Multi-Session SLAM with Differentiable Wide-Baseline Pose Optimization

Supplementary Material

A. Gradient Computation for the SED Solver

$$d \coloneqq l_x^2 + l_y^2 \tag{14}$$

$$\zeta \coloneqq l_x m_x + l_y m_y + l_z \tag{15}$$

$$\frac{\partial err(m,l)}{\partial l} = \begin{bmatrix} \left(\frac{-2l_x^2\zeta}{d^2} + \frac{l_x m_x}{d} + \frac{\zeta}{d}\right) & \left(\frac{-2l_x l_y \zeta}{d^2} + \frac{l_x m_y}{d}\right) & \left(\frac{l_x}{d}\right) \\ \left(\frac{-2l_y l_x \zeta}{d^2} + \frac{l_y m_x}{d}\right) & \left(\frac{-2l_y^2\zeta}{d^2} + \frac{l_y m_y}{d} + \frac{\zeta}{d}\right) & \left(\frac{l_y}{d}\right) \end{bmatrix} \tag{16}$$

Let $\bar{a}$ be the anchor location in homogenous coordinates. The partials for l w.r.t. $\mathbf{F}$ are

$$\frac{\partial l}{\partial \mathbf{F}} = \frac{\partial \mathbf{F}\bar{a}}{\partial \mathbf{F}} = \bar{a}^\top \otimes \mathbb{I} \in \mathbb{R}^{3\times(3\times3)} \tag{17}$$

The partials for the fundamental matrix w.r.t. the essential matrix, given the known calibration matrix K:

$$\frac{\partial \mathbf{F}}{\partial \mathbf{E}} = \frac{\partial\left[(K^{-1})^\top \mathbf{E} K^{-1}\right]}{\partial \mathbf{E}} = (K^{-1})^\top \otimes K^{-1} \in \mathbb{R}^{(3\times3)\times(3\times3)} \tag{18}$$

The 3×3 essential matrix in terms of the input rotation $\mathbf{R}$ and translation $\mathbf{t}$, and the local updates $\xi_\mathbf{R}$ and $\xi_\mathbf{t}$, is

$$\mathbf{E} = (\mathbf{e}^{\xi_\mathbf{R}}\mathbf{R})^\top [\mathbf{e}^{\xi_\mathbf{t}}\mathbf{t}]_\times \tag{19}$$

The derivatives of $\xi_\mathbf{R}$ and $\xi_\mathbf{t}$ are taken at the identity, so they are equal to 0 when treated as a constant. To compute the partial of this essential matrix w.r.t. $\xi_\mathbf{R}$ at 0:

$$\mathbf{tx} \coloneqq [\mathbf{t}]_\times \in \mathbb{R}^{3\times3} \tag{20}$$

$$\frac{\partial \mathbf{E}}{\partial \xi_\mathbf{R}} = \frac{\partial (e^{\xi_\mathbf{R}}\mathbf{R})^\top [\mathbf{t}]_\times}{\xi_\mathbf{R}} = \begin{bmatrix} \mathbf{R}^\top [\mathbf{t}\mathbf{x}_{c1}]_\times \\ \mathbf{R}^\top [\mathbf{t}\mathbf{x}_{c2}]_\times \\ \mathbf{R}^\top [\mathbf{t}\mathbf{x}_{c3}]_\times \end{bmatrix}^\top \in \mathbb{R}^{(3\times3)\times3} \tag{21}$$

Similarly, the partial w.r.t. $\xi_\mathbf{t}$ at 0:

$$\frac{\partial \mathbf{E}}{\partial \xi_\mathbf{t}} = \frac{\partial \mathbf{R}^\top [\mathbf{e}^{\xi_\mathbf{t}}\mathbf{t}]_\times}{\partial \xi_\mathbf{t}} = \frac{\partial \mathbf{R}^\top [\mathbf{e}^{\xi_\mathbf{t}}\mathbf{t}]_\times}{\partial (e^{\xi_\mathbf{t}}\mathbf{t})} \frac{\partial (e^{\xi_\mathbf{t}}\mathbf{t})}{\partial \xi_\mathbf{t}} = \mathbf{R}^\top \frac{\partial [\vec{n}]_\times}{\partial \vec{n}} (-\mathbf{tx}) \in \mathbb{R}^{(3\times3)\times3} \tag{22}$$

Putting it together with the chain rule:

$$\frac{\partial err(m,l)}{\partial \xi_\mathbf{R}} = \frac{\partial err(m,l)}{\partial l} \frac{\partial l}{\partial \mathbf{F}} \frac{\partial \mathbf{F}}{\partial \mathbf{E}} \frac{\partial \mathbf{E}(e^{\xi_\mathbf{R}}\mathbf{R}, e^{\xi_\mathbf{t}}\mathbf{t})}{\partial \xi_\mathbf{R}} \in \mathbb{R}^{2\times3} \tag{23}$$

$$\frac{\partial err(m,l)}{\partial \xi_\mathbf{t}} = \frac{\partial err(m,l)}{\partial l} \frac{\partial l}{\partial \mathbf{F}} \frac{\partial \mathbf{F}}{\partial \mathbf{E}} \frac{\partial \mathbf{E}(e^{\xi_\mathbf{R}}\mathbf{R}, e^{\xi_\mathbf{t}}\mathbf{t})}{\partial \xi_\mathbf{t}} \in \mathbb{R}^{2\times3} \tag{24}$$

$$\tag{25}$$

B. Extracting Rotation and Translation from the Essential Matrix

To obtain the relative pose given the essential matrix $\mathbf{E}$, we follow the procedure prescribed in [13].

$$W := \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (26) \qquad Z := \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (27) \qquad U, \Sigma, V^\top = \text{SVD}(\mathbf{E}) \quad (28)$$

$$t = UZU^\top \quad (29) \qquad R1 = UWV^\top \quad (30) \qquad R2 = UW^\top V^\top \quad (31)$$

The four plausible solutions are

$$[(t, R1), (t, R2), (-t, R1), (-t, R2)] \quad (32)$$

During training, we choose the solution closest to the ground-truth. During inference, we choose the pose which triangulates the most points in front of the camera.

C. Architecture Details

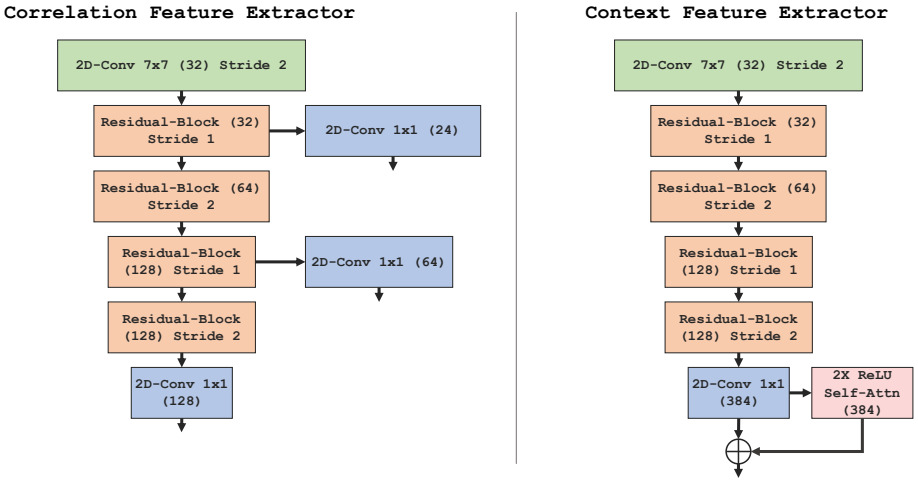

Figure 11. The architecture of our context and correlation feature extractors. Both feature extractors are residual networks. The context feature extractor also uses ReLU self-attention to propagate information across the image. The correlation features are used to evaluate visual similarity at multiple spatial resolutions. The numbers in parenthesis are the output feature dimensions.

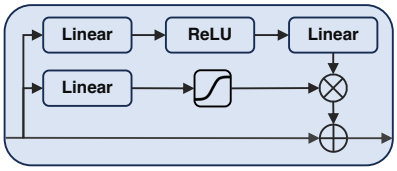

Figure 12. The gated residual unit.

MLP-GRU We depict the gated residual unit in Fig. 12. This is the same design from DPVO [45].

Feature Extractors We visualize the feature extractors in Fig. 11. The correlation features are produced at $1/2$, $1/4$ and $1/8$ the image resolution. The context features are extracted only at $1/8$ resolution, and have additional self-attention layers to propagate information over the image.

D. Additional Qualitative Results

Figure 13. Additional Qualitative results on Scannet.