Direct Visual SLAM

Instructor - Simon Lucey

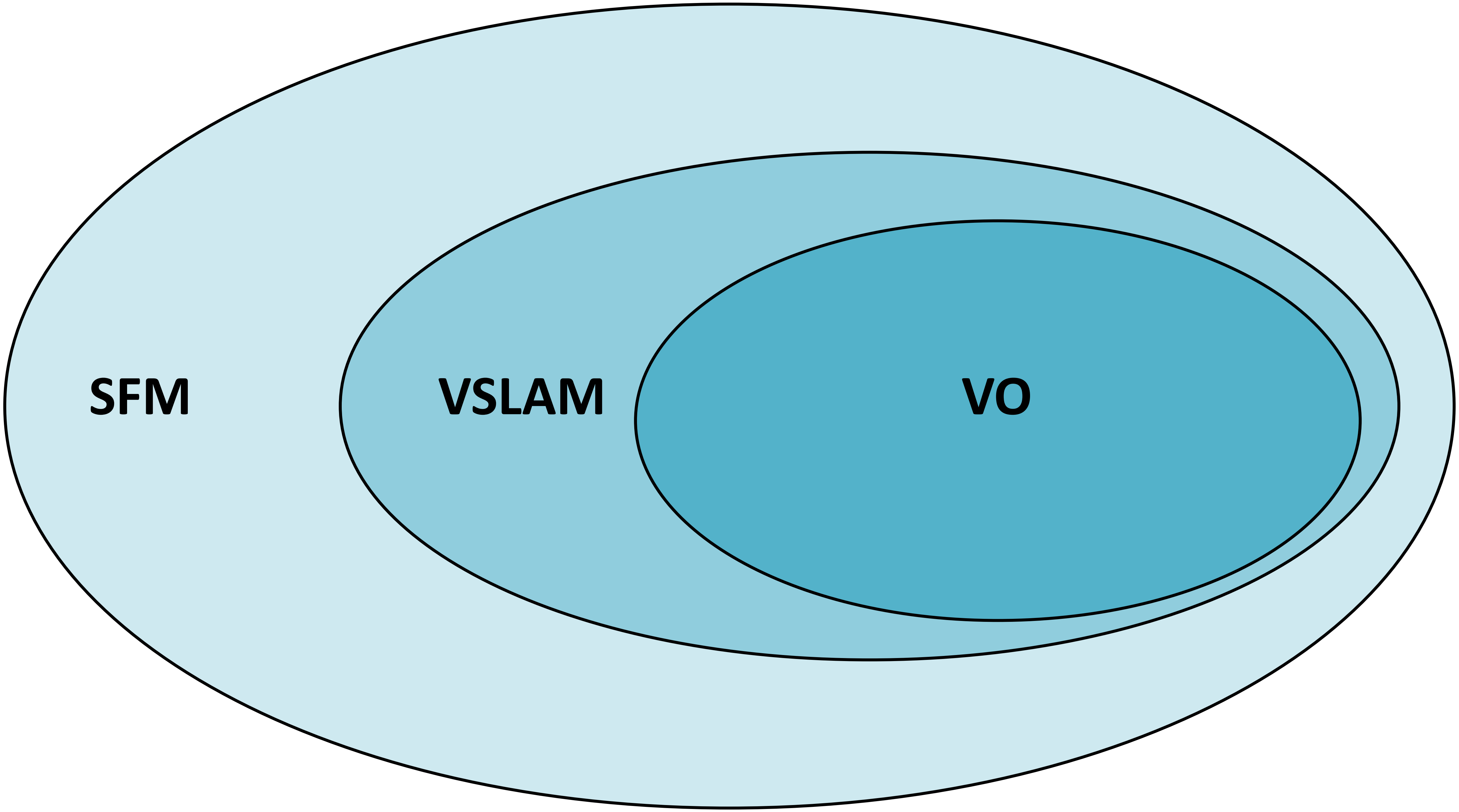
16-623 - Designing Computer Vision Apps

Carnegie Mellon
THE ROBOTICS INSTITUTE

Reminder: SLAM

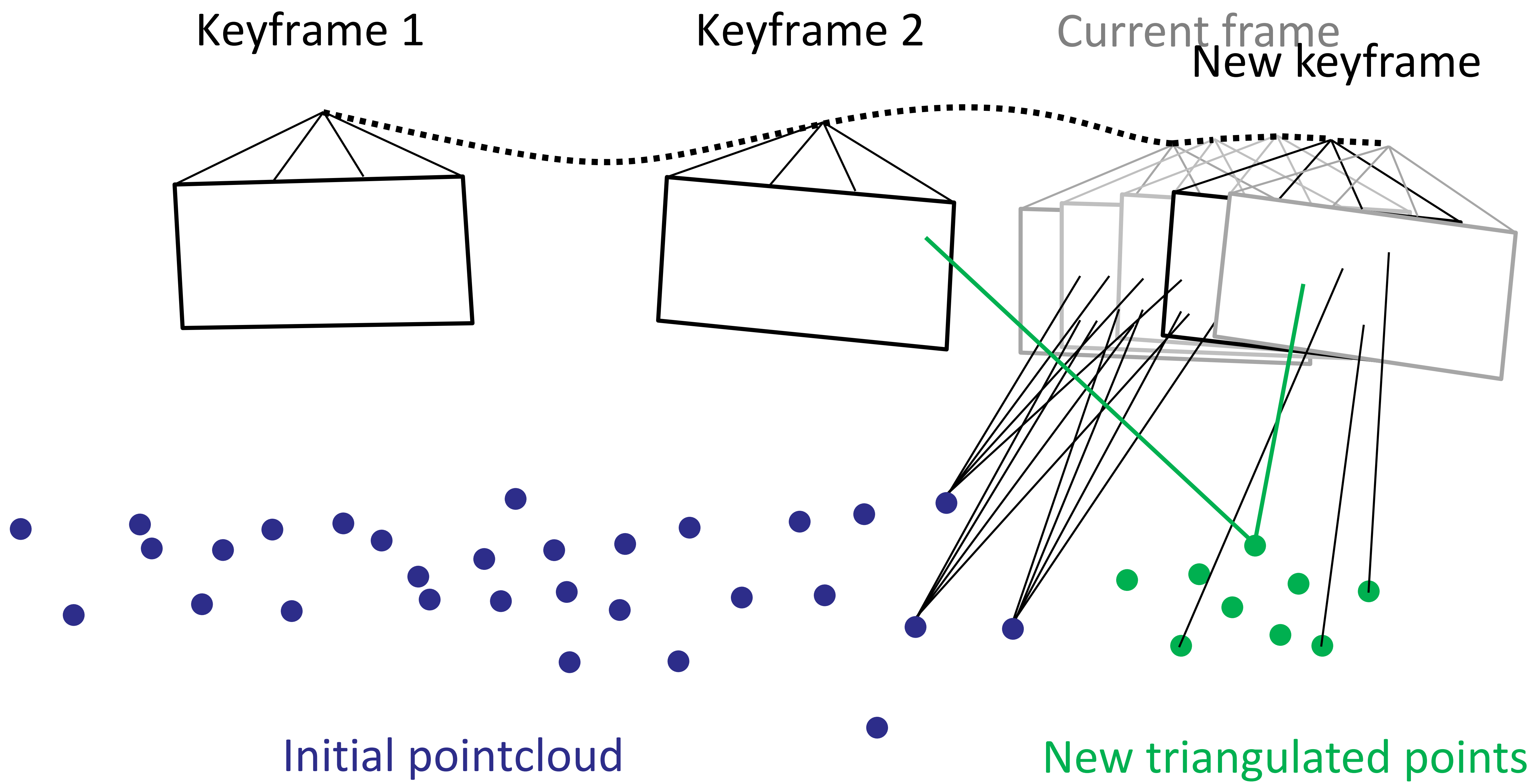
- **S**imultaneous **L**ocalization **a**nd **M**apping.
- On mobile interested primarily in Visual SLAM (VSLAM).
- Sometimes called Mono SLAM if there is only one camera.
- Can be viewed as an online SfM problem.

Reminder: VO vs VSLAM vs SFM



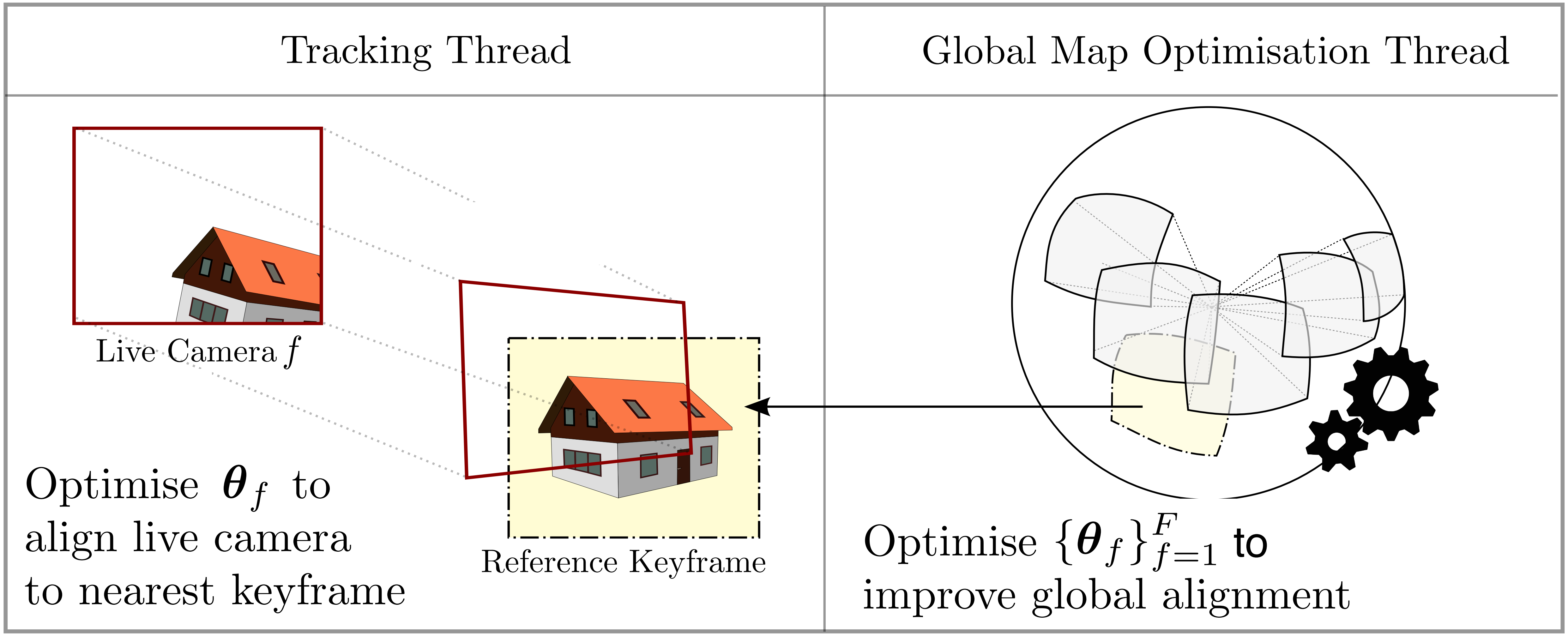
Taken from D. Scaramuzza “Tutorial on Visual Odometry”.

Reminder: Keyframe-based SLAM



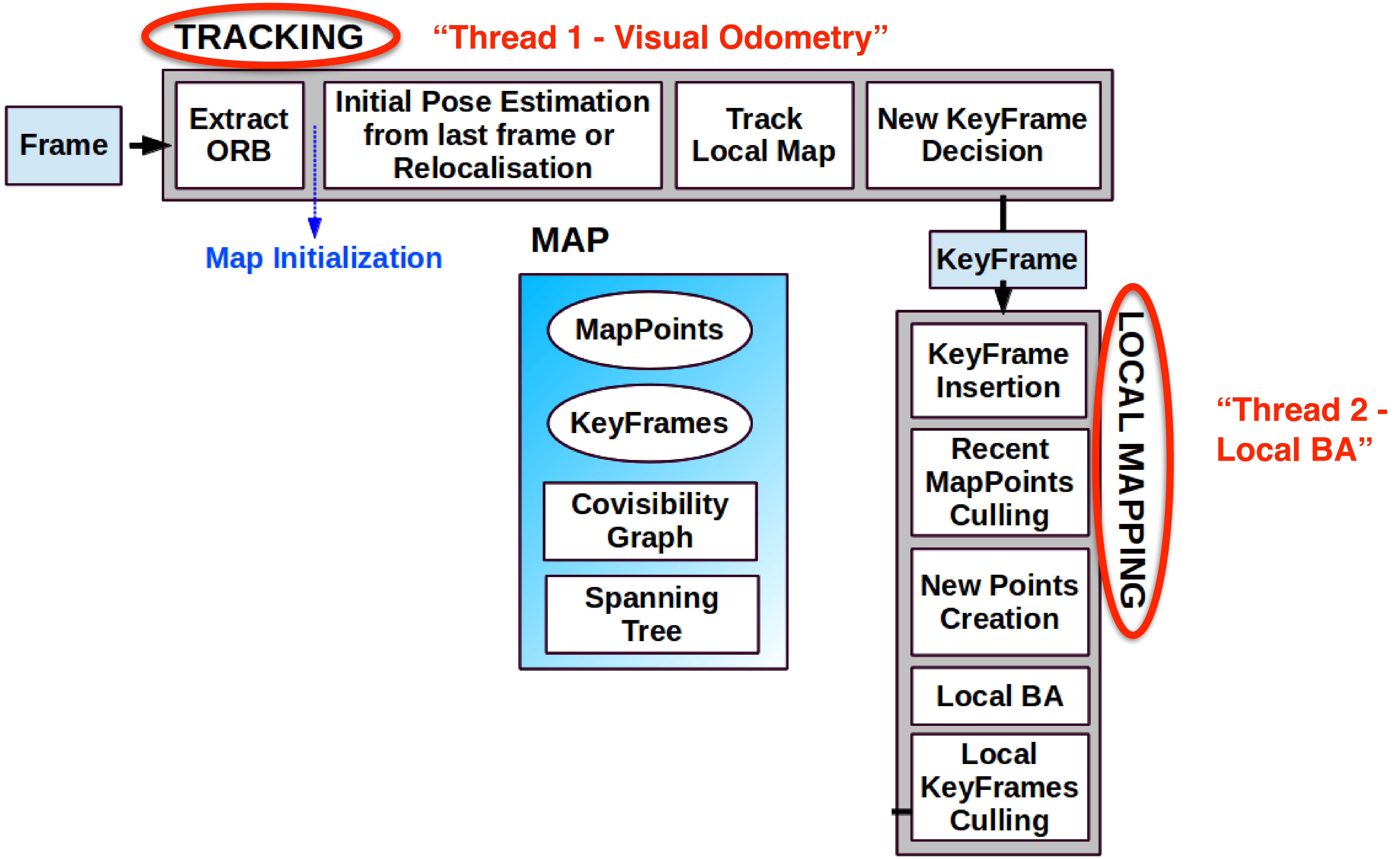
Taken from D. Scaramuzza “Tutorial on Visual Odometry”.

A Tale of Two Threads



Adapted from S. Lovegrove & A. J. Davison "Real-Time Spherical Mosaicing using Whole Image Alignment", ECCV 2010.

Example - ORB SLAM



R. Mur-Artal, J. M. M. Montiel, J. D. Tardos, "ORB-SLAM: a Versatile and Accurate Monocular SLAM System" IEEE Trans. Robotics 2015.

Today

- Direct vs. Feature based methods
- Dense SLAM
- Semi-Dense SLAM

ECCV 1999

All About Direct Methods

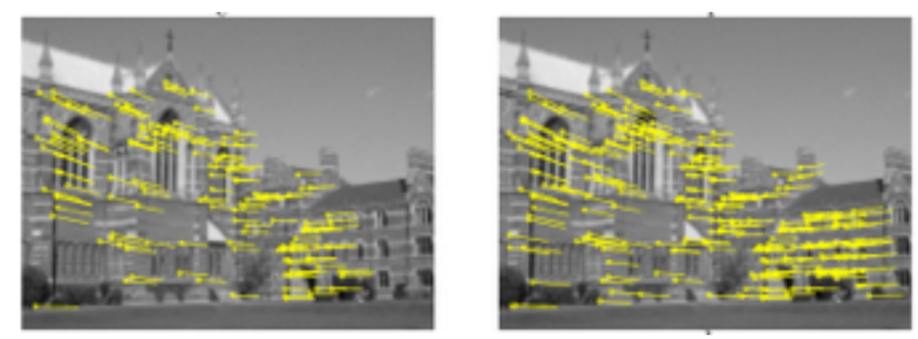
M. Irani[1] and P. Anandan[2]

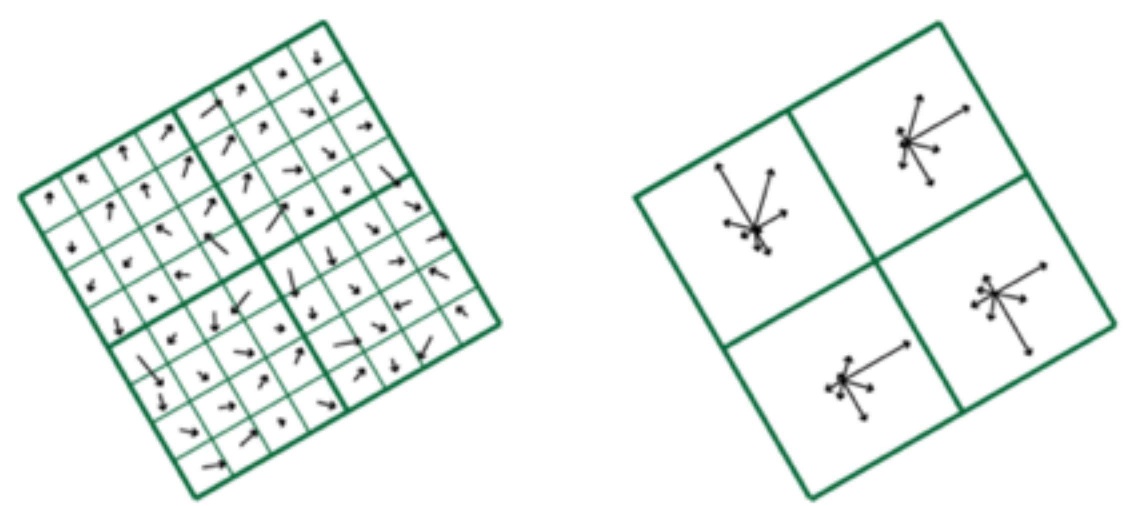
[1] Dept. of Computer Science and Applied Mathematics,
The Weizmann Inst. of Science, Rehovot, Israel.
irani@wisdom.weizmann.ac.il
[2] Microsoft Research, One Microsoft Way,
Redmond, WA 98052, USA.
anandan@microsoft.com

Feature Based Methods for Structure and Motion Estimation

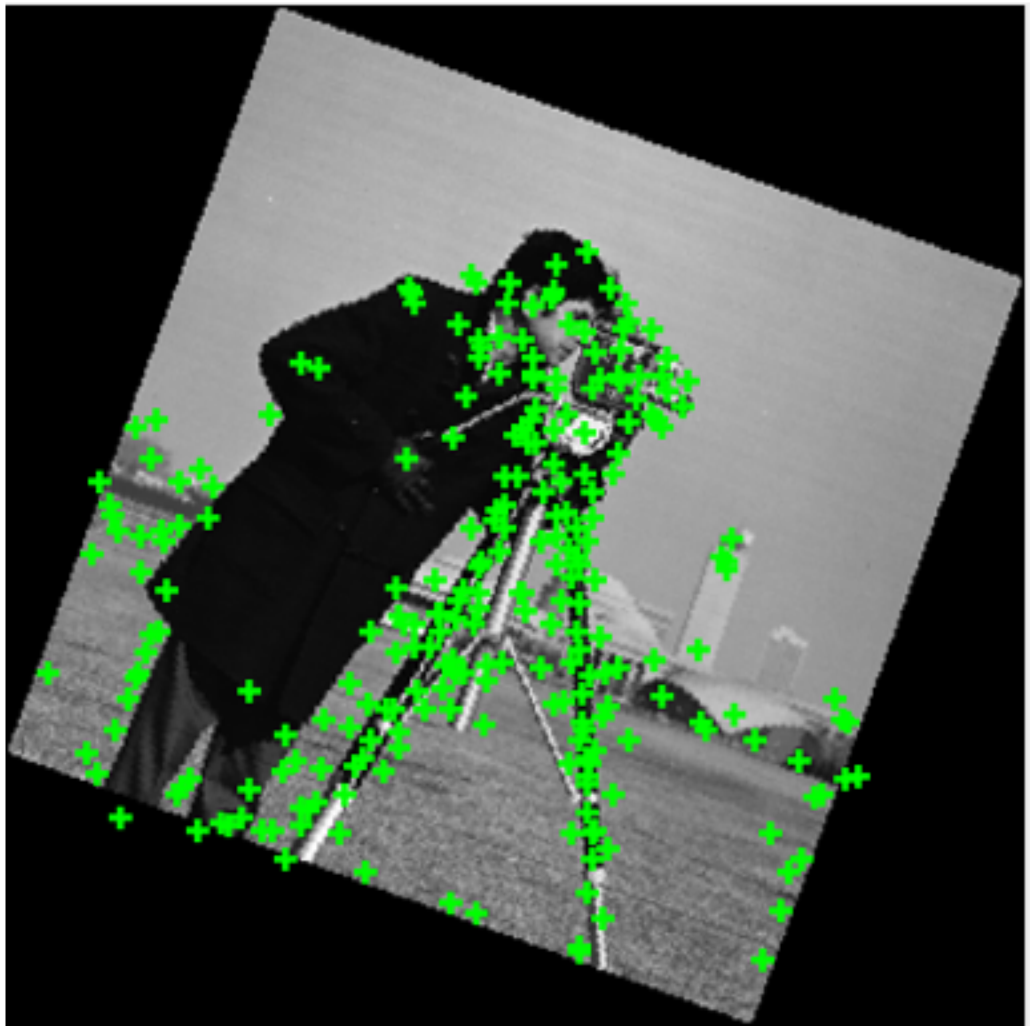
P. H. S. Torr[1] and A. Zisserman[2]

[1] Microsoft Research Ltd, 1 Guildhall St
Cambridge CB2 3NH, UK
philtorr@microsoft.com
[2] Department of Engineering Science, University of Oxford
Oxford, OX1 3PJ, UK
az@robots.ox.ac.uk

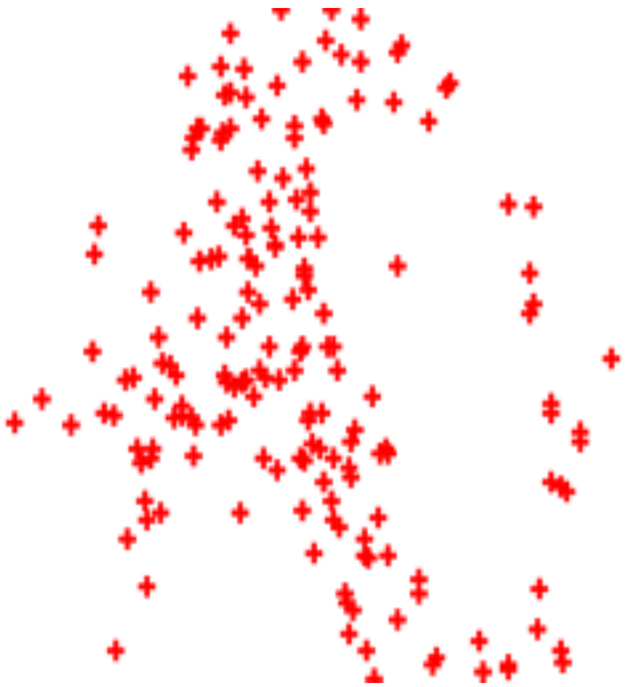




Feature-Based Methods



Feature-Based Methods



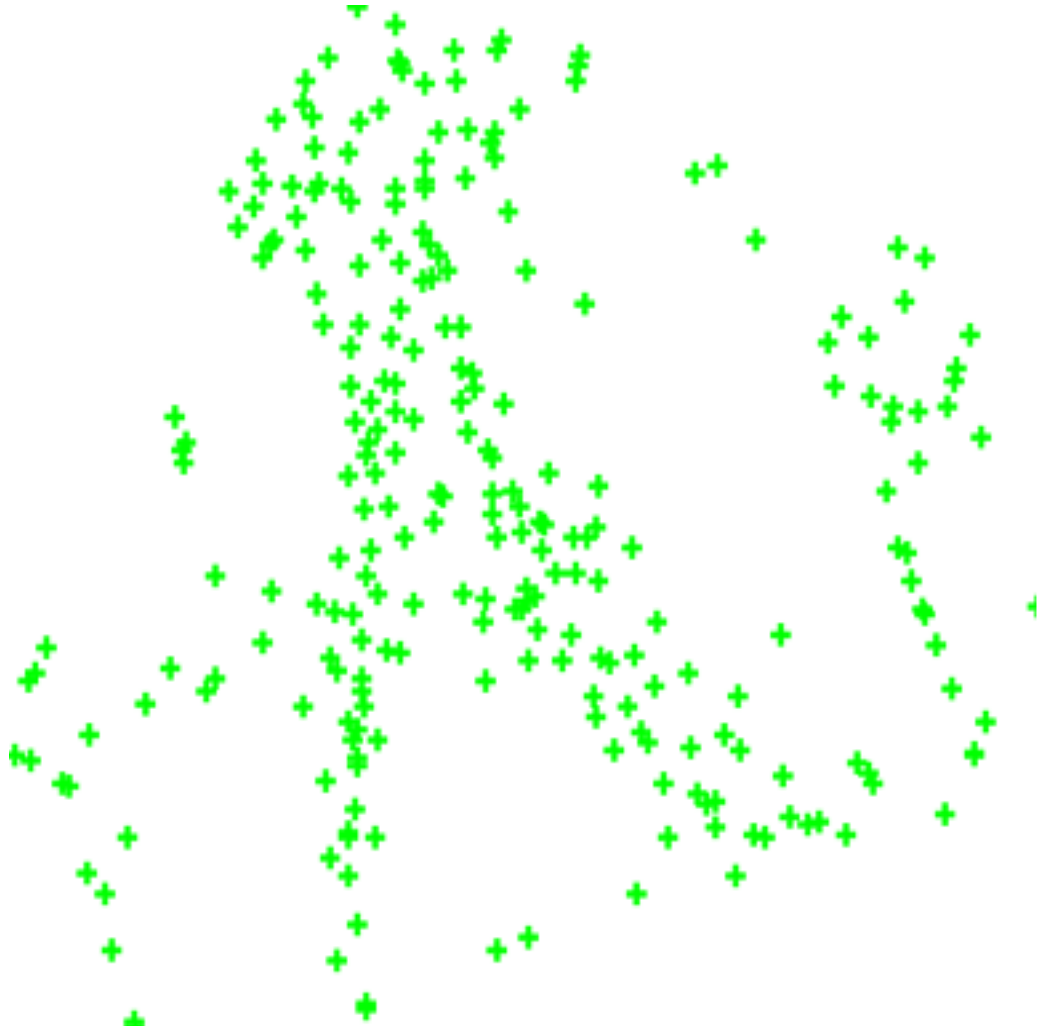
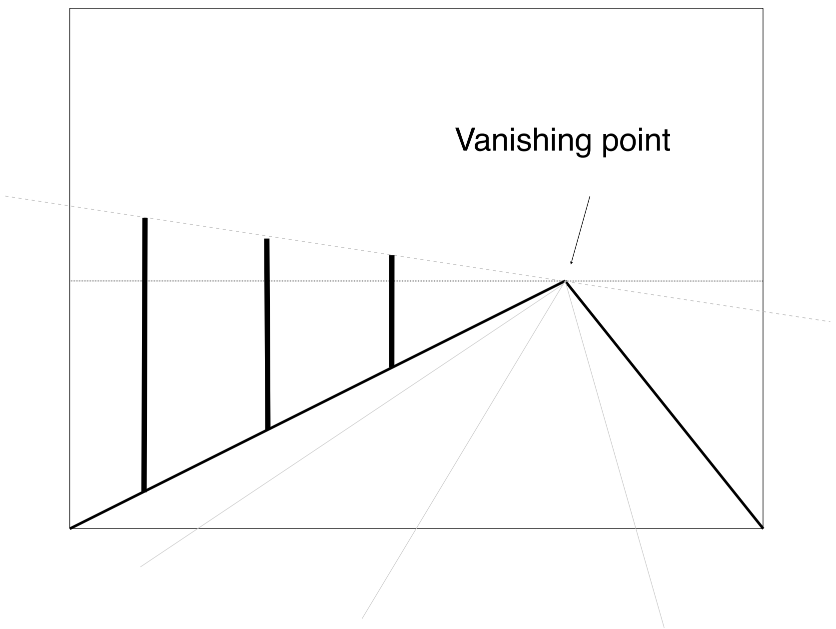


Image is reduced to a sparse set of **keypoints**
Usually matched with feature **descriptors**

Feature-Based Advantages



Easier transition from images to geometry

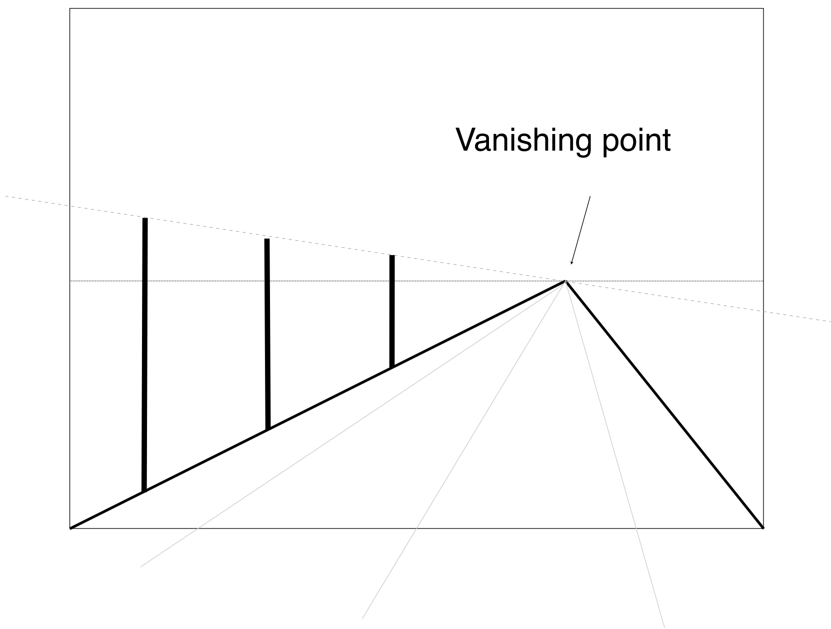
Mikolajczyk, 2007

Wide-baseline matching

Mikolajczyk, 2007

Illumination invariance

Feature-Based Advantages



Easier transition from images to geometry

Mikolajczyk, 2007

Wide-baseline matching

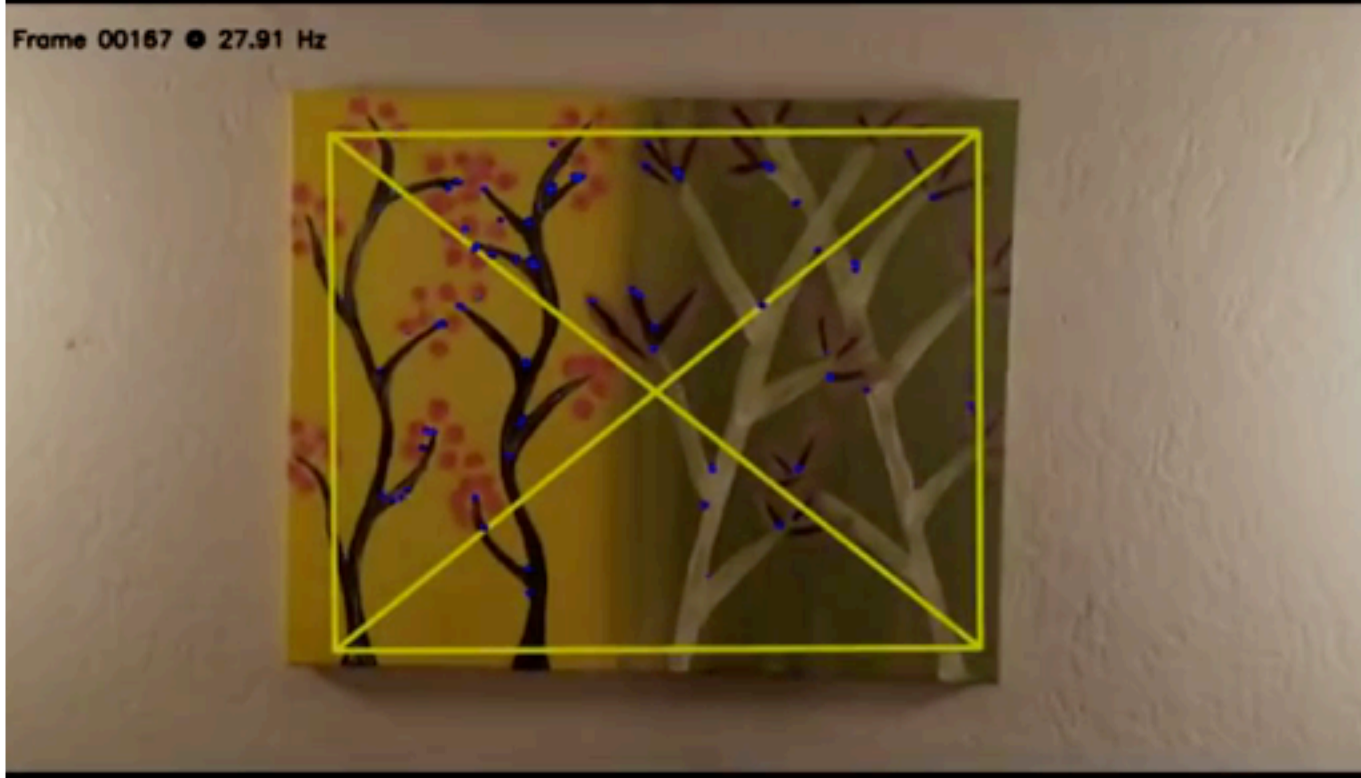
Mikolajczyk, 2007

Illumination invariance

Using invariant descriptors

Feature-Based Challenges

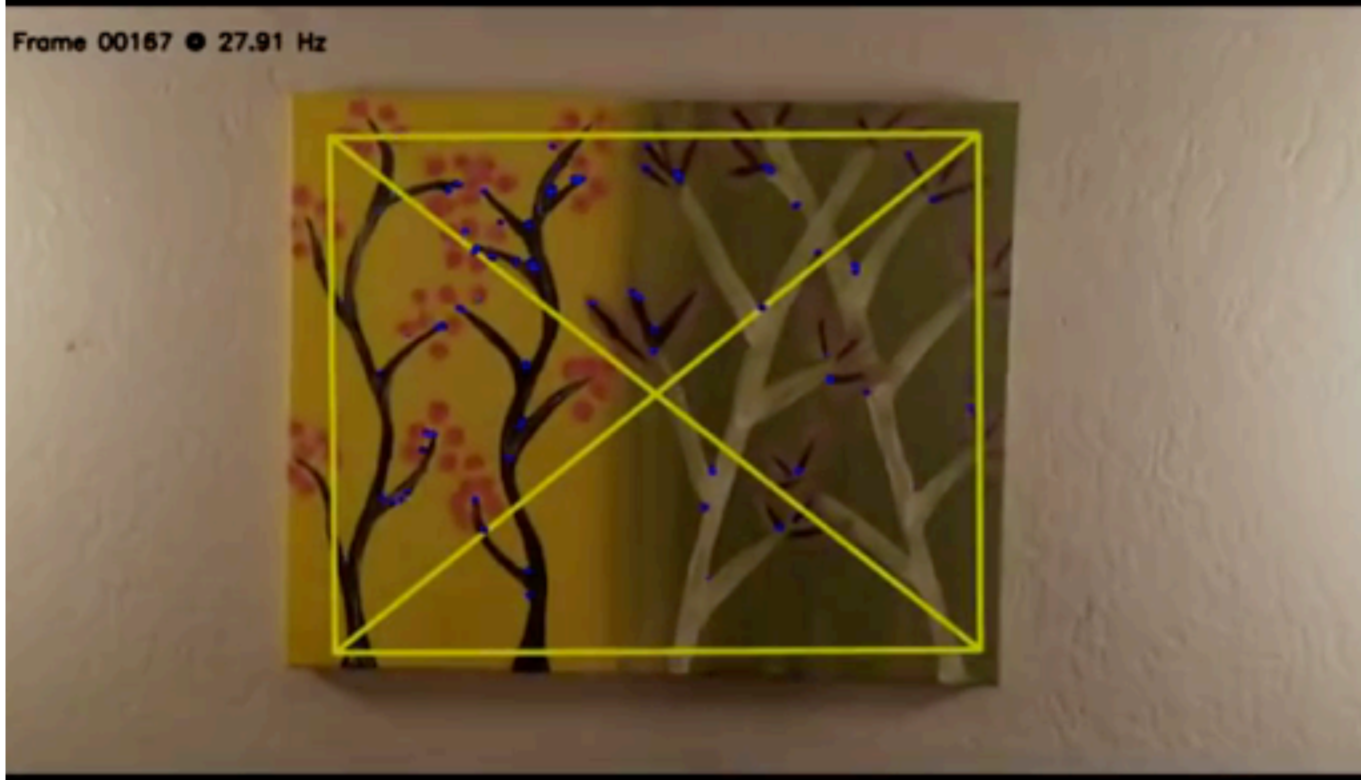
- Creates only a sparse map of the world.
- Does not sample across all available image data - edges & weak intensities.
- Needs high-resolution camera mode (bad for efficiency and battery life).



Feature-Based Method
(ORB+RANSAC)

Feature-Based Challenges

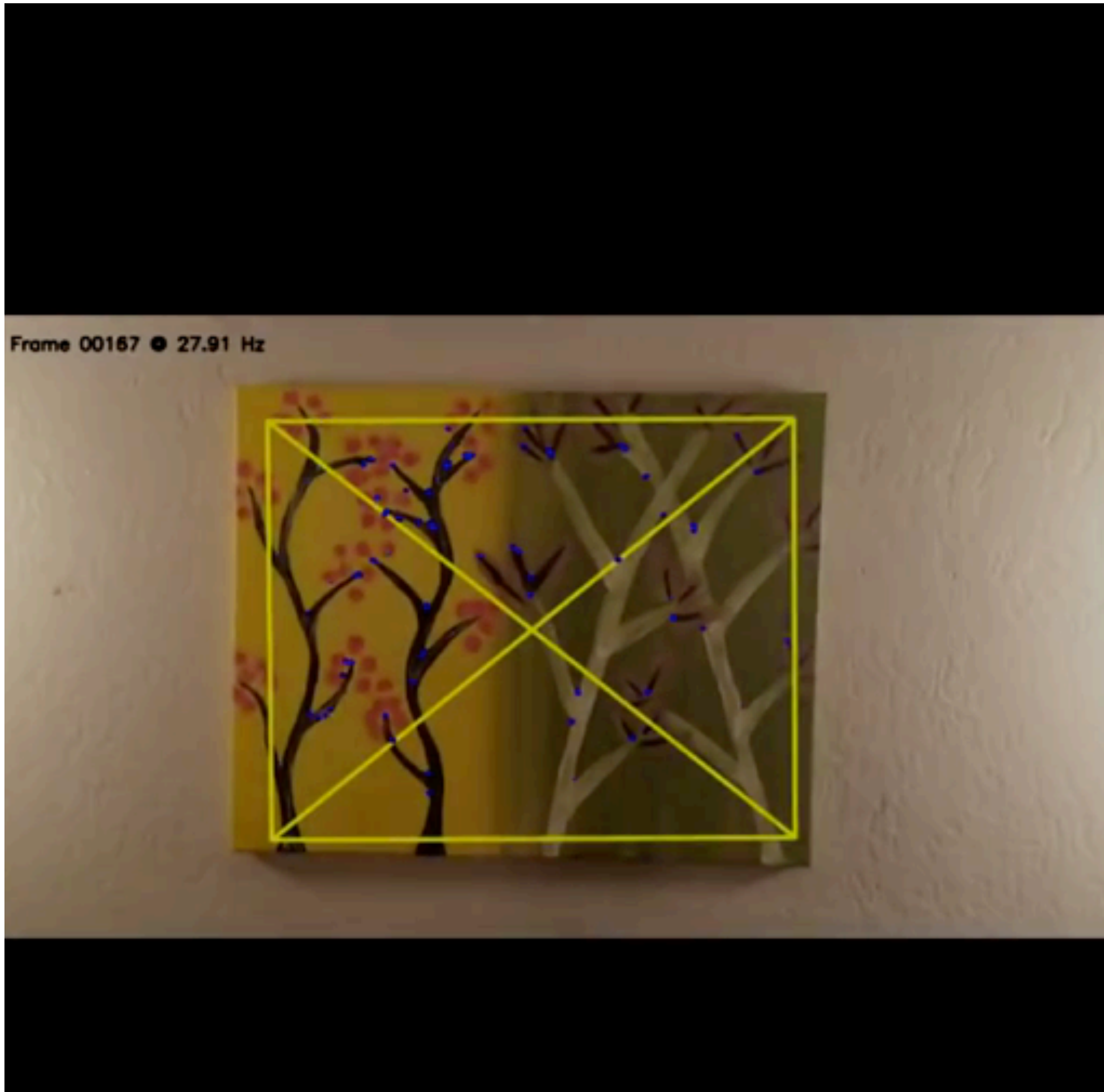
- Creates only a sparse map of the world.
- Does not sample across all available image data - edges & weak intensities.
- Needs high-resolution camera mode (bad for efficiency and battery life).



Feature-Based Method
(ORB+RANSAC)

Feature-Based Challenges

- Creates only a sparse map of the world.
- Does not sample across all available image data - edges & weak intensities.
- Needs high-resolution camera mode (bad for efficiency and battery life).

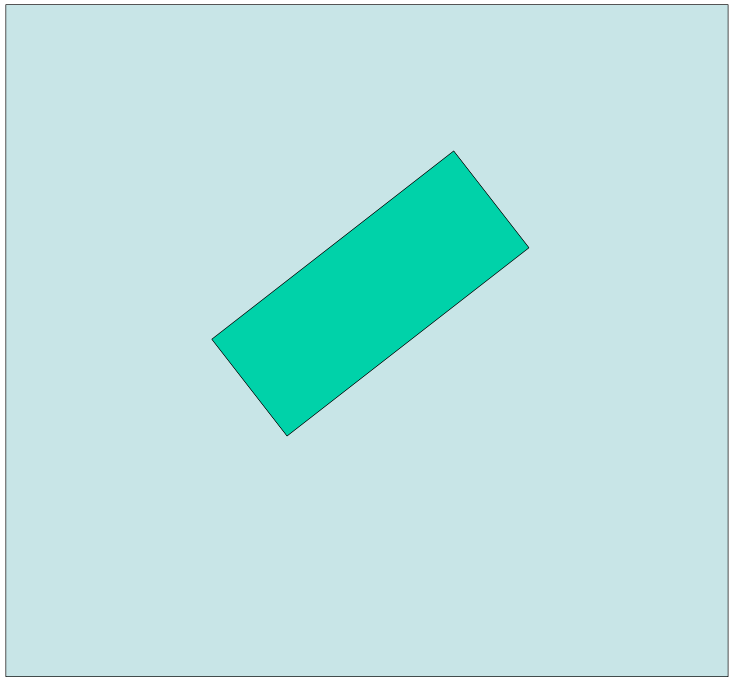


Feature-Based Method
(ORB+RANSAC)

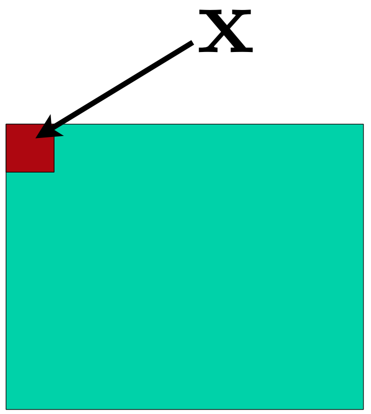
Today

- Direct vs. Feature based methods
- Dense SLAM
- Semi-Dense SLAM

Reminder: Warp Functions

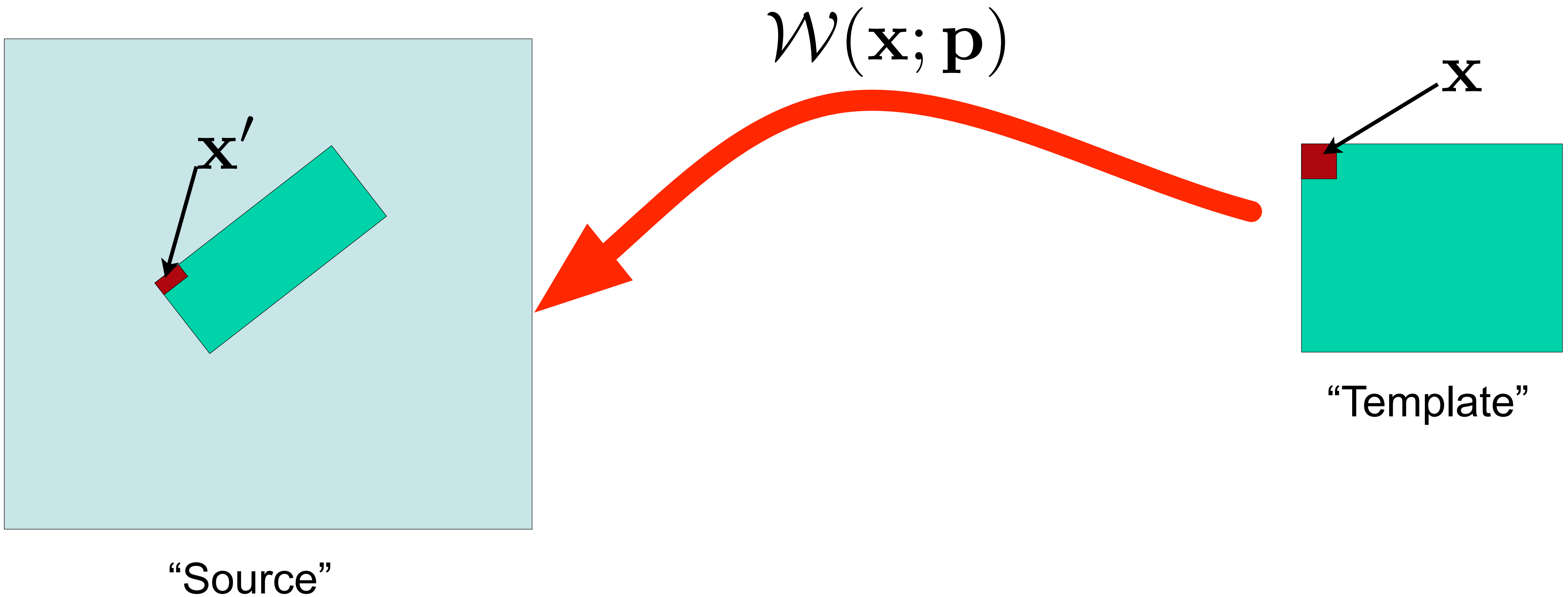


“Source”



“Template”

Reminder: Warp Functions



Our goal is to find the warp parameter vector $\mathbf{p}$!

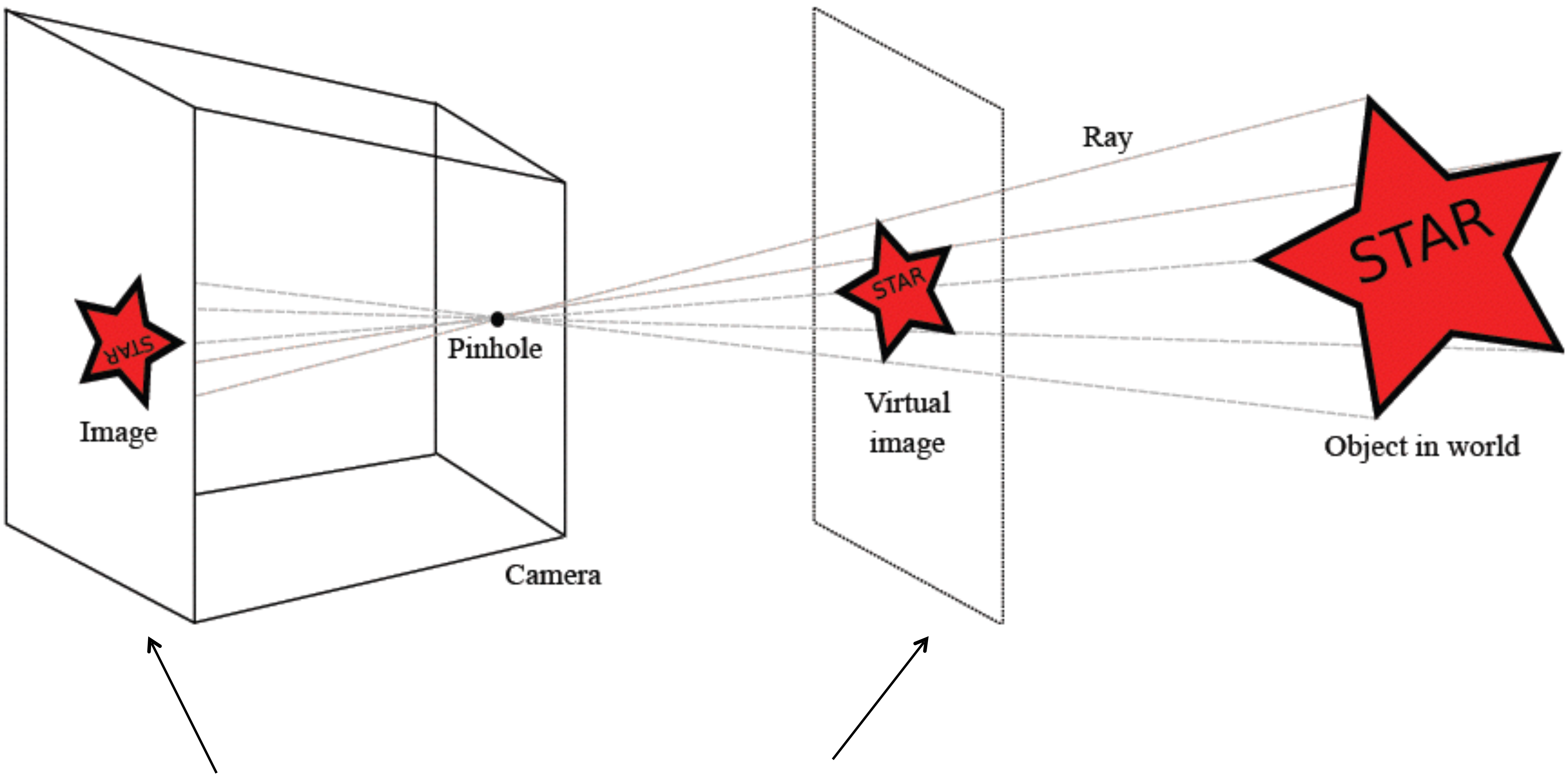
$\mathbf{x}$ = coordinate in template $[x, y]^T$

$\mathbf{x}'$ = corresponding coordinate in source $[x', y']^T$

$\mathcal{W}(\mathbf{x};\mathbf{p})$ = warping function such that $\mathbf{x}' = \mathcal{W}(\mathbf{x};\mathbf{p})$

$\mathbf{p}$ = parameter vector describing warp

Review: Pinhole Camera



Real camera image
is inverted

Instead model impossible but more convenient
virtual image

Adapted from: Computer vision: models, learning and inference. Simon J.D. Prince

Relating Points between Views

First camera:

$$\lambda_1 \tilde{\mathbf{x}}_1 = \mathbf{w}$$

Second camera:

$$\lambda_2 \tilde{\mathbf{x}}_2 = \boldsymbol{\Omega}\mathbf{w} + \boldsymbol{\tau}$$

Substituting:

$$\lambda_2 \tilde{\mathbf{x}}_2 = \lambda_1 \boldsymbol{\Omega} \tilde{\mathbf{x}}_1 + \boldsymbol{\tau}$$

Pinhole Warp Function

- One can represent the relationship of points between views of pinhole cameras as a warp function,

$$\mathcal{W}(\mathbf{x};\boldsymbol{\theta},\lambda) = \pi(\lambda\boldsymbol{\Omega}\tilde{\mathbf{x}} + \boldsymbol{\tau})$$ "warp function"

$$\pi\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} u/w \\ v/w \end{pmatrix}$$ "pinhole projection"

$$\mathbf{T} = \begin{bmatrix} \boldsymbol{\Omega} & \boldsymbol{\tau} \\ \mathbf{0}^T & 1 \end{bmatrix} \in \mathrm{SE}(3)$$ "pose parameters"

Pinhole Warp Function

- One can represent the relationship of points between views of pinhole cameras as a warp function,

$$\mathcal{W}(\mathbf{x}; \boldsymbol{\theta}, \lambda) = \pi(\lambda \boldsymbol{\Omega} \tilde{\mathbf{x}} + \boldsymbol{\tau})$$ "warp function"

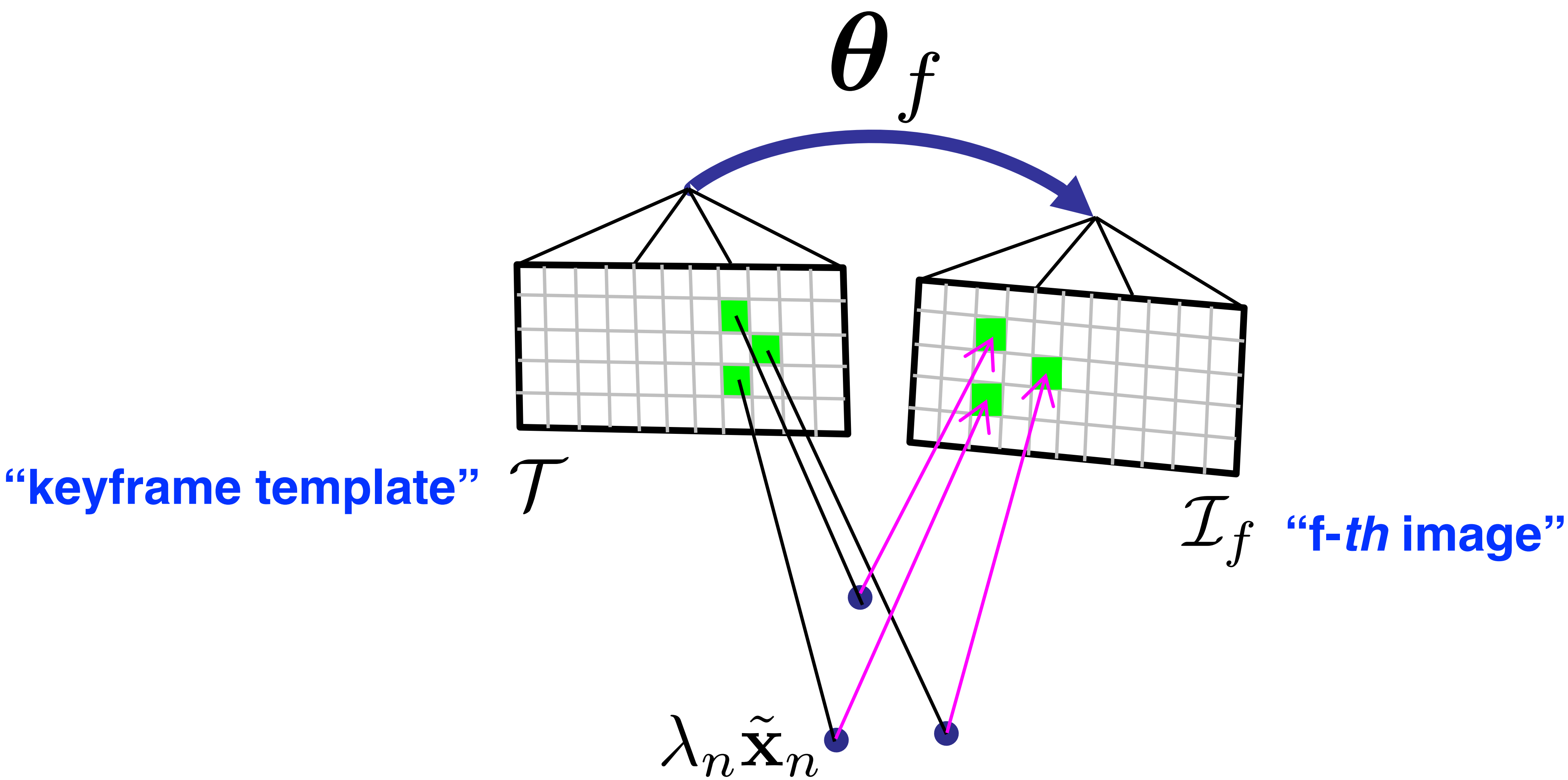
$$\pi \begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} u/w \\ v/w \end{pmatrix}$$ "pinhole projection"

$$\mathbf{T}(\boldsymbol{\theta}) = \exp\left(\sum_{i=1}^{6} \theta_i \mathbf{A}_i\right) \in \mathrm{SE}(3)$$ "pose parameters"

Photometric Relationship

- We can employ this warp function to now express the problem as,

$$\mathcal{T}(\mathbf{x}_n) = \mathcal{I}(\mathcal{W}\{\mathbf{x}_n; \boldsymbol{\theta}_f, \lambda_n\})$$



Linearizing the Image for Pose

$$\mathcal{T}(\mathbf{x}_n) = \mathcal{I}_f(\mathcal{W}\{\mathbf{x}_n; \boldsymbol{\theta}_f \circ \Delta\boldsymbol{\theta}, \lambda_n\})$$

$$\approx \mathcal{I}_f(\mathcal{W}\{\mathbf{x}_n; \boldsymbol{\theta}_f, \lambda_n\}) + \mathbf{A}_n^f \Delta\boldsymbol{\theta}_f$$

Baker, Simon, and Iain Matthews. "Equivalence and efficiency of image alignment algorithms." CVPR 2001.

Linearizing the Image for Pose

$$\mathcal{T}(\mathbf{x}_n) = \mathcal{I}_f(\mathcal{W}\{\mathbf{x}_n; \boldsymbol{\theta}_f \circ \Delta\boldsymbol{\theta}, \lambda_n\})$$

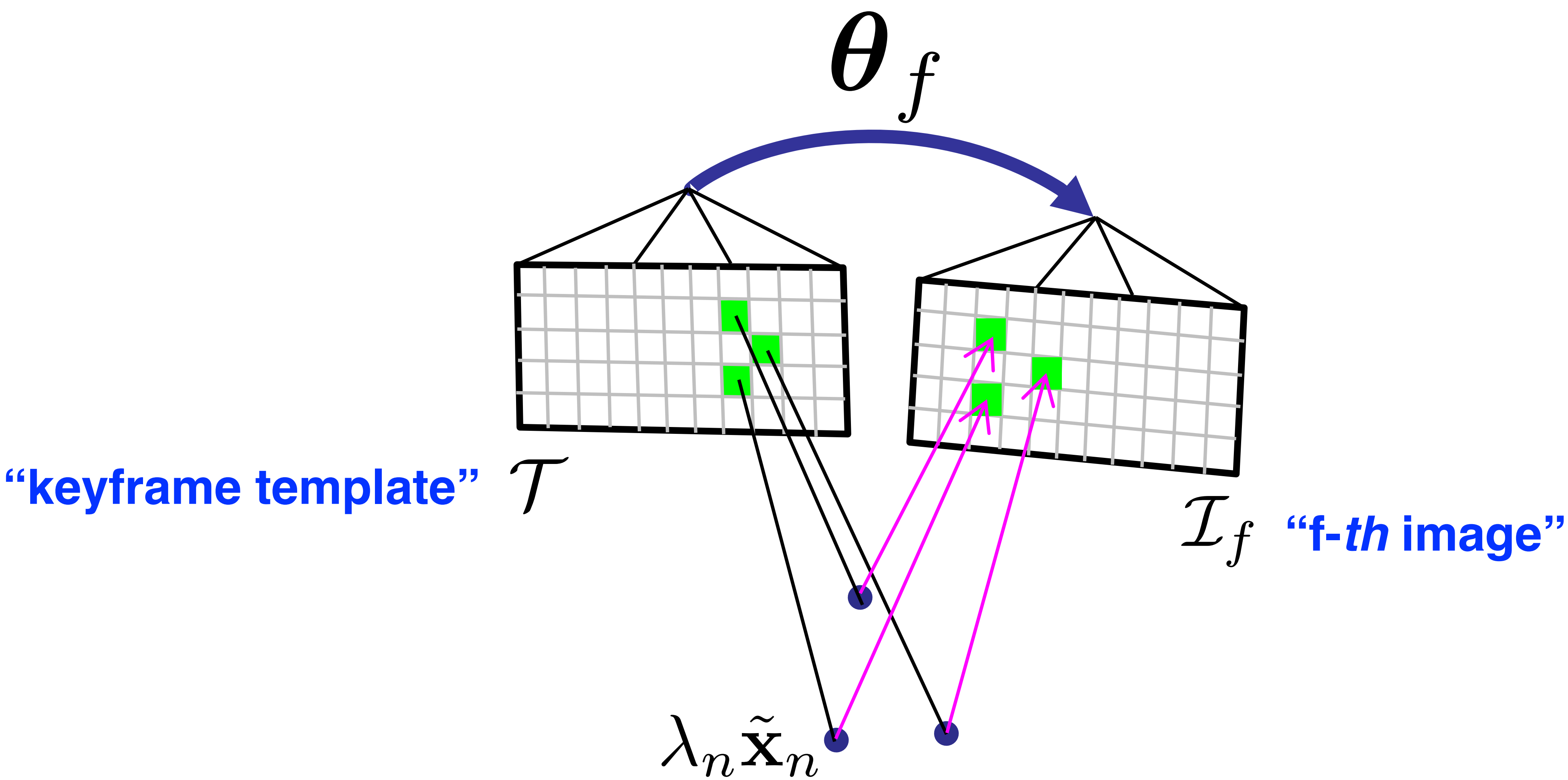
$$\approx \mathcal{I}_f(\mathcal{W}\{\mathbf{x}_n; \boldsymbol{\theta}_f, \lambda_n\}) + \mathbf{A}_n^f \Delta\boldsymbol{\theta}_f$$

Baker, Simon, and Iain Matthews. "Equivalence and efficiency of image alignment algorithms." CVPR 2001.

Direct Camera Tracking

- Assuming known depths $\{\lambda_n\}_{n=1}^{N}$,

$$\arg\min_{\Delta\boldsymbol{\theta}_f} \sum_{n=1}^{N} ||\mathcal{T}(\mathbf{x}_n) - \mathcal{I}_f(\mathcal{W}\{\mathbf{x}_n;\boldsymbol{\theta}_f,\lambda_n\}) - \mathbf{A}_n^f \Delta\boldsymbol{\theta}_f||_2^2$$



Direct Camera Tracking

- Most methods employ a variant of the Lucas-Kanade algorithm for estimating camera pose.
- Engel et al. demonstrated using a “dense” number of points does not improve the performance of camera tracking (i.e pose estimation).
- Advantage of density stems mainly from the map estimation.

J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. arXiv preprint arXiv:1607.02565, 2016.
J. Engel, T. Schops, and D. Cremers. LSD-SLAM: Large-scale direct monocular slam. In European Conference on Computer Vision, pages 834–849. Springer, 2014.

Direct Camera Tracking

- Most methods employ a variant of the Lucas-Kanade algorithm for estimating camera pose.
- Engel et al. demonstrated using a “dense” number of points does not improve the performance of camera tracking (i.e pose estimation).
- Advantage of density stems mainly from the map estimation.

How do we update the depths?

J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. arXiv preprint arXiv:1607.02565, 2016.
J. Engel, T. Schops, and D. Cremers. LSD-SLAM: Large-scale direct monocular slam. In European Conference on Computer Vision, pages 834–849. Springer, 2014.

Direct Map Estimation

- Assuming known pose parameters $\{\boldsymbol{\theta}_f\}_{f=1}^{F}$,
- Naively we could solve for the depths independently,

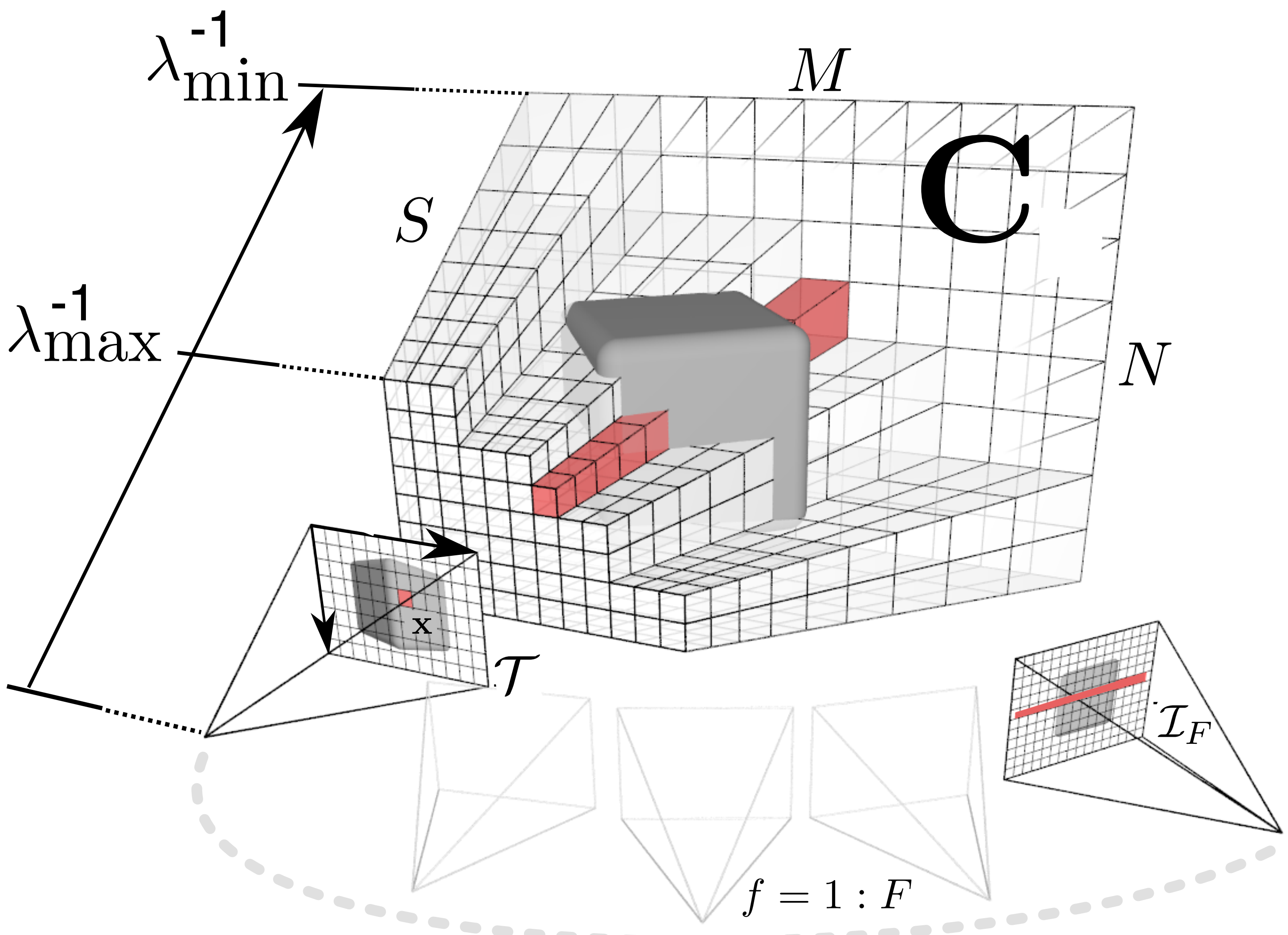
$$\lambda_n = \arg\min_{\lambda} \mathcal{C}(\mathbf{x}, \lambda)$$

$$\mathcal{C}(\mathbf{x}, \lambda) = \frac{1}{F}\sum_{f=1}^{F} ||\mathcal{T}(\mathbf{x}) - \mathcal{I}_f(\mathcal{W}\{\mathbf{x}; \boldsymbol{\theta}_f, \lambda\})||_1$$

R. A. Newcombe, S. J. Lovegrove and A. J. Davison "DTAM: Dense Tracking and Mapping in Real-Time", ICCV 2011.

DTAM

- Newcombe et al. proposed - **D**ense **T**racking **a**nd **M**apping.
- Attempted to substitute the feature based tracking and mapping modules of traditional VSLAM (e.g. PTAM) for dense methods.

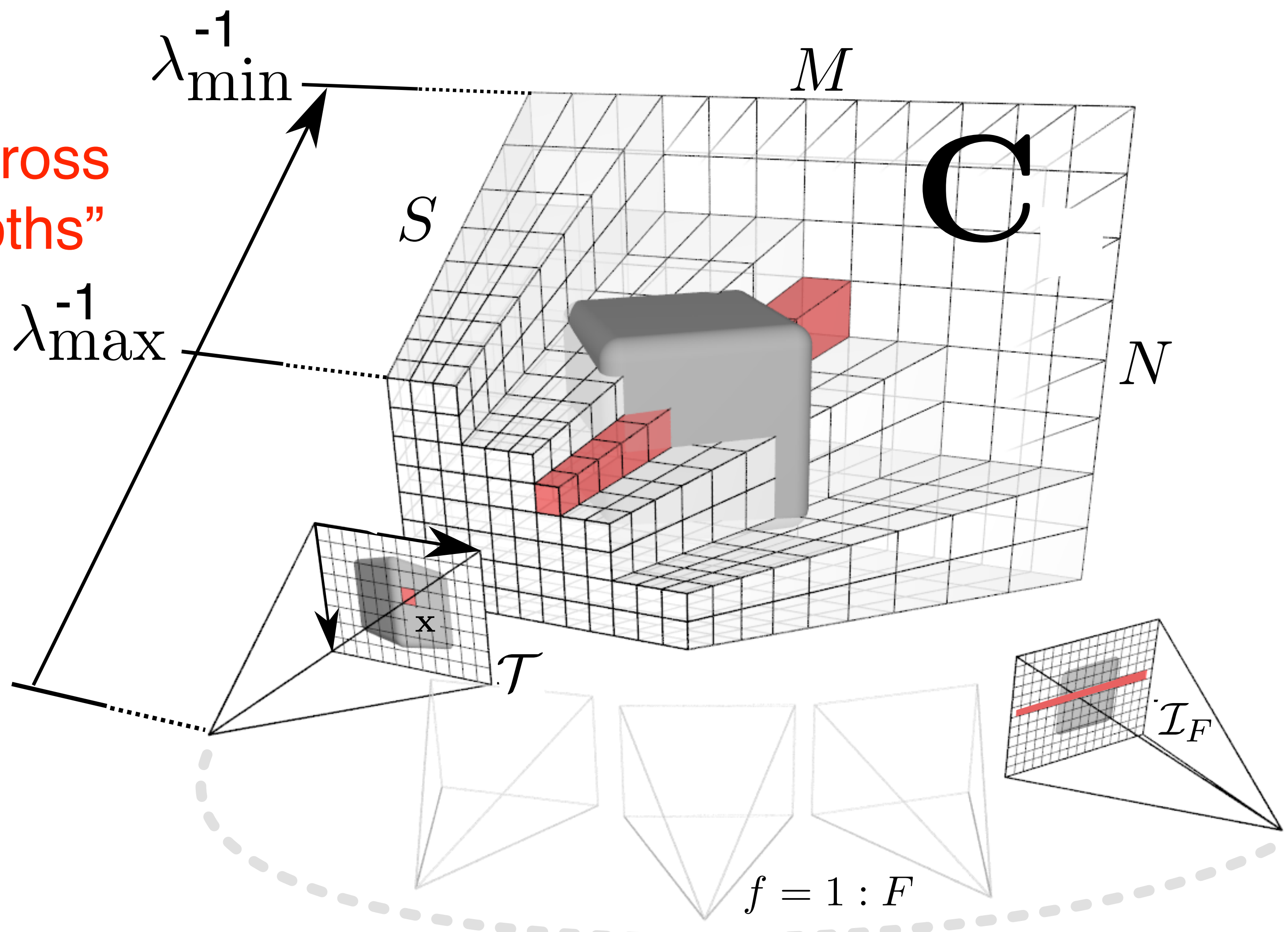


R. A. Newcombe, S. J. Lovegrove and A. J. Davison "DTAM: Dense Tracking and Mapping in Real-Time", ICCV 2011.

DTAM

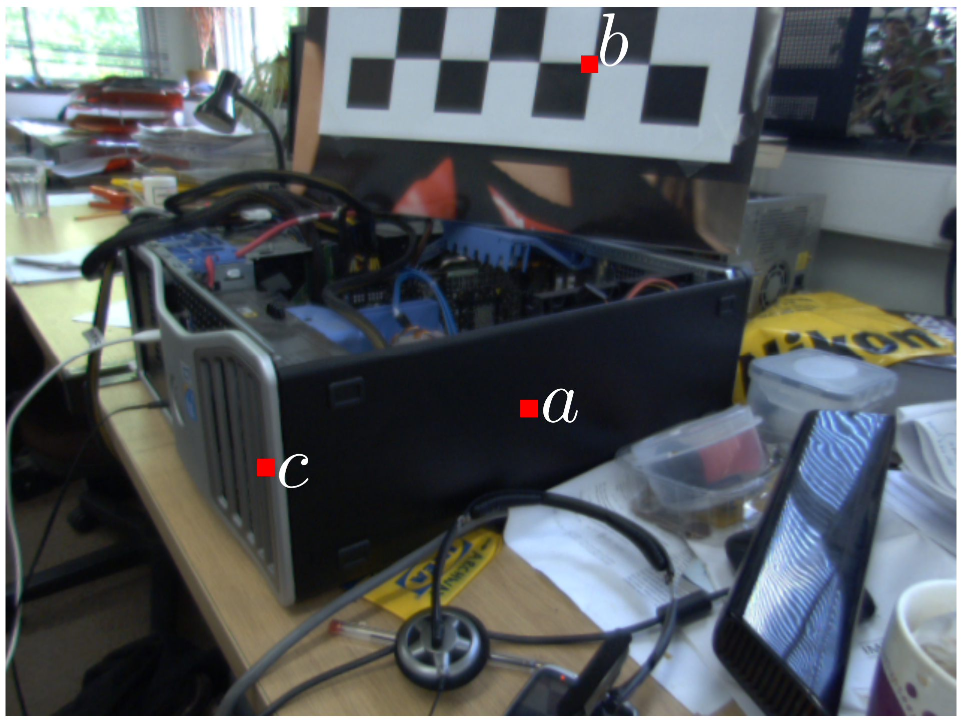
- Newcombe et al. proposed - **D**ense **T**racking **a**nd **M**apping.
- Attempted to substitute the feature based tracking and mapping modules of traditional VSLAM (e.g. PTAM) for dense methods.

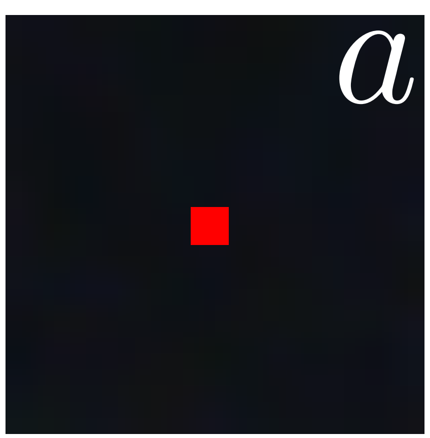
"Sample across inverse depths"

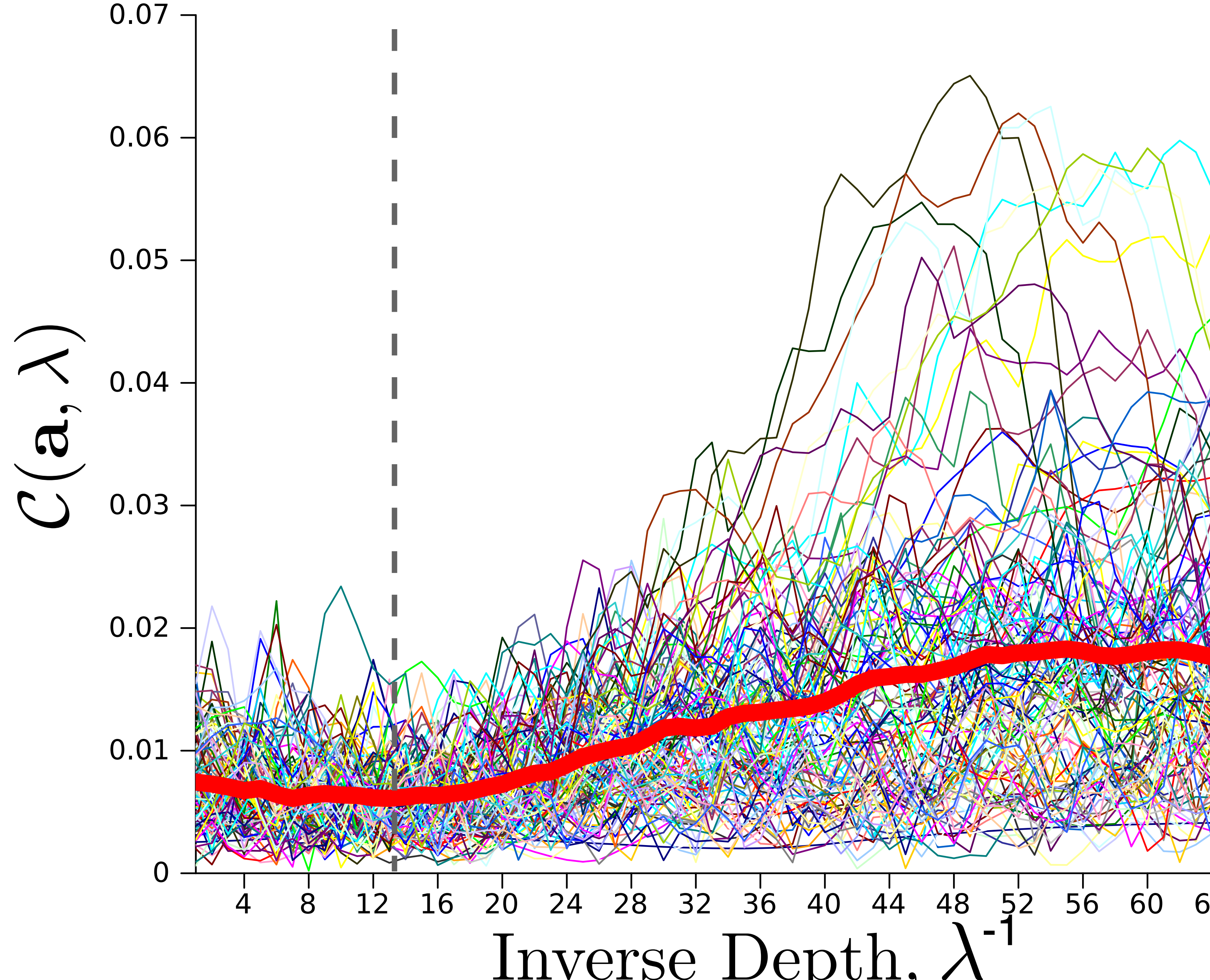


R. A. Newcombe, S. J. Lovegrove and A. J. Davison "DTAM: Dense Tracking and Mapping in Real-Time", ICCV 2011.

DTAM - Example

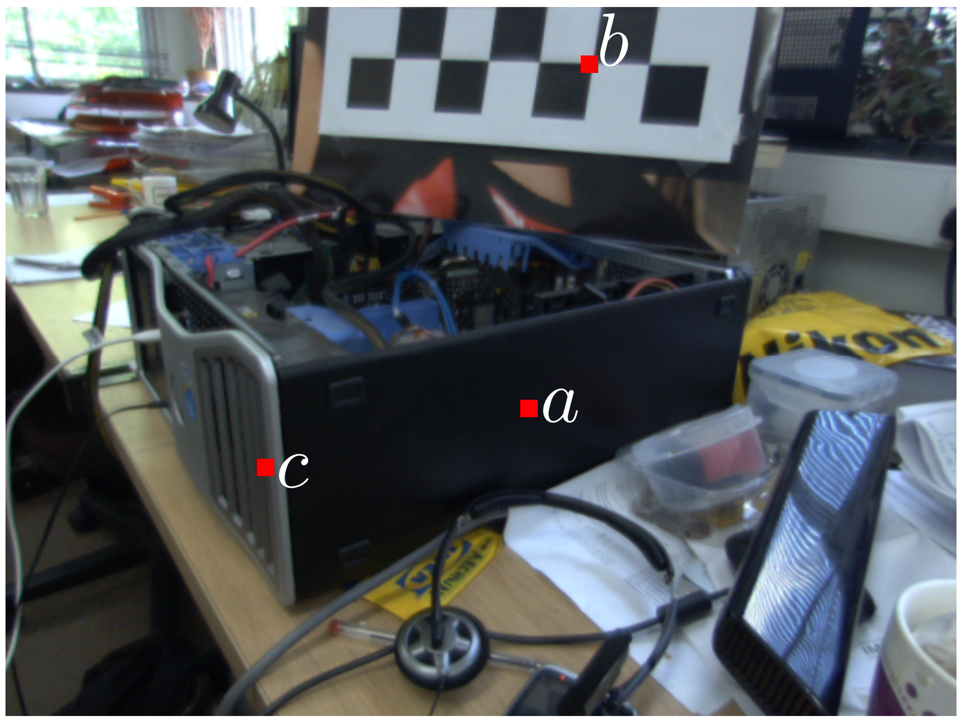


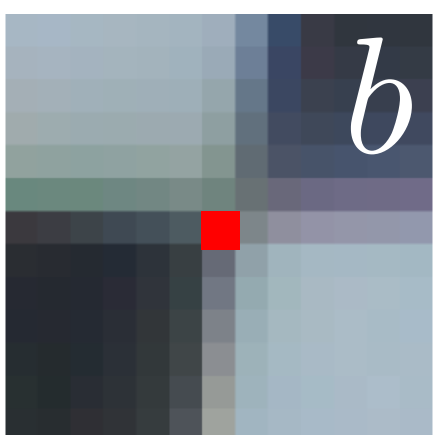


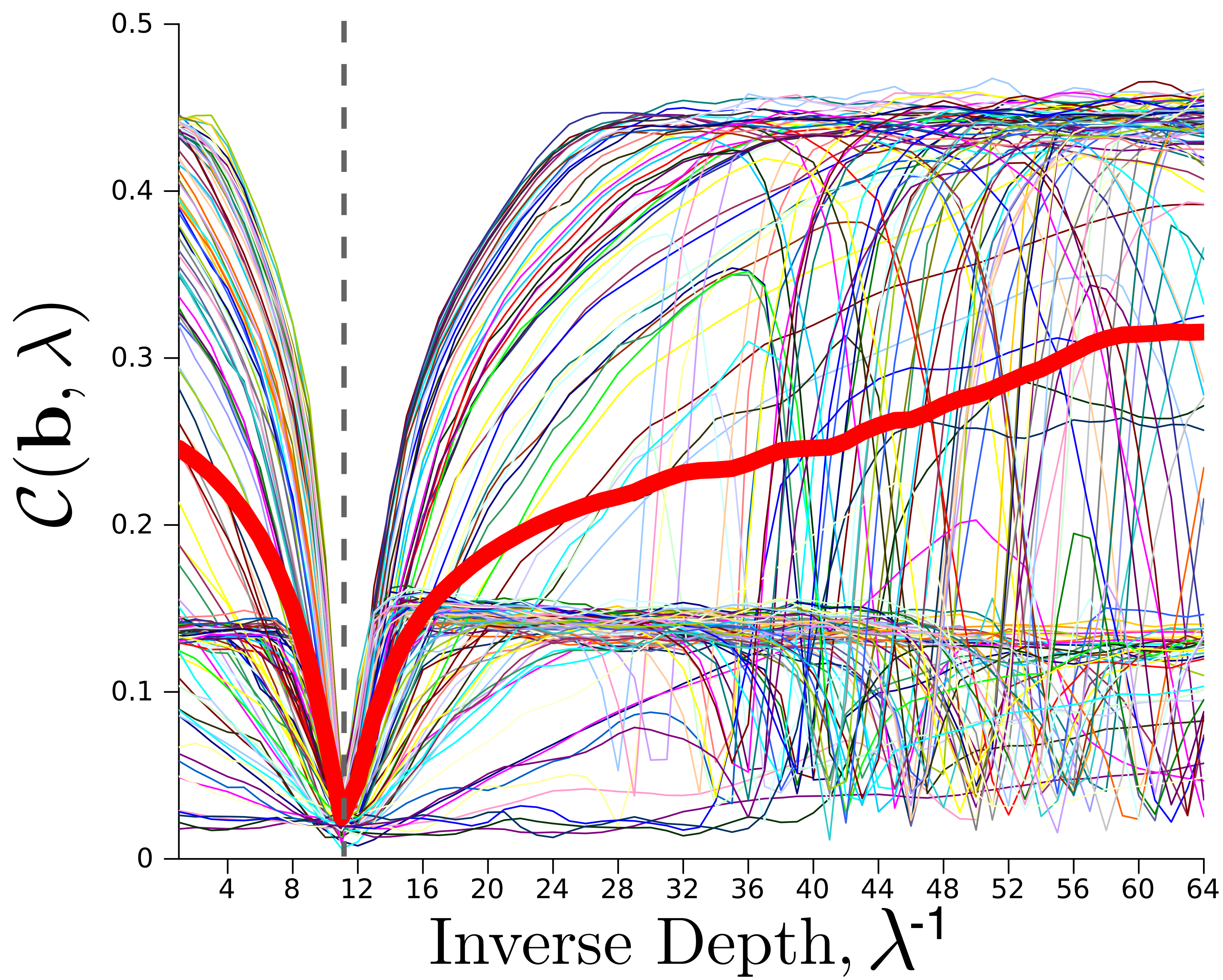


R. A. Newcombe, S. J. Lovegrove and A. J. Davison "DTAM: Dense Tracking and Mapping in Real-Time", ICCV 2011.

DTAM - Example

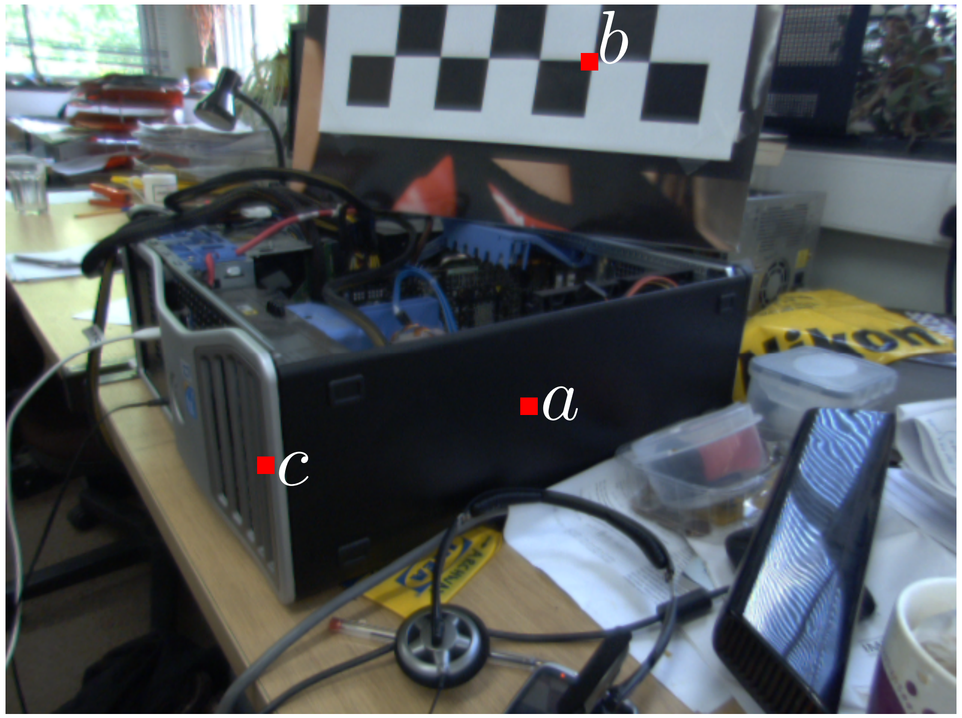


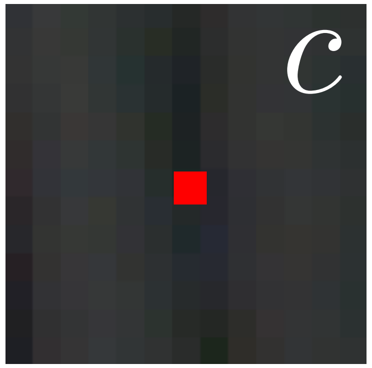


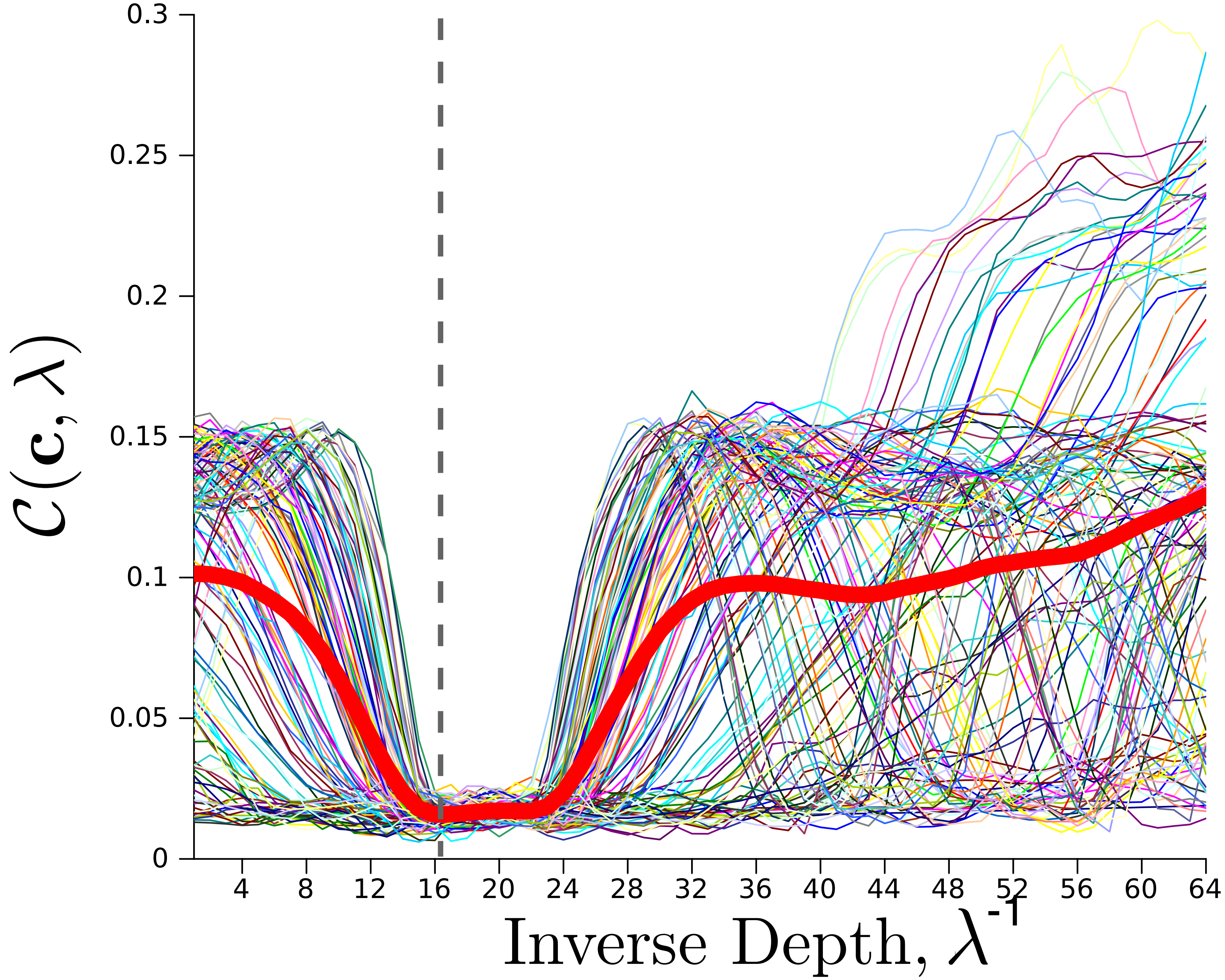


R. A. Newcombe, S. J. Lovegrove and A. J. Davison "DTAM: Dense Tracking and Mapping in Real-Time", ICCV 2011.

DTAM - Example







R. A. Newcombe, S. J. Lovegrove and A. J. Davison "DTAM: Dense Tracking and Mapping in Real-Time", ICCV 2011.

DTAM - Geometric Prior

- Newcombe et al. proposed the employment of a geometric prior on depths,

$$\arg\min_{\boldsymbol{\lambda}} \sum_{n=1}^{N} C(\mathbf{x}_n, \lambda_n) + g(\mathbf{x}_n)||\nabla * \lambda_n^{-1}||_\epsilon$$

$$g(\mathbf{x}) = \exp(-\alpha||\nabla T(\mathbf{x})||_2^\beta)$$

DTAM - Geometric Prior

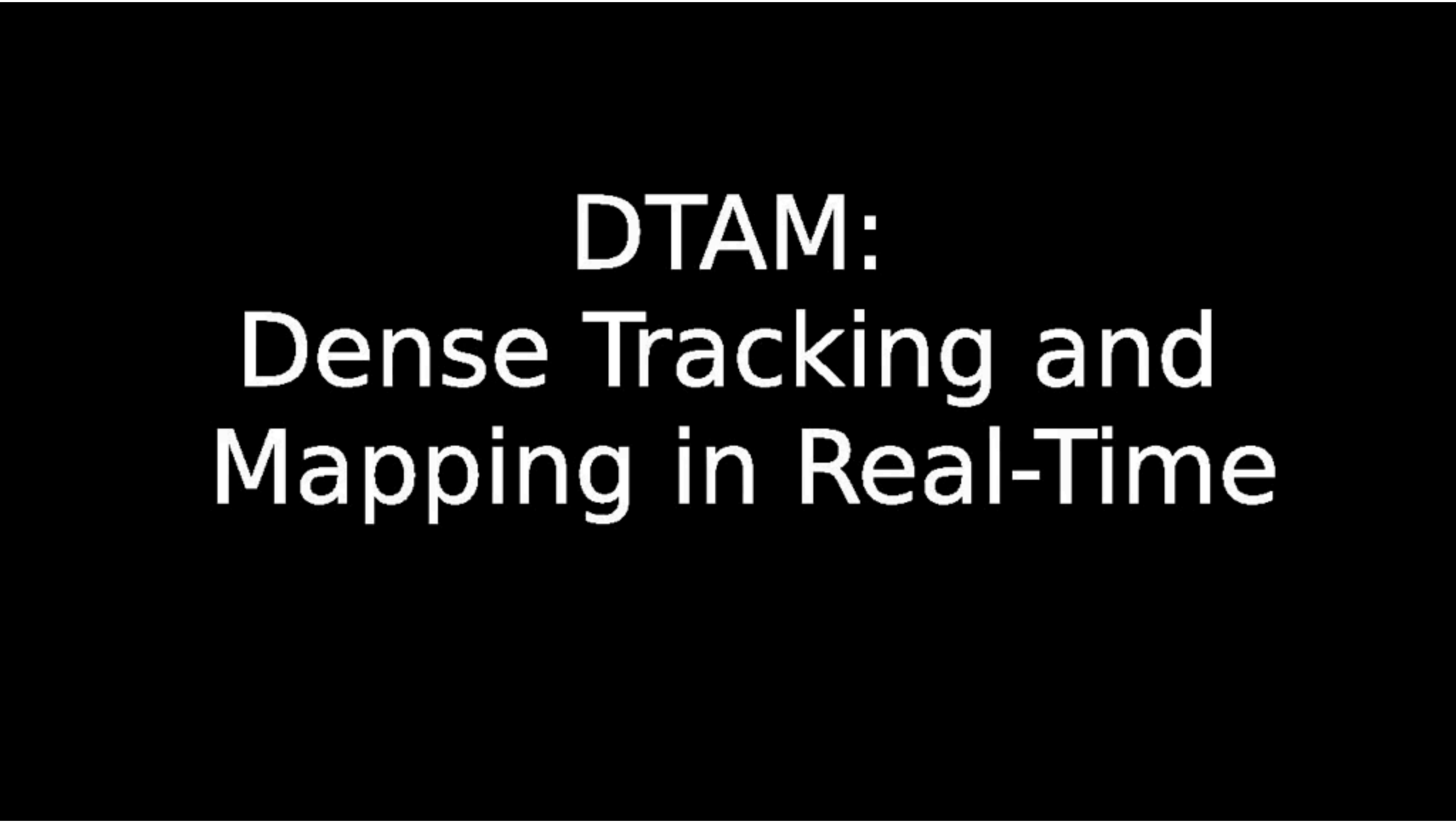
- Newcombe et al. proposed the employment of a geometric prior on depths,

$$\arg\min_{\boldsymbol{\lambda}} \sum_{n=1}^{N} C(\mathbf{x}_n, \lambda_n) + g(\mathbf{x}_n)||\nabla * \lambda_n^{-1}||_\epsilon$$

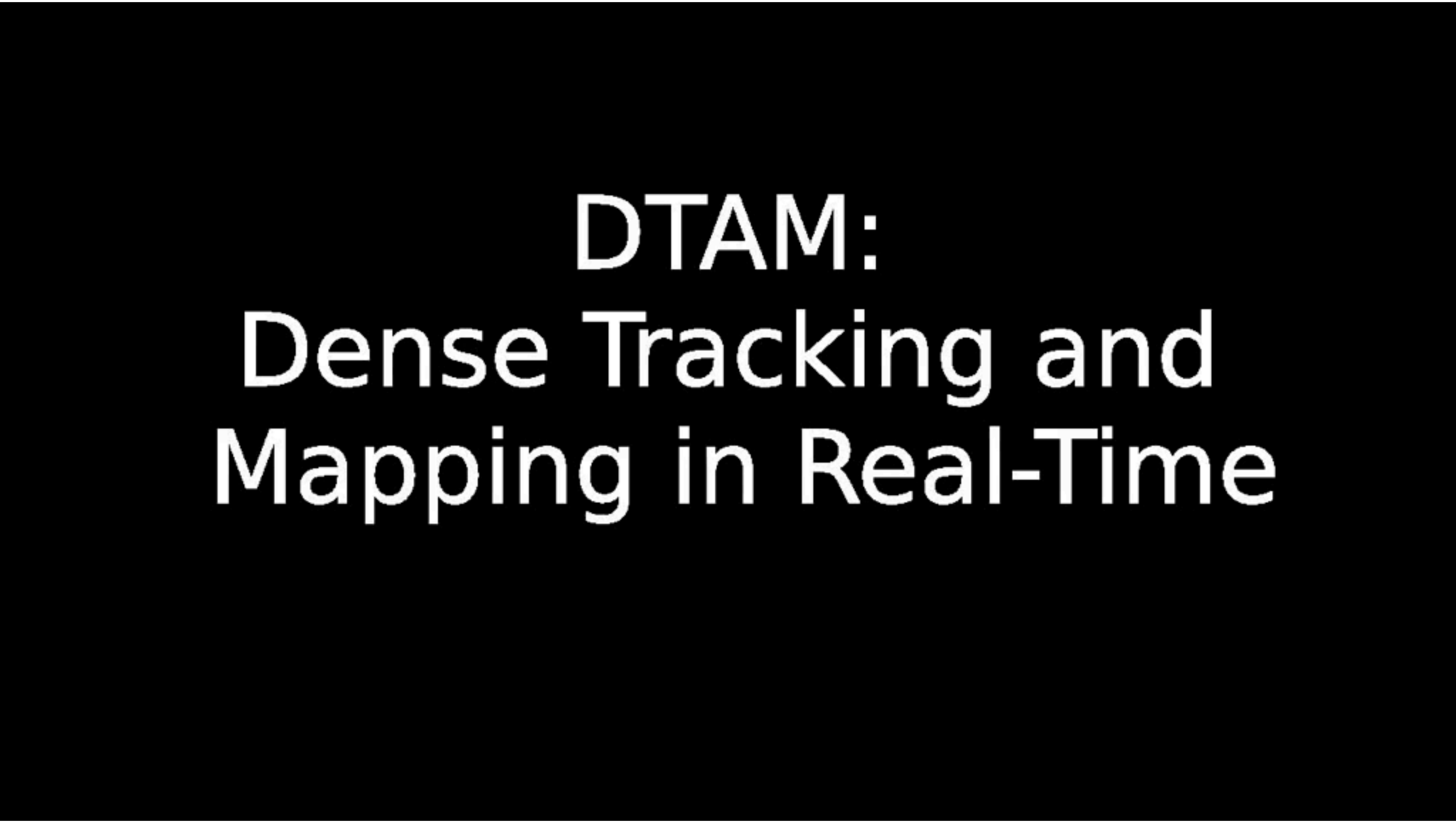
$$g(\mathbf{x}) = \exp(-\alpha||\nabla T(\mathbf{x})||_2^\beta)$$

What do you think the prior is doing?

DTAM - Video

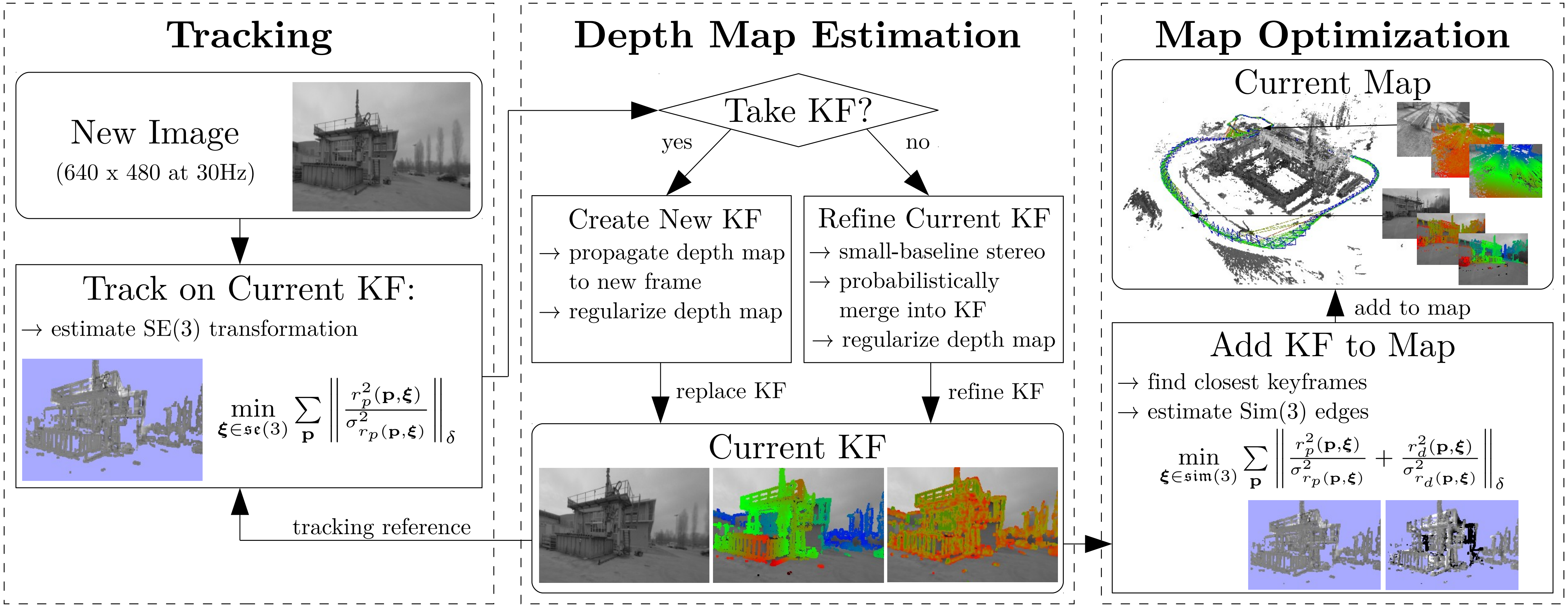


DTAM - Video



LSD SLAM

- A drawback to DTAM is that the depth estimation is a volumetric method and therefore requires state of the art GPU to run in real-time.
- Engel et al. recently proposed **L**arge-**S**cale **D**irect Monocular SLAM that circumvents this limitation.



J. Engel, T. Schops, and D. Cremers. LSD-SLAM: Large-scale direct monocular slam. In European Conference on Computer Vision, pages 834–849. Springer, 2014.

LSD SLAM

- Depth map can instead be represented as a Gaussian distribution.

$$\mathcal{C}(\mathbf{x}, \lambda) = \sigma(\mathbf{x})^{-2} ||\lambda^{-1} - \mu(\mathbf{x})||_2^2$$

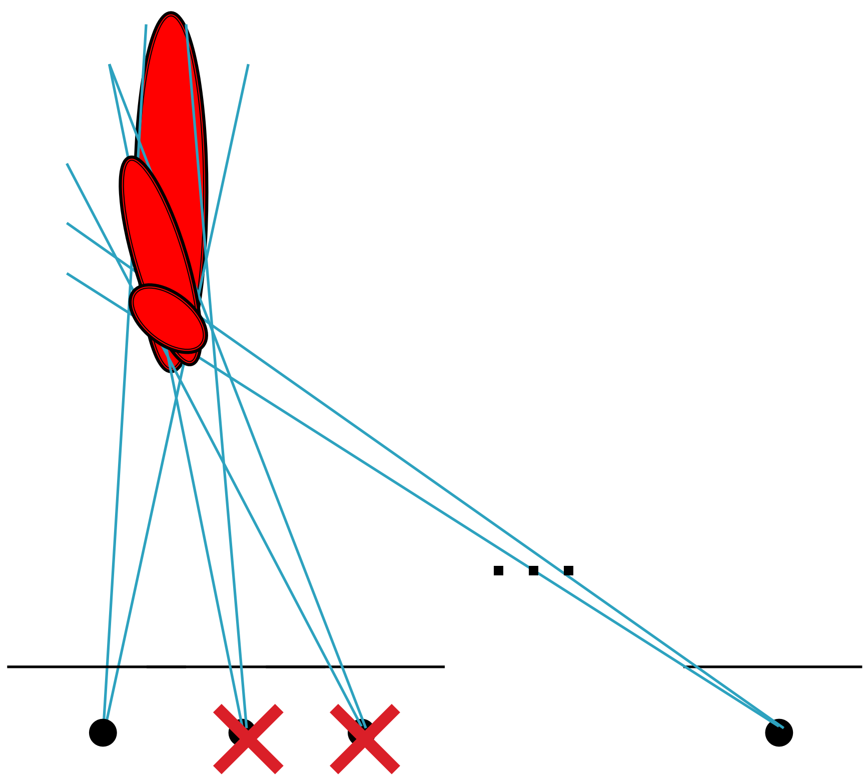
- Much more efficient than DTAM's volumetric approach.
- Engel et al. also used a similar (but more efficient) geometric prior to DTAM.

J. Engel, T. Schops, and D. Cremers. LSD-SLAM: Large-scale direct monocular slam. In European Conference on Computer Vision, pages 834–849. Springer, 2014.

Reminder: Keyframe Selection

- Rule of thumb: add a keyframe when,

$$\frac{\textit{keyframe distance}}{\textit{average-depth}} > \textit{threshold } (\sim 10\text{-}20\ \%)$$



Taken from D. Scaramuzza "Tutorial on Visual Odometry".

Depths across Keyframes

First camera:

$$\lambda_1 \tilde{\mathbf{x}}_1 = \mathbf{w}$$

Second camera:

$$\lambda_2 \tilde{\mathbf{x}}_2 = \boldsymbol{\Omega}\mathbf{w} + \boldsymbol{\tau}$$

Substituting:

$$\lambda_2 \tilde{\mathbf{x}}_2 = \lambda_1 \boldsymbol{\Omega} \tilde{\mathbf{x}}_1 + \boldsymbol{\tau}$$

- Depth from keyframe 1 can be propagated to keyframe 2.

LSD SLAM - Details

- To boot-strap LSD slam it is sufficient to initialize a random depth map with large variance.
- Given sufficient translation camera motion in the first seconds of operation the algorithm “locks” to a good configuration.
- Map is continuously optimized in the background using pose graph optimization.

J. Engel, T. Schops, and D. Cremers. LSD-SLAM: Large-scale direct monocular slam. In European Conference on Computer Vision, pages 834–849. Springer, 2014.

LSD SLAM - Details

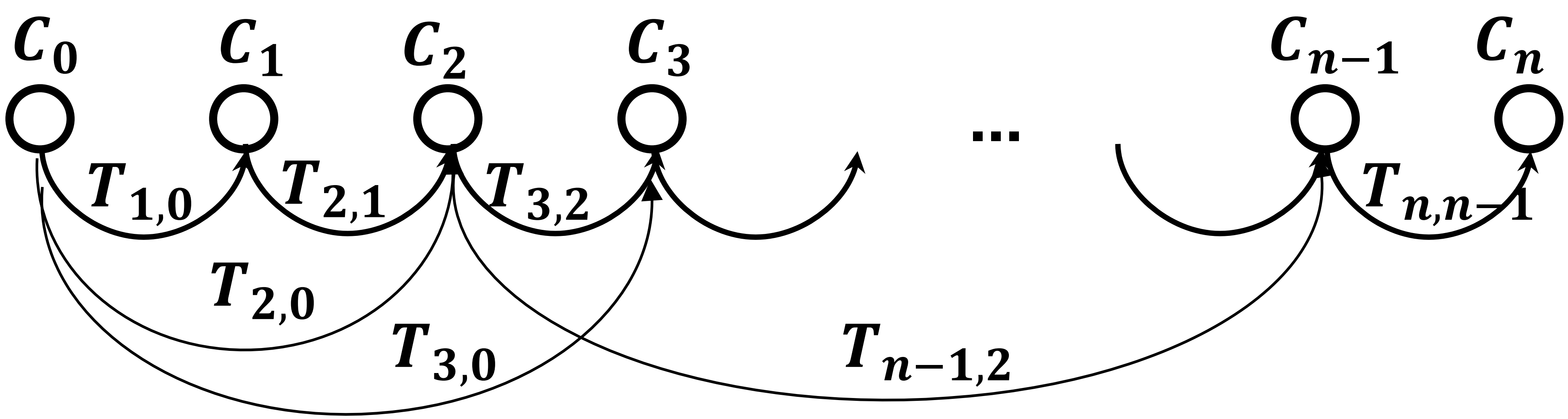
- To boot-strap LSD slam it is sufficient to initialize a random depth map with large variance.
- Given sufficient translation camera motion in the first seconds of operation the algorithm “locks” to a good configuration.
- Map is continuously optimized in the background using pose graph optimization.

Why is BA not employed?

J. Engel, T. Schops, and D. Cremers. LSD-SLAM: Large-scale direct monocular slam. In European Conference on Computer Vision, pages 834–849. Springer, 2014.

Pose Graph Optimization

- Similar to BA, but does not optimize over 3D points.
- Employs knowledge that transformations can be computed between non-adjacent frames.



$$\sum_i \sum_j \left\| C_i - T_{ij} C_j \right\|^2$$

Taken from D. Scaramuzza "Tutorial on Visual Odometry".

LSD SLAM - Details

- Source code to LSD SLAM can be found at,

  https://github.com/tum-vision/lsd_slam

- ROS is only used for input and output, facilitating easy portability to other platforms.

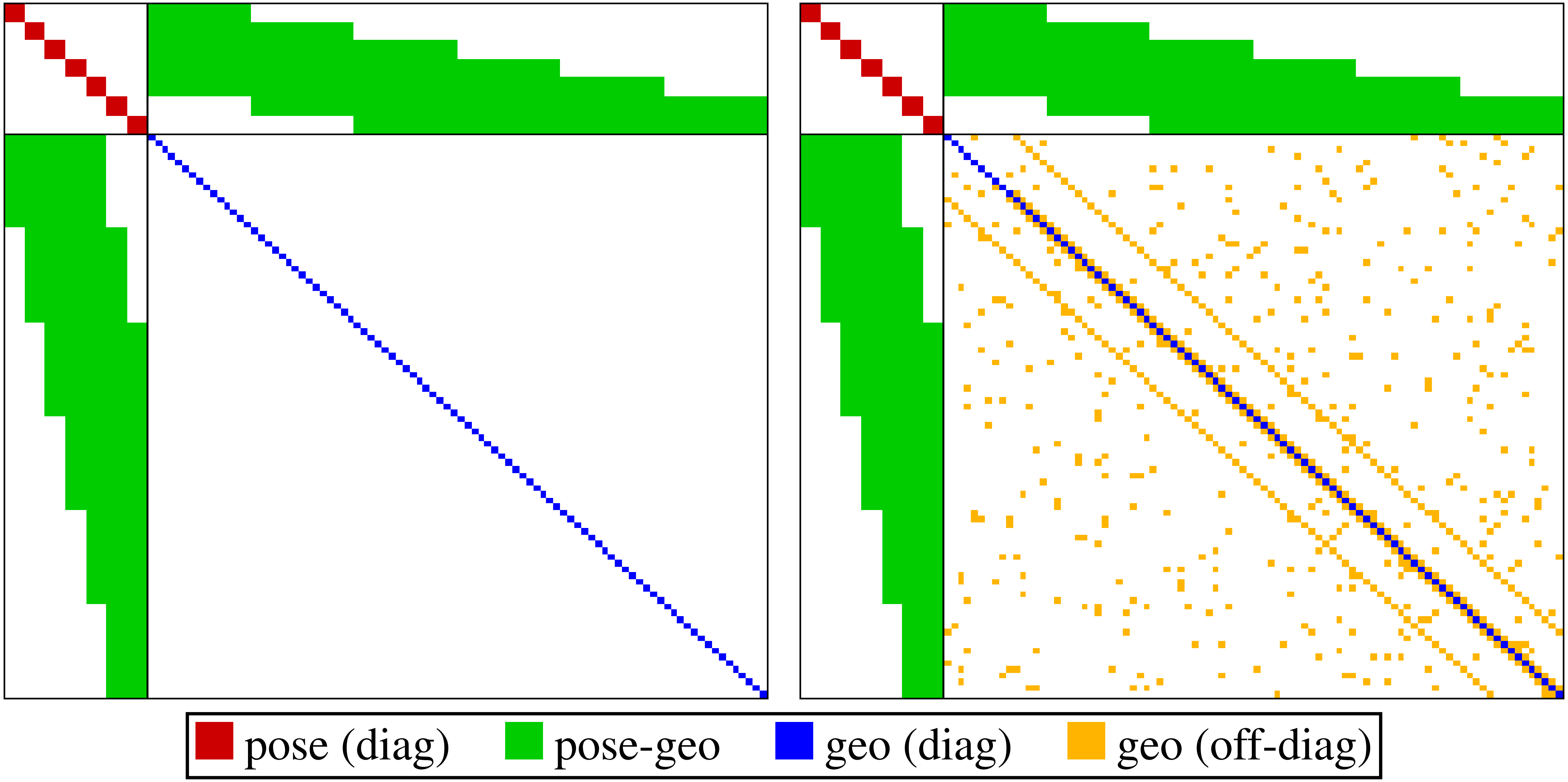
J. Engel, T. Schops, and D. Cremers. LSD-SLAM: Large-scale direct monocular slam. In European Conference on Computer Vision, pages 834–849. Springer, 2014.

Today

- Direct vs. Feature based methods
- Dense SLAM
- Semi-Dense SLAM
- Photometric Bundle Adjustment

Drawbacks to Geometric Prior

- Geometric prior used in DTAM and LSD slam can have unwanted effects when solving BA problem.



J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. arXiv preprint arXiv:1607.02565, 2016.

Drawbacks to Geometric Prior

- While geometric prior makes 3D reconstruction denser, locally more accurate and visual appealing.

- Has additional drawbacks as it can introduce bias and thereby reduce long-term, large-scale accuracy.

J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. arXiv preprint arXiv:1607.02565, 2016.

Semi-Dense SLAM

- Recently, the community has been exploring the idea of semi-dense direct SLAM.
- In this new approach ALL parameters are solved simultaneously within a photometric bundle adjustment framework.
- Can naturally sample all parts of image that contain image gradient information.

$$\begin{aligned}\mathcal{T}(\mathbf{x}_n) &= \mathcal{I}_f(\mathcal{W}\{\mathbf{x}_n; \boldsymbol{\theta}_f \circ \Delta\boldsymbol{\theta}_f, \lambda_n + \Delta\lambda_n\}) \\ &\approx \mathcal{I}_f(\mathcal{W}\{\mathbf{x}_n; \boldsymbol{\theta}_f, \lambda_n\}) + [\mathbf{A}_n^f, \mathbf{B}_n^f] \begin{bmatrix} \Delta\boldsymbol{\theta}_f \\ \Delta\lambda_n \end{bmatrix}\end{aligned}$$

J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. arXiv preprint arXiv:1607.02565, 2016.
H. Alismail, B. Browning and S. Lucey "Photometric Bundle Adjustment for Vision-based SLAM", ACCV 2016.

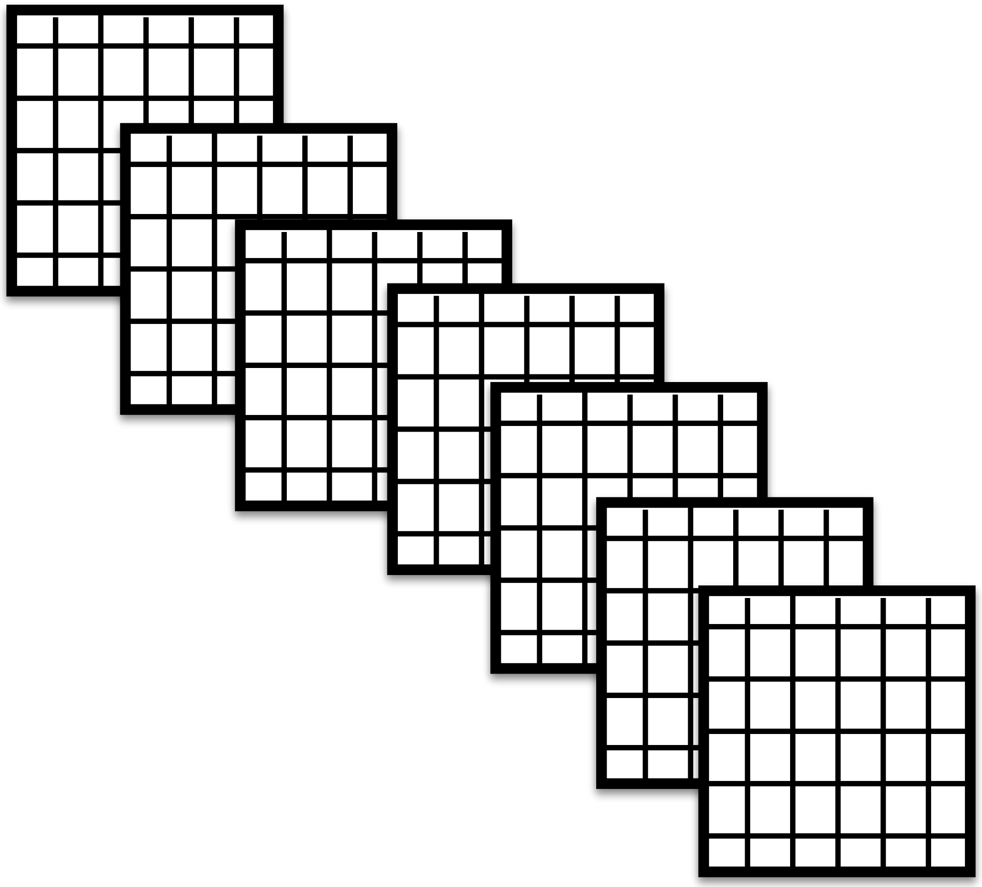
Semi-Dense SLAM

- Recently, the community has been exploring the idea of semi-dense direct SLAM.
- In this new approach ALL parameters are solved simultaneously within a photometric bundle adjustment framework.
- Can naturally sample all parts of image that contain image gradient information.

$$
\begin{aligned}
\mathcal{T}(\mathbf{x}_n) &= \mathcal{I}_f(\mathcal{W}\{\mathbf{x}_n; \boldsymbol{\theta}_f \circ \Delta\boldsymbol{\theta}_f, \lambda_n + \Delta\lambda_n\}) \\
&\approx \mathcal{I}_f(\mathcal{W}\{\mathbf{x}_n; \boldsymbol{\theta}_f, \lambda_n\}) + [\mathbf{A}_n^f, \mathbf{B}_n^f] \begin{bmatrix} \Delta\boldsymbol{\theta}_f \\ \Delta\lambda_n \end{bmatrix}
\end{aligned}
$$

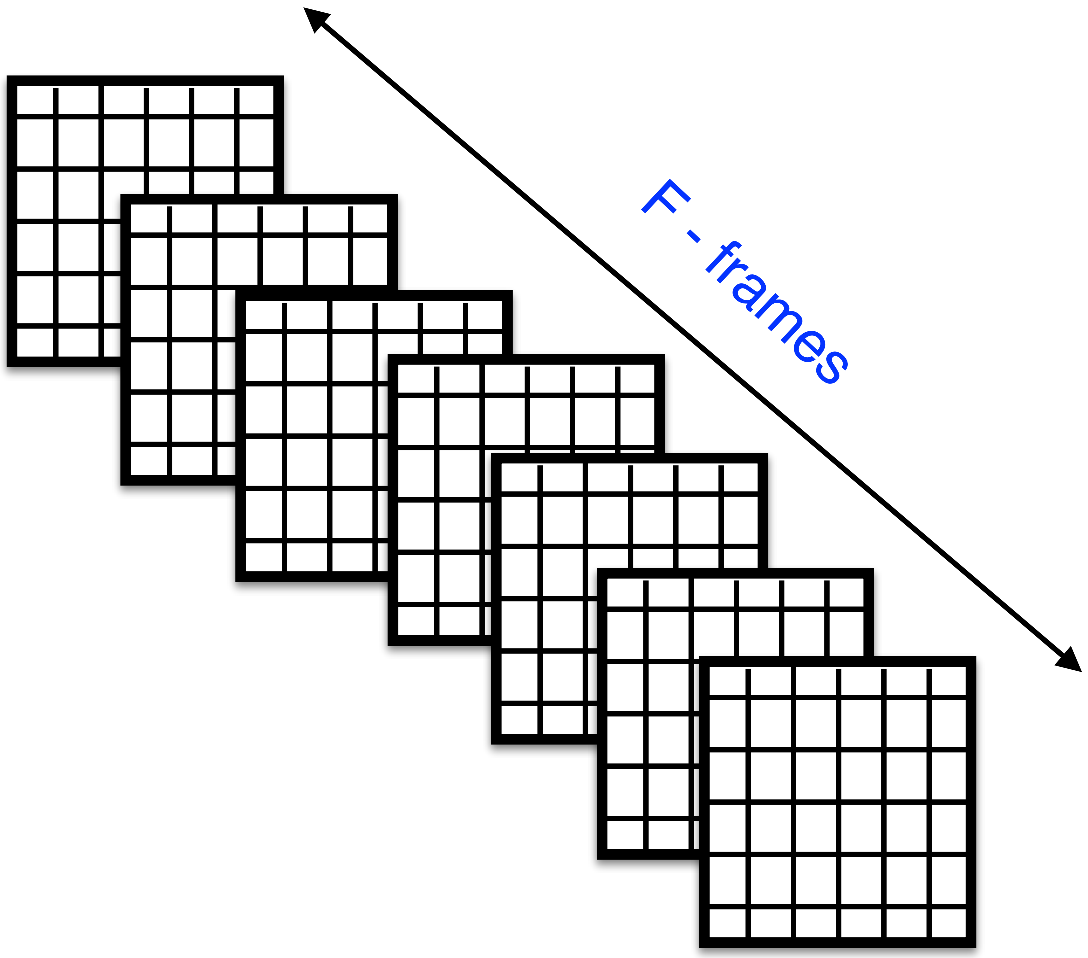
J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. arXiv preprint arXiv:1607.02565, 2016.
H. Alismail, B. Browning and S. Lucey "Photometric Bundle Adjustment for Vision-based SLAM", ACCV 2016.

Photometric Bundle Adjustment



J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. arXiv preprint arXiv:1607.02565, 2016.
H. Alismail, B. Browning and S. Lucey “Photometric Bundle Adjustment for Vision-based SLAM”, ACCV 2016.

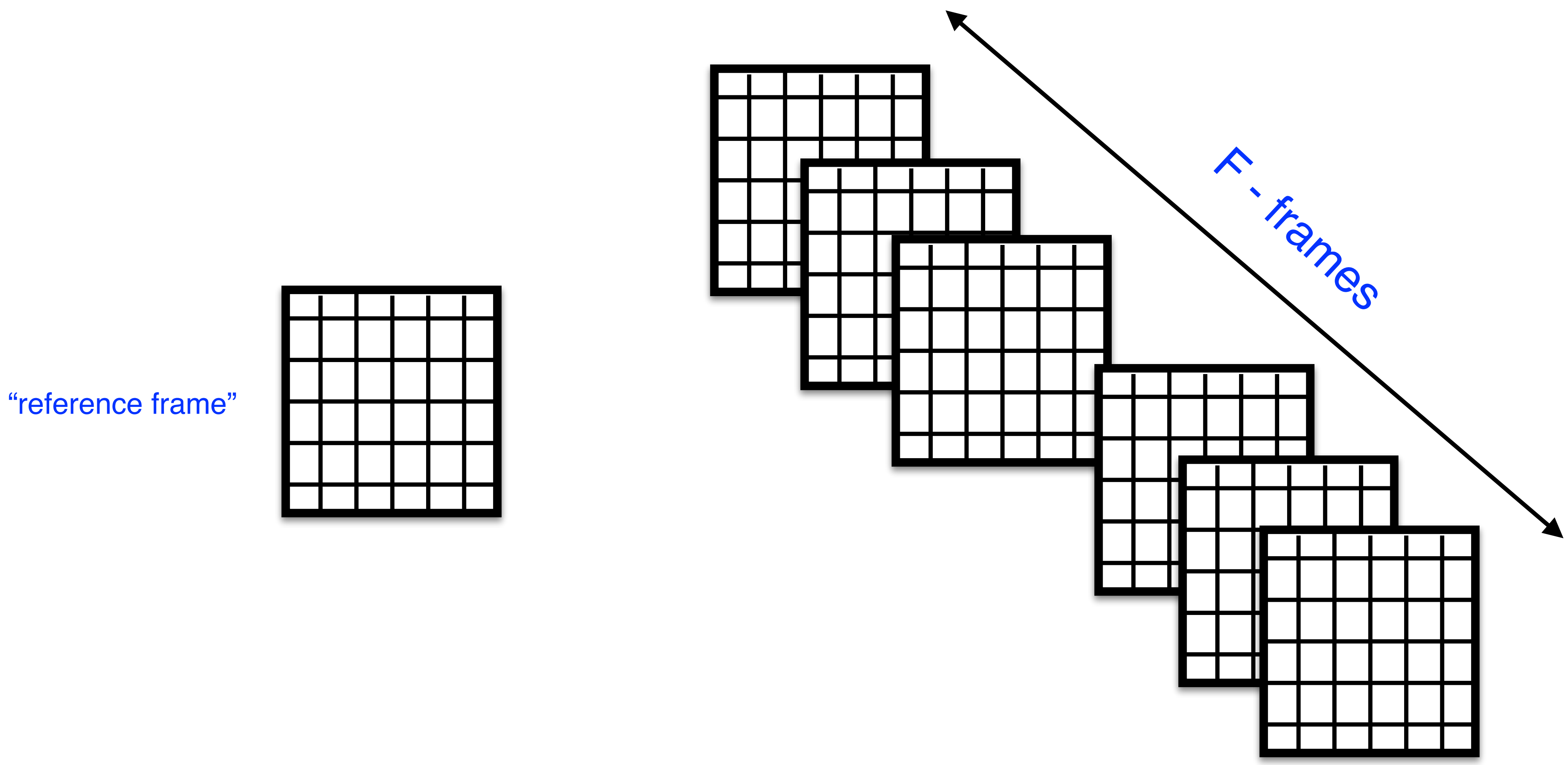
Photometric Bundle Adjustment



$$\sum_{r=1}^{F}$$

J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. arXiv preprint arXiv:1607.02565, 2016.
H. Alismail, B. Browning and S. Lucey "Photometric Bundle Adjustment for Vision-based SLAM", ACCV 2016.

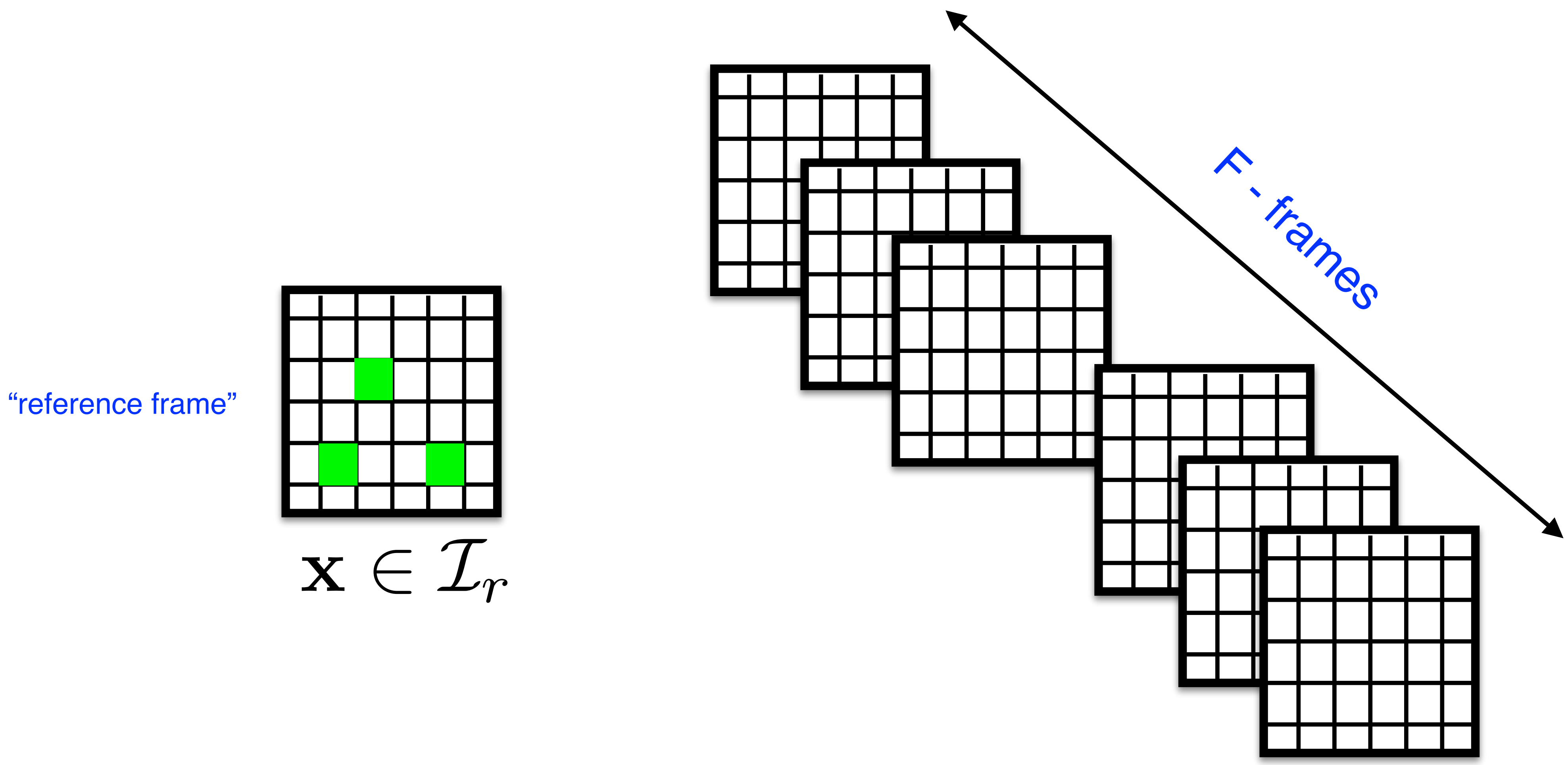
Photometric Bundle Adjustment



$$\sum_{r=1}^{F}$$

J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. arXiv preprint arXiv:1607.02565, 2016.
H. Alismail, B. Browning and S. Lucey "Photometric Bundle Adjustment for Vision-based SLAM", ACCV 2016.

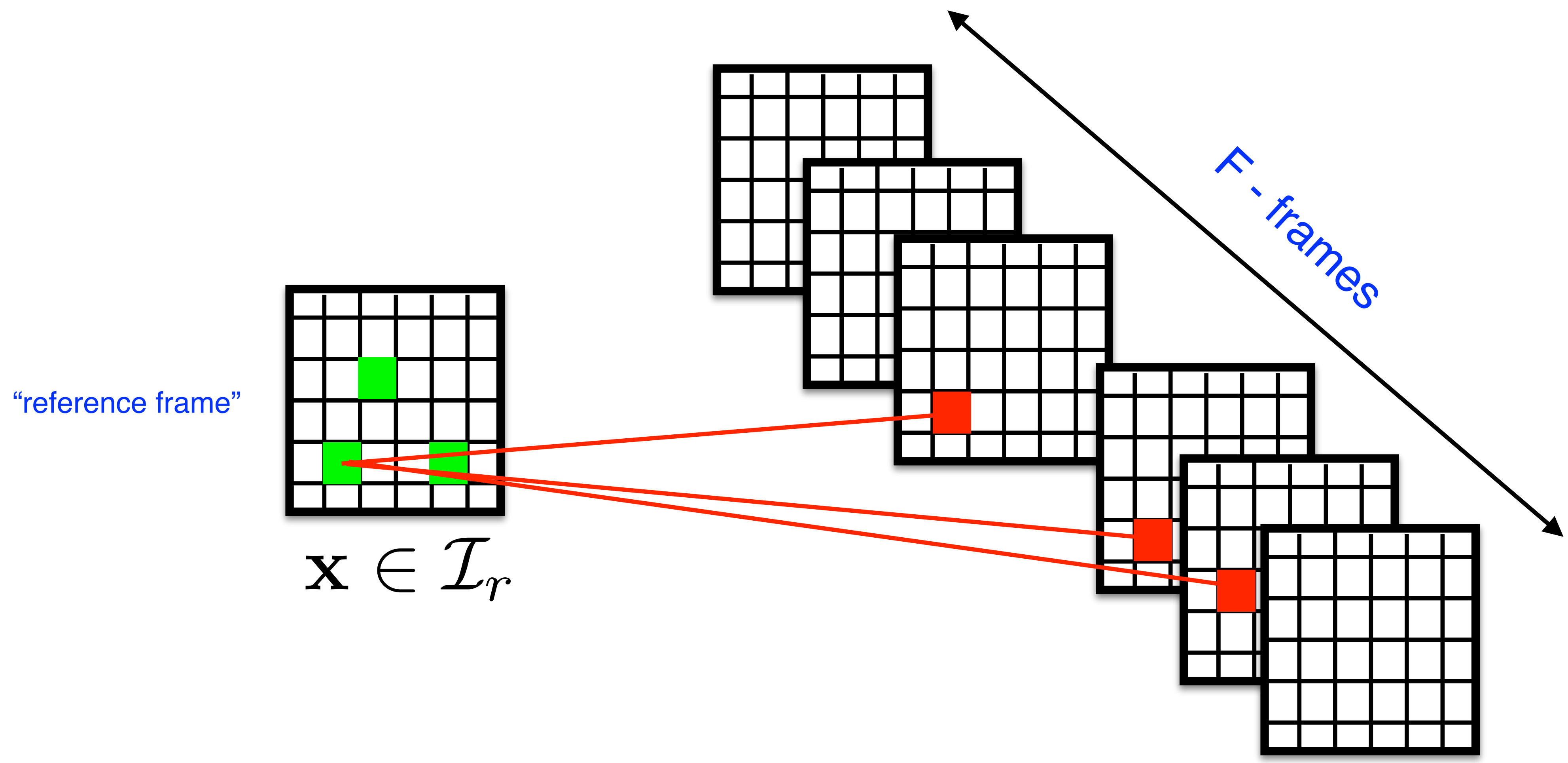
Photometric Bundle Adjustment



$$\sum_{r=1}^{F} \sum_{\mathbf{x} \in \mathcal{I}_r}$$

J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. arXiv preprint arXiv:1607.02565, 2016.
H. Alismail, B. Browning and S. Lucey "Photometric Bundle Adjustment for Vision-based SLAM", ACCV 2016.

Photometric Bundle Adjustment

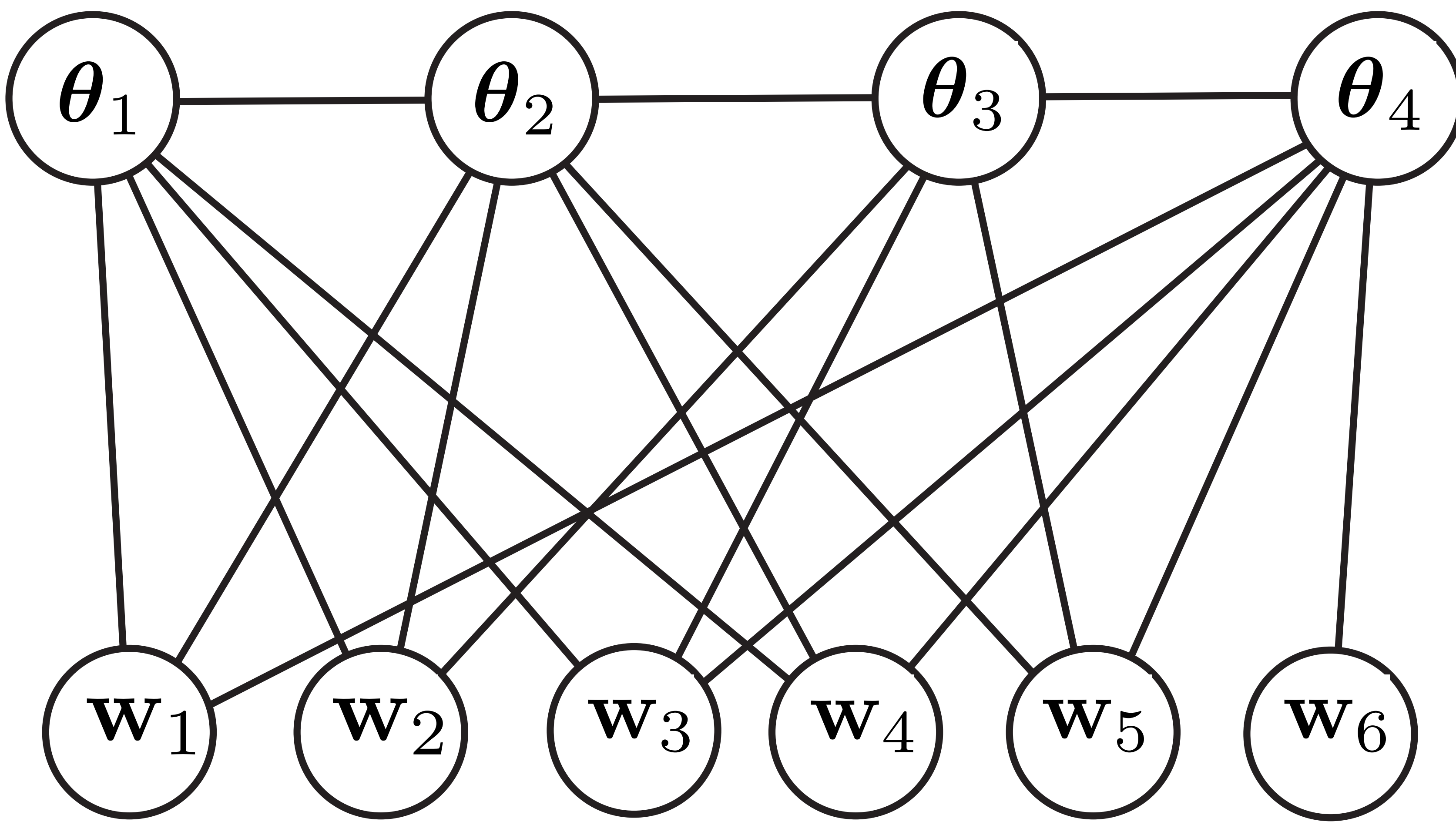


$$\arg\min_{\boldsymbol{\lambda},\boldsymbol{\theta}} \sum_{r=1}^{F} \sum_{\mathbf{x}\in\mathcal{I}_r} \sum_{f\in\mathrm{obs}(\mathbf{x})} ||\mathcal{I}_r(\mathbf{x}) - \mathcal{I}_f(\mathcal{W}\{\mathbf{x};\boldsymbol{\theta}_f,\lambda_r(\mathbf{x})\})||_2^2$$

J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. arXiv preprint arXiv:1607.02565, 2016.
H. Alismail, B. Browning and S. Lucey "Photometric Bundle Adjustment for Vision-based SLAM", ACCV 2016.

Reminder: SLAM = BA

- One can view the problem of SfM - Bundle Adjustment as doing inference on a Markov Random Field (MRF).
- **Problem** - becomes exponentially harder as times goes on.



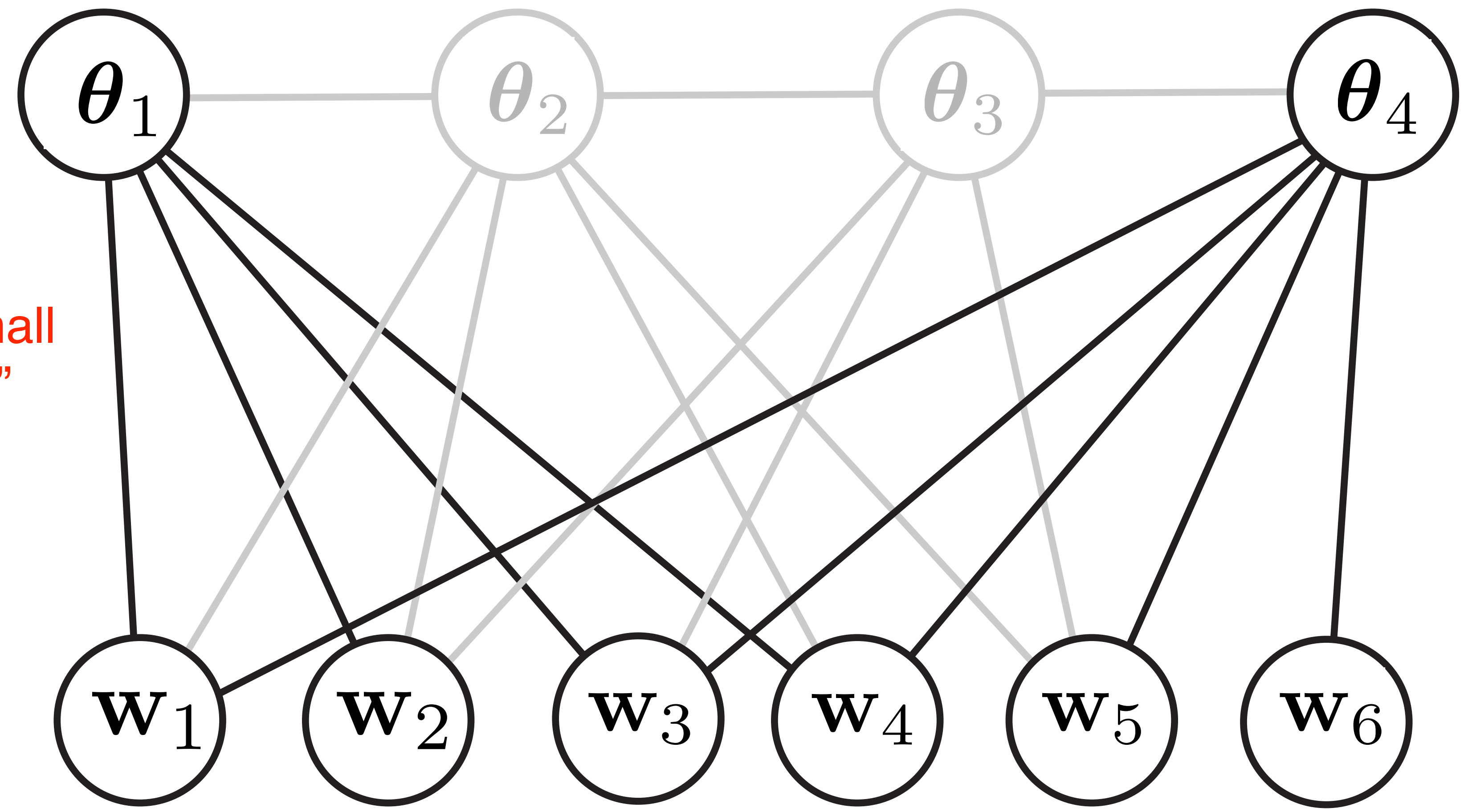
"edges based on visibility"

H. Strasdat, J. M. M. Montiel, and A. J. Davison, "Visual SLAM: Why filter?" Image and Vision Computing, vol. 30, no. 2, pp. 65–77, 2012.

Reminder: Keyframe

- A better strategy is to employ **keyframe BA**.
- Made popular by Klein & Murray's - Parallel Tracking and Mapping (PTAM) algorithm.

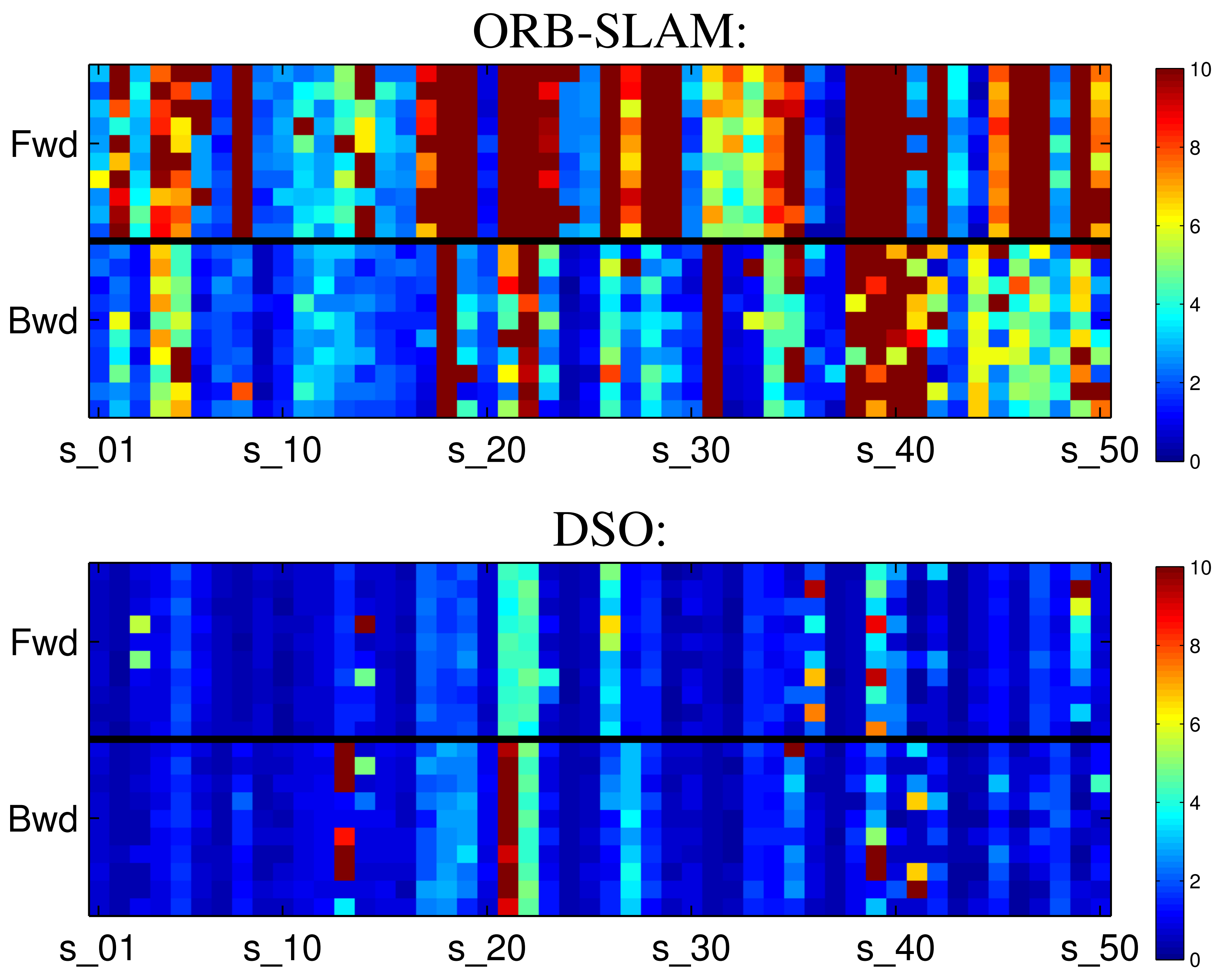
“remove **all** but a small subset of keyframes”



G. Klein and D. Murray, “Parallel tracking and mapping for small AR workspaces”, ISMAR 2007.

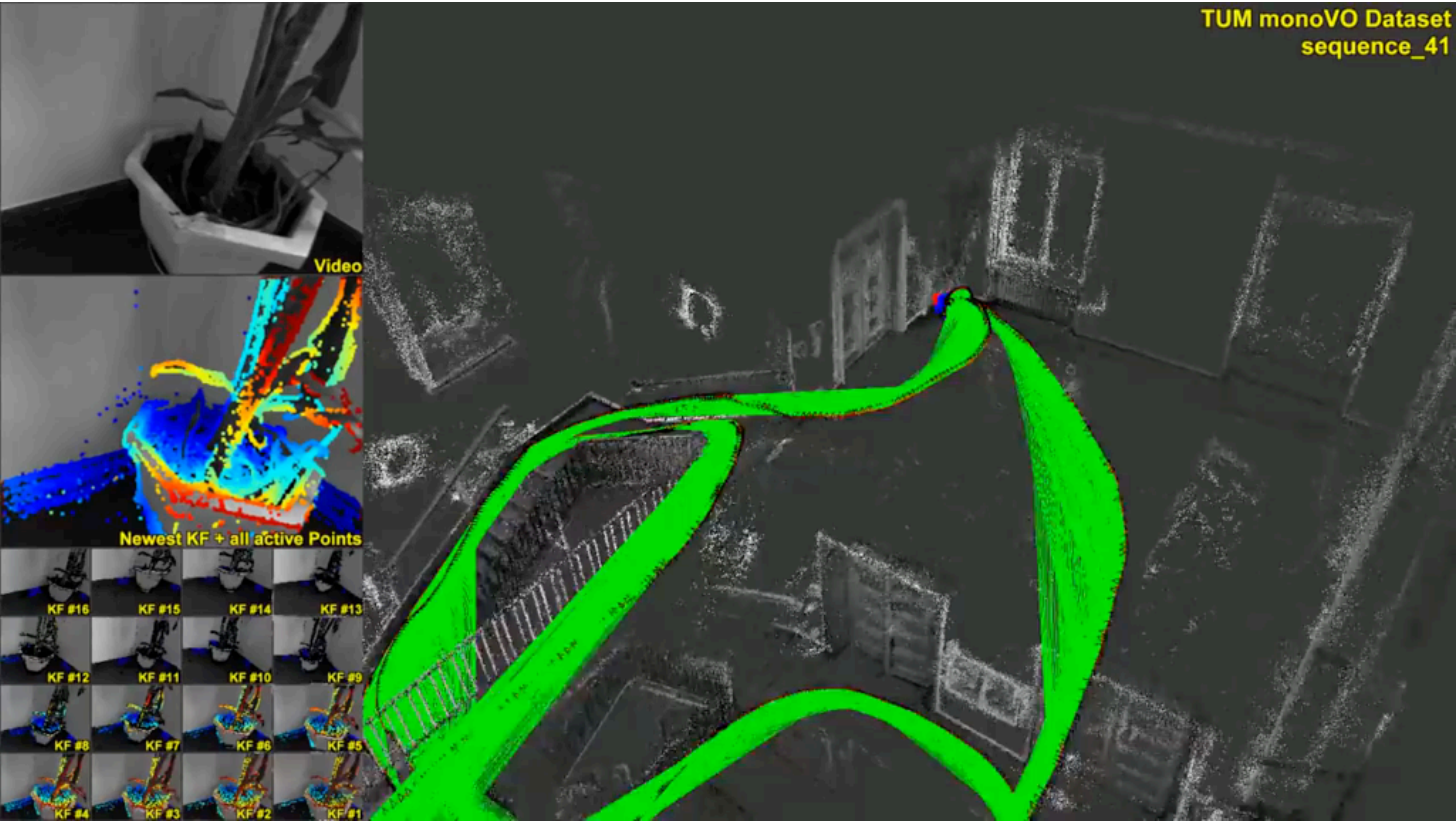
H. Strasdat, J. M. M. Montiel, and A. J. Davison, “Visual SLAM: Why filter?” Image and Vision Computing, vol. 30, no. 2, pp. 65–77, 2012.

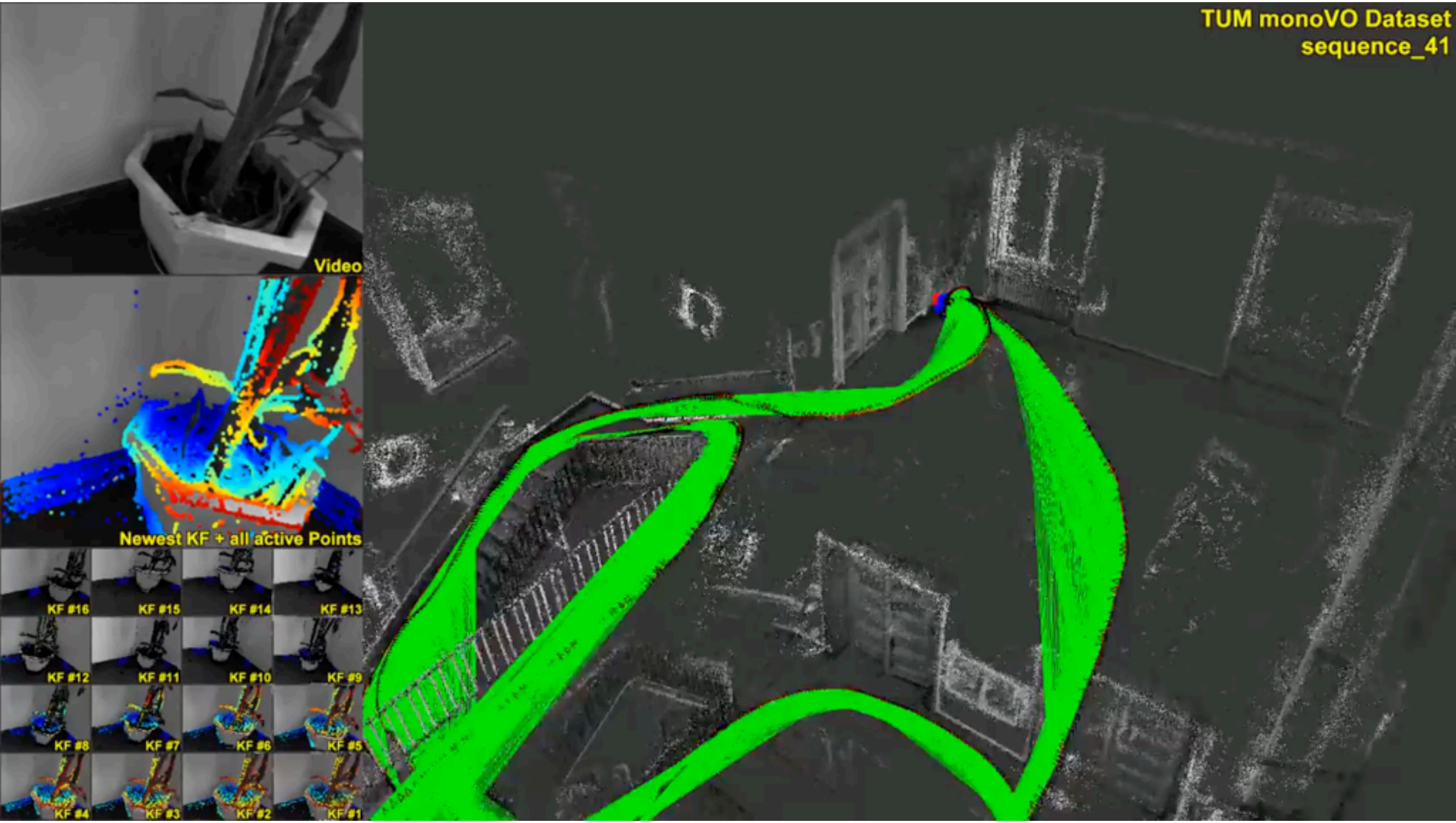
DSO SLAM



All error values for the TUM- monoVO dataset.

J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. arXiv preprint arXiv:1607.02565, 2016.


TUM monoVO Dataset
sequence_41
Video
Newest KF + all active Points
KF #16
KF #15
KF #14
KF #13
KF #12
KF #11
KF #10
KF #9
KF #8
KF #7
KF #6
KF #5
KF #4
KF #3
KF #2
KF #1


TUM monoVO Dataset
sequence_41
Video
Newest KF + all active Points
KF #16
KF #15
KF #14
KF #13
KF #12
KF #11
KF #10
KF #9
KF #8
KF #7
KF #6
KF #5
KF #4
KF #3
KF #2
KF #1