BEV-SLD: Self-Supervised Scene Landmark Detection for Global Localization with LiDAR Bird's-Eye View Images
Supplementary Material

David Skuddis Vincent Ress Wei Zhang Vincent Ofosu Nyako Norbert Haala
Institute for Photogrammetry and Geoinformatics, University of Stuttgart, Germany
{firstname.lastname, vincent.ofosu-nyako}@ifp.uni-stuttgart.de

1. Summary

In this supplementary material, we provide additional details on the proposed method. We describe how the landmark density is determined, outline the training procedure, present the learned landmarks across different datasets, and discuss failure cases and practical considerations for deployment.

2. Landmark Density

As stated in the main paper, the number of landmarks remains constant during training. Thus, the landmark density is determined when the initial landmarks are created. Landmark initialization consists of two steps. First, we iterate over all keyframe poses of the reference sequence. For each pose, we divide the area covered by the corresponding BEV image into $d_P \times d_P$ patches, matching the number of patches used in the landmark learning loss function, and place landmark candidates at the patch centers. This process is illustrated in Fig. 1. Since the BEV images of neighboring keyframes overlap substantially, the accumulated landmark candidates have a much higher density than desired. Therefore, we apply a grid averaging filter with grid size s_{grid} to reduce the number of landmarks. The final number of landmarks is thus controlled by the grid size s_{grid}, with at most one landmark retained per grid cell of area $a_{grid} = s_{grid}^2$. Instead of using the grid size s_{grid} directly as a hyperparameter, we introduce the landmark density ρ_{lm}, defined as the ratio between the patch area $a_{patch} = l_{patch}^2$ and the grid area a_{grid}, where l_{patch} denotes the patch edge length. This allows the grid size to adapt automatically to other hyperparameters, such as the BEV image resolution or patch size, while preserving a defined landmark density.

$$\rho_{lm} = \frac{a_{patch}}{a_{grid}} = \frac{l_{patch}^2}{s_{grid}^2} \tag{1}$$

To be able to calculate the grid size as a function of the landmark density ρ_{lm}, we reformulate the formula:

$$s_{grid} = \frac{l_{patch}}{\sqrt{\rho_{lm}}} \tag{2}$$

We can now use ρ_{lm} as a control parameter and adaptively compute the grid size used to downsample the initial landmark candidates. For example, when $\rho_{lm} = 1.0$, the grid size matches the patch size. Empirically, however, we found that a lower landmark density leads to better results, and therefore we use $\rho_{lm} = 0.2$ in all experiments. This means that the grid size s_{grid} is approximately twice the patch edge length l_{patch}. Although this introduces a relative sparsity of landmarks during landmark position learning, it ultimately improves localization performance.

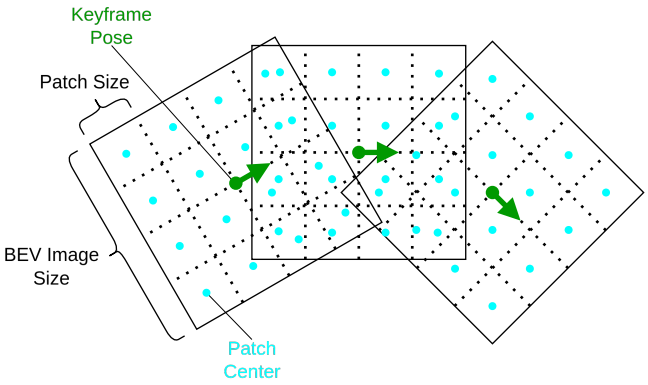

Figure 1. Landmark initialization visualized for three exemplary keyframe poses.

3. Training Setup

We implemented the proposed method in PyTorch [4]. We set the hyperparameter d_P to 16, which yields 16×16 patches in the loss function (see main paper Fig. 2). The loss weights are chosen as $\alpha = 1.0$, $\beta = 30.0$ and $\gamma = 3.0$. Optimization was performed using stochastic gradient descent (SGD) with a momentum of 0.9, combined with a StepLR

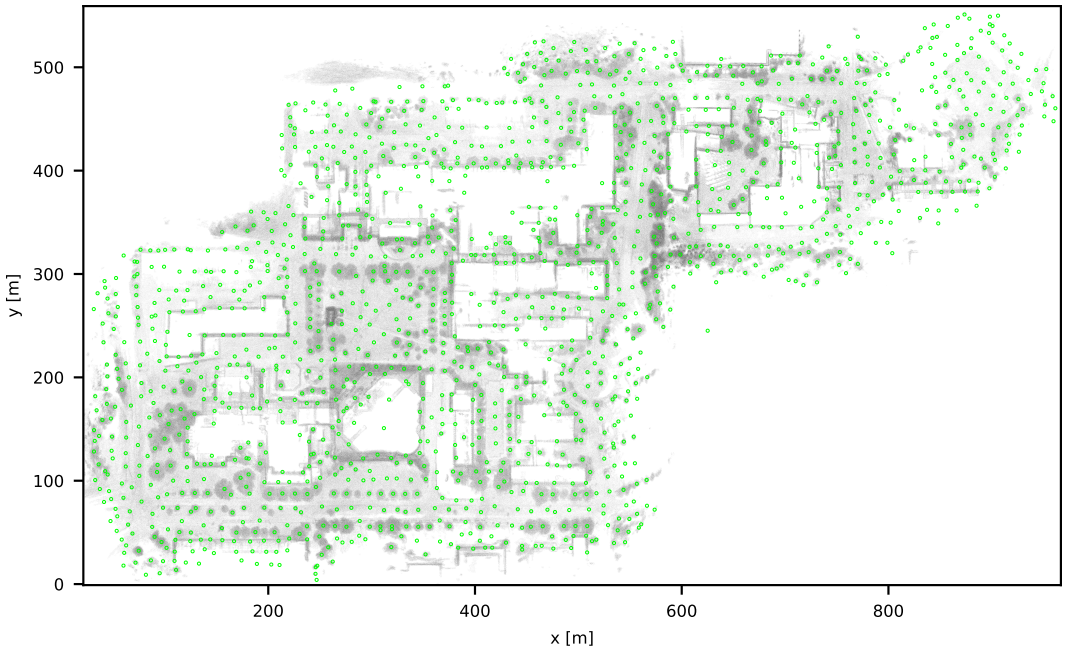

Figure 2. Landmarks in **green** after self-supervised joint landmark detection and positioning on sequence 2012-01-15 of the NCLT dataset [1] in a campus environment. In total, 1592 landmarks are utilized in this scene. A point cloud of the entire scene in the background is used for illustration purposes.

scheduler. The learning rate was initialized at 4×10^{-4} and decayed to 4×10^{-5}. For all experiments, 3 % of the reference sequence frames were used for validation.

4. Landmark Analysis on Datasets

In Figs. 2, 3, and Fig. 1 (main paper), we illustrate the learned landmarks for each dataset used for localization. For the MCD dataset [3] and the NCLT dataset [1], the method selects tree trunks and building corners among other things as landmarks. In contrast, for the Wild-Places dataset [2], it is challenging even for a human observer to provide an intuitive explanation for the landmark locations.

5. Failure Cases

In our experiments, suboptimal parameter settings degrade performance rather than cause complete failure. A low landmark density, $\rho_{lm} < 0.1$, may lead to landmark-free regions, whereas a high landmark density, $\rho_{lm} > 0.5$, may cause landmark collapse (co-located landmarks). While the latter does not affect the heatmaps, it can destabilize the correspondence loss. Performance degradation can also occur when training is stopped too early. In addition, the proposed method may fail in structureless or ambiguous scenes.

6. Practical Considerations for Deployment

In this section, we discuss practical considerations for deploying the proposed method in real-world robotic applications.

6.1. Gravity Alignment

LiDAR point clouds may require gravity alignment, particularly in non-planar environments. This can be achieved using IMU-based attitude estimation, which also enables 5 DoF pose estimation when combined with our method, which estimates 3 DoF.

6.2. Motion Distortion Correction

LiDAR point clouds recorded on dynamic platforms may capture the environment with motion-induced distortions. These distortions should be corrected, for example, by combining IMU angular velocity measurements with wheel-odometry-based velocity estimates.

6.3. Continuous Localization and Sensor Fusion

The proposed method focuses on LiDAR-based global localization. In our experiments, the global pose is estimated in a one-shot manner without prior information. For real-world deployment, however, sequential global localization estimates can be fused with IMU data and wheel odometry to enable continuous localization with improved robustness. In this setting, an additional inlier threshold may help reject

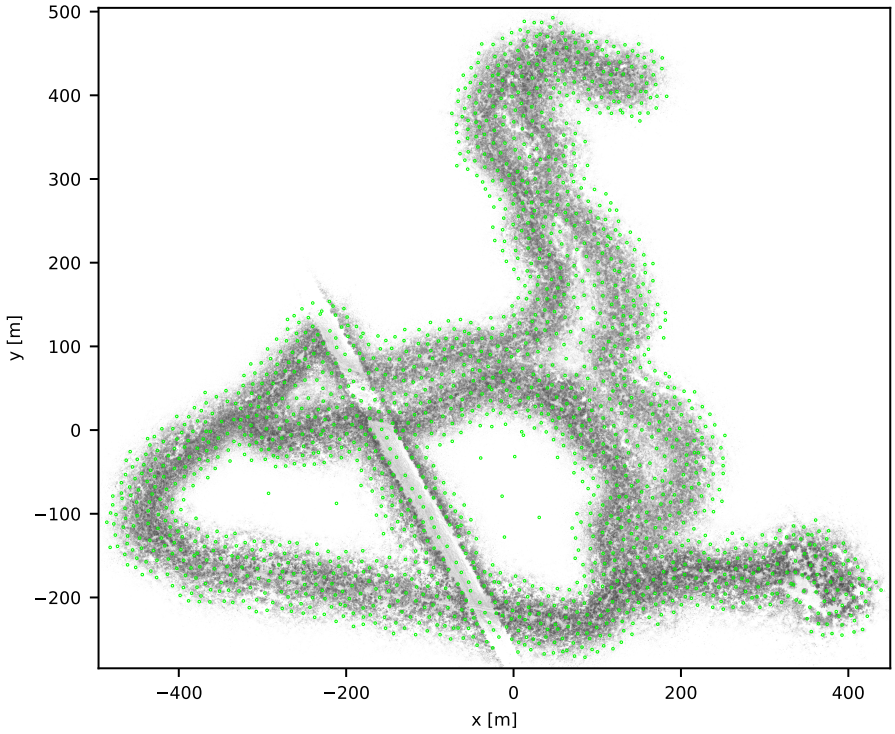

Figure 3. Landmarks in **green** after self-supervised joint landmark detection and positioning on sequence V-03 of the Wild-Places dataset [2] in a forest environment. In total, 1757 landmarks are utilized in this scene. A point cloud of the entire scene in the background is used for illustration purposes.

global pose estimates supported by too few landmark inliers.

6.4. Incremental Map Updates

The current framework does not directly support incremental map updates. However, a simple extension would be to initialize the network and the landmark list with additional placeholder landmarks. These placeholder landmarks could be masked in the loss and reinitialized when the map expands, followed by a fine-tuning stage. Since the network uses LeakyReLU activations, initially inactive correspondence outputs remain reactivatable. Overall, this extension could be realized with modest effort.

References

[1] Nicholas Carlevaris-Bianco, Arash K. Ushani, and Ryan M. Eustice. University of Michigan North Campus long-term vision and lidar dataset. *International Journal of Robotics Research*, 35(9):1023–1035, 2015. 2

[2] Joshua Knights, Kavisha Vidanapathirana, Milad Ramezani, Sridha Sridharan, Clinton Fookes, and Peyman Moghadam. Wild-places: A large-scale dataset for lidar place recognition in unstructured natural environments. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11322–11328, 2023. 2, 3

[3] Thien-Minh Nguyen, Shenghai Yuan, Thien Hoang Nguyen, Pengyu Yin, Haozhi Cao, Lihua Xie, Maciej Wozniak, Patric Jensfelt, Marko Thiel, Justin Ziegenbein, and Noel Blunder. Mcd: Diverse large-scale multi-campus dataset for robot perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22304–22313, 2024. 2

[4] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1