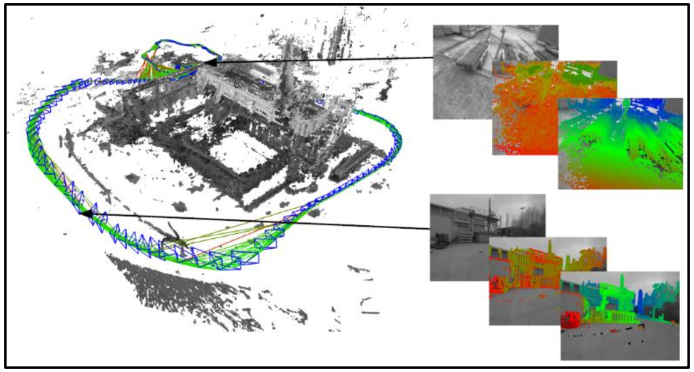
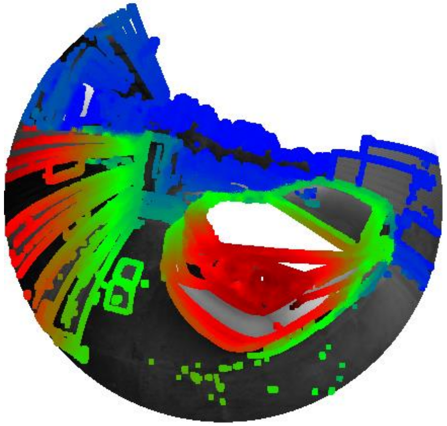
Semi-Dense Direct SLAM

Jakob Engel, Daniel Cremers
David Caruso, Thomas Schöps, Lukas von Stumberg,
Vladyslav Usenko, Jörg Stückler, Jürgen Sturm
Technical University Munich

Overview

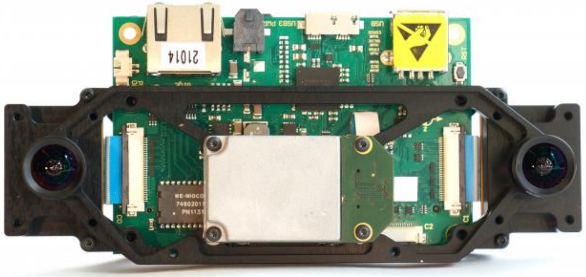


LSD-SLAM



Omnidirectional

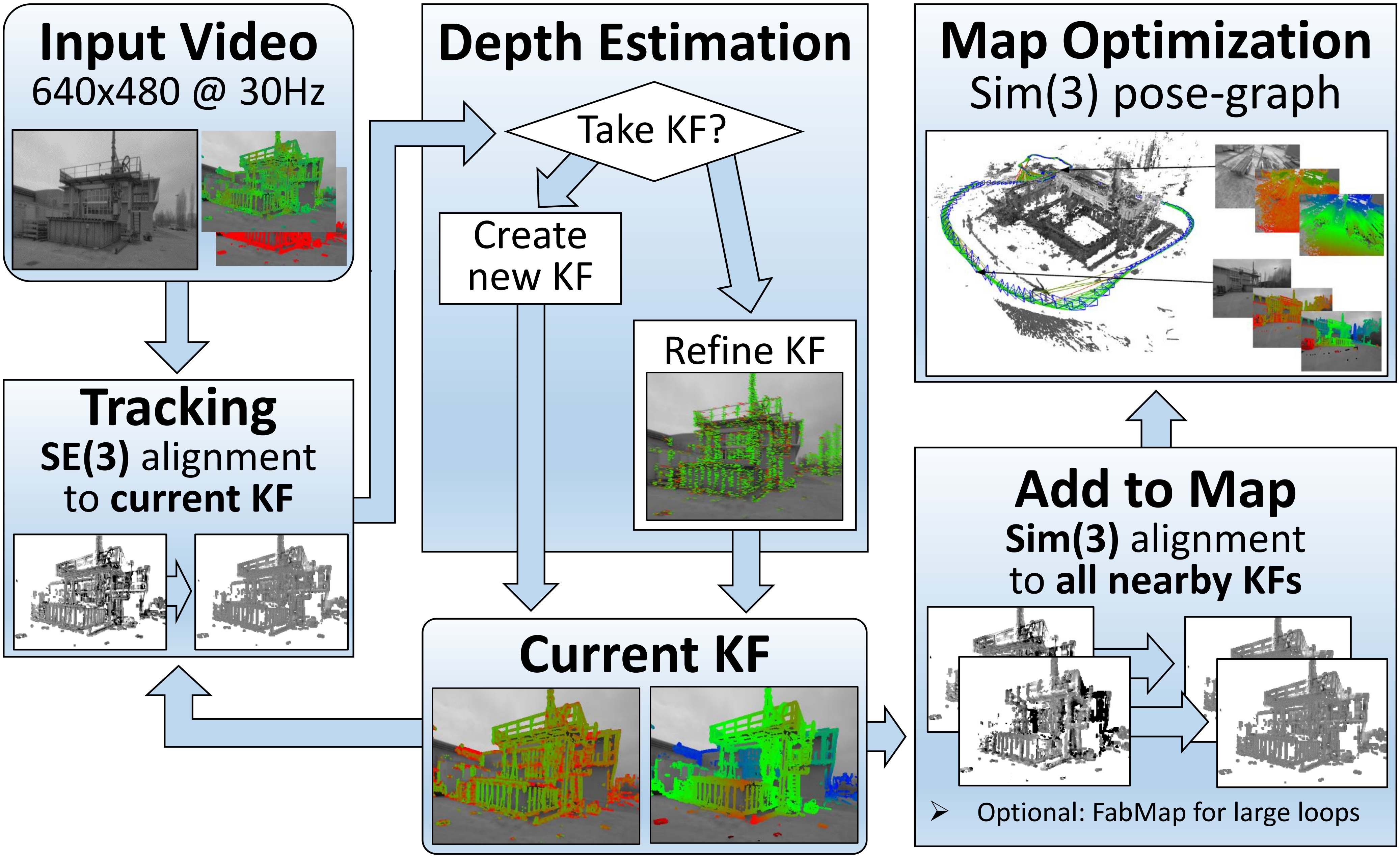
Stereo (+affine lighting)



Stereo + IMU

Quadcopter!

Overview

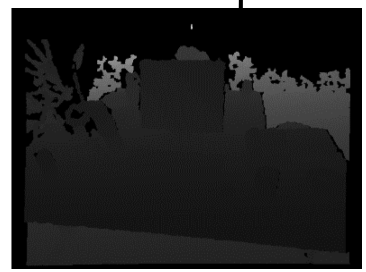

Input Video
640x480 @ 30Hz
Tracking
SE(3) alignment
to current KF
Depth Estimation
Take KF?
Create
new KF
Refine KF
Current KF
Map Optimization
Sim(3) pose-graph
Add to Map
Sim(3) alignment
to all nearby KFs
➢ Optional: FabMap for large loops

Direct Tracking

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

Camera Pose

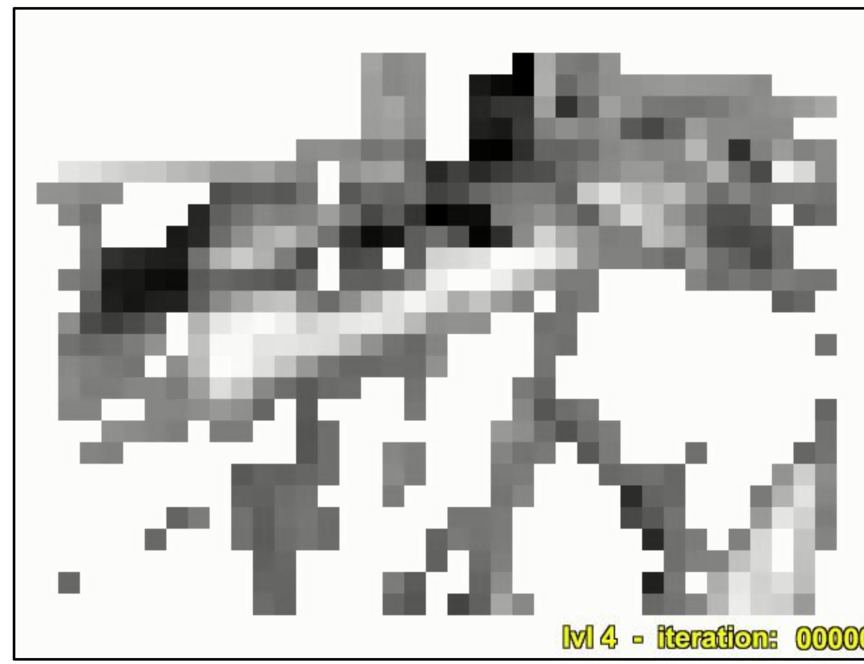
KF image



KF depth

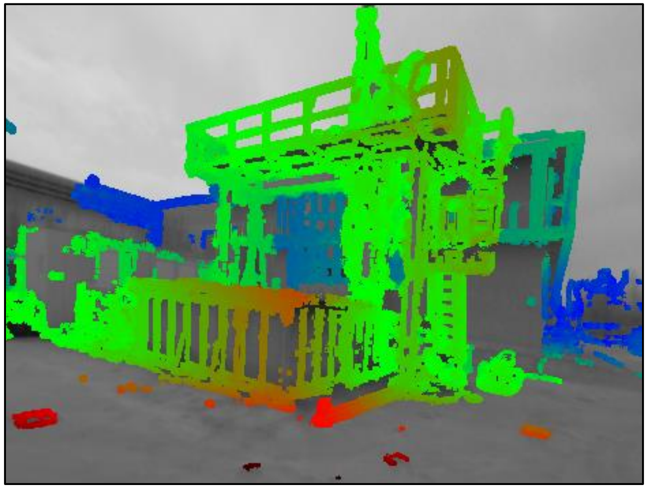
back-warped new frame

- minimize using **Gauss-Newton / LM** Algorithm
- Coarse-to-fine + Huber norm + statistical norm.

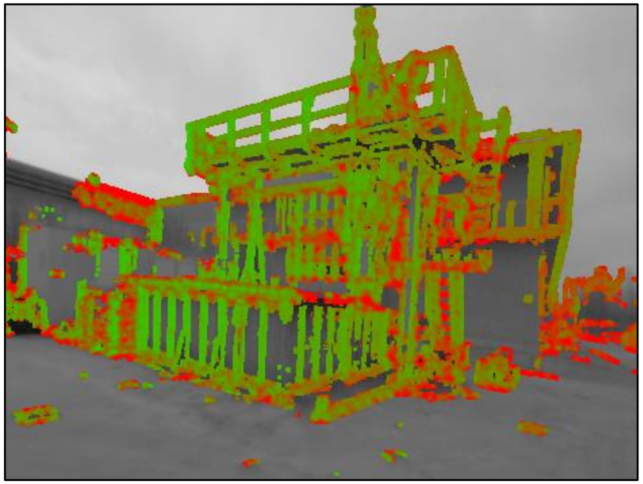


Depth Estimation

image



inverse depth

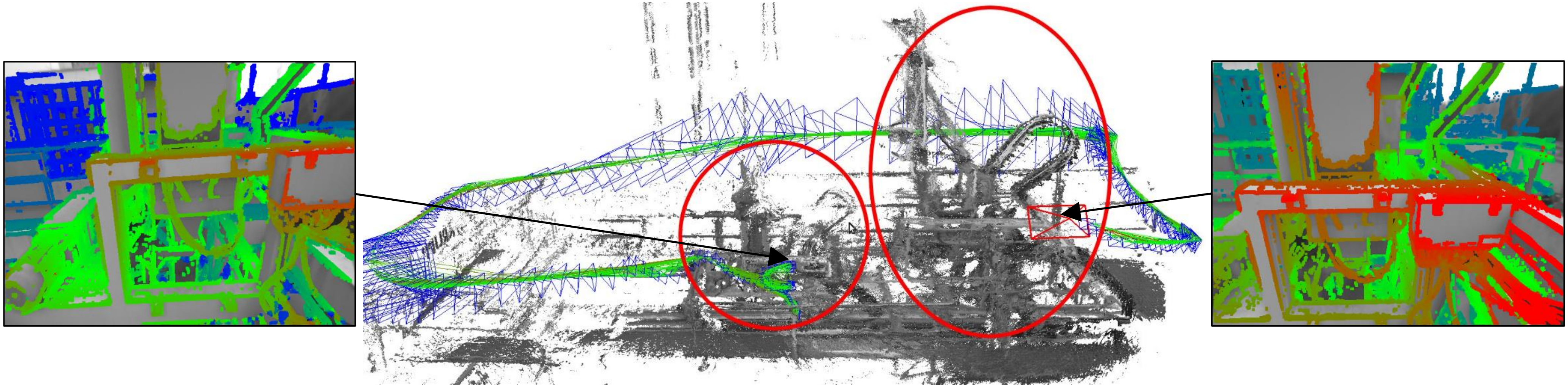


inverse depth variance

- filter over many (small-baseline) stereo-correspondences.
- small baseline + epipolar constraint + prior
  -> small search region („track “ instead of „detect “)
- only use „good“ (sufficiently constrained) pixel.
- Edge-preserving smoothing
- Distance-based KF selection

Engel, Sturm, Cremers; ICCV ´13

Global Mapping



- **Direct Tracking with scale (on Sim(3)):**

$$E(\boldsymbol{\xi}) = \sum_{\mathbf{x}\in\Omega_1}\Big(\big(I_1(\mathbf{x}) - I_2(\mathbf{x}')\big)^2 + \big([\mathbf{x}']_3 - D_2(\mathbf{x}')\big)^2 \Big)$$

$\cancel{\mathfrak{se}(3)}$

$\mathfrak{sim}(3)$ with $\mathbf{x}' := \omega(\mathbf{x}, D_1(\mathbf{x}), \boldsymbol{\xi})$ (warped point)

+ GN optimization + multi-resolution + Huber norm + statistical norm.

- **Optimize pose-graph on Sim(3)**

$$E(\boldsymbol{\xi}_{1W} \ldots \boldsymbol{\xi}_{nW}) := \sum_{(\boldsymbol{\xi}_{ij}, \boldsymbol{\Sigma}_{ij})\in\mathcal{E}} (\boldsymbol{\xi}_{ij} \circ \boldsymbol{\xi}_{iW}^{-1} \circ \boldsymbol{\xi}_{jW})^T \boldsymbol{\Sigma}_{ij}^{-1} (\boldsymbol{\xi}_{ij} \circ \boldsymbol{\xi}_{iW}^{-1} \circ \boldsymbol{\xi}_{jW}).$$

LSD-SLAM Result

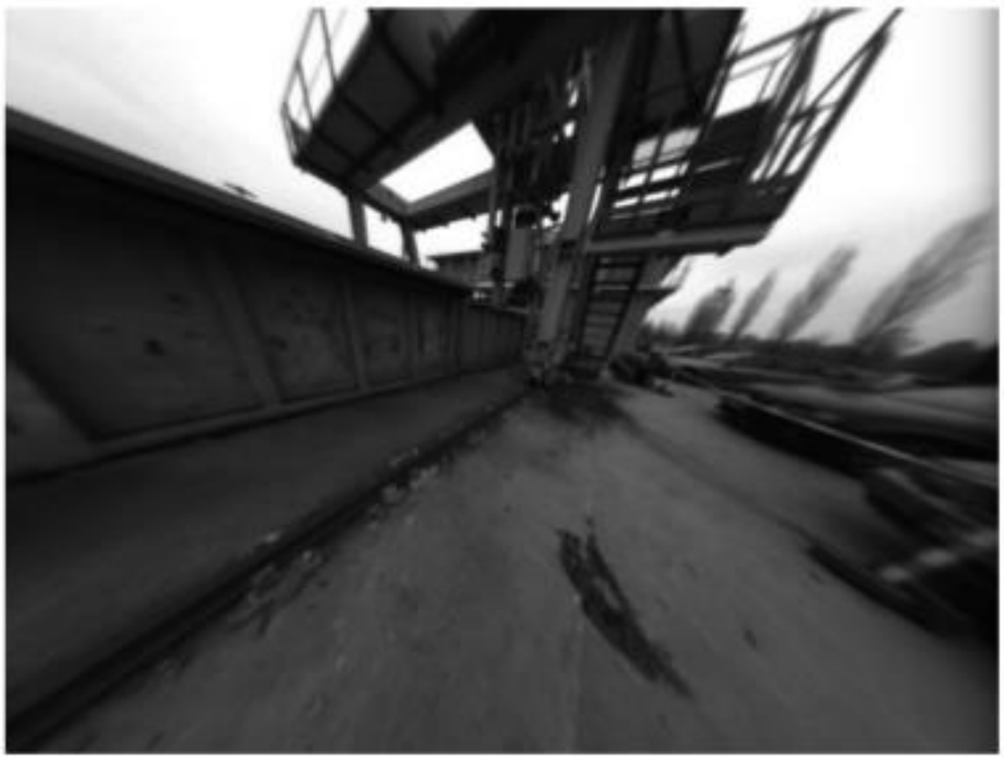
https://www.youtube.com/watch?v=isHXcv_AeFg

Engel, Schöps, Cremers; ECCV ´14

ECCV-sequence: 7 minutes, 640x480@50fps:

25.000 Tracked Frames, 450 Keyframes; 7.000 Constraints; 51 Million Points

A wider field of view is useful...
but pinhole model does not allow it.



Pinhole, 120° horiz. FoV

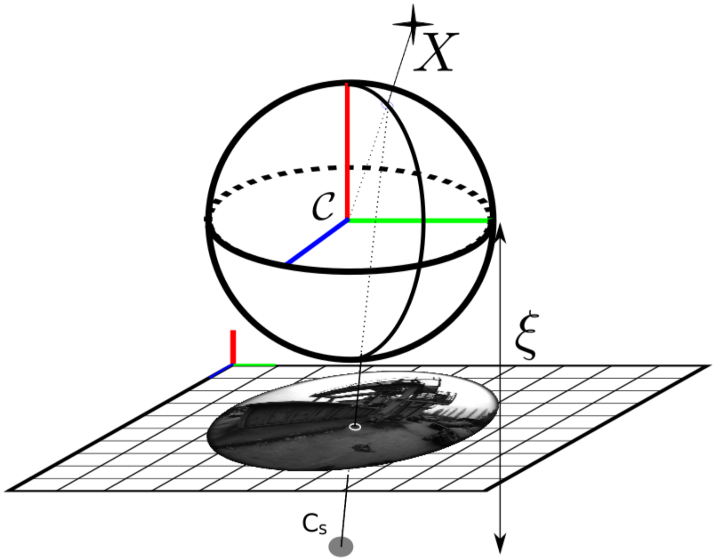
Can we use a different model?

- central system
- closed-form projection
- *closed form unprojection*
- large field of view (>180°)

Caruso, Engel, Cremers; IROS ´15

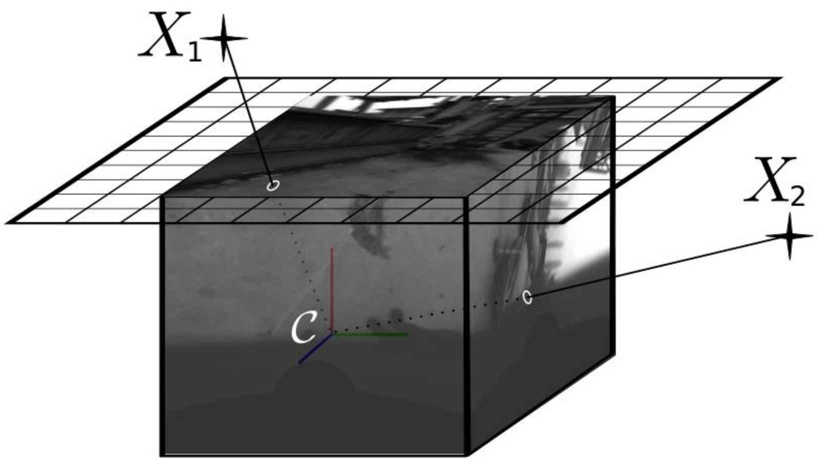
Omni LSD-SLAM

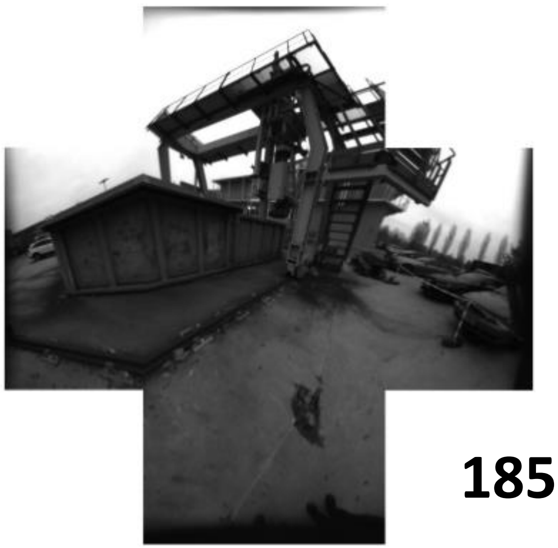
1. Unified Omnidirectional Model



185° FoV

2. Piecewise Pinhole Model





185° FoV

+ radio-tangential distortion, removed by pre-processing

Piecewise Pinhole Model:

Tracking: straight-forward
Mapping: ep. lines are straight
Ugly edge-handling, artificial model.

Unified Model:

Tracking: straight-forward (more costly!)
Mapping: ep. lines are conics -> costly.
Clean model, close to physical lenses.

1. estimate and filter **distance** instead of **depth**
2. **inverse compositional** LK instead of forward compositional

Caruso, Engel, Cremers; IROS ´15

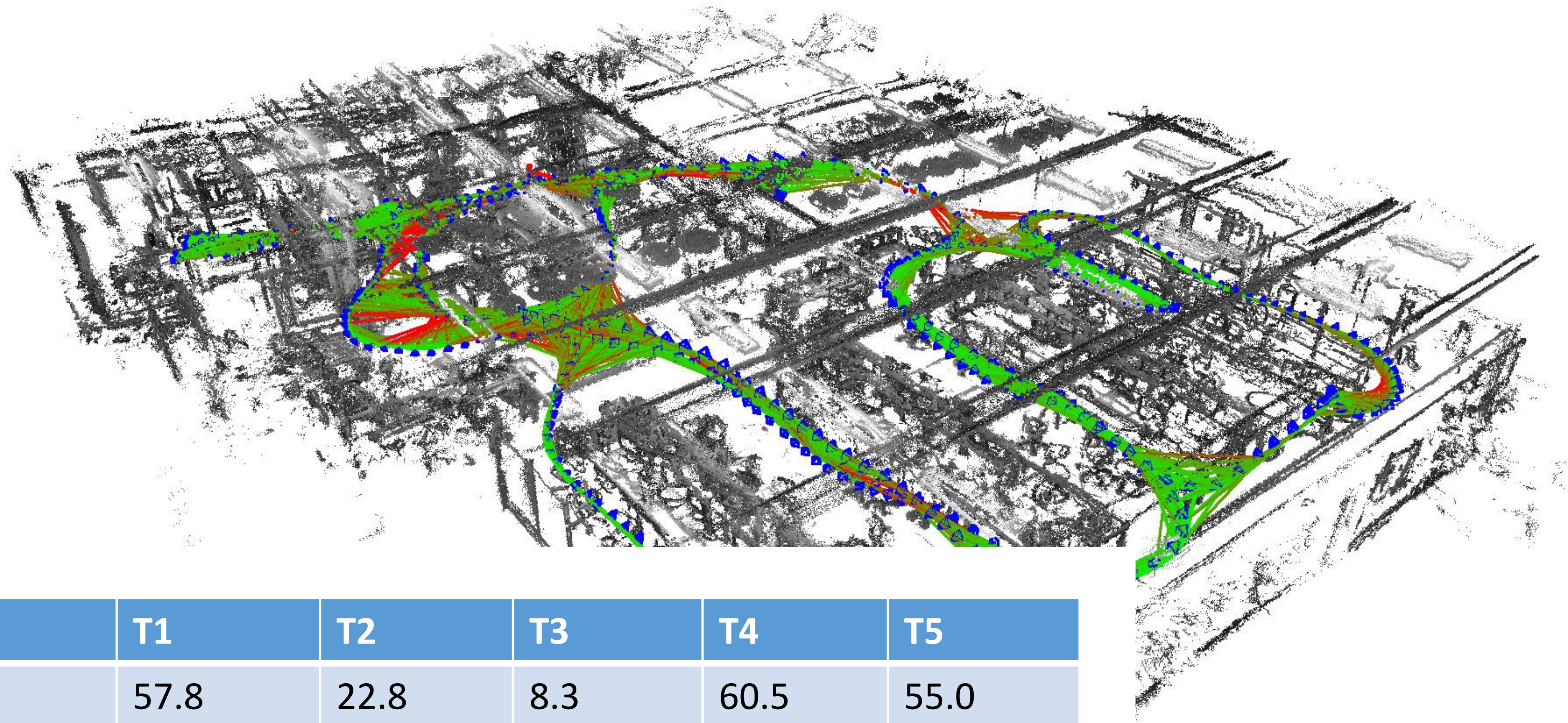
Omni LSD-SLAM

https://youtu.be/v0NqMm7Q6S8

Large-Scale Direct SLAM for Omnidirectional Cameras

Caruso, Engel, Cremers; IROS ´15

Omni LSD-SLAM



	T1	T2	T3	T4	T5
Pinhole	57.8	22.8	8.3	60.5	55.0
Multi-Pinhole	4.9	6.6	4.6	4.2	5.4
Unified	5.3	5.1	4.6	4.5	3.6

Absolute translational RMSE (cm) to MoCap groundtruth

dataset + groundtruth at

http://vision.in.tum.de/omni-lsdslam

Caruso, Engel, Cremers; IROS ´15

How about Stereo SLAM / VO?

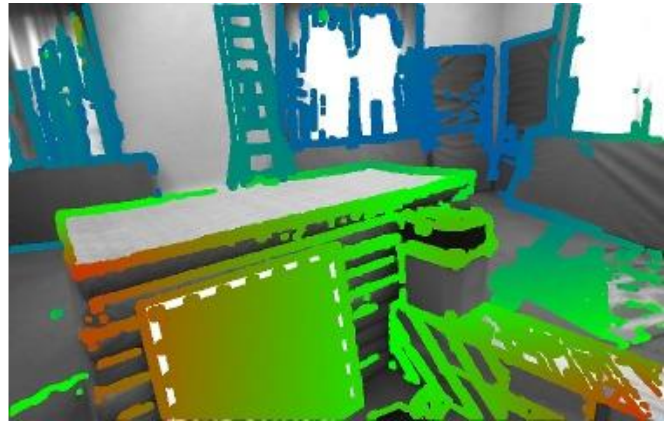
- Get absolute scale
- Initialization instantaneous
- No issues with strong rotation
- Often very practical

➢ Add stereo disparity observations
➢ Tracking still „monocular“
➢ Pose-graph again in SE(3)

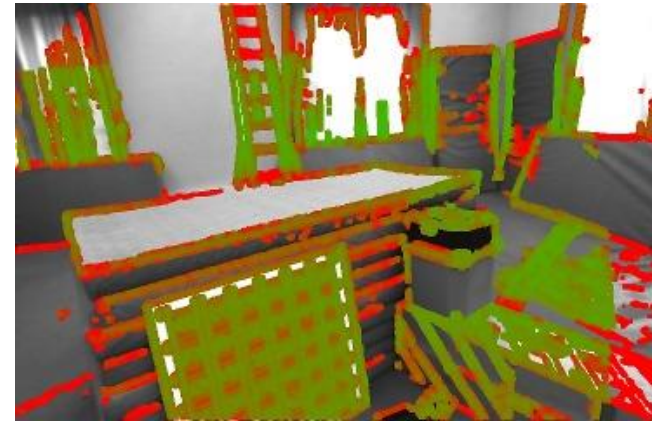
Engel, Stückler, Cremers; IROS ´15

Stereo LSD-SLAM

(left) image



(left) inverse depth



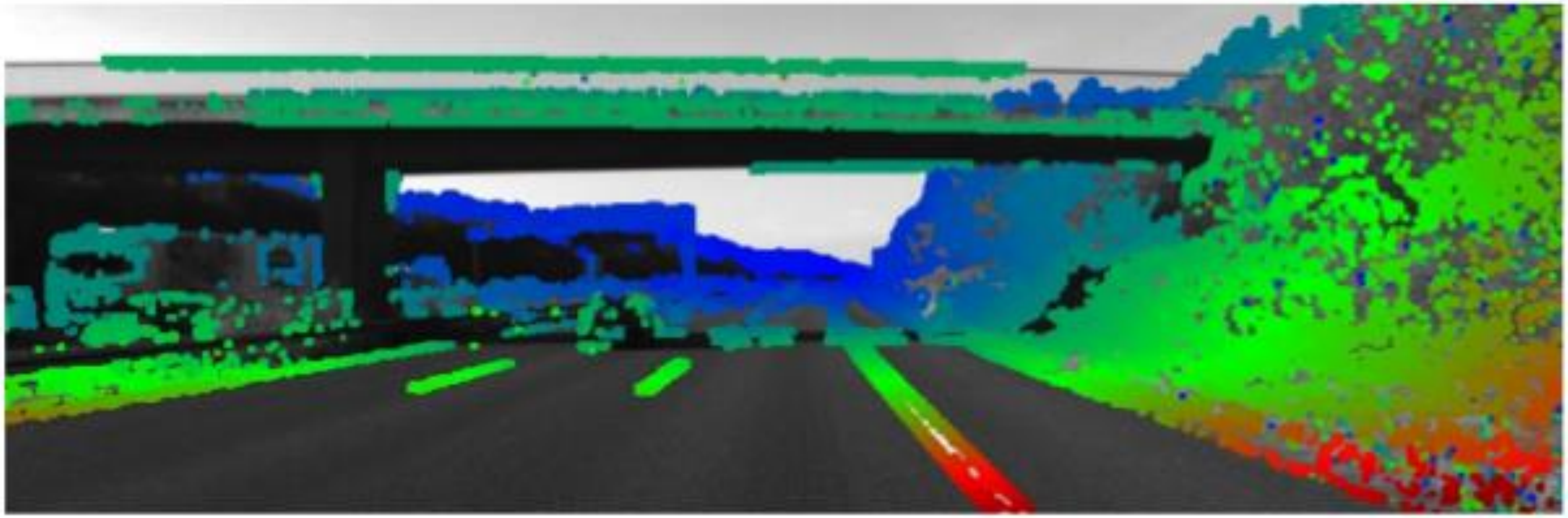
inverse depth variance

- Gaussian on inverse depth for each pixel in left image.
- Fusing stereo observations from
  - **Temporal Stereo** (left image I_l^k to next left image $I_l^{(k+n)}$)
  - **Static Stereo** (left image I_l^k to right image I_r^k)
  - **Temporal-Static Stereo** (static stero in next frame $I_l^{(k+n)}$ to $I_r^{(k+n)}$)
- Matching, Bayesian Fusion and Regularization as in LSD-SLAM.

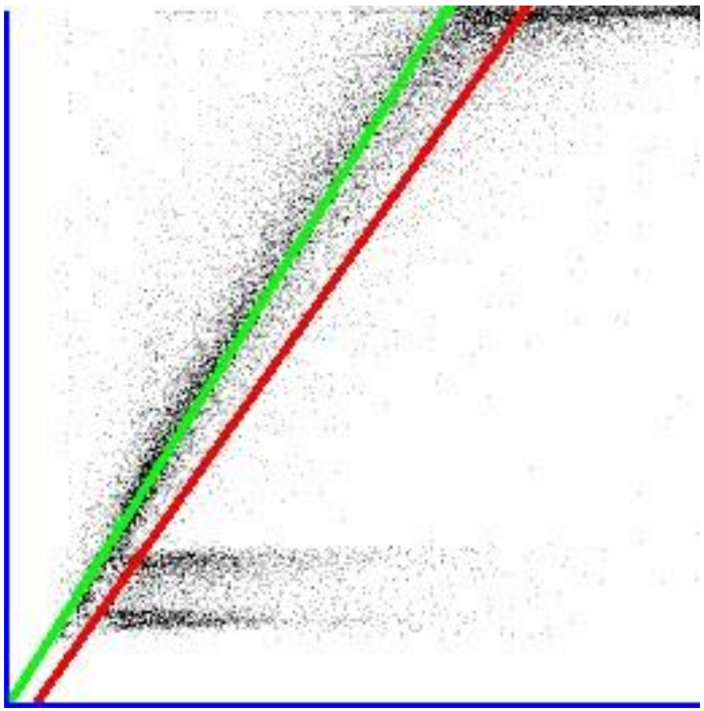
Stereo LSD-SLAM

no information from static stereo

no information from temporal stereo



Affine Lighting

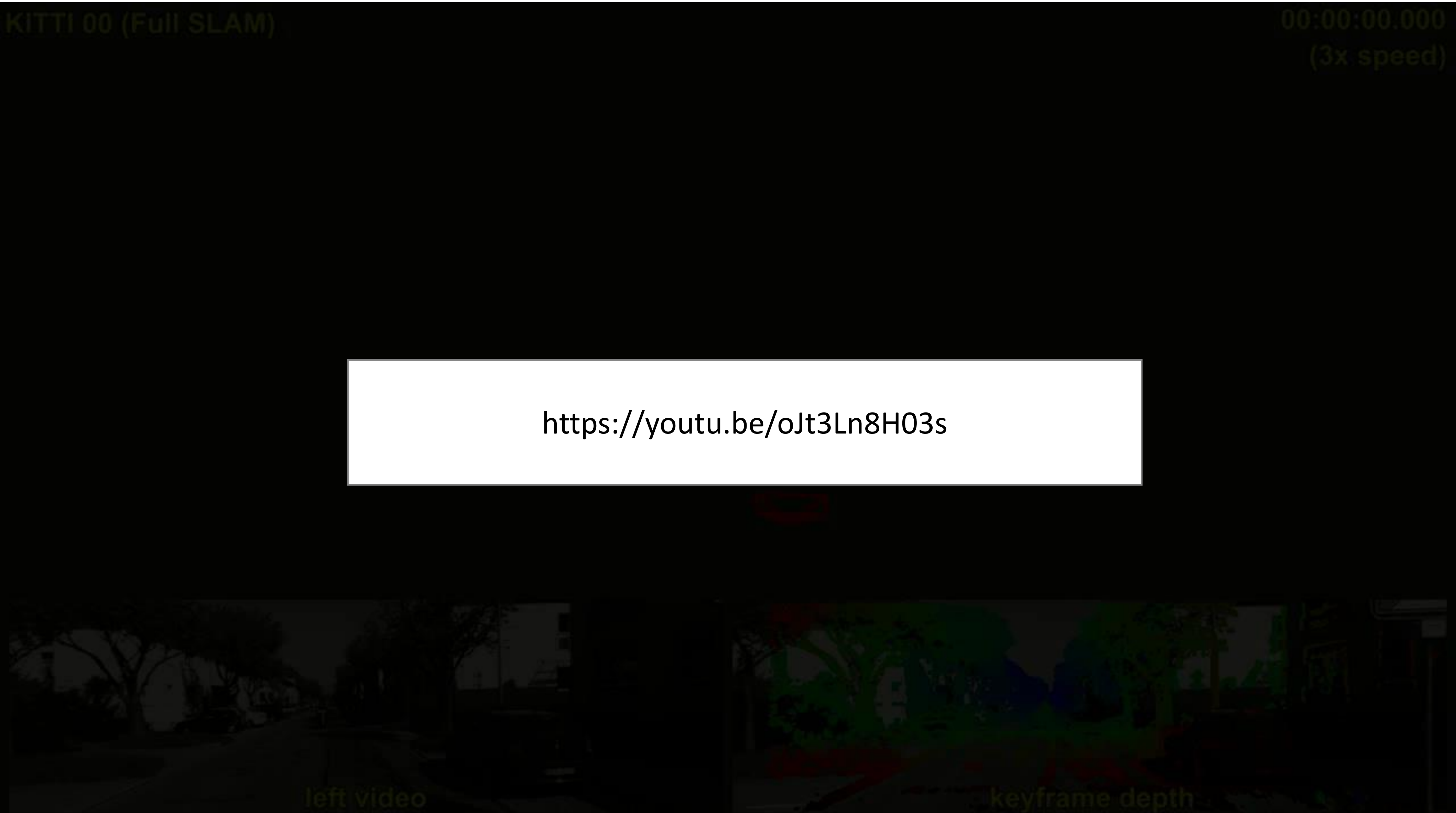


- **Approach:** Make error function invariant to affine lighting changes:

$$E(\boldsymbol{\xi}, {\color{red}a}, {\color{red}b}) = \sum_{\mathbf{x} \in \Omega_{\mathrm{KF}}} \left({\color{red}a} I^l_{\mathrm{KF}}[\mathbf{x}] + {\color{red}b} - I^l[\omega(\mathbf{x}, D_{\mathrm{KF}}(\mathbf{x}), \boldsymbol{\xi})] \right)^2$$

- **Need to be careful with outliers!**

Engel, Stückler, Cremers; IROS ´15



https://youtu.be/oJt3Ln8H03s

Large-Scale Direct SLAM with Stereo Cameras

Engel, Stückler, Cremers; IROS ´15

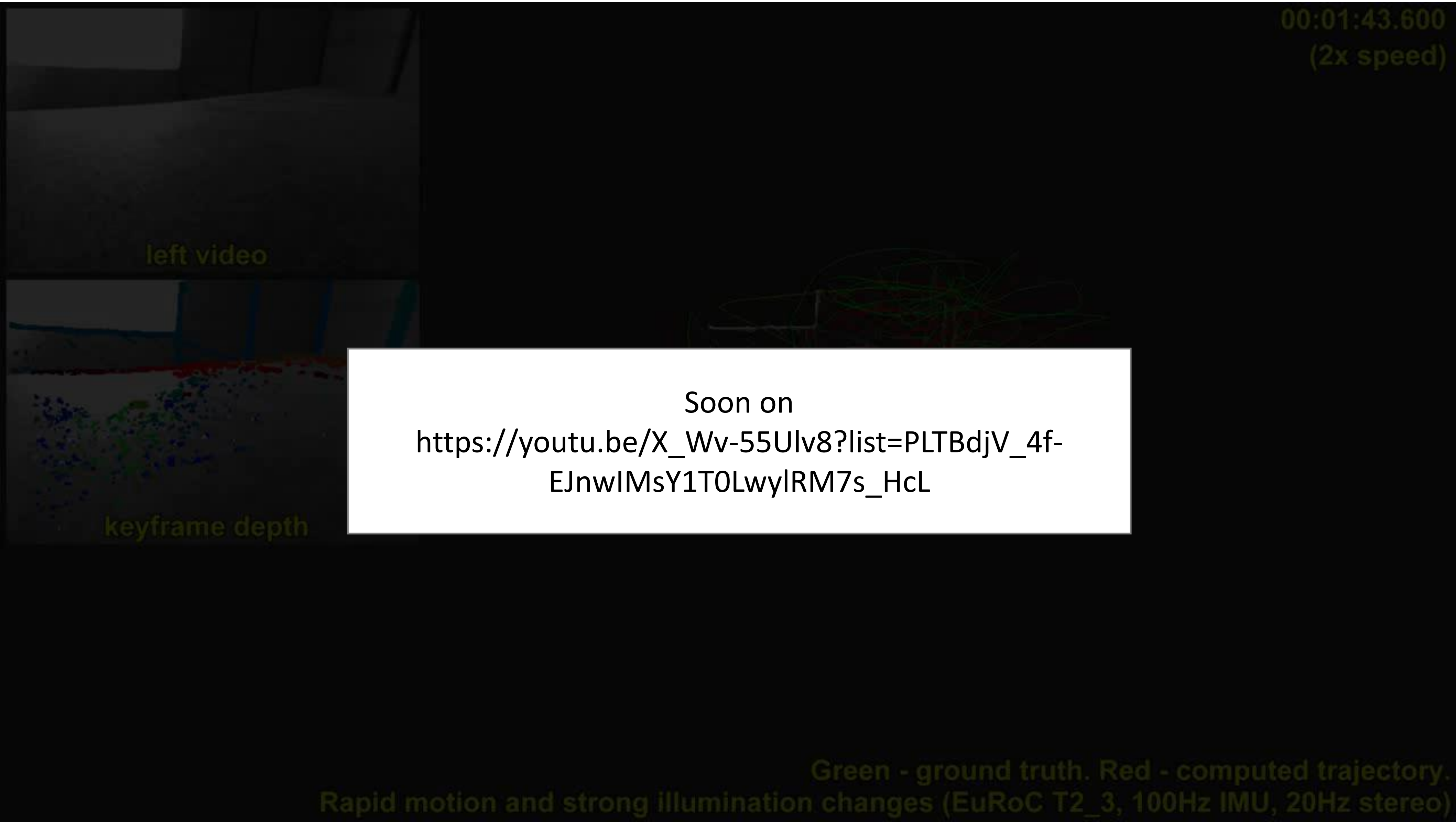
How about Visual-Inertial Integration?

Tight IMU integration:

$$E(\boldsymbol{s}) := \frac{1}{2}\sum_{i=1}^{N} E^{I}_{i\rightarrow \mathrm{ref}(i)}(\boldsymbol{\xi}_i, \boldsymbol{\xi}_{\mathrm{ref}(i)}) + \frac{1}{2}\sum_{i=2}^{N} E^{\mathrm{IMU}}(\boldsymbol{s}_{i-1}, \boldsymbol{s}_i)$$

+ windowed marginalization

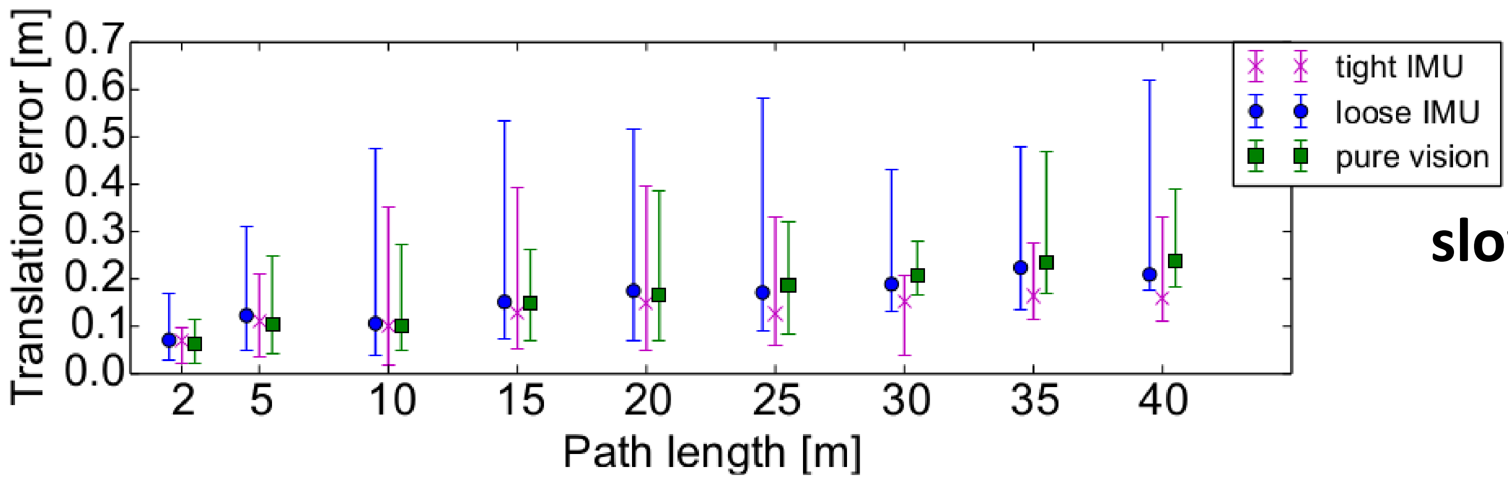
Usenko, Engel, Stückler, Cremers



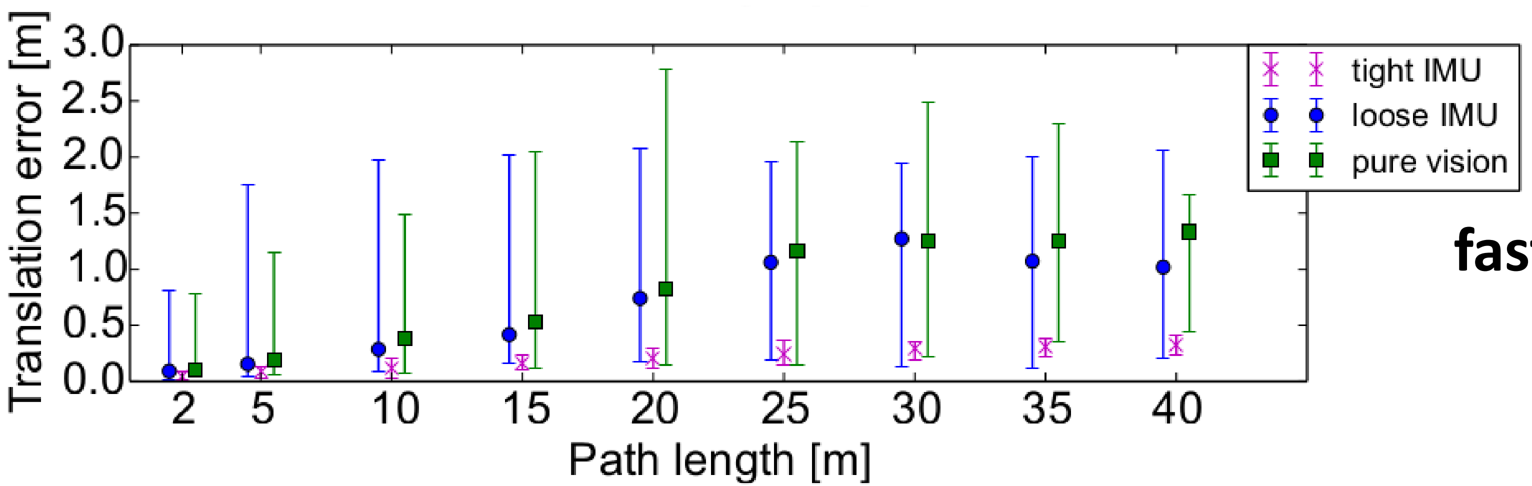
Soon on
https://youtu.be/X_Wv-55Ulv8?list=PLTBdjV_4f-EJnwIMsY1T0LwylRM7s_HcL

Stereo Camera + tight IMU integration

Usenko, Engel, Stückler, Cremers



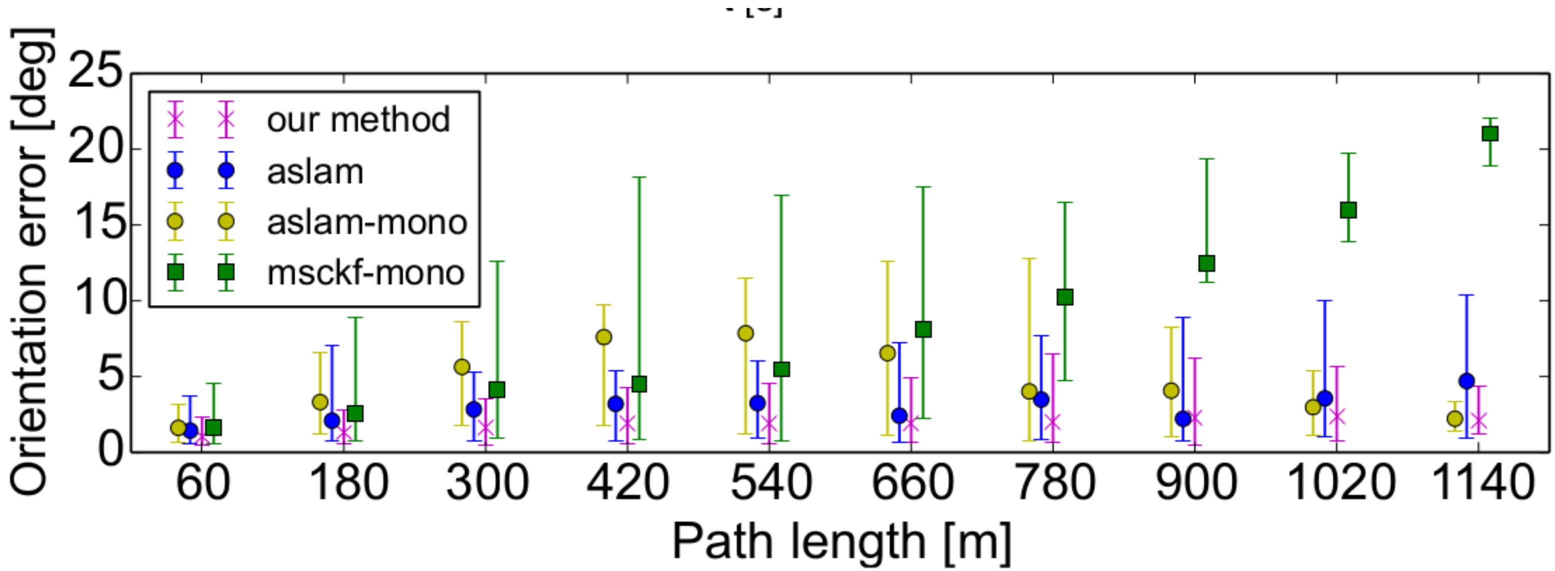
slow motion

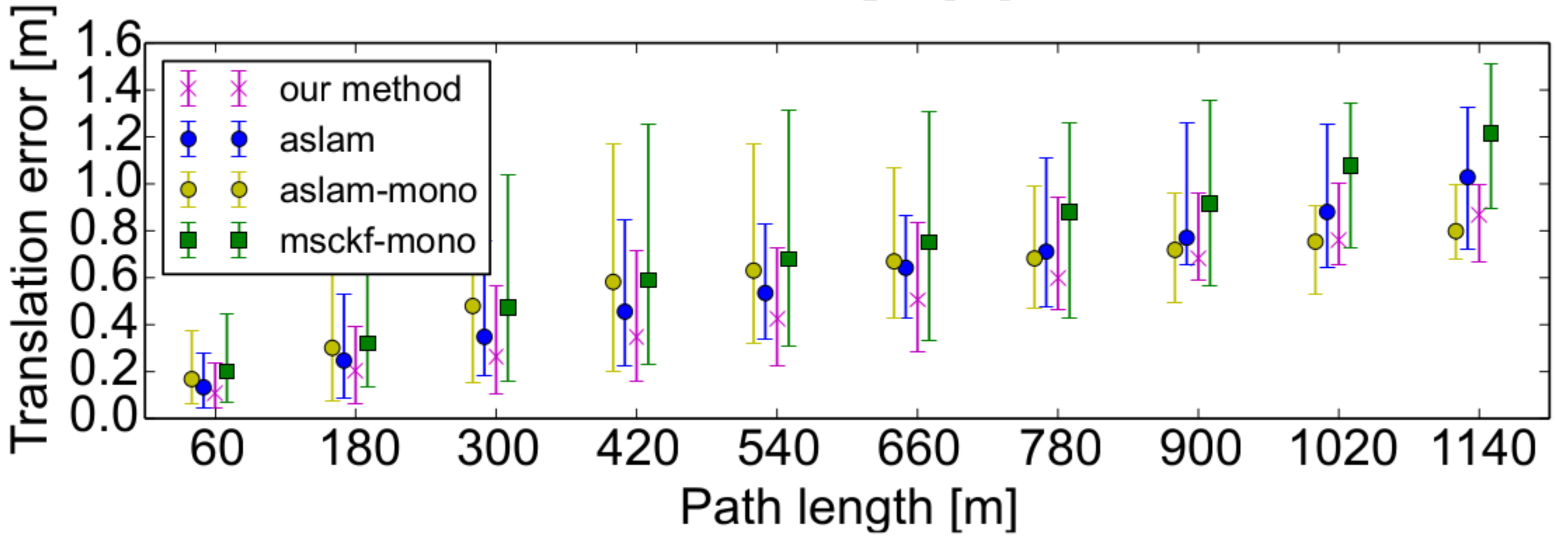


fast motion

Usenko, Engel, Stückler, Cremers

Stereo-Inertial LSD-SLAM





* Results, dataset & calibration from *„Keyframe-based visual–inertial odometry using nonlinear optimization"*, *Leutenegger, Lynen, Bosse, Siegwart, Furgale, IJRR'14*

Usenko, Engel, Stückler, Cremers

von Stumberg, Usenko, Engel, Stückler, Cremers

Comparison

Feature-Based

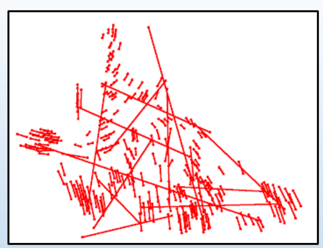
Input Images

Extract & Match Features

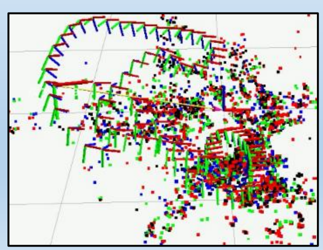
(SIFT / SURF / BRIEF /...)

abstract images to feature observations

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



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



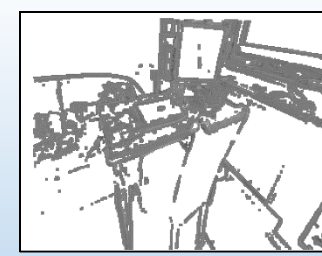
Chiuso '02, Nistér '04, Eade '06, Klein '06, Davison '07, Strasdat '10, Mur-Artal '14, ...

Direct

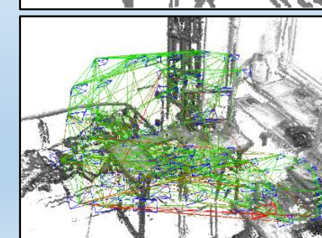
Input Images

keep full image

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



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

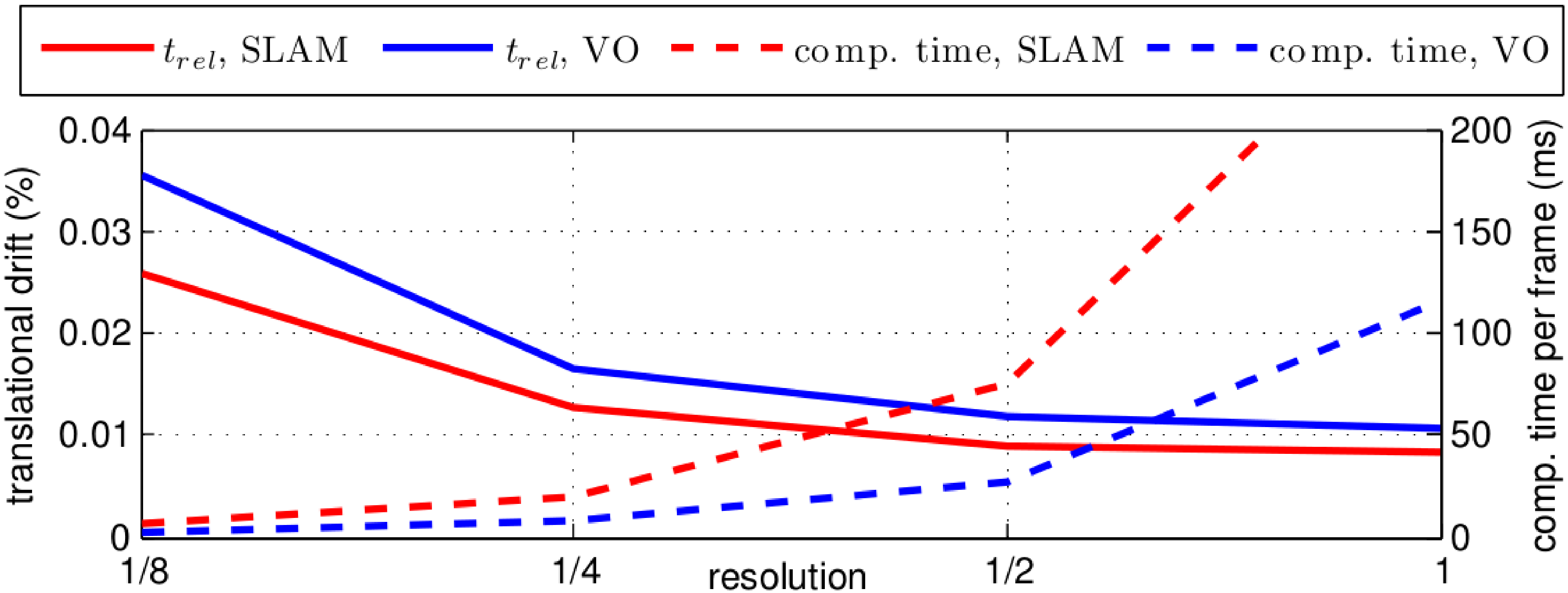


Matthies '88, Hanna '91, Comport '06, Newcombe '11, Engel '13, ...

Comparison

Feature-Based	Direct
can only use & reconstruct corners	can use & reconstruct whole image
faster	slower (but good for parallelism)
flexible: outliers can be removed retroactively.	inflexible: difficult to remove outliers retroactively.
robust to inconsistencies in the model/system (rolling shutter).	not robust to inconsistencies in the model/system (rolling shutter).
decisions (KP detection) based on less complete information.	decision (linearization point) based on more complete information.
no need for good initialization.	needs good initialization.
~20+ years of intensive research	~4 years of research (+5years 25 years ago)

Resolution vs. Accuracy



error vs. resolution on KITTI (00-10)

- Accuracy gracefully declines with resolution, while runtime greatly improves with resolution.
- *Where does the remaining error (1%) come from?*

Engel, Stückler, Cremers; IROS ´15

Questions