SLAM II: SLAM for robotic vision-based perception

Autonomous Mobile Robots

Margarita Chli

Martin Rufli, Roland Siegwart

SLAM II | today's lecture

Last time: how to do SLAM?

Today: what to do with SLAM?

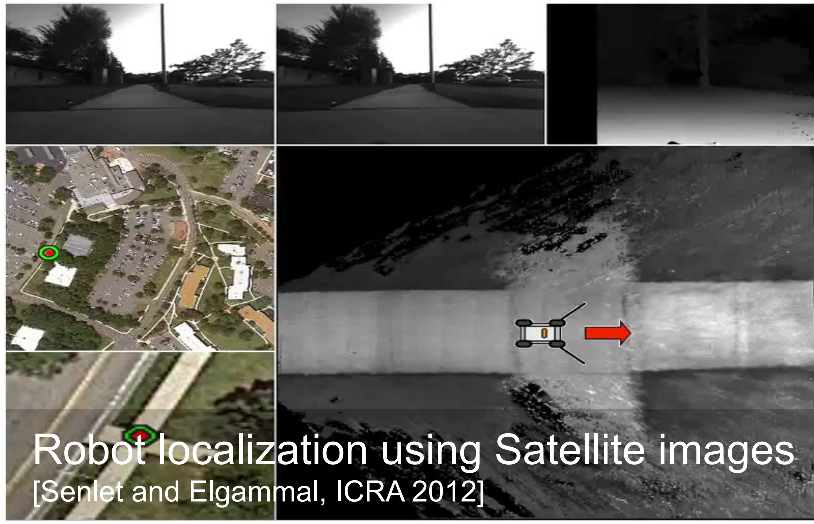
- Vision-based SLAM – state of the art
- Vision-based Robotic Perception:
  - Current Challenges
  - Overview of Research Activities in V4RL
  - Lifelong Place Recognition (Dr Zetao Chen)

Computer Vision meets Robotics| the SLAM problem

SLAM (SIMULTANEOUS LOCALIZATION AND MAPPING):

*"How can a body **navigate** in a previously unknown environment, while constantly building & updating a **map** of its workspace using onboard sensors & onboard computation only?"*

- **The backbone of spatial awareness of a robot**
- One of the most challenging problems in probabilistic robotics
  - **Pure localization with a known map.** SLAM: no a priori knowledge of the robot's workspace
  - **Mapping with known robot poses.** SLAM: the robot poses have to be estimated along the way



Robot localization using Satellite images
[Senlet and Elgammal, ICRA 2012]

Helicopter position given by Vicon tracker
ETH Zurich Flying Machine Arena, IDCS & NCCR DF, 2013

SLAM | how does it work?

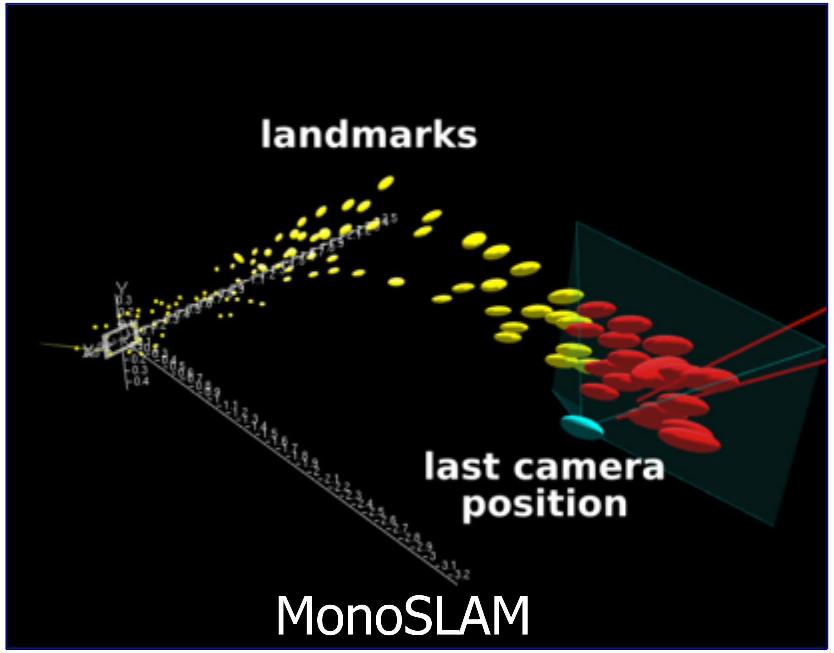
- Can we track the motion of a camera/robot while it is moving?

- Traditional SLAM:
  Pick natural scene features as landmarks, observe their motion & reason about robot motion

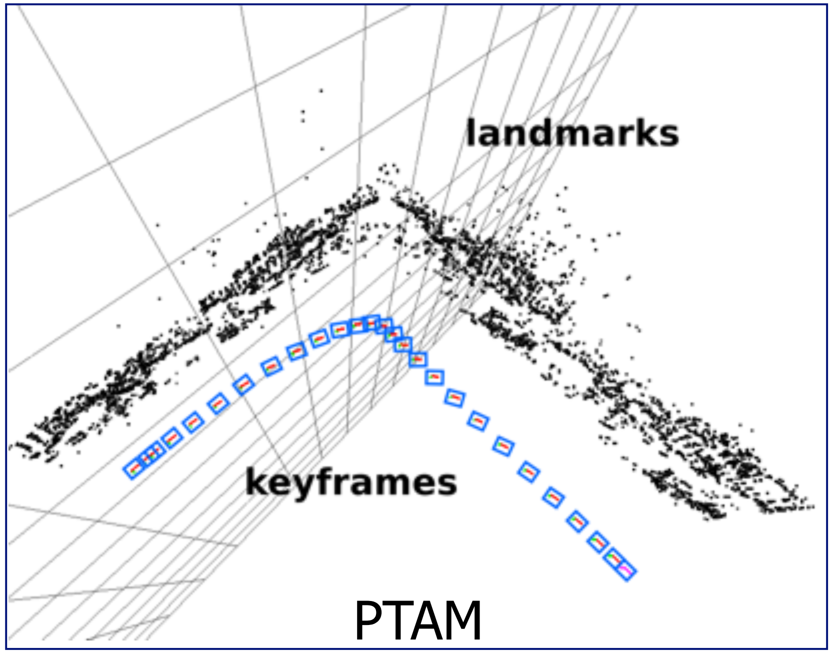
- Research into:
  - "Good" features to track, sensors, trackers, representations, assumptions
  - Ways of dealing with uncertainty in the processes involved

The videos are courtesy of A. Davison

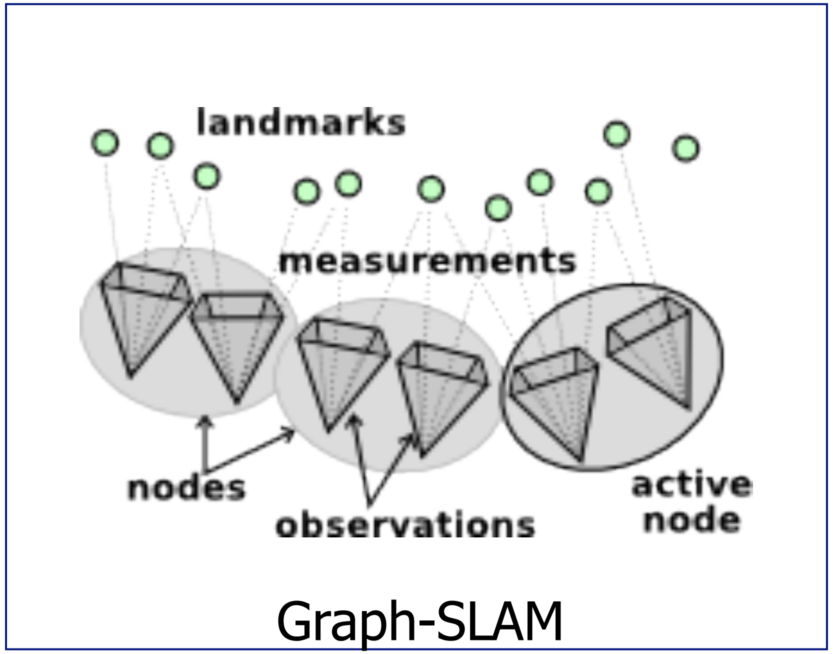
Monocular SLAM | milestone systems



[Davison et al. 2003, 2007]



[Klein, Murray 2007]



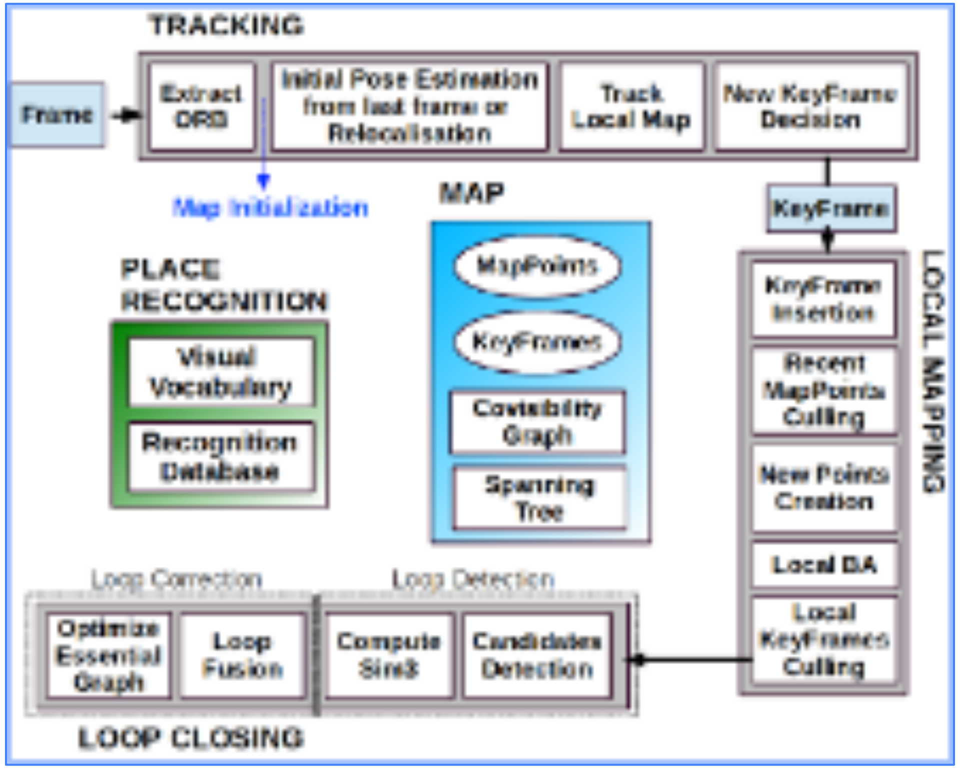
[Eade, Drummond 2007]

✓ revolutionary in the Vision & Robotics communities, but…

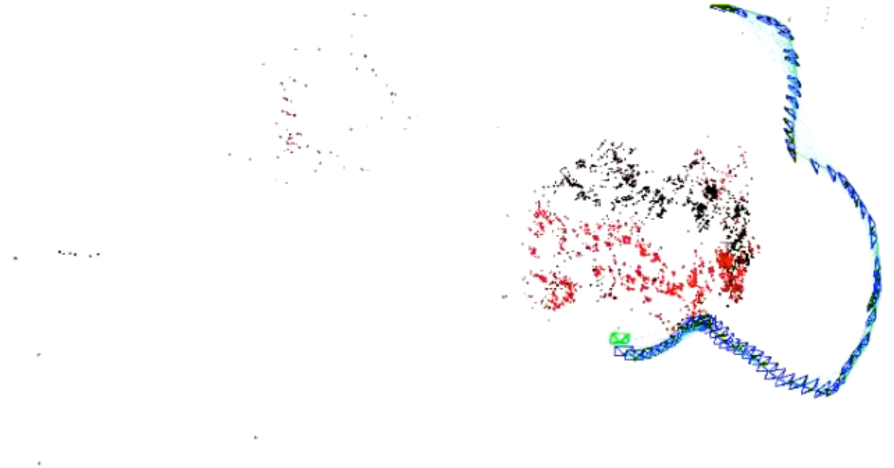
✗ not ready to perform tasks in general, uncontrolled environments

ORB-SLAM [Mur-Artal et al., TRO 2015]

- The most powerful monocular SLAM approach today
- Uses ORB features (binary) in a keyframe-based approach
- Binary place recognition

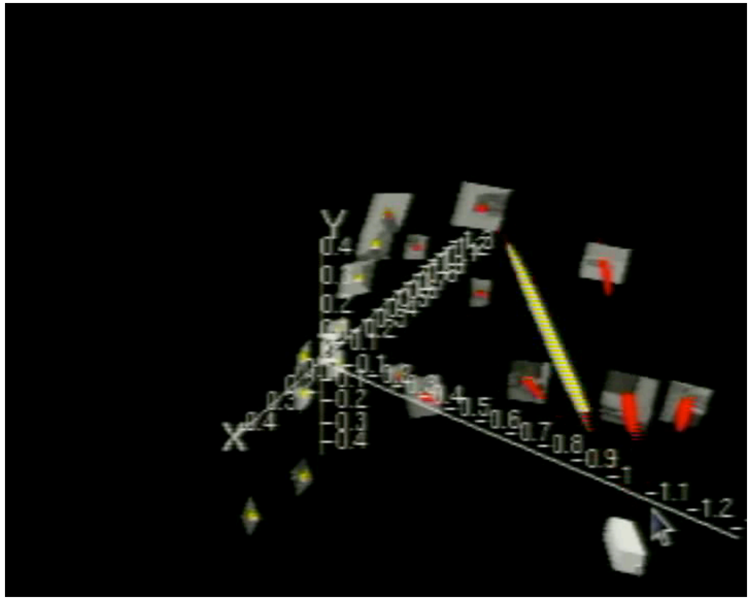


Code available on http://webdiis.unizar.es/~raulmur/orbslam/



Computer Vision meets Robotics | a very short history

2007: [**MonoSLAM**, Davison et al., PAMI]



2009: EU FP7, **sFly**

sFly | swarm of micro flying robots

aim:

Fully autonomous UAVs* to operate in and map an unknown environment in a search & rescue scenario.

*UAV= Unmanned Aerial Vehicle

Small UAVs | properties & challenges

Weight

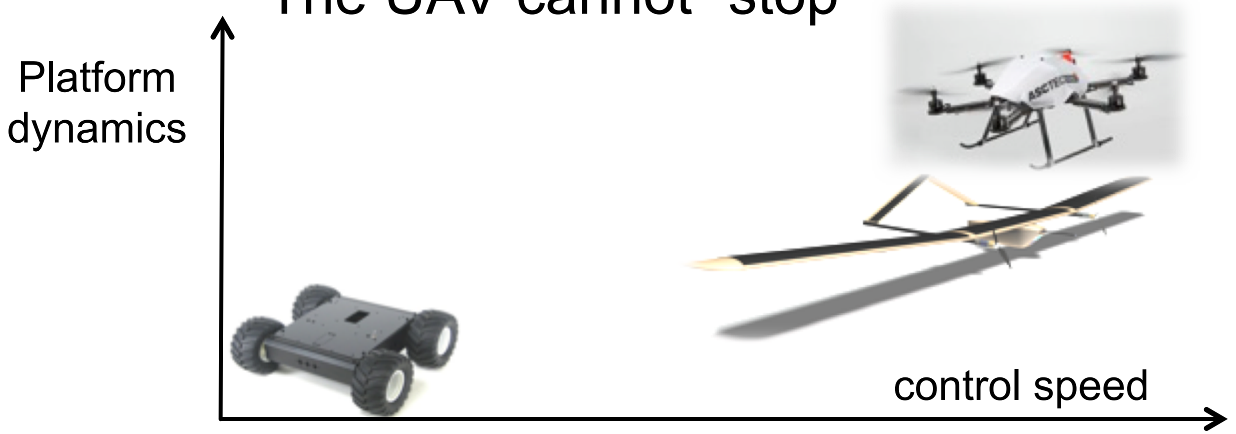
- Lightweight & safe(r) ⇨ easily deployable than larger robots
- Limited payload (<500g): 10g needs approx. 1W in hovering mode
  ➔ Limited computational power onboard ➔ choose sensors with high information density

Autonomy

- Low bandwidth / unreliable data links
  ⇨ onboard processing
- Limited battery life (~10mins)

Agility

- Highly agile (up to 8m/s)
- Fast, unstable dynamics
- High-rate real-time state estimation.
  The UAV cannot “stop”

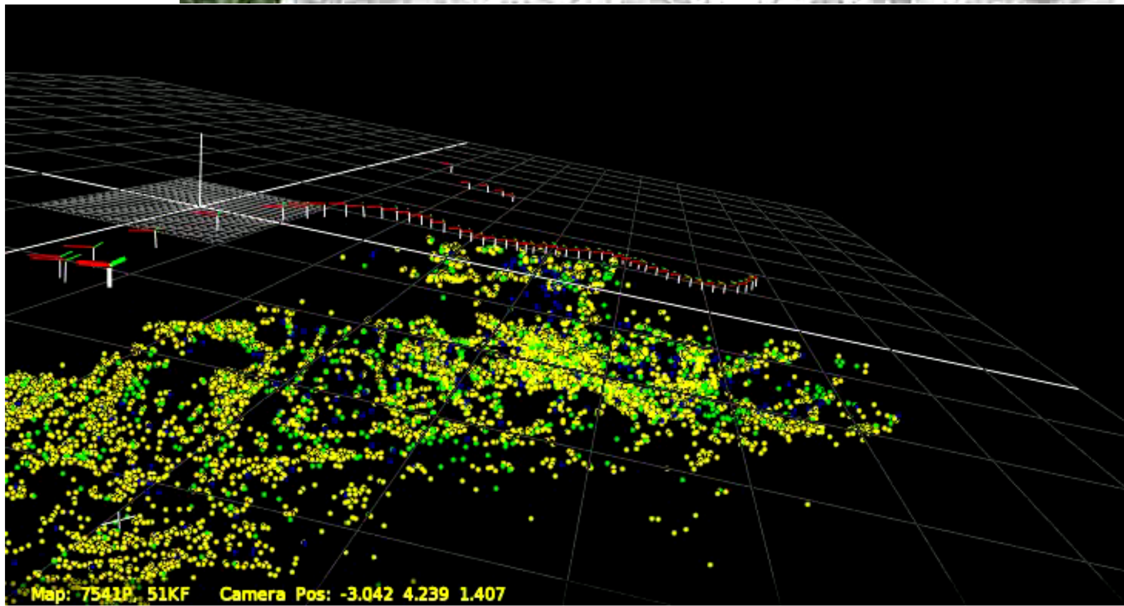


sFly | enabling UAV navigation

aim: autonomous vision-based flights in unknown environments

approach: minimal sensor setup
→ essentially fuse visual & inertial cues

- **Downward-looking camera**: bearing only measurements
  → Monocular SLAM (based on PTAM)
- **IMU:** Acceleration & angular velocity
- Loosely-coupled visual-inertial fusion



Failure detection

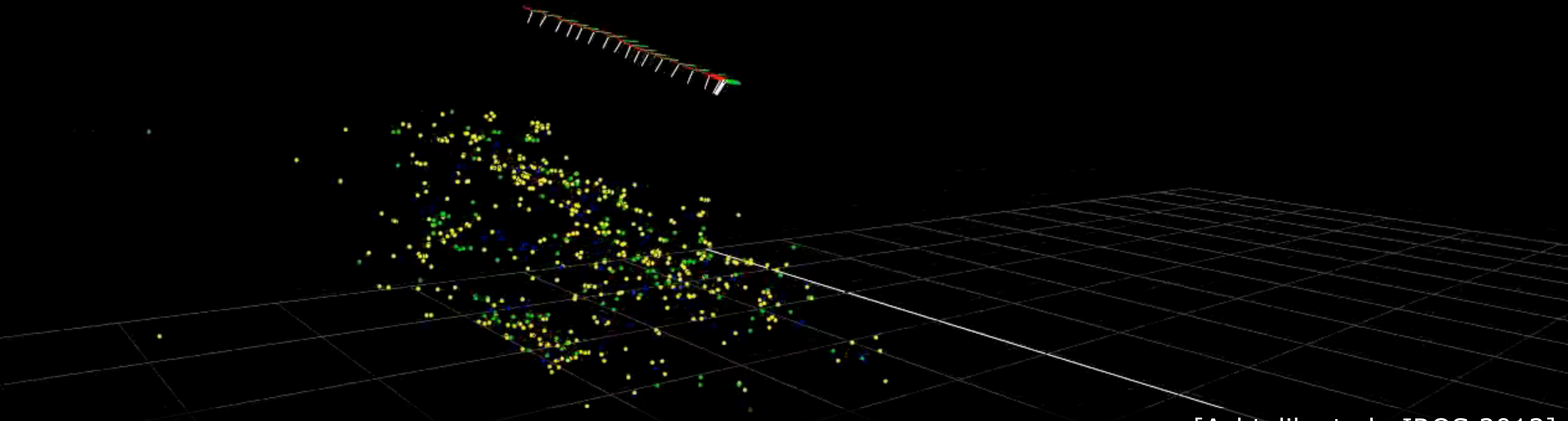
Monocular SLAM → unscaled pose + cov → **EKF**

IMU readings → acc gyro → **EKF**

EKF
State: pose, scale, biases
Complexity = constant

[Weiss et al., JFR 2013]

Flights controlled using visual & inertial cues



[Achtelik et al., IROS 2012]

Vision-based UAV navigation

- First UAV system capable of vision-based flights in such real scenarios
- Publicly available framework used by NASA JPL, UPenn, MIT, TUM,…

Are we there yet?

Photo credit: Francois Pomerleau

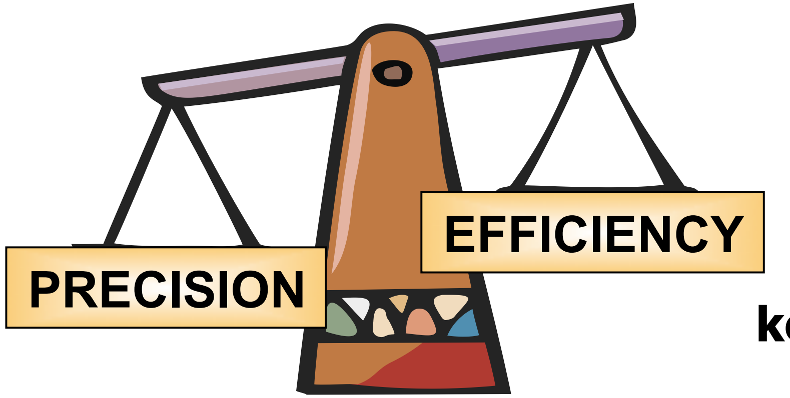
SLAM | current challenges

- Fast motion
- Large scales
- Robustness
- Rich maps
- Low computation for embedded apps
- Combination of multiple agents

↳ dynamic scenes, motion blur, lighting, …

1ms Target Tracking Vision Chip
by Ishikawa Komuro University

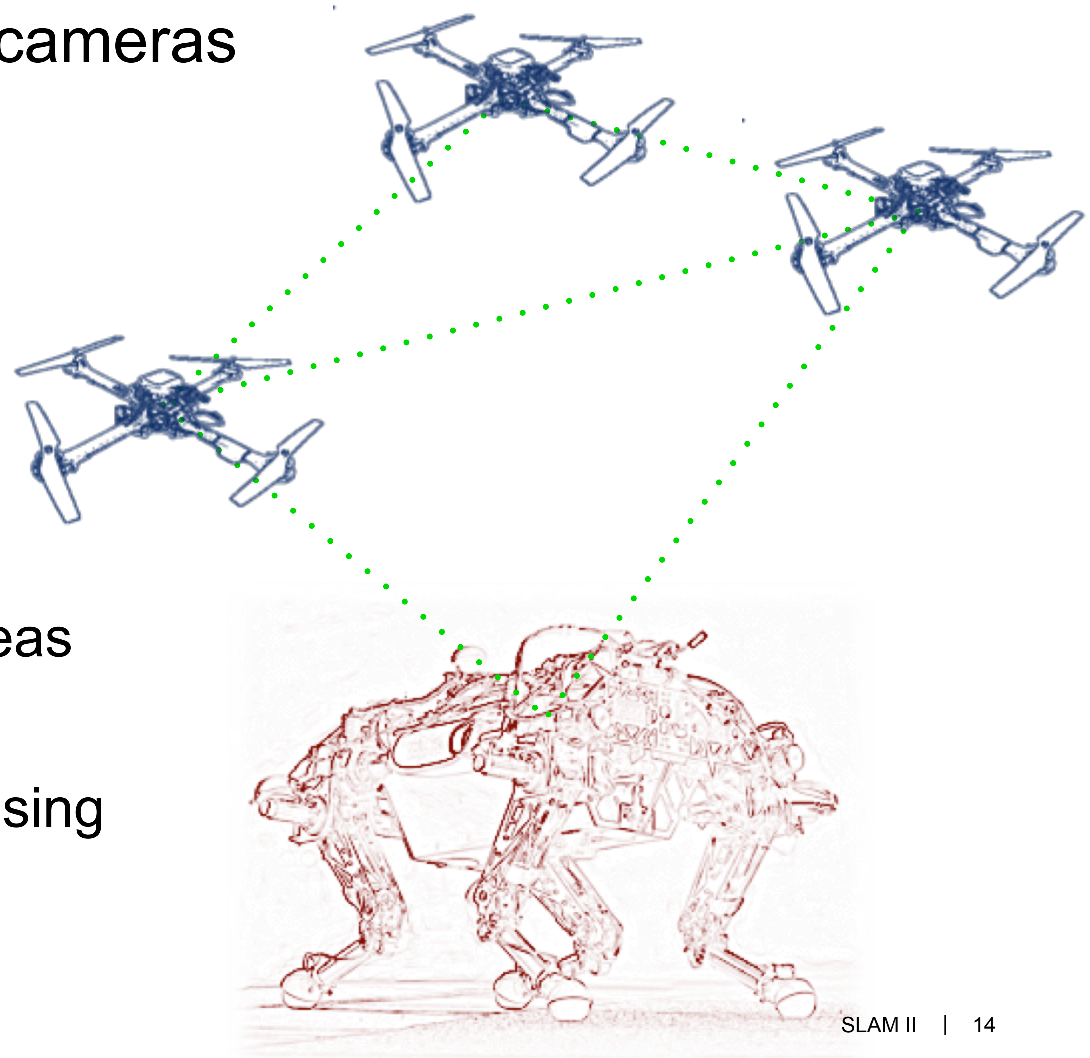
- Handle **larger** amounts of data more **effectively**
- Competing goals:



key: agile manipulation of information

Robotic Perception | what next?

- Employ team of aerial robots equipped with cameras
- Develop visual perception & intelligence to:
  - Navigate autonomously
  - Collaboratively build a 3D reconstruction of the surrounding area



UAVs : Unmanned Aerial Vehicles

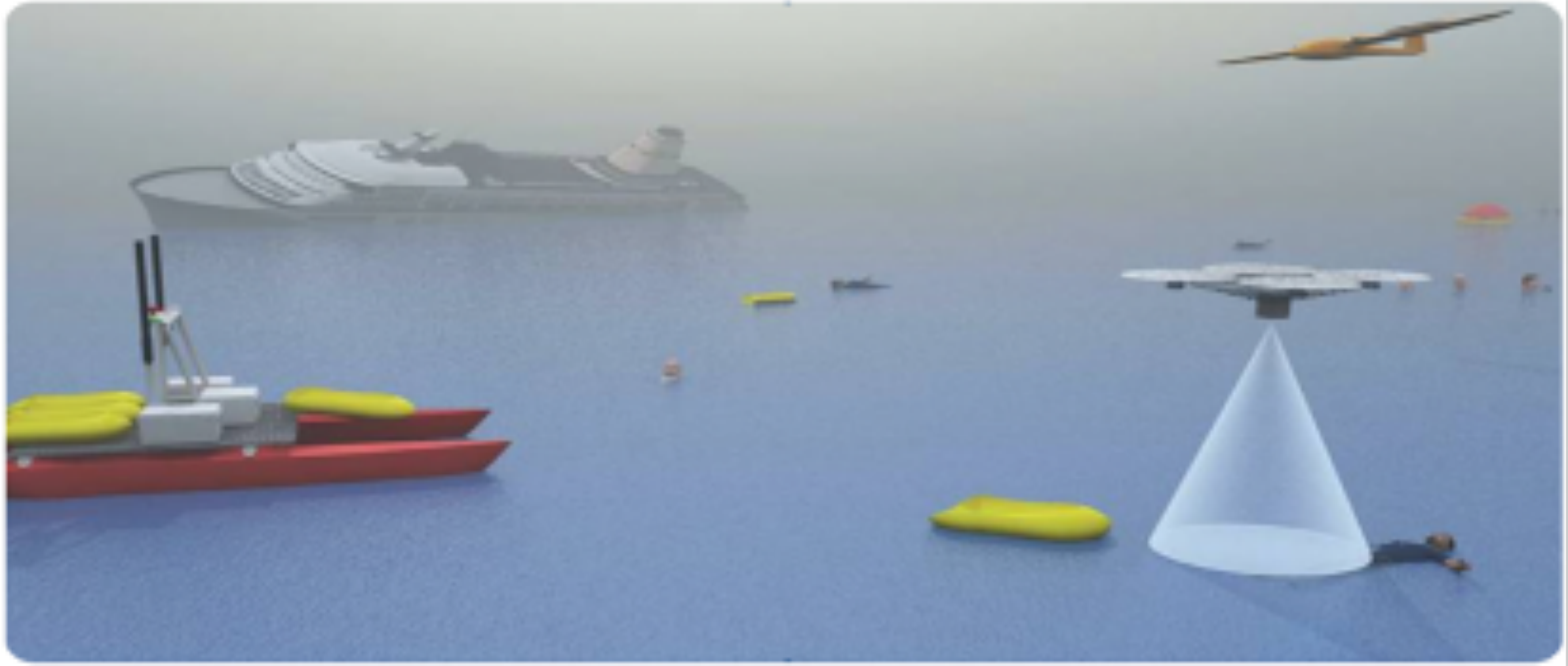
- Agile, easy access to scene overview & remote areas
- Dynamics hard to track, limited payload
  ⇨ *collaboration* is key to efficient sensing & processing
- Extension to additional platforms

AEROWORKS | EU project

- Team of small UAVs: each equipped with visual & inertial sensors and a manipulator
- **Aim**: collaboratively perceive the environment, develop autonomy in navigation and coordination to perform a common manipulation task
- V4RL: collaborative vision-based perception for navigation & 3D reconstruction
- 2015-2018, 9 partners

ICARUS| EU project

- **I**ntegrated **C**omponents for **A**ssisted **R**escue and **U**nmanned **S**earch operations (2012-2015), budget: 17.5 M€, 24 partners
- Search-and-rescue combining robotics for land, sea and air
- ETHZ: map generation, people detection, … from a UAV



SHERPA | EU project

- **S**mart collaboration between **H**umans and ground-a**E**rial **R**obots for im**P**roving rescuing activities in **A**lpine environments
- 11 M€, 10 partners, 2013-2017
- Sensor fusion (visible light and thermal cameras, IMU, …) for robust SLAM, environment reconstruction & victim localization

Vision-based Robotic Perception | the challenges

Challenge I: High-fidelity localization & mapping

- The backbone of perception of space & navigation autonomy

- Pioneering work in UAV navigation, but lacks scalability, robustness and deployability for application in real scenarios

Suitable keypoint detection/description

- Image keypoints suitable for robotics applications: for fast & robust detection and matching
- Rotation-, scale-invariant keypoints
- Binary descriptor: e.g. BRISK, ORB, BRIEF & variants

Descriptor	Run time [ms.]
SURF	117.1
SIFT	448.6
BRIEF	**3.8**
BRISK	**10.6**
ORB	**4.2**

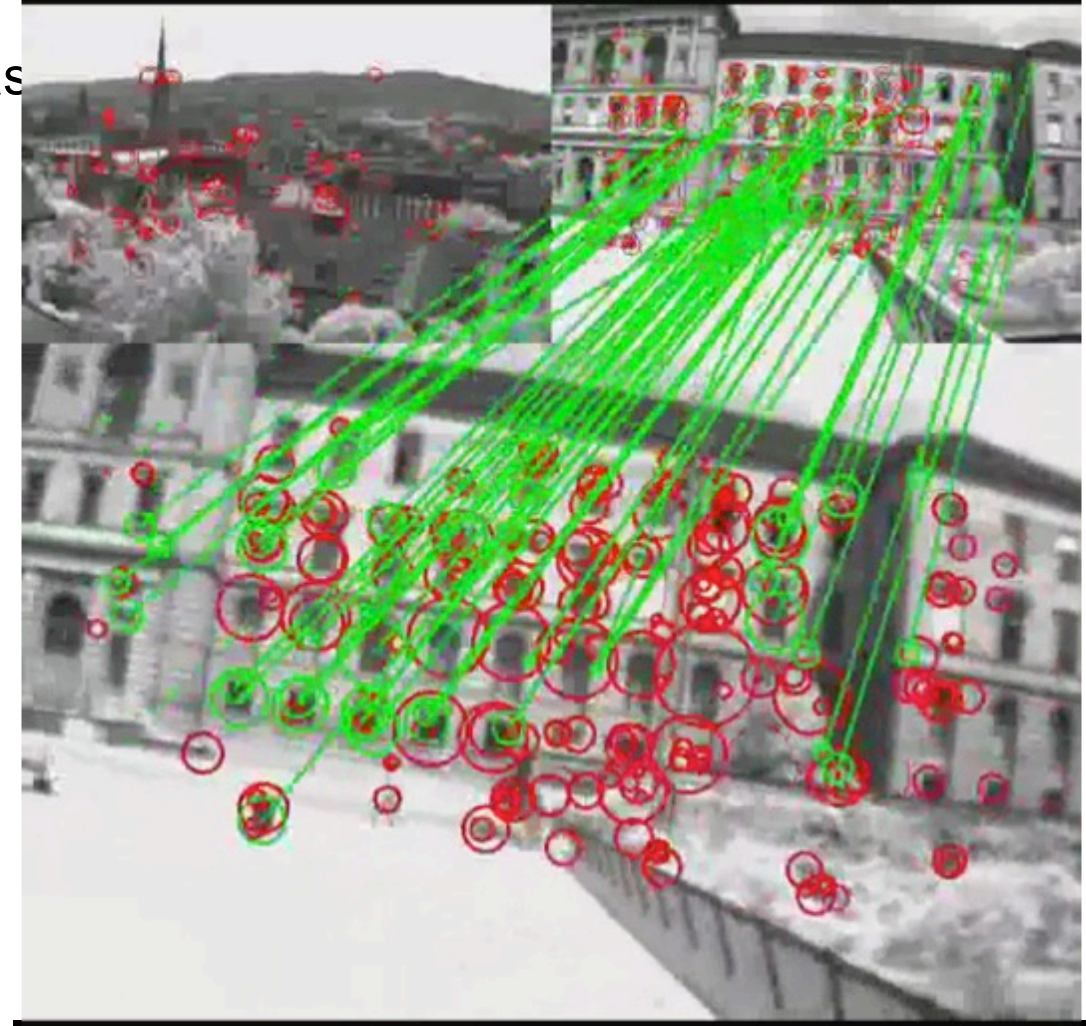
BRISK descriptor

1	0	0	1	1	1	...	0	0	1	0	0	1	1	1	0

SURF descriptor

0.4	1.0	0.1	...	0.3	0.4	0.7	0.6	0.11	2	5	0.1	0.7

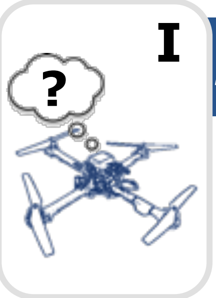
Test Frame 1 | Test Frame 2



[Leutenegger et al., ICCV 2011]

BRISK:

- Precision-Recall: comparable to SIFT & SURF
- ~10x faster than SURF
- Open-source, features in OpenCV



Sensor fusion for SLAM

- Visual-Inertial sensor:
  HW-synced stereo camera (global shutter) + IMU

"OKVIS": visual-inertial SLAM

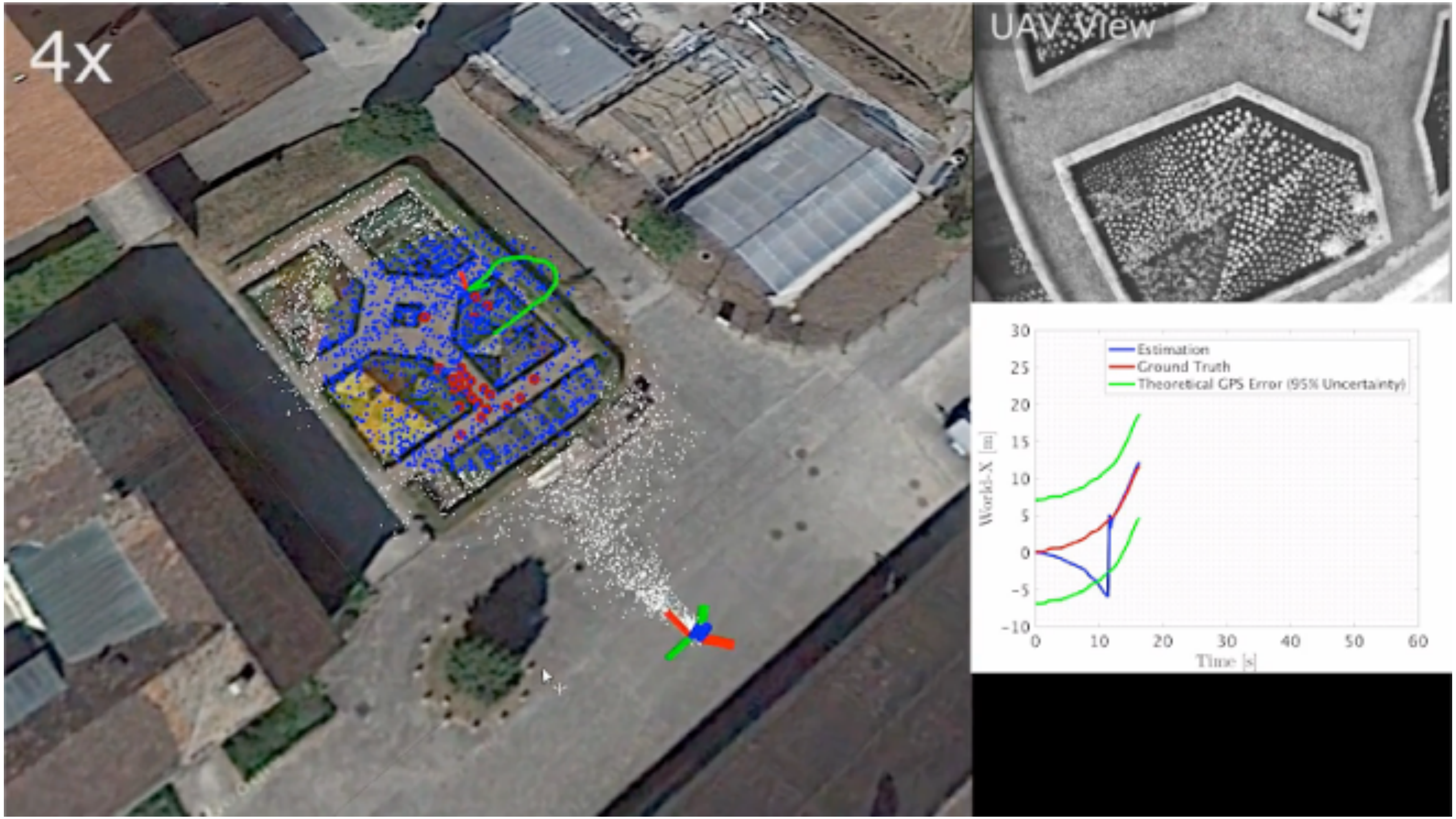
- Tight visual & inertial fusion: replace motion model with IMU constraints on the actual motion

- Visual cues: very descriptive, but sensitive to motion blur, lighting conditions…

- Inertial cues: accurate estimates for short-term motions, unsuitable for longer-term

- Open-source: http://ethz-asl.github.io/okvis_ros/

Robust VI SLAM for repetitive flights [Surber et al., ICRA 2017]

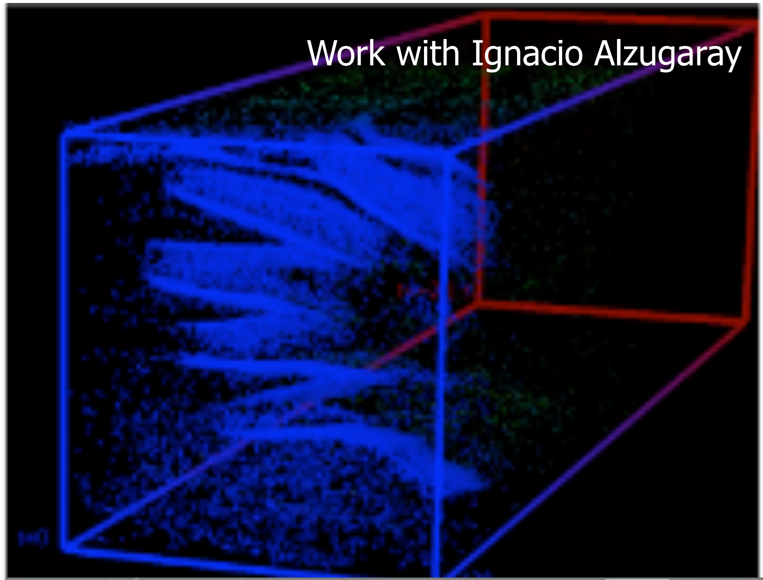


Event-based Cameras for Robot Navigation

V4RL VISION FOR ROBOTICS LAB

- Dynamic Vision Sensor (DVS)

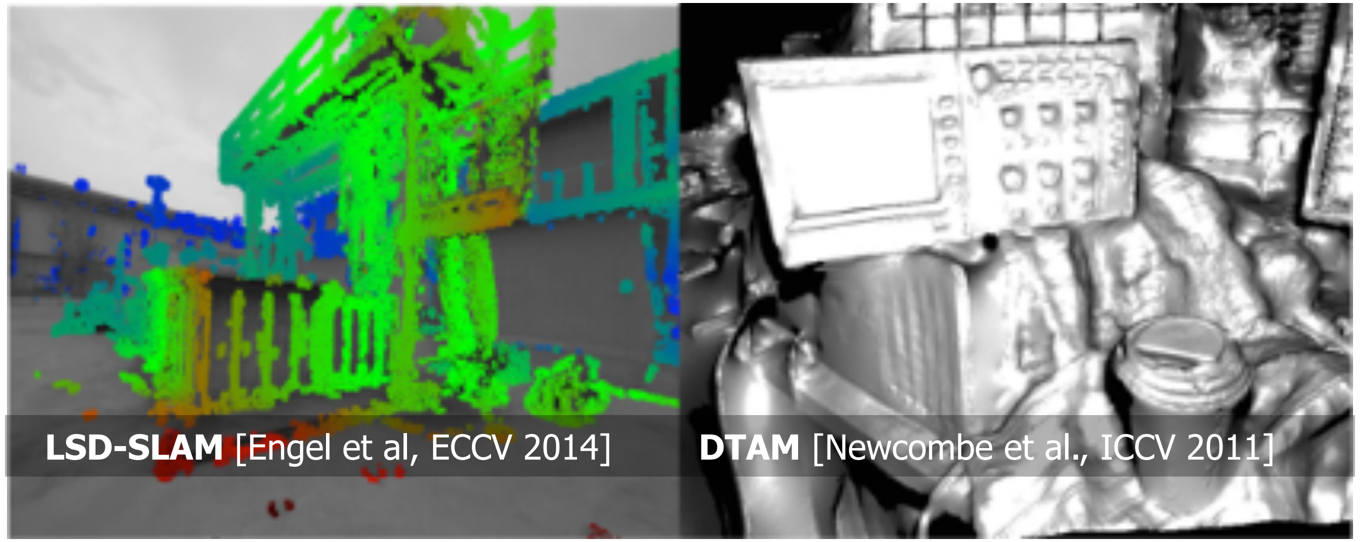
- Similar to the human retina:
  captures intensity changes asynchronously instead of capturing image frames at a fixed rate
  - ✓ Low power
  - ✓ High temporal resolution ➔ tackle motion blur
  - ✓ High dynamic range

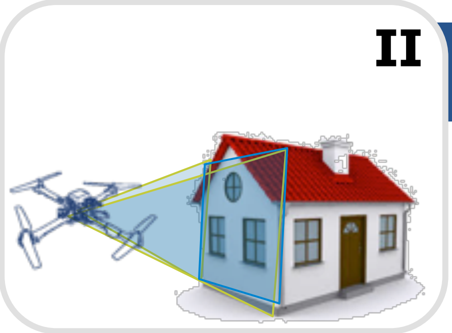


Vision-based Robotic Perception | the challenges

Challenge II: Dense scene reconstruction

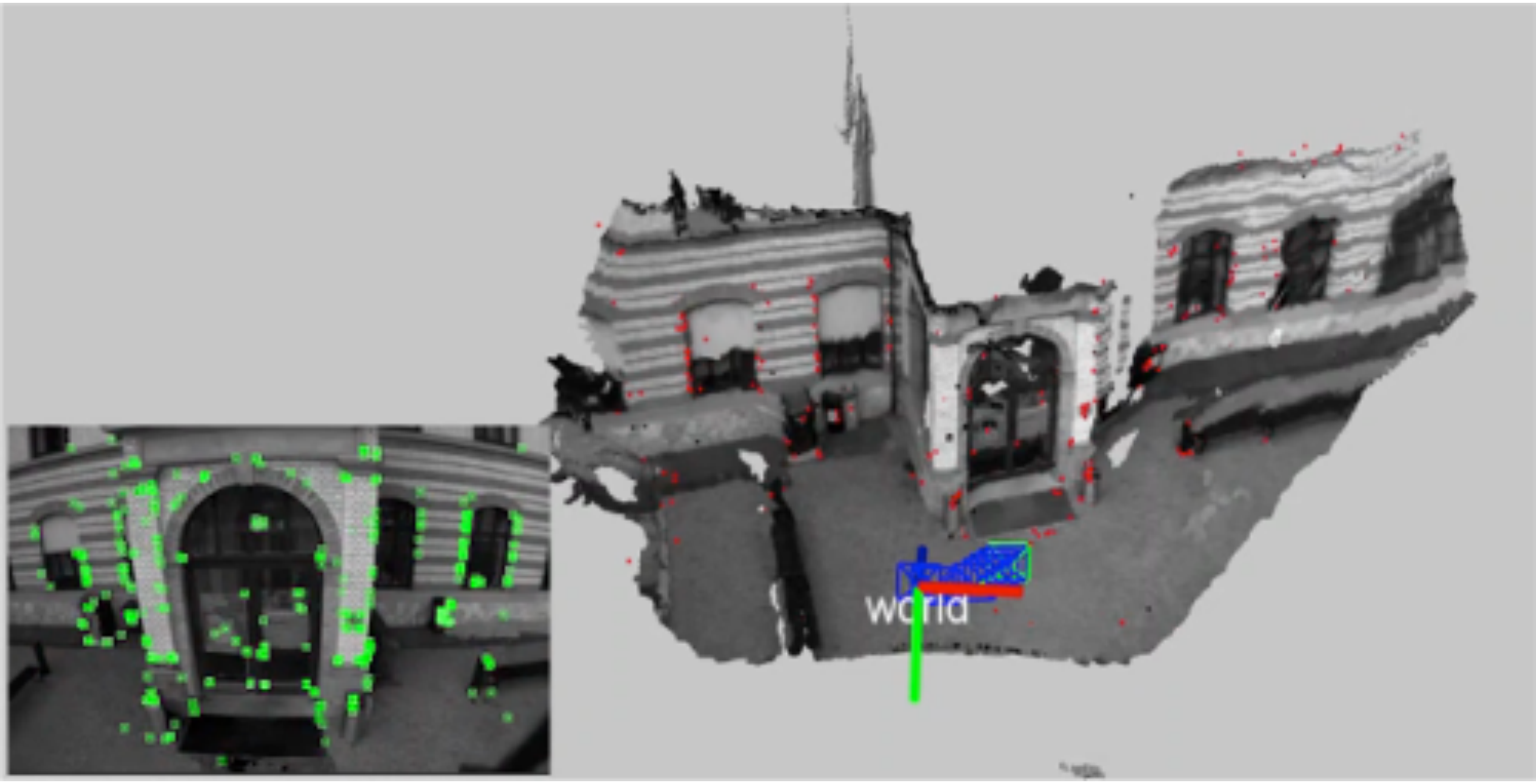
- Vital for robot interaction with its environment
- Trade-off: level of detail vs. computational cost
- Work towards both
  (a) online onboard and
  (b) scalable offboard functionality



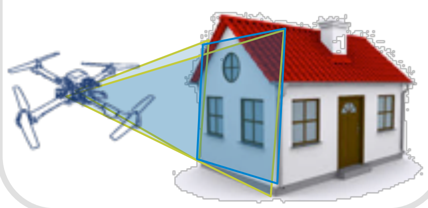


Towards low-cost, denser
3D reconstruction with a single camera

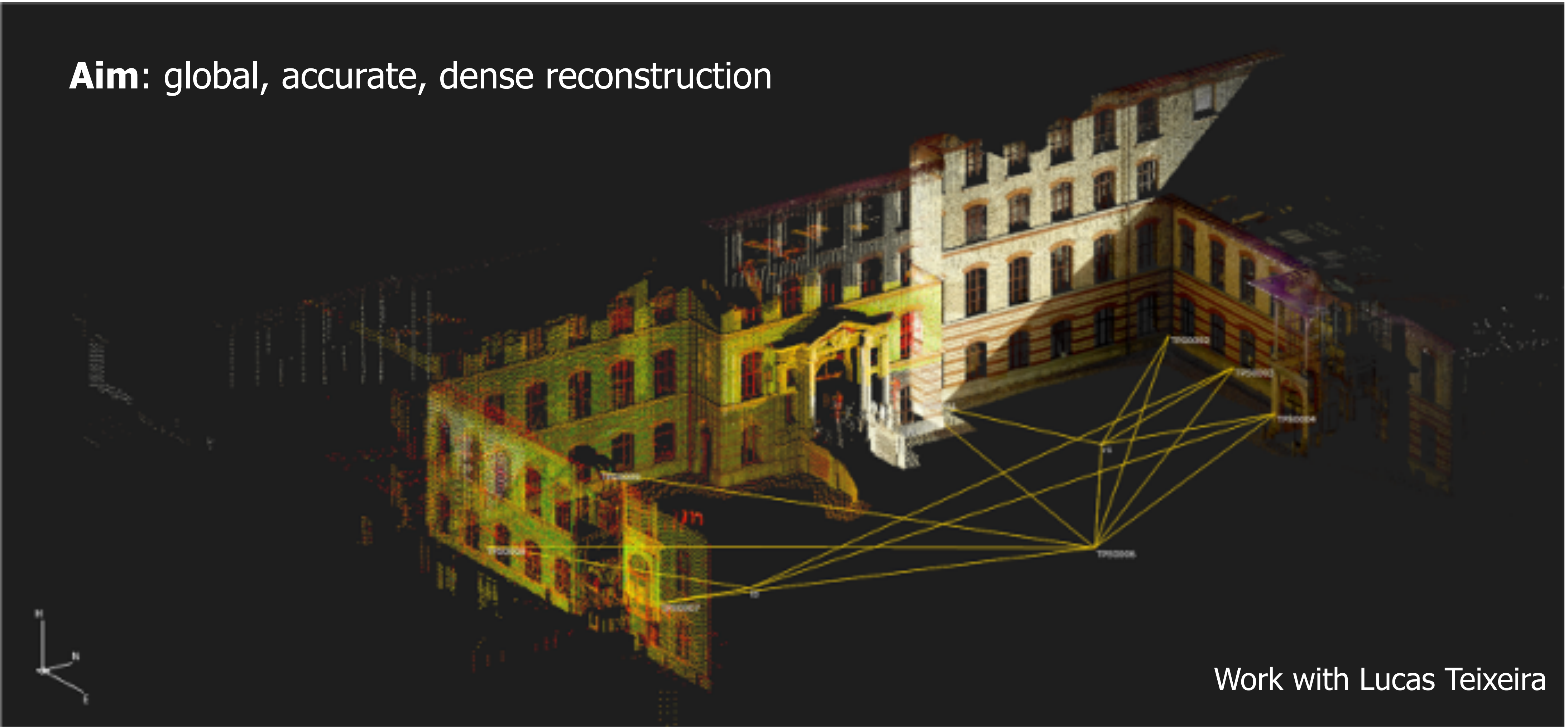
V4RL VISION FOR ROBOTICS LAB



[Teixeira & Chli, ICRA 2017]

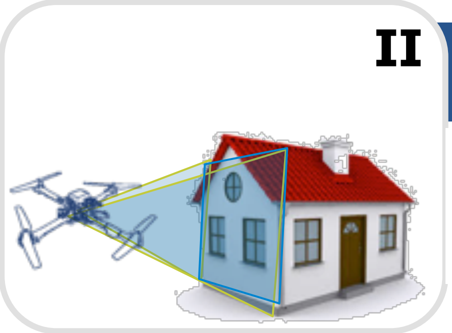


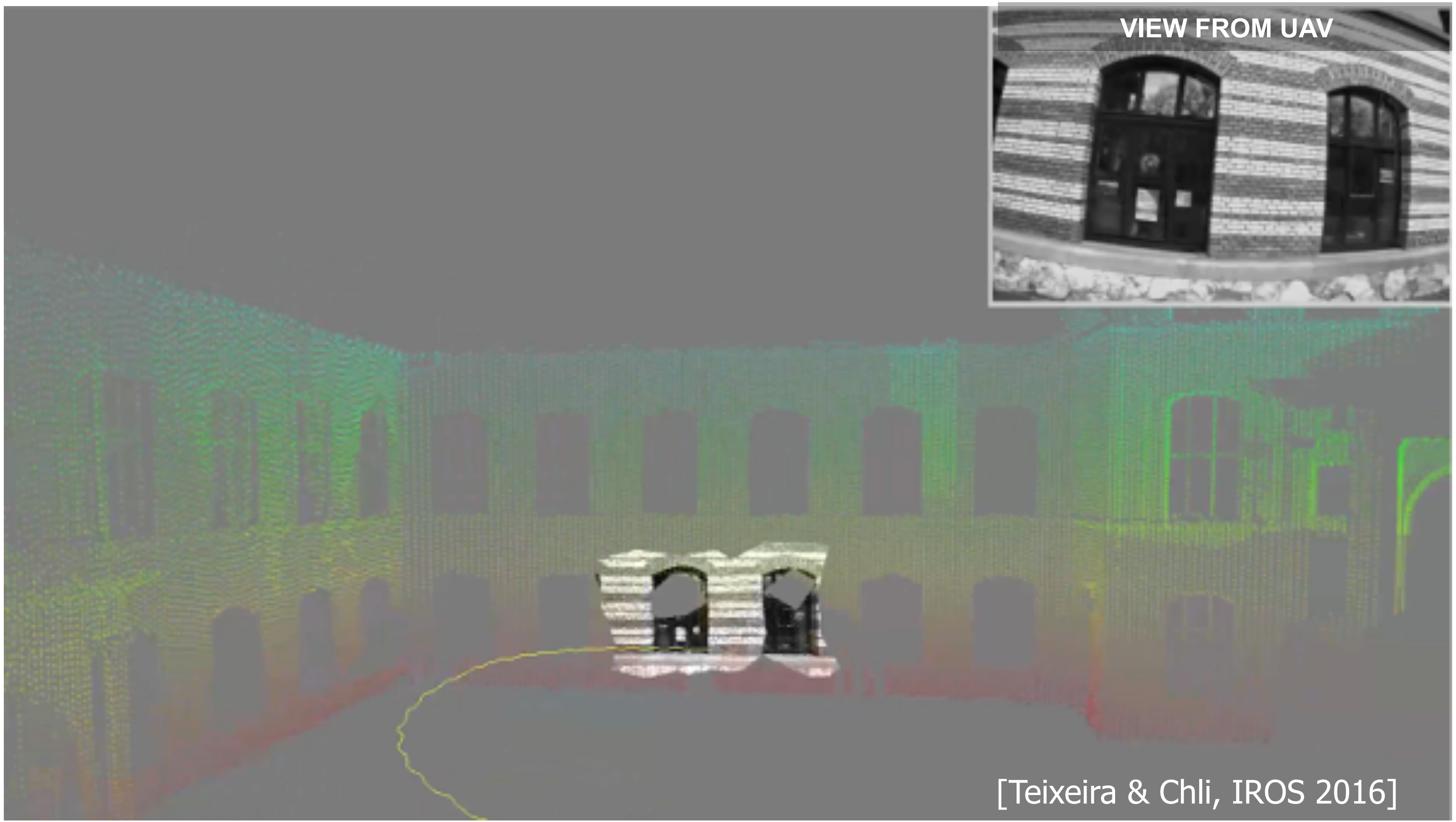
Towards low-cost, denser 3D reconstruction with a single camera



Leica Theodolite Total Station

Towards low-cost, denser 3D reconstruction with a single camera from a small UAV

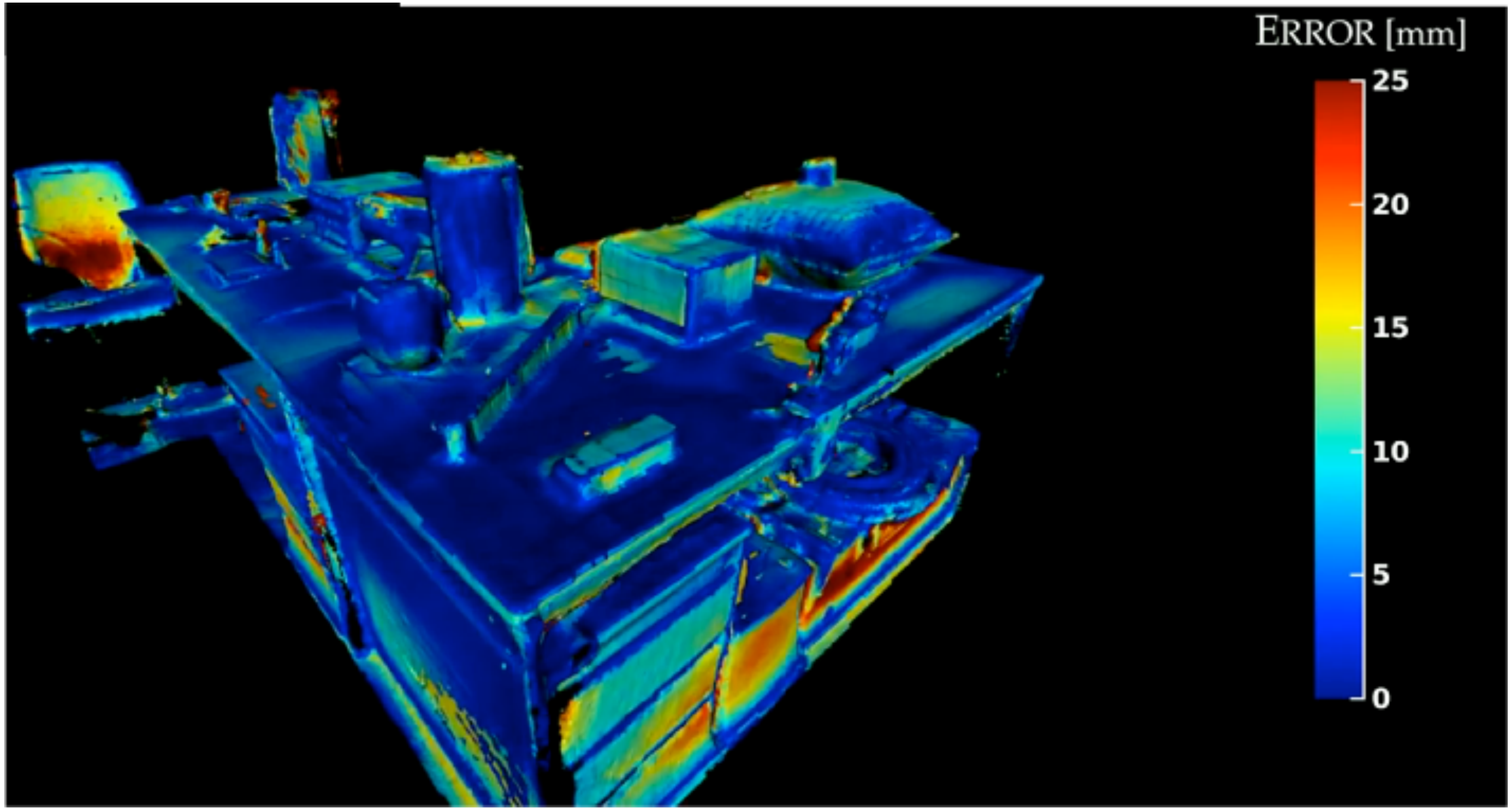




[Teixeira & Chli, IROS 2016]

- Monocular-inertial SLAM (OKVIS)
- Isolate reliable SLAM points ➔ form regular, “smooth” mesh
- Denser representation in < 8ms per frame
- Datasets & Code on www.v4rl.ethz.ch

Real-time Dense Surface Reconstruction for Manipulation



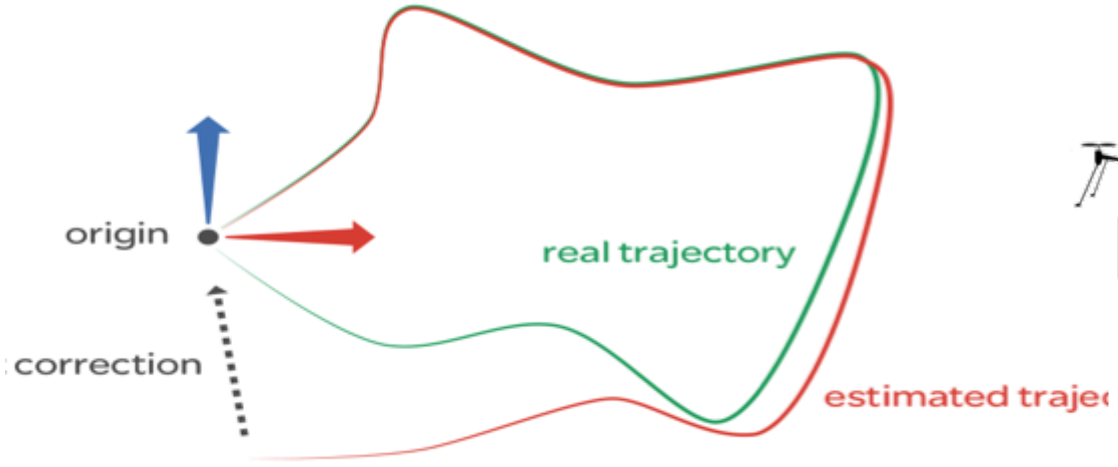
[Karrer et al., IROS 2016]

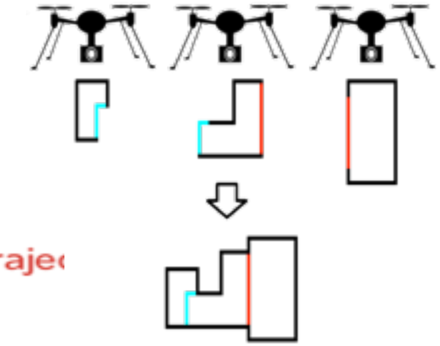
- Datasets & ground truth on www.v4rl.ethz.ch

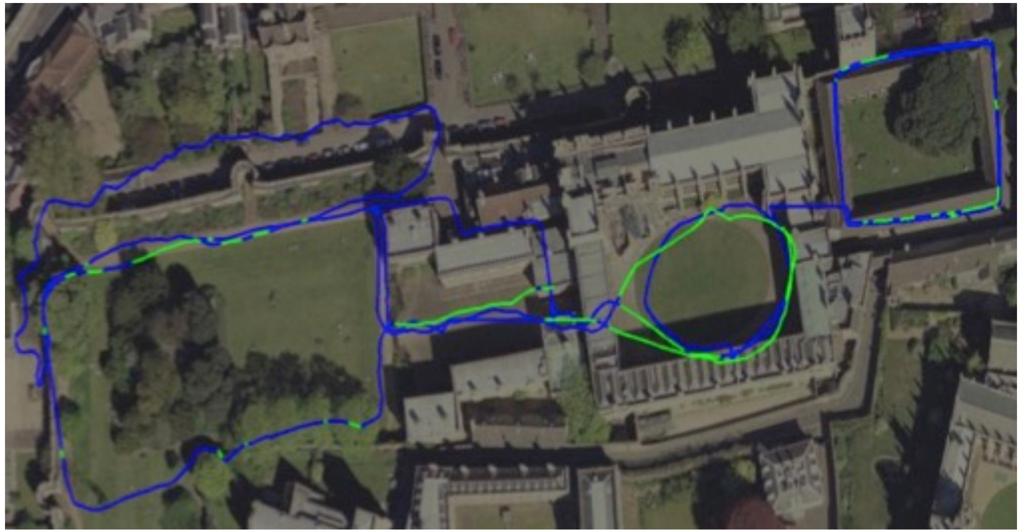
Vision-based Robotic Perception | the challenges

Challenge III: Place recognition

- Recognising when the robot visits a “known” location for:
  - Drift Correction
  - Trajectory / map merging

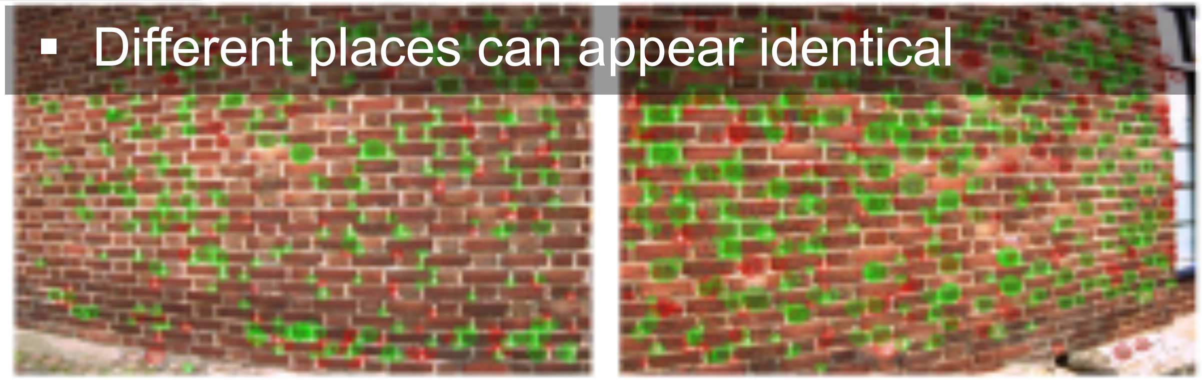






Vision-based Place recognition: common problems

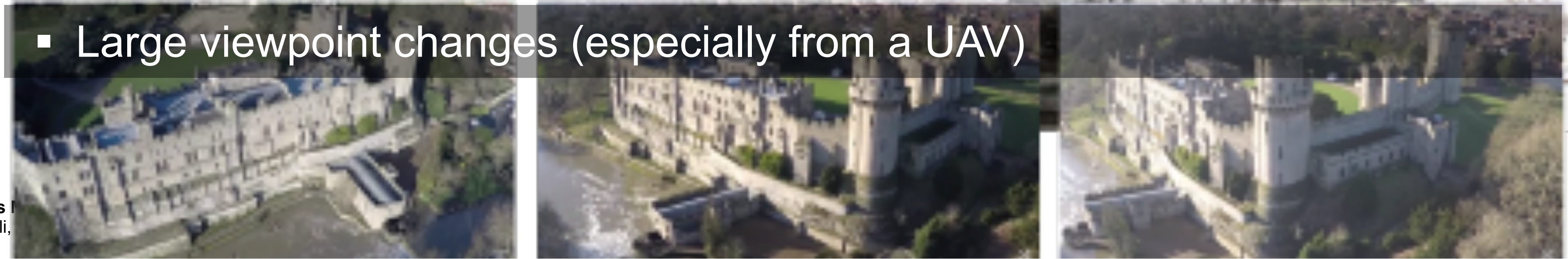
- Different places can appear identical



- Place appearance changes between visits

- Seasonal / Illumination changes

- Large viewpoint changes (especially from a UAV)



Towards lifelong place recognition

Dr Zetao Chen

Postdoctoral Research Fellow, V4RL, ETH Zurich

What is "Lifelong Place Recognition"?

*Lifelong Place Recognition is the process of identifying previously **visited locations over long time spans**, where the same location can undergo **dramatic condition** variations caused by illumination, seasons or weather.*

Lifelong Place Recognition | current systems

Current systems have come far, but are not there yet


Business

Google car is no match for snow and ice

Driverless vehicle can't yet detect winter road conditions, say experts who believe Google is decades away from a solution.

Tweet G+1 12 reddit this!

ERIC RISBERG/THE ASSOCIATED PRESS

A Google self-driving car: snow remains an issue.

Lifelong Place Recognition | spot the similarities

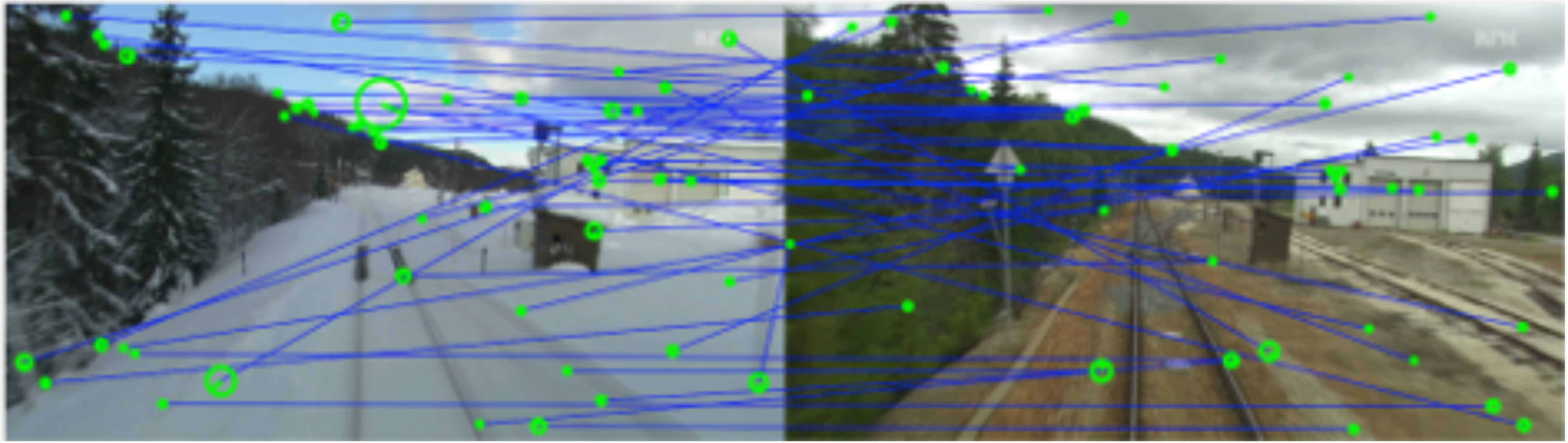
Is this the same place? Why?

Winter Image

Summer Image

Lifelong Place Recognition | spot the similarities

Extract SIFT features in each image & match them



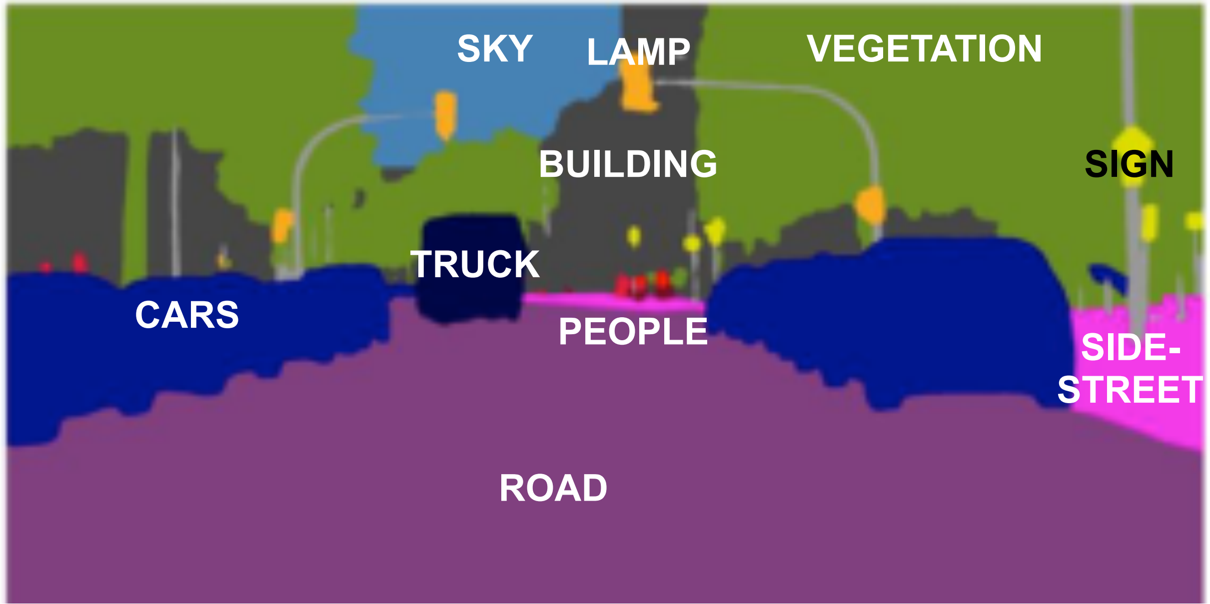
- Many false matches, because SIFT only looks at local patch gradients, which are not robust under strong condition variations.

Lifelong Place Recognition | spot the similarities

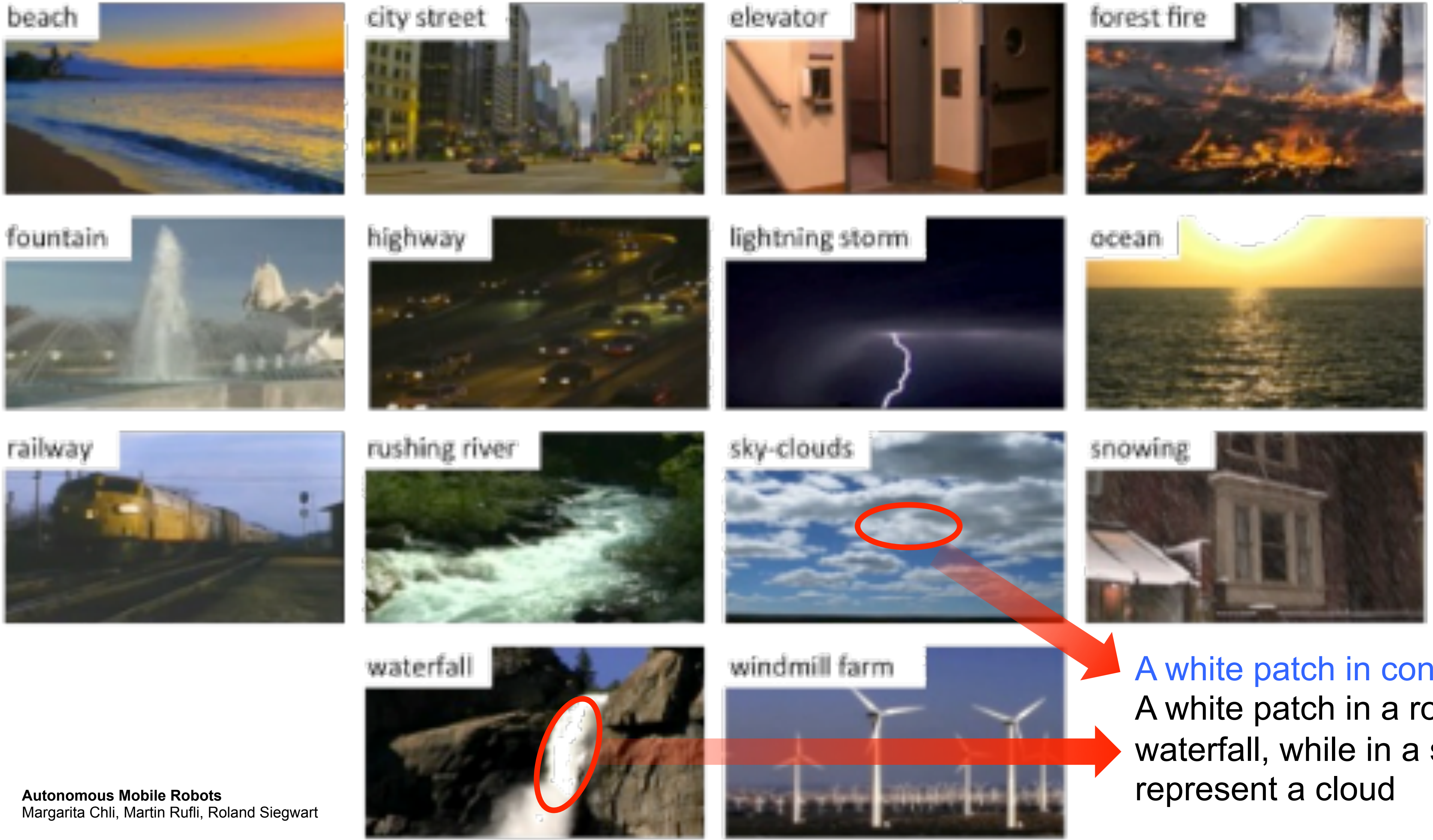
- We look for the underlying, basic scene structure, e.g. buildings, railways, vegetation
- \+ We instinctively predict Conditional Changes, e.g. green trees in the summer may turn white in winter

Lifelong Place Recognition | how to do it?

▪ In order to localize against strong condition variations, we need **high-level semantic context**, such as what the scene is about, for example, via image segmentation to **assign a class-label to each pixel** in the scene, etc.



Semantic Context I | scene type recognition



A white patch in context:

A white patch in a rocky scene can be a waterfall, while in a sky scene it can represent a cloud

Semantic Context II | scene segmentation

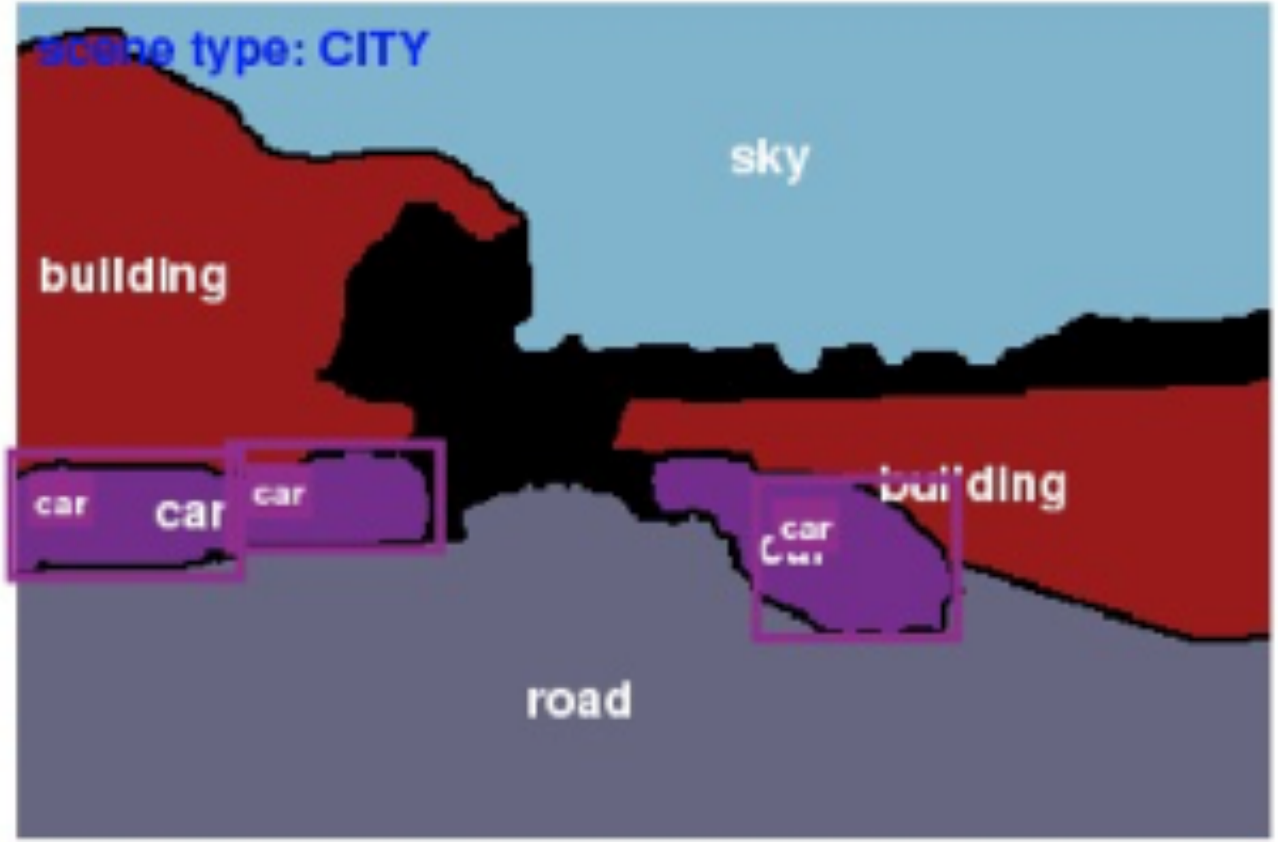


Image from [Yao et al., CVPR 2012]

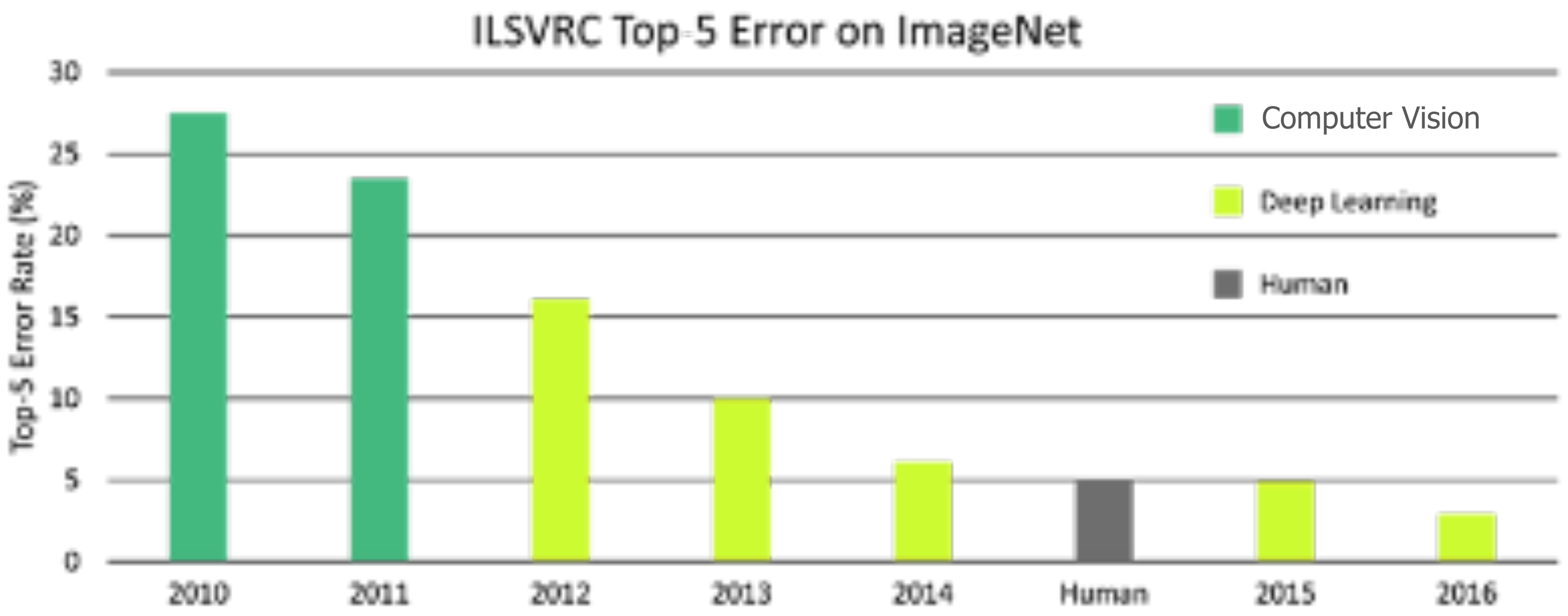
- Scene segmentation can be used to predict the class label for each pixel in the image

Obtaining Semantic Context | the deep learning approach

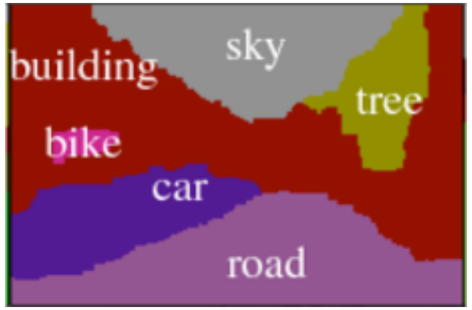
- Deep learning models have achieved state-of-the-art performance in various image semantic tasks, such as scene recognition, object detection, scene segmentation, etc.

- Use of deep learning enables **end-to-end training directly on the task**, without manual tuning on system parameters.

Deep Learning approaches have been **dominating the top scoring performances** in the ongoing "ImageNet" image recognition challenge over the last 4 years!



Obtaining Semantic Context | an illustrative example

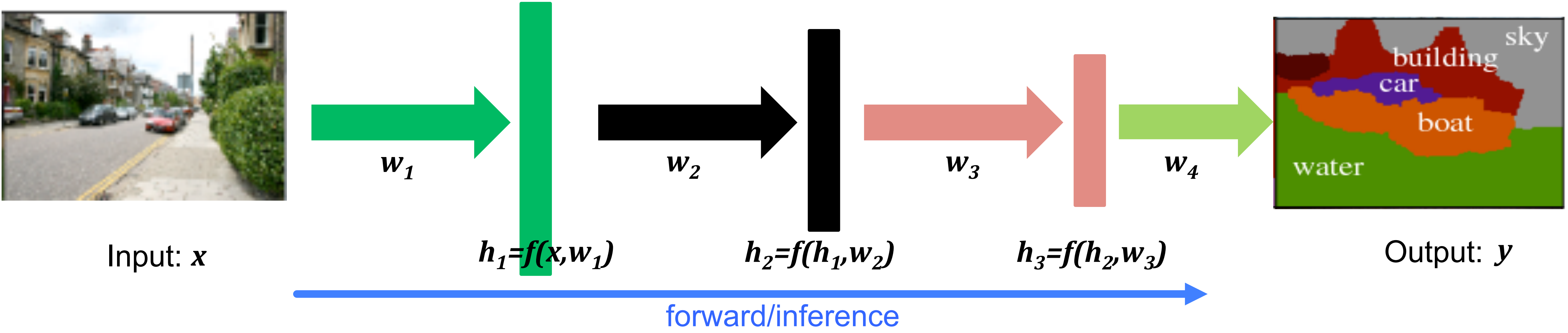


Input: $\boldsymbol{x}$

Output: $\boldsymbol{y}$

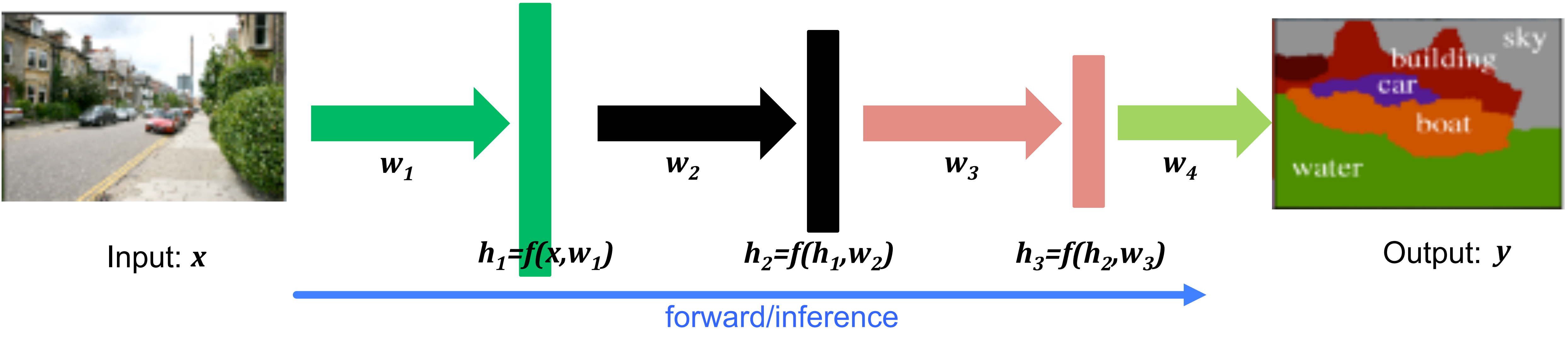
- We need a model in the middle, which takes the input image on the left and generates the semantic segmentation & labeling of each pixel in that image as shown on the right.

Obtaining Semantic Context | an illustrative example



- The 1st model layer, which is parameterized by w_1, takes x as input and outputs $h_1=f(x,w_1)$
- Th 2nd model layer, which is parameterized by w_2, takes h_1 as input and outputs $h_2=f(h_1, w_2)$
- The 3rd model layer, which is parameterized by w_3, takes h_2 as input and outputs $h_3=f(h_2, w_3)$
- The last model layer, which is parameterized by w_4, takes h_3 as input and outputs $y=f(h_3, w_4)$
- A forward inference stage completes!

Obtaining Semantic Context | an illustrative example

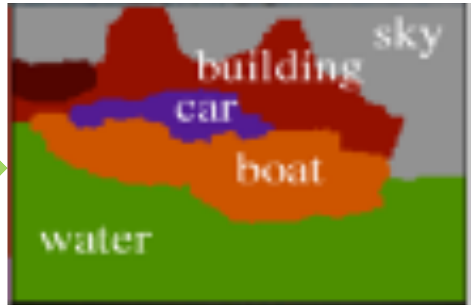


- The parameter set $W=\{w_1, w_2, w_3, w_4\}$ encodes the mapping from x to y. How does it learn that?

Obtaining Semantic Context | an illustrative example

Input: x

w_1 $\quad h_1=f(x,w_1)$ $\quad w_2$ $\quad h_2=f(h_1,w_2)$ $\quad w_3$ $\quad h_3=f(h_2,w_3)$ $\quad w_4$

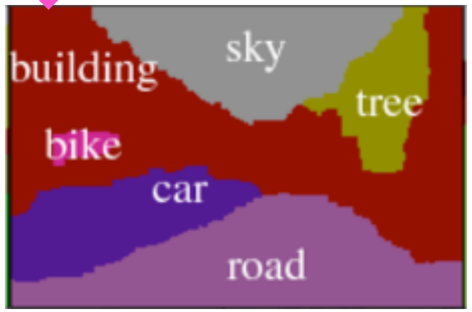


Output: y

\- difference

forward/inference

backward/learning



Ground truth

- The parameter set $W=\{w_1, w_2, w_3, w_4\}$ encodes the mapping from x to y. How does it learn that?
- Actually, at the 1st forward stage, the output may be quite different from the ground truth
- A backward learning stage can then back-propagate their difference and update W to minimize this difference gradually
- This process iterates until the difference between the output and the ground truth is smaller than a pre-defined threshold

Lifelong Place Recognition | ongoing work at V4RL

Construction of a condition-varying dataset to train a deep learning network

- We gather images captured from **static cameras around the world**
- Each camera observes **the same scene constantly and over several years**
- 2500 cameras selected at the locations above(red dots)

Lifelong Place Recognition | ongoing work at V4RL

Dataset examples: diversity of scenes

Lifelong Place Recognition | ongoing work at V4RL

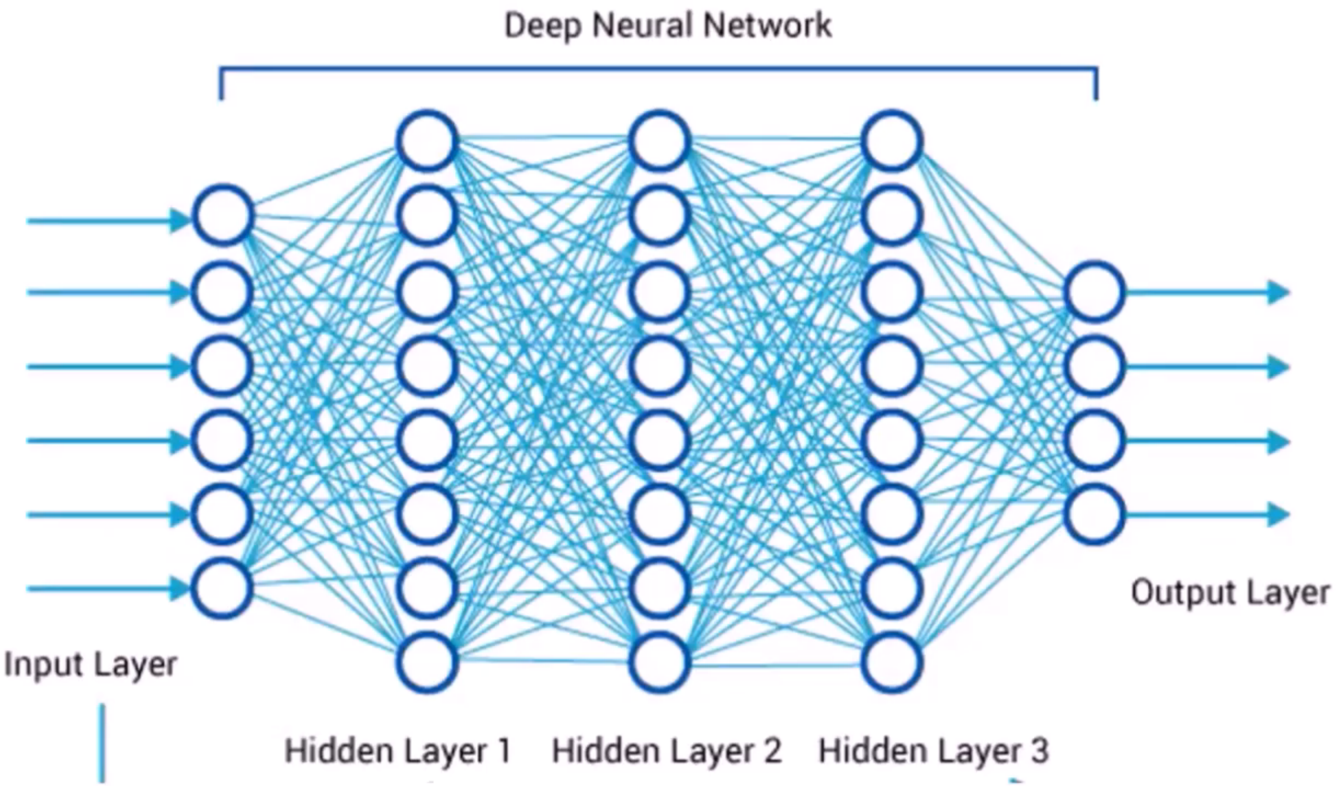
VISION FOR ROBOTICS LAB

Large condition variations in each scene

Lifelong Place Recognition | ongoing work at V4RL

Network Training

Feeds from different cameras:



Lifelong Place Recognition | conclusion

- Currently, the use of deep learning-based approaches onboard a UAV is unrealistic due to their:
  - High computational cost – cannot run in real-time on a typical UAV processor
  - High power consumption
  - Need for bigger onboard memory to host most existing deep learning models

- Open research questions:
  - How can we compress deep learning models?
  - Could we reuse image features that are typically extracted onboard UAVs in combination with deep learning approaches e.g. via the help of a ground station?

Vision-based Robotic Perception | the challenges

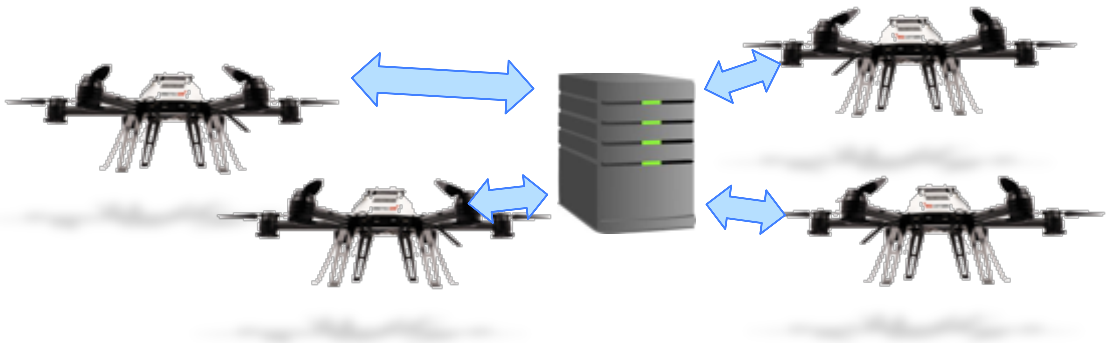
Challenge IV: Collaborative robot sensing & mapping

- Exploit presence of multiple UAVs (occlusions, accuracy, time efficiency)

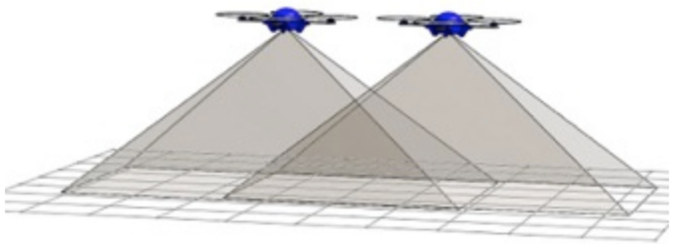
Vision-based Robotic Perception | the challenges

Challenge IV: Collaborative robot sensing & mapping

- Exploit presence of multiple UAVs (occlusions, accuracy, time efficiency)



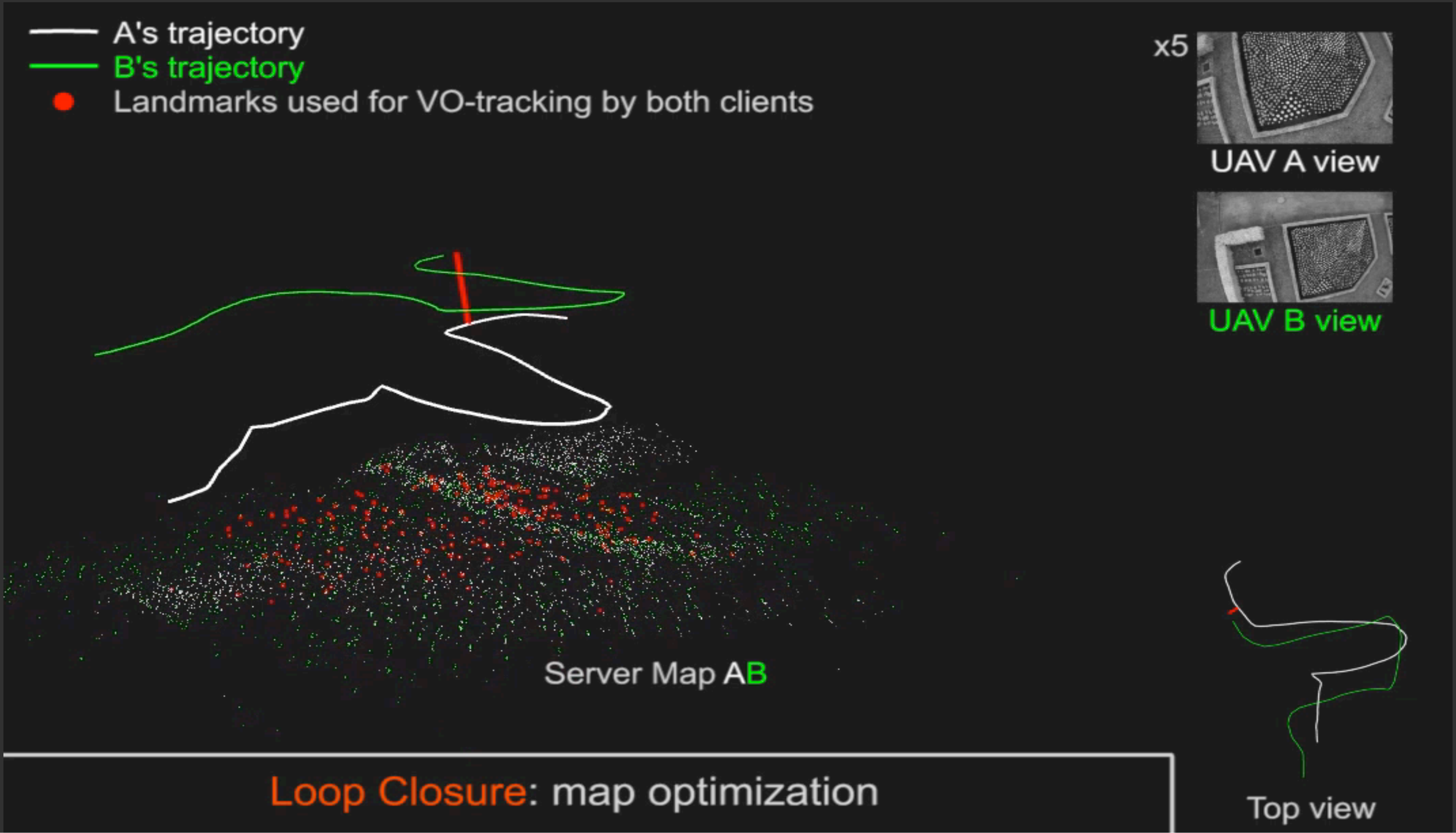
- Flight-critical tasks on client
- Computationally expensive tasks on server
- What information needs to be shared?



Variable-baseline stereo from 2 UAVs
[Achtelik et al, IROS 2011]

Work with Patrik Schmuck
& Marco Karrer

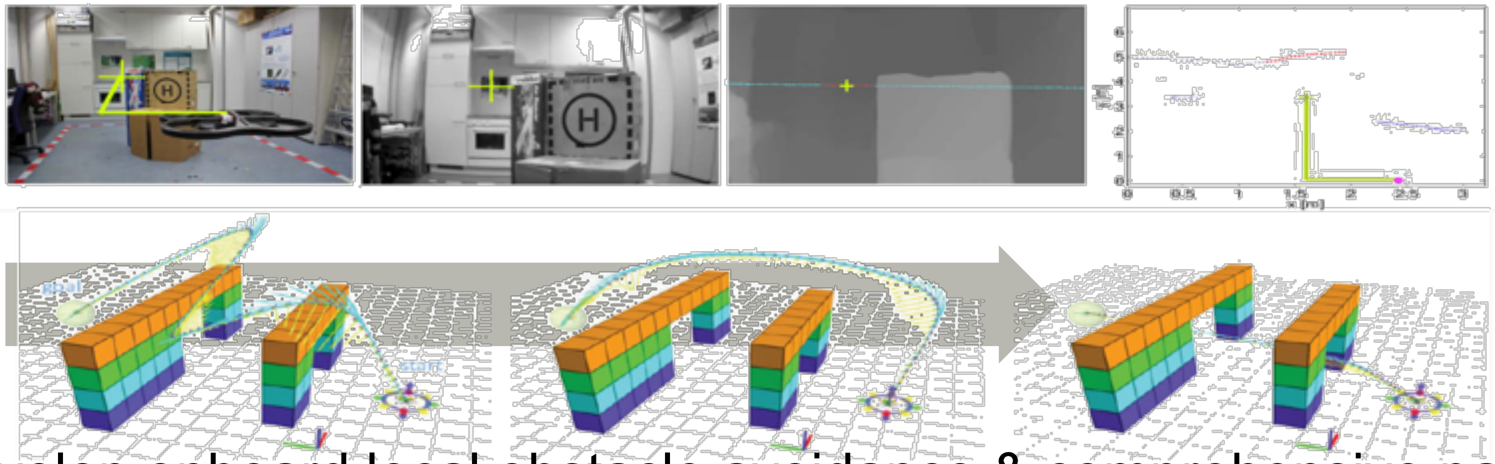
[Schmuck & Chli, ICRA 2017]



Vision-based Robotic Perception | the challenges

Challenge V: Navigation Strategies – obstacle avoidance & path planning

- Complete the navigation loop
- Existing: mostly off-board solutions



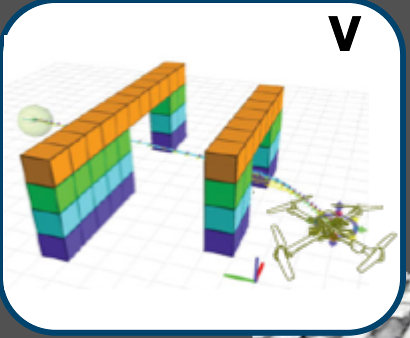
[Alvarez et al, ISER 2014]

Collision avoidance with a camera and offboard GPU processing

[Achtelik et al, JFR 2014]

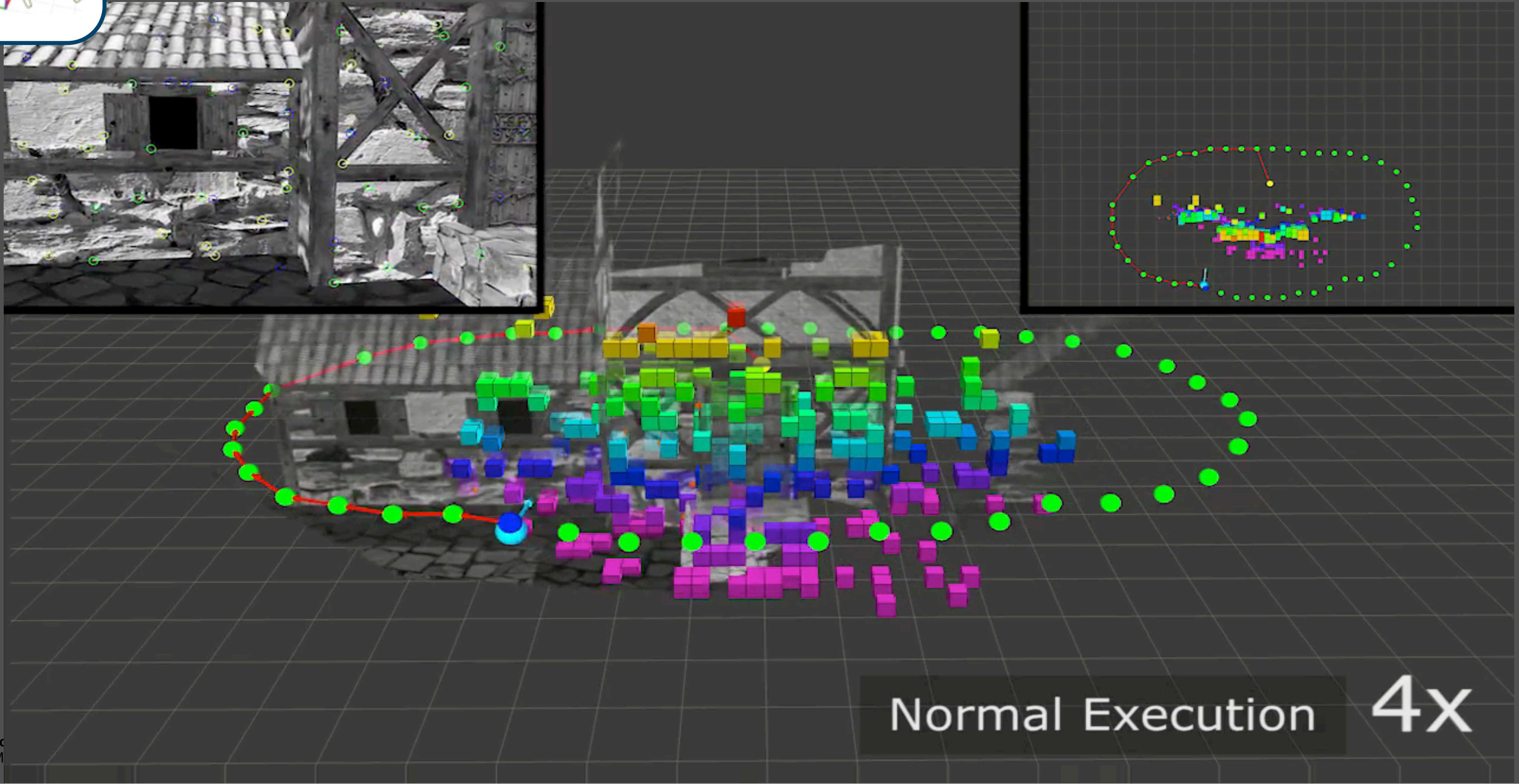
Intermediate & final paths computed in simulation

- Develop onboard local obstacle avoidance & comprehensive path planning



UAV path planning with VI-SLAM in the loop

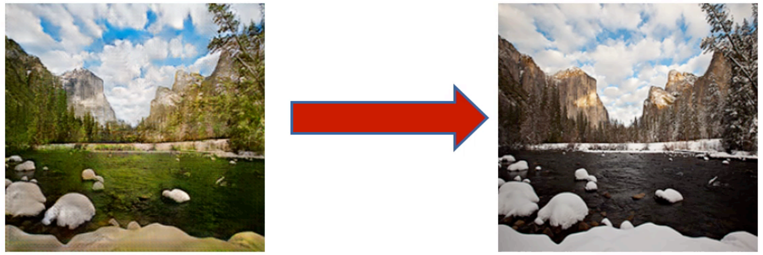
[Alzugaray et al., ICRA 2017]



Master/Semester Projects @ V4RL

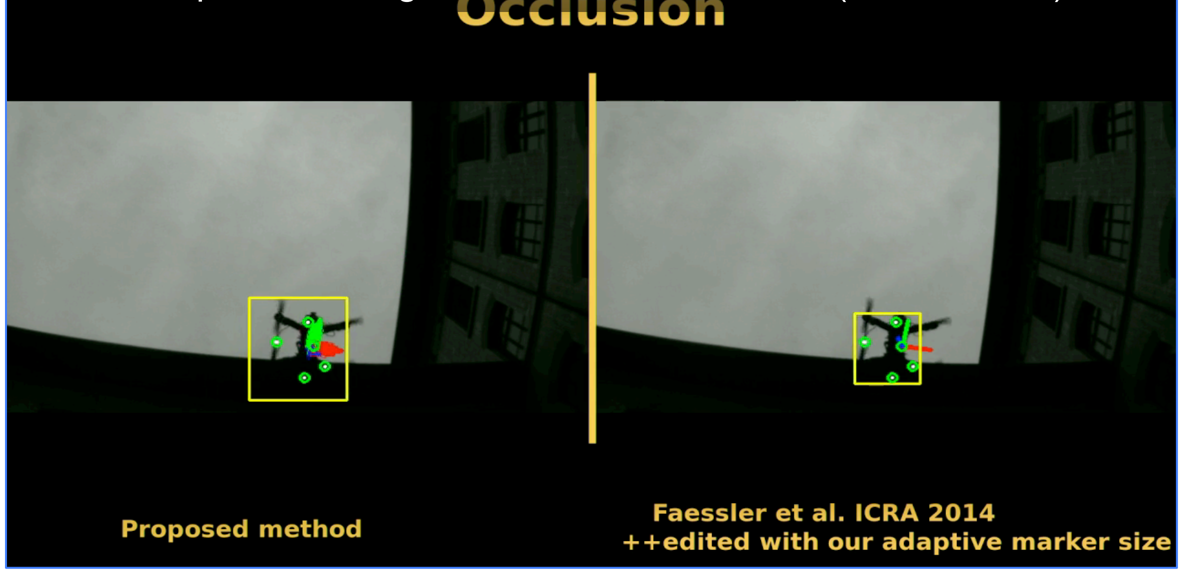
Long-Term Place Recognition with Generative Adversarial Nets (Tianshu Hu)

Current results

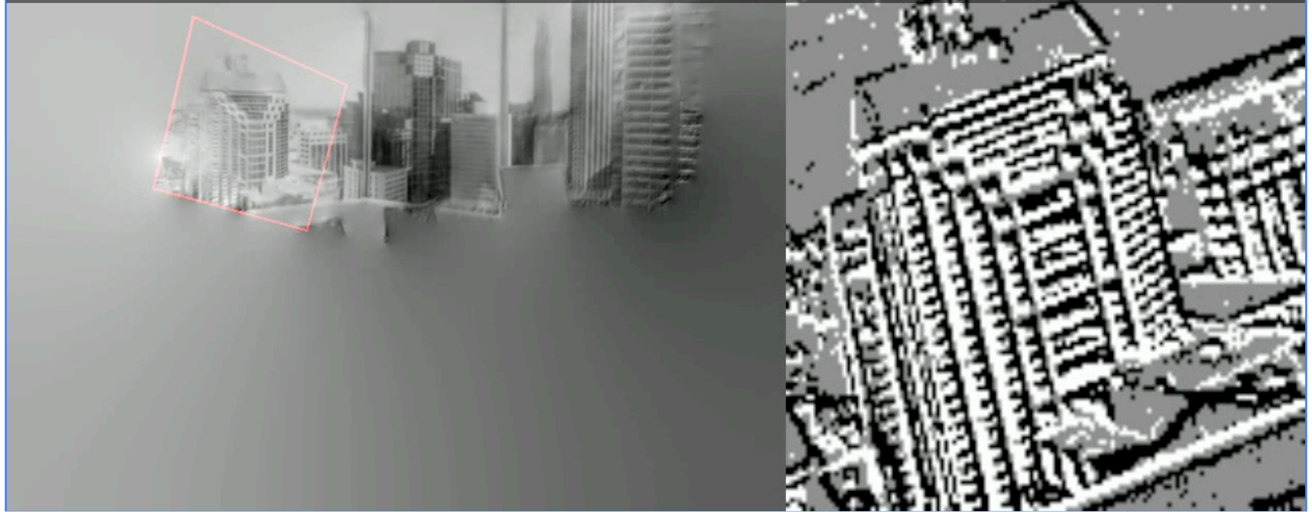


Summer Winter

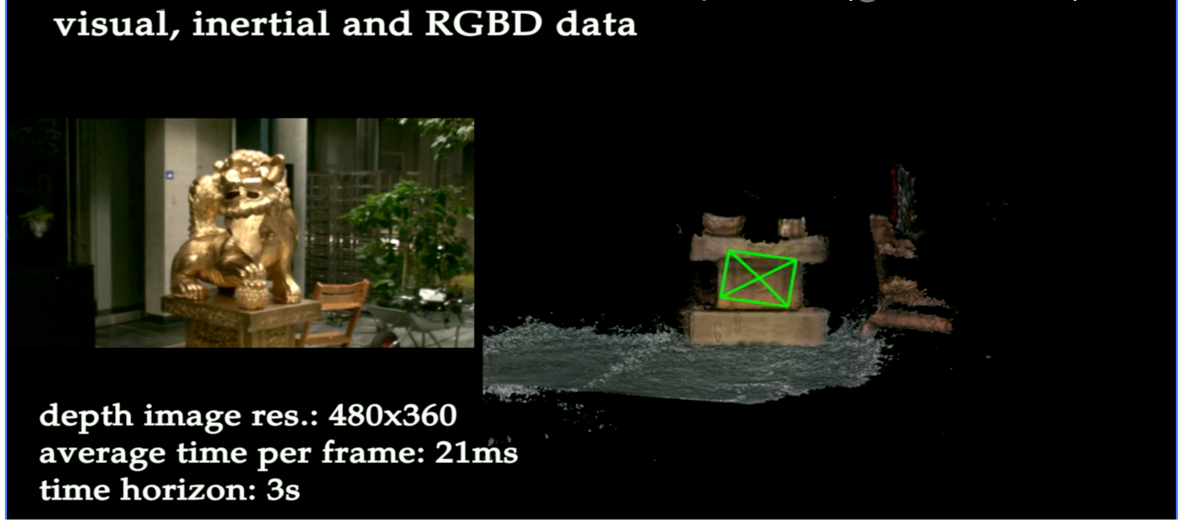
Real-time pose tracking with an external camera (Marco Moos)



Scene Reconstruction from a DVS camera (Wilko Schwarting)



Dense 3D Reconstruction for Aerial Manipulation (Marco Karrer)



Conclusion & Impact

Vision-based SLAM:

- has come a long way: from handheld to vision-stabilised flights of UAVs
- key to spatial awareness of robots ⇨ bridges the gap between Computer Vision and Robotics

Perception + Collaboration are central to Robotics today:

  - Large sums of research funds in the area (e.g. SHERPA €11M, ICARUS €17.5M)

- Still work to be done before robots are ready for real missions
- Potential for great impact in the way we perceive/employ robots today