Learning Less is More - 6D Camera Localization via 3D Surface Regression

Eric Brachmann and Carsten Rother
Heidelberg University (HCI/IWR)

Visual Learning Lab Heidelberg

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

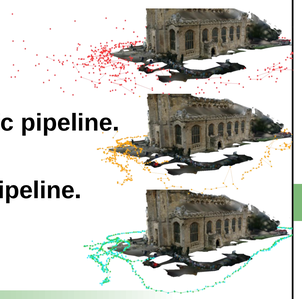
Code and trained models:

Problem Statement

Estimate the **6D camera pose** (position + orientation) relative to a known scene **from a single RGB image**.

We show that learning less is more. See right:

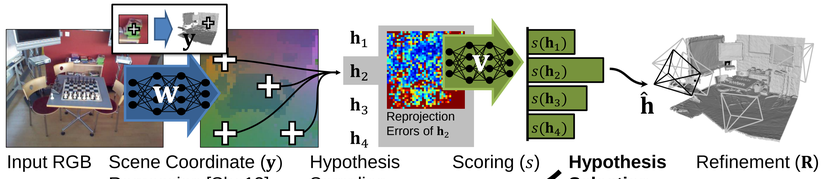
- **Red: Learn everything.** CNN predicts pose directly.
- **Orange: Learn two components of a geometric pipeline.** Our previous work [Bra17].
- **Cyan: Learn one component of a geometric pipeline.** This work.
- **Green: Ground truth** camera path.



Contributions

- **Fully differentiable**, **robust pose optimization** without learnable parameters **on top of learned scene coordinate regression**
- **Learning** scene coordinate regression **without a 3D scene model or depth maps**
- **Stable end-to-end training** due to new approximation of refinement gradients, and controlling the entropy of pose hypotheses
- We **exceed state-of-the-art** on camera localization **on three datasets** (indoor and outdoor)

Previous Work: Differentiable RANSAC (DSAC) [Bra17]



Probabilistic Pose Selection

$$\hat{\mathbf{h}}^{\mathbf{w}} = \mathbf{h}_j^{\mathbf{w}}, \text{where } j \sim P(j|\mathbf{w})$$

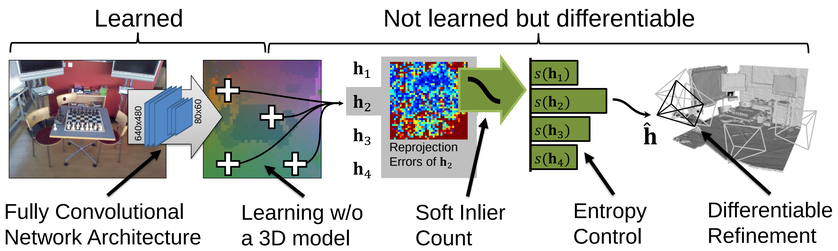
$$P(j|\mathbf{w}) = \exp\left(s(\mathbf{h}_j^{w})\right) / \sum_k \exp(s(\mathbf{h}_k^{\mathbf{w}}))$$

DSAC Learning Objective

$$\frac{\partial}{\partial \mathbf{w}} \mathbb{E}_{j\sim P(j|\mathbf{w},\mathbf{v})}\left[\ell(\mathbf{R}(\mathbf{h}_j^{\mathbf{w}}, \mathbf{w}), \mathbf{h}^*)\right] =$$

$$\mathbb{E}_{j\sim P(j|\mathbf{w})}\left[\ell(.)\frac{\partial}{\partial \mathbf{w}} \log P(j|\mathbf{w}) + \frac{\partial}{\partial \mathbf{w}}\ell(.)\right]$$

Updated Pipeline



Learning without a 3D Scene Model

Training scene coordinate regression in 3 stages:

1) $\min \sum_i \left\| \mathbf{y}_{i(\mathbf{w})} - \mathbf{y}_i^* \right\|$, with $\mathbf{y}_i^* = \mathbf{h}^* \left[\frac{dx_i}{f}, \frac{dy_i}{f}, d, 1\right]^T$

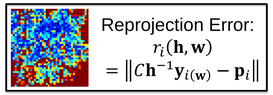
2) $\min \sum_i \left\| C\mathbf{h}^{*-1}\mathbf{y}_{i(\mathbf{w})} - \mathbf{p}_i \right\|$

Hypothesis Score: Soft Inlier Count

Previously: learned $s(\mathbf{h})$ - hard to regularize, overfits

Inlier Count: $s(\mathbf{h}) = \sum_i \mathbb{1}[\tau - r_i(\mathbf{h}, \mathbf{w})]$ - not differentiable

Soft In. Count: $s(\mathbf{h}) = \sum_i \text{sig}(\tau - \beta r_i(\mathbf{h}, \mathbf{w}))$ – differentiable



Reprojection Error: $r_i(\mathbf{h}, \mathbf{w}) = \left\| C\mathbf{h}^{-1}\mathbf{y}_{i(\mathbf{w})} - \mathbf{p}_i \right\|$

[Sho13] "Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images", Shotton et al., CVPR'13
[Ken15] "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization" Kendall et al., ICCV'15
[Ken17] "Geometric Loss Functions for Camera Pose Regression with Deep Learning" Kendall and Cipolla, CVPR 2017
[Sat16] "Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization" Sattler et al., PAMI 2016
[Bra17] "DSAC - Differentiable RANSAC for Camera Localization", Brachmann et al., CVPR'17

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 647769)

European Research Council Established by the European Commission

European Commission

Horizon 2020 European Union funding for Research & Innovation

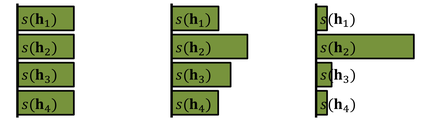
Hypothesis Score: Entropy Control

Hypothesis Distribution:

$$P(j|\mathbf{w}, \alpha) = \frac{\exp(\alpha s(\mathbf{h}_j, \mathbf{w}))}{\sum_k \exp(\alpha s(\mathbf{h}_k, \mathbf{w}))}$$

Entropy:

$$S(\alpha) = -\sum_j P(j|\mathbf{w}, \alpha)\ \log P(j|\mathbf{w}, \alpha)$$



High Entropy: No separation, Instable

Medium Entropy: Rich Learning Signals, Stable

Low Entropy: Few Learning Signals, Instable

Keep target entropy S^* during training by adjusting α: $\text{argmin}_\alpha |S(\alpha) - S^*|$

Differentiable Refinement

Refinement $\mathbf{R}$ optimizes re-projection errors $\mathbf{r}_\mathcal{I}$ of inlier set $\mathcal{I}$: $\mathbf{R}(\mathbf{h}) = \text{argmin}_{\mathbf{h}'} \|\mathbf{r}_\mathcal{I}(\mathbf{h}', \mathbf{w})\|^2$

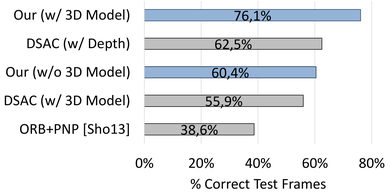
Gauss-Netwon update step: $\mathbf{R}^{t+1} = \mathbf{R}^t - (J_\mathbf{r}^T J_\mathbf{r})^{-1} J_\mathbf{r}^T \mathbf{r}_\mathcal{I}(\mathbf{R}^t, \mathbf{w})$

Last update: $\mathbf{R}(\mathbf{h}) = \mathbf{h}_\mathrm{O} - (J_\mathbf{r}^T J_\mathbf{r})^{-1} J_\mathbf{r}^T \mathbf{r}_\mathcal{I}(\mathbf{h}_\mathrm{O}, \mathbf{w})$, with $\mathbf{h}_\mathrm{O} = \mathbf{R}^{t=\infty}(\mathbf{h})$

Gradient approximation: $\frac{\partial}{\partial \mathbf{w}} \mathbf{R}(\mathbf{h}) \approx -(J_\mathbf{r}^T J_\mathbf{r})^{-1} J_\mathbf{r}^T \frac{\partial}{\partial \mathbf{w}} \mathbf{r}_\mathcal{I}(\mathbf{h}_\mathrm{O}, \mathbf{w})$

Results

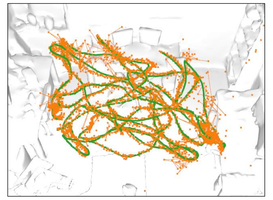
7Scenes [Sho13] Results (Error < 5cm,5°)



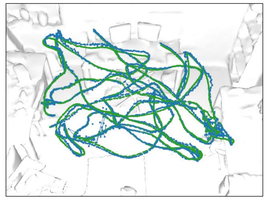
Cambridge Landmarks [Ken15] Results

Avg. Median Err.	w/ 3D Model	w/o 3D Model
PoseNet [Ken17]	1.43m, 2.9°	1.63m, 2.8°
Active Search [Sat16]	0.29m, 0.6°	-
DSAC [Bra17]	0.31m, 0.8°	-
Our	**0.14m, 0.3°**	**0.19m, 0.5°**

DSAC (w/ 3D Model)



Our (w/ 3D Model)



Our (w/o 3D Model)