



Robotic Vision and Navigation

Part III Visual SLAM

Lourdes Agapito

Email:

l.agapito@ucl.ac.uk

Slide acknowledgements: some slides have been adapted from S. Julier,
D. Scaramuzza, J. Engel, R. Newcombe

Motivation: Monocular Camera SLAM



What is Visual Odometry (VO) ?

VO is the process of incrementally estimating the pose of the vehicle by examining the changes that motion induces on the images of its onboard cameras

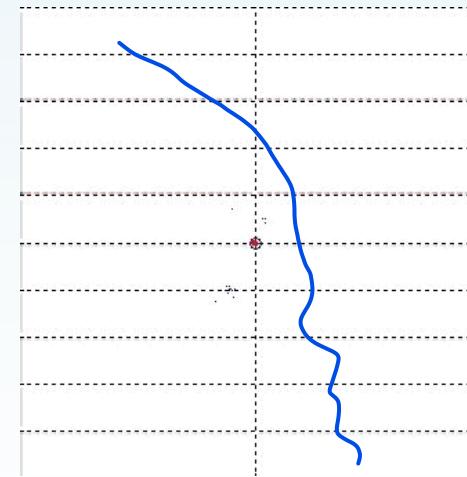
input



Image sequence (or video stream)
from one or more cameras attached to a moving vehicle



output



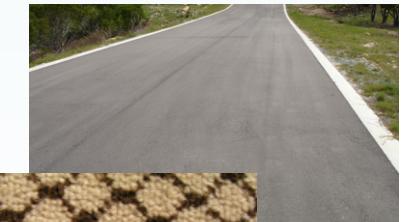
R_0, R_1, \dots, R_i

t_0, t_1, \dots, t_i

Camera trajectory (3D structure is a plus):

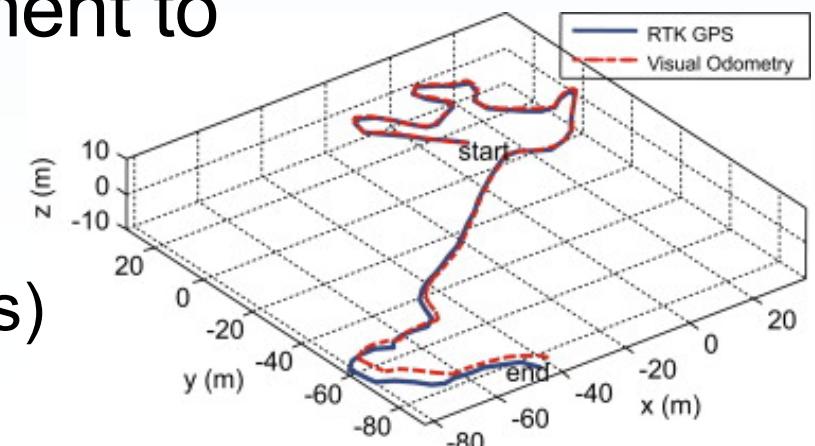
Assumptions

- Sufficient illumination in the environment
- Dominance of static scene over moving objects
- Enough texture to allow apparent motion to be extracted
- Sufficient scene overlap between consecutive frames

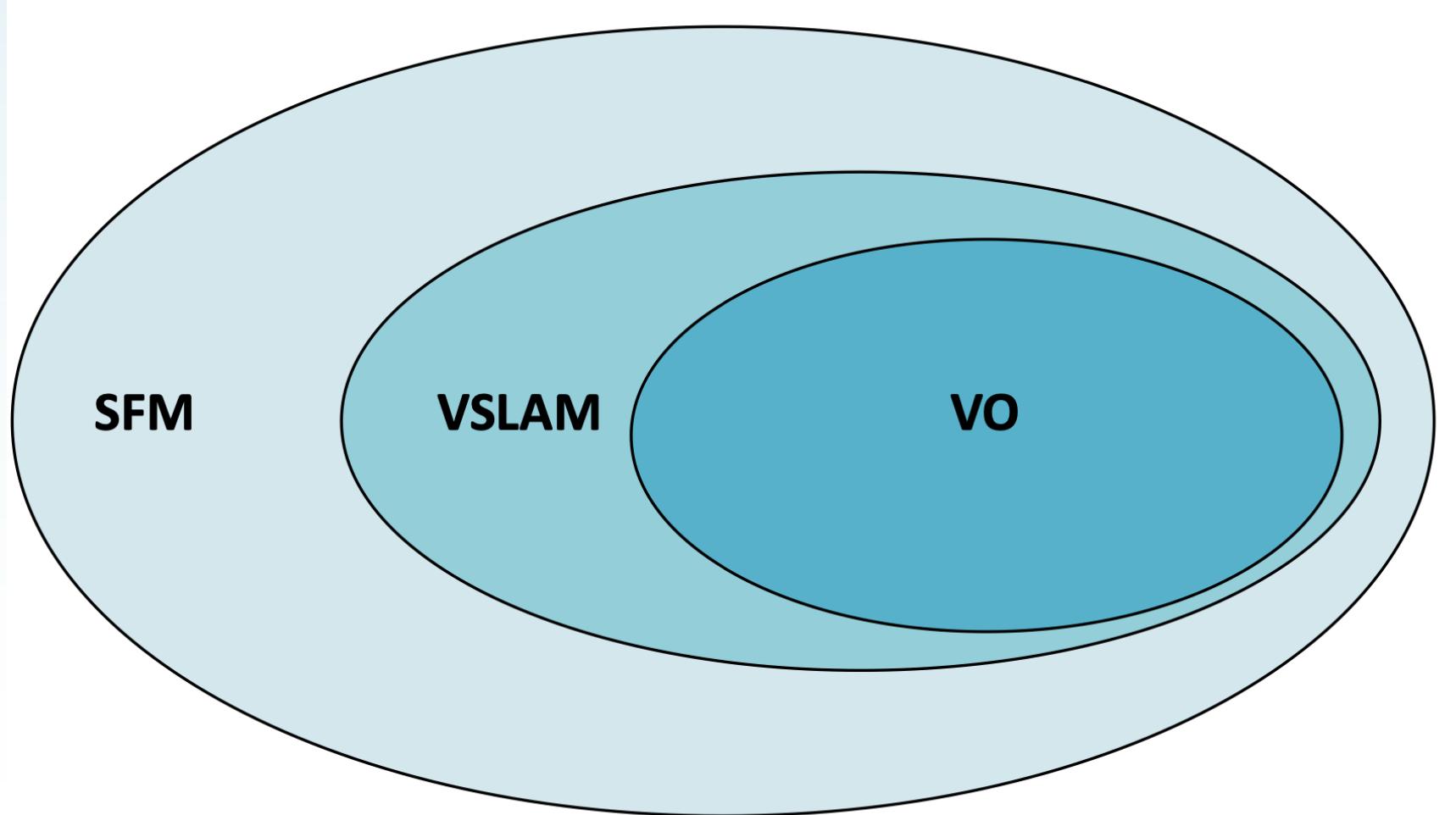


Why VO ?

- Contrary to wheel odometry, VO is not affected by wheel slip in uneven terrain or other adverse conditions.
- More accurate trajectory estimates compared to wheel odometry (relative position error 0.1% – 2%)
- VO can be used as a complement to
 - wheel odometry
 - GPS
 - inertial measurement units (IMUs)
 - laser odometry

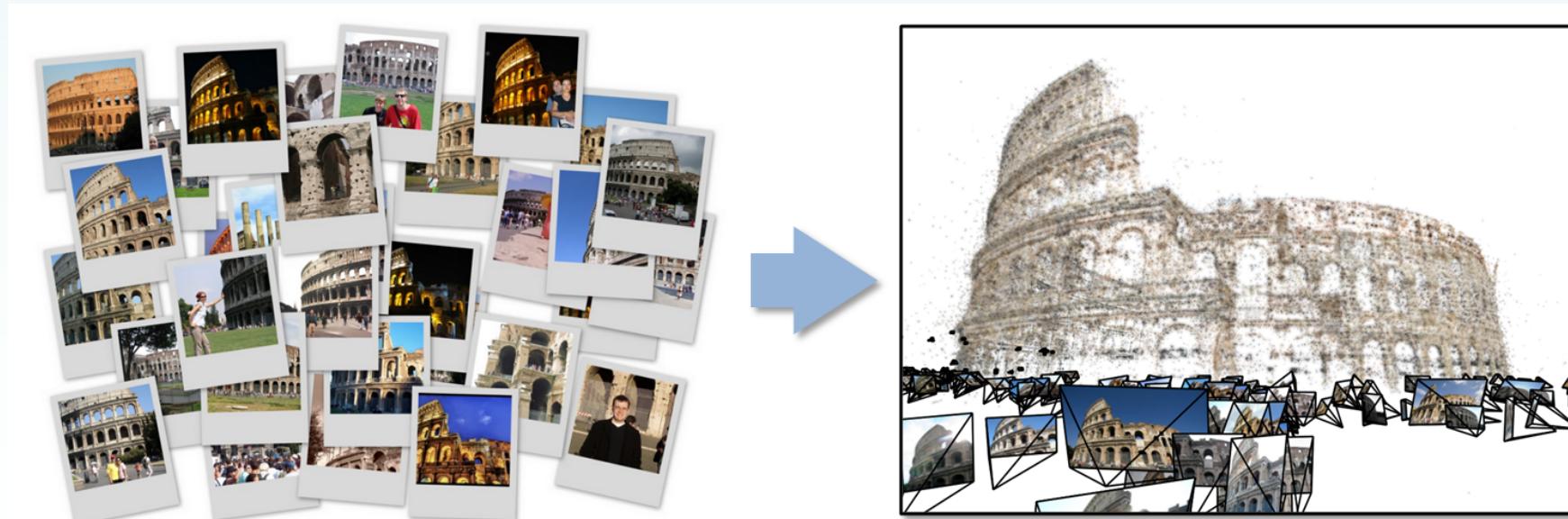


Visual Odometry vs Visual SLAM vs Structure from Motion



SfM from Unordered Internet Images

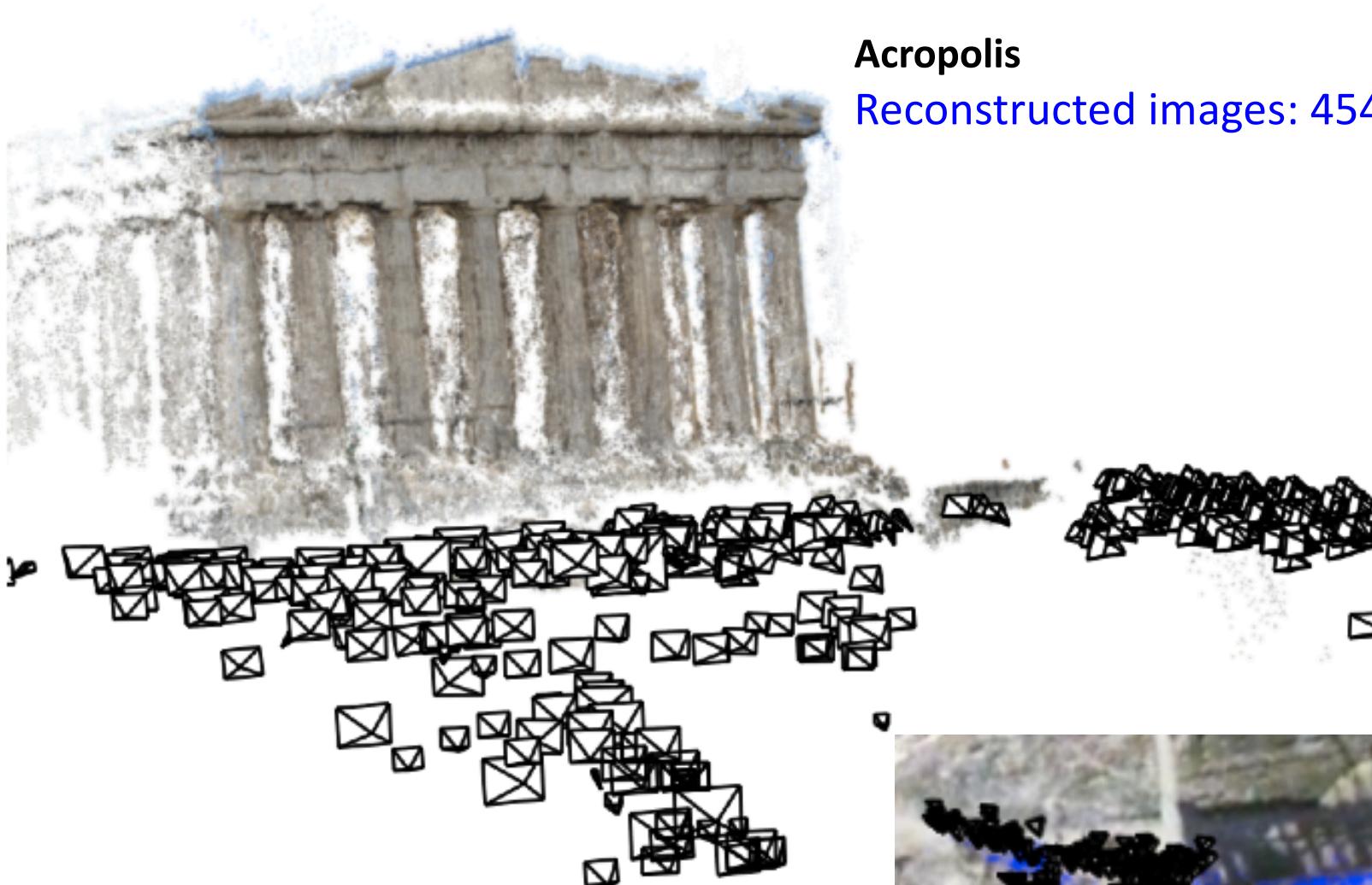
- Building 3D models from large, unordered image collections
 - [Snavely06], [Li08], [Agarwal09], [Frahm10], Microsoft's PhotoSynth, ...



- SfM is a key part of these reconstruction pipelines

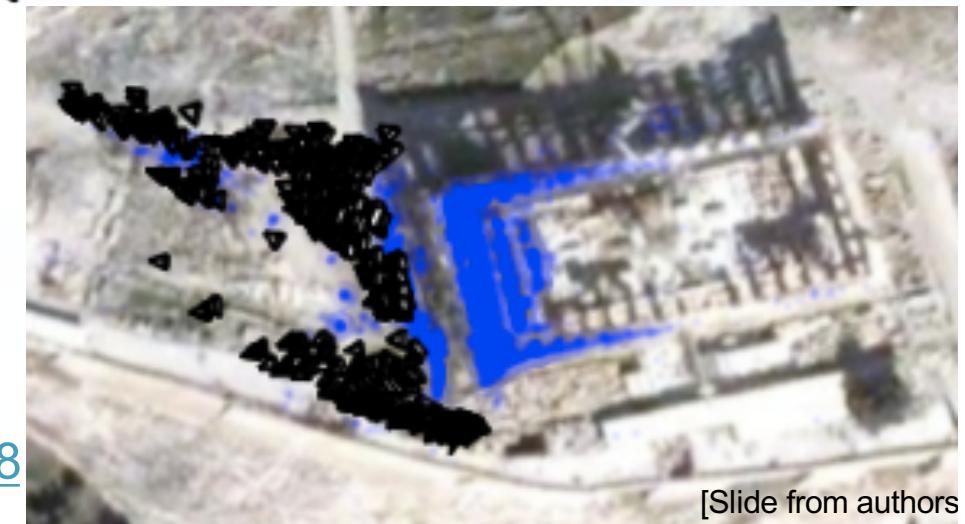
<https://www.cs.cornell.edu/projects/bigsfm/>

[Slide from authors]



Acropolis

Reconstructed images: 454



Videos

<http://youtu.be/kxtQqYLRaSQ>

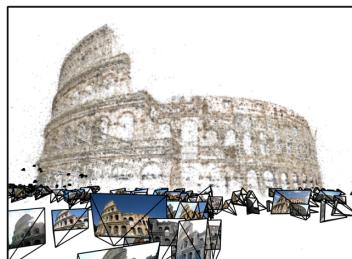
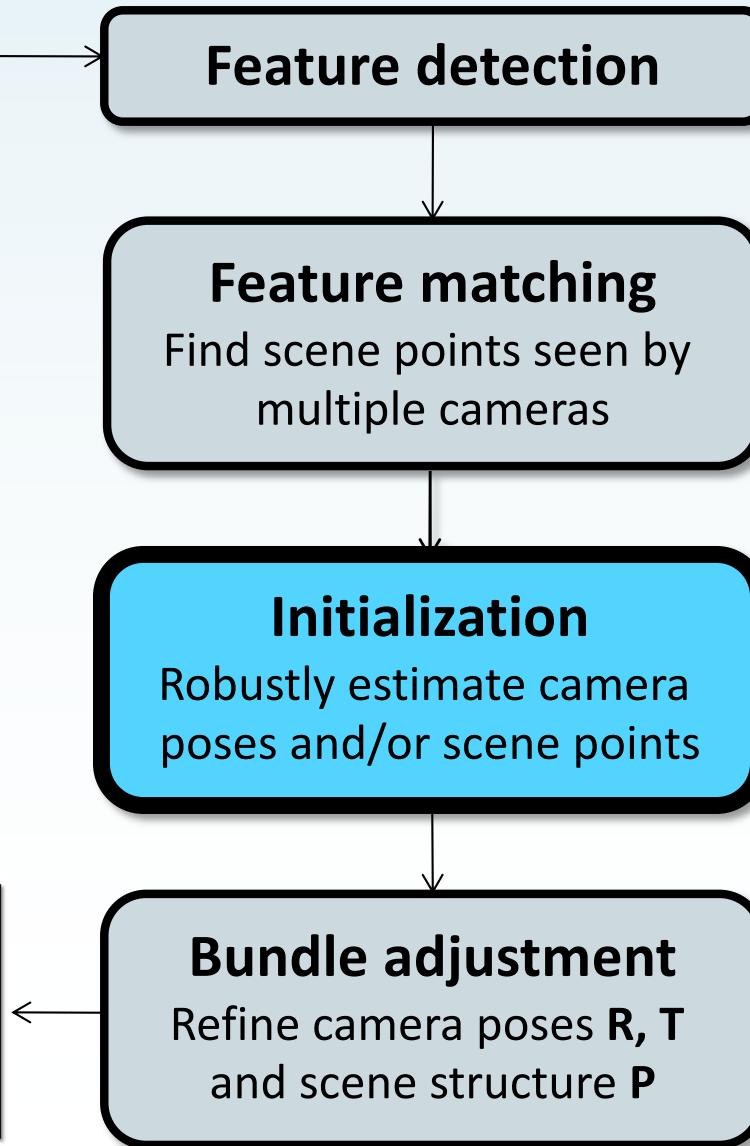
<https://www.youtube.com/watch?v=i7ierVkXYa8>

[Slide from authors]

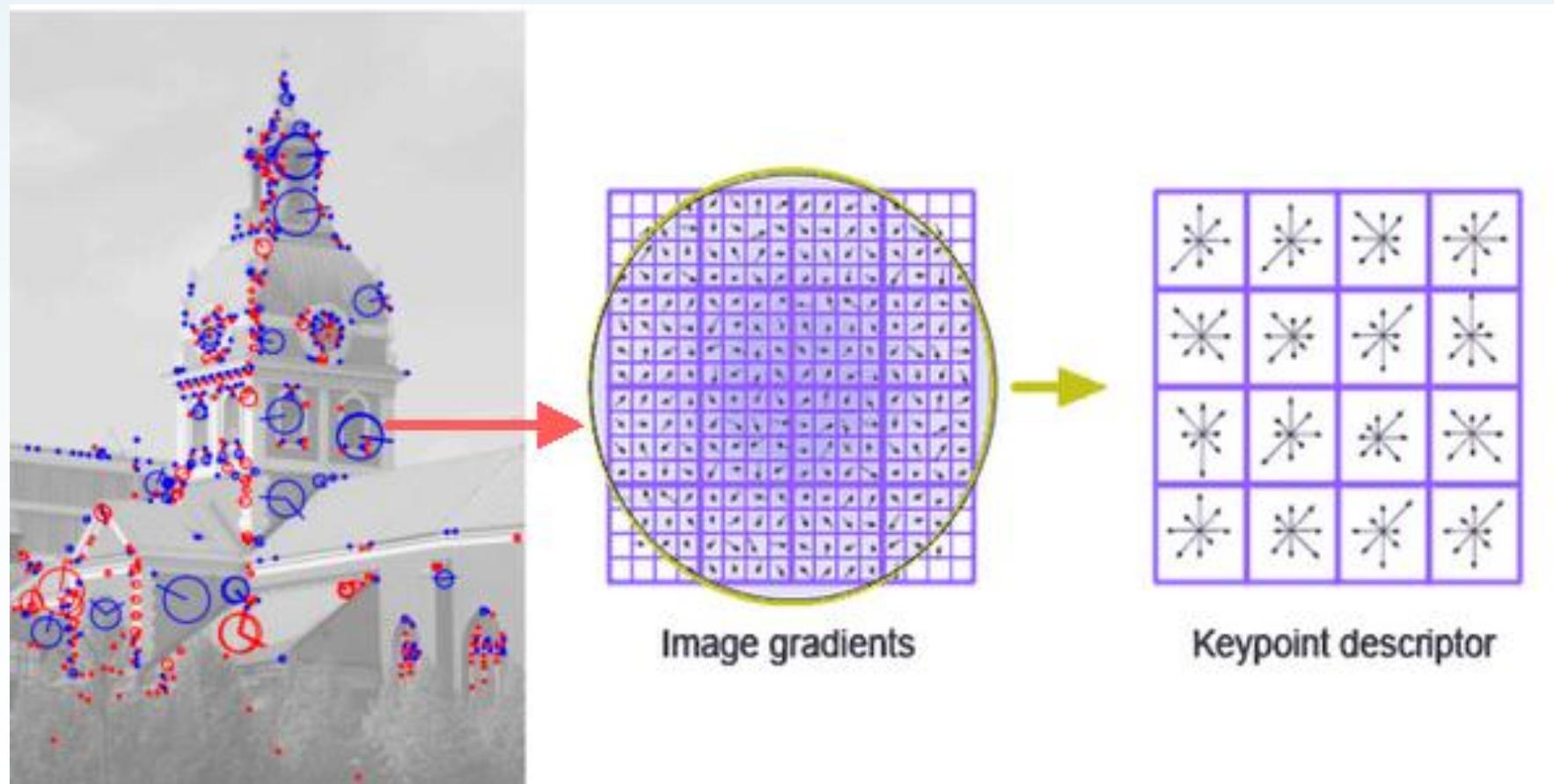
Reconstruction pipeline



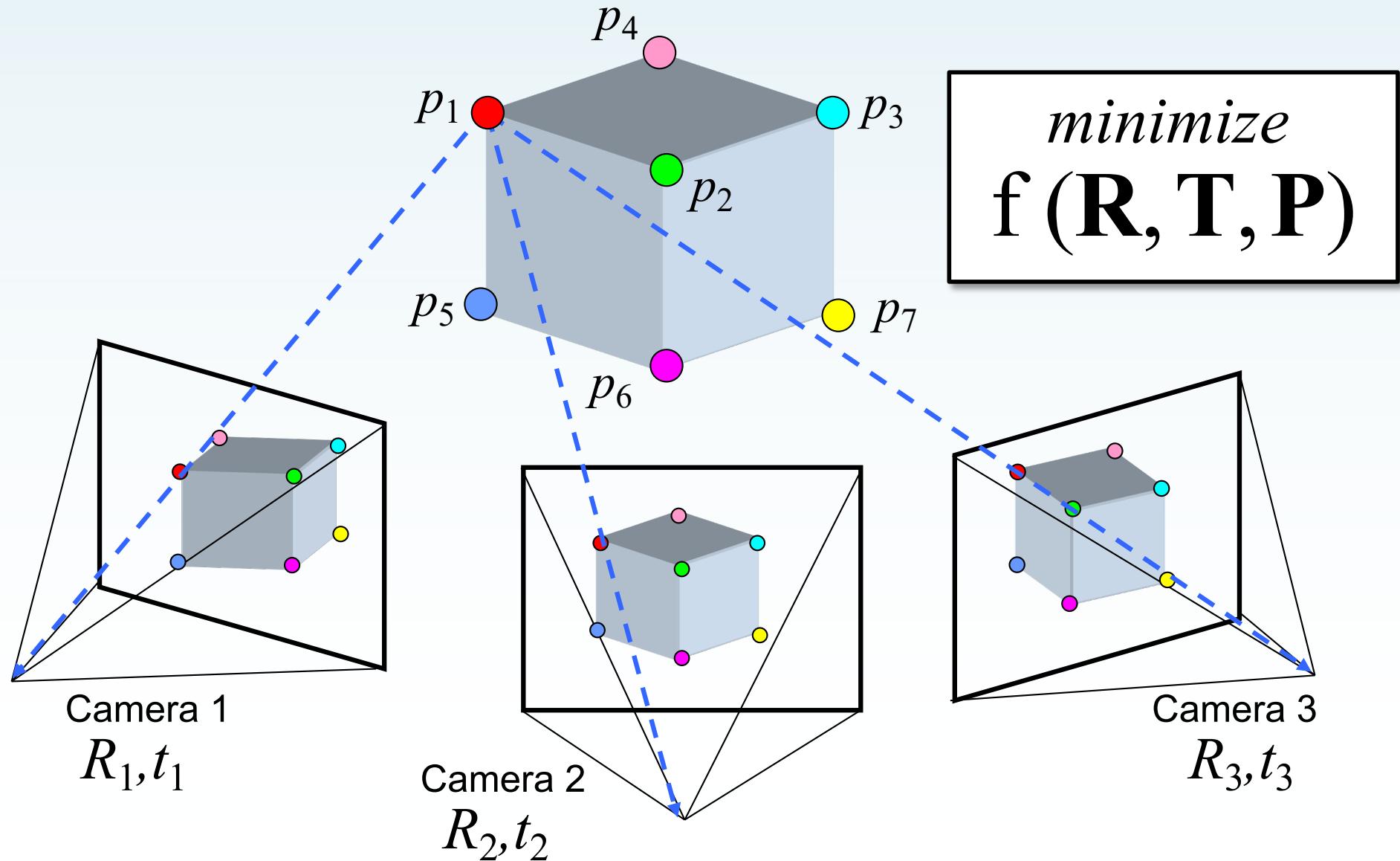
<https://www.cs.cornell.edu/projects/bigsfm/>



SIFT Features



Structure from Motion – Bundle Adjustment



Visual Odometry: Problem Formulation

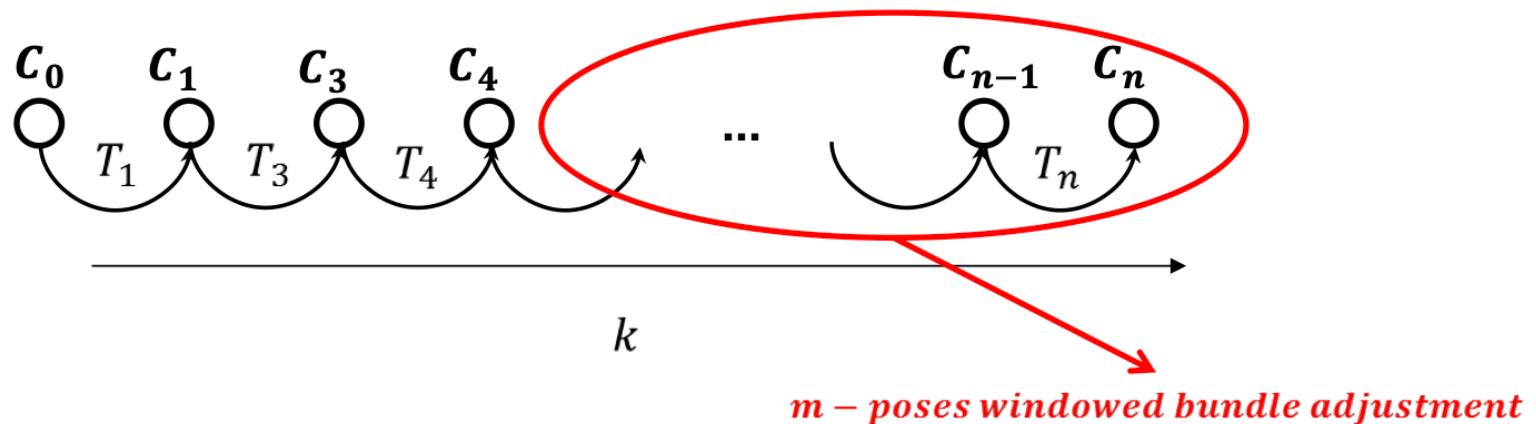
1. Compute the relative motion T_k from images I_{k-1} to image I_k

$$T_k = \begin{bmatrix} R_{k,k-1} & t_{k,k-1} \\ 0 & 1 \end{bmatrix}$$

2. Concatenate them to recover the full trajectory

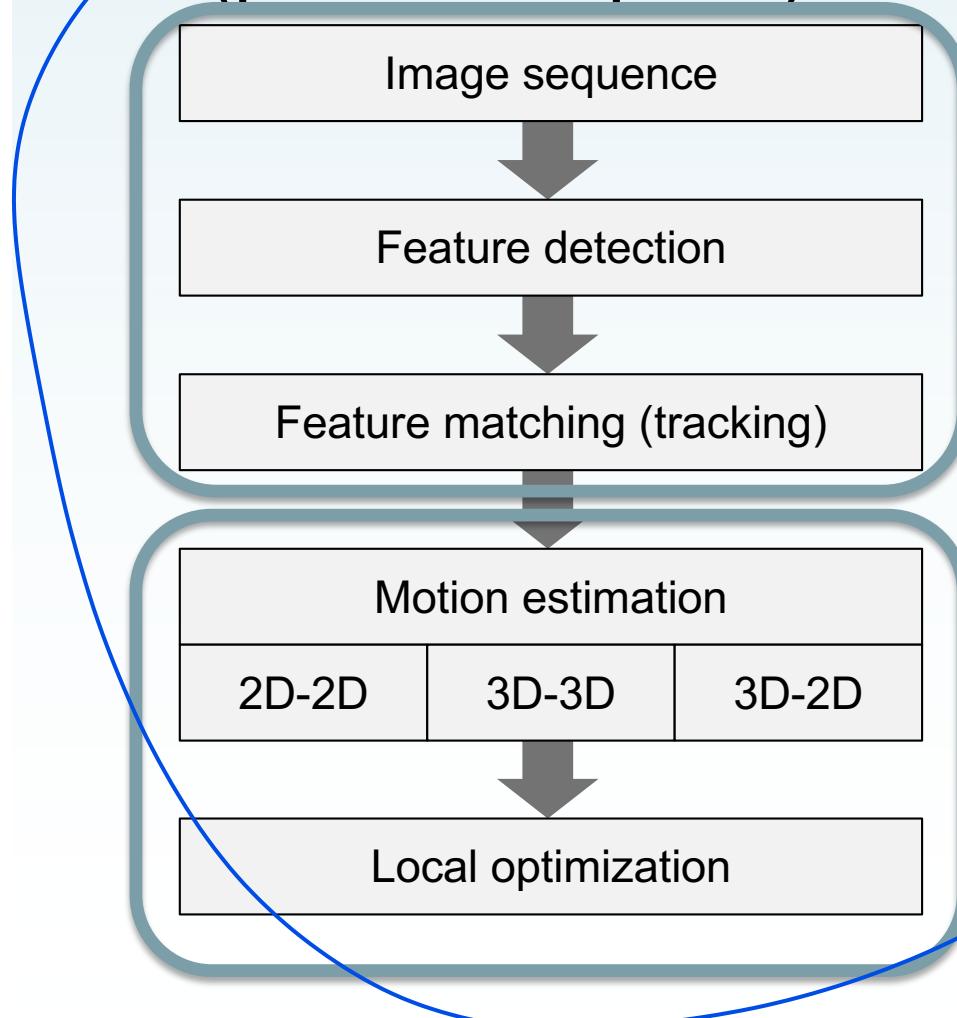
$$C_n = C_{n-1} T_n$$

3. An optimization over the last m poses can be done to refine locally the trajectory (Pose-Graph or Bundle Adjustment)



VO Flow Chart

- VO computes the camera path incrementally (pose after pose)



Front
End

SIFT features tracks

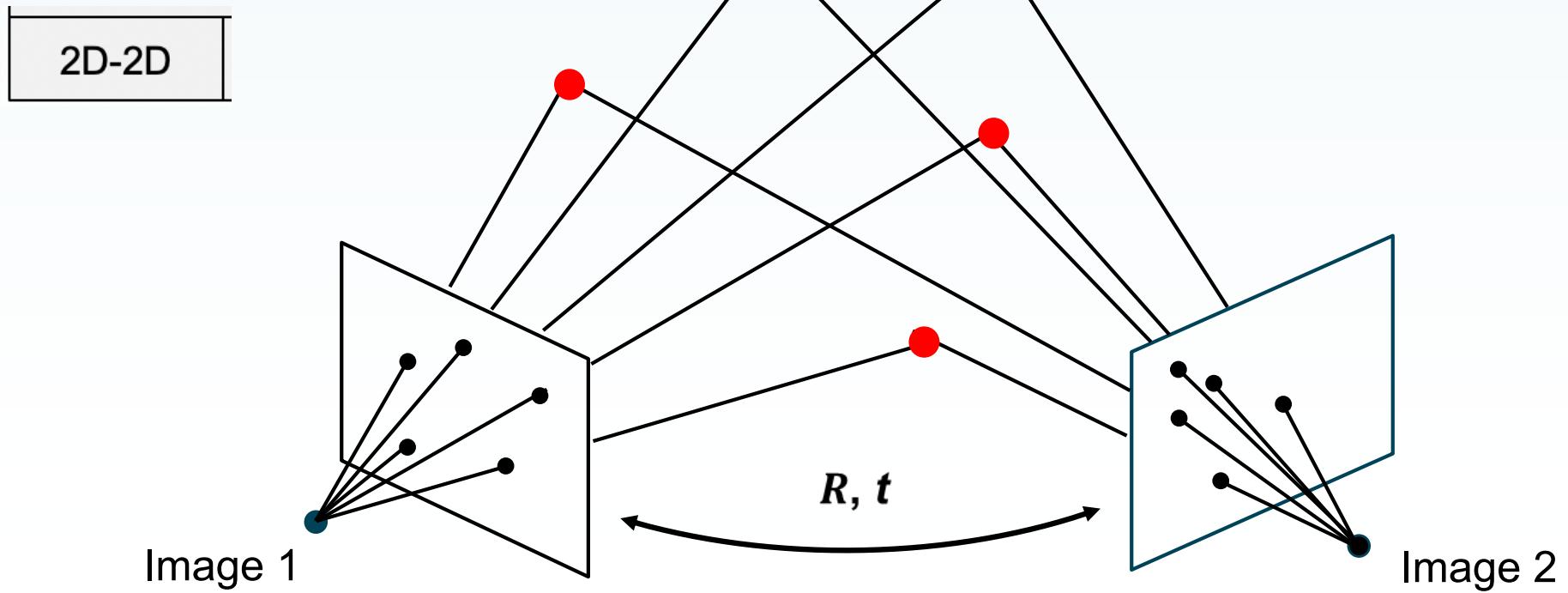
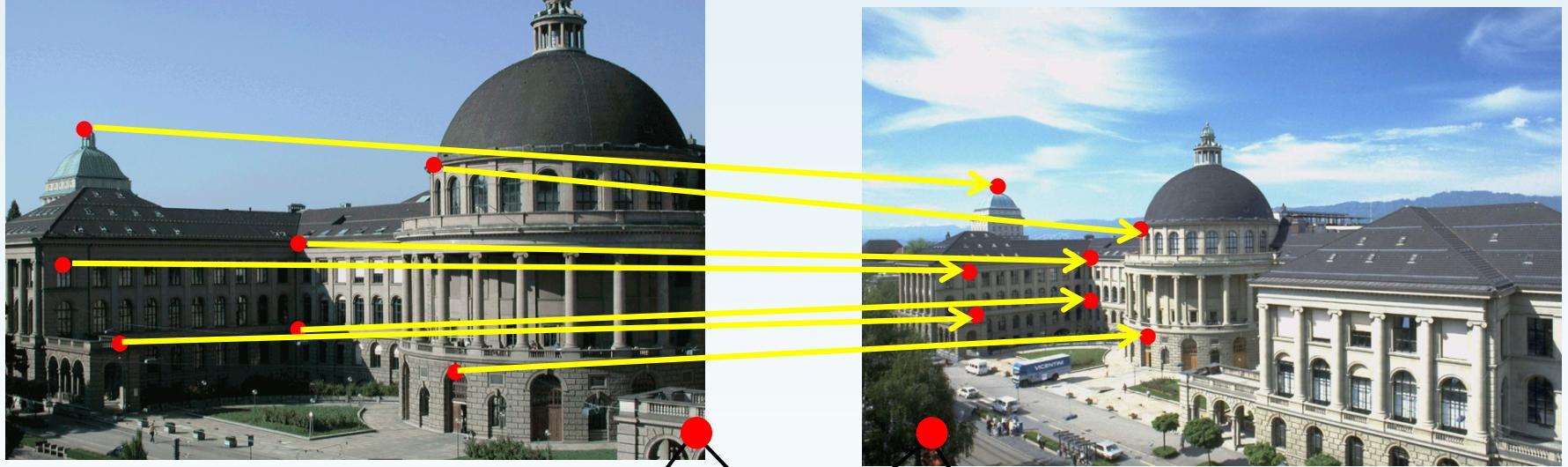
Back
End

Front-End vs Back-End in VO/SLAM

- The *front-end* is responsible for:
 - Feature extraction, matching and outlier rejection
 - Loop closure detection

- The *back-end* is responsible for:
 - Camera pose + 3D structure optimization (e.g. factor-graph, Kalman filter, bundle-adjustment)

Relative motion estimation: 2-view geometry



2D-to-2D

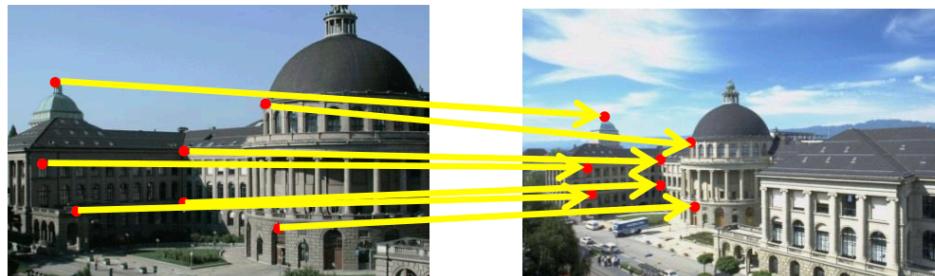
Motion estimation		
2D-2D	3D-2D	3D-3D

Motion from Image Feature Correspondences

- Both feature points f_{k-1} and f_k are specified in 2D
- The minimal-case solution involves **5-point** correspondences
- The solution is found by minimizing the reprojection error:

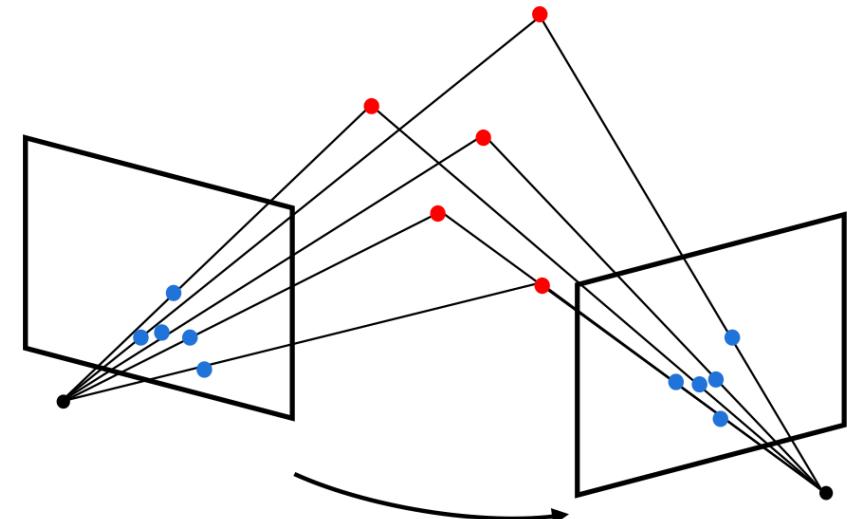
$$T_k = \begin{bmatrix} R_{k,k-1} & t_{k,k-1} \\ 0 & 1 \end{bmatrix} = \arg \min_{X^i, C_k} \sum_{i,k} \|p_k^i - g(X^i, C_k)\|^2$$

- Popular algorithms: 8- and 5-point algorithms [Hartley'97, Nister'06]



I_{k-1}

I_k



3D-to-2D

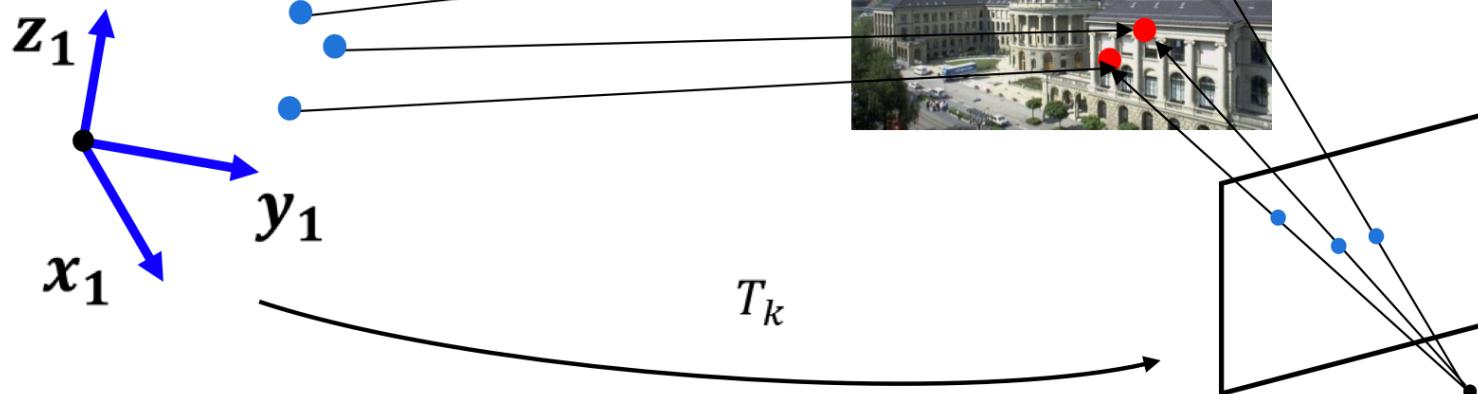
Motion estimation		
2D-2D	3D-2D	3D-3D

Motion from 3D Structure and Image Correspondences

- f_{k-1} is specified in 3D and f_k in 2D
- This problem is known as *camera resection* or PnP (perspective from n points)
- The minimal-case solution involves **3 correspondences** (+1 for disambiguating the 4 solutions)
- The solution is found by minimizing the reprojection error:

$$T_k = \begin{bmatrix} R_{k,k-1} & t_{k,k-1} \\ 0 & 1 \end{bmatrix} = \arg \min_{T_k} \sum_i \|p_k^i - \hat{p}_{k-1}^i\|^2$$

- Popular algorithms: **P3P** [Gao'03, Kneip'11]



3D-to-3D

Motion estimation

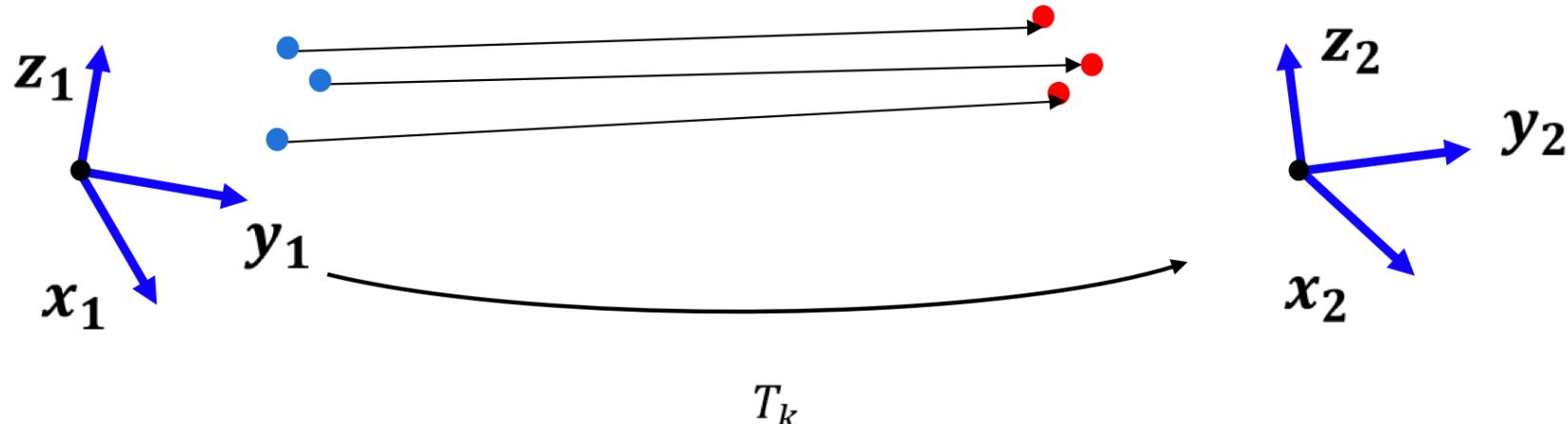
2D-2D 3D-2D 3D-3D

Motion from 3D-3D Point Correspondences (point cloud registration)

- Both f_{k-1} and f_k are specified in 3D. To do this, it is necessary to triangulate 3D points (e.g. use a stereo camera)
- The minimal-case solution involves **3 non-collinear correspondences**
- The solution is found by minimizing the 3D-3D Euclidean distance:

$$T_k = \begin{bmatrix} R_{k,k-1} & t_{k,k-1} \\ 0 & 1 \end{bmatrix} = \arg \min_{X^i, C_k} \sum_{i,k} \|p_k^i - g(X^i, C_k)\|^2$$

- Popular algorithm: [Arun'87] for global registration, ICP for local refinement or Bundle Adjustment (BA)

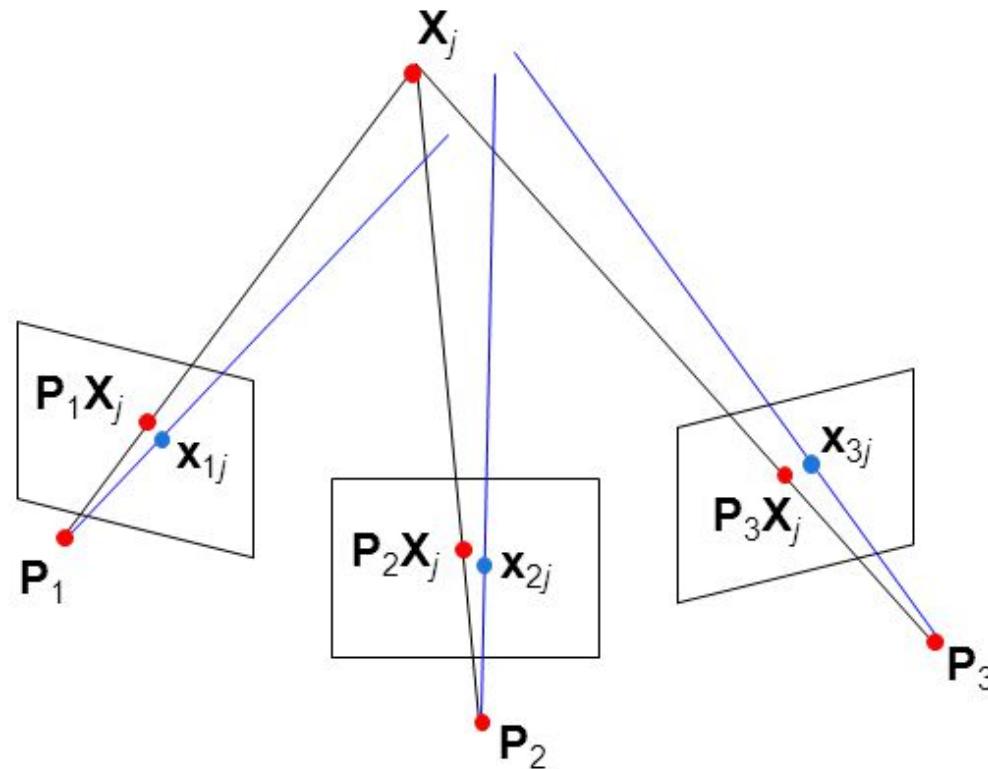


T_k

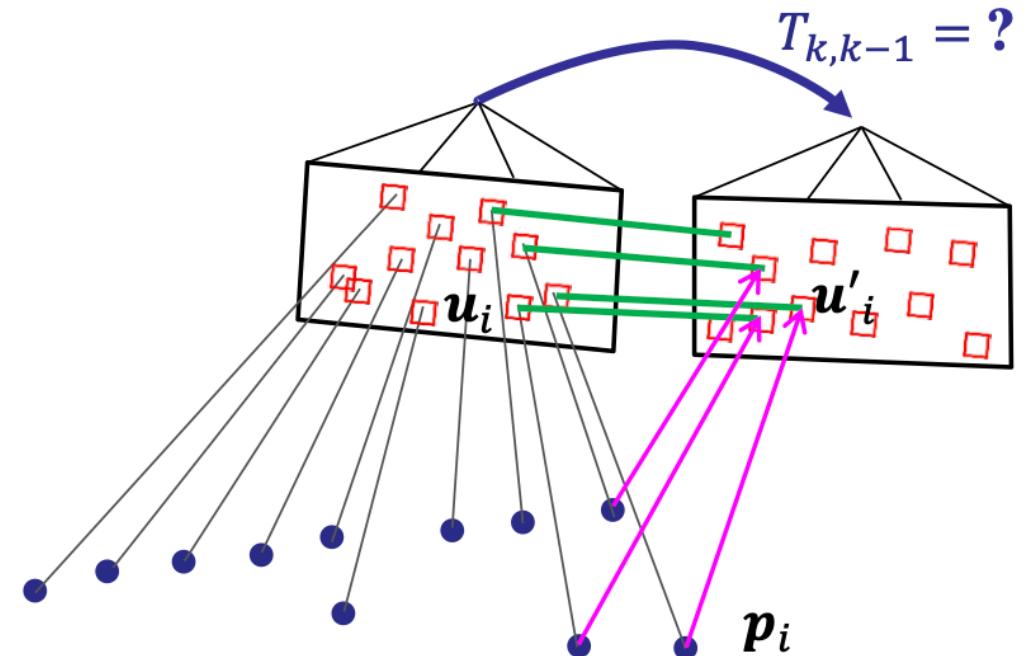
Bundle adjustment

- Non-linear method for refining structure and motion
- Minimizing reprojection error

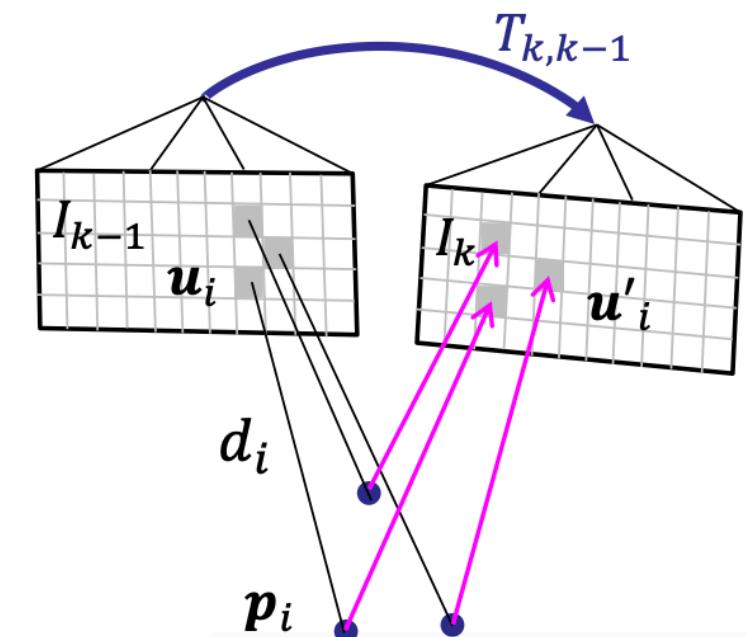
$$E(\mathbf{P}, \mathbf{X}) = \sum_{i=1}^m \sum_{j=1}^n D(\mathbf{x}_{ij}, \mathbf{P}_i \mathbf{X}_j)^2$$



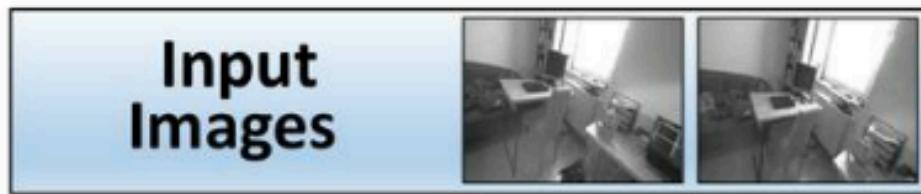
Feature-based methods



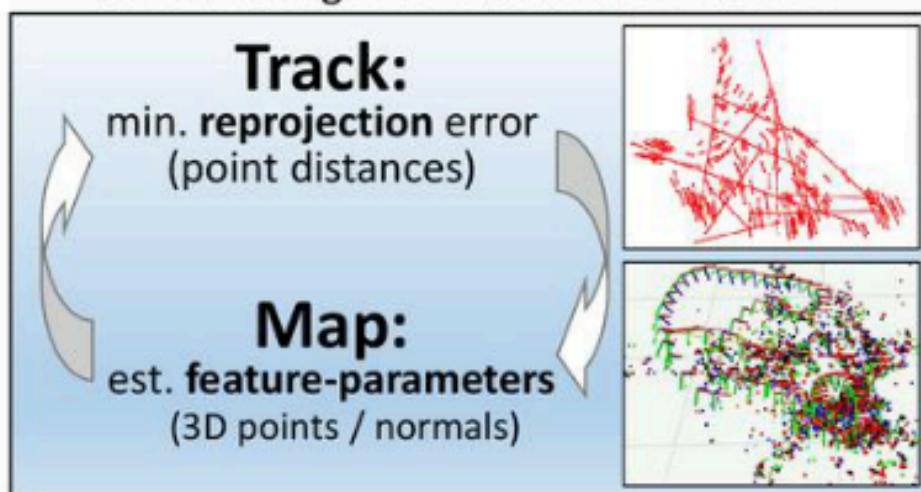
Direct methods



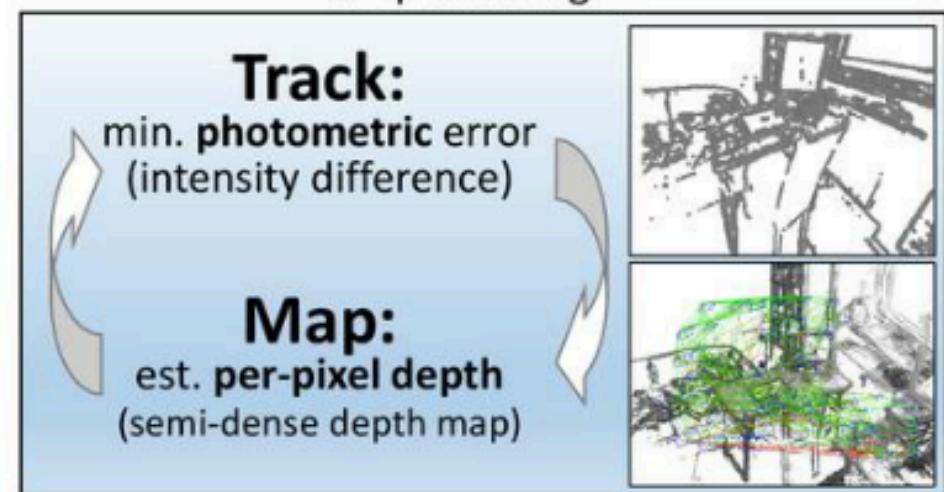
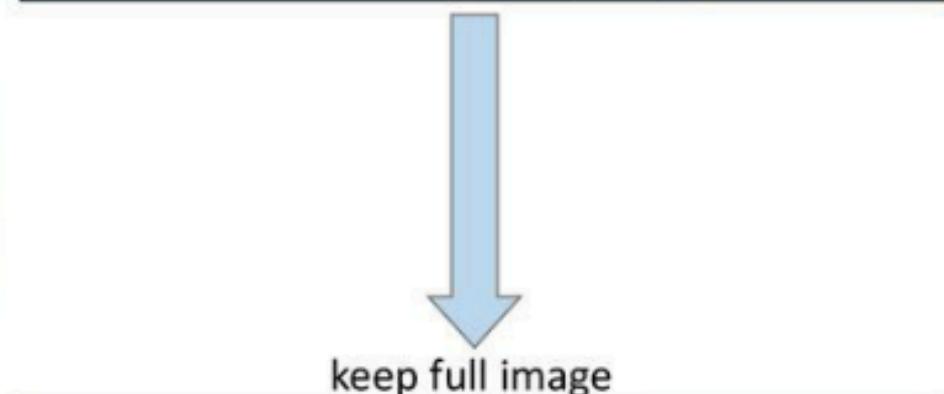
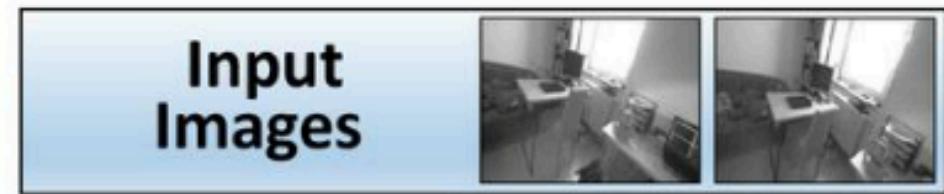
Keypoint-Based



abstract images to feature observations



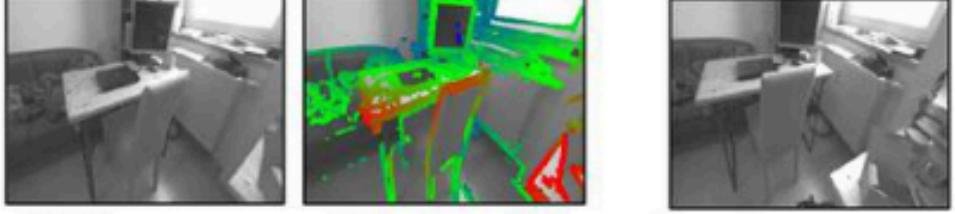
Direct Dense SLAM



Direct Methods minimise photometric cost

$$E(\xi) = \sum_{\mathbf{x} \in \Omega_{\text{KF}}} (I_{\text{KF}}(\mathbf{x}) - \underbrace{I(\omega(\mathbf{x}, D_{\text{KF}}(\mathbf{x}), \xi)))^2}_{\|\mathbf{r}(\xi)\|_2^2}$$

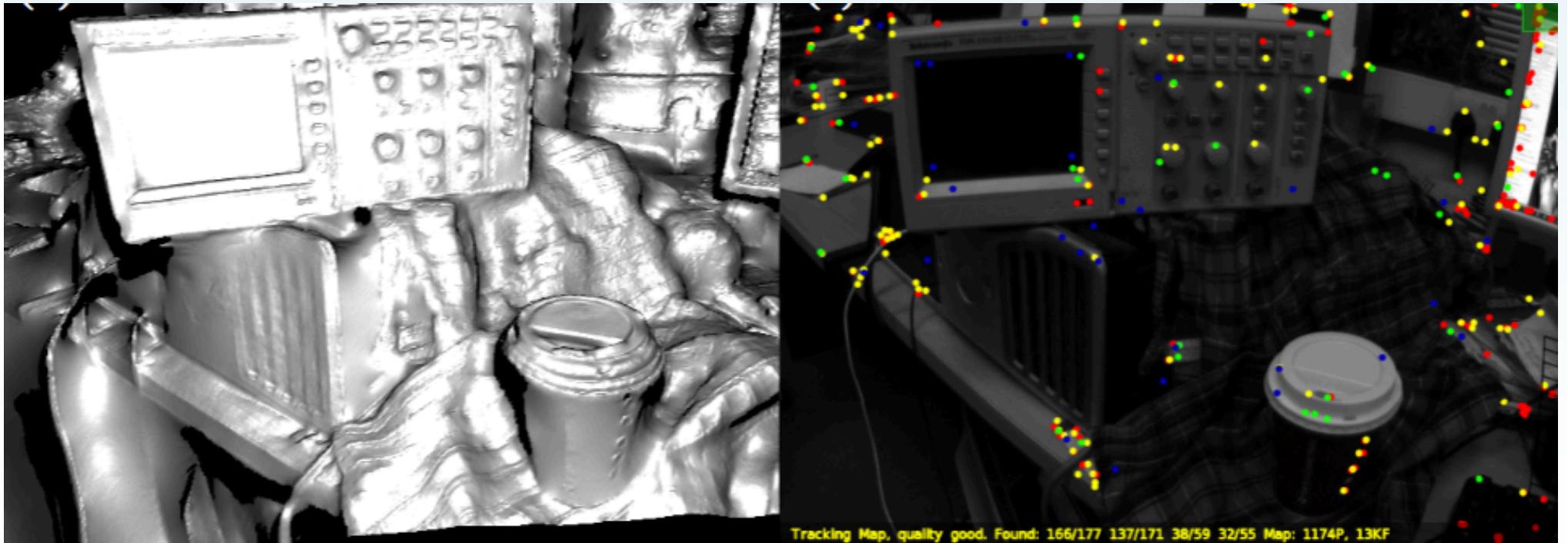
↑
Camera Pose
in $\mathfrak{se}(3)$



KF image **KF depth** **back-warped
new frame**

- minimize using **Gauss-Newton Algorithm**
(≈ forward-compositional Lucas-Kanade)

Why direct methods?



DTAM– Dense
Newcombe et al.
ICCV'11

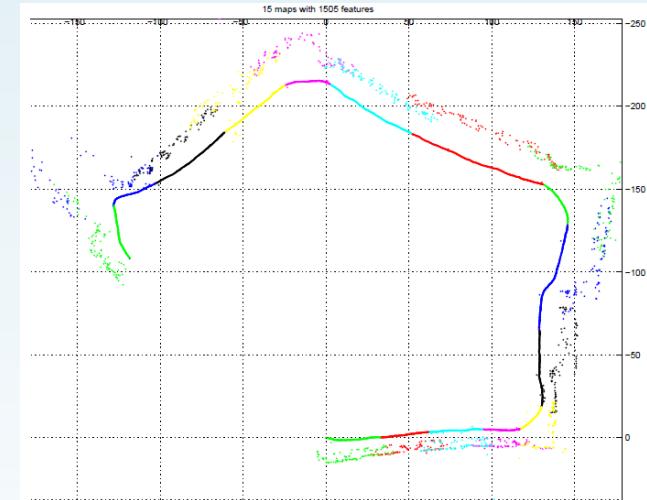
PTAM– Sparse
Klein and Murray,
ISMAR'07

Visual Odometry vs. Visual SLAM

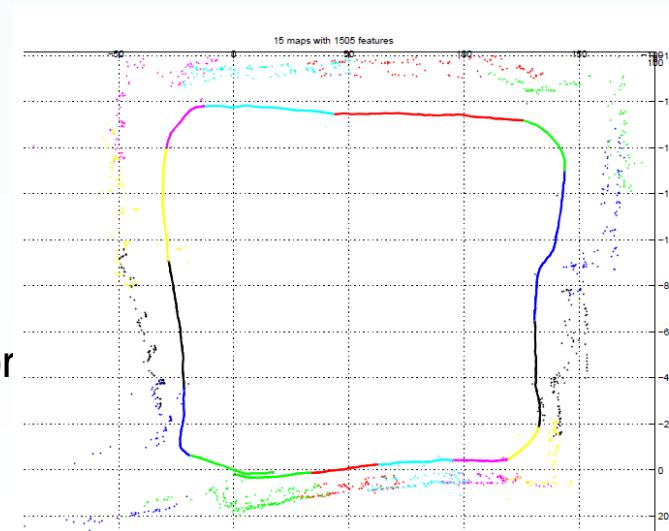
- Main difference between Visual Odometry (VO) and Visual SLAM (V-SLAM) is that while VO cares mostly about **local** estimation of the **camera trajectory**, V-SLAM also estimates an accurate globally consistent map.
- VO focuses on estimating the 3D motion of the camera (trajectory) sequentially (as a new frame arrives) and in real time.
- Instead, V-SLAM also focuses on estimating a map of the 3D environment.
- Bundle adjustment can be used in VO to refine the local estimate of the trajectory, but it is "optional".
- Often said that the difference between VO and V-SLAM is that V-SLAM incorporates loop closure.
- Both are strongly related to Structure from Motion (SFM)

VO vs. Visual SLAM (2/2)

- VO only aims to the local consistency of the trajectory
- SLAM aims to the global consistency of the trajectory and of the map
- VO can be used as a building block of SLAM
- VO is SLAM before closing the loop!
- The choice between VO and V-SLAM depends on the tradeoff between performance and consistency, and simplicity in implementation.
- VO trades off consistency for real-time performance, without the need to keep track of all the previous history of the camera.



Visual odometry



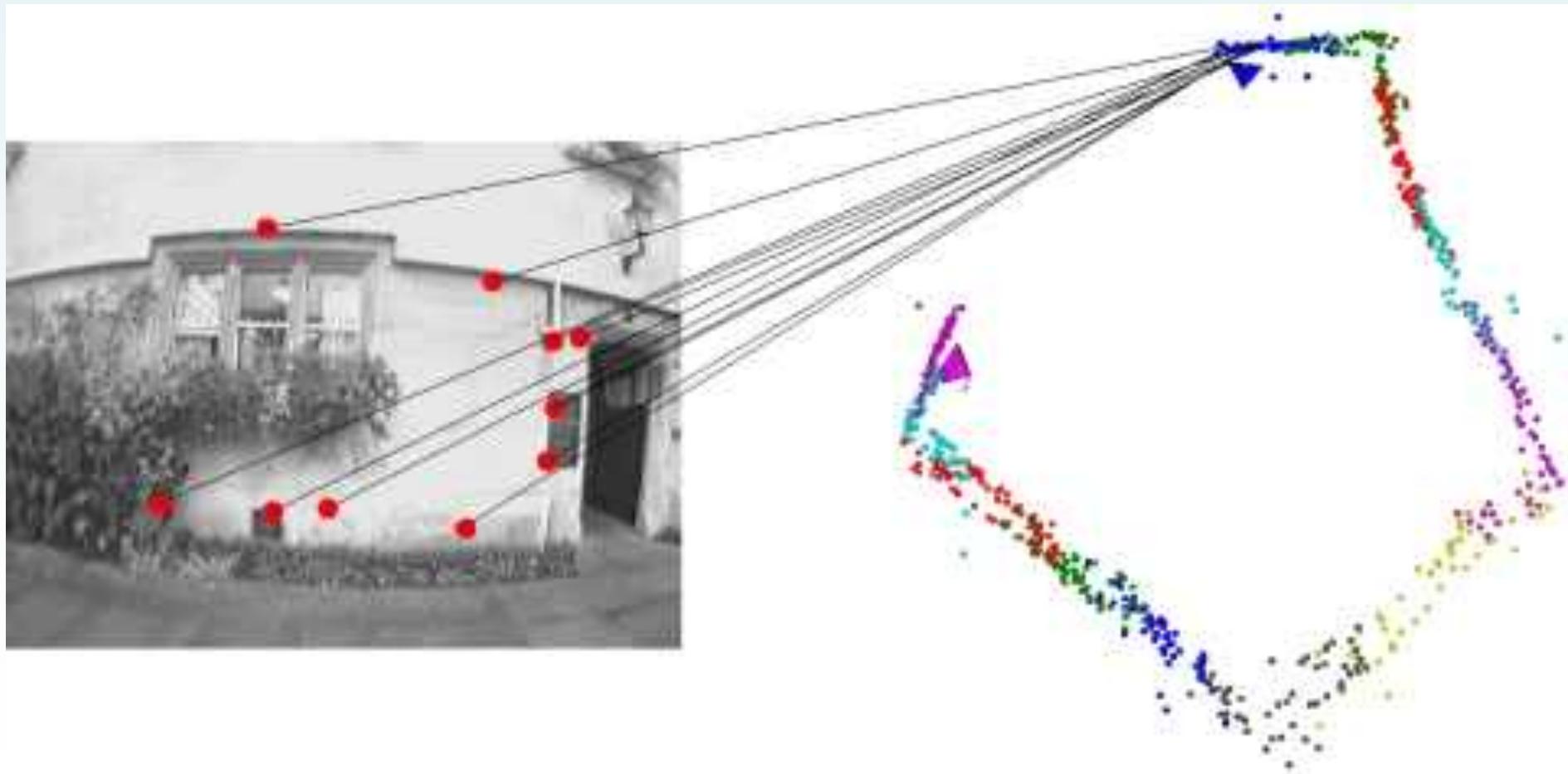
Visual SLAM

Image courtesy of Clemente et al. RSS'07

Example Loop Closing Problem



Example Loop Closing Problem



Example Loop Closing Problem

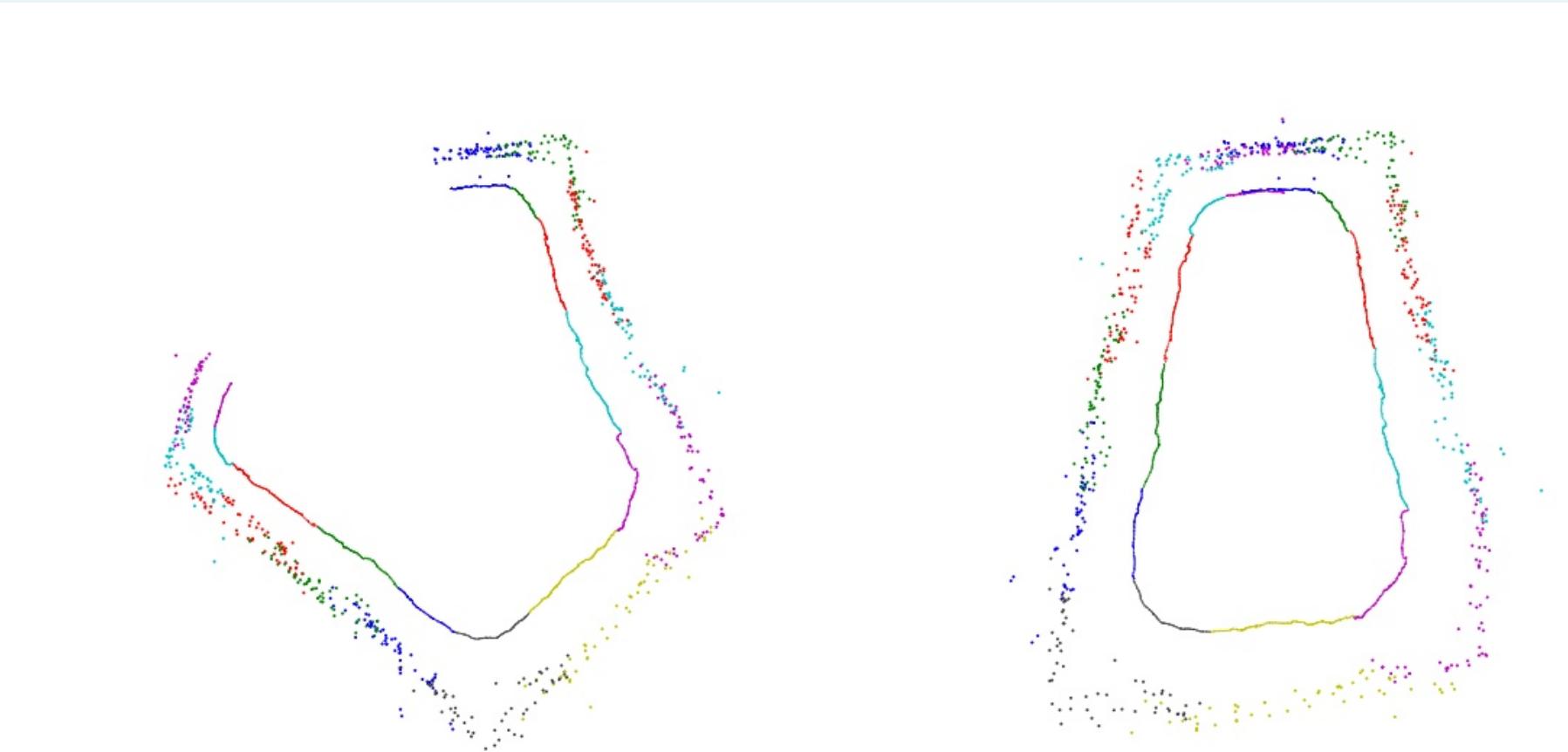


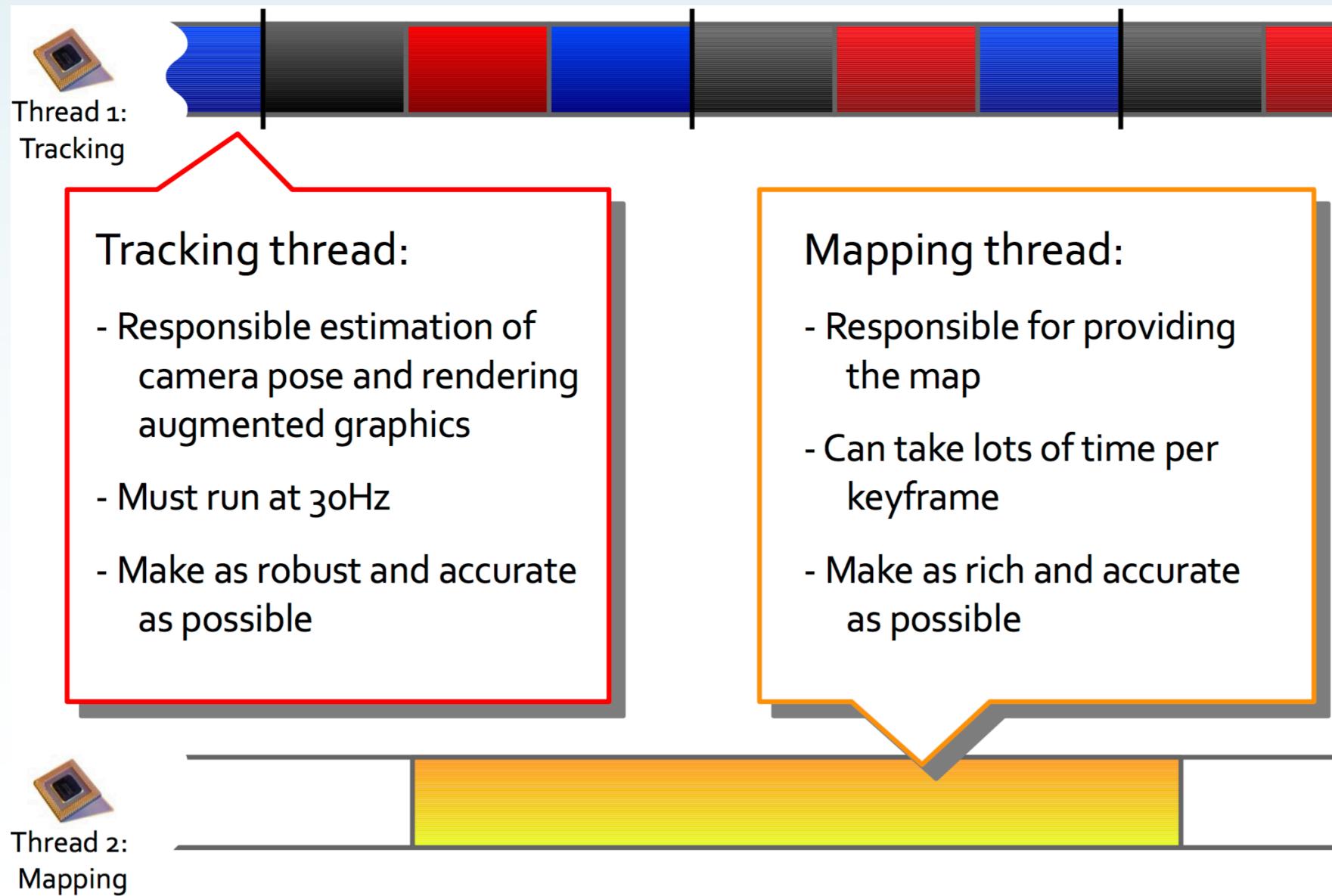
Table 1 List of different visual SLAM system. Non-filter-based approaches are highlighted in a gray color.

Year	Name	Method	Type	Reference
2003	Real-time simultaneous localization and mapping with a single camera	filter	indirect	Davison (2003)
2004	Simultaneous localization and mapping using multiple view feature descriptors	filter	indirect	Meltzer et al (2004)
2004	Real-Time 3D SLAM with Wide-Angle Vision	filter	indirect	Davison et al (2004)
2005	Real-Time Camera Tracking Using a Particle Filter	filter	indirect	Pupilli and Calway (2005)
2005	CV-SLAM	filter	indirect	Jeong and Lee (2005)
2006	Real-time Localization and 3D Reconstruction	non-filter	indirect	Mouragnon et al (2006)
2006	Scalable Monocular SLAM	filter	indirect	Eade and Drummond (2006)
2006	Real-Time Monocular SLAM with Straight Lines	filter	indirect	Smith et al (2006)
2007	Monocular-vision based SLAM using Line Segments	filter	indirect	Lemaire and Lacroix (2007)
2007	MonoSLAM	filter	indirect	Davison et al (2007)
2007	Parallel Tracking and Mapping (PTAM)	non-filter	indirect	Klein and Murray (2007)
2007	Monocular SLAM as a Graph of Coalesced Observations	filter	indirect	Eade and Drummond (2007)
2007	Mapping Large Loops with a Single Hand-Held Camera	filter	indirect	Clemente et al (2007)
2007	Dimensionless Monocular SLAM	filter	indirect	Civera et al (2007)
2008	A Square Root UKF for visual monoSLAM	filter	indirect	Holmes et al (2008)
2008	An Efficient Direct Approach to Visual SLAM	non-filter	direct	Silveira et al (2008)
2008	Efficient View-Based SLAM Using Visual Loop Closures	filter	indirect	Mahon et al (2008)
2008	Large-Scale SLAM Building Conditionally Independent Local Maps: Application to Monocular Vision	filter	indirect	Pinies and Tardos (2008)
2009	Towards a robust visual SLAM approach: Addressing the challenge of life-long operation	filter	indirect	Hochdorfer and Schlegel (2009)
2009	Use a Single Camera for Simultaneous Localization And Mapping with Mobile Object Tracking in dynamic environments	filter	indirect	Migliore et al (2009)
2010	On Combining Visual SLAM and Visual Odometry	filter	indirect	Williams and Reid (2010)
2010	Scale Drift-Aware Large Scale Monocular SLAM	non-filter	indirect	Strasdat et al (2010a)
2010	Live dense reconstruction with a single moving camera	non-filter	hybrid	Newcombe and Davison (2010)
2010	Monocular SLAM with locally planar landmarks via geometric rao-blackwellized particle filtering on Lie groups	filter	indirect	Kwon and Lee (2010)

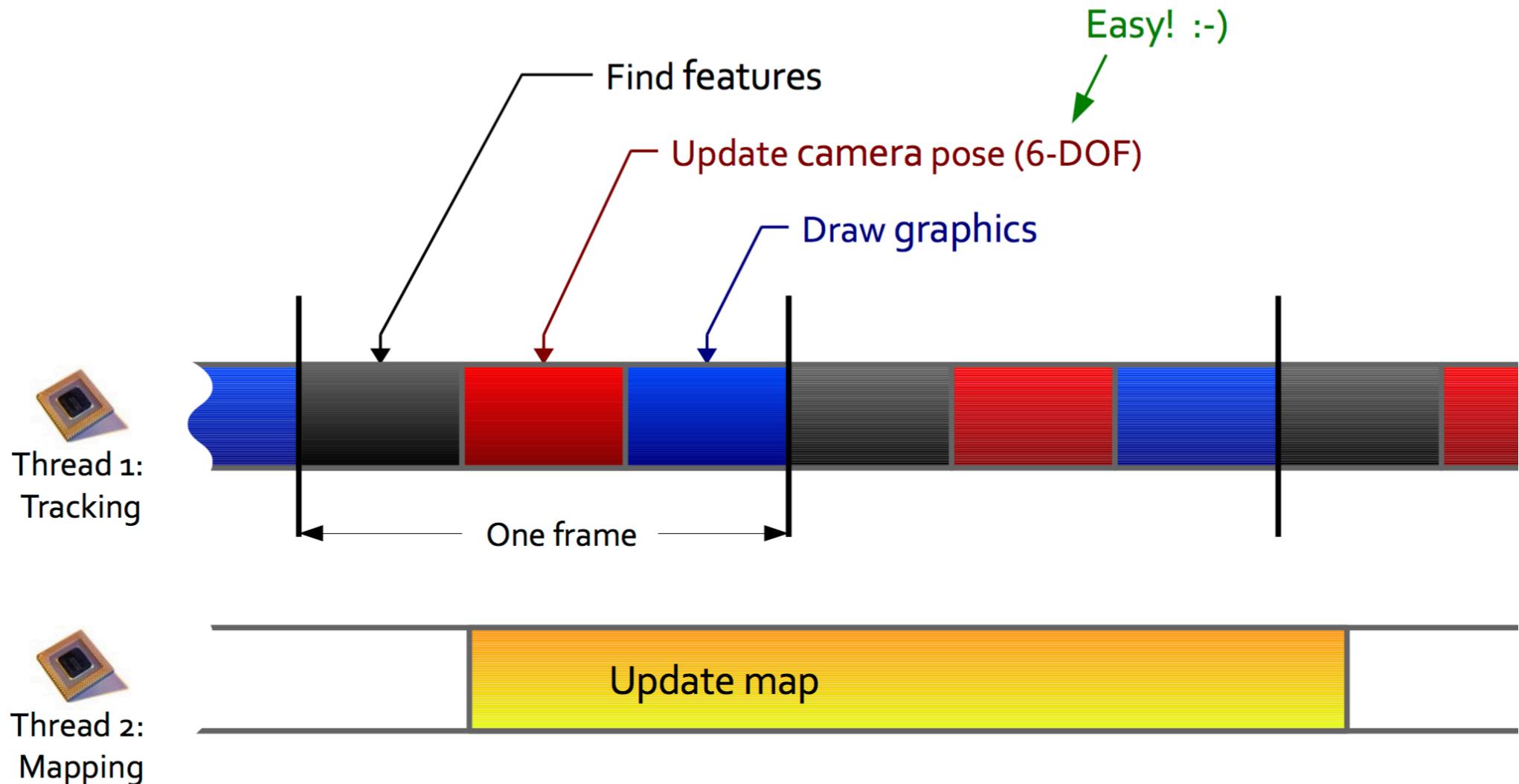
Visual SLAM

2011	Dense Tracking and Mapping (DTAM)	non-filter	direct	Newcombe et al (2011)
2011	Omnidirectional dense large-scale mapping and navigation based on meaningful triangulation	non-filter	direct	Pretto et al (2011)
2011	Continuous localization and mapping in a dynamic world (CD SLAM)	non-filter	indirect	Pirker et al (2011)
2011	Online environment mapping	non-filter	indirect	Lim et al (2011)
2011	Homography-based planar mapping and tracking for mobile phones	non-filter	indirect	Pirchheim and Reitmayr (2011)
2013	Robust monocular SLAM in Dynamic environments (RD SLAM)	non-filter	indirect	Tan et al (2013)
2013	Handling pure camera rotation in keyframe-based SLAM (Hybrid SLAM)	non-filter	indirect	Pirchheim et al (2013)
2013	Monocular Vision SLAM for Indoor Aerial Vehicles	filter	indirect	Celik and Somanı (2013)
2014	Visual SLAM for Handheld Monocular Endoscope	filter	indirect	Grasa et al (2014)
2014	Real-time camera tracking using a particle filter combined with unscented Kalman filters	filter	indirect	Lee (2014)
2014	Semi-direct Visual Odometry (SVO)	non-filter	hybrid	Forster et al (2014)
2014	Large Scale Direct monocular SLAM (LSD SLAM)	non-filter	direct	Engel et al (2014)
2014	Deferred Triangulation SLAM (DT SLAM)	non-filter	indirect	Herrera et al (2014)
2014	Real-Time 6-DOF Monocular Visual SLAM in a Large Scale Environment	non-filter	indirect	Lim et al (2014)
2015	StructSLAM: Visual SLAM With Building Structure Lines	filter	indirect	Zhou et al (2015)
2015	Robust large scale monocular Visual SLAM	non-filter	indirect	Bourmaud and Megret (2015)
2015	ORB SLAM	non-filter	indirect	Mur-Artal et al (2015)
2015	Dense Piecewise Parallel Tracking and Mapping (DPPTAM)	non-filter	direct	Concha and Civera (2015)

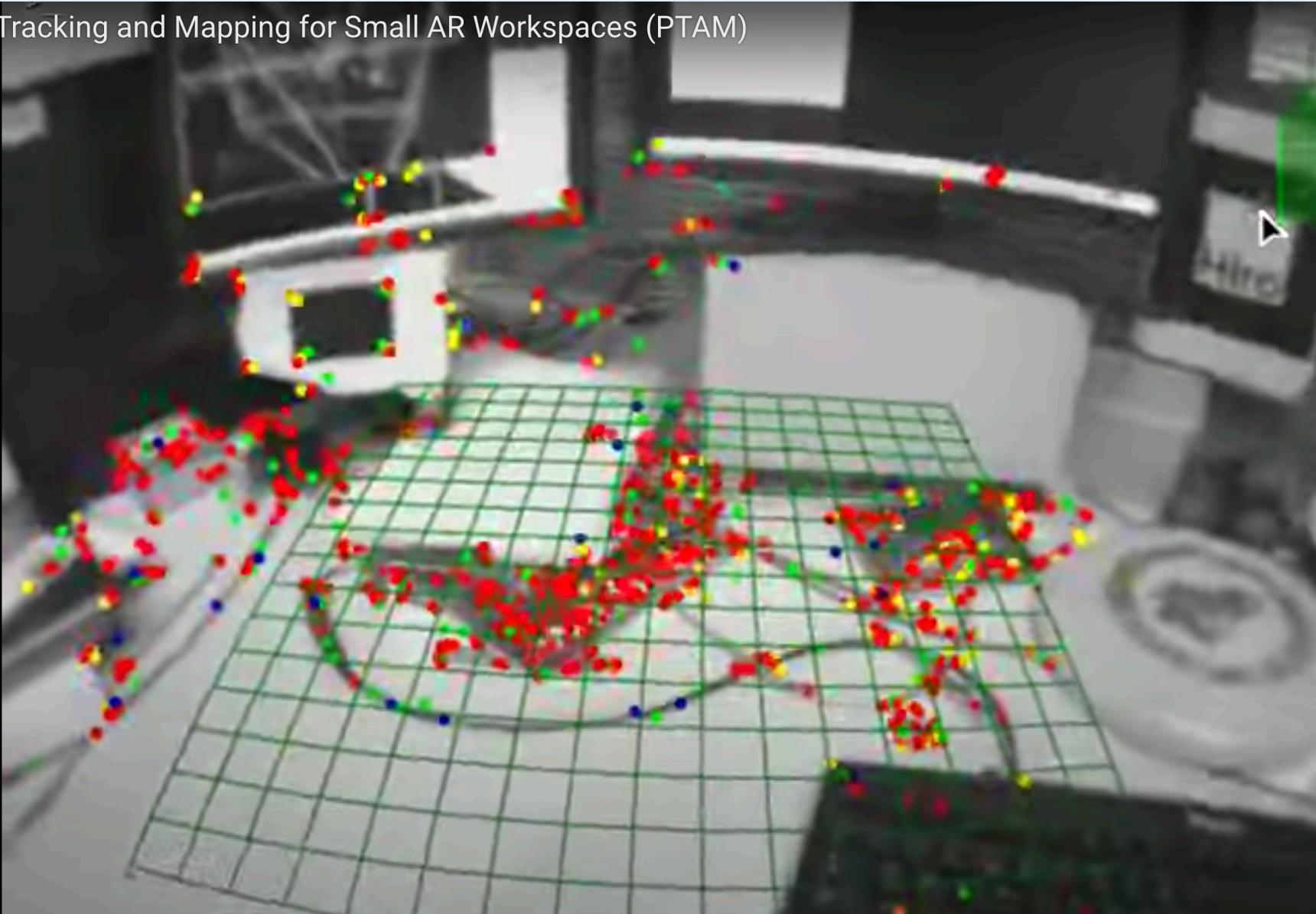
PTAM



PTAM

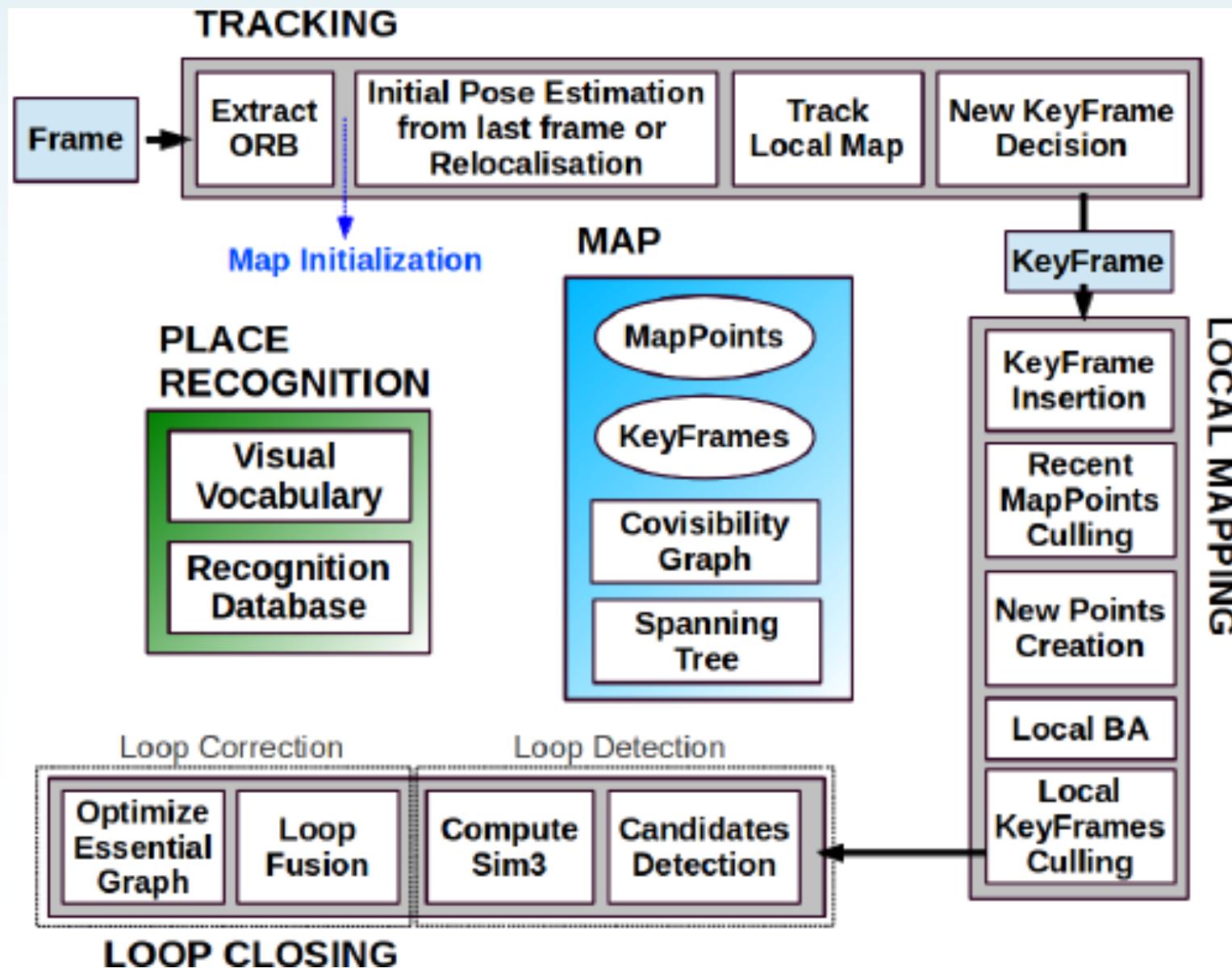


Parallel Tracking and Mapping for Small AR Workspaces (PTAM)



<https://www.youtube.com/watch?v=F3s3M0mokNc>

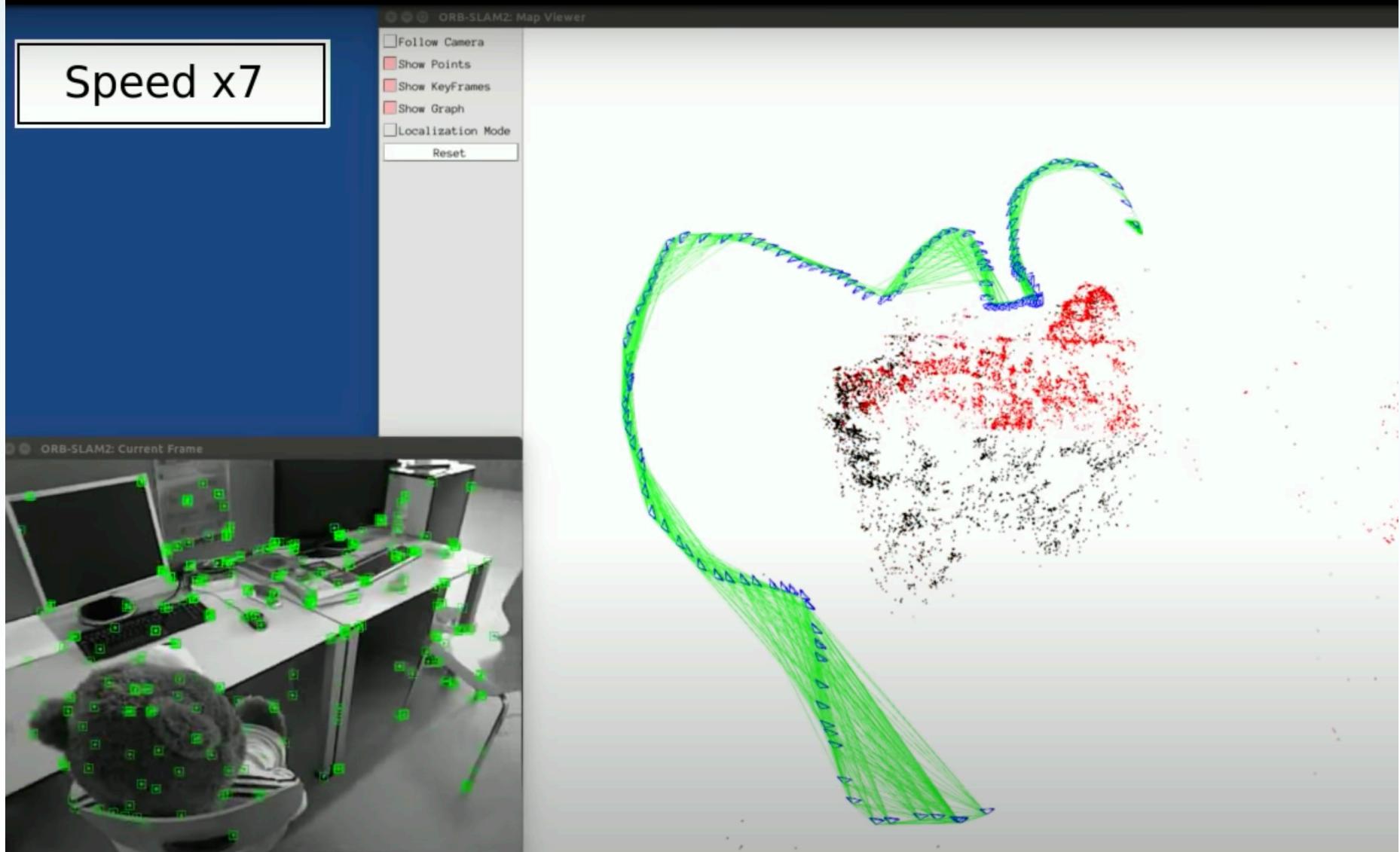
ORB-SLAM

<https://www.youtube.com/watch?v=ufvPS5wJAx0>

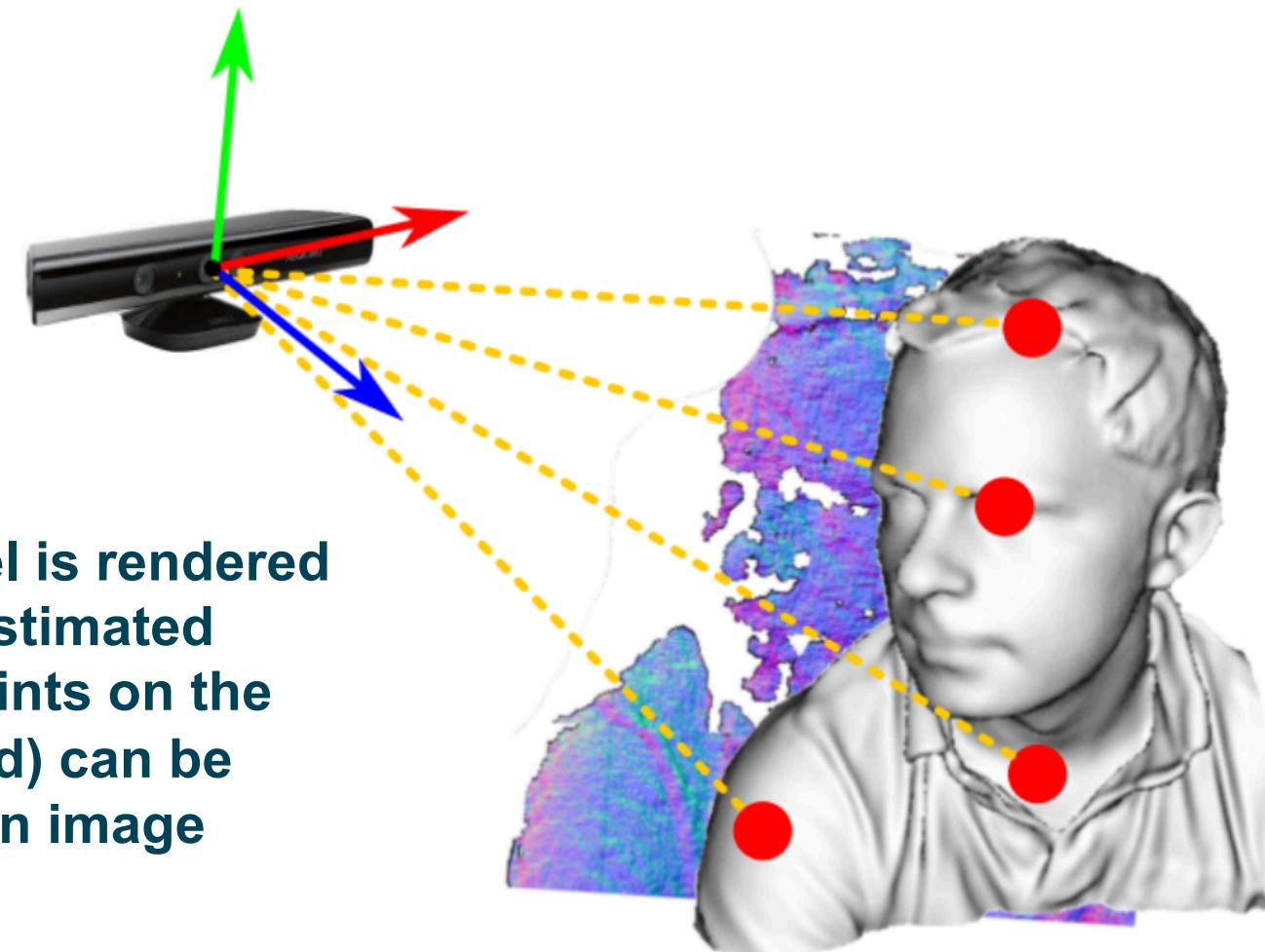
ORB-SLAM

<https://www.youtube.com/watch?v=ufvPS5wJAx0>

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



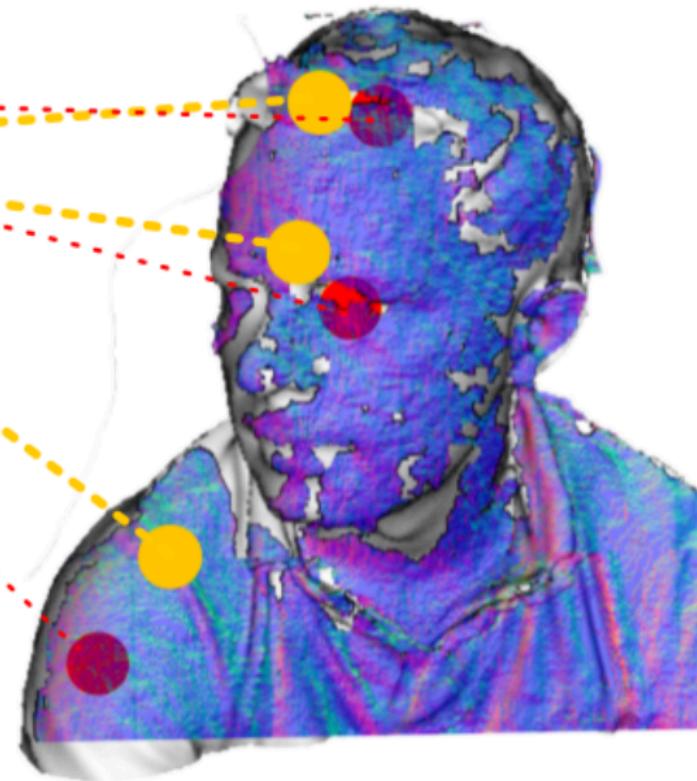
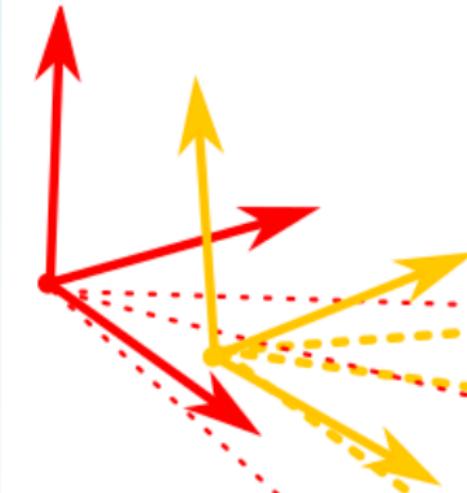
More Extensions: Depth Only Camera Tracking (ICP)



The model is rendered into the estimated frame. Points on the model (red) can be sampled in image space.

Slide acknowledgement R. Newcombe

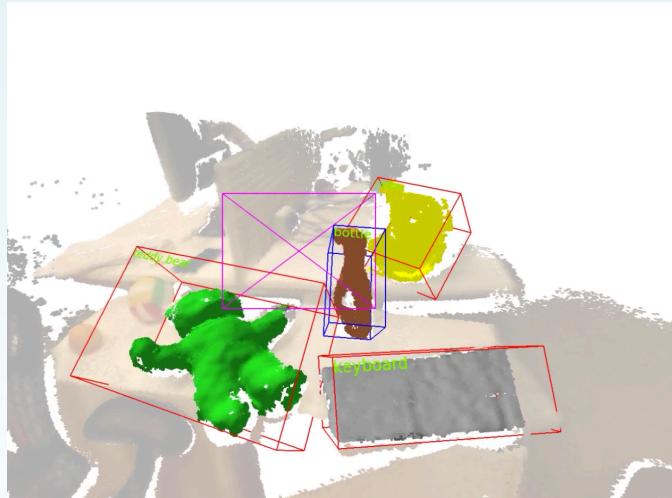
Depth Only Camera Tracking (ICP)



New pose is estimated by performing iterative Gauss-Newton minimisation of the sum of the square errors.

Slide acknowledgement R. Newcombe

Recognition meets 3D Reconstruction



Semantic + shape priors

FroDO: From Detections to Objects

Runz et al. CVPR'20



⋮

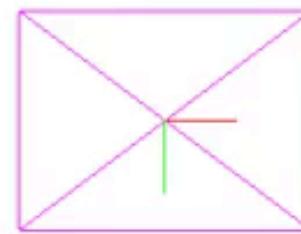
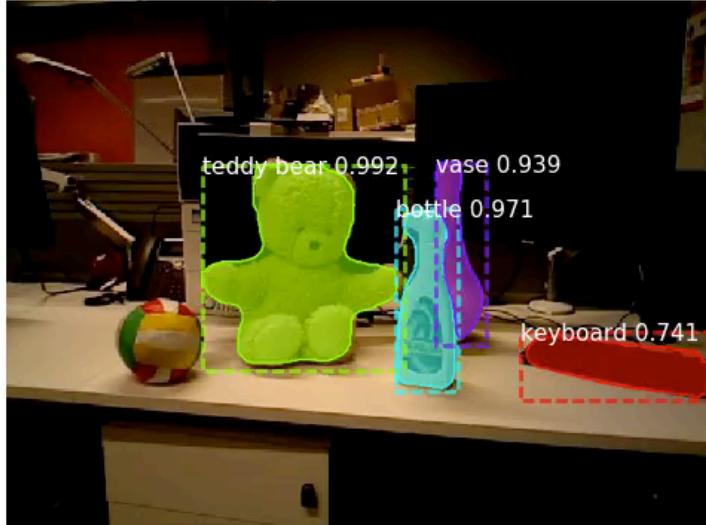


Input

MaskFusion: Real-Time Recognition, Tracking and Reconstruction of Multiple Moving Objects

Martin Runz, Maud Buffier, Lourdes Agapito ISMAR'18

Contribution - Object Based Dynamic SLAM





Monocular Dense Direct and Deformable



with R. Yu, N. Campbell and C. Russell (ICCV'15)

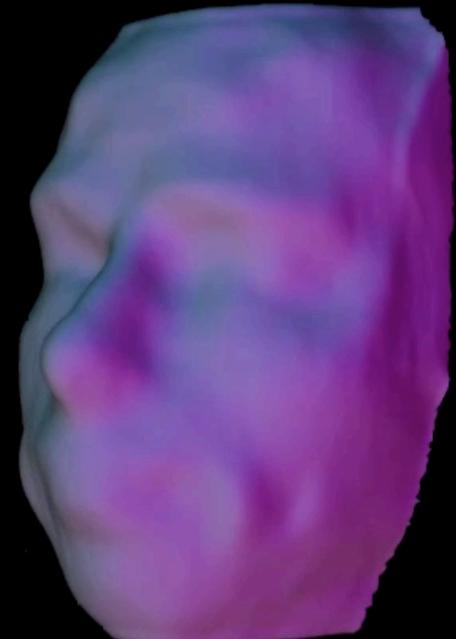
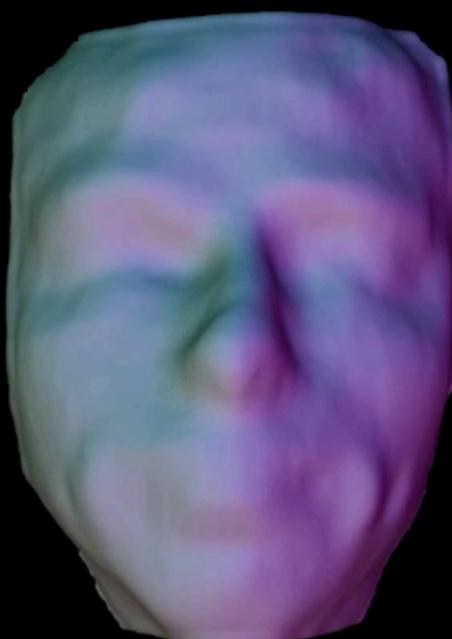
Online Non-rigid Reconstruction



Input

Normals + Shading

Novel View



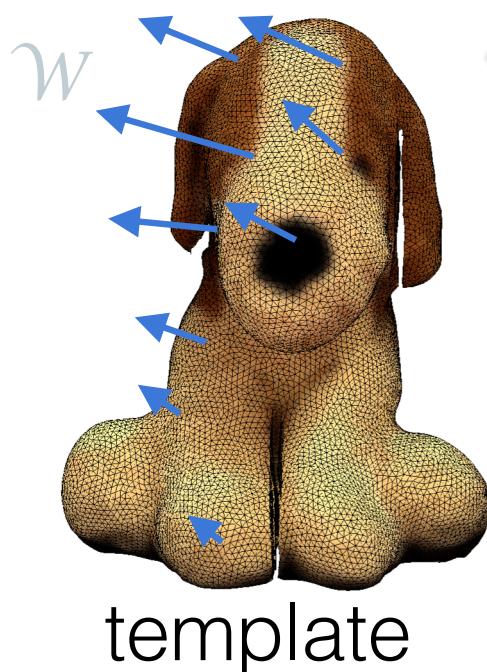
Monocular Dense Direct and Deformable



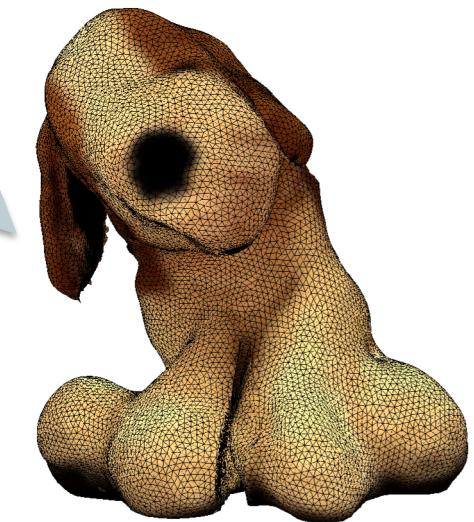
with R. Yu, N. Campbell and C. Russell (ICCV'15)



current live frame



w
3D warp field



$$E_{(w,R,t)} = E_{\text{data}} + E_{\text{reg}}$$

