A New Approach to SLAM: Flow-based Learning Paradigm

Hongbin Zha
Key Laboratory of Machine Perception (MOE)
Peking University
zha@cis.pku.edu.cn

Outline

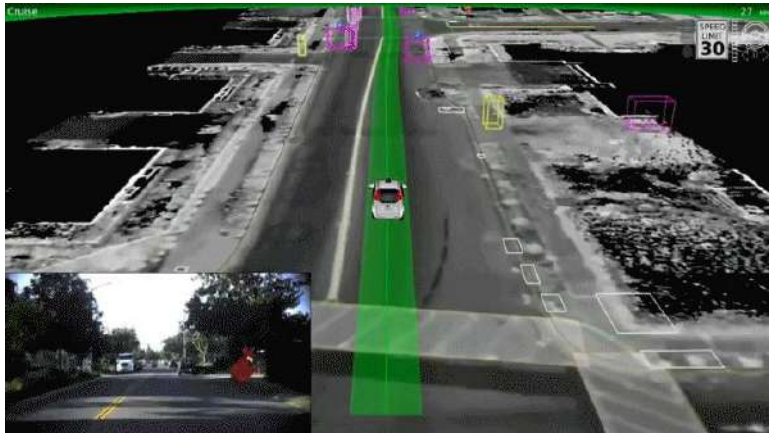
- **Introduction to Flow-Based Learning for SLAM**
- Related Research Topics:
  - Deep visual odometry using RNN with long-term dependency
  - Line Flow based SLAM
- Conclusions

Localization and Mapping

◆ Localization

Odometry of sensors or robot systems

Amazon warehouse robot



Google autonomous driving car

◆ Mapping

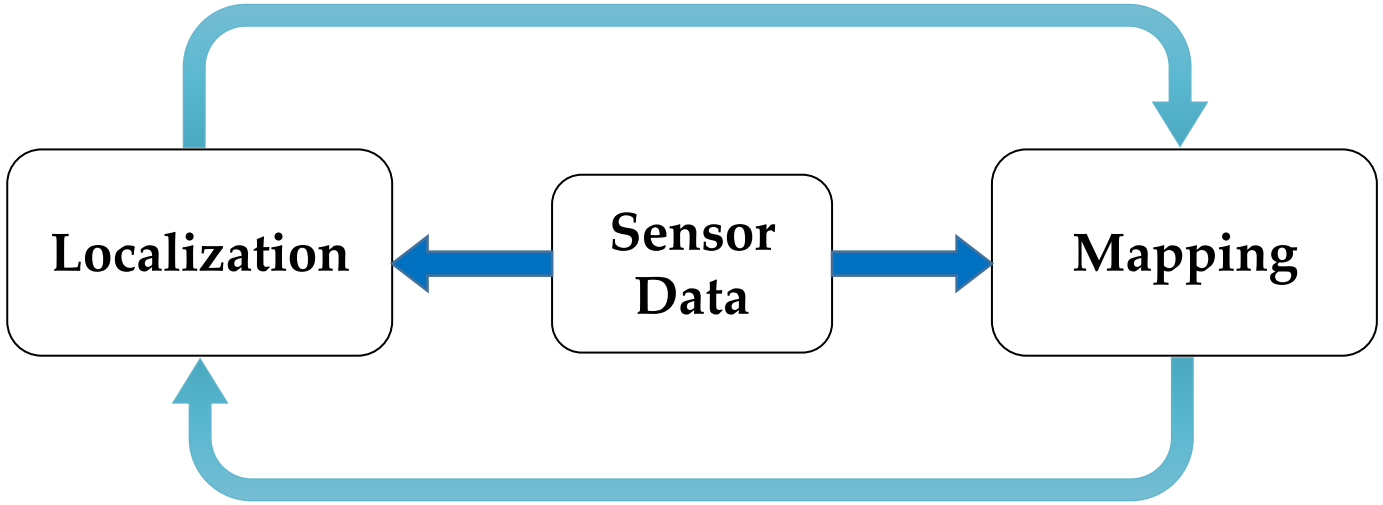
3D reconstruction of environments

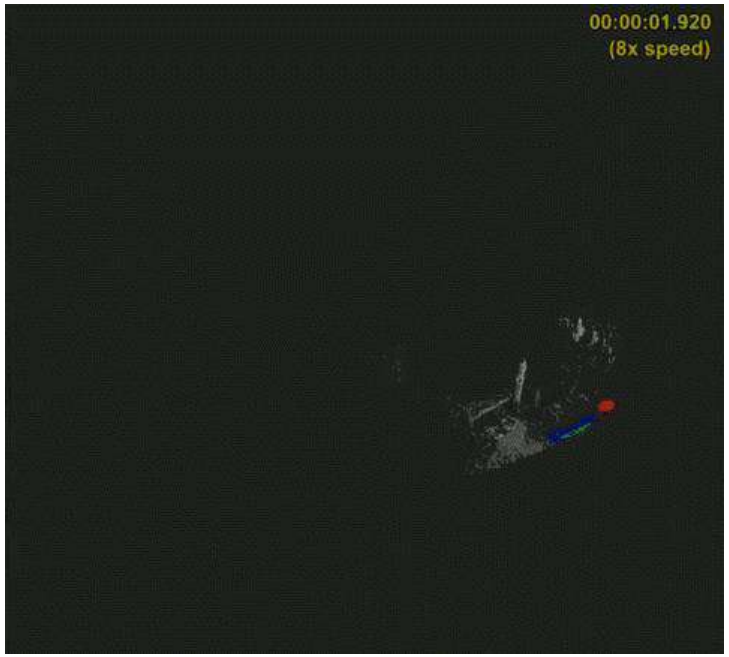
Google earth

Microsoft Holoportation

SLAM: Simultaneous Localization And Mapping

- Tight coupling of **localization** and **mapping**



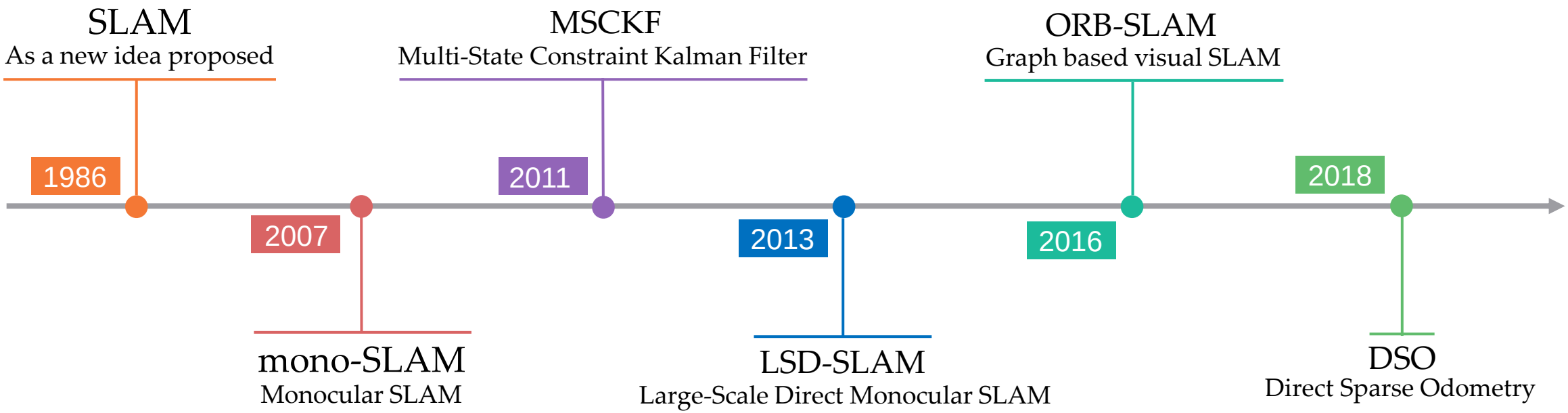


LSD-SLAM

A fundamental function of dynamic vision systems as for human

History of SLAM Research

ORB: Oriented FAST and Rotated BRIEF

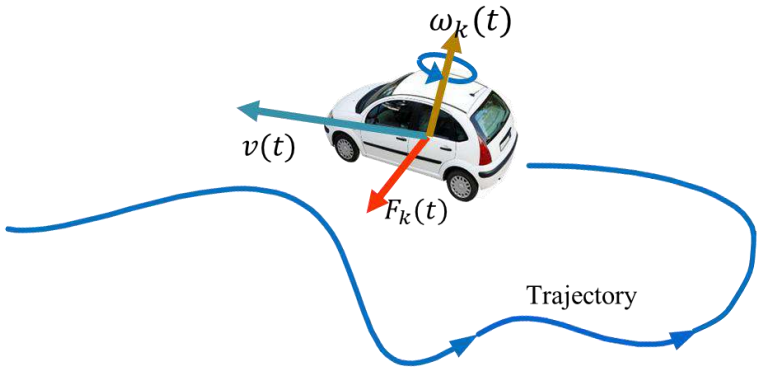


- Make good use of explicit geometrical relationships between consecutive frames: multi-view geometry
- Enhance the performance by fusing different kinds of sensors
- Work well in limited environments for specific tasks

Is Current SLAM Good Enough?

◆ Pay Little Attention to Temporal Continuity

- Disregarding spatial-temporal consistency inherent in SLAM
- **A big source of accumulated error**
- **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

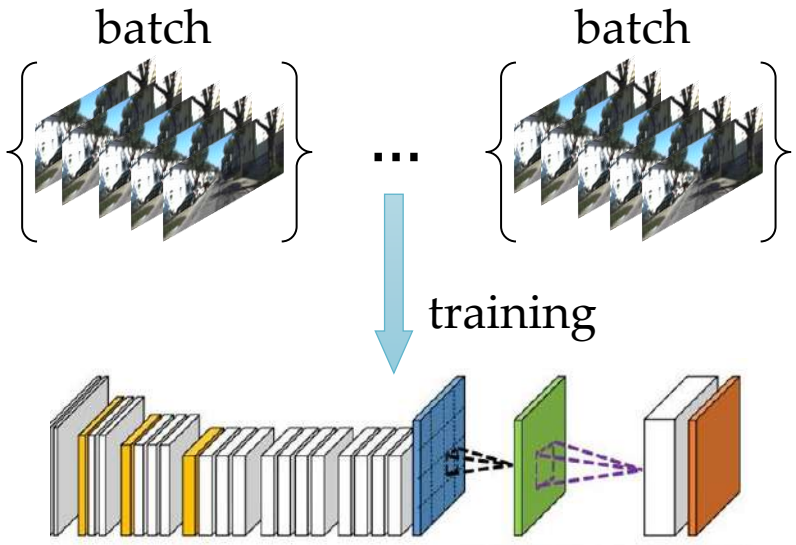
Is Current SLAM Good Enough?

◆ High Computational Cost

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

◆ Learning Approach and SLAM

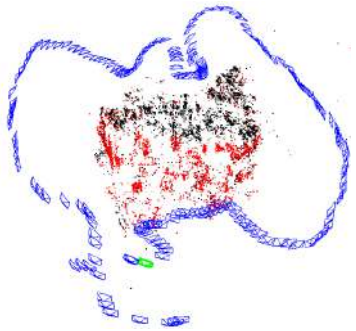
- Off-line batch training instead of online learning
- Supervised learning requiring massive labeled data
- Tedious parameter tuning
- Poor generalization ability



Critical Problem

◆ Lack of Systematic Formulation

- **Require delicate hand-crafted design and ad hoc strategies**
- **Various optimizations are proposed 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

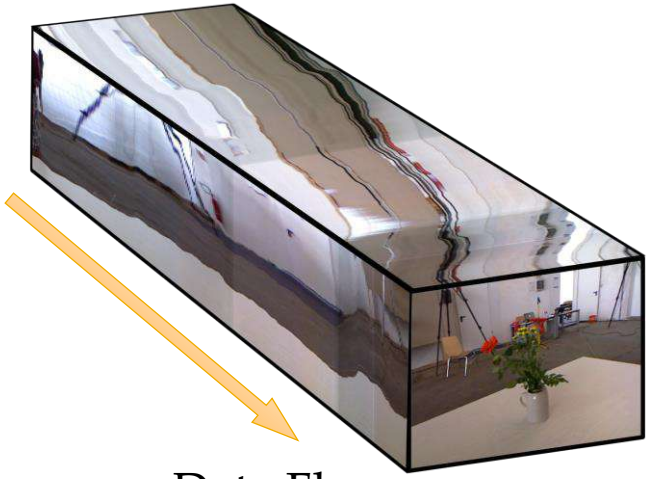
DSO

- ➢ Requires photometric calibration
- ➢ Not robust to illumination changes

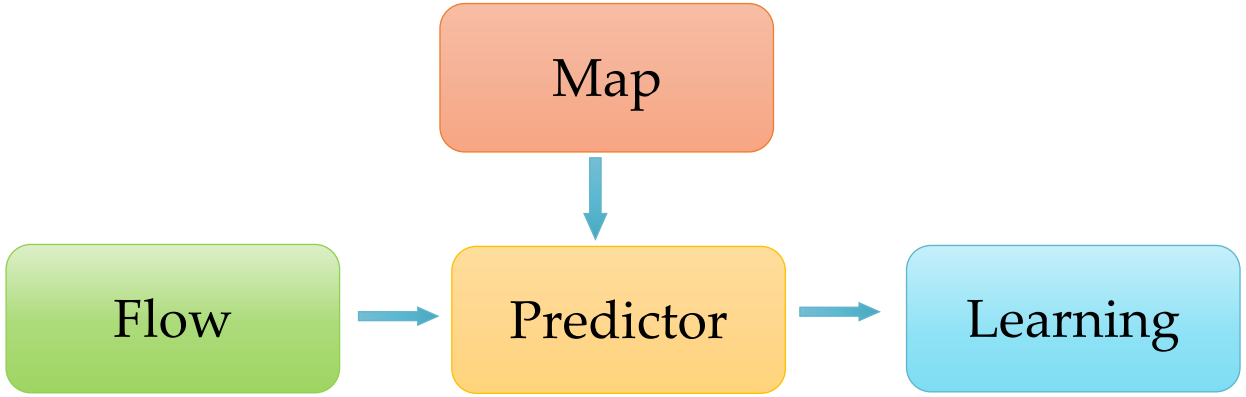
What is Flow?

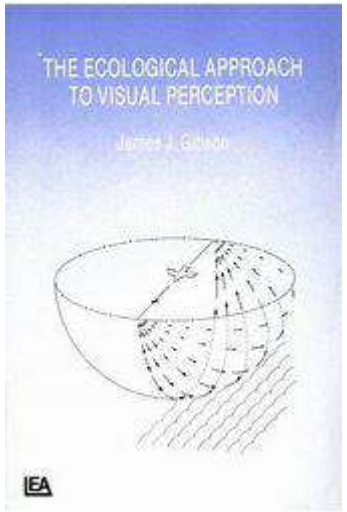
Sensor Data Flow:

- 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



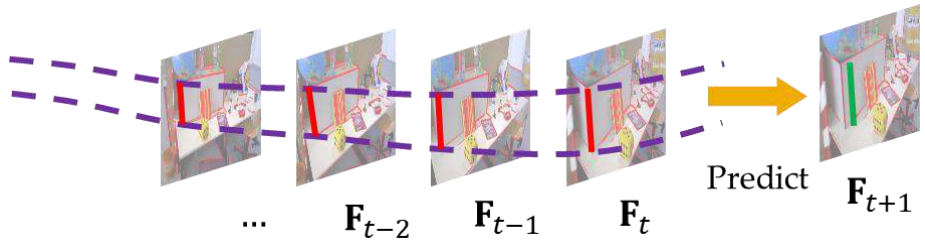


James Gibson: The Ecological Approach to Visual Perception, 1986

Predictor and Map

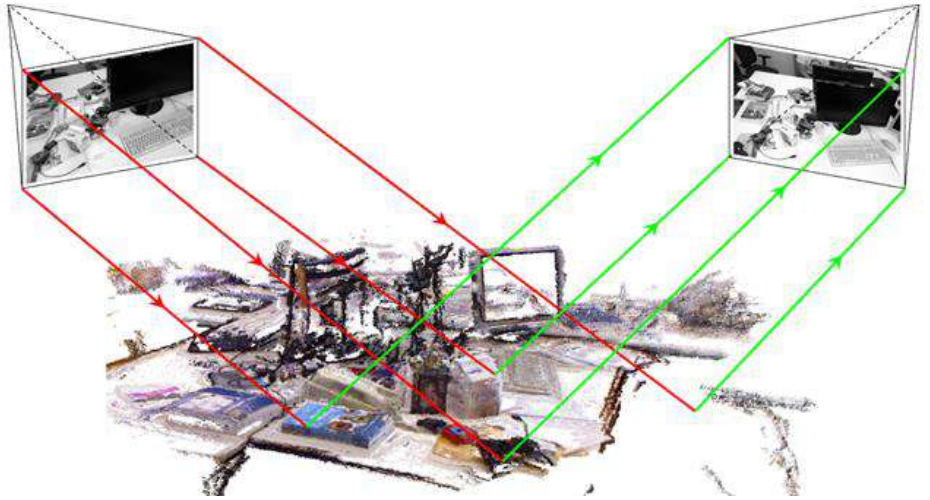
◆ Predictor: the engine of efficient SLAM

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



◆ Map: a global, invariant representation

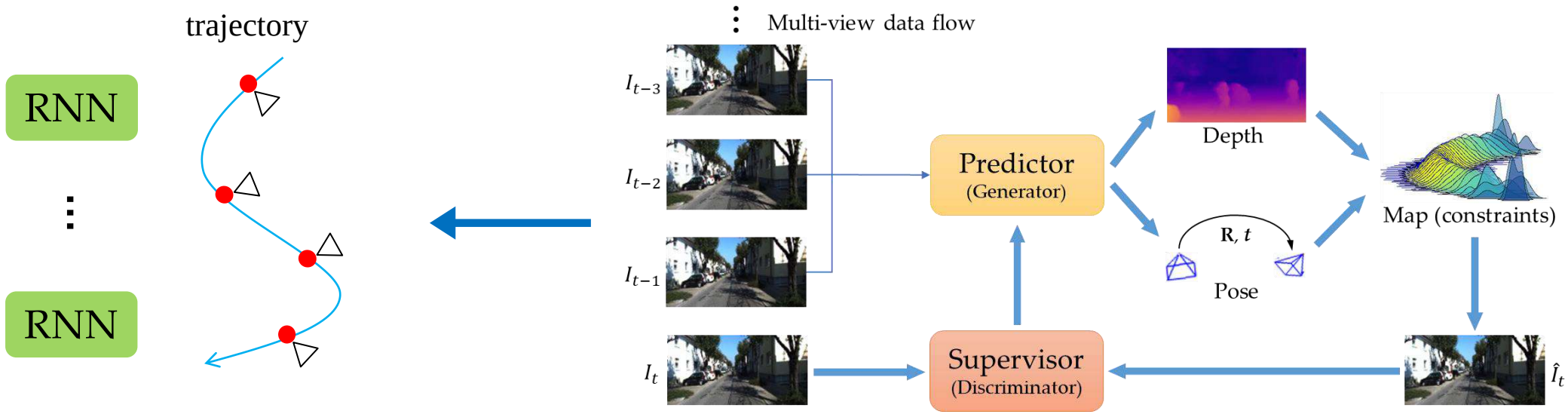
- Implicit/explicit representation of physical world
- Provide global constraints as regularization
- Supervisory information for prediction



Learning: A Systematic Solution

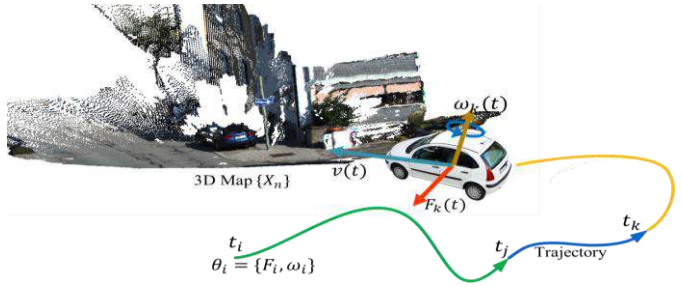
- Learning camera pose in a supervised manner
- Modelling the data flow using RNN

- Self-supervised, online learning
- Generative Adversarial Networks (GAN)

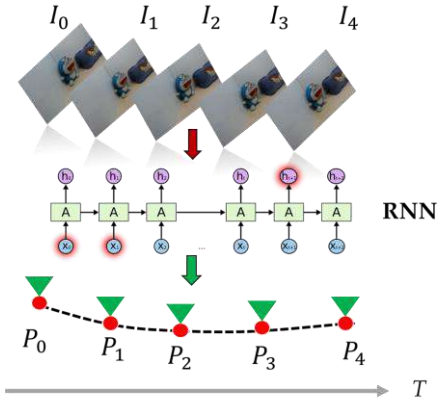


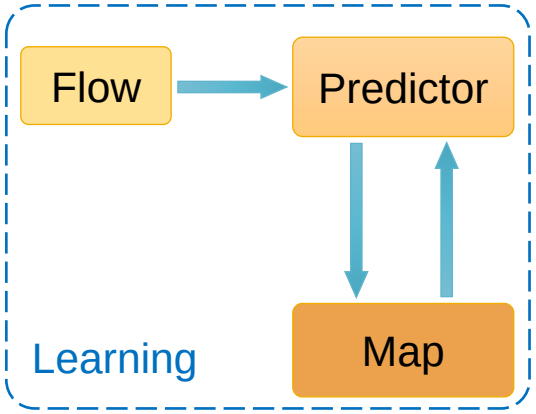
Our Related Research

◆ Dynamics Model



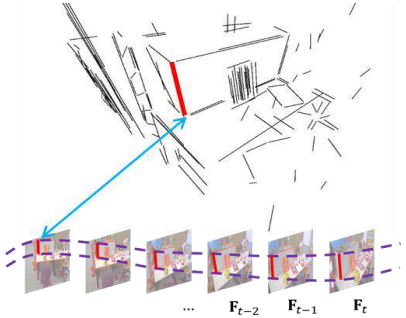
◆ RNN Learning



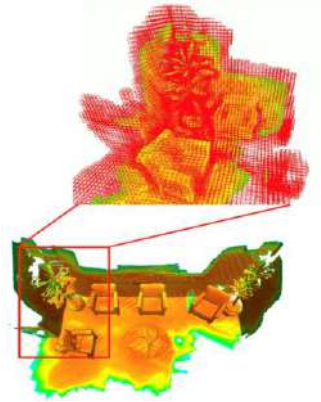


Goal: **Unsupervised online learning** for SLAM

◆ Line Flow



◆ Probabilistic Map Representation



ICCV'19 （2篇）, CVPR'19, ECCV'18
ACCV'18（2篇）, ICRA'18, ICPR'18 (Track Best Paper), ICRA'17, IROS'17, BMVC'16

Outline

- Introduction to Flow-Based Learning for SLAM
- **Related Research Topics:**
  - **Deep visual odometry using RNN with long-term dependency**
  - Line Flow based SLAM
- Conclusions

Beyond Tracking: Selecting Memory and Refining Poses for Deep Visual Odometry (VO)

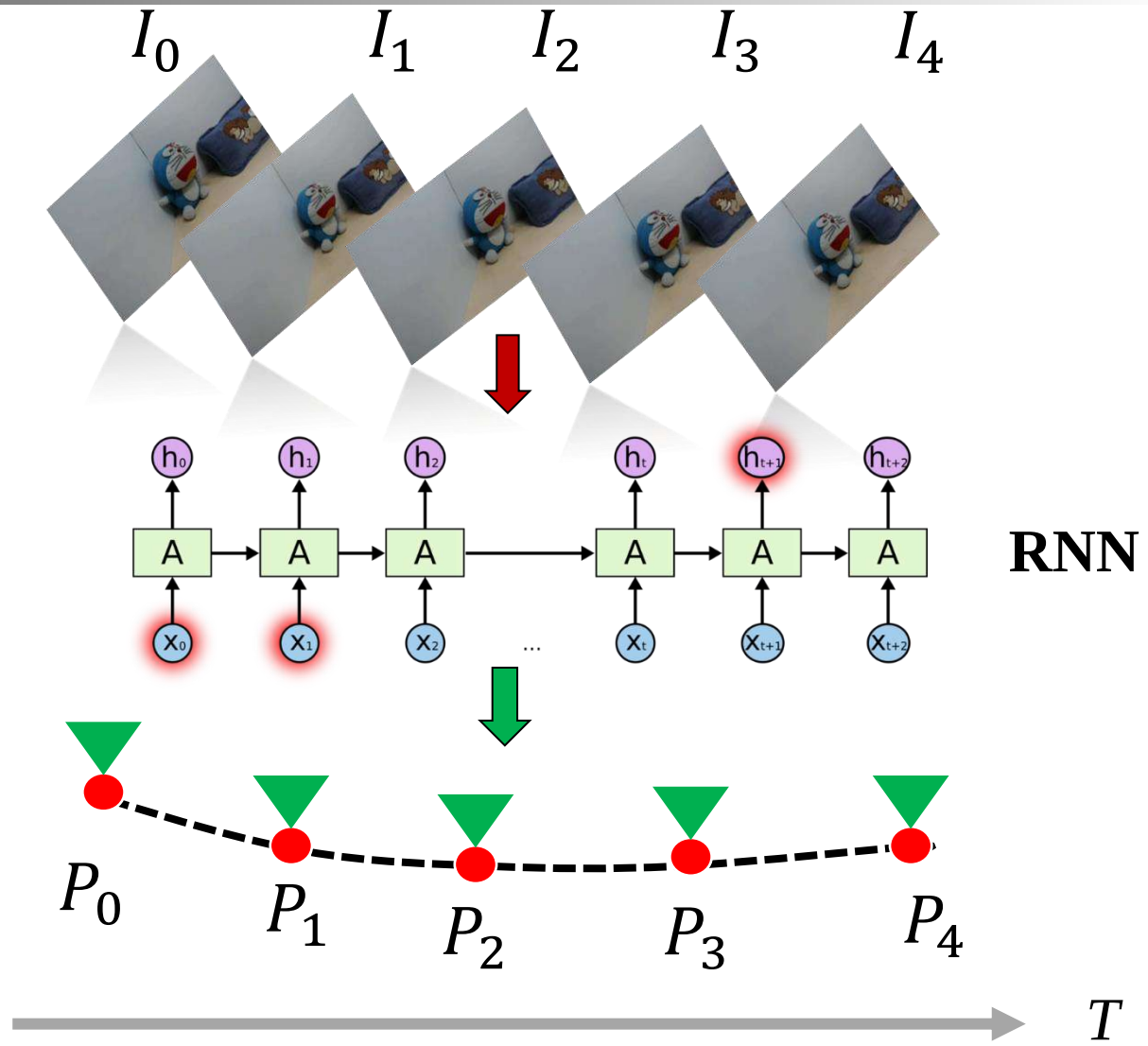
- Fei Xue, Xin Wang, Shunkai Li, Qiuyuan Wang, Junqiu Wang, Hongbin Zha

- CVPR 2019 (oral presentation)

Problem Definition

- **Input**
  - **Image flow** $|I_t - I_{t-1}| < \sigma_{small}$
- **Output**
  - **Camera pose** $|P_t - P_{t-1}| < \alpha_{small}$

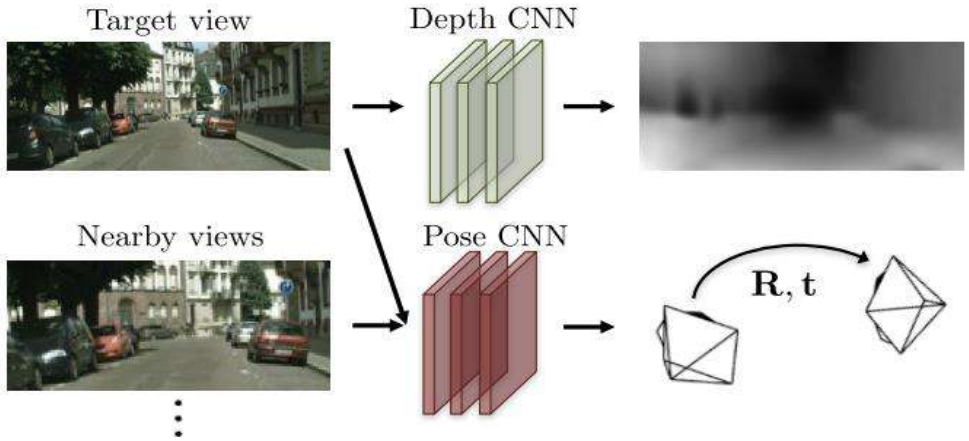
- **Learning camera pose incrementally**
- **Sequence modeling using RNN**
- **<u>Long-term dependency</u>**



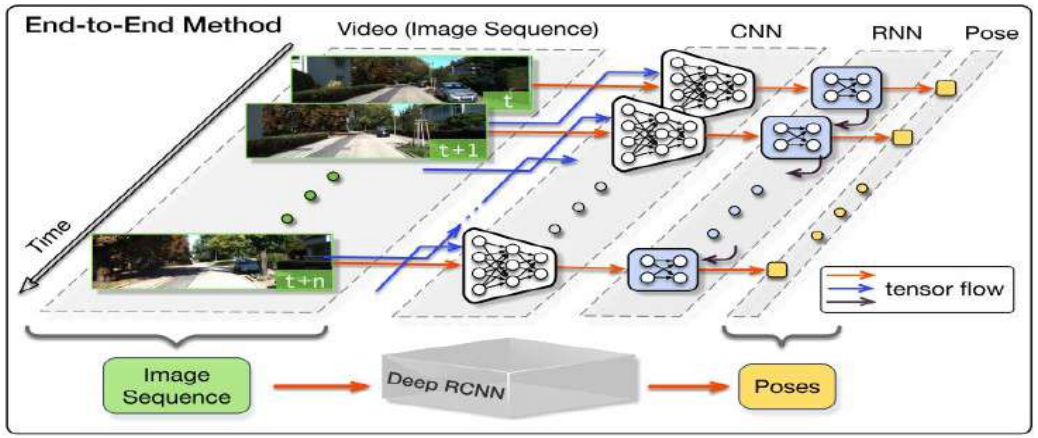
Learning-based VO Methods: Previous Work

- **Mimicking structure from motion**
  - **Processing image snippets**
  - **Learning ego-motion jointly with depth**
- **Modeling the sequence with LSTM**
  - **Estimating relative camera poses**
  - **Relying on local historical knowledge**

- **Treat VO as a pure tracking problem**



SfmLearner (Zhou, et al., CVPR, 2017)



DeepVO (Wang, et al., ICRA, 2017)

Challenges

- Severe error accumulation

$$P_t = \Pi_{i=1}^{t} P_{i,i-1} * P_0$$

- LSTM cannot remember long-term dependencies
  - Historical knowledge is forgotten

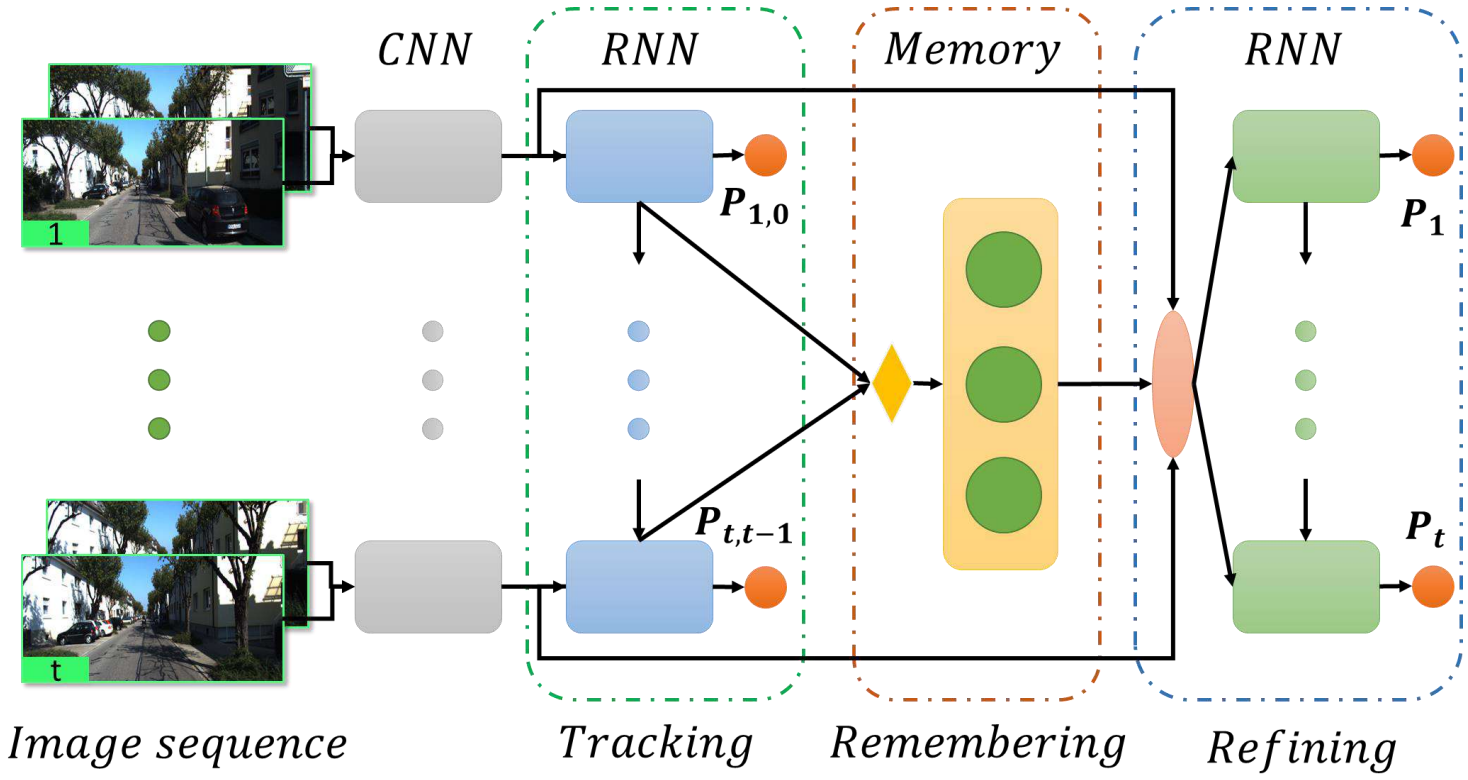
- The contributions of coming observations are ignored
  - New data is supposed to refine previous poses

Our Contributions

- A novel VO framework consisting of Tracking, Remembering and Refining
  - Remembering global information
  - Refining previous results with new observations
- An adaptive and efficient strategy for memory selection
  - Motion-sensitive selection
- A spatial temporal attention mechanism for feature distilling
  - Co-visibility based correlation

System Overview

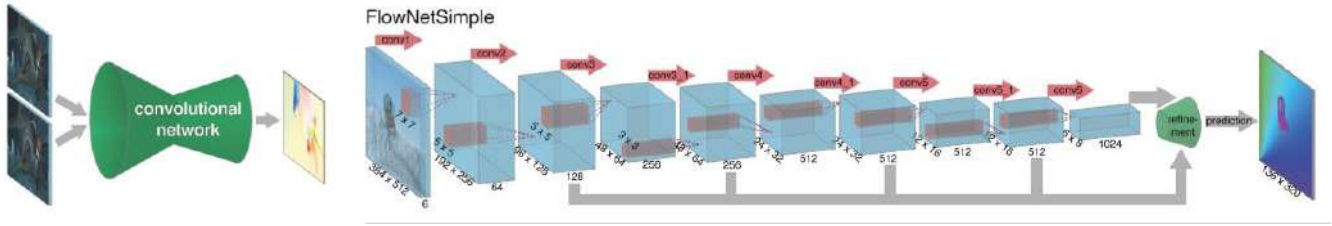
- Encoder
- Tracking
- Remembering
- Refining



Architecture

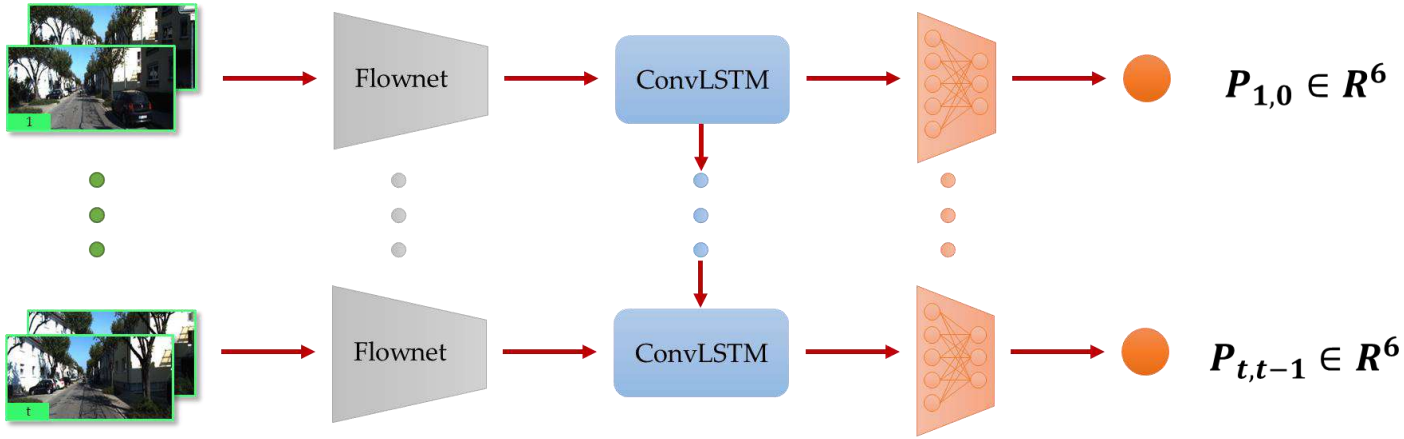
Encoder and Tracking

- Encoder
  - Optical flow estimation
  - Flownet (ICCV, 2015)



Flownet

- Tracking
  - ConvLSTM (NIPS, 2014)
    - Spatial connections preserved
  - Relative pose estimation



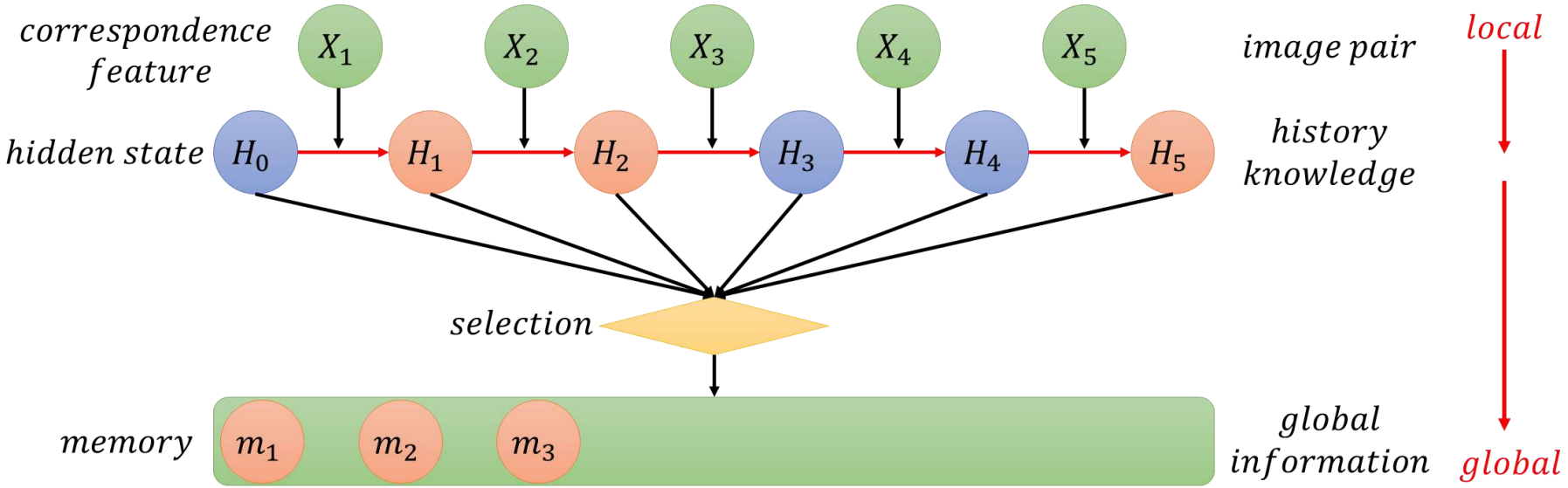
Tracking module

Remembering

- **Taking hidden state as "local map"**
- **Hierarchical map representation**
- **Motion based selection**

$$\left\| Rot_{m_i} - Rot_{m_{i-1}} \right\|_2 \leq \theta_{Rot}$$

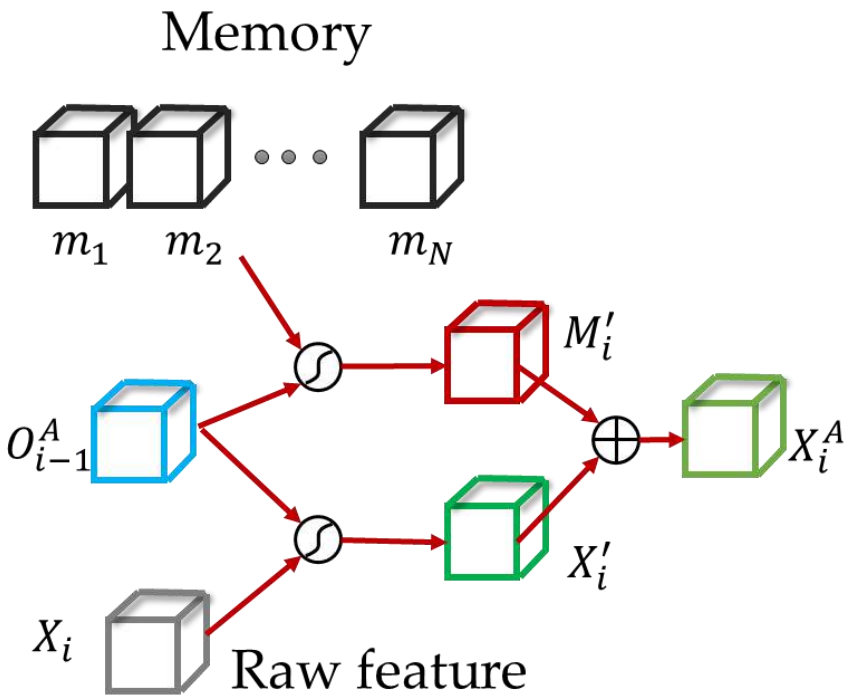
$$\left\| Tans_{m_i} - Tans_{m_{i-1}} \right\|_2 \leq \theta_{Trans}$$

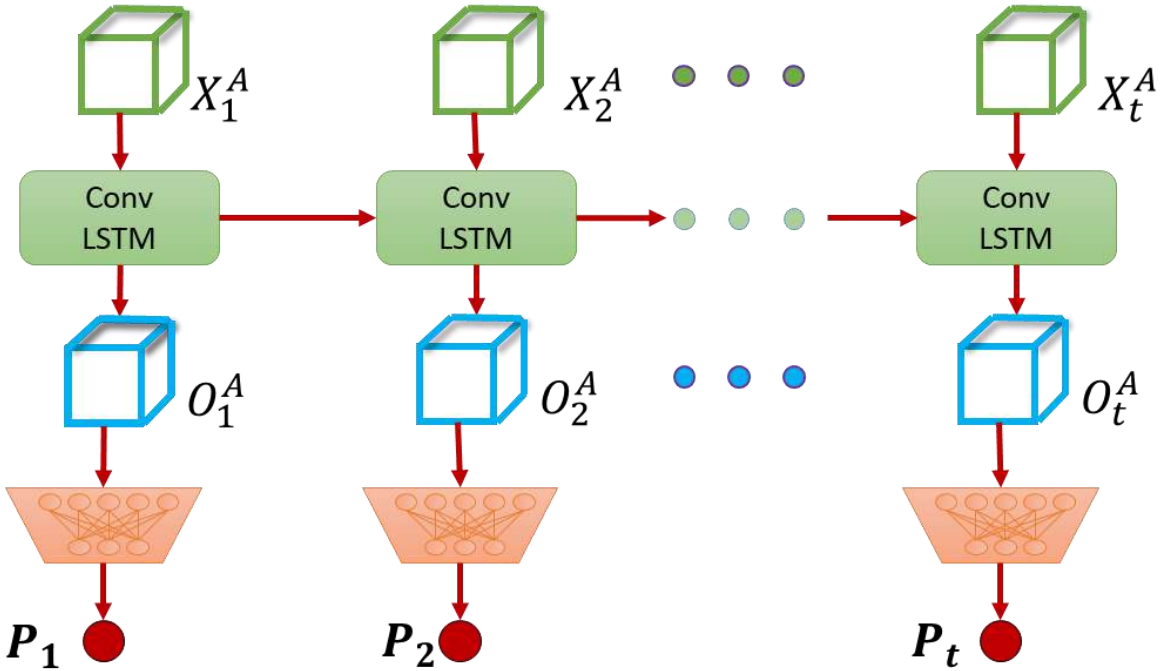


Remembering module

Refining

- **Attention-based feature selection**
- **Propagate refined results using the ConvLSTM** $O_t^A, H_t^A = U^A(X_t^A, H_{t-1}^A)$

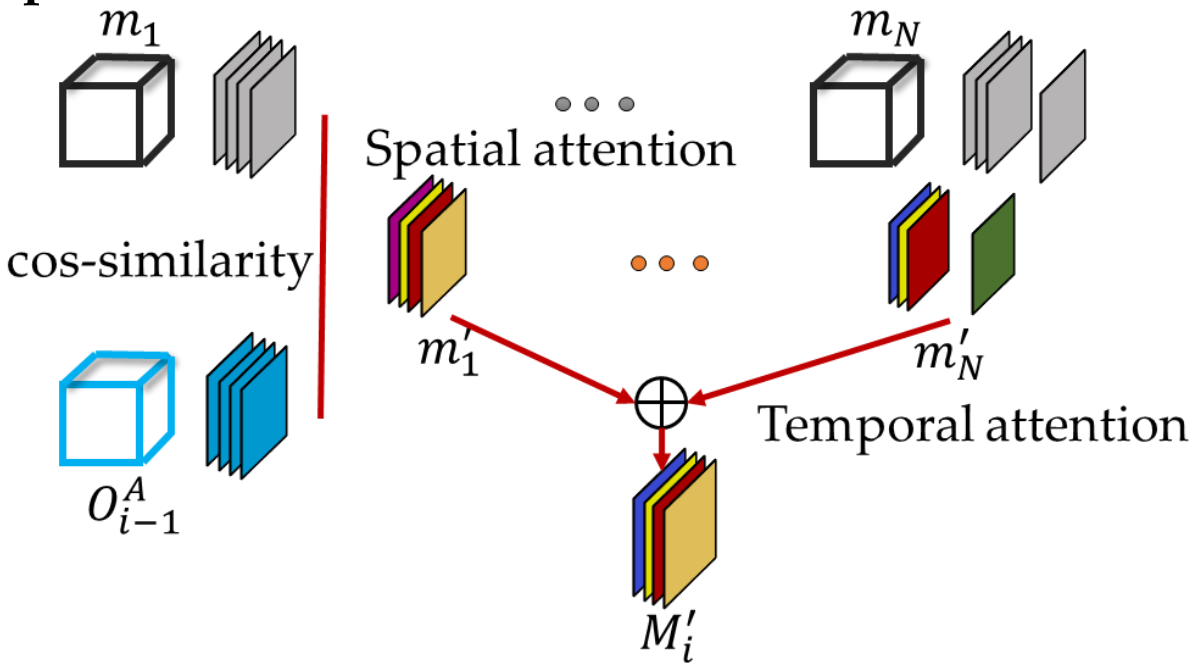


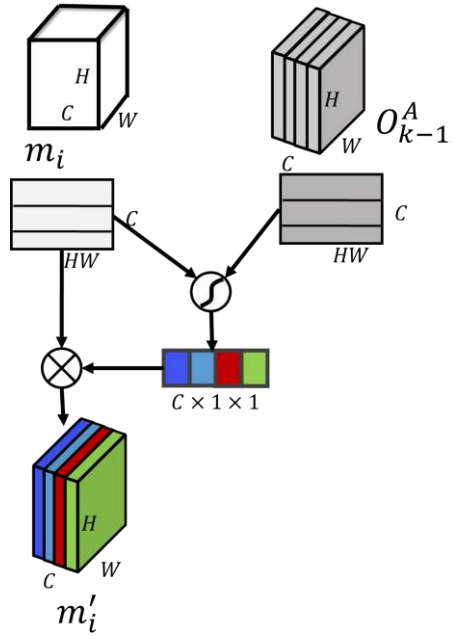


The last output as guidance

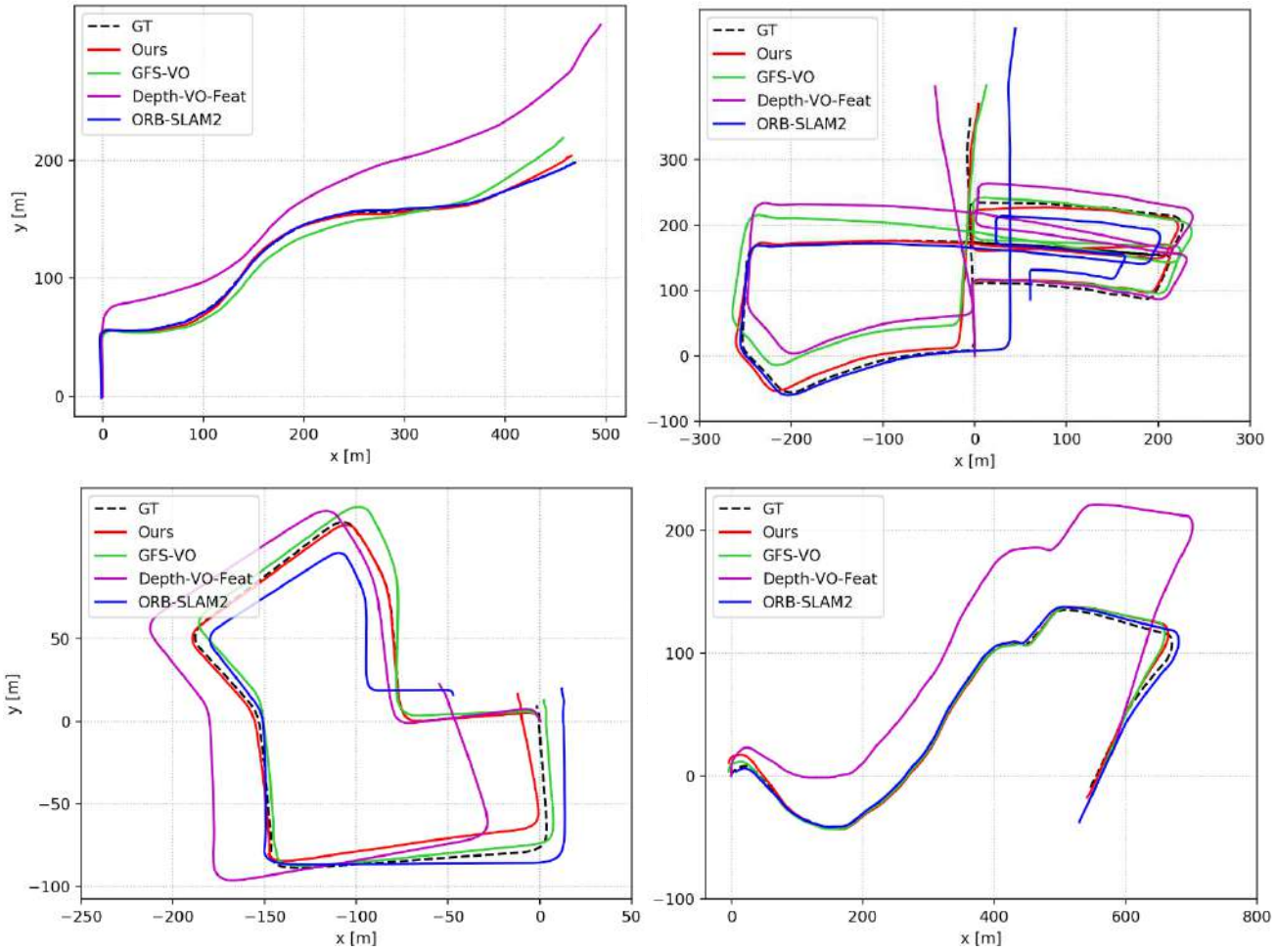
Spatial-Temporal Attention

- **Different hidden states contribute discriminatively**
  - **Temporal attention**
- **Different visual cues contribute discriminatively**
  - **Spatial attention**





Experiments-KITTI Dataset



Seq 03, 05, 07, 10

Comparison with learning-based methods

Method	Sequence 03		04		05		06		07		10		Avg	
	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
UnDeepVO [26]	5.00	6.17	5.49	2.13	3.40	1.50	6.20	1.98	3.15	2.48	10.63	4.65	5.65	3.15
Depth-VO-Feat [24]	15.58	10.69	2.92	2.06	4.94	2.35	5.80	2.07	6.48	3.60	12.45	3.46	7.98	4.04
GeoNet [25]	19.21	9.78	9.09	7.54	20.12	7.67	9.28	4.34	8.27	5.93	20.73	9.04	13.12	7.38
Vid2Depth [27]	27.02	10.39	18.92	1.19	51.13	21.86	58.07	26.83	51.22	36.64	21.54	12.54	37.98	18.24
SfmLearner [23]	10.78	3.92	4.49	5.24	18.67	4.10	25.88	4.80	21.33	6.65	14.33	3.30	15.91	4.67
NeuralBundler [47]	4.51	2.82	**2.30**	**0.87**	3.91	1.64	**4.60**	2.85	3.56	2.39	12.90	3.17	5.30	2.29
Ours	**3.32**	**2.10**	2.96	1.76	**2.59**	**1.25**	4.93	**1.90**	**3.07**	**1.76**	**3.94**	**1.72**	**3.47**	**1.75**
DeepVO [28]	8.49	6.89	7.19	6.97	2.62	3.61	5.42	5.82	3.91	4.60	8.11	8.83	5.96	6.12
ESP-VO [29]	6.72	6.46	6.33	6.08	3.35	4.93	7.24	7.29	3.52	5.02	9.77	10.2	6.15	6.66
CL-VO [53]	8.12	3.47	7.57	2.61	5.77	2.00	7.66	**1.66**	6.79	3.00	8.29	2.94	7.37	2.67
GFS-VO-RNN [30]	6.36	3.62	5.95	2.36	5.85	2.55	14.58	4.98	5.88	2.64	7.44	3.19	7.68	3.22
GFS-VO [30]	5.44	3.32	2.91	1.30	3.27	1.62	8.50	2.74	3.37	2.25	6.32	2.33	4.97	2.26
Ours	**3.32**	**2.10**	2.96	1.76	**2.59**	**1.25**	**4.93**	1.90	**3.07**	**1.76**	**3.94**	**1.72**	**3.47**	**1.75**

t_{rel} : average translational RMSE drift (%) on length from 100, 200 to 800 m.
r_{rel} : average rotational RMSE drift (°/100m) on length from 100, 200 to 800 m.

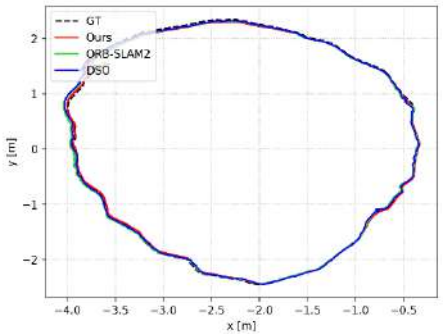
Comparison with classic methods

Seq	Ours		VISO2-M [2]		ORB-SLAM2 [1]		ORB-SLAM2 (LC) [1]	
	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
03	3.32	2.10	8.47	8.82	2.28	0.40	2.17	0.39
04	2.96	1.76	4.69	4.49	1.41	0.14	1.07	0.17
05	2.59	1.25	19.22	17.58	13.21	0.22	1.86	0.24
06	4.93	1.90	7.30	6.14	18.68	0.26	4.96	0.18
07	3.07	1.76	23.61	19.11	10.96	0.37	1.87	0.39
10	3.94	1.72	41.56	32.99	3.71	0.30	3.76	0.29
Avg	3.47	1.75	17.48	16.52	8.38	0.28	2.62	0.28

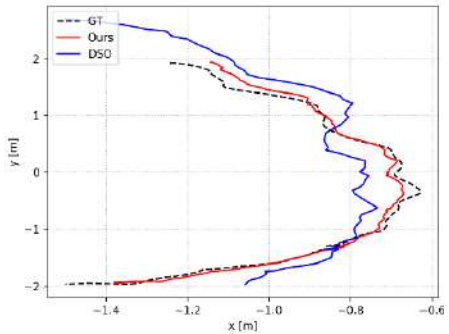
Experiments-TUM-RGBD Dataset

- **Baselines**
  - **ORB-SLAM2, DSO**
- **Various conditions**

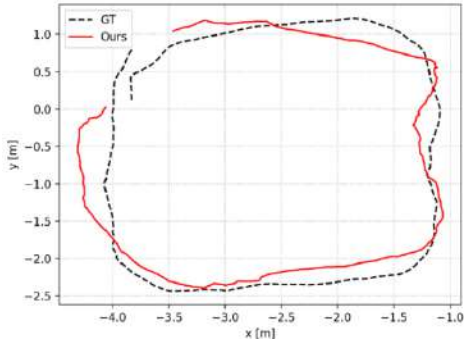
Sequence	Desc. str/tex/a.m.	Frames	ORB-SLAM2 [1]	DSO [5]	Ours (tracking)	Ours (w/o temp atten)	Ours (w/o spat atten)	Ours
fr2/desk	Y/Y/N	2965	**0.041**	X	0.183	0.164	0.159	0.153
fr2/360_kidnap	Y/Y/N	1431	**0.184**	0.197	0.313	0.225	0.224	0.208
fr2/pioneer_360	Y/Y/Y	1225	X	X	0.241	0.1338	0.076	**0.056**
fr2/pioneer_slam3	Y/Y/Y	2544	X	0.737	0.149	0.1065	0.085	**0.070**
fr2/large_cabinet	Y/N/N	1011	X	X	0.193	0.193	0.177	**0.172**
fr3/sitting_static	Y/Y/N	707	X	0.082	0.017	0.018	0.017	**0.015**
fr3/nstr_ntex_near_loop	N/N/N	1125	X	X	0.371	0.195	0.157	**0.123**
fr3/nstr_tex_near_loop	N/Y/N	1682	0.057	0.093	0.046	0.011	0.010	**0.007**
fr3/str_ntex_far	Y/N/N	814	X	0.543	0.069	0.047	0.039	**0.035**
fr3/str_tex_far	Y/Y/N	938	**0.018**	0.040	0.080	0.049	0.046	0.042



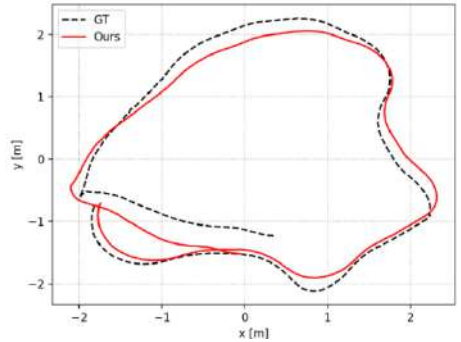
fr3/str_tex_far



fr3/str_ntex_far



fr3/nstr_ntex_near_loop



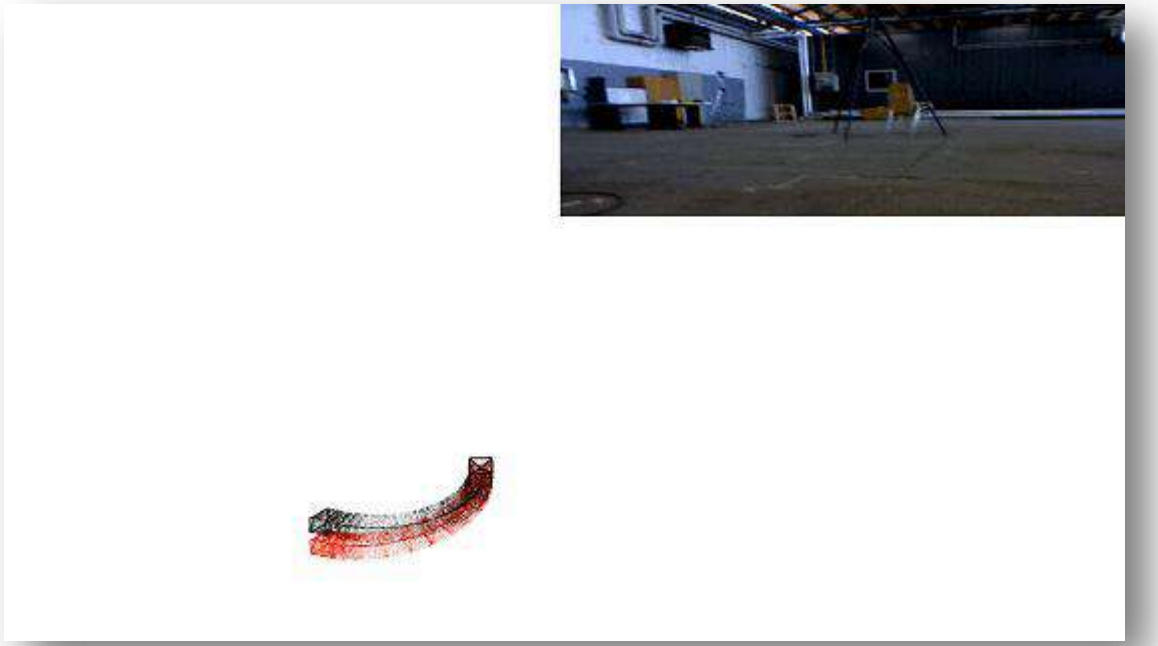
fr2/pioneer_360

Experiments

Ours

Ground-truth

KITTI Seq 07



fr2/pioneer_360

Take-home Message

- **A novel VO framework consisting of *Tracking*, *Remembering* and *Refining* components**
- **An adaptive and efficient strategy for memory selection**
- **A spatial temporal attention mechanism for feature distilling**

- **Future work**
  - **Focus on learning a unified map representation**
  - **Build a full SLAM system using neural networks**

Outline

- Introduction to Flow-Based Learning for SLAM
- **Related Research Topics:**
  - Deep visual odometry using RNN with long-term dependency
  - **Line Flow based SLAM**
- Conclusions

Line Flow based SLAM

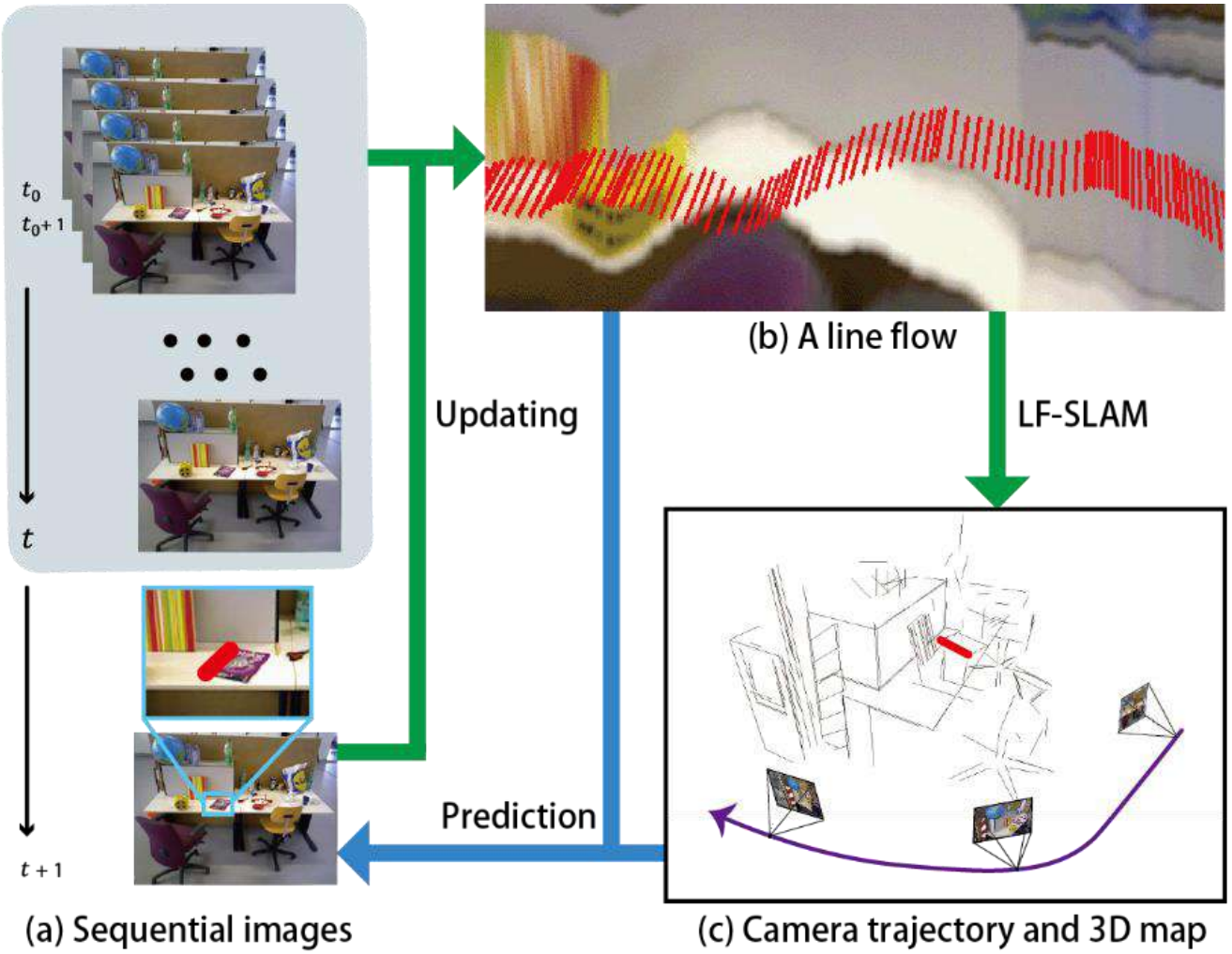
Qiuyuan Wang, Zike Yan, Fei Xue, Wei Ma,

Junqiu Wang, Xin Wang, Hongbin Zha

wangqiuyuan@pku.edu.cn

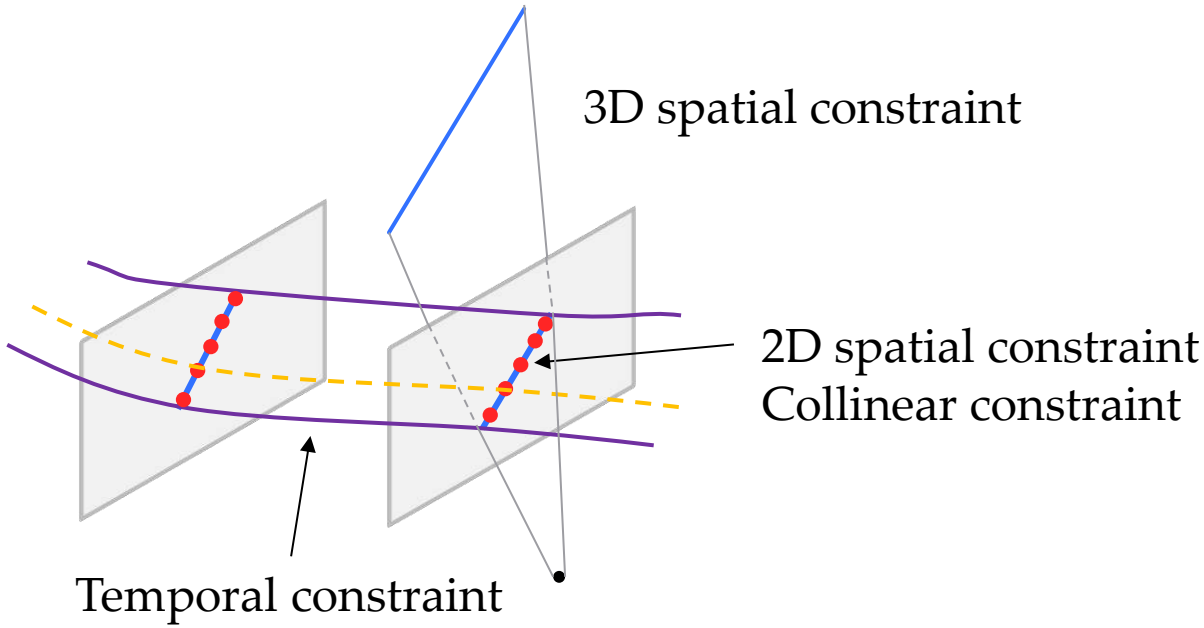
Problem Definition

- Input
  - Sequential images
- Output
  - Camera pose
  - 3D line map
- Motivation
  - Line segment
  - Spatial-temporal constraint
  - Localization
    - Texture-less scenes
    - Man-made building



Problem Definition

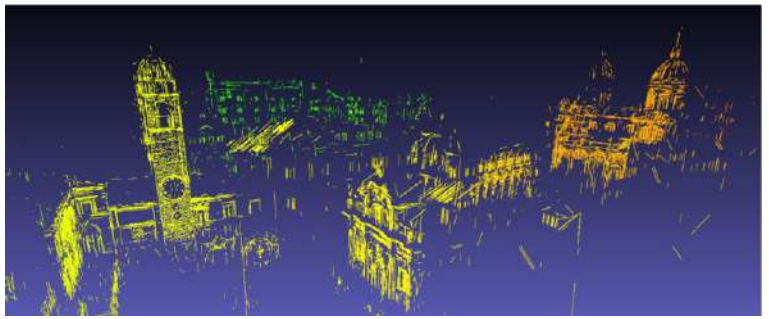
- **Input**
  - **Sequential images**
- **Output**
  - **Camera pose**
  - **3D line map**
- **Motivation**
  - **Line segment**
  - **Spatial-temporal constraint**
  - **Localization**
    - **Texture-less scenes**
    - **Man-made building**



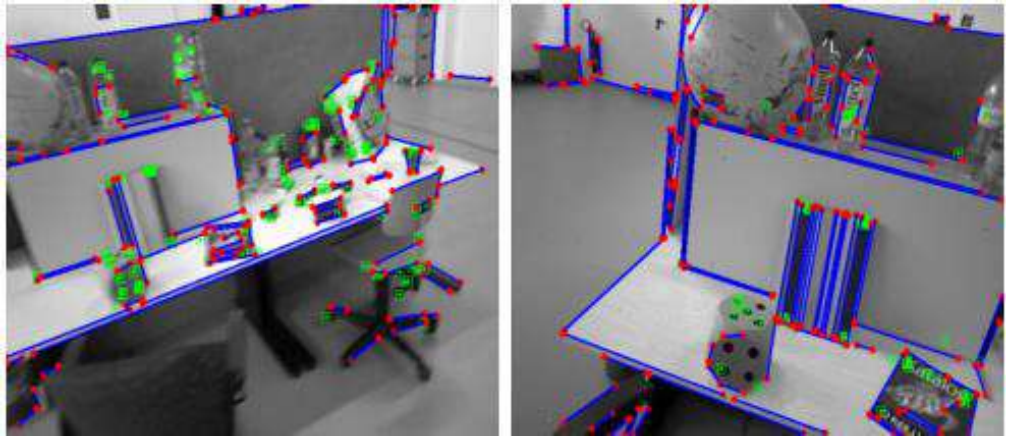
Corridor

Related Works

- **Multi-perspective pose estimation (Miraldo, *et al.*, ECCV 2018)**
- **Stereo relative pose estimation (Vakhitov *et al.* ECCV, 2018)**
- **Line3D++ (CVIU, 2017)**
- **<u>Line features for SLAM systems</u>**
  - **RGBD Odemetry (Lu, et al., ICCV 2015)**
  - **StructSLAM (Zhou, et al., VT 2015)**
  - **StereoSLAM (Zhang, et al., T-RO 2015)**
  - **PL-SLAM (Pumarola, et al., ICRA 2017)**
  - **…**



Line3D++ (CVIU, 2017)



PL-SLAM (Pumarola, et al., ICRA, 2017)

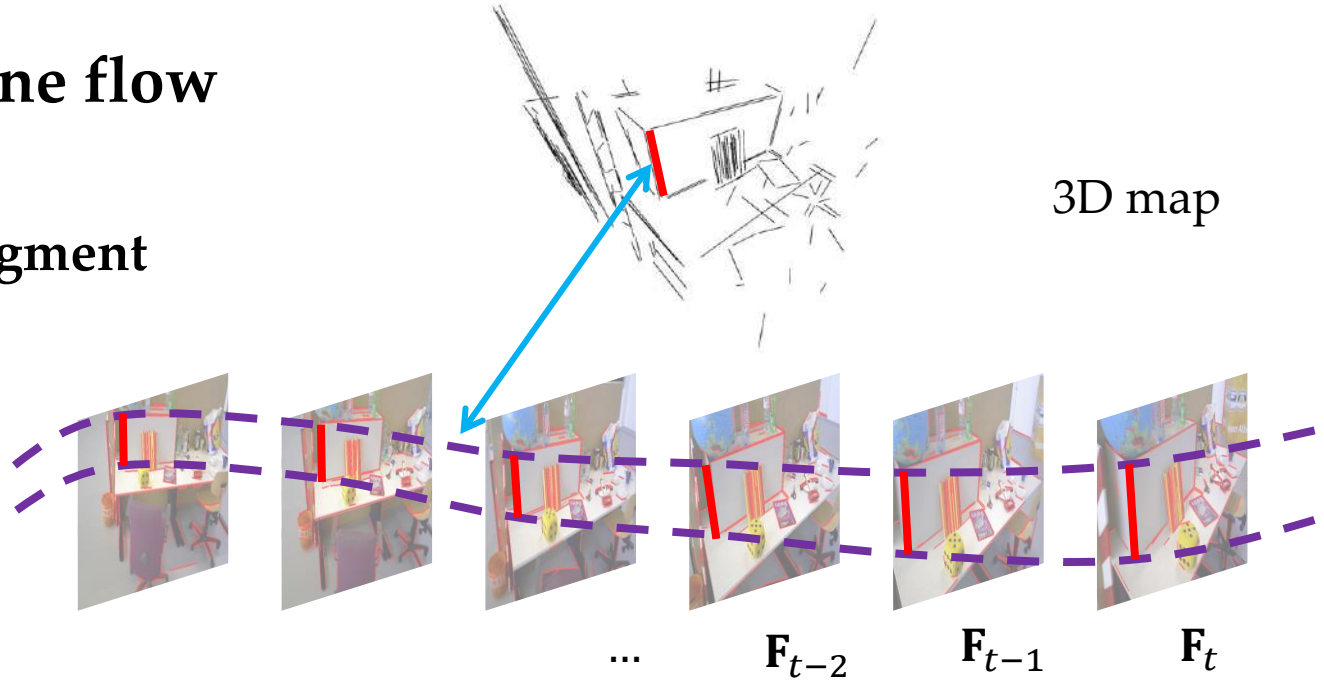
Challenge

- Image blur, occlusion, repetitive structures
- 2D/3D Line segments
  - Extraction
    - Unstable endpoints, easy to broken
  - Matching
    - Similar appearance, heavy computation cost
    - Many-to-many problem due to broken lines
- Real time requirement

Detected line segments in a video

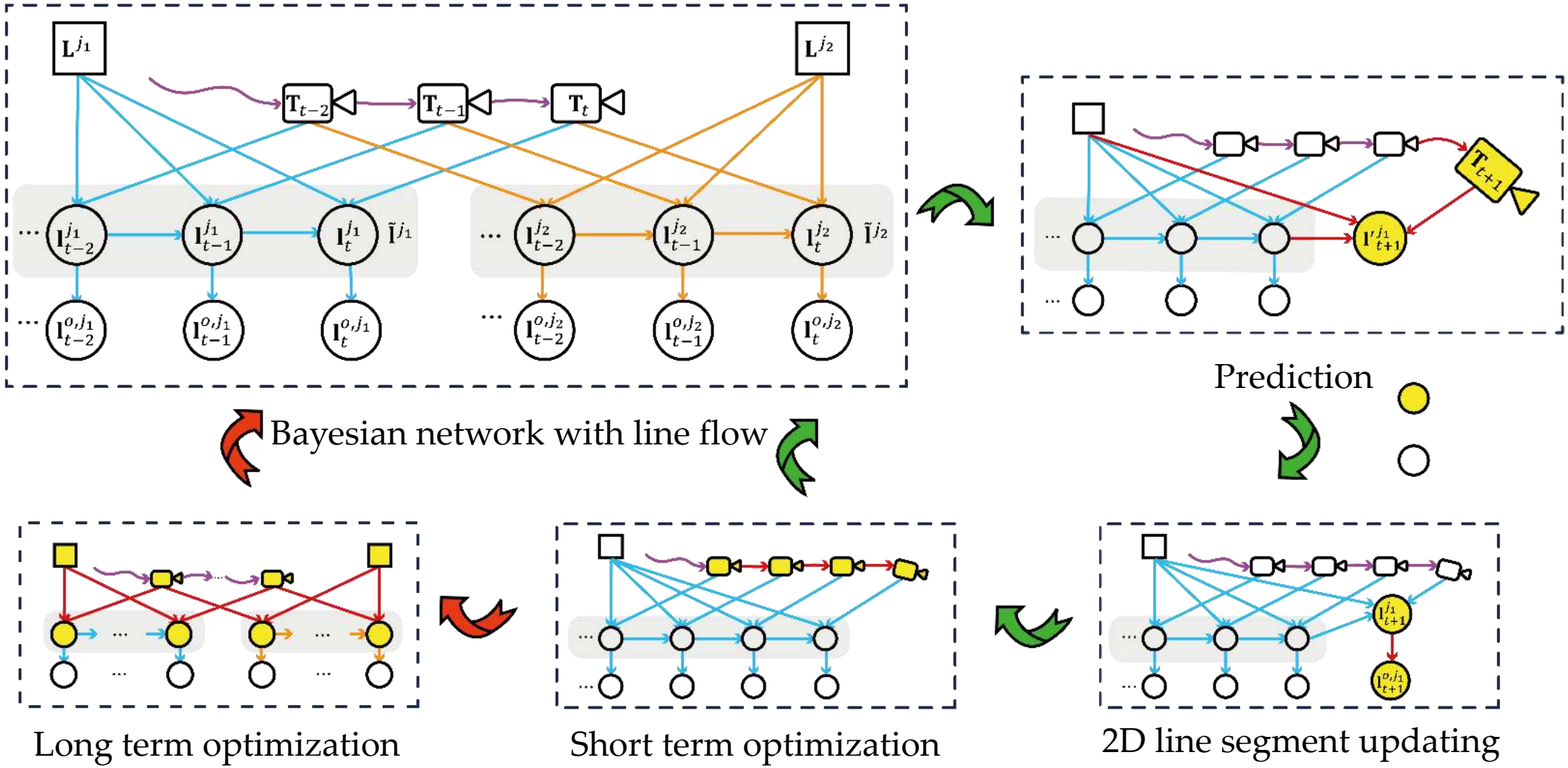
Our Contribution

- **Propose a new representation – Line flow**
  - **Sequential line segments**
  - **Corresponding to the same 3D line segment**
  - **Encoding**
    - **Spatial-temporal 2D constraint**
    - **3D spatial constraint**
- **Line flow based SLAM algorithm**
  - **Bayesian network in an incremental prediction-updating fashion**
  - **Temporally maintained reliable line coherence**



A line flow in a video

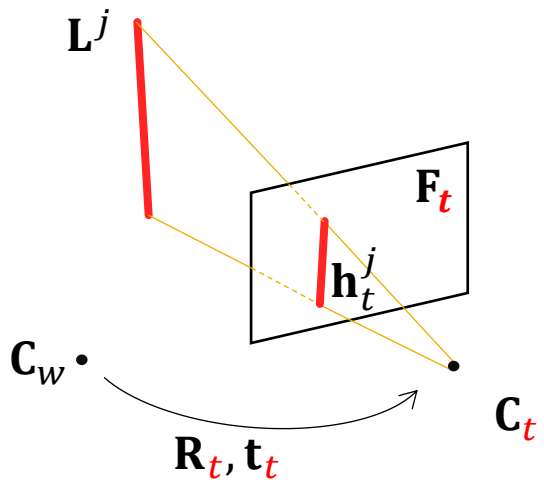
System Framework



Prediction

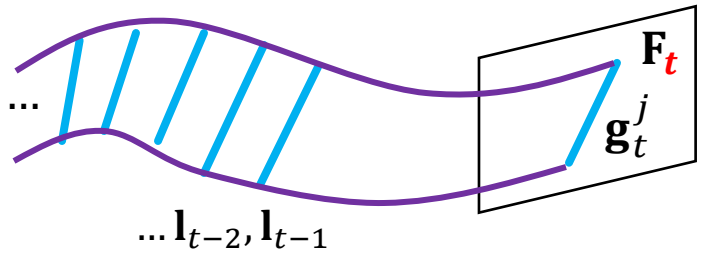
- **3D Spatial constraint**

$$\mathbf{h}_t^j = \mathbf{K}(\mathbf{R}_t \quad [\mathbf{t}_t]_\times \mathbf{R}_t)\mathbf{L}^j$$



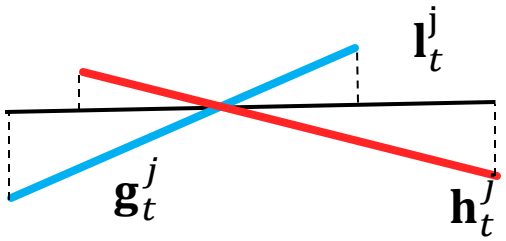
- **Spatial-temporal constraint**

$$\mathbf{g}_t^j = \mathbf{l}_t^j + \mathbf{V}_{\tilde{\mathbf{l}}}^j$$



- **Collinear constraint**

$$\mathbf{l}_t^j = fused\left(\mathbf{h}_t^j, \mathbf{g}_t^j\right)$$

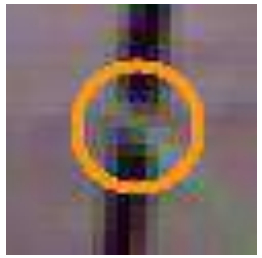


Pose motion: $\mathbf{V_T}$

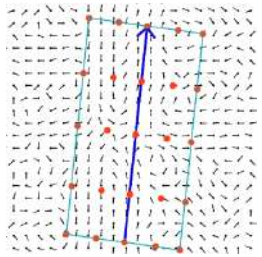
Line flow motion: $\mathbf{V}_{\tilde{\mathbf{l}}}^j$

Updating

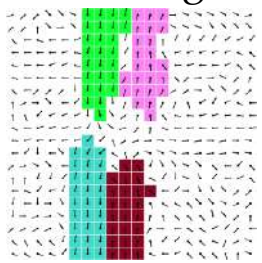
- 2D line segment updating
  - Guided line segment detection
  - Line flow management
- Short term optimization
  - Line re-projection error
  - Pose motion constraint
- Long Term Optimization
  - Keyframe strategy
  - 3D line segment management
  - Local bundle adjustment
  - Loop closure



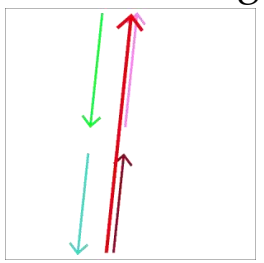
RGB Image



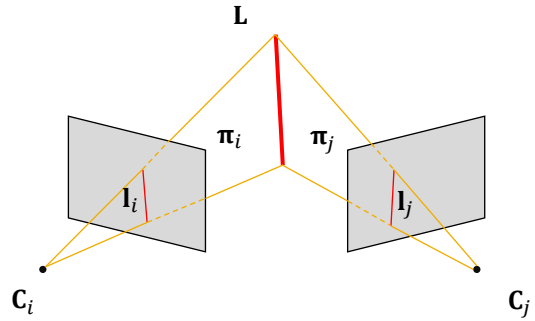
Predicted line segment



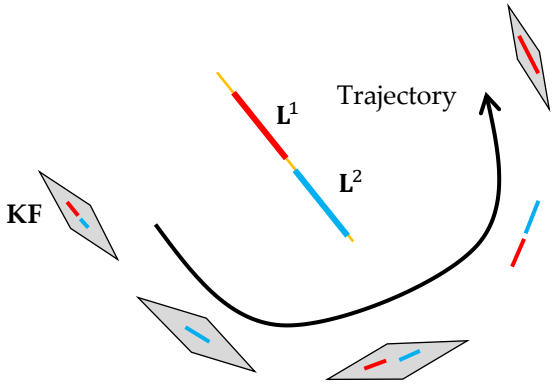
Region growing



Observed line segment



Line flow triangulation



Line flow merging

Experiment

- Platform：PC with 3.6GHz i7-7700 CPU and 16G Memory
- Dataset：
  - TUM RGBD dataset (only use RGB images)
  - 7 Scenes RGBD dataset (only use RGB images)
  - EuRoC dataset (only use left images)
  - Peking university dataset (author-collected dataset)
- Experiment content
  - Line flow evaluation and visualization
  - Localization evaluation
  - Reconstruction evaluation

Localization Evaluation

- **TUM RGBD dataset**
- **7 Scenes dataset**

Method / Dataset	LF-SLAM	PL-SLAM [67] (Re-imp)	PL-SLAM [67] (Ori)	ORB-SLAM [63]	DSO† [18]	DLGO [53]
fr1_xyz	1.14	1.21	1.21	**0.90**	6.30	5.38
fr1_floor	**2.34**	3.91	7.59	2.99	5.25	-
fr2_xyz	**0.27**	0.42	0.43	0.30	0.98	6.53
fr2_360_kidnap	**2.57**	4.60	3.92	3.81	4.12	80.50
fr2_desk_with_person	**0.71**	1.49	1.99	0.88	-	41.20
fr3_str_tex_far	0.91	1.00	0.89	**0.77**	1.36	74.4
fr3_str_tex_near	**0.83**	1.48	1.25	1.58	7.26	-
fr3_nostr_tex_near	**1.21**	1.39	2.06	1.39	7.30	6.00
fr3_sit_halfsph	**1.29**	1.91	1.31	1.34	3.57	-
fr3_long_office	**1.34**	1.40	1.97	3.45	10.11	116.80
fr3_walk_xyz	**1.16**	1.38	1.54	1.24	14.14	27.50
fr3_walk_halfsph	1.64	**1.40**	1.60	1.74	31.86	37.4
average	**1.28**	1.80	2.15	1.70	8.39	43.97

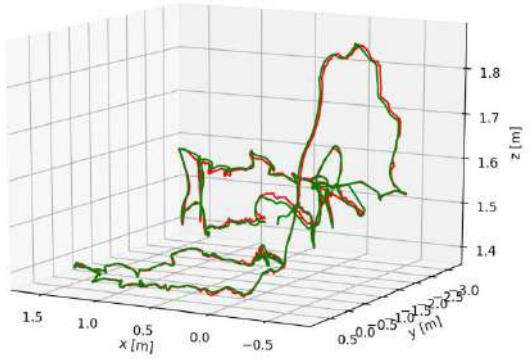
TUM RGBD dataset

Scene_no	LF-SLAM	PL-SLAM [18] (Re-imp)	ORB-SLAM† [15]
chess_01	**4.15**	4.58	5.06
chess_02	4.21	**3.58**	4.06
chess_03	**3.60**	7.58	7.74
fire_01	**4.12**	4.60	4.94
fire_02	**2.26**	3.08	3.10
fire_03	4.15	**3.53**	3.56
heads_01	5.99	**4.26**	6.66
heads_02	**3.66**	4.05	4.98
office_01	**9.38**	10.10	9.74
office_02	**8.33**	12.69	11.44
office_03	**7.95**	7.97	11.80
stairs_01	**12.71**	19.31	48.04
stairs_02	**9.10**	-	-
stairs_03	**5.06**	-	-
stairs_04	**6.42**	-	-
pumpkin_01	**9.91**	13.22	13.25
pumpkin_02	**1.97**	3.33	4.62
pumpkin_03	**8.03**	10.14	10.21
redkitchen_01	**3.92**	5.77	6.03
redkitchen_02	**13.04**	13.24	15.63
redkitchen_03	**4.87**	7.01	7.53
average	**6.33**	7.67	9.86

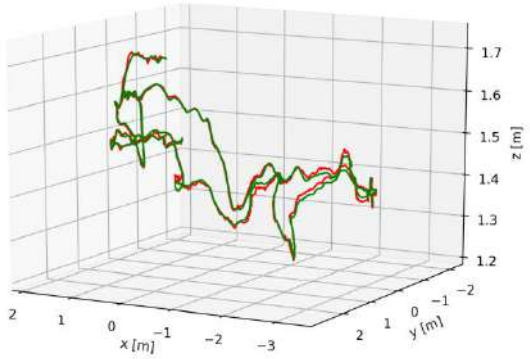
7 Scenes dataset

Localization Visualization

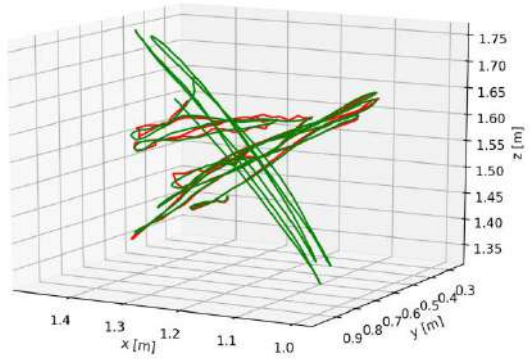
- **TUM RGBD dataset**



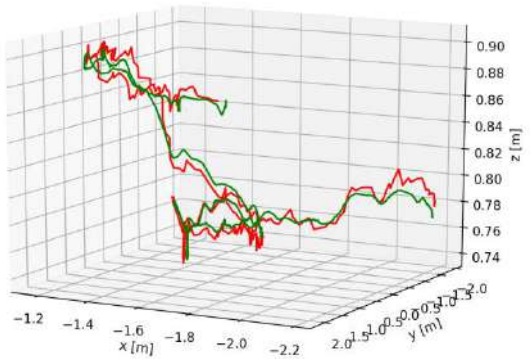
fr2_desk_with_person



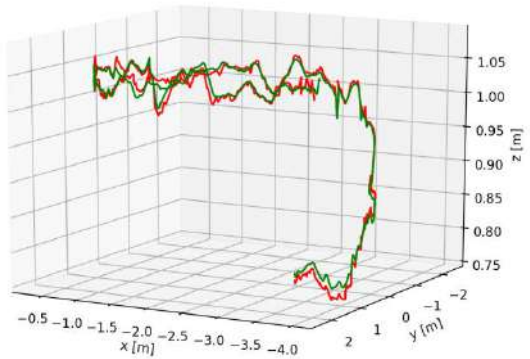
fr3_long_office



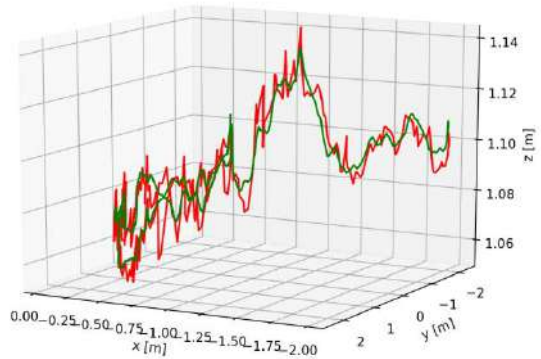
fr1_xyz



fr3_str_tex_near

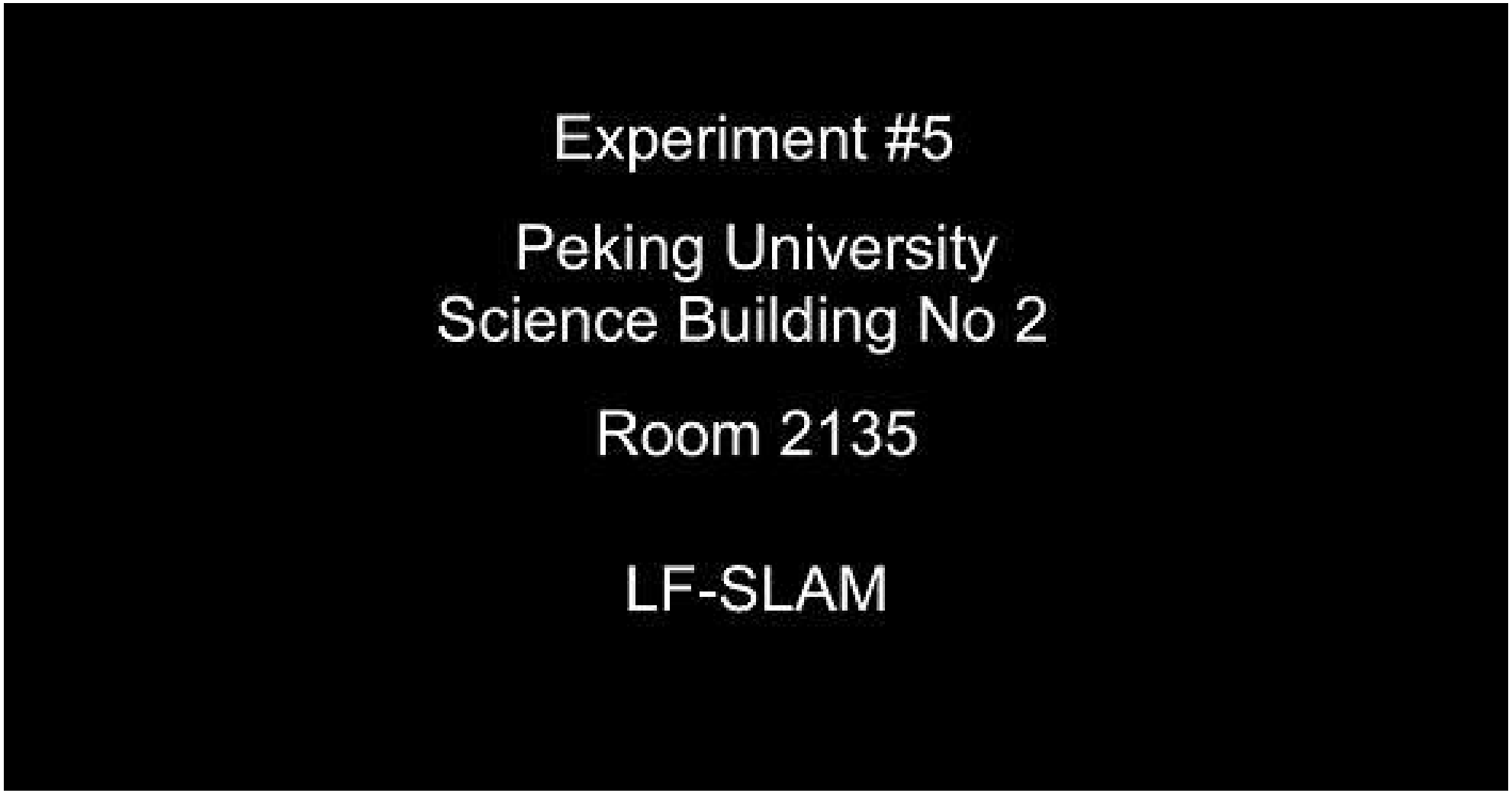


fr3_nostr_tex_near



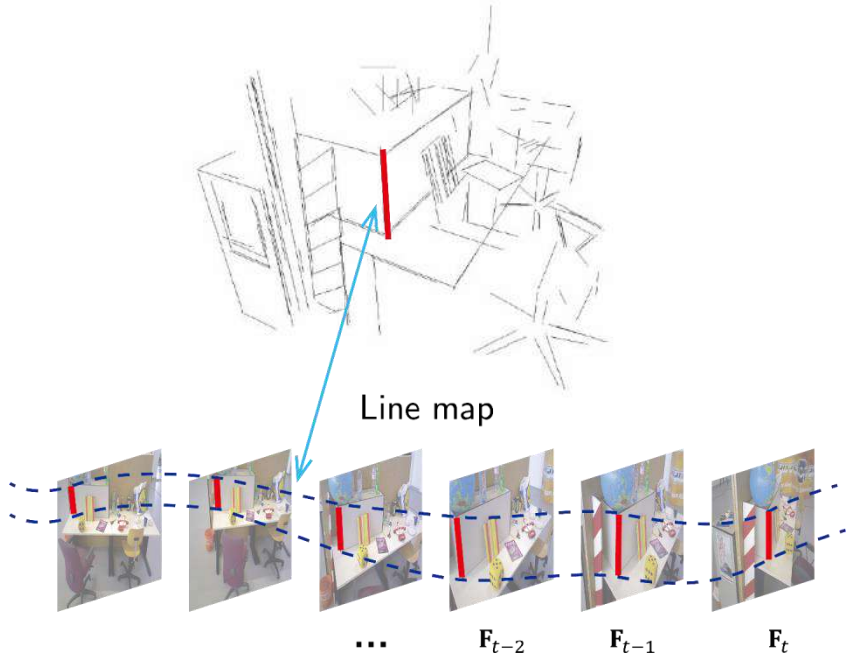
fr3_str_tex_far

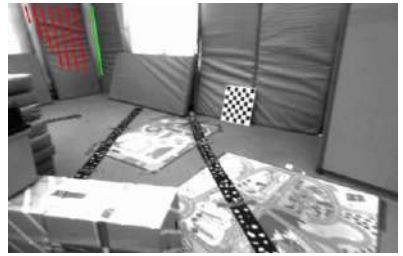
Peking University Dataset



Take-home Message

- Propose a new concept – Line flow
  - Sequential line segments
    - 2D/3D Spatial-temporal constraint
- Line flow based SLAM algorithm
  - Bayesian network in an incremental prediction-updating fashion
  - Temporally maintained reliable line coherence
- Runtime (Comparison with LSD + LBD)
  - LSD + LBD： 37.97 ms
  - LSD + Line flow matching： 21.01 ms
  - Line flow extraction： 16.91 ms
- Total time: 25-30 FPS





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 memory and attentions, as recent cognitive science suggested
- On-line processing performance will be a critical element, thus requiring powerful on-line learning techniques

Thank you!