Loosely-Coupled Semi-Direct Monocular SLAM

I3A, University of Zaragoza, Spain

Seong Hun Lee and Javier Civera

unizar

Background

Visual Simultaneous Localization and Mapping (SLAM) can be categorized as:

(1) Direct:
- Minimize photometric errors.
- They can use semi-dense points.
- Robust to low textures.

(2) Feature-based:
- Minimize reprojection errors.
- They can track large motions and recognize previously visited places.

Contribution

We **loosely couple** direct odometry and feature-based SLAM, such that

(1) Locally, a direct method tracks the real-time camera pose wrt a short-term semi-dense map.

(2) Globally, a feature-based method builds a globally consistent sparse feature map from the keyframes.

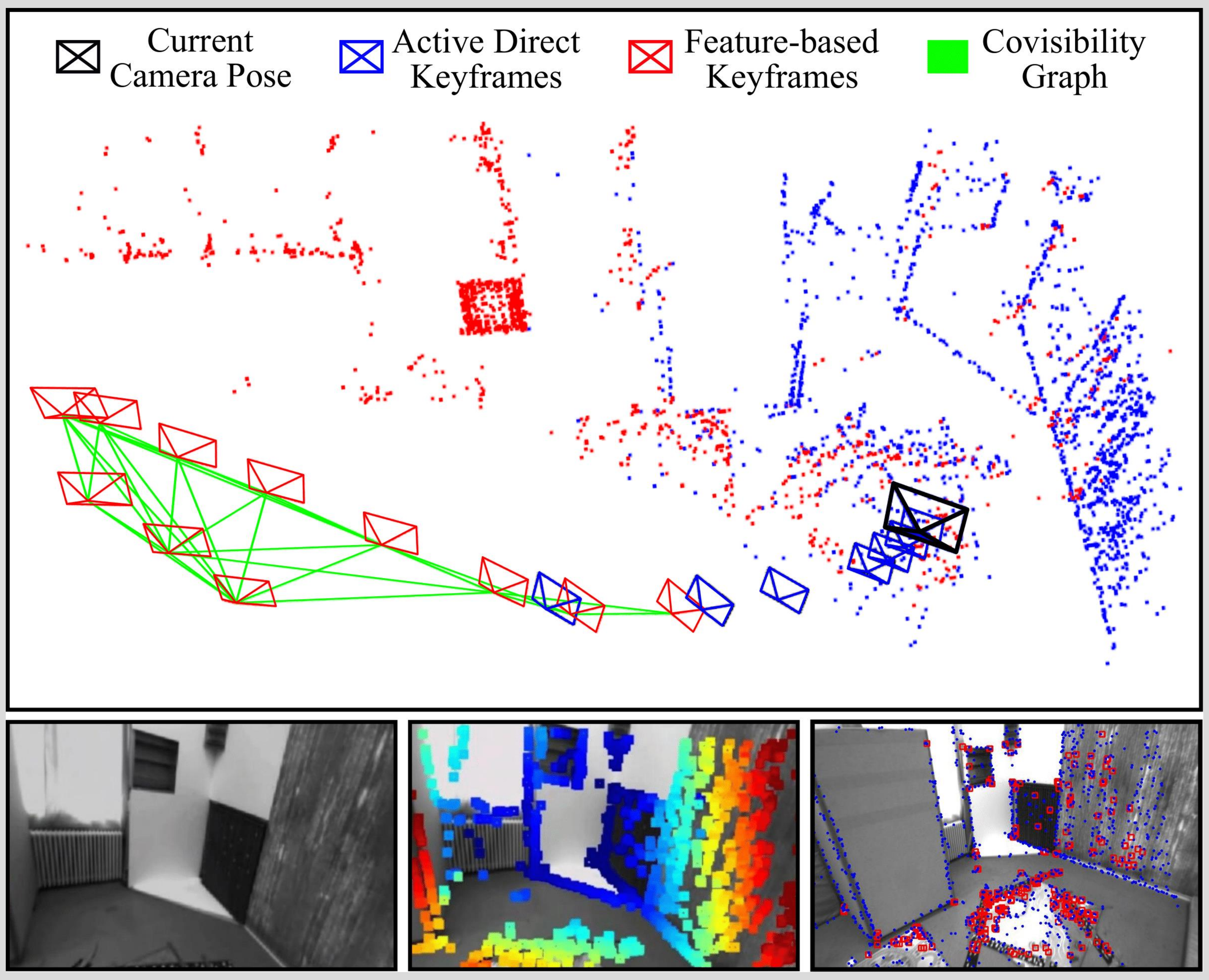
System Overview

Our system consists of two separate asynchronous modules:
- Direct module based on DSO [1],
- Feature-based module based on ORB-SLAM [2].

New Frame → **Direct module**: Direct image alignment; Photo-metric BA

↓ Marginalized Keyframes & Points

Feature-based module: 3D Keypoints generation; Geometric BA; Loop Closing



[Top] Blue points: Short-term local map for direct tracking. Red points: Long-term global map for reuse.

[Bottom] Left: Current frame, Mid: Direct keyframe with color-coded depths, Right: Feature-based keyframe.

Triple-Window Optimization

1. Photometric BA in a sliding window

$$E^{\mathbf{p}}_{ij} := \sum_{\tilde{\mathbf{p}}\in\mathcal{N}_{\mathbf{p}}} \omega_{\tilde{\mathbf{p}}} \left\| I_j[\tilde{\mathbf{p}}'] - b_j - \frac{t_j e^{a_j}}{t_i e^{a_i}}\left(I_i[\tilde{\mathbf{p}}] - b_i\right)\right\|_\gamma$$

$$\text{with} \quad \omega_{\tilde{\mathbf{p}}} := \frac{c^2}{c^2 + \|\nabla I_i(\tilde{\mathbf{p}})\|_2^2},$$

$$\tilde{\mathbf{p}}' = \mathbf{\Pi_c}\left(\mathbf{T}_{jw}^{-1}\mathbf{T}_{iw}\mathbf{\Pi_c}^{-1}(\tilde{\mathbf{p}}, d_{\mathbf{p}})\right)$$

$$E_{\text{photo}} := \sum_{i\in\mathcal{F}}\sum_{\mathbf{p}\in\mathcal{P}_i}\sum_{j\in\text{obs}(\mathbf{p})} E^{\mathbf{p}}_{ij} + \sum_{i\in\mathcal{F}}\left(\lambda_a a_i^2 + \lambda_b b_i^2\right)$$

- $E^{\mathbf{p}}_{ij}$: Photometric error of point **p** (with inverse depth $d_{\mathbf{p}}$) observed in keyframe *i* and *j*.
- t, a, b : Exposure time and brightness parameters.
- λ's are set to some constant when t is known. Otherwise, $\lambda_a = \lambda_b = 0$ and $t_i = t_j = 1$.

2. Geometric BA based on the covisibility

$$E_{\text{reproj}} = \sum_{i\in\mathcal{F}_{\text{local}}}\sum_{\mathbf{x}\in\mathcal{P}_i}\sum_{j\in\text{obs}(\mathbf{x})}\left\|\frac{\mathbf{p}_{j,\mathbf{x}} - \mathbf{\Pi_c}(\mathbf{T}_{iw}\mathbf{x}_w)}{\sigma_{\mathbf{x}}^2}\right\|_\gamma$$

$$\text{with} \quad \sigma_{\mathbf{x}}^2 := (\lambda_{\text{pyr}})^{2L_{\text{pyr},\mathbf{x}}},$$

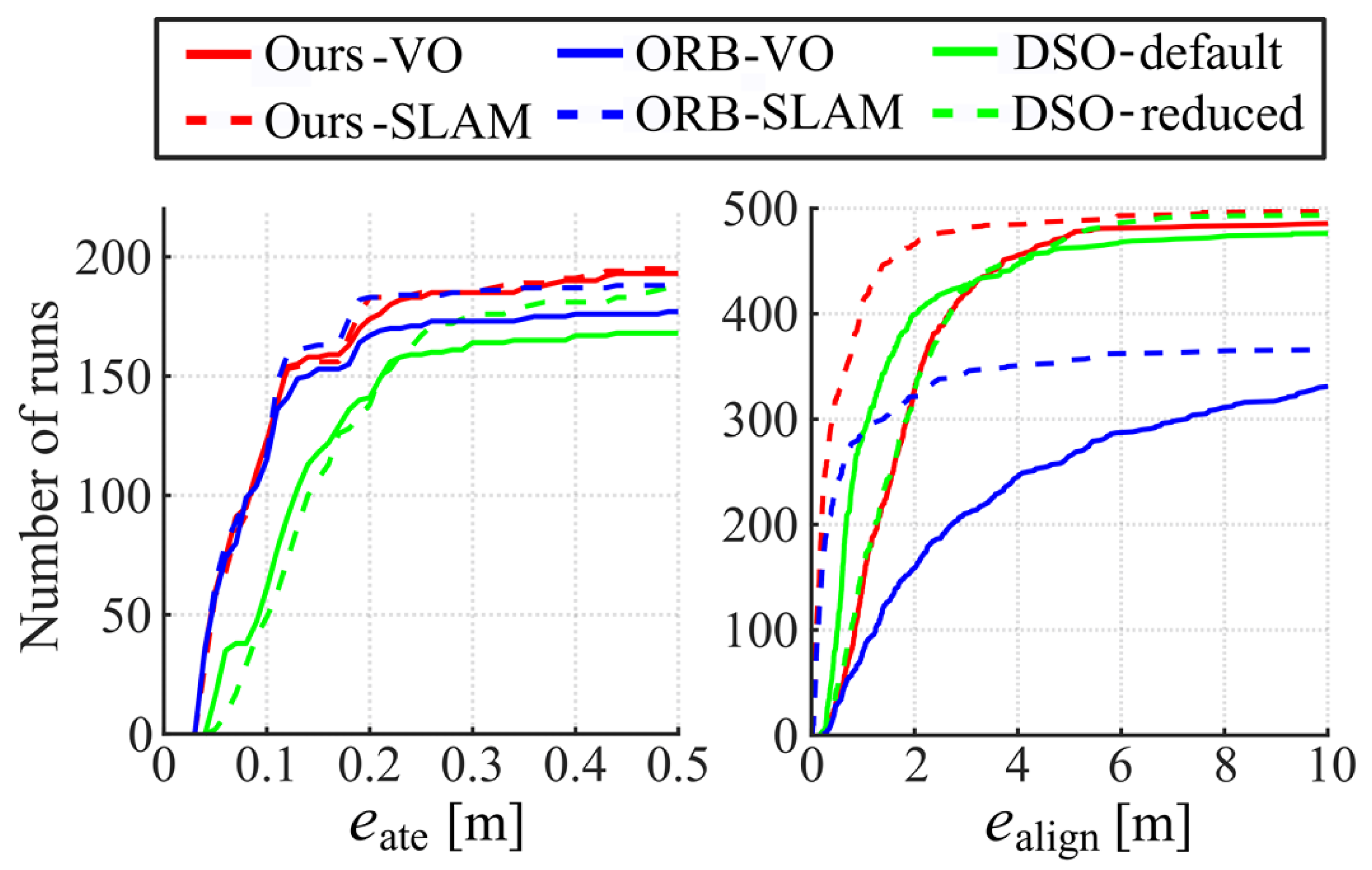
- $\mathbf{p}_{j,\mathbf{x}}$: The match in frame *j*.
- λ_{pyr} , $L_{\text{pyr},\mathbf{x}}$: Scale factor and level of the image pyramid at which **x** was detected.

3. Pose graph optimization

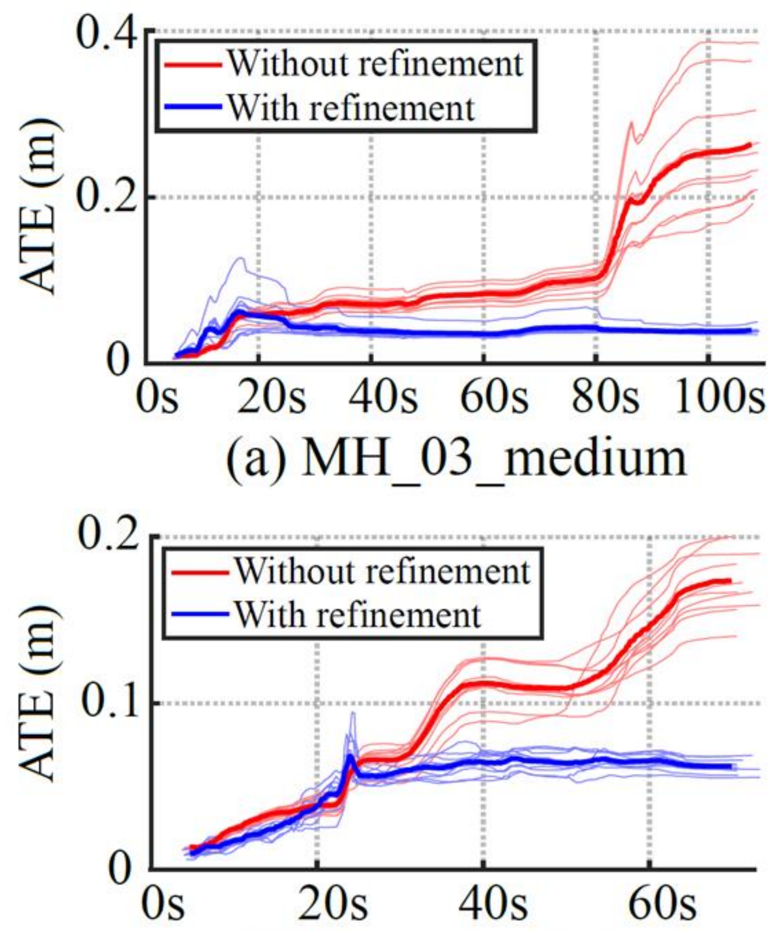
$$E_{\text{graph}} = \sum_{(i,j)\in\mathcal{E}_{\text{edge}}}\left\|\log_{\text{Sim}(3)}\left(\mathbf{S}_{ij,0}\,\mathbf{S}_{jw}\,\mathbf{S}_{iw}^{-1}\right)\right\|_2^2$$

- $\mathcal{E}_{\text{edge}}$: Edges in the essential graph [2].
- $\mathbf{S}_{ij,0} = \mathbf{S}_{iw,0}\mathbf{S}_{jw,0}^{-1}$: Fixed similarity transformation (with the scale 1) prior to the optimization.

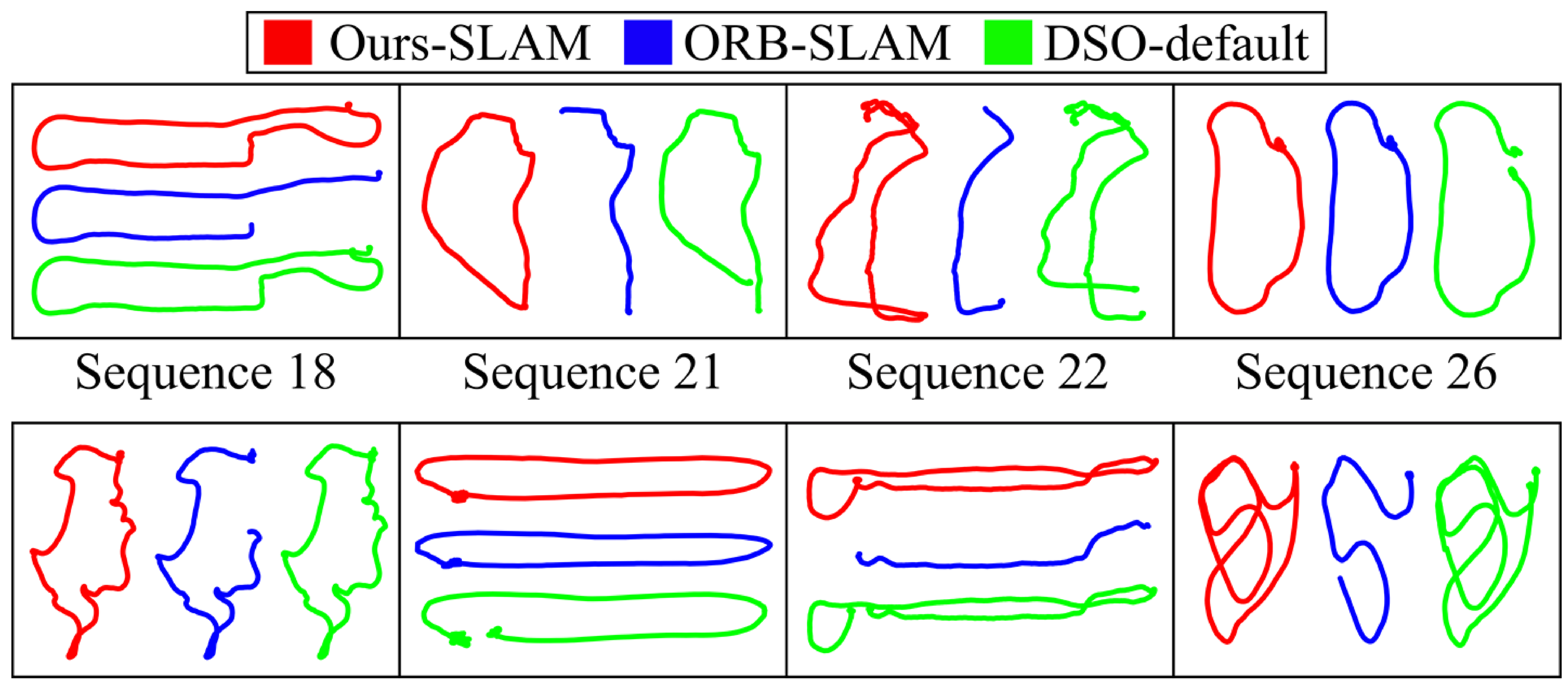
Evaluation Results



(a) EuRoC MAV (b) TUM monoVO



(a) MH_03_medium (b) MH_04_difficult



Sequence 18, Sequence 21, Sequence 22, Sequence 26, Sequence 31, Sequence 36, Sequence 40, Sequence 41

References:

[1] J. Engel, V. Koltun, D. Cremers, "Direct Sparse Odometry", TPAMI, 2018

[2] R. Mur-Artal, J.M.M. Montiel, J.D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system", TRO, 2015

Video Demo