DSAC – Differentiable RANSAC for Camera Localization
– Supplementary Materials –

Eric Brachmann[1], Alexander Krull[1], Sebastian Nowozin[2]
Jamie Shotton[2], Frank Michel[1], Stefan Gumhold[1], Carsten Rother[1]
[1] TU Dresden, [2] Microsoft

This document contains additional information on the derivative of the task loss function (resp. the expectation thereof) for the SoftAM and DSAC learning strategies. Furthermore, we illustrate some difficulties of camera localization on the 7-Scenes dataset to motivate the usage of a RANSAC schema for this problem. Finally, we discuss the running time of our pipeline, and potential benefits of predicting multi-modal scene coordinate distributions as future work.

1. Derivatives

1.1. Soft argmax Selection (SoftAM)

To learn our camera localization pipeline in an end-to-end fashion, we have to calculate the derivatives of the task loss function $\ell(\mathbf{R}(\mathbf{h}^{\mathbf{w},\mathbf{v}}_{\mathrm{SoftAM}}, Y^{\mathbf{w}}), \mathbf{h}^*)$ w.r.t. to learnable parameters. In the following, we show the derivative w.r.t. parameters $\mathbf{w}$, but derivation w.r.t. parameters $\mathbf{v}$ works similarly.

Applying the chain rule and calculating the total derivative of $\mathbf{R}$, we get:

$$\frac{\partial}{\partial \mathbf{w}} \ell(\mathbf{R}(\mathbf{h}^{\mathbf{w},\mathbf{v}}_{\mathrm{SoftAM}}, Y^{\mathbf{w}}), \mathbf{h}^*) = \frac{\partial \ell}{\partial \mathbf{R}} \left(\frac{\partial \mathbf{R}}{\partial \mathbf{h}^{\mathbf{w},\mathbf{v}}_{\mathrm{SoftAM}}} \frac{\partial \mathbf{h}^{\mathbf{w},\mathbf{v}}_{\mathrm{SoftAM}}}{\partial \mathbf{w}} + \frac{\partial \mathbf{R}}{\partial Y^{\mathbf{w}}} \frac{\partial Y^{\mathbf{w}}}{\partial \mathbf{w}} \right) \tag{1}$$

Since $\mathbf{h}^{\mathbf{w},\mathbf{v}}_{\mathrm{SoftAM}}$ is a weighted average of hypothesis (see Eq. 4 of the main paper) we can differentiate it as follows:

$$\frac{\partial}{\partial \mathbf{w}} \mathbf{h}^{\mathbf{w},\mathbf{v}}_{\mathrm{SoftAM}} = \sum_J \left(\left(\frac{\partial}{\partial \mathbf{w}} P(J|\mathbf{v},\mathbf{w}) \right) \mathbf{h}^{\mathbf{w}}_J + P(J|\mathbf{v},\mathbf{w}) \frac{\partial}{\partial \mathbf{w}} \mathbf{h}^{\mathbf{w}}_J \right) \tag{2}$$

Weights $P(J|\mathbf{v},\mathbf{w})$ follow a softmax distribution of hypothesis scores (see Eq. 5 of the main paper). Hence, we can differentiate as follows:

$$\frac{\partial}{\partial \mathbf{w}} P(J|\mathbf{v},\mathbf{w}) = P(J|\mathbf{v},\mathbf{w}) \left(\frac{\partial}{\partial \mathbf{w}} s(\mathbf{h}^{\mathbf{w}}_J, \mathbf{v}) - \mathbb{E}_{J' \sim P(J'|\mathbf{v},\mathbf{w})} \left[\frac{\partial}{\partial \mathbf{w}} s(\mathbf{h}^{\mathbf{w}}_{J'}, \mathbf{v}) \right] \right) \tag{3}$$

1.2. Probabilistic Selection (DSAC)

Using the DSAC strategy, we learn our camera localization pipeline by minimizing the expectation of the task loss function:

$$\frac{\partial}{\partial \mathbf{w}} \mathbb{E}_{J \sim P(J|\mathbf{v},\mathbf{w})} \left[\ell(\cdot) \right] = \mathbb{E}_{J \sim P(J|\mathbf{v},\mathbf{w})} \left[\ell(\cdot) \frac{\partial}{\partial \mathbf{w}} \log P(J|\mathbf{v},\mathbf{w}) + \frac{\partial}{\partial \mathbf{w}} \ell(\cdot) \right], \tag{4}$$

where we use $\ell(\cdot)$ as a stand-in for $\ell(\mathbf{R}(\mathbf{h}^{\mathbf{w},\mathbf{v}}_J, Y^{\mathbf{w}}), \mathbf{h}^*)$. We differentiate $\frac{\partial}{\partial \mathbf{w}} \ell(\cdot)$ following Eq. 1 of this document, and log probabilities $\log P(J|\mathbf{v},\mathbf{w})$ as:

$$\frac{\partial}{\partial \mathbf{w}} \log P(J|\mathbf{v},\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}} s(\mathbf{h}^{\mathbf{w}}_J, \mathbf{v}) - \mathbb{E}_{J' \sim p(J'|\mathbf{v},\mathbf{w})} \left[\frac{\partial}{\partial \mathbf{w}} s(\mathbf{h}^{\mathbf{w}}_{J'}, \mathbf{v}) \right]. \tag{5}$$

2. Further Discussions

Difficulty of the 7-Scenes Dataset. Please see Fig. 1 for examples of difficult situations in the 7-Scenes dataset. In our experiments, inlier ratios of scene coordinate predictions range from 5% to 85%. See Fig. 2 (left) for the inlier ratio distribution over the complete 7-Scenes dataset. In accordance to [2, 1], we consider a scene coordinate prediction an inlier if it is within 10cm of the ground truth scene coordinate. In Fig. 2 (right) we plot the performance of DSAC against the ratio of inliers. For comparison we plot the performance of a naive approach without RANSAC (pose fit to all scene coordinate predictions).

Test Time. The scene coordinate prediction takes ∼0.5s on a Tesla K80 GPU. Pose optimization takes ∼1s. The runtime of argmax hypothesis selection (RANSAC) or probabilistic selection (DSAC) is identical and negligible.

Multi-Modality. Compared to Brachmann *et al*. [1], our pipeline performs not as well on the *Stairs* scene (see Table 1 of the main paper). We account this to the fact that the Coordinate CNN predicts only uni-modal point estimates, whereas the random forest of [1] predicts multi-modal scene coordinate distributions. The *Stairs* scene contains many repeating structures, so we expect multi-modal predictions to help. We also expect bad performance of the SoftAM strategy in case pose hypothesis distributions are multi-modal, because an average is likely to be a bad representation of either mode. In contrast, DSAC can probabilistically select the correct mode. We conclude that multi-modality in scene coordinate predictions and pose hypothesis distributions is a promising direction for future work.

References

[1] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother. Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image. In *CVPR*, 2016.

[2] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *CVPR*, 2013.

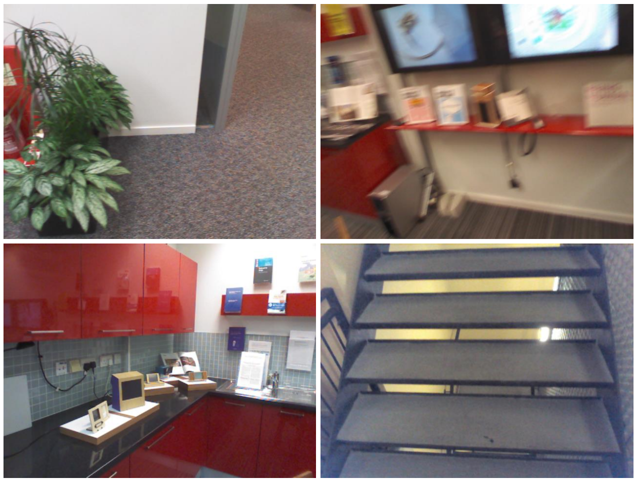

Figure 1. Difficult frames within the 7-Scenes dataset: Textureless surfaces **(upper left)**, motion blur **(upper right)**, reflections **(lower left)**, and repeating structures **(lower right)**. DSAC estimates the correct pose in all 4 cases.

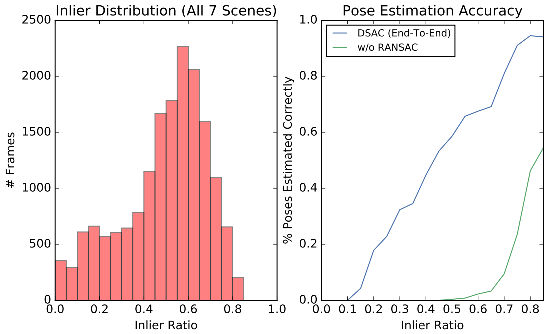

Figure 2. Distribution of inlier ratios of our scene coordinate predictions **(left)**, and corresponding pose estimation accuracy of DSAC compared to a naive approach without RANSAC **(right)**.