ROBOT INTELLIGENCE LAB

Simultaneous Localization And Mapping

Vision-based SLAM

Sungjoon Choi, Korea University

kakaobrain

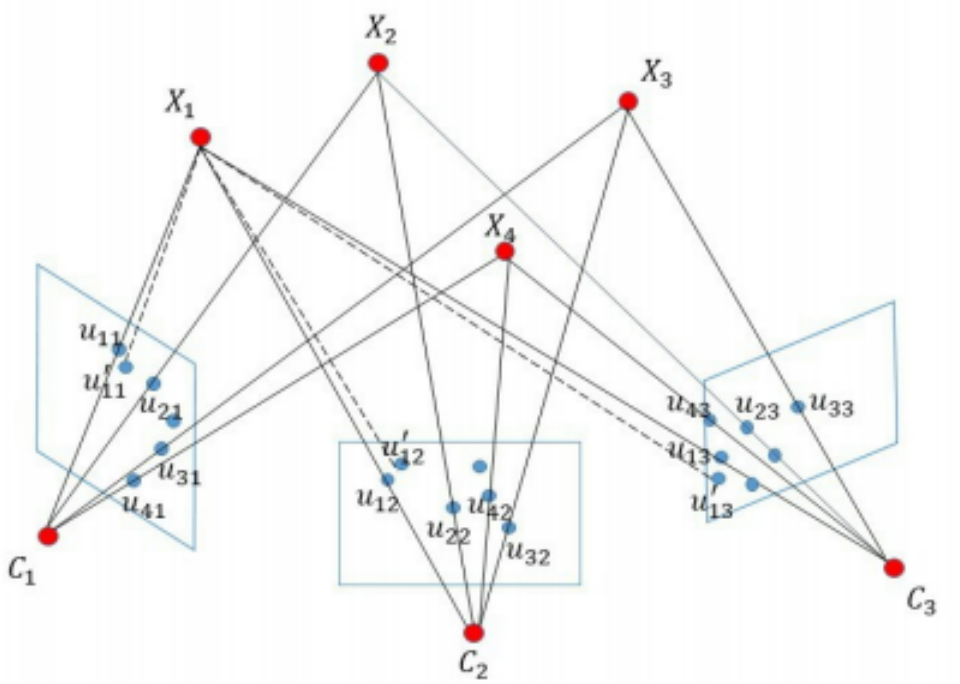
ROBOT INTELLIGENCE LAB

Multi-View Geometry

Bundle Adjustment

ROBOT INTELLIGENCE LAB

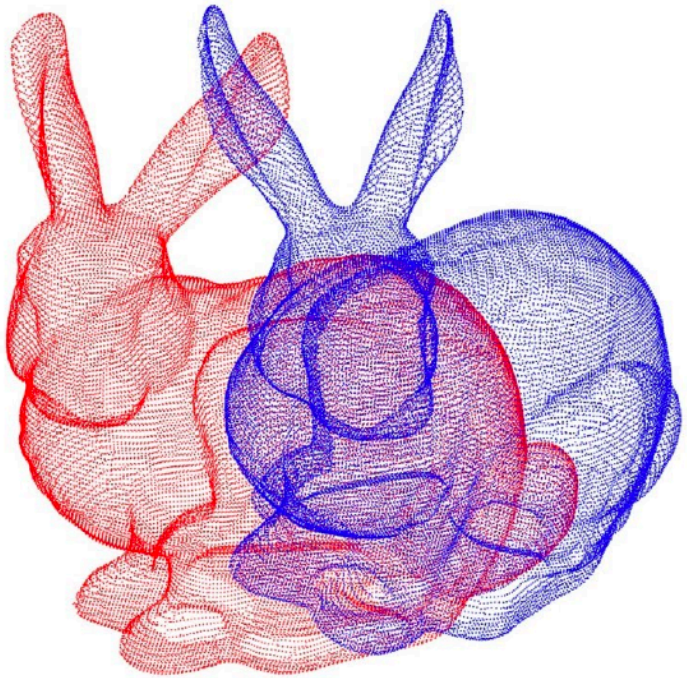
- Bundle Adjustment (BA)
  - BA estimates the 3D feature points and the position of the camera in $SE(3)$.
  - $SE(3)$: special Euclidean group consists of (x, y, z) and (r, p, y).
  - It minimizes the re-projection error using least squares.



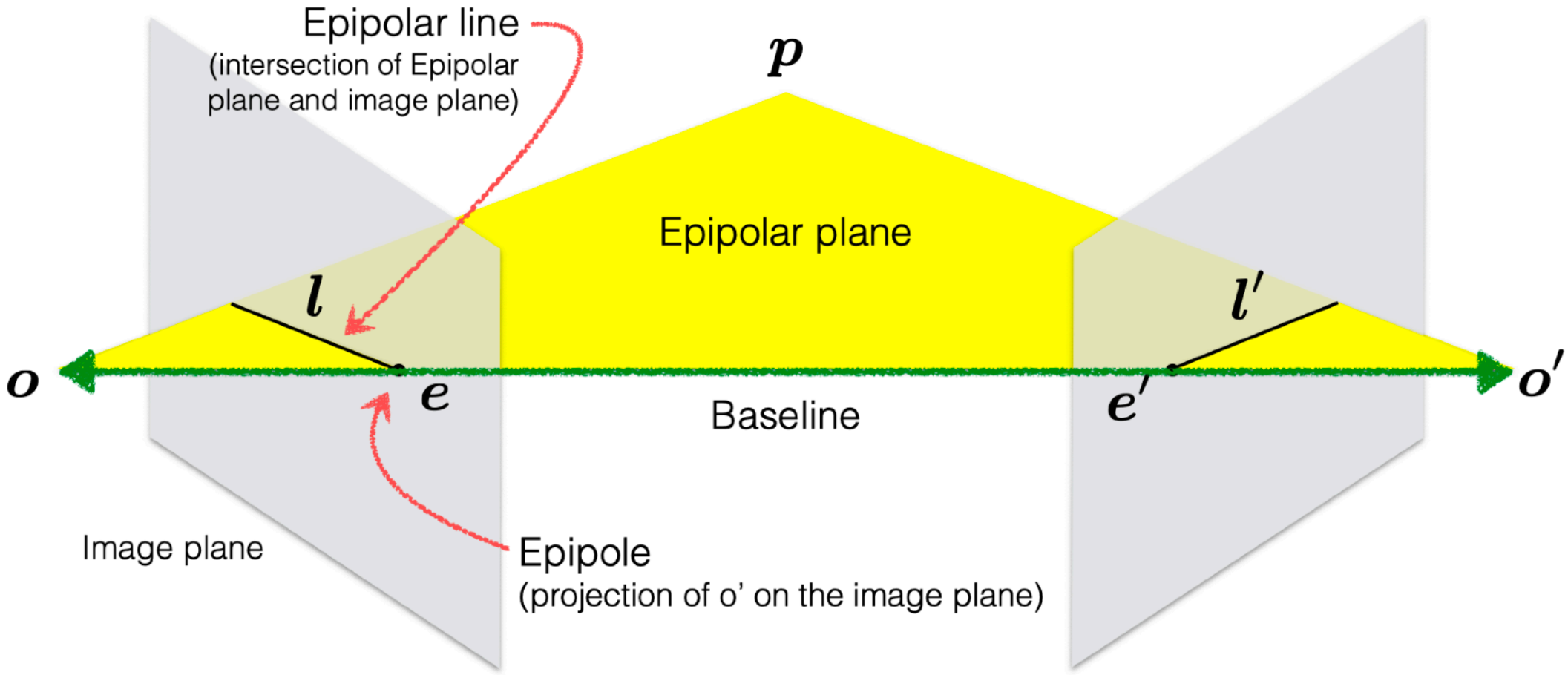
Iterative Closest Point

ROBOT INTELLIGENCE LAB

- Iterative Closest Point (ICP)
  - ICP minimizes the difference between two point clouds by estimating the translation information between two point sets.



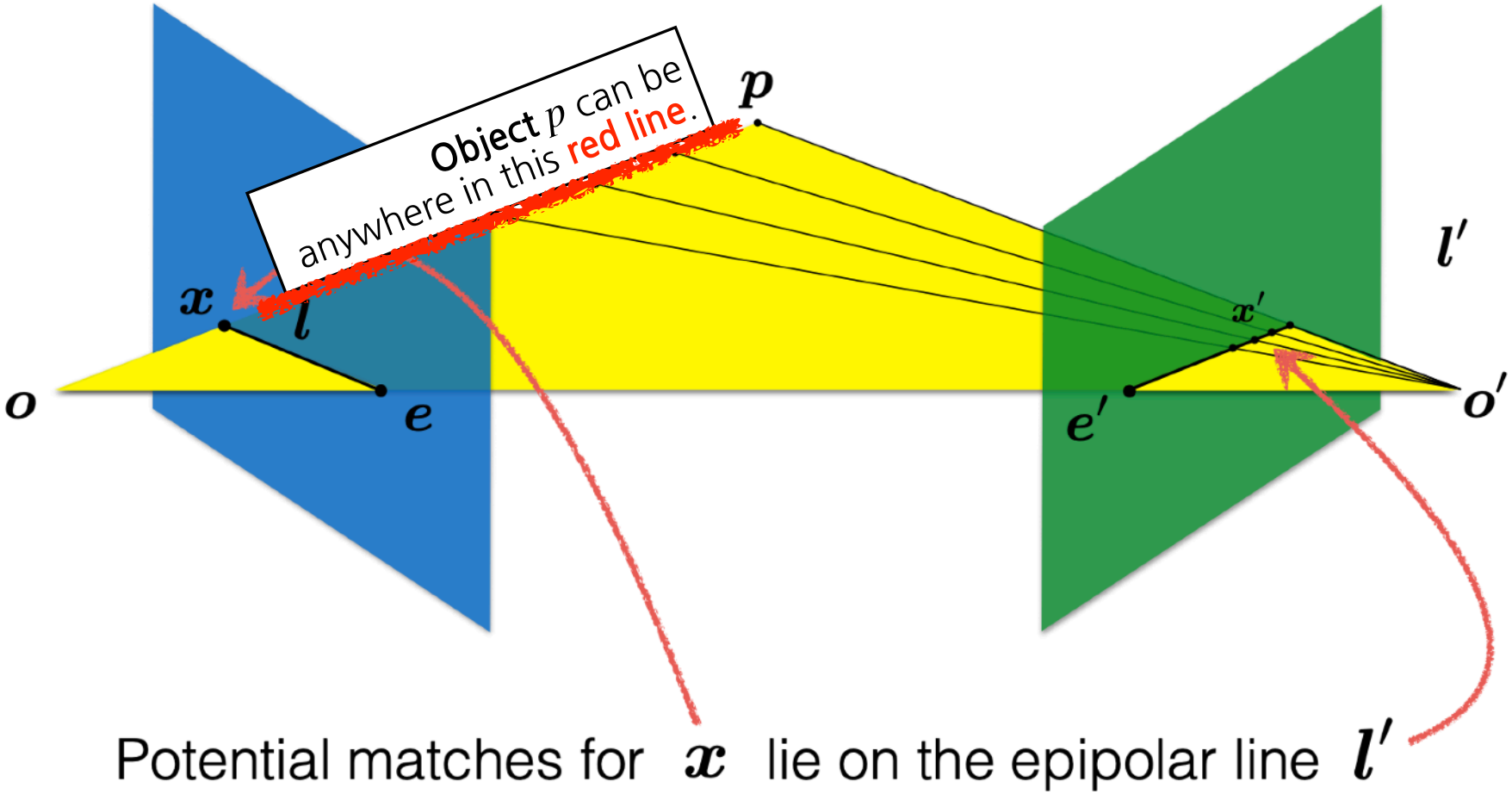
Epipolar Geometry



Essential Matrix

ROBOT INTELLIGENCE LAB

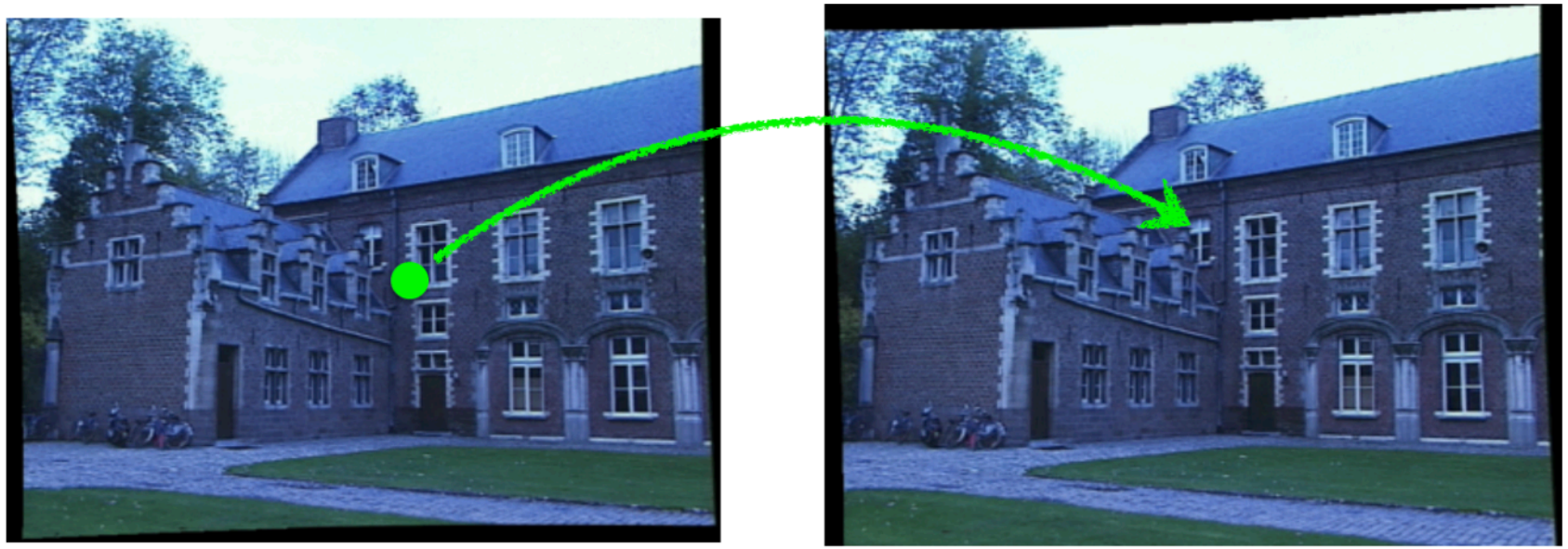
- Epipolar constraint



Essential Matrix

ROBOT INTELLIGENCE LAB

- **Task**: Match a point in the left image to the point in the right image.



Left image

Right image

- How would you do it?
- Epipolar constraints (line) reduces the search to a single line!
- How do you compute the epipolar line?

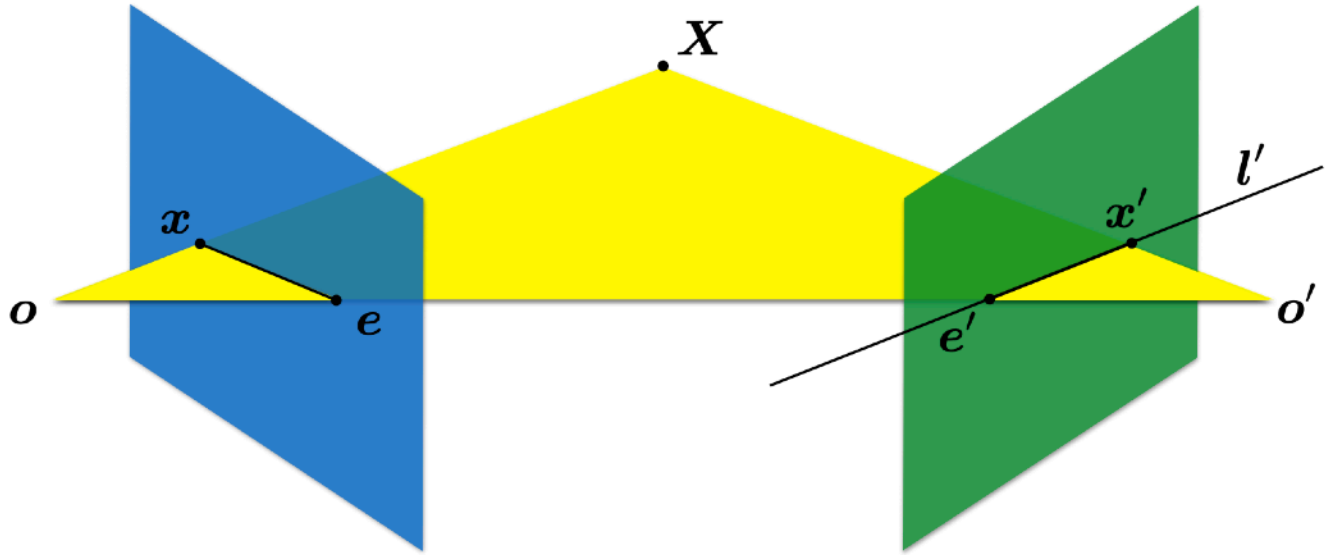
Essential Matrix

ROBOT INTELLIGENCE LAB

- Essential matrix E is a 3×3 matrix that encodes epipolar geometry.

Given a point in one image,
multiplying by the **essential matrix** will tell us
the **epipolar line** in the second view.

$$\mathbf{E}x = l'$$



Essential Matrix

ROBOT INTELLIGENCE LAB

Longuet-Higgins equation $\boldsymbol{x}'^{\top}\mathbf{E}\boldsymbol{x} = 0$

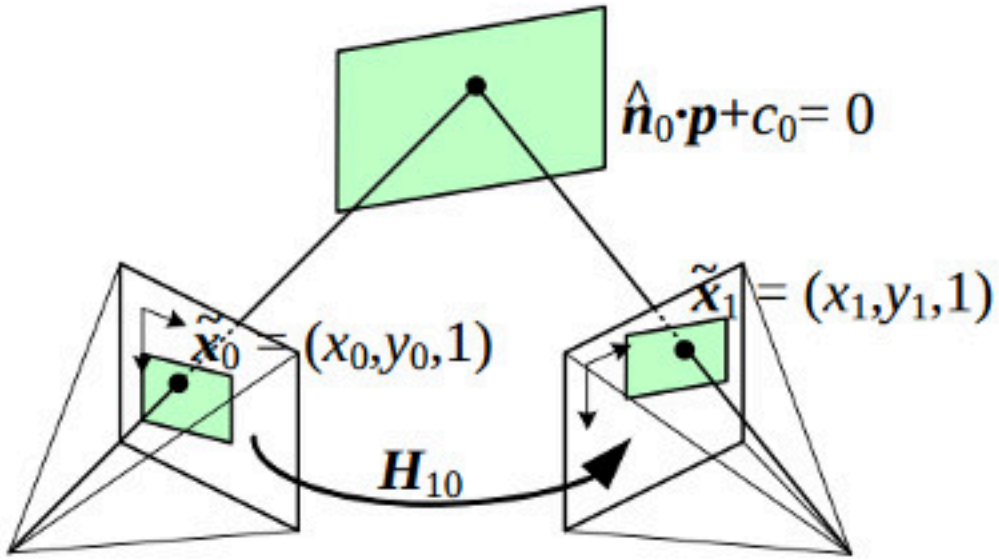
Epipolar lines $\boldsymbol{x}^{\top}\boldsymbol{l} = 0$ $\quad \boldsymbol{x}'^{\top}\boldsymbol{l}' = 0$

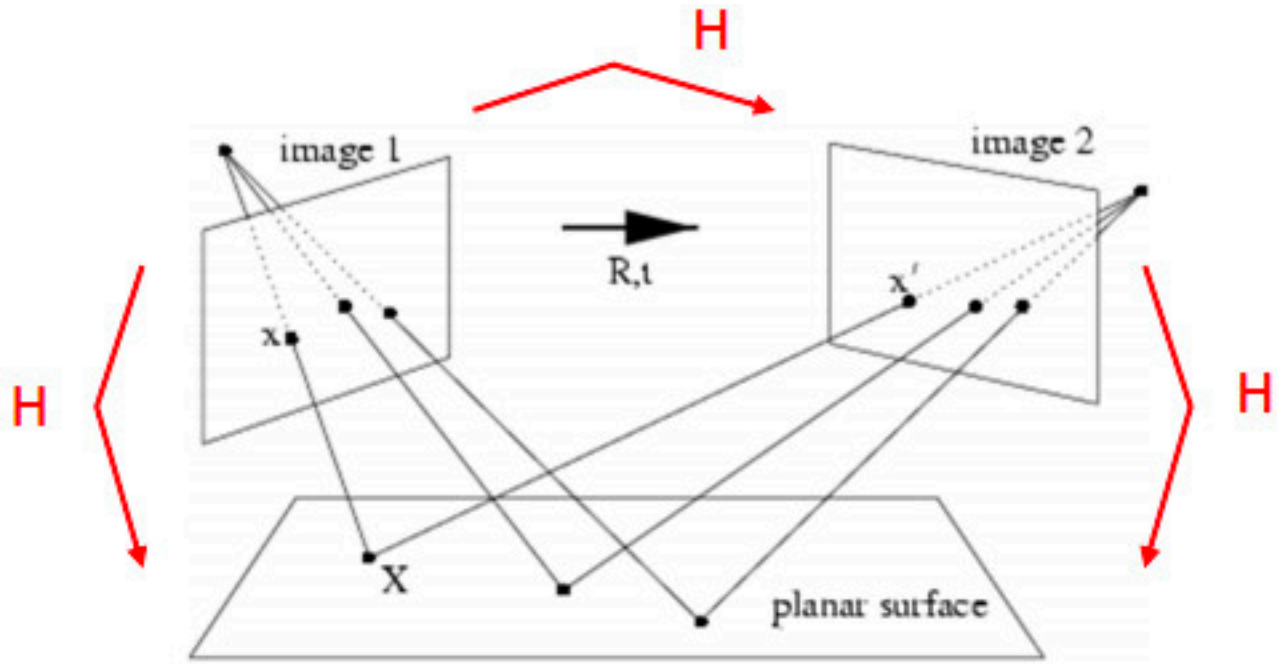
$\boldsymbol{l}' = \mathbf{E}\boldsymbol{x}$ $\quad \boldsymbol{l} = \mathbf{E}^{T}\boldsymbol{x}'$

Homography Matrix

ROBOT INTELLIGENCE LAB

- The homography matrix is a 3×3 matrix which relates the transformation between two planes (up to a scale factor).

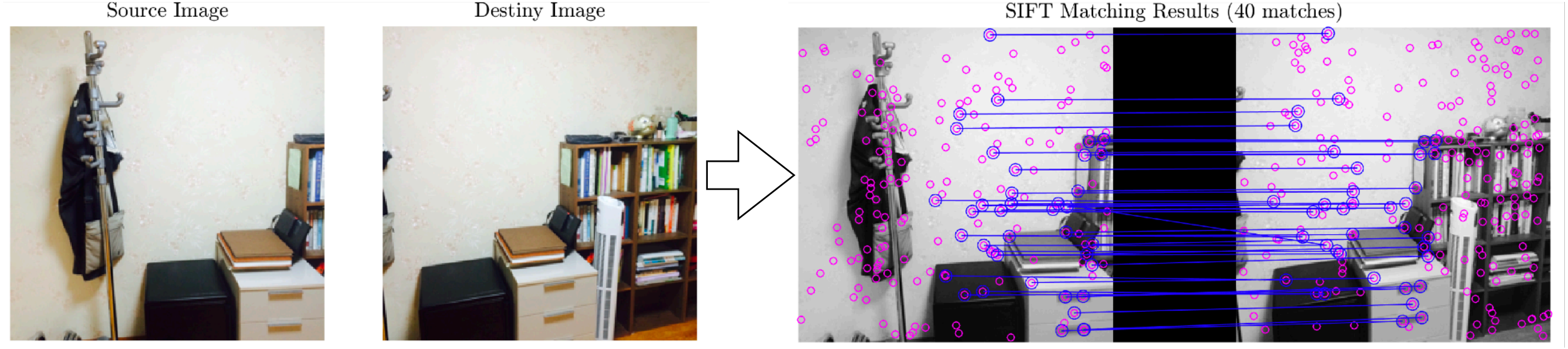




Homography Matrix

ROBOT INTELLIGENCE LAB

- Given two images, a feature matching is performed (e.g., SIFT matching).



Homography Matrix

ROBOT INTELLIGENCE LAB

- Then, we compute the homography matrix using RANSAC.

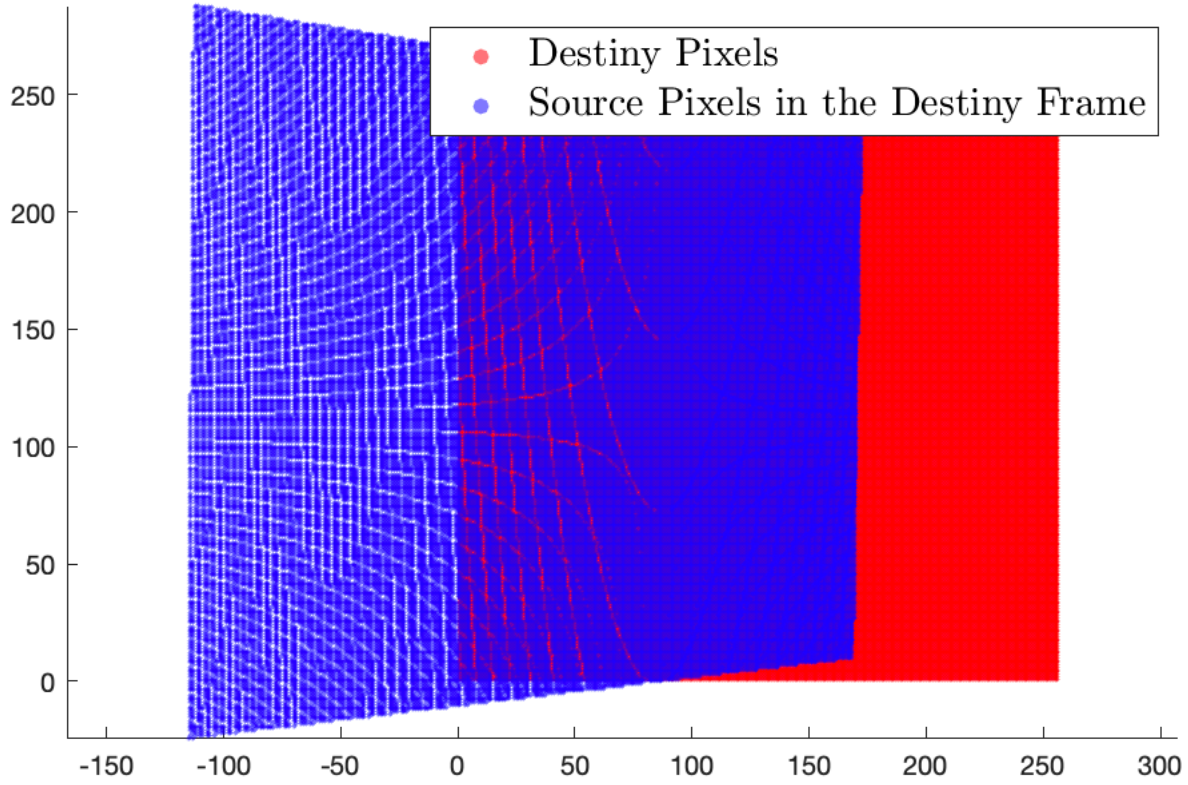
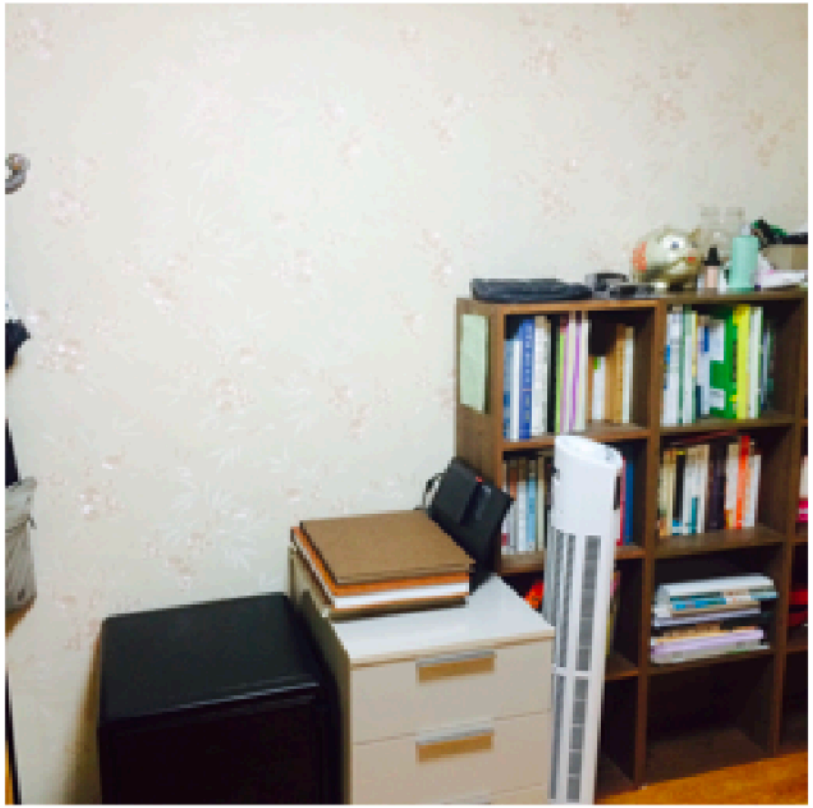


Image Stitching

ROBOT INTELLIGENCE LAB

Source Image

Destiny Image



Mosaiced Image

ROBOT INTELLIGENCE LAB

MonoSLAM

"MonoSLAM: Real-Time Single Camera SLAM," 2007 (Oxford, ICL, AIST)

MonoSLAM

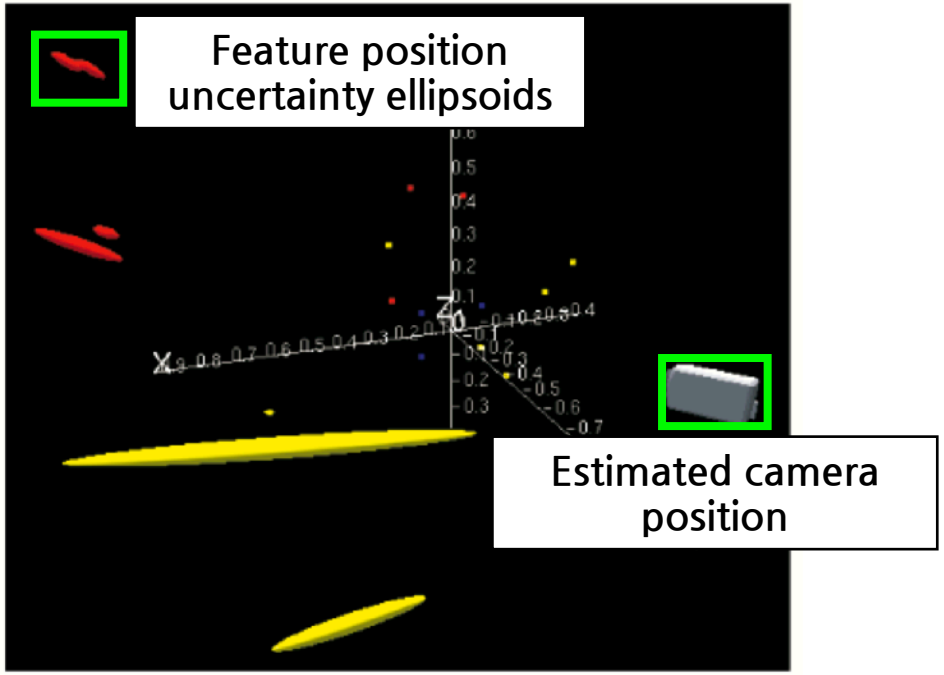
ROBOT INTELLIGENCE LAB

- "MonoSLAM: Real-Time Single Camera SLAM," 2007 (Oxford, ICL, AIST)
  - The first successful application of SLAM from mobile robotics to the "pure vision" domain of a single uncontrolled camera, achieving real time but drift-free performance.
  - The core is the online creation of a sparse but persistent map of **natural landmarks** within a probabilisitic framework.
  - It combines visual feature template matching and EKF.

MonoSLAM

ROBOT INTELLIGENCE LAB

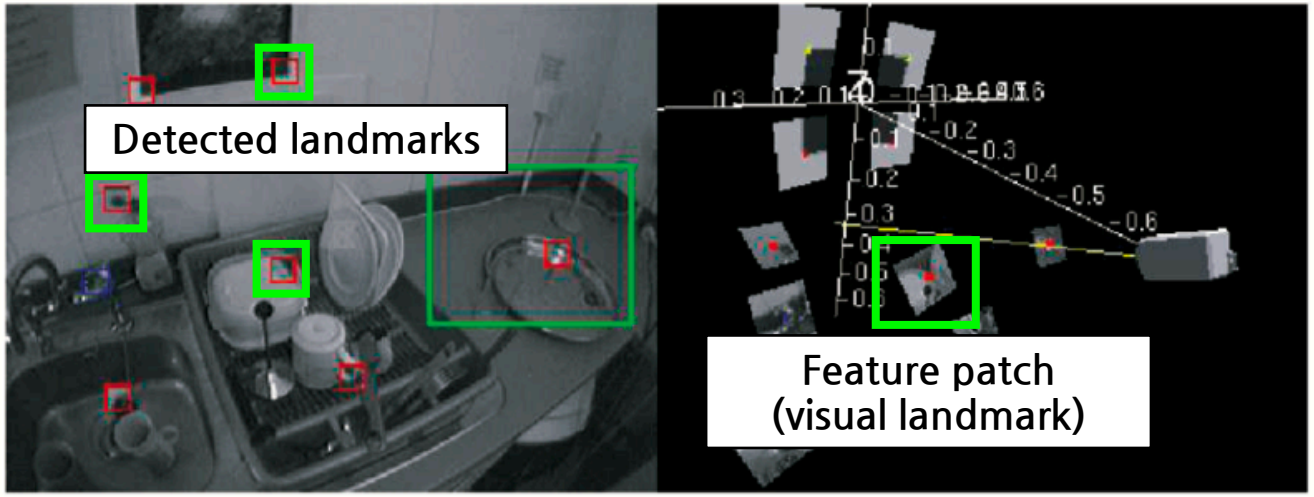
- Probabilistic 3D Map
  - MonoSLAM incorporates a probabilistic **feature-based** map, representing a snapshot of the current estimates of the state of the camera and the uncertainty in these estimates.
  - The map is updated by the extended Kalman filter (EKF).



MonoSLAM

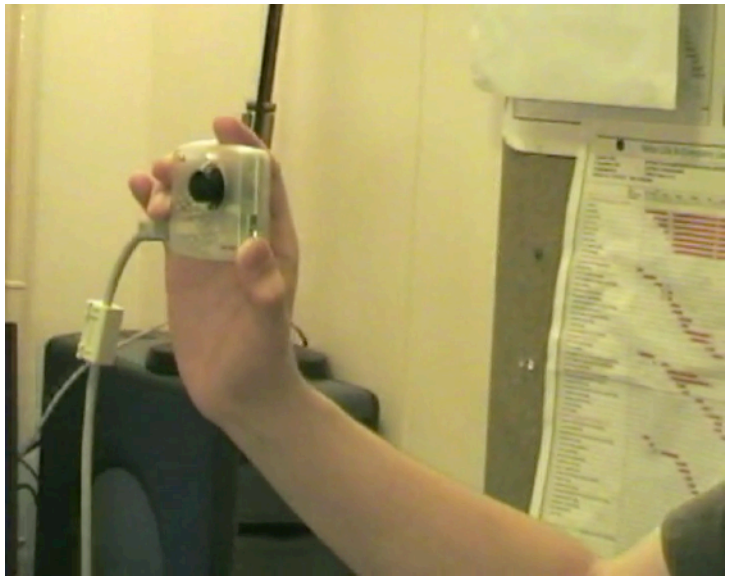
ROBOT INTELLIGENCE LAB

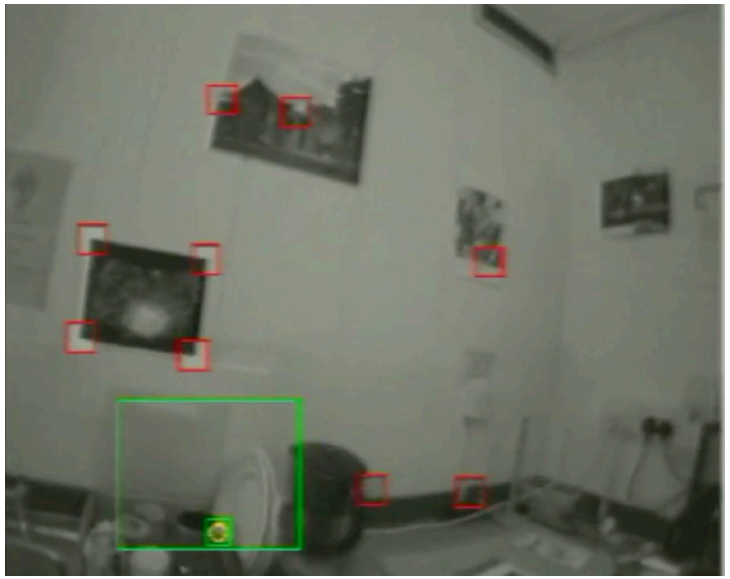
- Natural Visual Landmarks
  - The role of the map is to permit real-time localization rather than serve as a complete scene description, and therefore aim to capture a sparse set of high-quality **landmarks**.
  - Relatively large (11x11 pixels) image patches work as **long-term landmark** features.

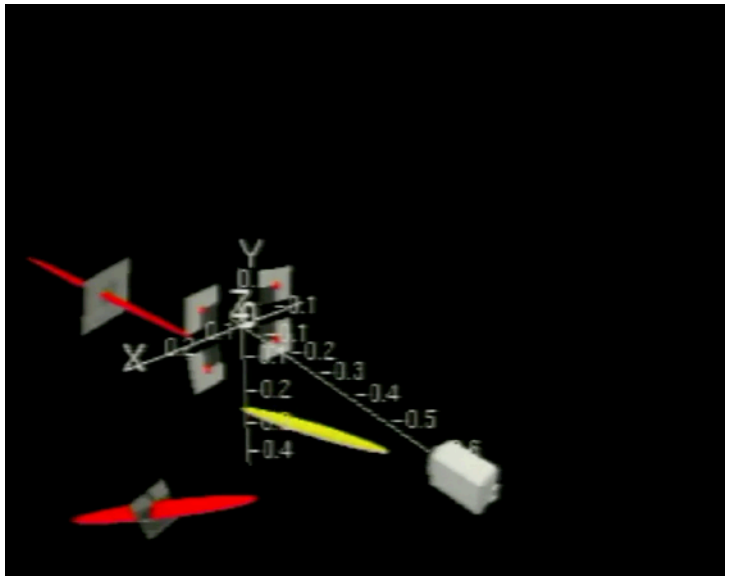


MonoSLAM

ROBOT INTELLIGENCE LAB







https://youtu.be/mimAWVm-0qA

PTAM

"Parallel Tracking and Mapping for Small AR Workspaces," 2007 (Oxford)

PTAM

ROBOT INTELLIGENCE LAB

- "Parallel Tracking and Mapping for Small AR Workspaces," 2007 (Oxford)
  - It is specifically designed to track a hand-held camera in a small AR workspace named PTAM (parallel tracking and mapping).
  - The core is to **split tracking and mapping** into two separate tasks which allows the use of (computationally-expensive) batch optimization for mapping.
    - No more SLAM wiggle (i.e., no additional smoothing for incremental update)
  - But, it has a clear limitation in terms of general performances.

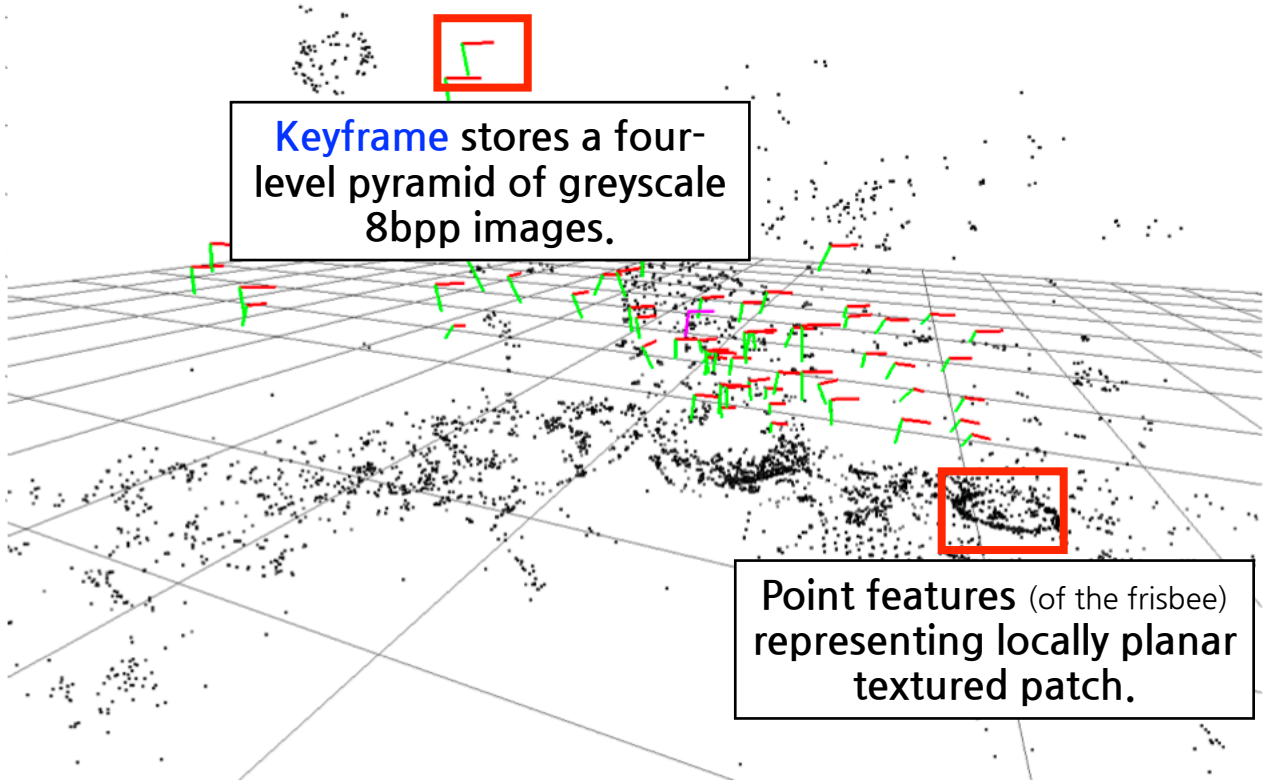
PTAM

ROBOT INTELLIGENCE LAB

- Overview of PTAM
  - Tracking and Mapping are separated.
    - Tracking (of camera positions) is no longer probabilistically slaved to the map-making procedure.
  - Mapping is based on **keyframes**, which are processed using Bundle Adjustments.
  - The map is densely initialized from a stereo pair (5-point algorithm).
  - New points are initialized with an epipolar search.

PTAM

ROBOT INTELLIGENCE LAB



PTAM

https://youtu.be/Y9HMn6bd-v8

ROBOT INTELLIGENCE LAB

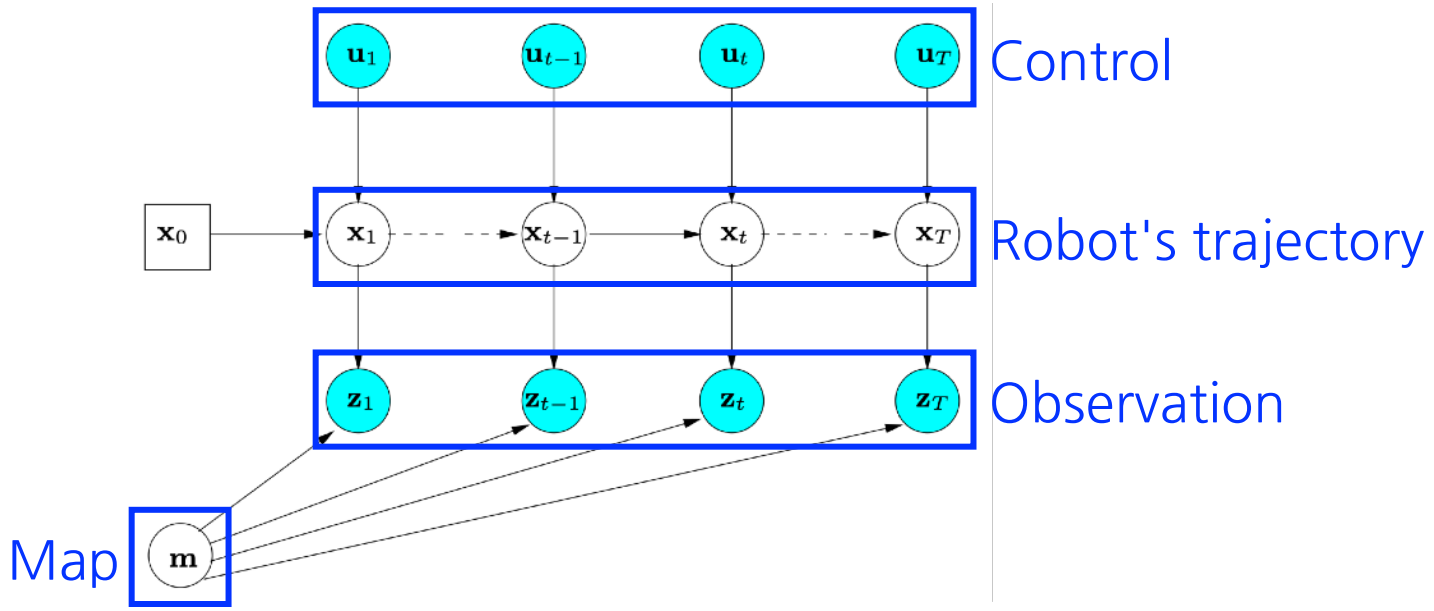
Graph-based SLAM

"A Tutorial on Graph-Based SLAM," 2010

Graph-based SLAM

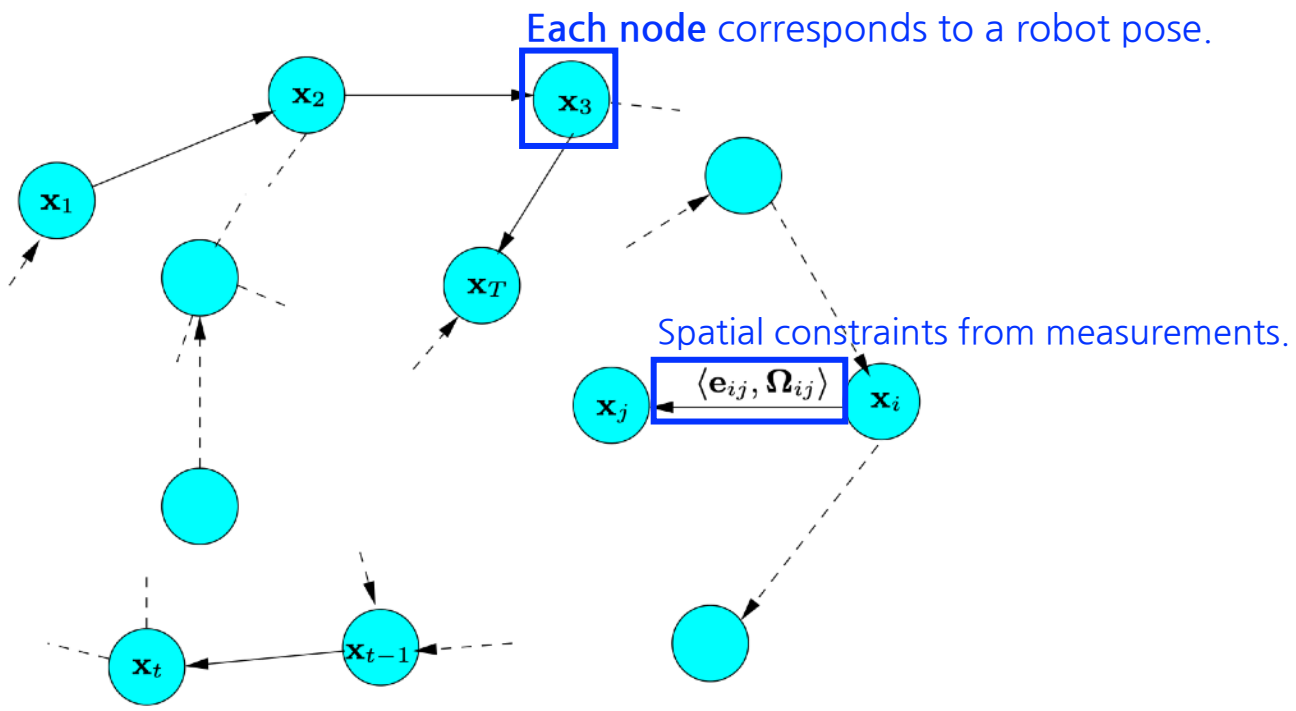
ROBOT INTELLIGENCE LAB

- "A Tutorial on Graph-Based SLAM," 2010
  - It uses a graph whose nodes correspond to the poses of the robot at different points in time and whose edges represent constraints between the poses (obtained from observations and control inputs).

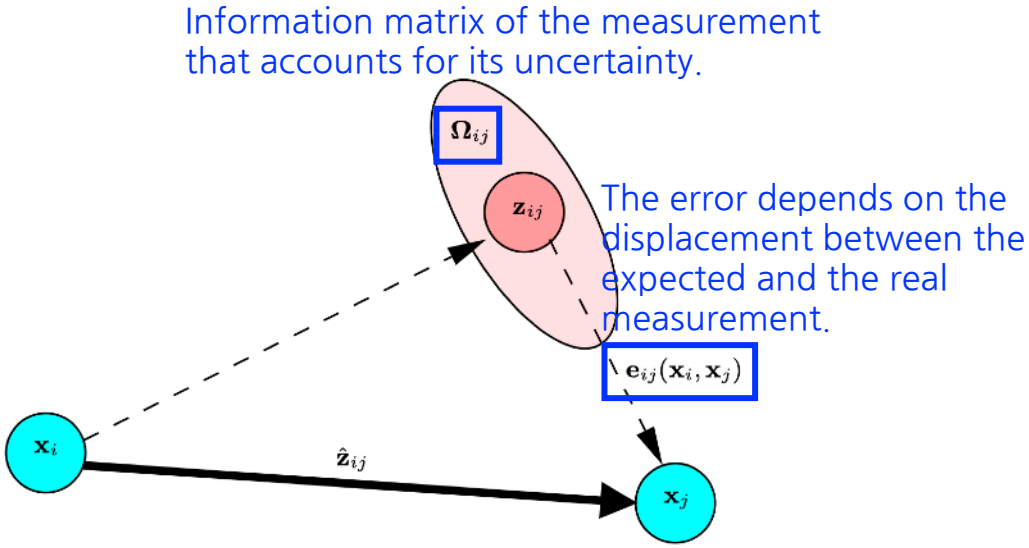


Graph-based SLAM

ROBOT INTELLIGENCE LAB



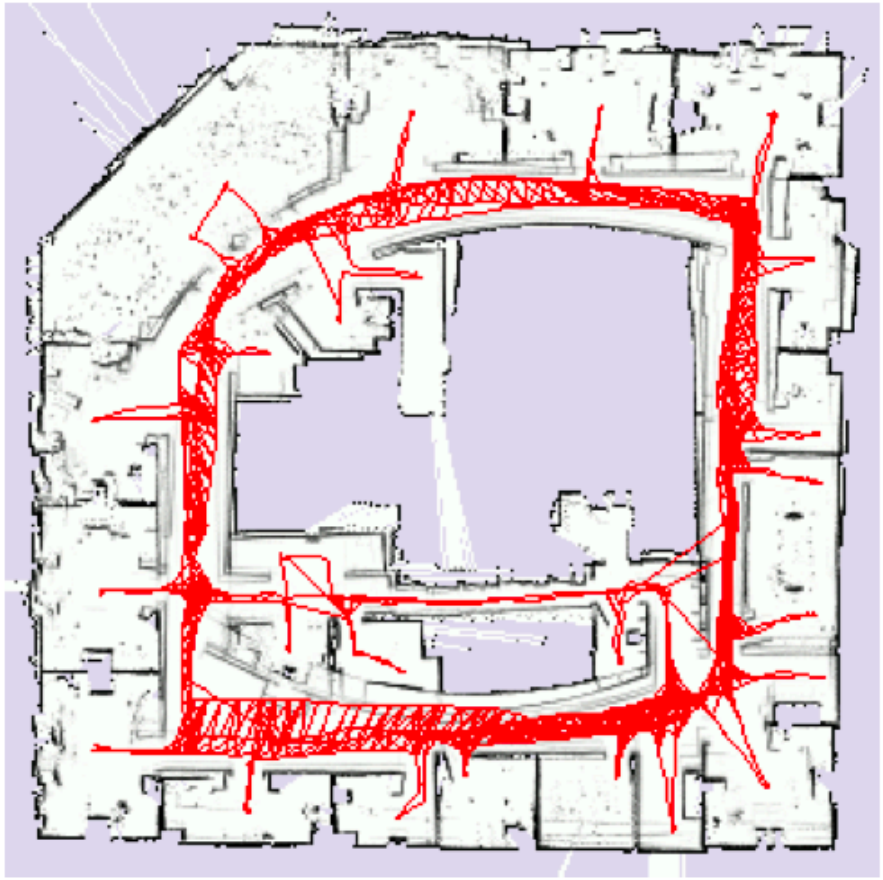
A pose-graph representation of SLAM



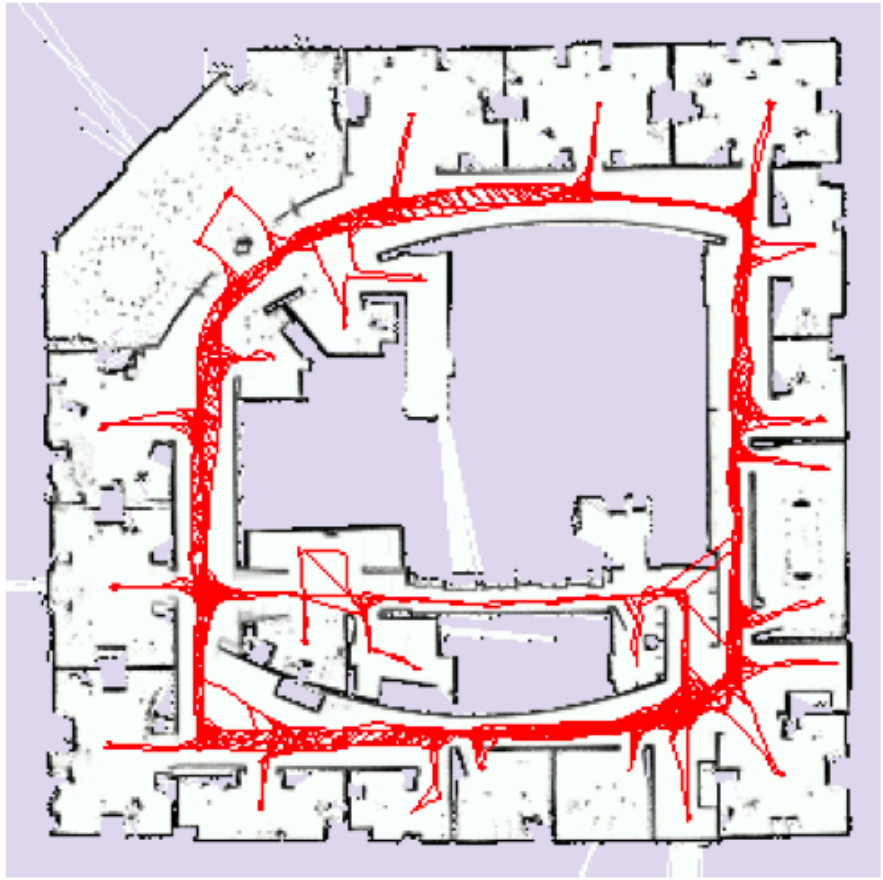
Aspects of an edge in SLAM

Graph-based SLAM

ROBOT INTELLIGENCE LAB



Unoptimized pose graph



Optimized pose graph

ROBOT INTELLIGENCE LAB

LSD-SLAM

"LSD-SLAM: Large-Scale Direct Monocular SLAM," 2014 (TUM)

LSD-SLAM

ROBOT INTELLIGENCE LAB

- "LSD-SLAM: Large-Scale Direct Monocular SLAM," 2014 (TUM)
  - It proposes a **direct** feature-less monocular SLAM algorithm to build large-scale, consistent maps of the environment.
  - There are two different approaches in Monocular SLAM
    1. **Feature**-based methods: Extract features from images and compute scene geometry from these features only.
    2. **Direct** methods: direct visual odometry (VO) optimizes the geometry directly on the image intensities.
  - LSD-SLAM uses **direct image alignment** with filtering-based estimation of semi-dense maps where the global map is represented as a pose graph.

Feature-Based (Indirect) vs. Direct

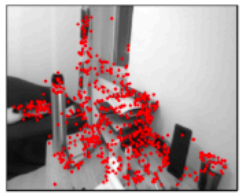
ROBOT INTELLIGENCE LAB

Feature-Based

Input Images

Feature descriptions and matching is required!

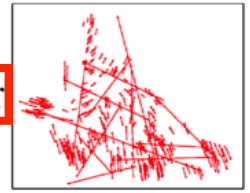
Extract & Match Features
(SIFT / SURF / ...)



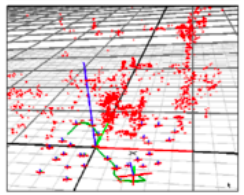
abstract image to feature observations

Bundle adjustment

Track:
min. **reprojection** error
(point distances)



Map:
est. feature-parameters
(3D points / normals)



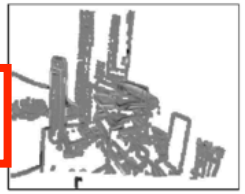
Direct

Input Images

keep full images (no abstraction)

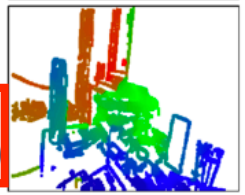
Intensities of pixels matter

Track:
min. **photometric** error
(intensity differences)



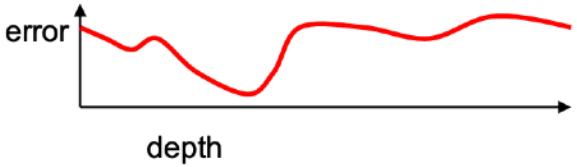
Depth per each pixel!

Map:
est. per-pixel depth
(semi-dense depth map)



Photometric Error

ROBOT INTELLIGENCE LAB

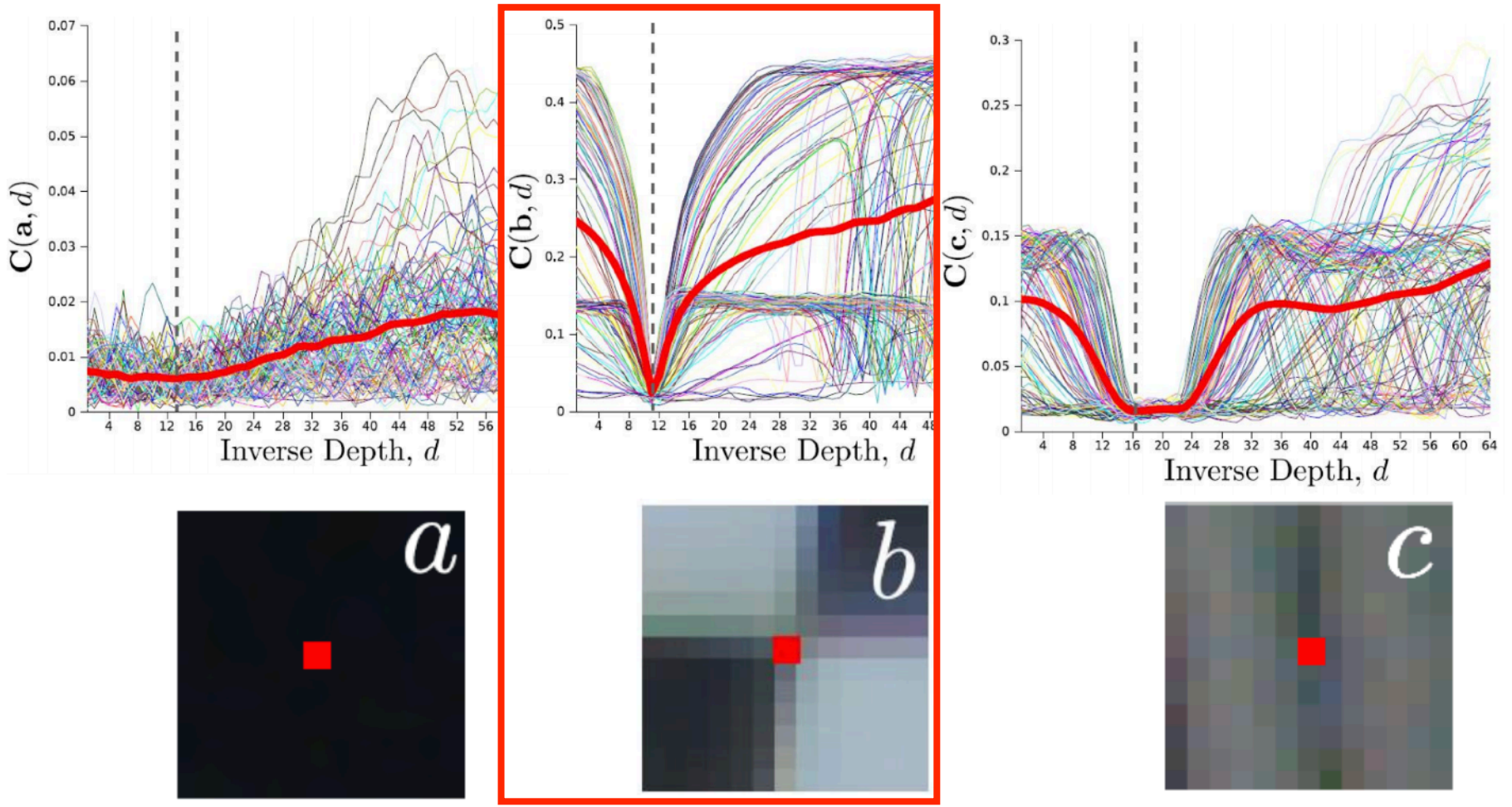


This error plot is derived for every combination of the reference image and any further image

We assume that the **relative positions** and **orientations between cameras** are **known**.

IDEA: the optimal depth minimizes the photometric error in all the images as a function of the depth in the first image

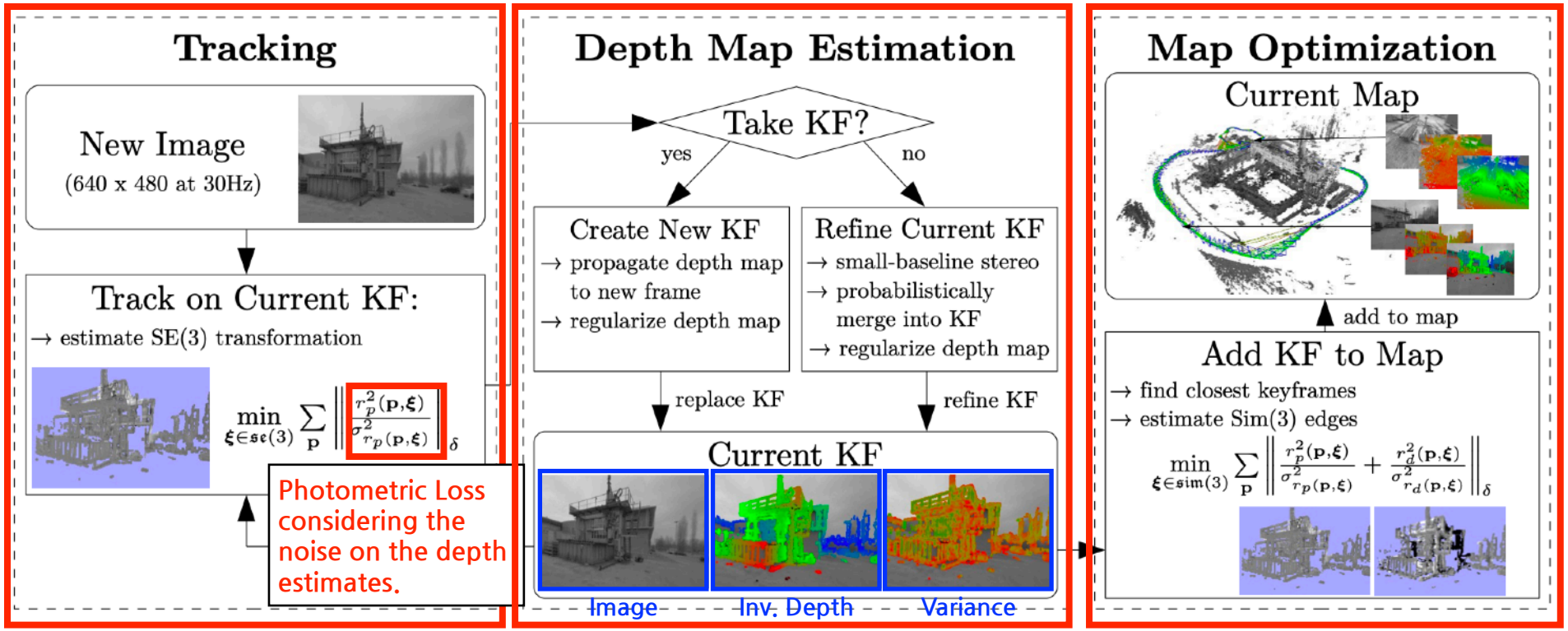
Photometric Error



For distinctive pixels, the aggregated photometric error has typically one clear minimum.

LSD-SLAM

ROBOT INTELLIGENCE LAB



Tracking component continuously tracks new camera images by estimating the rigid body pose (p and R) w.r.t the current keyframe.

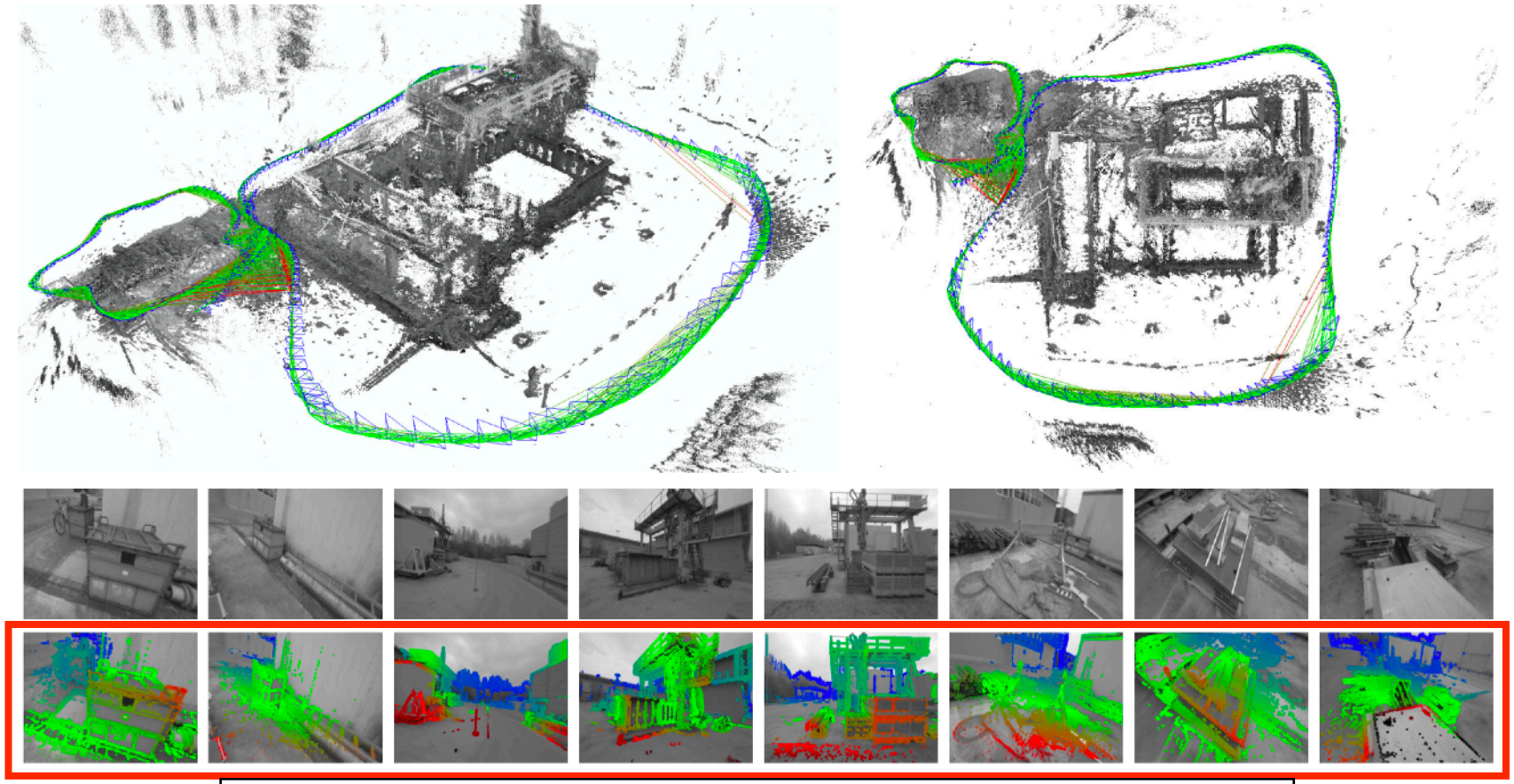
Depth map is estimated by projecting points from the previous keyframe into the new keyframe.

To handle the scale ambiguity, the mean of the inverse depth is scaled to be one.

Finally, graph optimization is used to estimate the global map.

LSD-SLAM

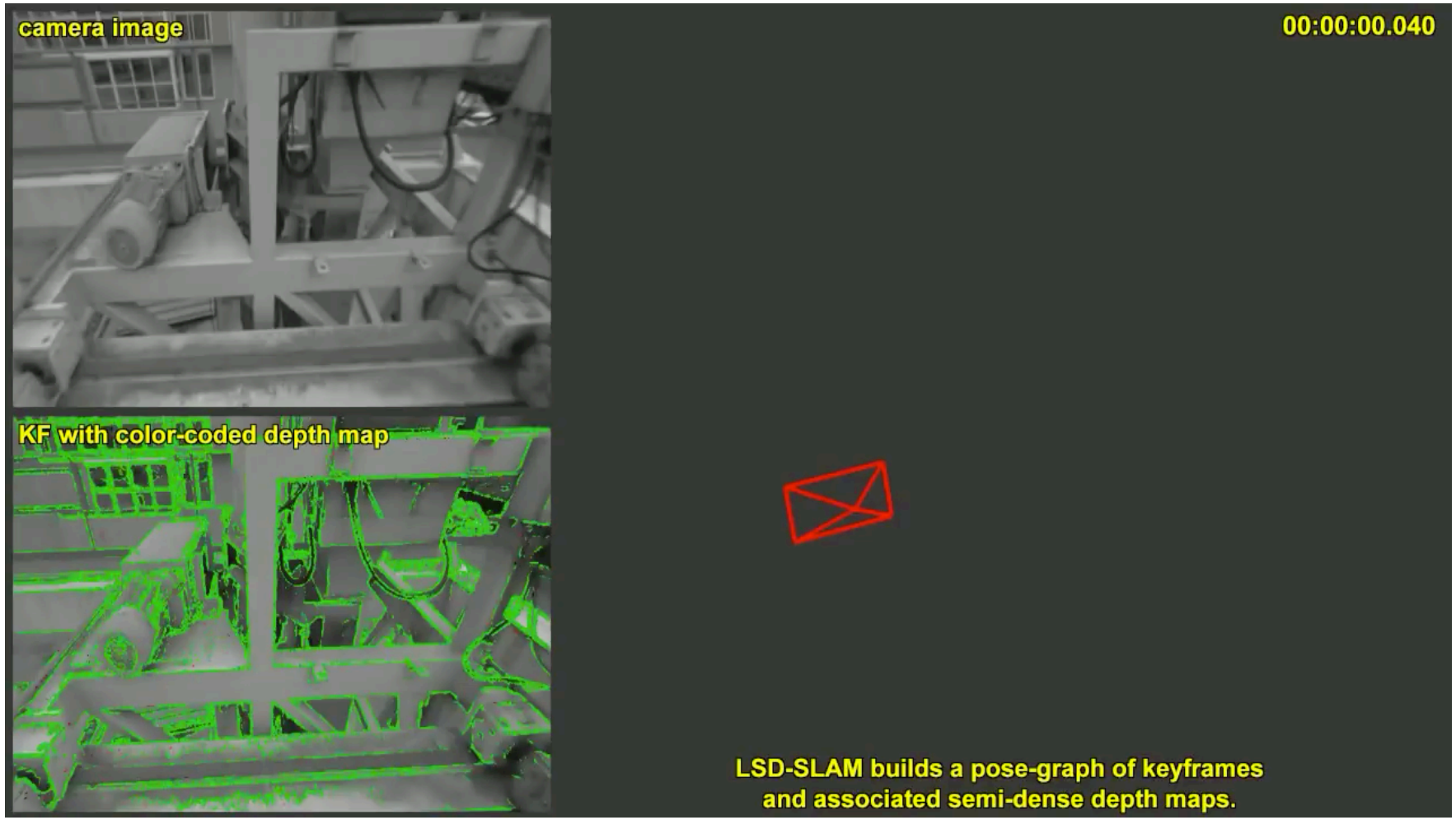
ROBOT INTELLIGENCE LAB



Keyframes with color-coded semi-dense inverse depth map

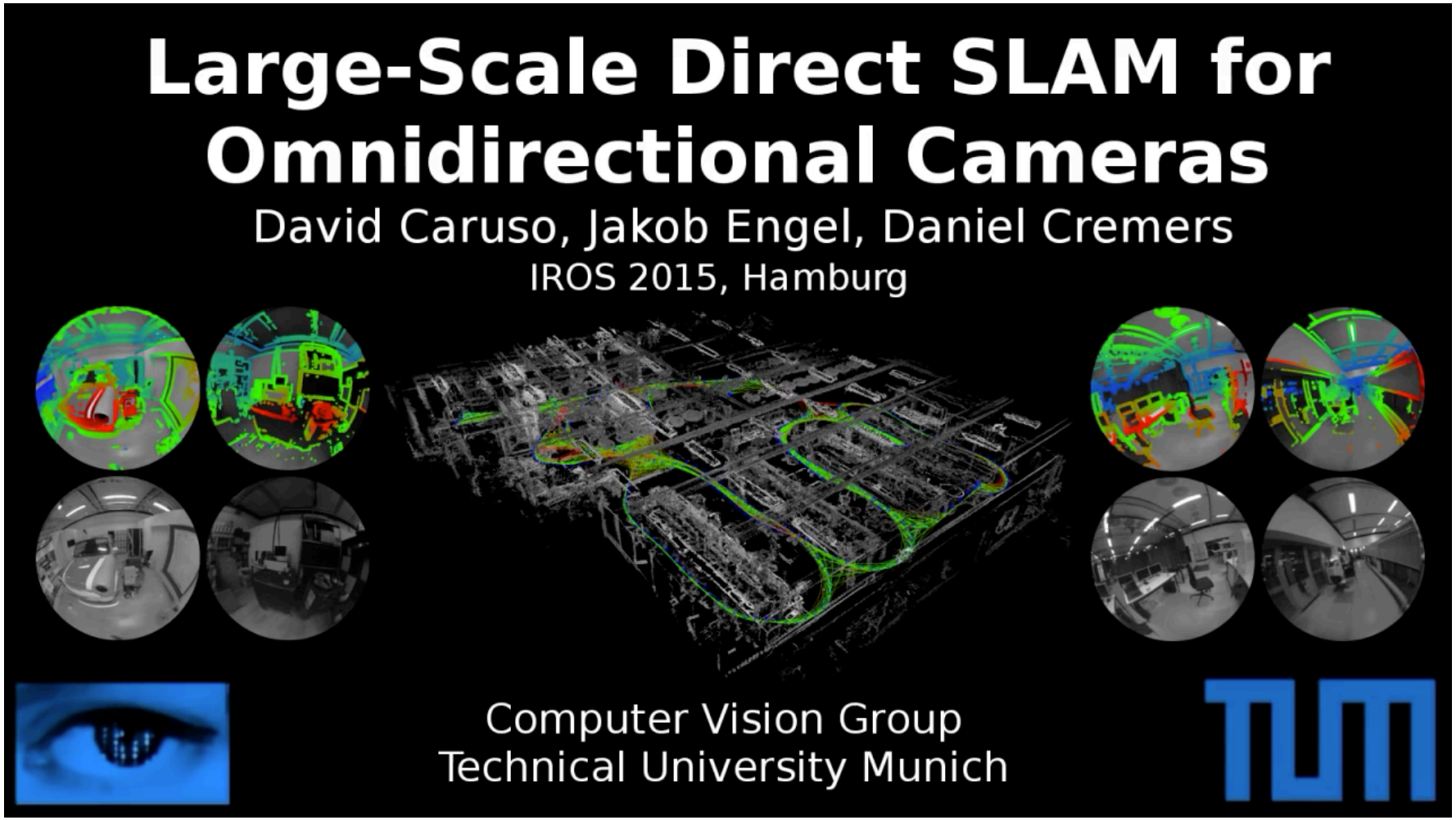
LSD-SLAM

ROBOT INTELLIGENCE LAB



https://youtu.be/oJt3Ln8H03s

Omni-directional LSD-SLAM



https://youtu.be/v0NqMm7Q6S8

ROBOT INTELLIGENCE LAB

Monocular Visual SLAM

"Real-Time 6-DOF Monocular Visual SLAM in a Large-Scale Environment," 2014 (SNU)

Monocular Visual SLAM

ROBOT INTELLIGENCE LAB

- "Real-Time 6-DOF Monocular Visual SLAM in a Large-Scale Environment," 2014 (SNU)
  - It continuously computes the current 6-DOF camera pose and 3D landmark position from a monocular video sequence.
  - A binary **feature descriptor** (FAST detector + BRIEF descriptor) is used for low-cost computation and metric-topological mapping is presented.
  - It considers loop-closure which is the key challenge for real-time visual SLAM in a large-scale environment.

Monocular Visual SLAM

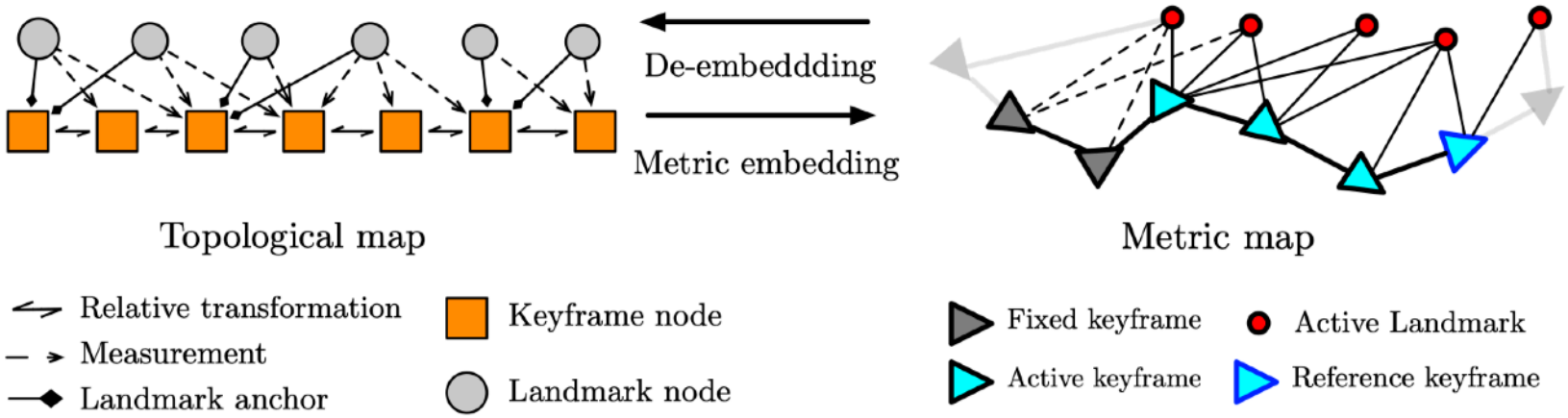
ROBOT INTELLIGENCE LAB

- Outline of the Algorithm
  - **Feature** extraction (BRIEF)
  - Feature tracking - Associate new features to previous extracted features.
  - Keyframe addition - Add a keyframe if the number of tracked features is small or the motion of the camera is larger than a threshold.
  - Pose computation - If enough number of features are matched to the most recent keyframe, the relative pose of the current frame is computed by **RANSAC**.

Monocular Visual SLAM

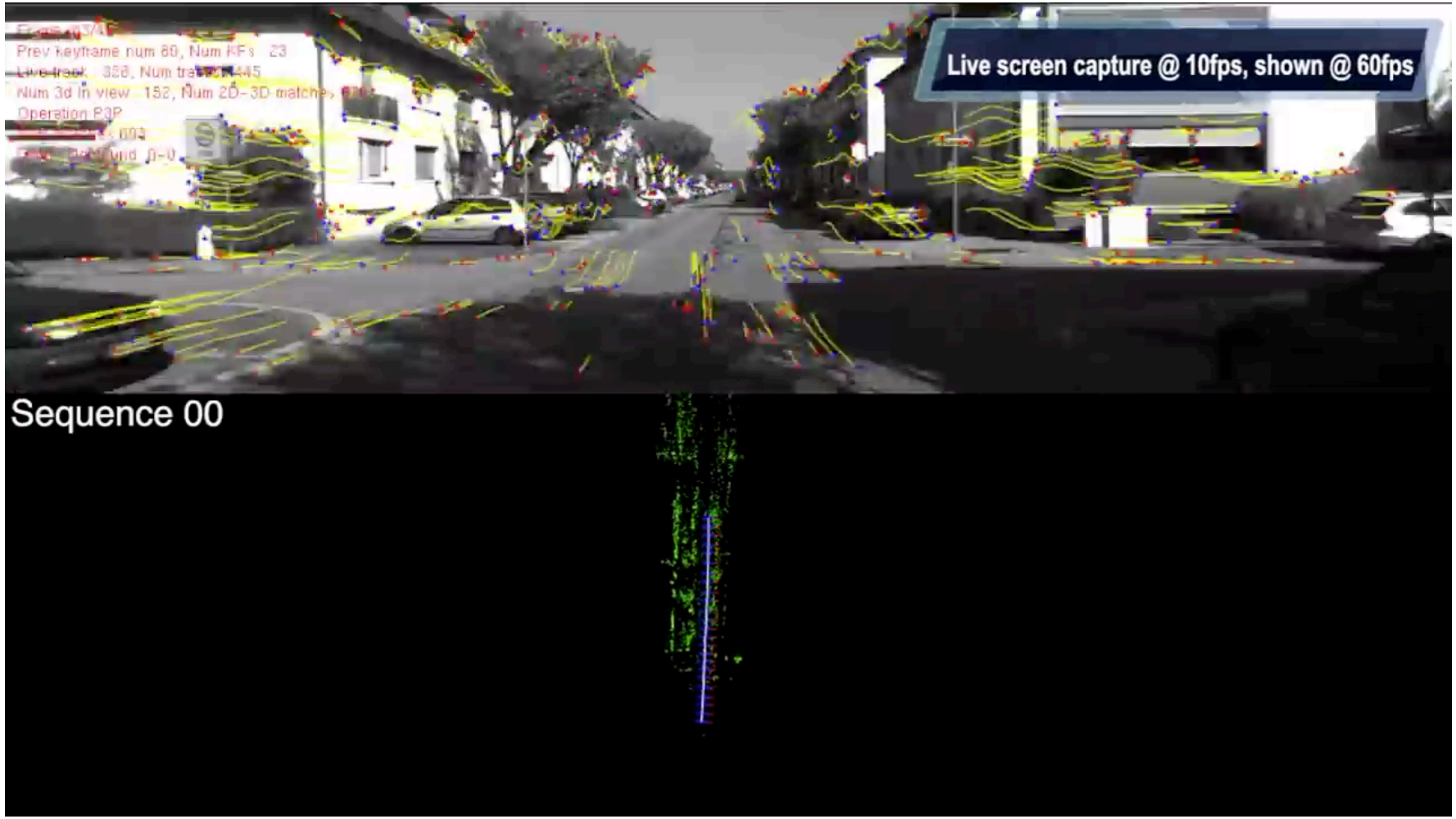
ROBOT INTELLIGENCE LAB

- **Metric-topological map**
  - The **topological** map only stores relative information in edges while the **metric** map contains the location of nodes (for loop closure).
  - As the **topological** map lacks the metric information for optimization, a metric embedding operation is performed.



Monocular Visual SLAM

ROBOT INTELLIGENCE LAB



https://youtu.be/JyG1EeqCmHY

ROBOT INTELLIGENCE LAB

SVO

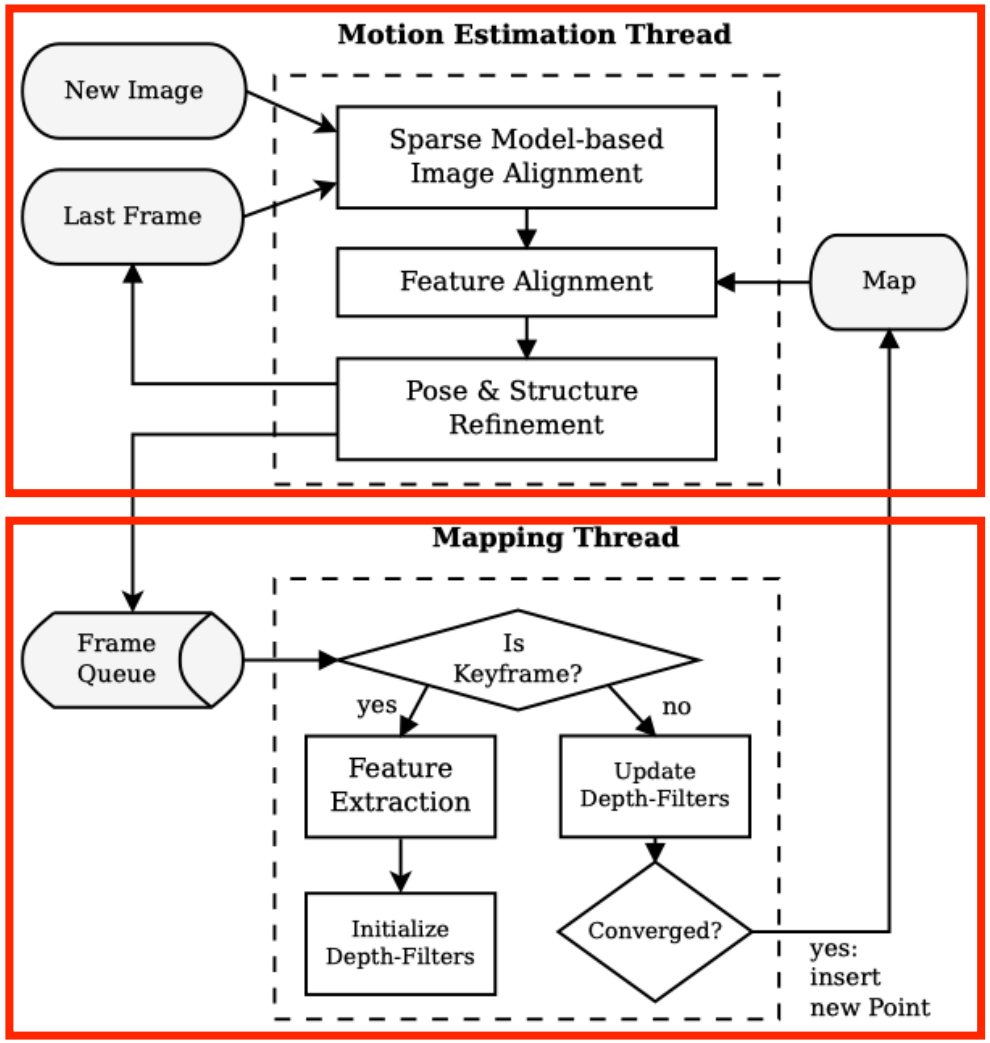
"SVO: Fast Semi-Direct Monocular Visual Odometry," 2014 (UZH)

SVO

ROBOT INTELLIGENCE LAB

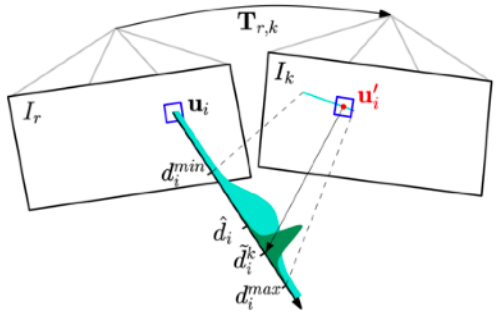
- "SVO: Fast Semi-Direct Monocular Visual Odometry," 2014 (UZH)
  - SVO eliminates the need of costly **feature** extraction and robust matching for motion estimation.
  - It operates **directly** on pixel intensities, which results in sub-pixel precision at high frame-rates with probabilistic mapping methods.
  - SVO uses feature-correspondences, but it is an implicit result of direct motion estimation rather than of explicit feature extraction and matching.
  - Thus, feature extraction is only required for initializing new 3D points.

SVO



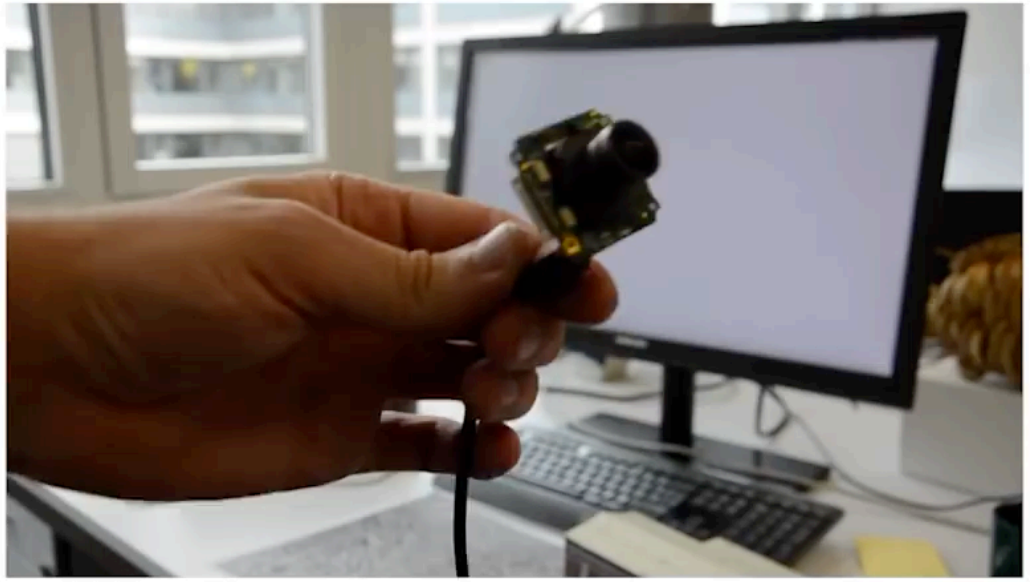
The camera pose relative to the previous frame is found through minimizing the photometric error between patches.

The depth of 2D features is estimated.



SVO

ROBOT INTELLIGENCE LAB



https://youtu.be/2YnIMfw6bJY

ROBOT INTELLIGENCE LAB

DynamicFusion

"DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time," 2015

DynamicFusion

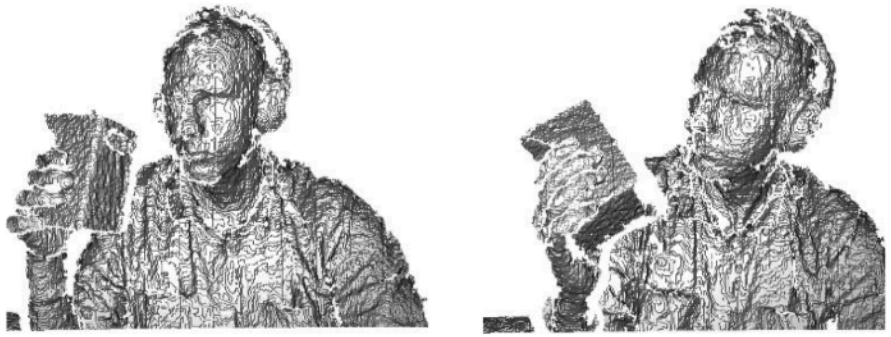
ROBOT INTELLIGENCE LAB

- "DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time," 2015
  - It presents the first dense SLAM system capable of reconstruction non-rigidly deforming scenes in real-time by fusing together RGBD scans.
  - DynamicFusion reconstructs scene geometry while simultaneously estimating a dense volumetric 6D motion field that warps the estimated geometry into a live frame.

DynamicFusion

ROBOT INTELLIGENCE LAB

(a) Initial Frame at $t = 0s$



(b) Raw (noisy) depth maps for frames at $t = 1s, 10s, 15s, 20s$

(c) Node Distance

(d) Canonical Model

(e) Canonical model warped into its live frame

(f) Model Normals

DynamicFusion

ROBOT INTELLIGENCE LAB

Live Input Depth Map

Live Model Output

Live RGB Image (unused)

Canonical Model Reconstruction

Warped Model

https://youtu.be/i1eZekcc_lM

ROBOT INTELLIGENCE LAB

ORB-SLAM

"ORB-SLAM: a Versatile and Accurate Monocular SLAM System," 2015

ORB-SLAM

ROBOT INTELLIGENCE LAB

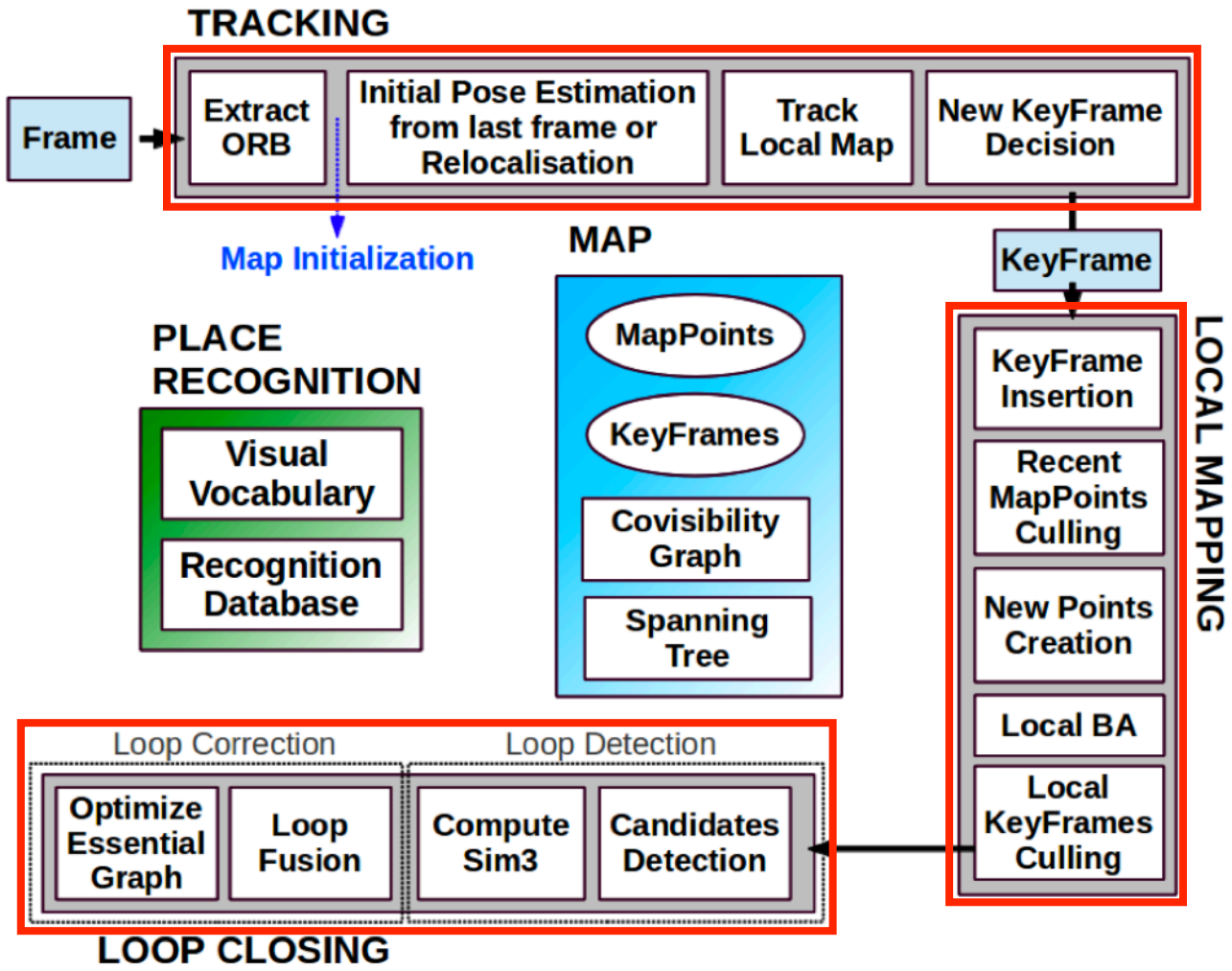
- "ORB-SLAM: a Versatile and Accurate Monocular SLAM System," 2015
  - It presents a **feature**-based monocular SLAM system that operates in real time in versatile environments.
  - It is built on the main ideas of **PTAM**.
  - ORB-SLAM uses the same features for all SLAM tasks: tracking, mapping, re-localization, and loop closing.
  - **ORB feature (binary) descriptor** [1] is used (which is rotation invariant and resistant to noise and faster than SIFT or SURF).

ORB-SLAM

ROBOT INTELLIGENCE LAB

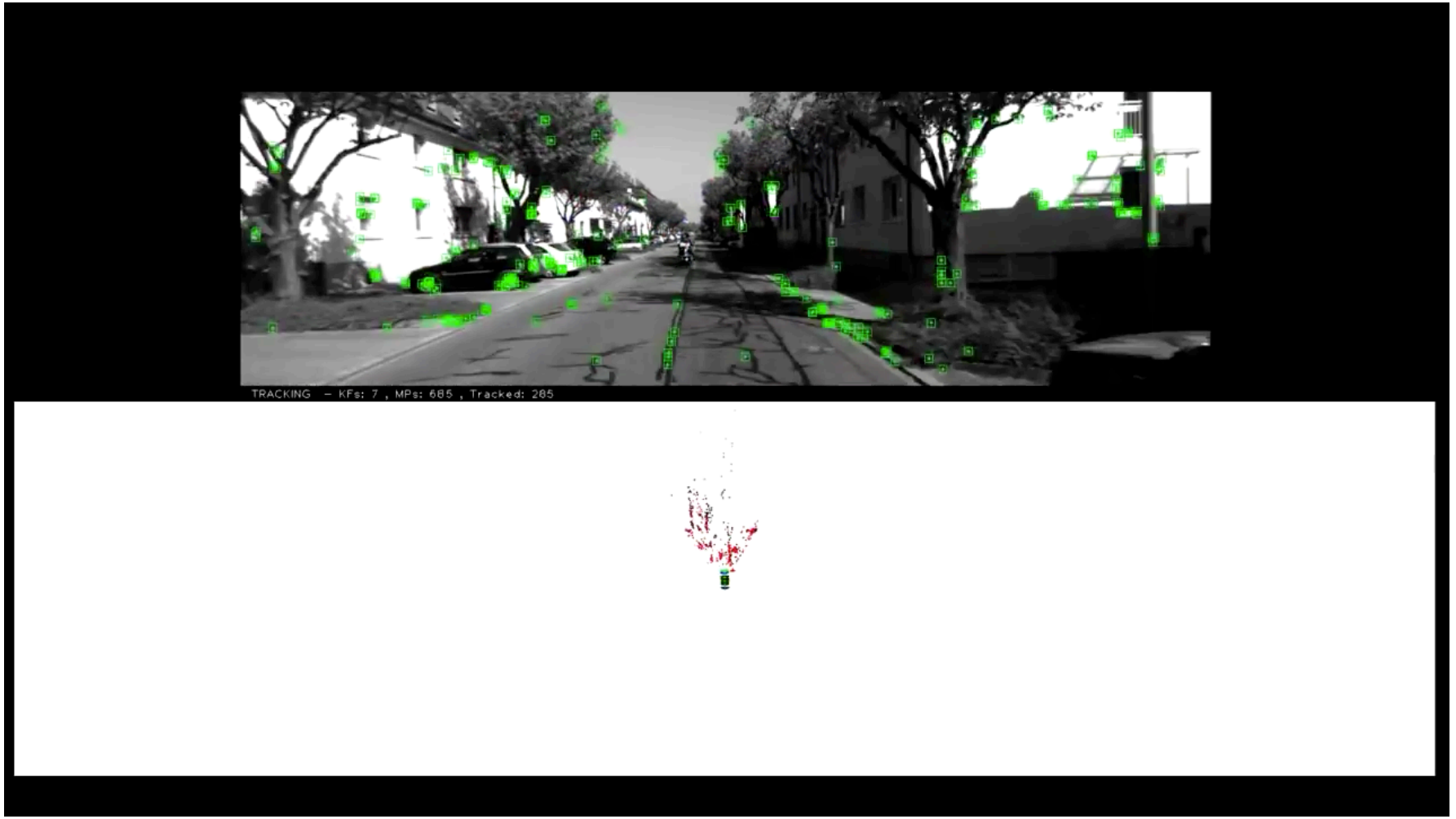
- Main contributions
  - Use of the same **features** (ORB features).
  - Real time loop closing based on the Essential Graph.
  - Real time camera relocalization.
  - Novel initialization technique.

ORB-SLAM



- Three threads:
  - Tracking
  - Local Mapping
  - Loop Closing
- It outperforms **PTAM** (indirect) and **LSD-SLAM** (direct).

ORB-SLAM



https://youtu.be/8DISRmsO2YQ

ROBOT INTELLIGENCE LAB

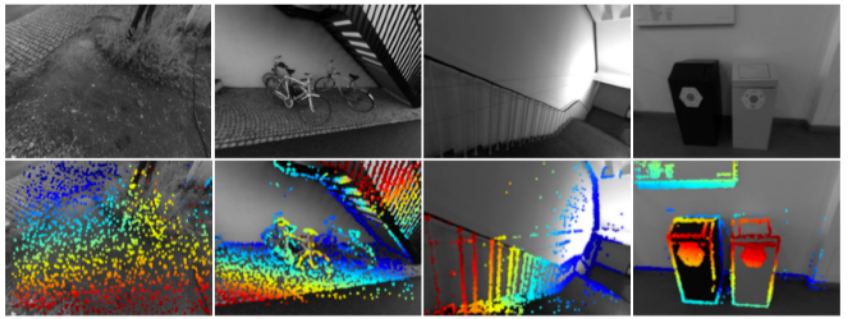
Direct Sparse Odometry

"Direct Sparse Odometry," 2016 (TUM)

Direct Sparse Odometry

ROBOT INTELLIGENCE LAB

- "Direct Sparse Odometry," 2016 (TUM)
  - **Direct** (feature-less) vs. Indirect and Sparse vs.Dense
    - Sparse+Indirect: Most widely-used, estimating 3D geometry from a set of key-point matches (monoSLAM, PTAM, ORB-SLAM)
    - Dense+Indirect: Estimate 3D geometry from a dense optical flow field.
    - Dense+**Direct**: Employs a photometric error with a geometric prior (DTAM, LSD-SLAM)
    - Sparse+**Direct**: The one proposed here.
  - Same author of LSD-SLAM (**direct**).



Example depth maps used for initial frame tracking.

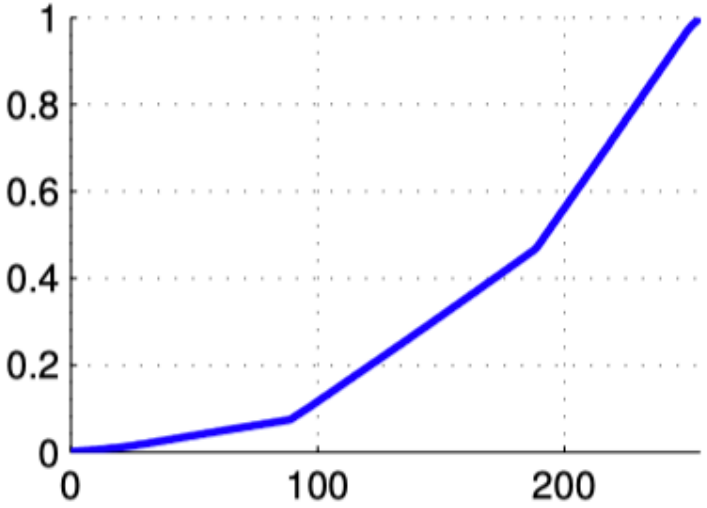
Direct Sparse Odometry

- Why **direct** and sparse?
  - **Direct**: Whereas one of the main benefits of key-points (indirect) is their robustness, direct methods does not require a point to be recognizable by itself, hence allows a more finely grained geometry representation (e.g., pixel-wise inverse depth).
  - Sparse: Less restrictive to geometric prior.
- Unlike SVO that takes hybrid approaches (semi-direct), it is the first fully direct method that jointly optimizes camera poses, camera intrinsics, and geometry parameters (inverse depth).

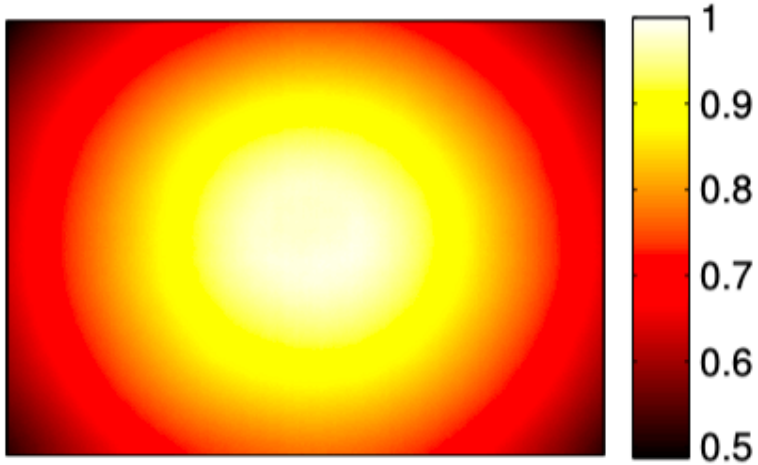
Direct Sparse Odometry

ROBOT INTELLIGENCE LAB

- Photometric Calibration
  - DSO leverages extensive photometric calibration to achieve fully-Direct SLAM.



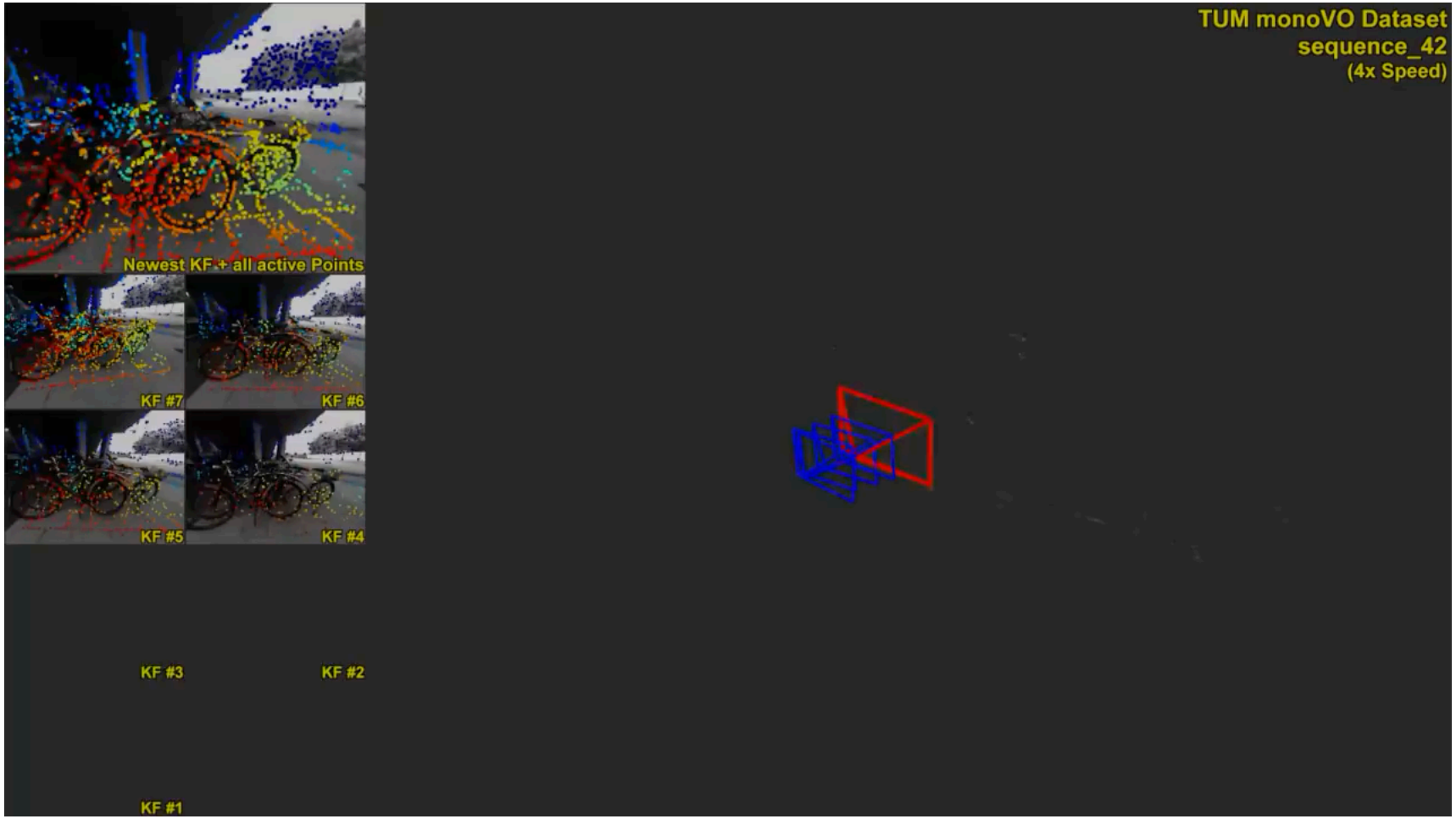
Inverse response function G^{-1}
(pixel intensity $[0,255]$ to $[0,1]$)



Lens attenuation model
$V : \Omega \rightarrow [0,1]$

Direct Sparse Odometry

ROBOT INTELLIGENCE LAB



https://youtu.be/C6-xwSOOdqQ

ROBOT INTELLIGENCE LAB

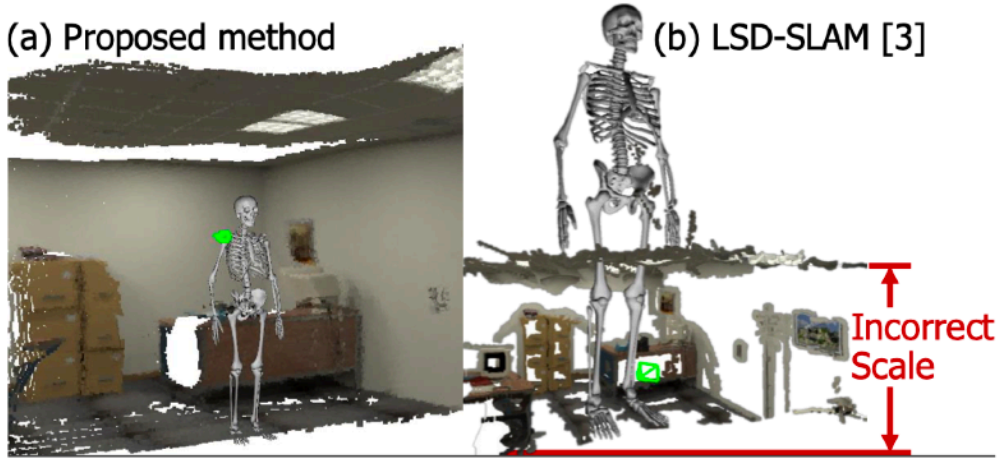
CNN-SLAM

"CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction," 2017

CNN-SLAM

ROBOT INTELLIGENCE LAB

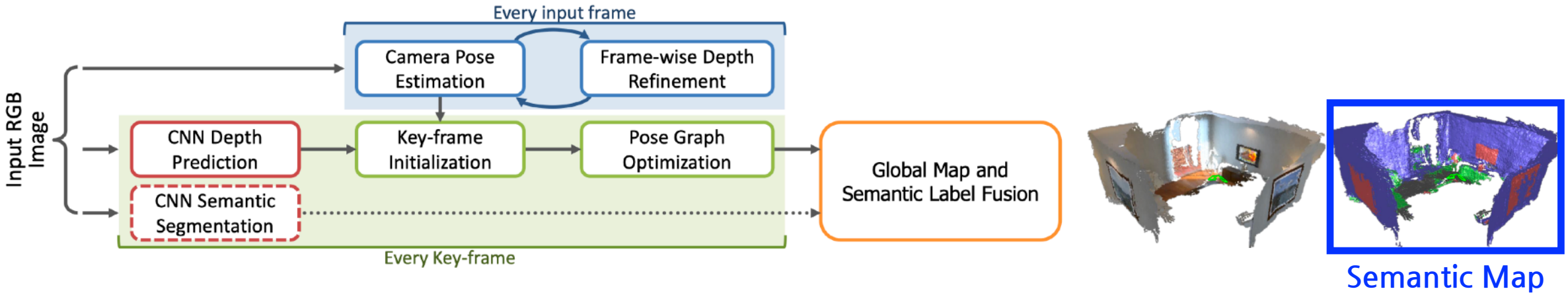
- "CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction," 2017
  - This paper investigates how predicted depth maps (from CNN) can be deployed from accurate and dense monocular reconstruction.
  - CNN-predicted dense depth maps are fused with depth measurements obtained from direct monocular SLAM overcoming existing limitations (e.g., low-textured regions).
  - The use of depth prediction enables better absolute scale estimations.



CNN-SLAM

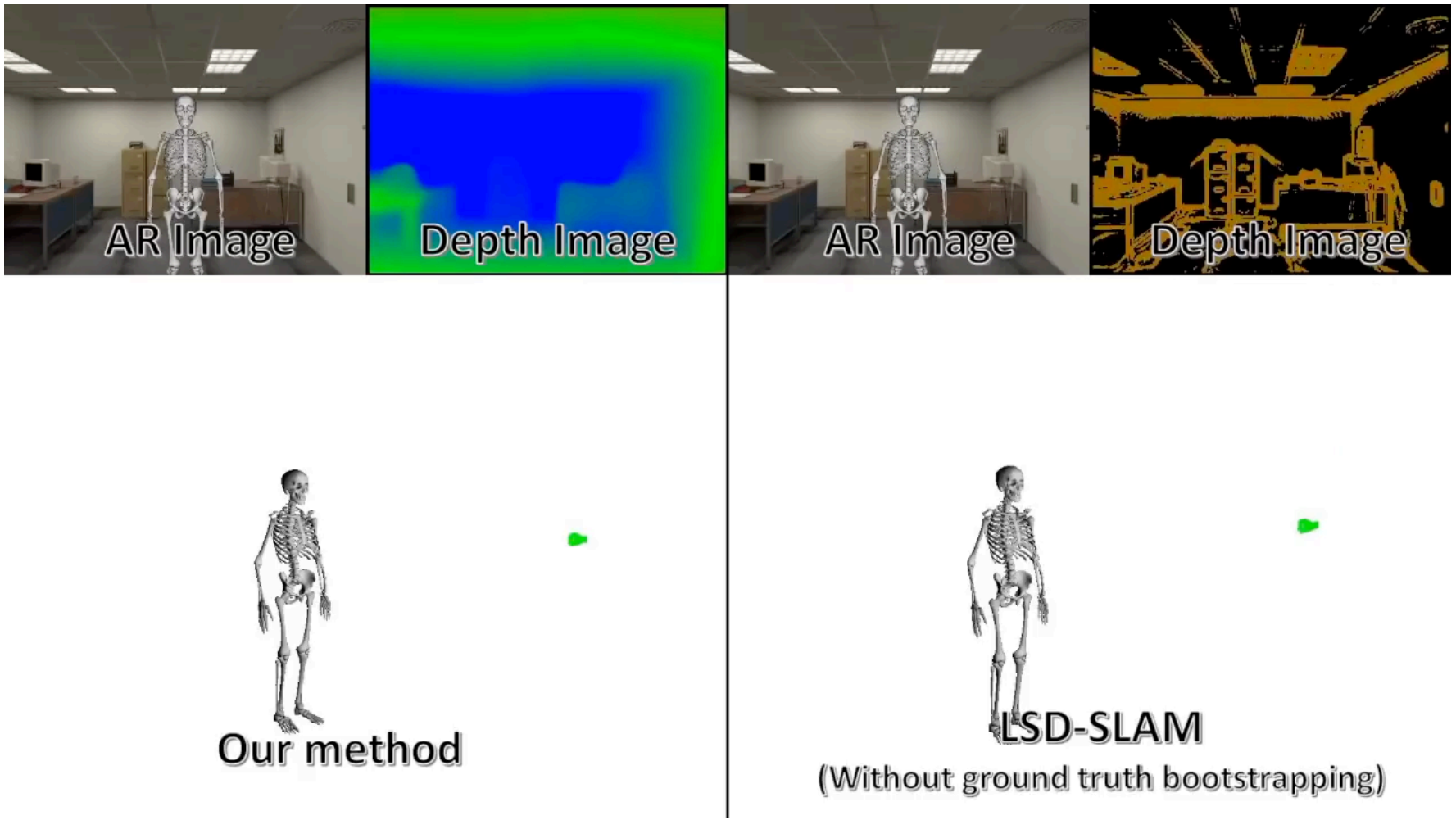
ROBOT INTELLIGENCE LAB

- It also fuses semantic labels from Semantic Segmentation yielding semantically coherent scene reconstruction from a single view.
- CNN-SLAM shows superior performances compared to LSD-SLAM and ORB-SLAM with respect to estimating the absolute scale.



CNN-SLAM

ROBOT INTELLIGENCE LAB



https://youtu.be/z_NJxbkQnBU

ROBOT INTELLIGENCE LAB

ORB-SLAM2

"ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras," 2017

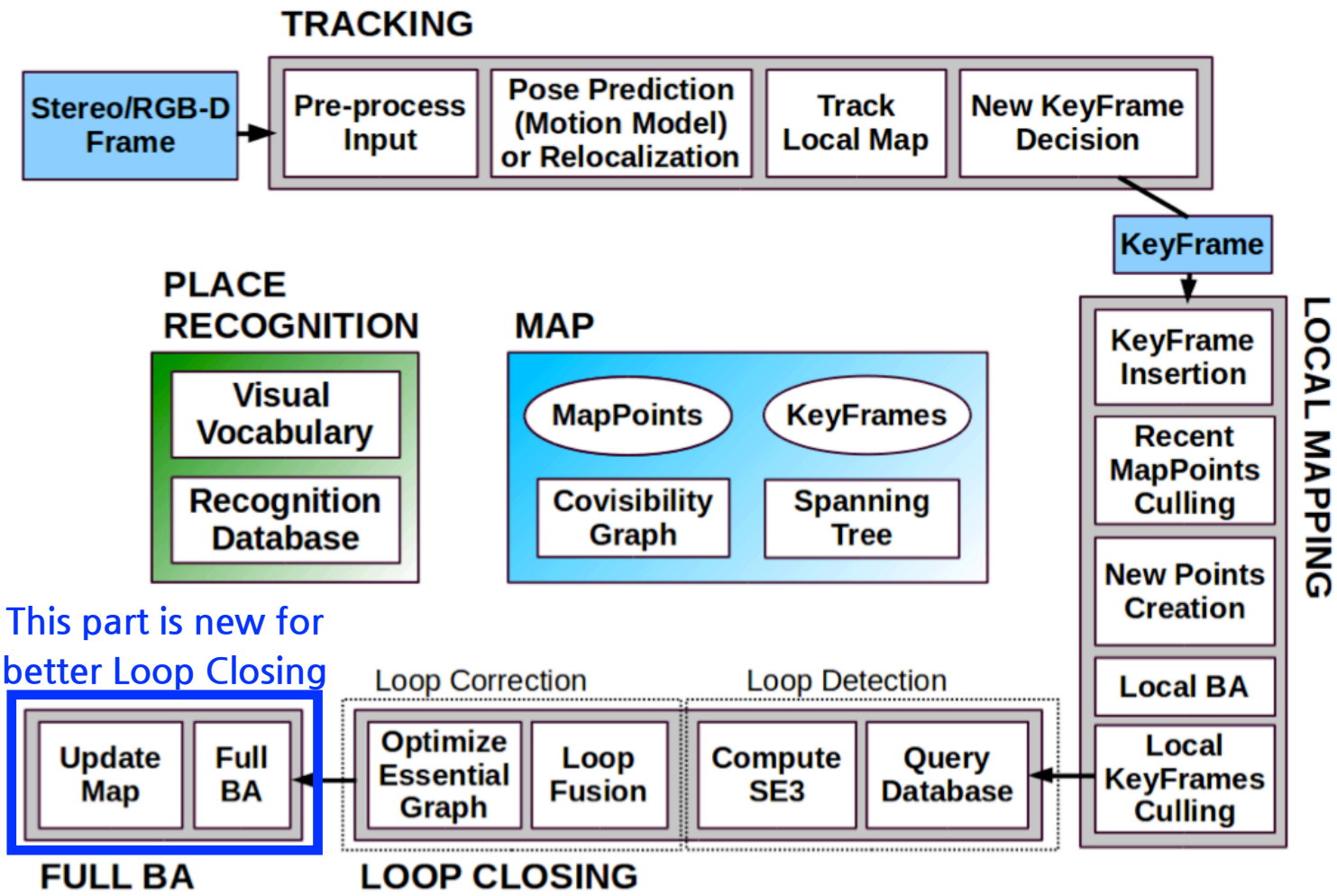
ORB-SLAM2

ROBOT INTELLIGENCE LAB

- "ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras," 2017
  - This work presents a complete SLAM system for monocular, stereo and RGB-D cameras, including map reuse, loop closing and re-localization capabilities and works in real-time on standard CPUs.
  - It is the first open-source SLAM system for monocular, stereo, and RGB-D cameras.
  - Of course, ORB-SLAM2 also leverages ORB features.

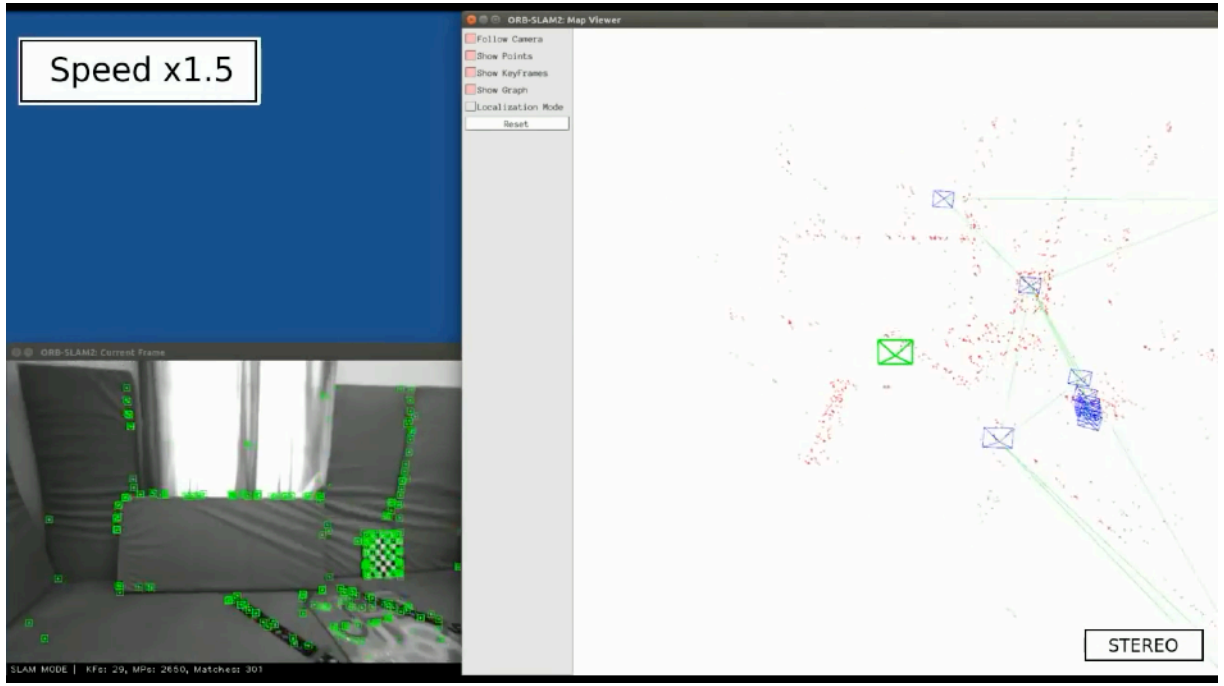
ORB-SLAM2

ROBOT INTELLIGENCE LAB

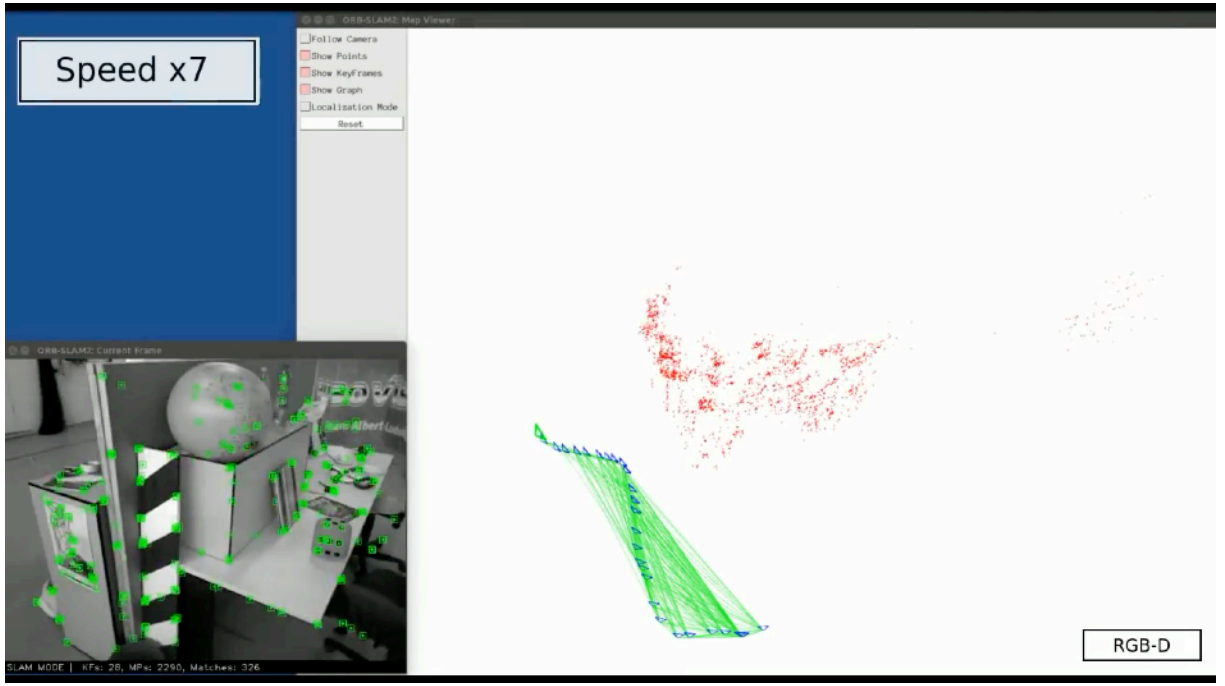


ORB-SLAM2

ROBOT INTELLIGENCE LAB



Stereo Camera



RGBD Camera

ORB-SLAM2

ROBOT INTELLIGENCE LAB

Dense RGB-D reconstruction **without** any fusion.

ROBOT INTELLIGENCE LAB

ProSLAM

"ProSLAM: Graph SLAM from a Programmer's Perspective," 2017

ProSLAM

ROBOT INTELLIGENCE LAB

- "ProSLAM: Graph SLAM from a Programmer's Perspective," 2017
  - It presents a lightweight stereo visual SLAM system designed with simplicity in mind highlighting the data structures and the algorithmic aspects.
  - **Feature** based SLAM: FAST detector + BRIEF descriptor
  - Eigen (matrix calculation), OpenCV (feature extraction), g2o (pose graph optimization)

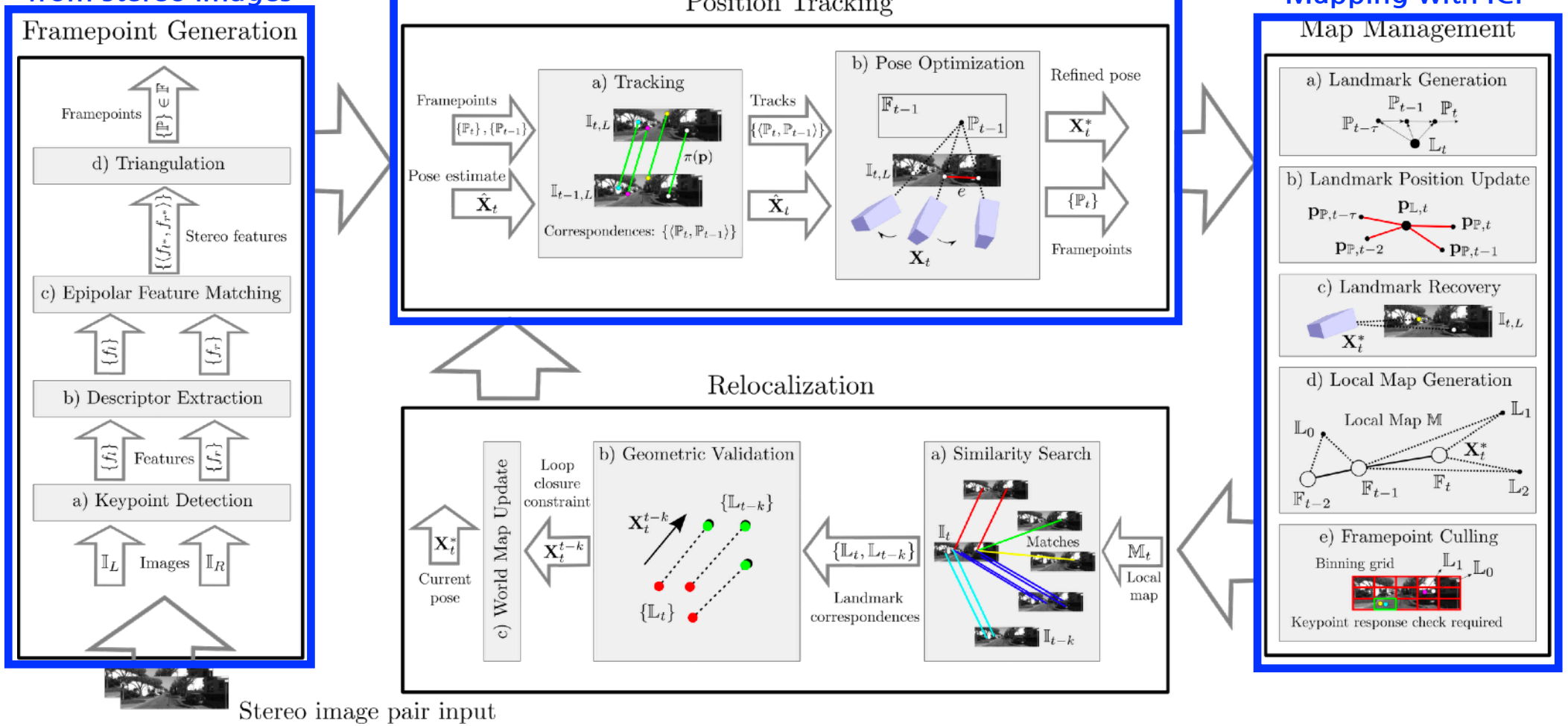
ProSLAM

ROBOT INTELLIGENCE LAB

Generate 3D point cloud from stereo images

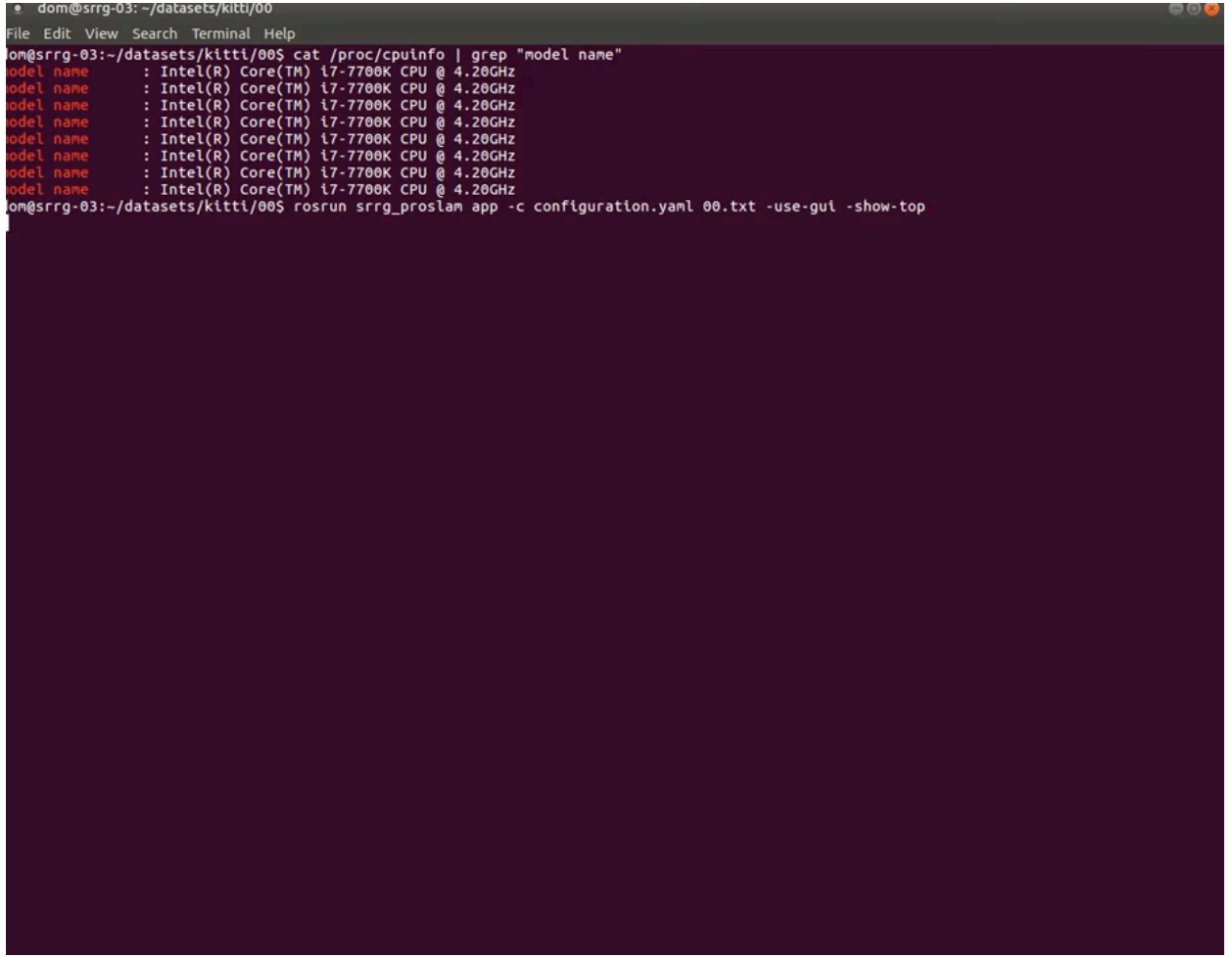
Camera position tracking

Mapping with ICP



ProSLAM

ROBOT INTELLIGENCE LAB



✦ **ProSLAM** outperformed **LSD-SLAM** in 7/11 and **ORB-SLAM2** in 3/11 cases in KITTI SLAM evaluation.

ROBOT INTELLIGENCE LAB

Geometric Deep SLAM

"Toward Geometric Deep SLAM," 2017

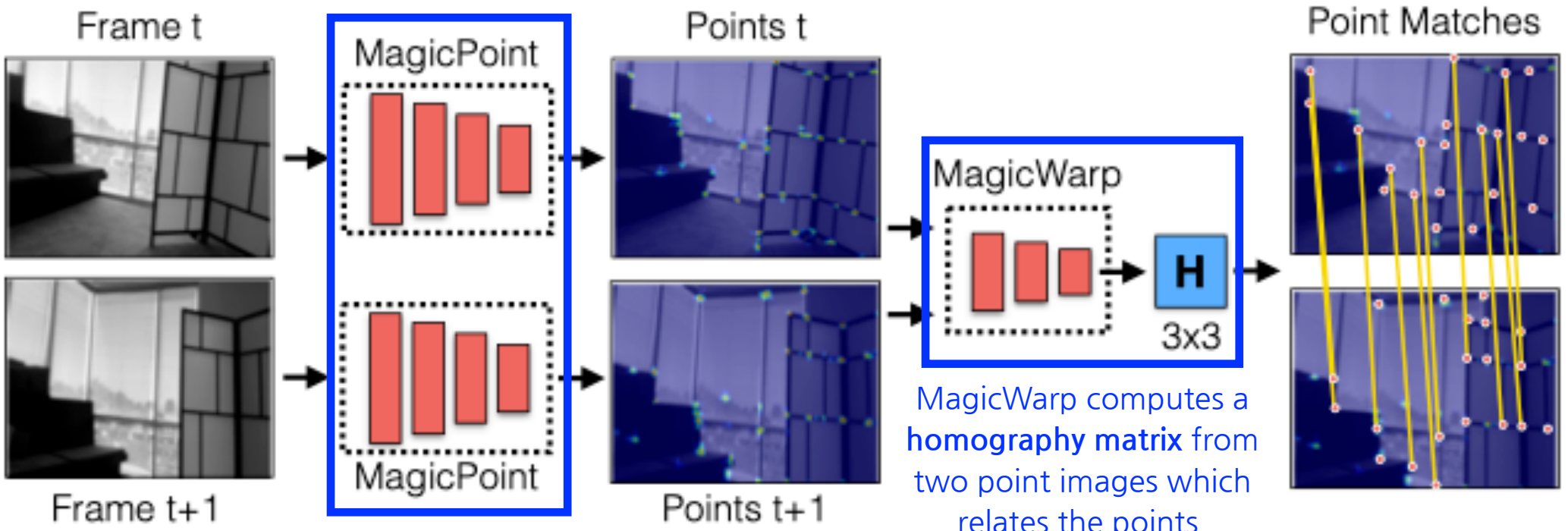
Geometric Deep SLAM

ROBOT INTELLIGENCE LAB

- "Toward Geometric Deep SLAM," 2017
  - It presents a point tracking system by two deep CNNs:
    - **MagicPoint** extracts salient 2D points on single images
    - **MagicWarp** estimates the homography from the pairs of point images (output of MagicPoint).
  - It shows its robustness in the presence of image noise and runs 30+ FPS on a single CPU as it is free from local point descriptors.

Geometric Deep SLAM

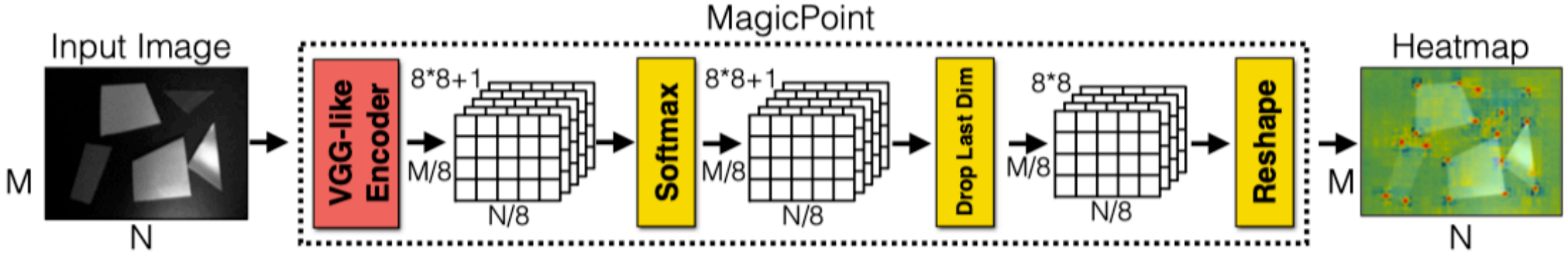
ROBOT INTELLIGENCE LAB



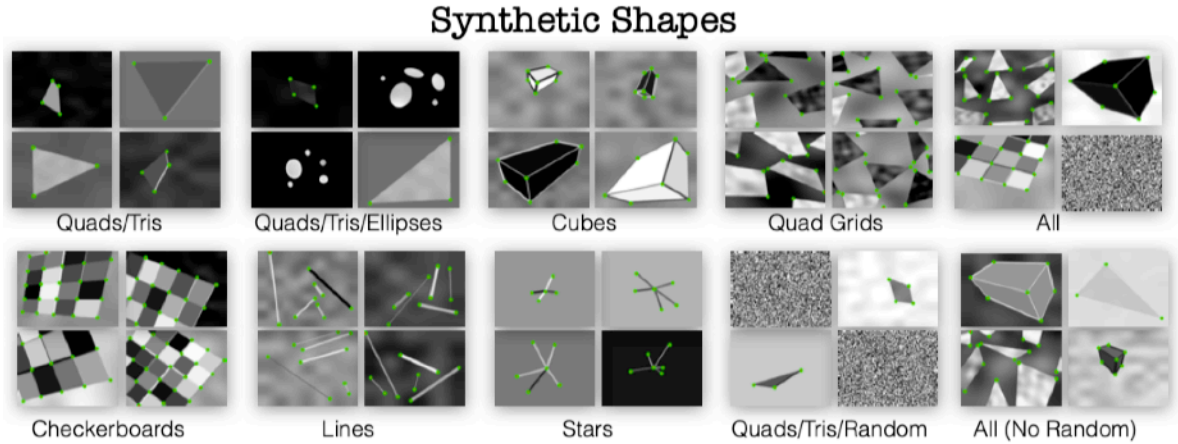
Magic Point is trained to detect salient corners in the image.

MagicWarp computes a **homography matrix** from two point images which relates the points.

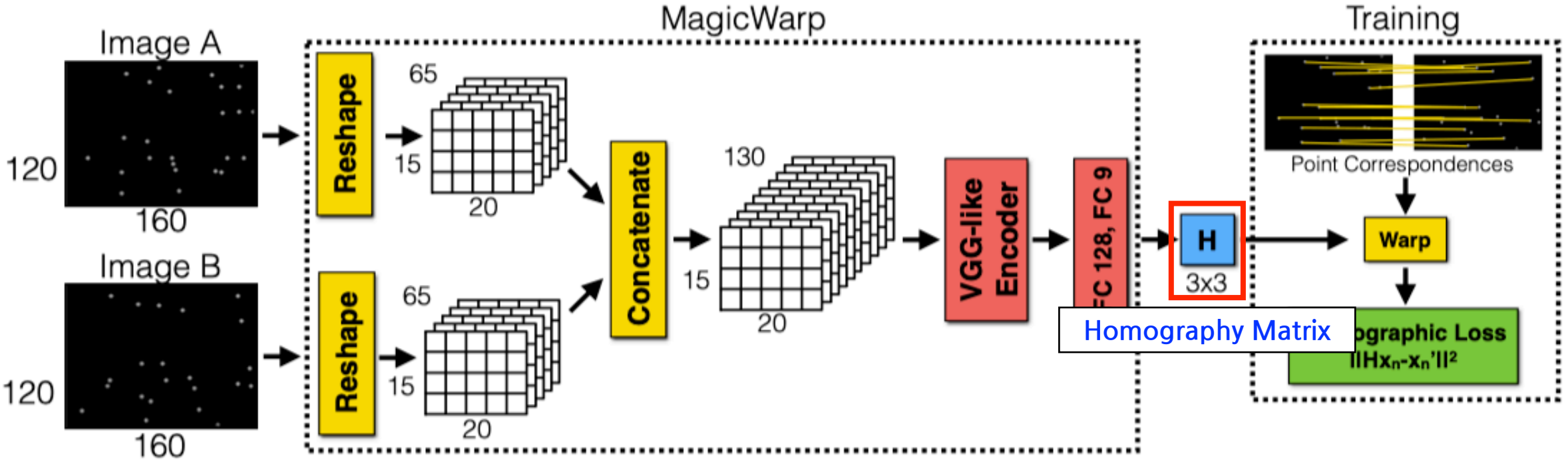
Geometric Deep SLAM



- **MagicPoint** is trained with synthetic dataset to output a "point-ness" probability for each pixel (corner detection).



Geometric Deep SLAM



- **MagicWarp** is optimized to estimate the **homography matrix** by minimizing the **Homographic Loss**.

Super Glue

ROBOT INTELLIGENCE LAB

- Geometric Deep SLAM does not have a video available. Following is **SuperGlue** from the same author:

https://github.com/magicleap/SuperGluePretrainedNetwork

Thank You