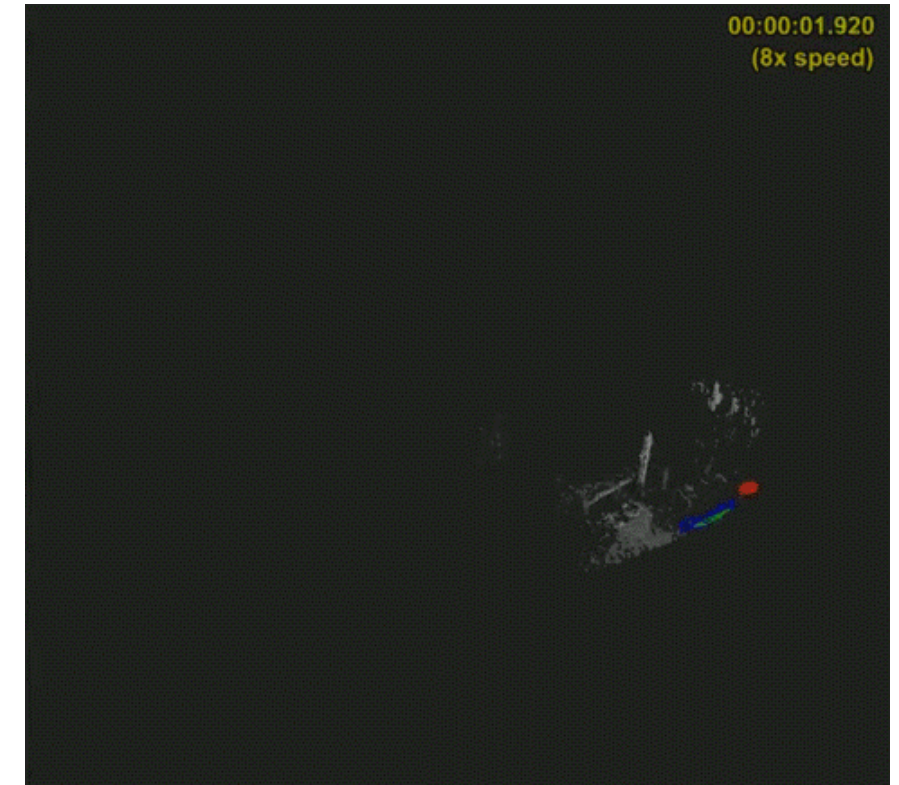
ICCV 2019

Background — VO/SLAM

SLAM: tight coupling of localization and mapping



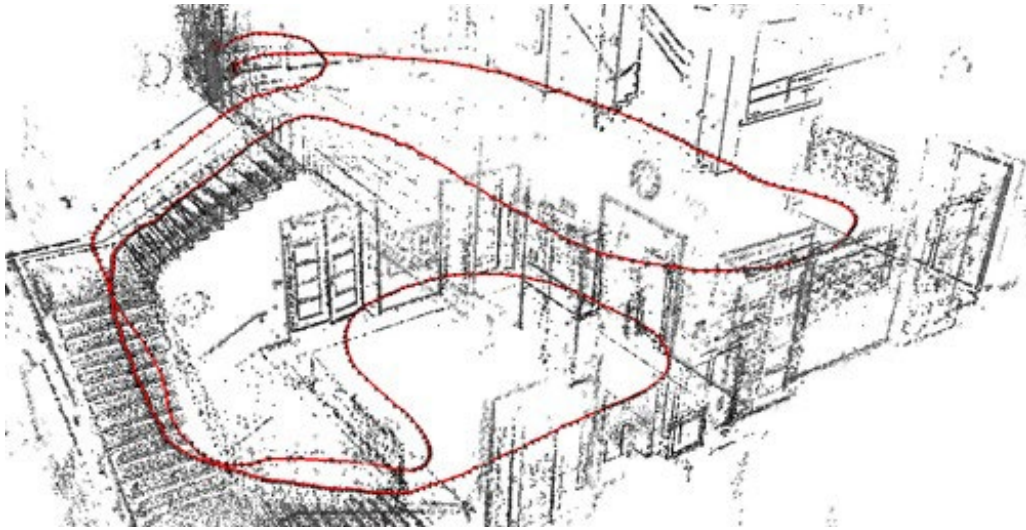
- **Input:** monocular video
- **Output:** depth map and camera pose
- **Applications:** autonomous driving, AR/VR, navigation, robotics



LSD-SLAM

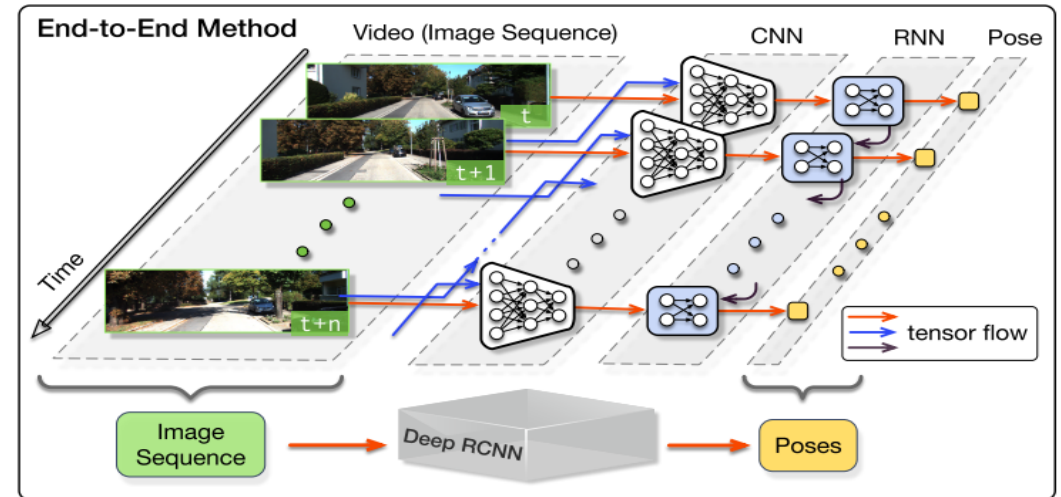
Related Works

- **Classic SLAM/VO**
 - Hard-coded, fixed
 - Rely on low-level features
 - Not robust to complex situations



DSO (PAMI 2018)

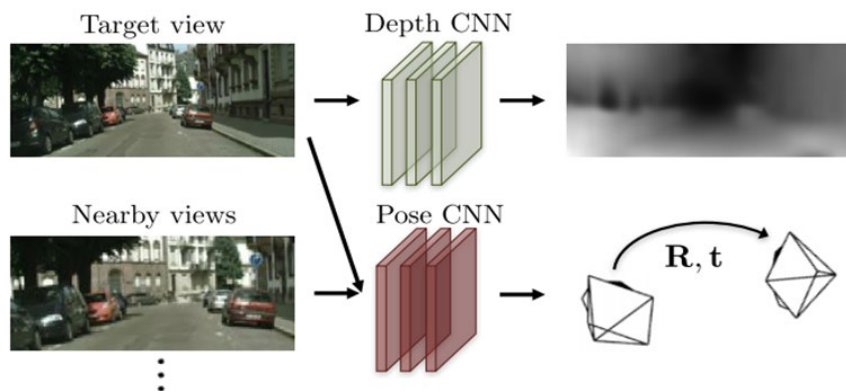
- **Learning-based SLAM/VO**
 - Data driven, adaptive
 - High-level deep features
 - Perform well in complex situations



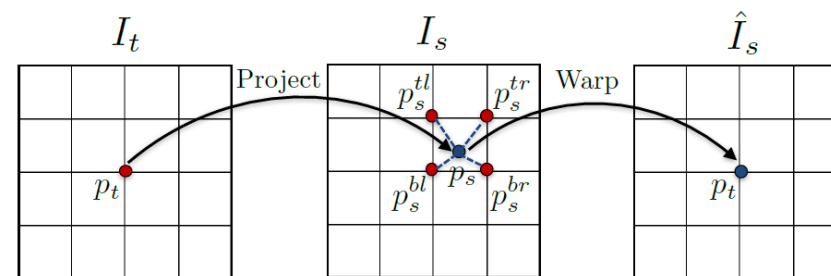
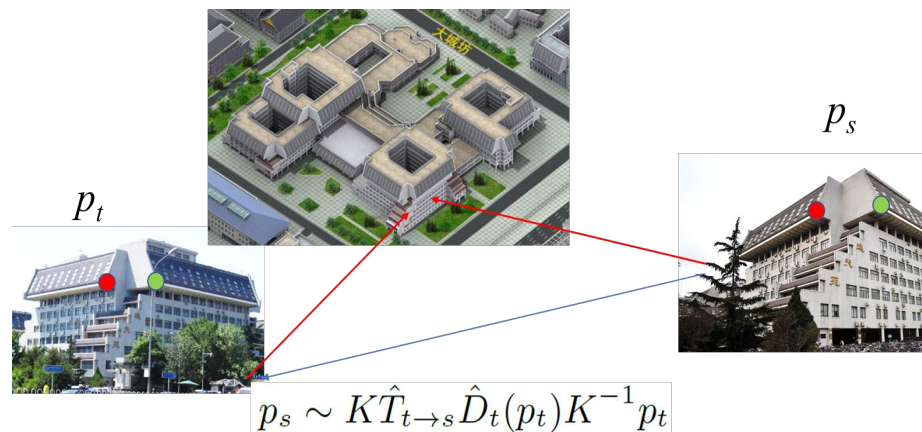
DeepVO (ICRA 2018)

Learning-Based Approach

- Supervised SLAM/VO
 - Good performance
 - Rely on large dataset with labeled data
- Self-supervised VO
 - Exploit geometric correlations of image pairs
 - Minimize photometric loss between predicted and real images



SfMLearner (CVPR 2017 oral)



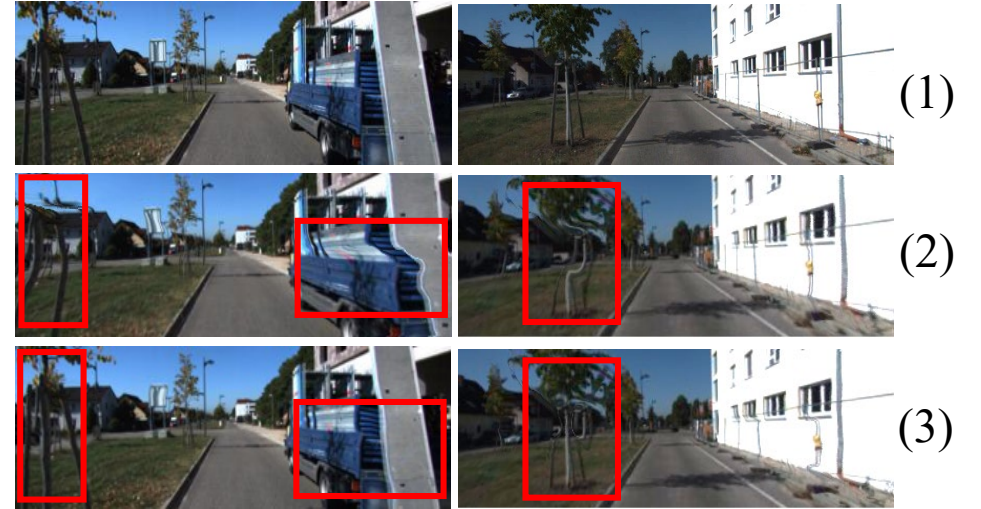
View synthesis by warping

Photometric error

$$\mathcal{L}_{vs} = \sum_s \sum_p |I_t(p) - \hat{I}_s(p)|$$

Challenges

- **Expensive computation for refinement**
 - Dense depth: high dimensionality
 - Require expensive refinement for depth and pose
- **Limitations of photometric loss**
 - Multiple local minima, pixel-level judgement
 - Require global minimum and structural perception



Original images (1) and images synthesized by minimizing **photometric** loss (2) and **GAN** loss (3)

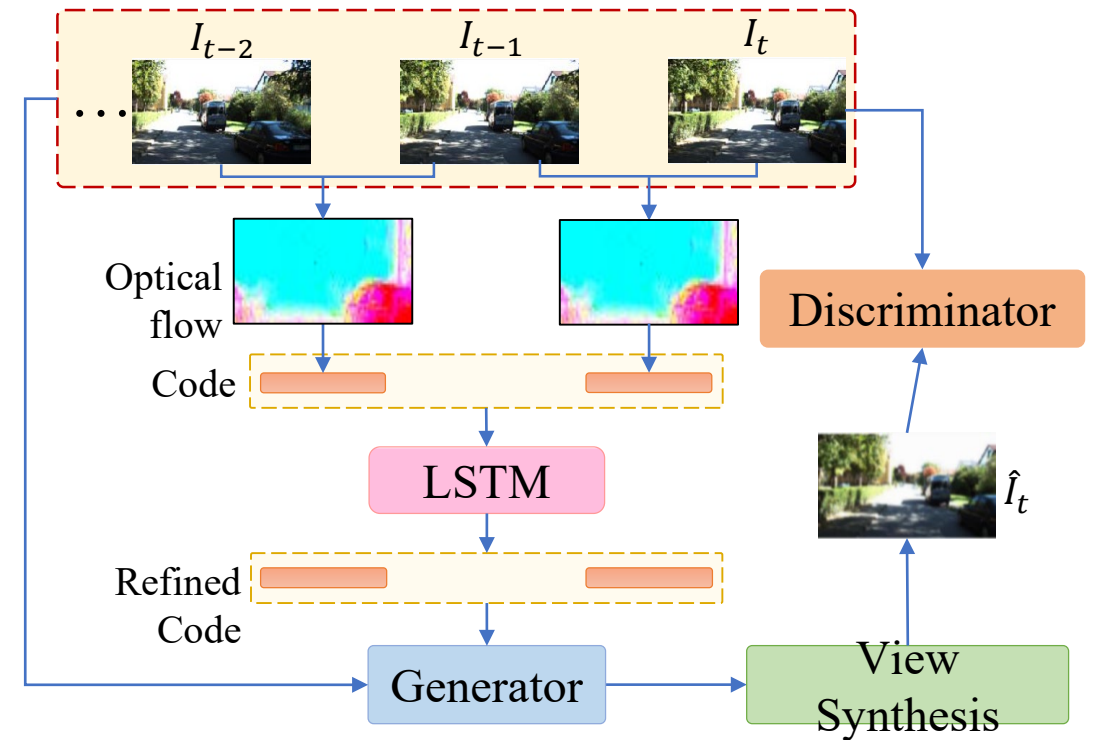
Contributions

- Exploit spatial-temporal correlations over long sequence
 - Reduce accumulated trajectory errors significantly
 - Temporal prediction helps refine current estimation, and vice versa
- Efficient refinement
 - Depth \rightarrow 128D codes
 - Refine dense depth by updating codes via LSTM
- Regard VO in a GAN paradigm
 - Formulate VO as a self-supervised **generator**
 - Discriminator: **learn** structural features via adversarial training
 - Overcome problems in pixel-level photometric loss

Overview

Main processes:

- Estimate depth from 2 images by optical flow
- Encoding depth into 128D codes
- Refining previous estimations by LSTM
- VO: Generator
- Learned loss: Discriminator



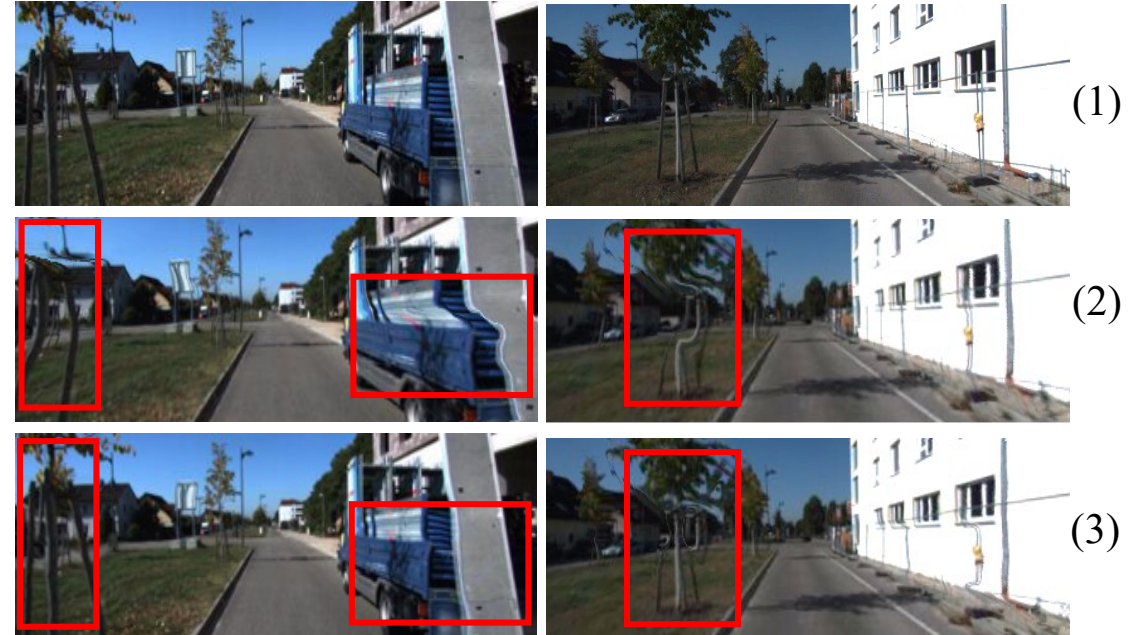
VO as a Generative Model

VO: a generator conditioned on:

- 1) adjacent images
- 2) estimated poses
- 3) continuous latent depth 128D code

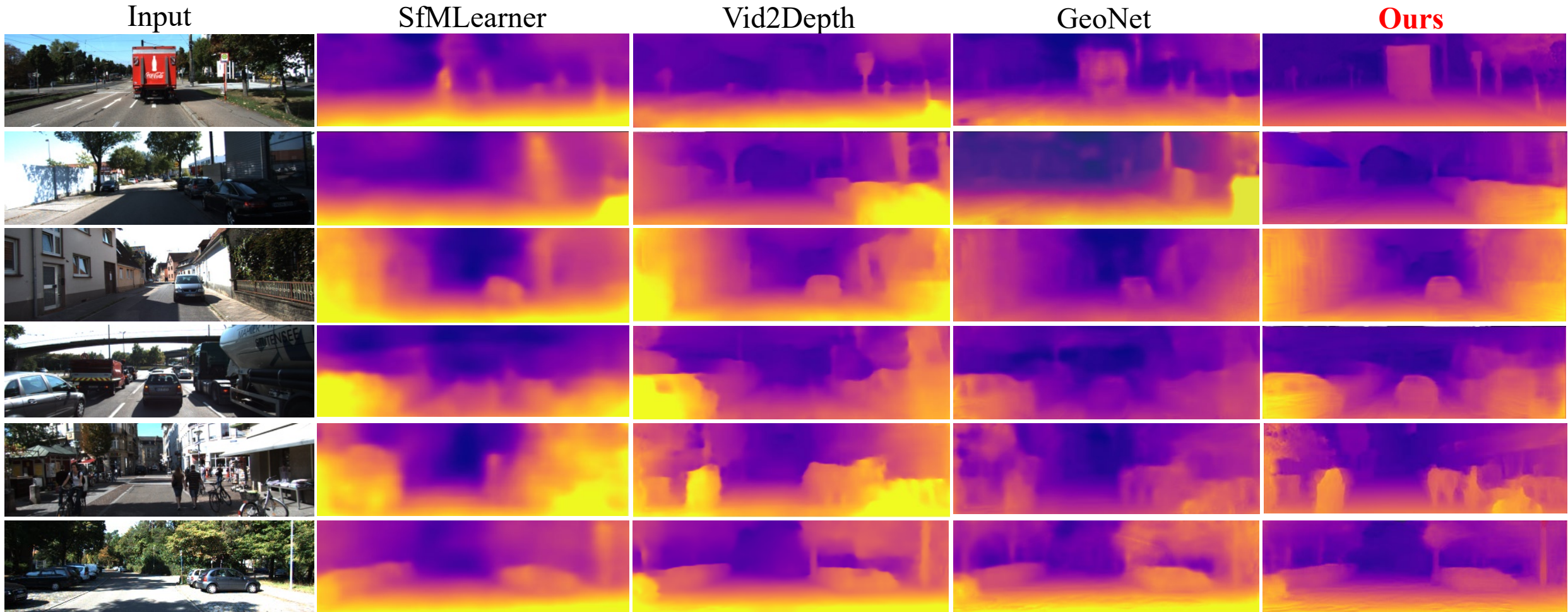
Problem: How to **design** a proper loss to assess reconstruction quality considering occlusions, moving objects and illumination changes

Solution: **learn** a loss with structural perception by adversarial learning



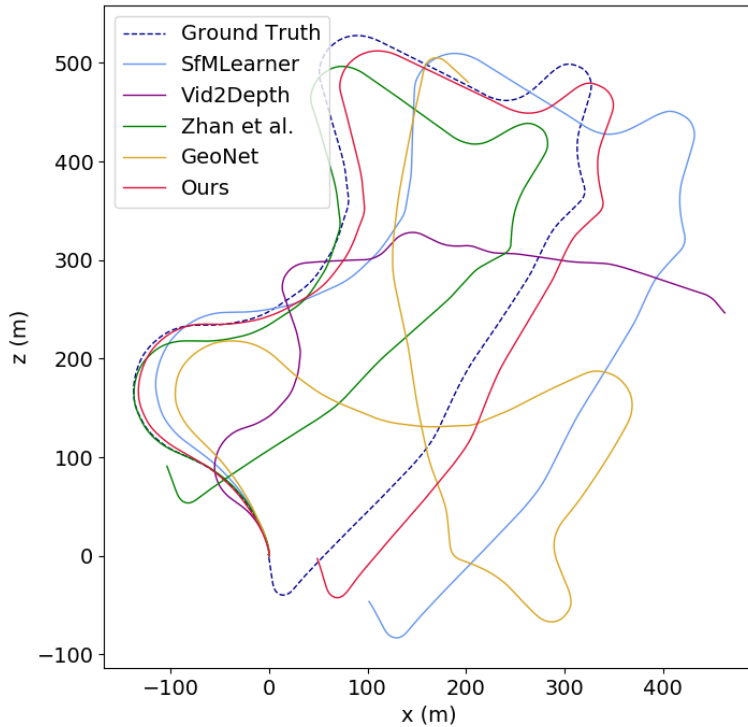
Original images (1) and images synthesized by minimizing **photometric** loss (2) and **GAN** loss (3)

Experiments: Depth

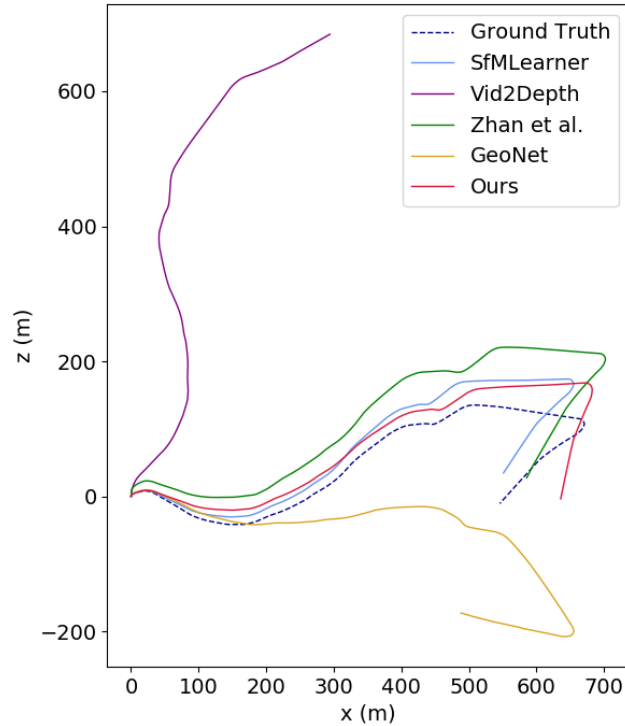


Experiments: Pose

KITTI Seq. 09



KITTI Seq. 10



Absolute Trajectory Error (ATE)

Method	Seq.09	Seq.10
ORB-SLAM [28] (short)	0.064±0.141	0.064±0.130
ORB-SLAM [28] (full)	0.014±0.008	0.012±0.011
SfMLearner [39]	0.021±0.017	0.020±0.015
SfMLearner [39] modified ¹	0.016±0.009	0.013±0.009
Zhan <i>et al.</i> [37]	0.013±0.009	0.013±0.008
Vid2Depth [27]	0.013±0.010	0.012±0.011
GeoNet [36]	0.012±0.007	0.012±0.009
Ours	0.0029±0.0012	0.0028±0.0012

Take Home Message

Contributions:

- **Sequential**: exploit spatial-temporal correlations over long sequence to reduce accumulated error and achieve efficient refinement by LSTM
- **Adversarial**: formulate VO as a generative model and learn a loss by adversarial training

Impact and future work:

- Accurate and explainable video prediction
- Self-supervised bundle adjustment (BA)
- Probabilistic multi-view depth refinement

Conclusions

- ◆ Dynamic vision and SLAM are central topics in 3D vision
- ◆ **The temporal consistency of sensor data and the incremental characteristics of processing** have to be used effectively
- ◆ New learning approaches to dynamic vision have to be developed, which are different from the current deep-learning methodology
- ◆ We should embed in the systems mechanisms such as **self-adaptation**, as recent cognitive science suggested
- ◆ On-line processing performance will be a critical element, thus requiring powerful **on-line learning techniques**



Thank you!