

# **A New Approach to SLAM: Online Learning Paradigm**

**Hongbin Zha**

Key Laboratory of Machine Perception (MOE)

Peking University

[zha@cis.pku.edu.cn](mailto:zha@cis.pku.edu.cn)

# Outline

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

# Outline

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

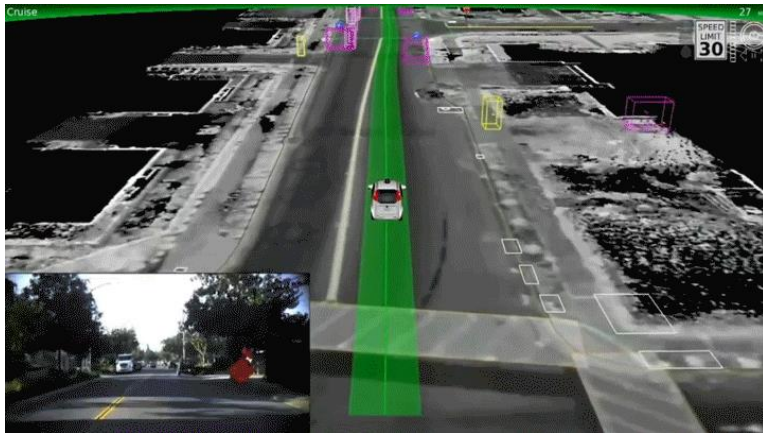
# Localization and Mapping

## ◆ Localization

Odometry of sensors or robot systems



Amazon warehouse robot



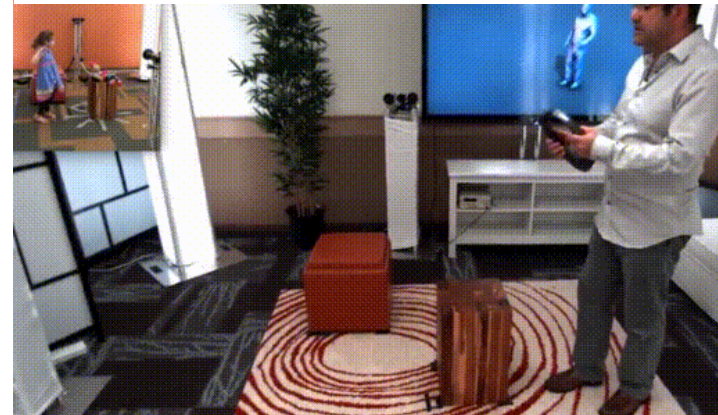
Google autonomous driving car

## ◆ Mapping

3D reconstruction of environments



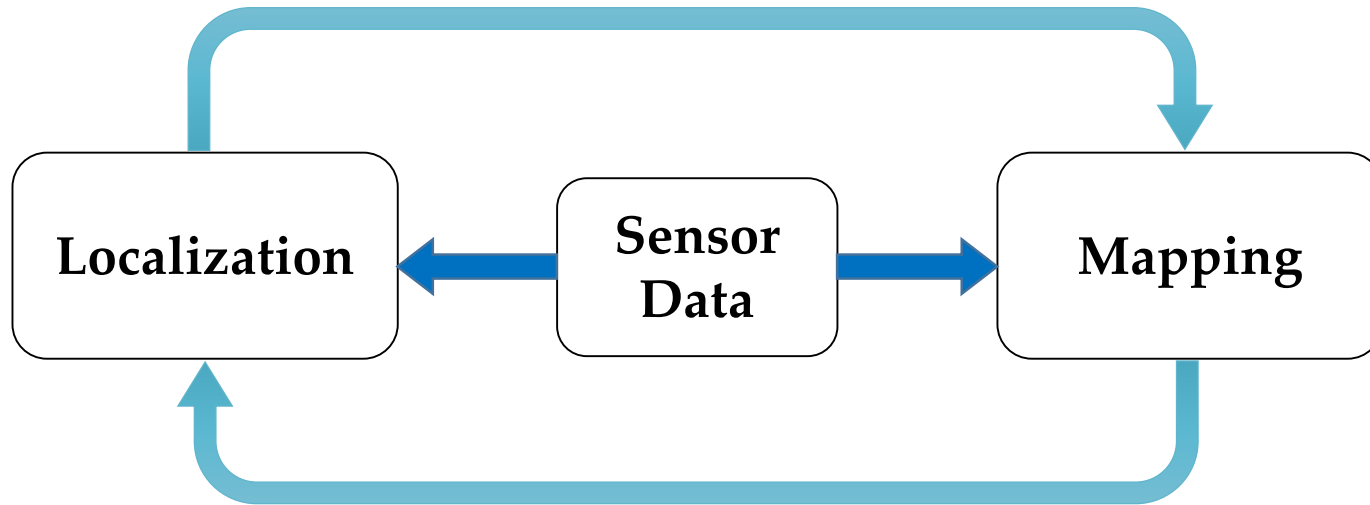
Google earth



Microsoft Holoportation

# SLAM: Simultaneous Localization And Mapping

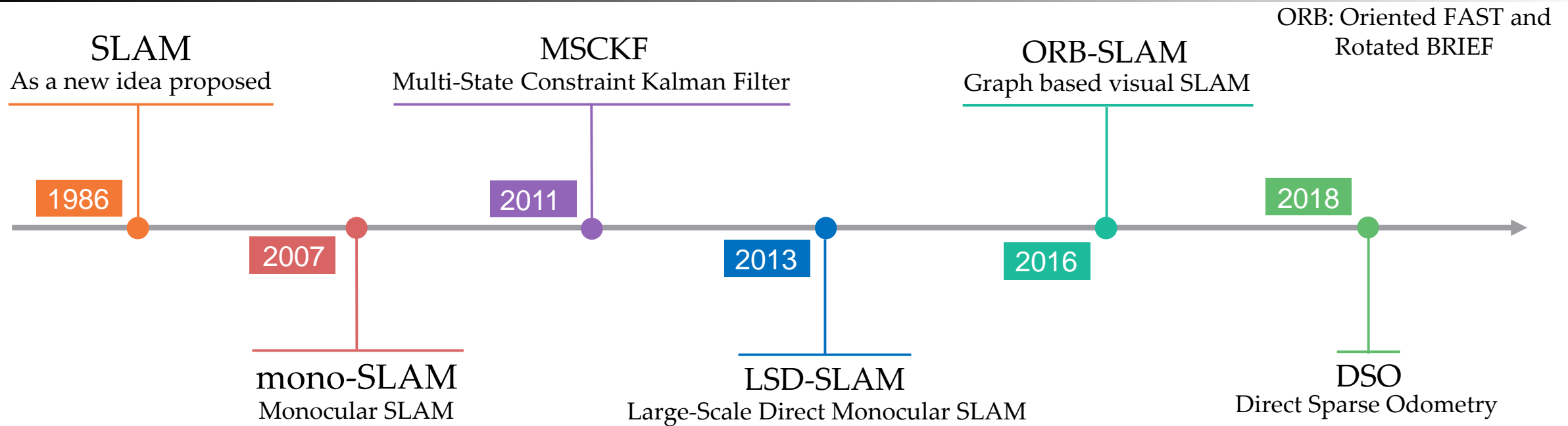
- ◆ Tight coupling of localization and mapping



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

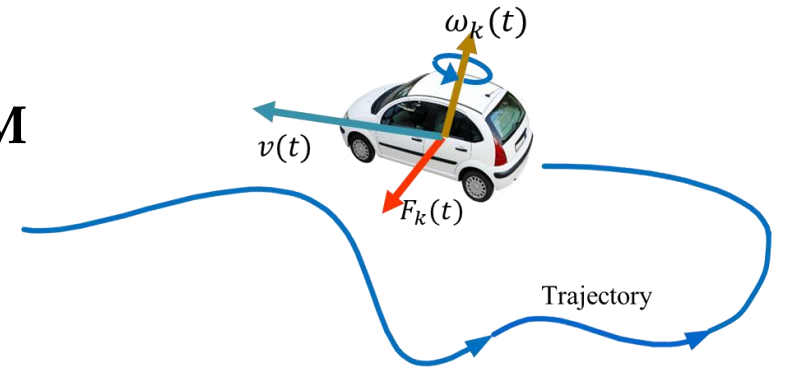
# Is Current SLAM Good Enough?

## ◆ Pay Little Attention to Temporal Continuity

- Disregarding spatial-temporal consistency inherent in SLAM
- A big source of accumulated errors
- Low robustness for feature tracking

## ◆ Rely Too Strongly on Pixel Correspondence

- Unable to use scene structures: line, surfaces, super-pixels
- Poor performance on texture-less scenes
- Difficult to transform 3D maps to structural descriptions



# Is Current SLAM Good Enough?

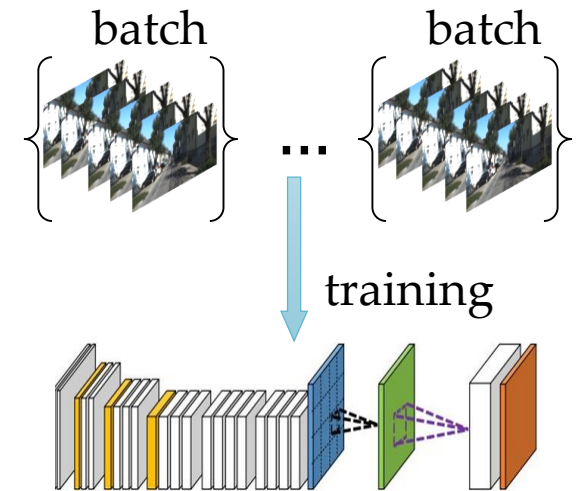
## ◆ High Computational Cost

- Complicated optimization processes
- High demand for hardware
- Limited real-time applications



## ◆ Learning Approach and SLAM

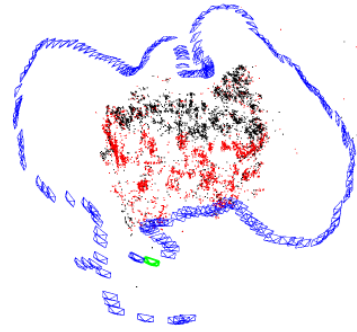
- Off-line batch training instead of online adaptation
- Supervised learning requiring massive labeled data
- Tedious parameter tuning
- Poor generalization ability



# Critical Problem

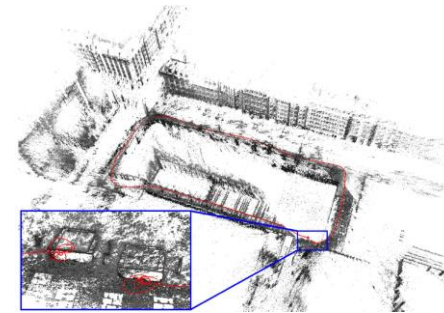
## ◆ Lack of Systematic Formulation

- 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
- Extensive computation caused by optimization



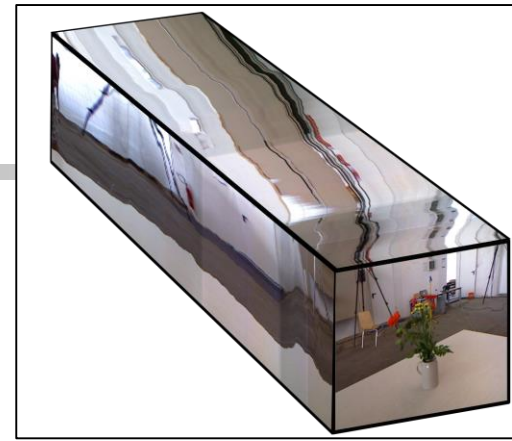
DSO

- Requires photometric calibration
- Not robust to illumination changes

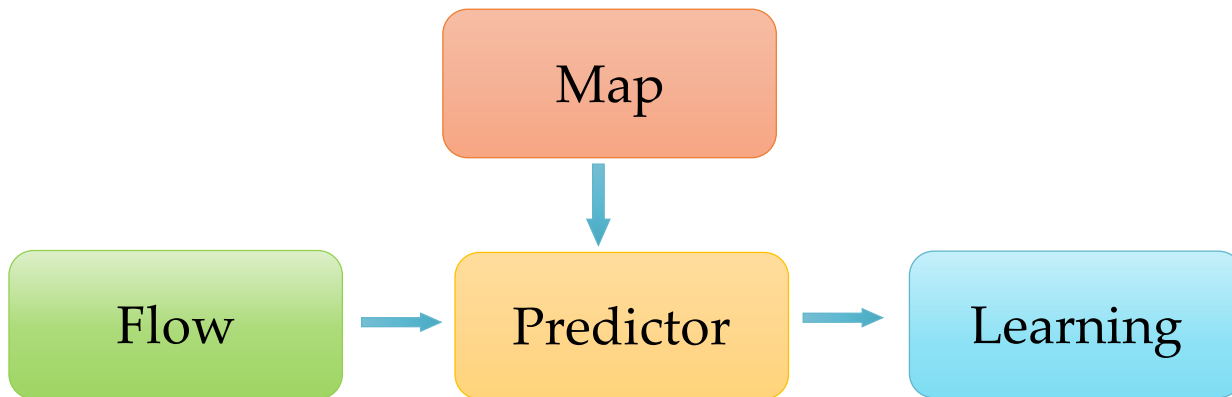
# What is Flow?

## Sensor Data Flow:

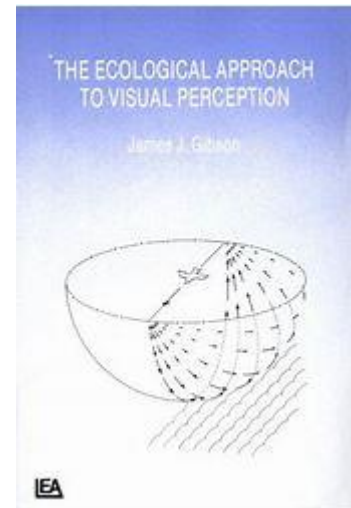
- Continuous motion patterns of time varying sequential data
- Explicit representation of temporal consistency of input data
- Comply to regular patterns according to laws of physics
- **Make the unpredictable, predictable**



Data Flow



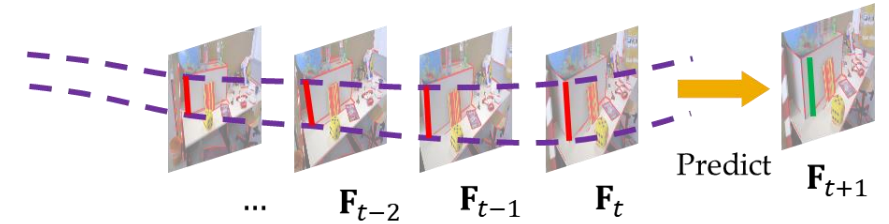
James Gibson: The Ecological Approach to Visual Perception, 1986



# Predictor and Map

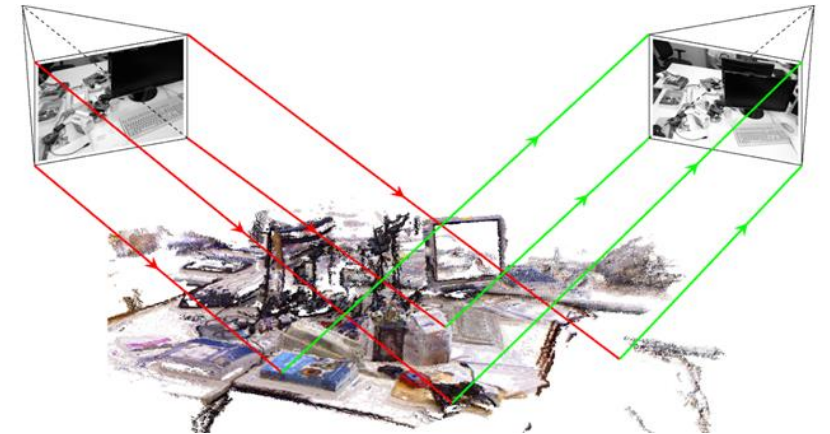
## ◆ Predictor: the engine of efficient SLAM

- Recurrent state inference: a generative model
- Infer the current state from history
- Provide guidance to reduce computational cost



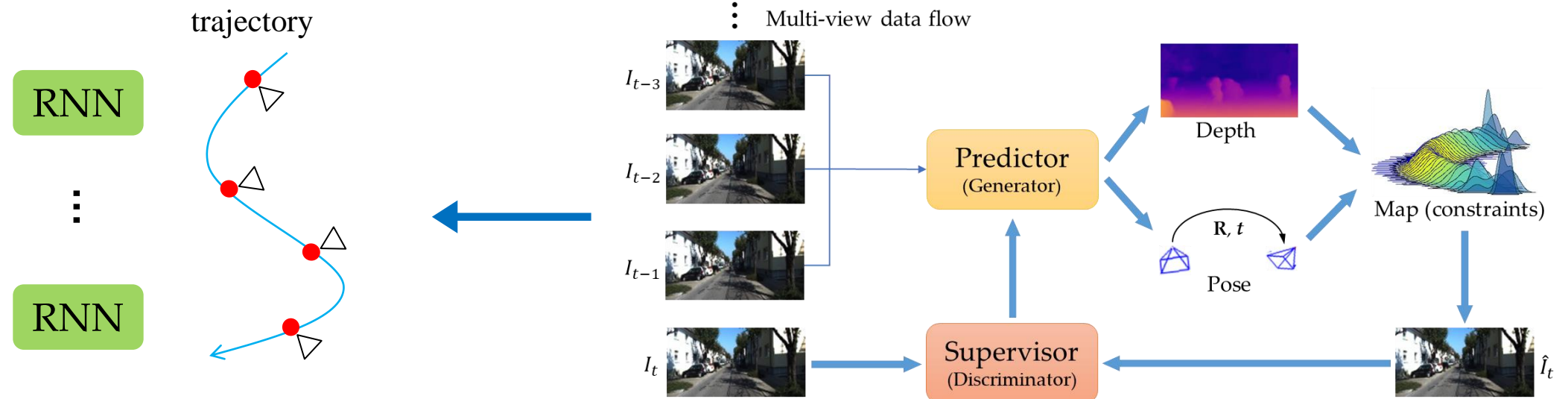
## ◆ Map: a global, invariant representation

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



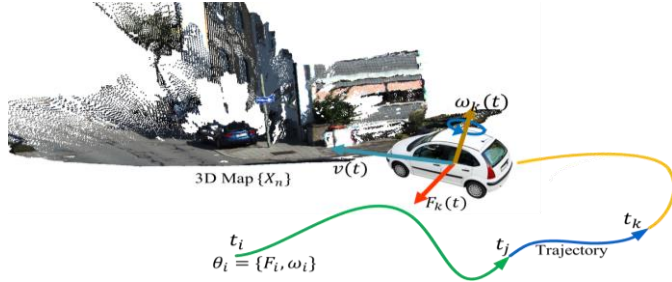
# Learning: A Systematic Solution

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

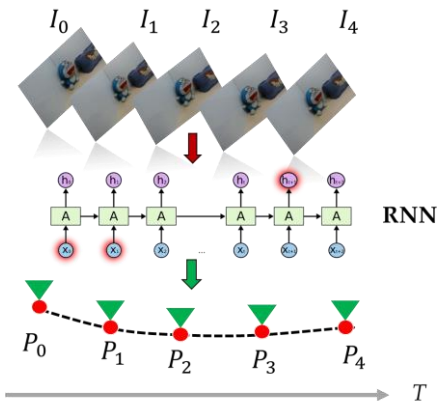


# Our Related Research

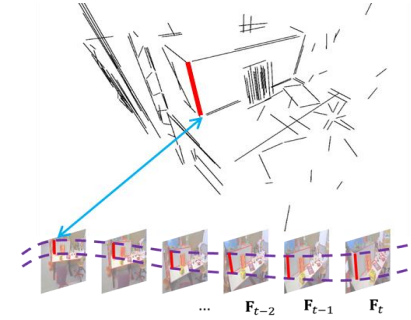
## ◆ Dynamics Model



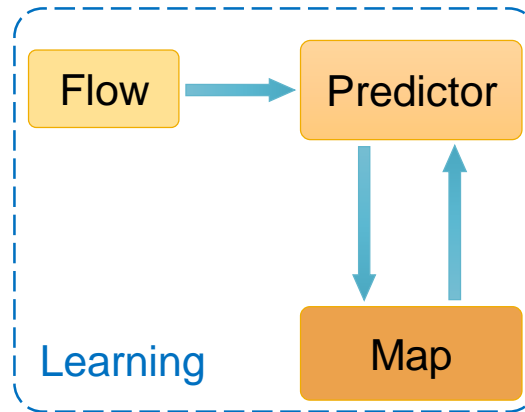
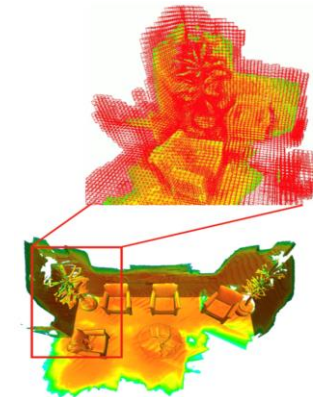
## ◆ RNN Learning



## ◆ Line Flow



## ◆ Probabilistic Map Representation



Goal: **Unsupervised online learning** for SLAM

ICCV'20, ICCV'19 (2篇), CVPR'19, ECCV'18, ICRA'18, ICRA'17  
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
  - ◆ Self-Supervised Deep Visual Odometry with Online Adaptation
- ◆ Conclusions

---

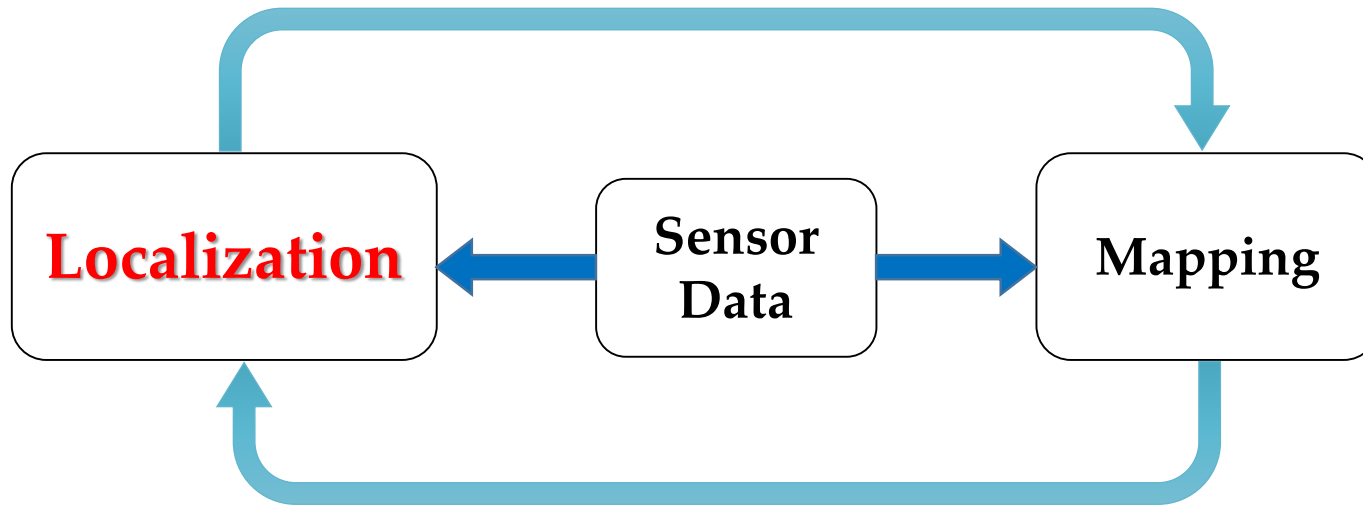
# **Sequential Adversarial Learning for Self-Supervised Deep Visual Odometry**

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

ICCV 2019

# Background — VO/SLAM

SLAM: tight coupling of localization and mapping



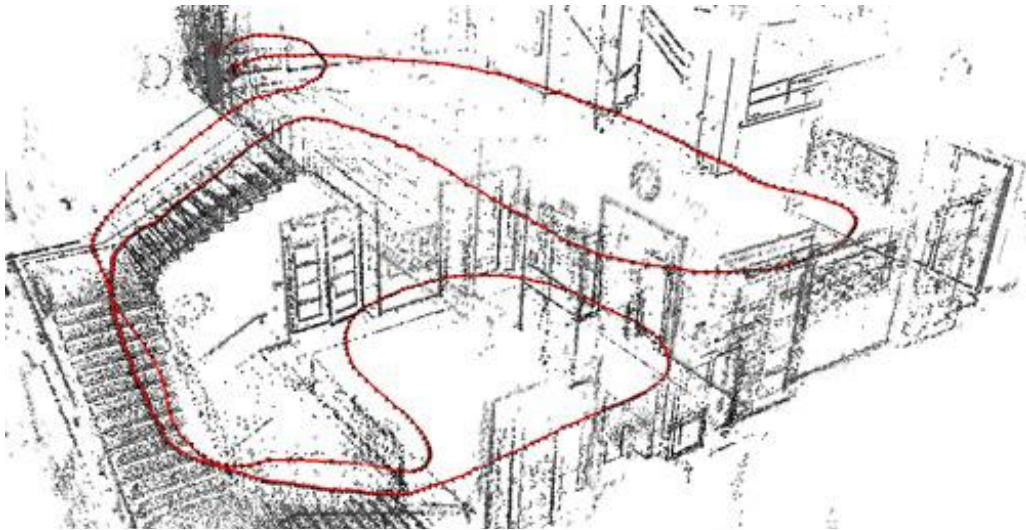
- **Input:** monocular video
- **Output:** depth map and camera pose
- **Applications:** autonomous driving, AR/VR, navigation, robotics



LSD-SLAM

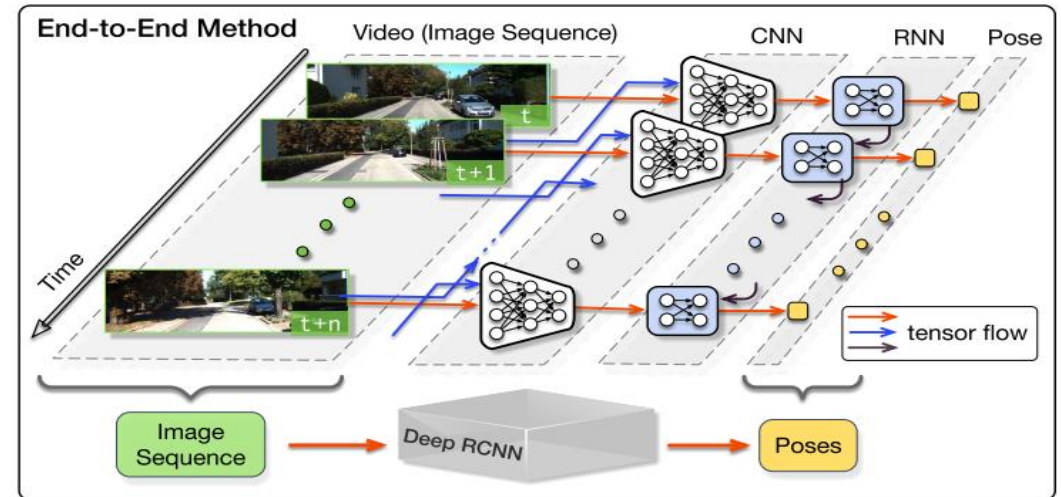
# Related Works

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



DSO (PAMI 2018)

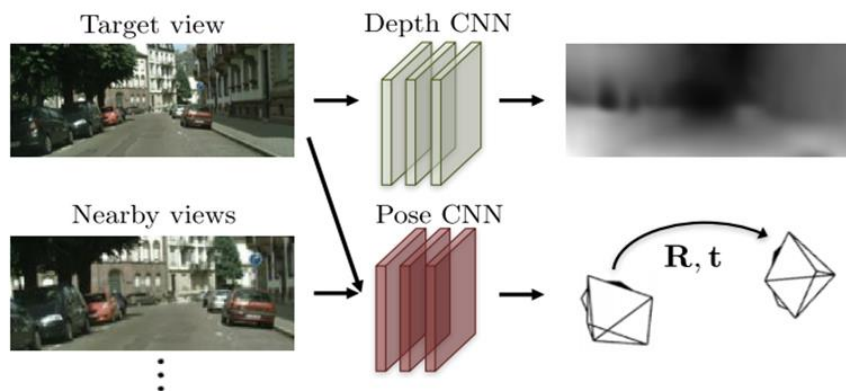
- **Learning-based SLAM/VO**
  - Data driven, adaptive
  - High-level deep features
  - Perform well in challenging situations



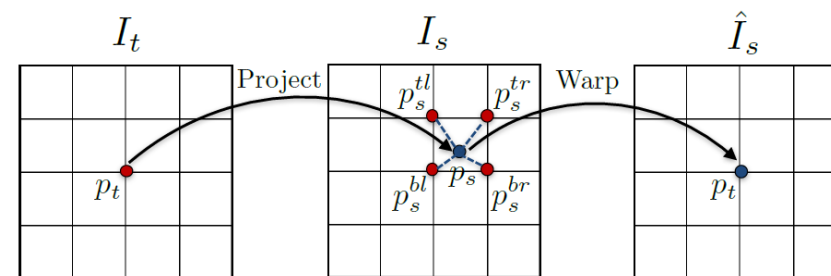
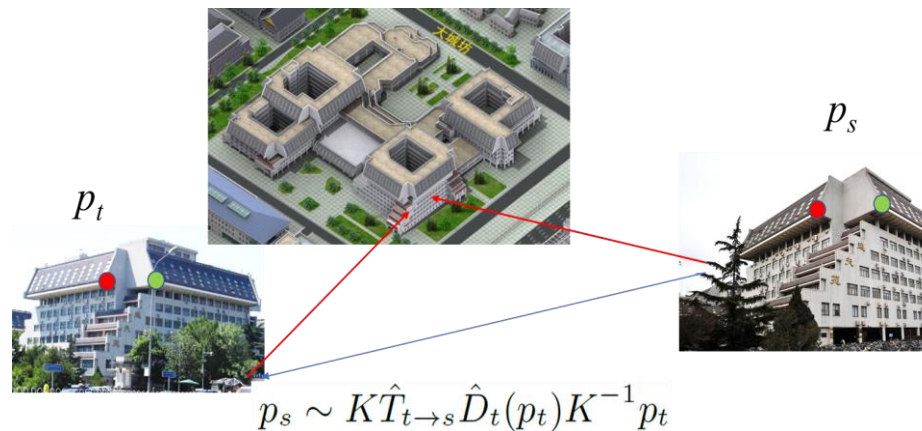
DeepVO (ICRA 2018)

# 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)



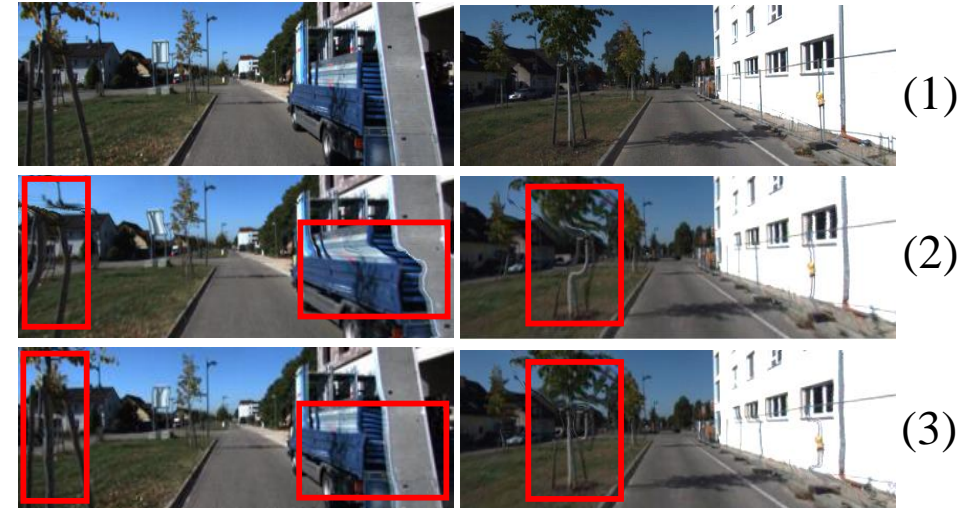
View synthesis by warping

## 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 efficient refinement for depth and pose
- **Limitations of photometric loss**
  - Multiple local minima, pixel-level judgement
  - Require global minimum and structural perception



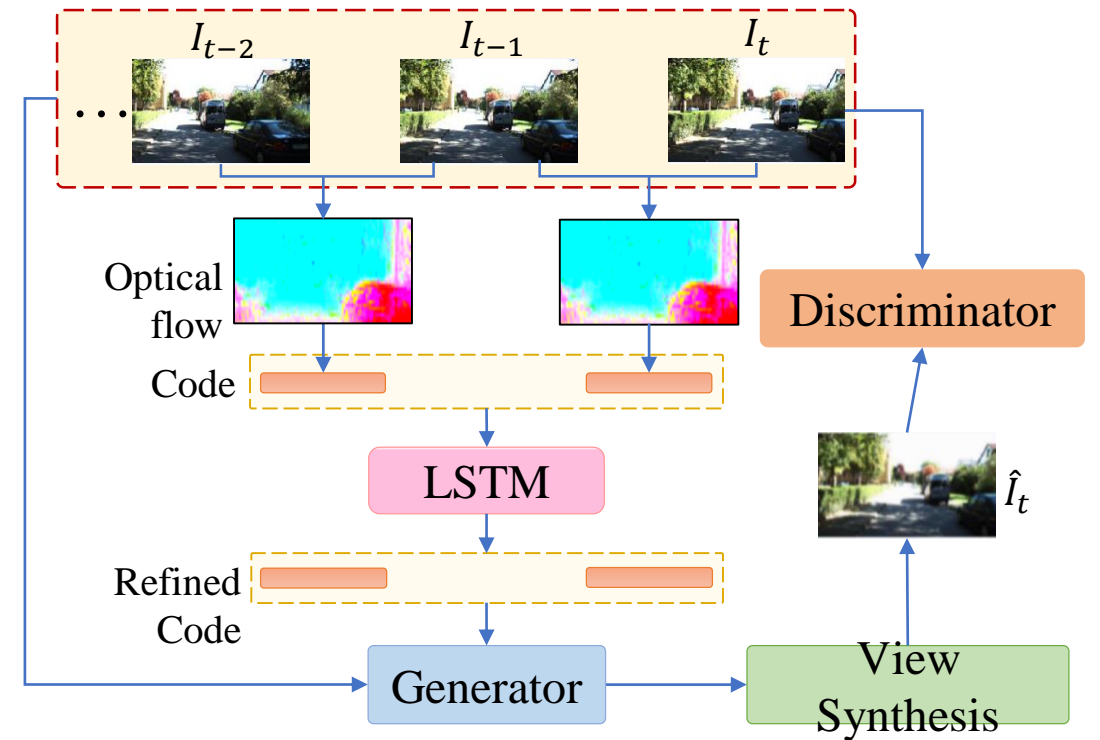
# Contributions

- Exploit spatial-temporal correlations over long sequence
  - Reduce accumulated trajectory errors significantly
  - Historical estimations help 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



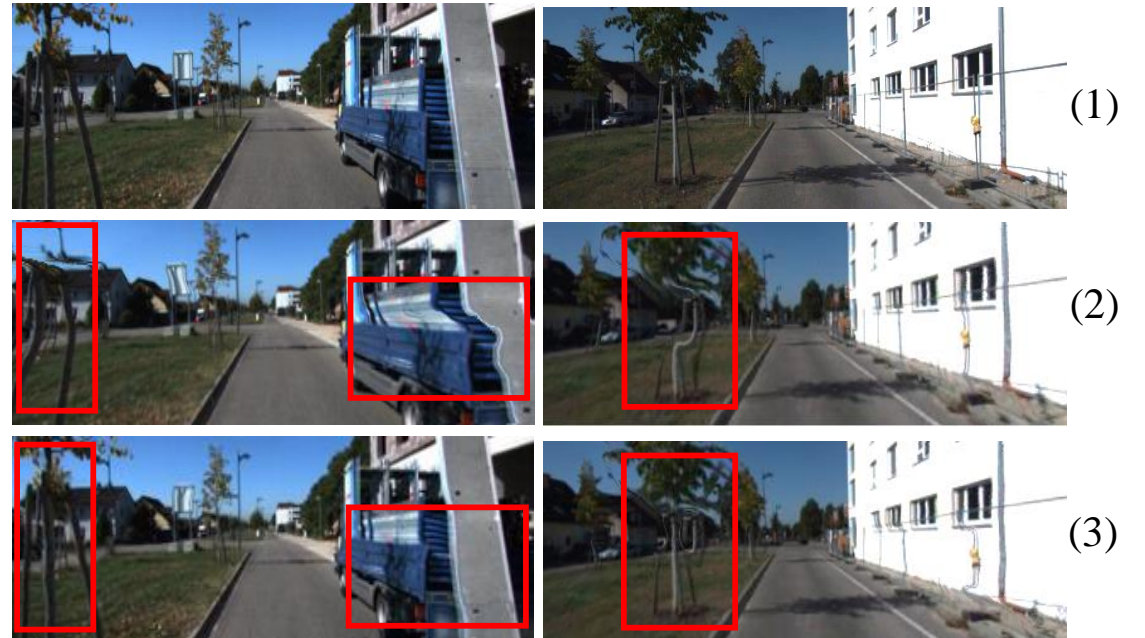
# VO as a Generative Model

VO: a generator conditioned on:

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

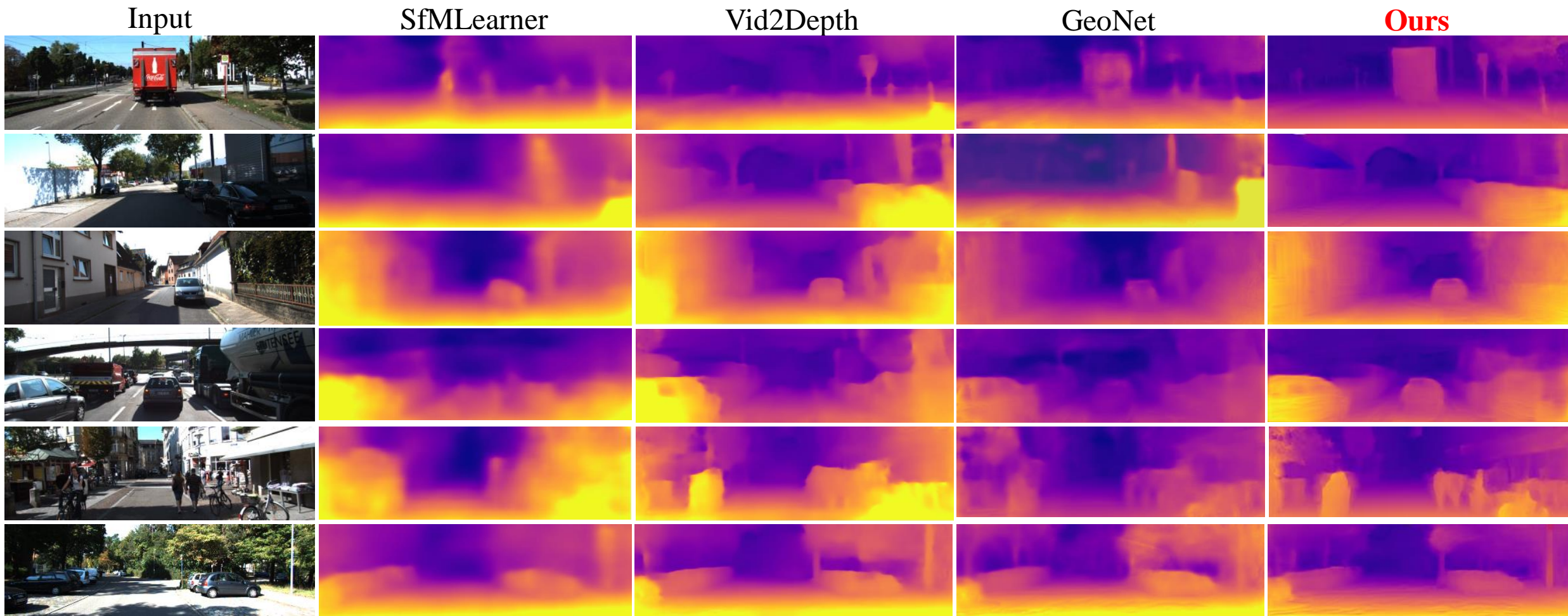
**Problem:** How to **design** a proper loss to assess reconstruction quality considering occlusions, moving objects and illumination changes

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



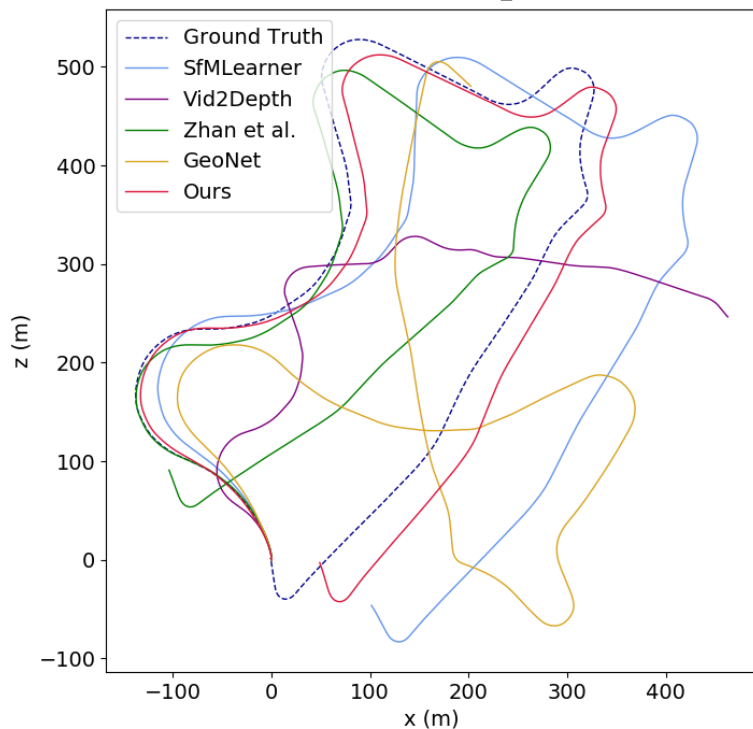
Original images (1) and images synthesized by minimizing **photometric** loss (2) and **GAN** loss (3)

# Experiments: Depth

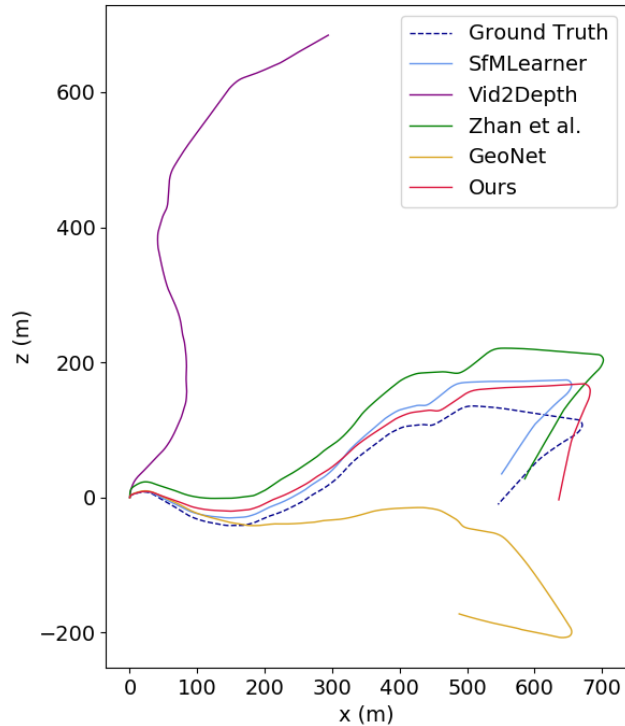


# Experiments: Pose

## 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 <sup>1</sup>	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	<b>0.0029±0.0012</b>	<b>0.0028±0.0012</b>

# 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

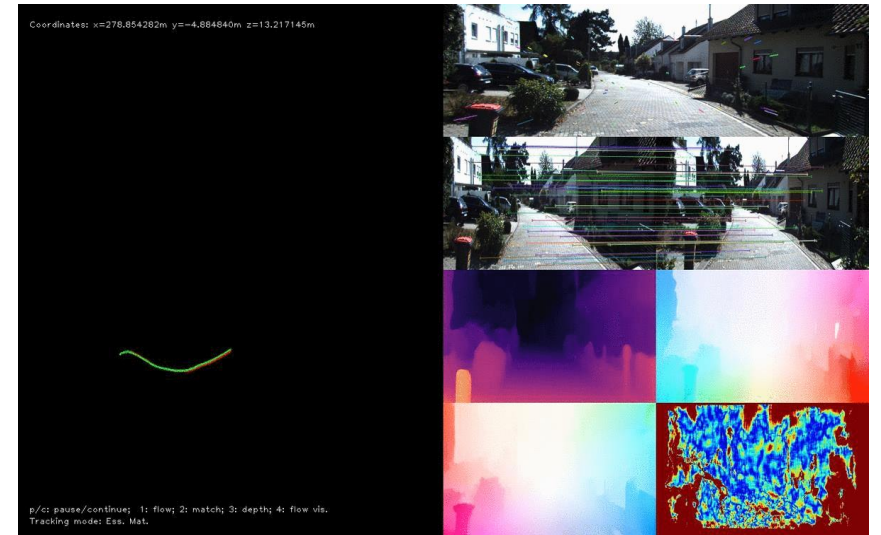
---

# **Self-Supervised Deep Visual Odometry with Online Adaptation**

Shunkai Li, Xin Wang, Yingdian Cao, Fei Xue, Zike Yan, Hongbin Zha  
CVPR 2020 (oral)

# Background

- Learning-based VO has achieved great success in recent years
- In new situations, however, many applications can only use networks pre-trained in datasets different from the situations
- Existing methods often perform bad in this case, and thus require efficient online adaptation

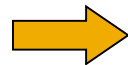


DF-VO (ICRA 2020)

Virtual Carla simulator



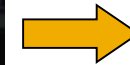
Pretrain



Outdoor KITTI



Online test



Indoor TUM



Online test

# Online Learning

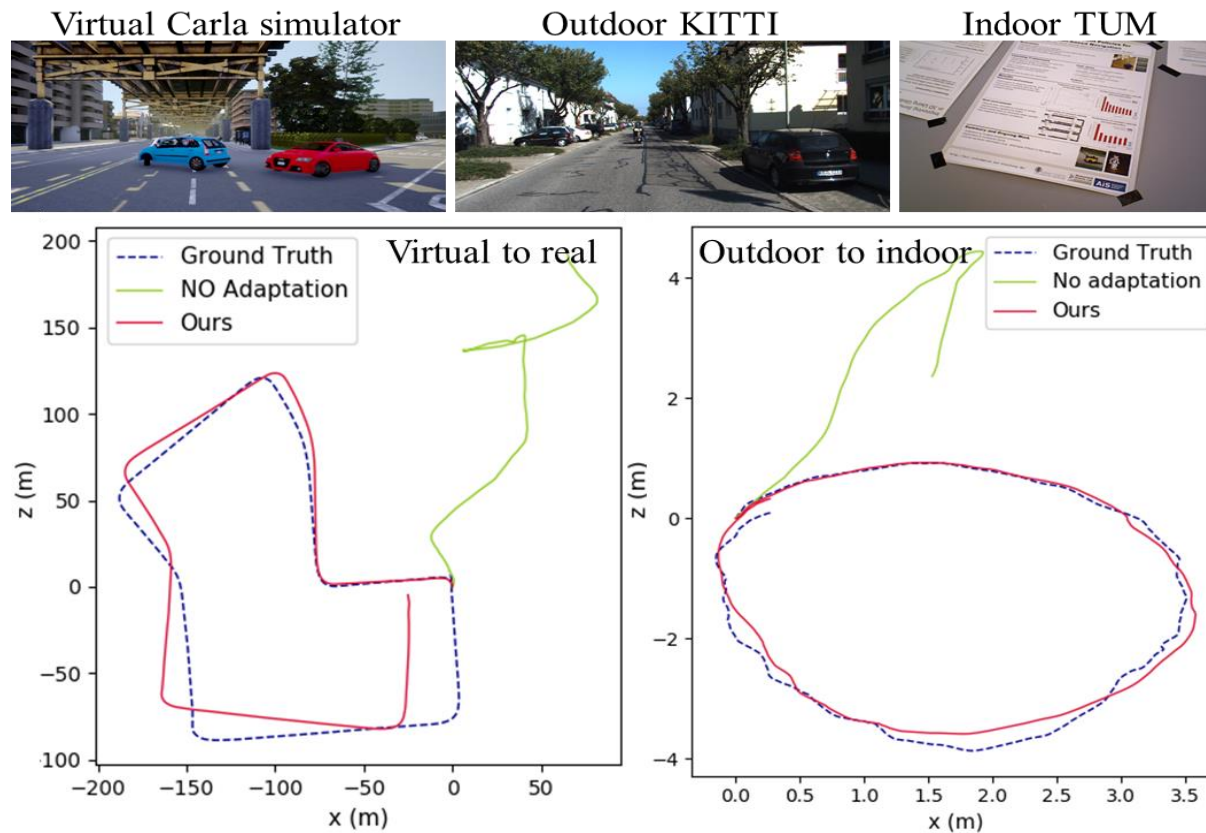
- Basic assumption of machine learning: the training and test data are i.i.d. sampled from the same distribution
- However, the online captured frames no longer satisfy i.i.d. assumption as the camera is continuously moving in the changing environment
- We thus should use online learning to deal with changing data distributions for domain adaptation



# Challenges

Naïve online learning suffers from

- **slow convergence**
- **over-fitting**
- **catastrophic forgetting**

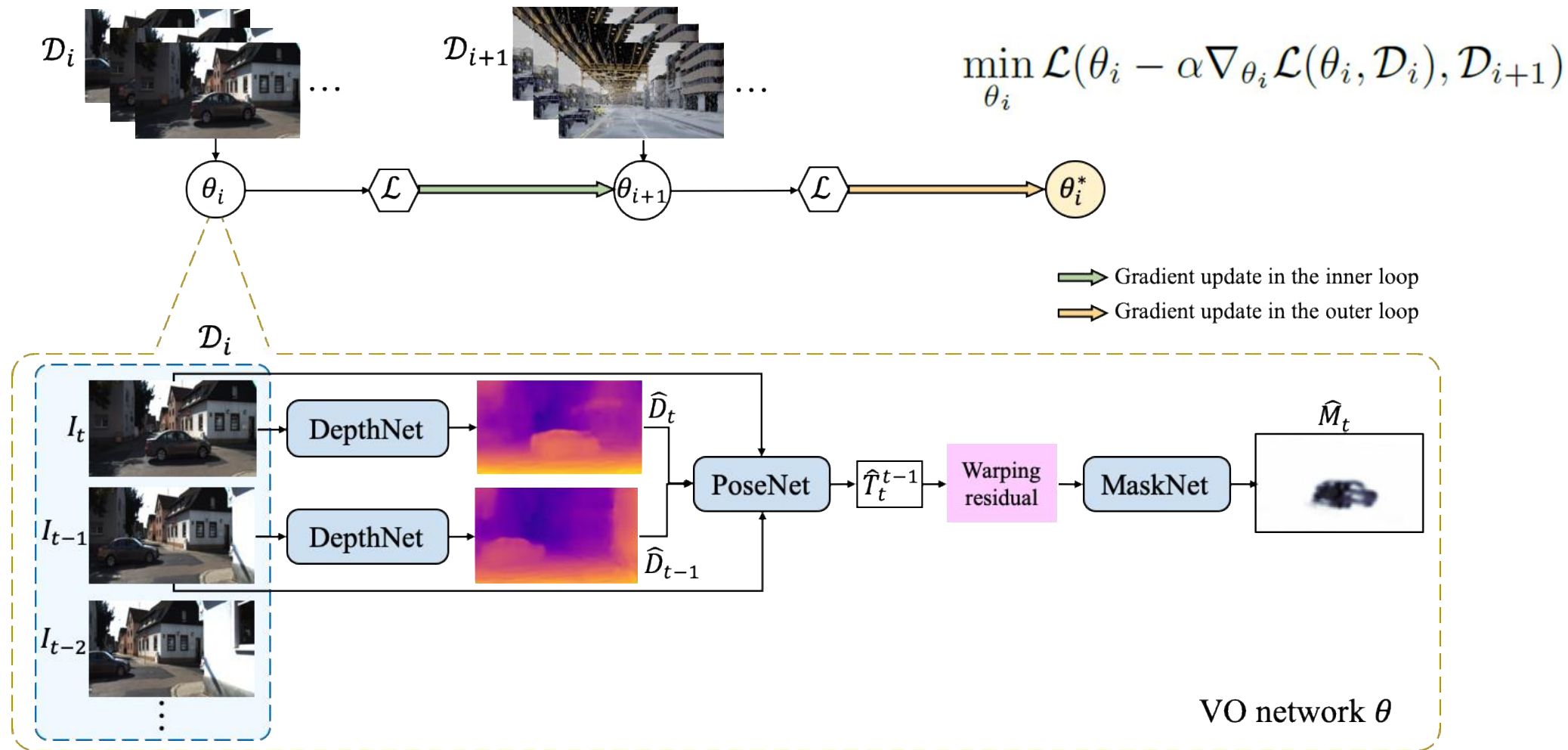


# Contributions

---

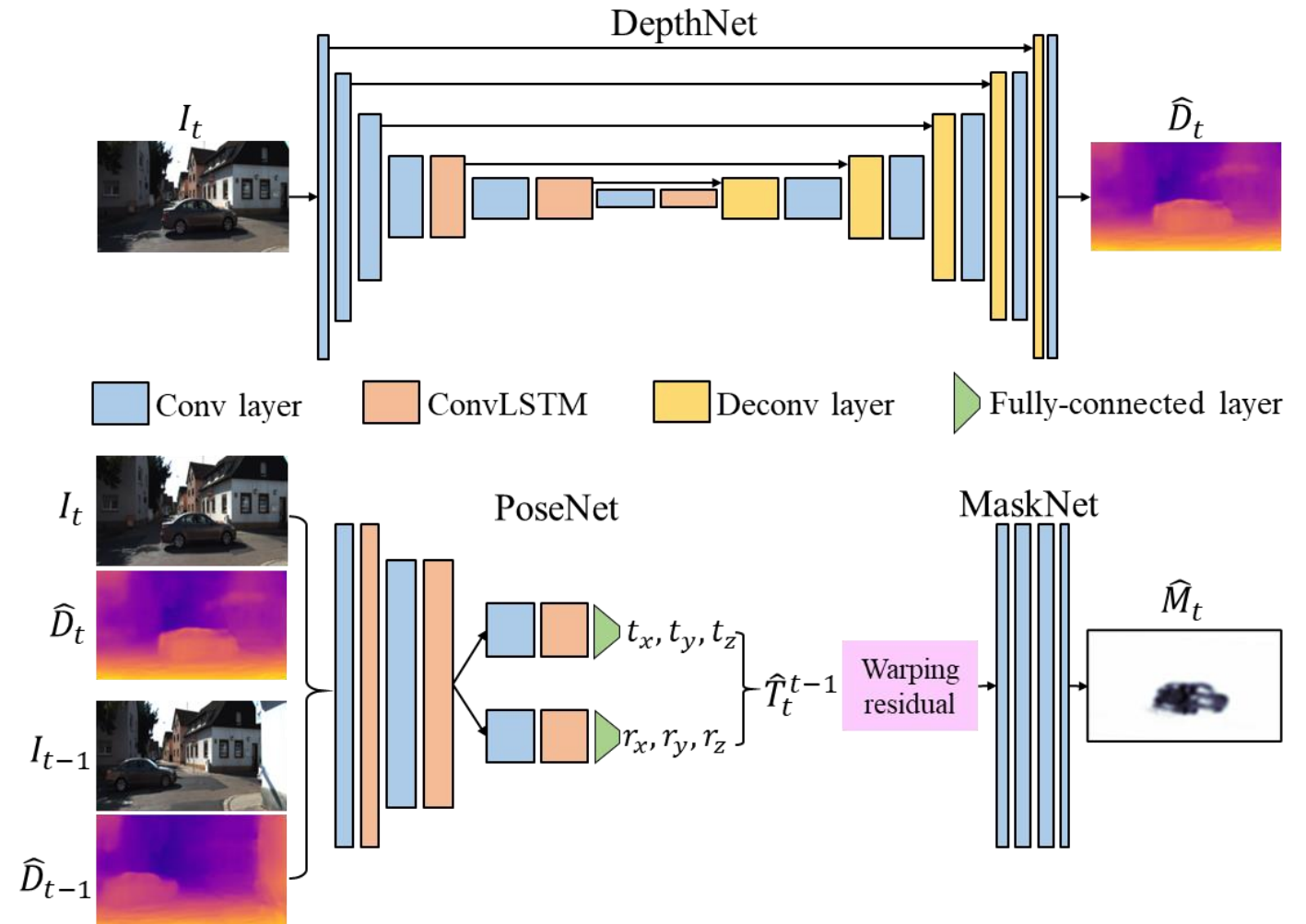
- **An online meta-learning algorithm to continuously adapt to unseen environments in a self-supervised manner**
- **Utilize past experience by convLSTM to achieve better estimation and adapt quickly to the current frame**
- **Online feature alignment to maintain long-term structures in the open world to avoid over-fitting and forgetting**

# Overview



# Spatial-Temporal Aggregation

- Utilize past experience by convLSTM for better online prediction
- Achieve better estimation and adapt quickly to the current frame



# Online Feature Alignment

- Structure maintenance in long-term sequences: deal with delicate changing in the open world
- Similar to: instance normalization, style transfer in GANs



Night

Online feature alignment

Pre-trained  
day prior

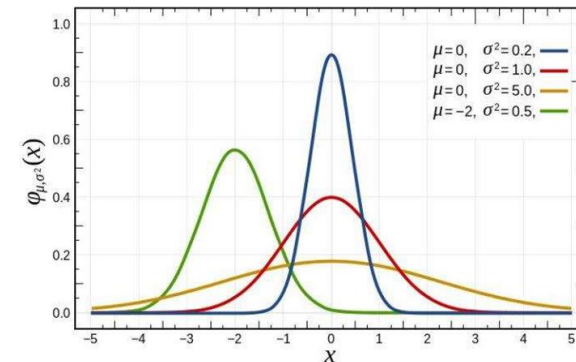


Day

$$\mu_s = \frac{1}{n} \sum_{j=1}^n f_j, \quad \sigma_s^2 = \frac{1}{n} \sum_{j=1}^n (f_j - \mu_s)^2$$

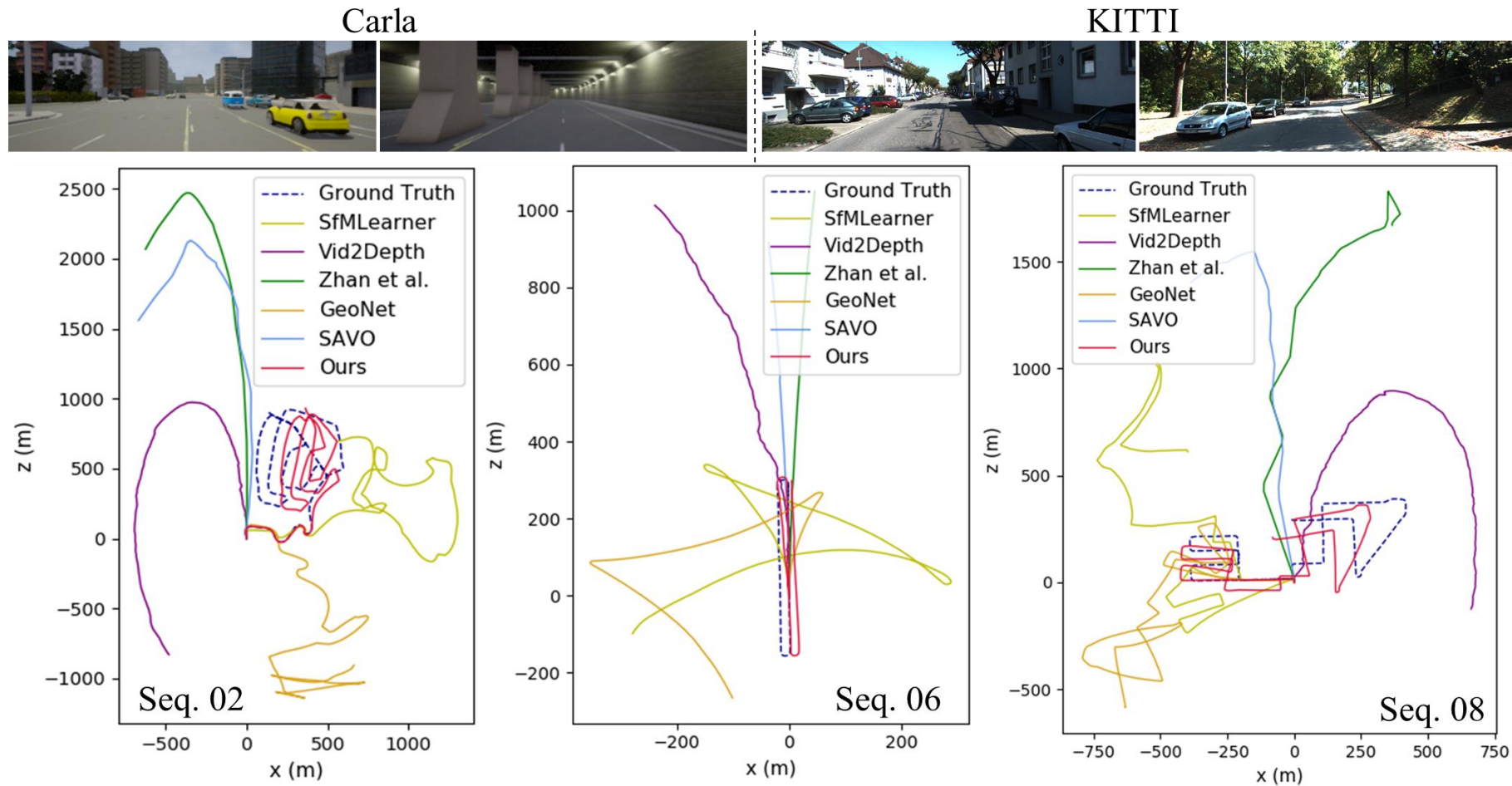
$$\mathcal{F}_s = (\mu_s, \sigma_s^2)$$

$$\mu_i = (1 - \beta)\mu_{i-1} + \beta\hat{\mu}_i \quad \sigma_i^2 = (1 - \beta)\sigma_{i-1}^2 + \beta\hat{\sigma}_i^2$$



# Experiments: Virtual to Real

Pre-train on synthetic Carla simulator, test on real-world KITTI dataset



# Experiments: Outdoor to Indoor

Pre-train on  
KITTI

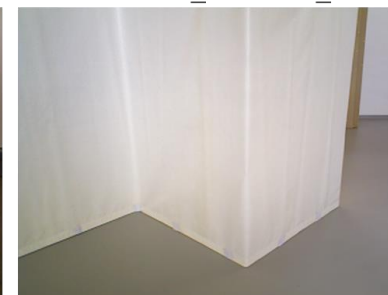
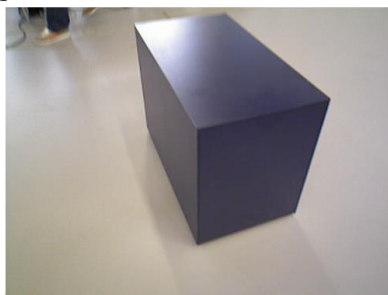


fr3/nostructure\_texture\_near\_withloop

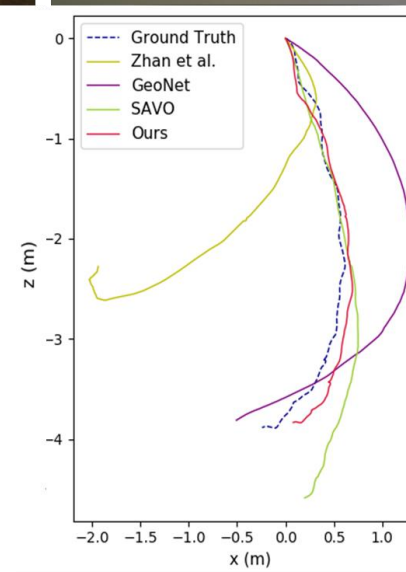
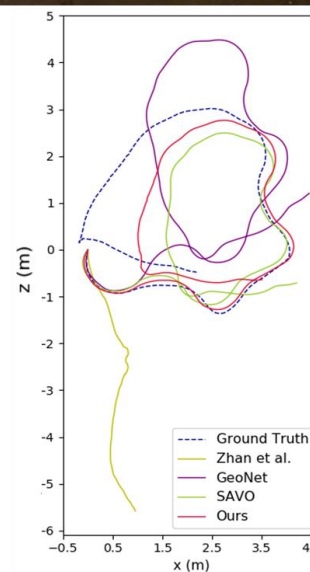
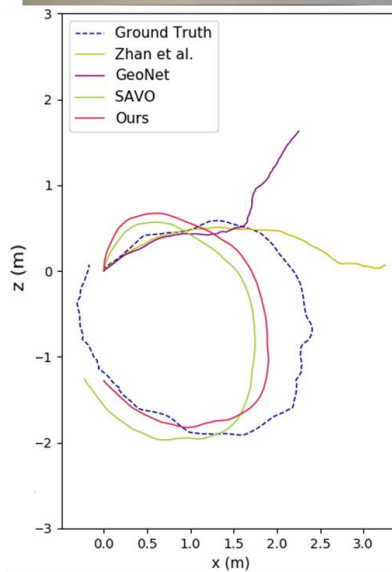
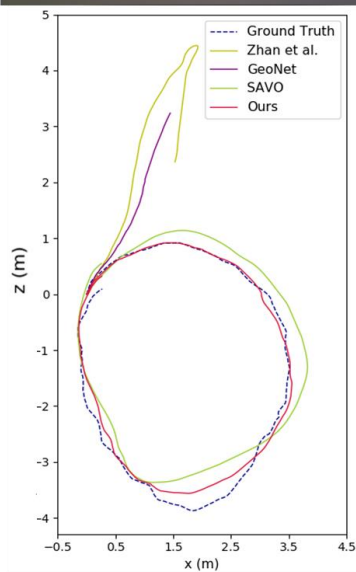
fr3/cabinet

fr2/pioneer\_360

fr3/structure\_notexture\_near



Test on TUM



# Take Home Message

## Contributions:

- Online meta-learning for VO to achieve self-supervised online adaptation
- Utilize past experience by convLSTM to achieve better estimation and quick adaptation
- Online feature alignment to maintain long-term structures in the open world to avoid over-fitting and forgetting

## Future work:

- Combination of geometry computation and learning for faster online adaptation
- Map-centric VO with online adaptation

# Conclusions

- ◆ Dynamic vision is a central topic in computer 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!