Received: 1 August 2018 | Revised: 16 March 2019 | Accepted: 24 March 2019

DOI: 10.1002/rob.21870

REGULAR ARTICLE

WILEY

Appearance-based landmark selection for visual localization

Mathias Bürki[1] | Cesar Cadena[1] | Igor Gilitschenski[2] | Roland Siegwart[1] | Juan Nieto[1]

[1]Autonomous Systems Lab, ETH Zürich, Zürich, Switzerland

[2]Computer Science and Artificial Intelligence Lab, MIT, Cambridge, Massachusetts

Correspondence
Mathias Bürki, Autonomous Systems Lab, ETH Zürich, Zürich, Switzerland.
Email: mathias.buerki@mavt.ethz.ch

Funding information
Bundesbehörden der Schweizerischen Eidgenossenschaft, Grant/Award Number: 15.0284; H2020 European Institute of Innovation and Technology, Grant/Award Number: 688652; EU H2020 research project "UP-Drive", Grant/Award Number: 688652; Swiss State Secretariat for Education, Research and Innovation, Grant/Award Number: 15.0284

Abstract

Visual localization in outdoor environments is subject to varying appearance conditions rendering it difficult to match current camera images against a previously recorded map. Although it is possible to extend the respective maps to allow precise localization across a wide range of differing appearance conditions, these maps quickly grow in size and become impractical to handle on a mobile robotic platform. To address this problem, we present a landmark selection algorithm that exploits appearance co-observability for efficient visual localization in outdoor environments. Based on the appearance condition inferred from recently observed landmarks, a small fraction of landmarks useful under the current appearance condition is selected and used for localization. This allows to greatly reduce the bandwidth consumption between the mobile platform and a map backend in a shared-map scenario, and significantly lowers the demands on the computational resources on said mobile platform. We derive a landmark ranking function that exhibits high performance under vastly changing appearance conditions and is agnostic to the distribution of landmarks across the different map sessions. Furthermore, we relate and compare our proposed appearance-based landmark ranking function to popular ranking schemes from information retrieval, and validate our results on the challenging *University of Michigan North Campus long-term vision and LIDAR data sets* (*NCLT*), including an evaluation of the localization accuracy using ground-truth poses. In addition to that, we investigate the computational and bandwidth resource demands. Our results show that by selecting 20–30% of landmarks using our proposed approach, a similar localization performance as the baseline strategy using all landmarks is achieved.

KEYWORDS

landmark selection, long-term localization, multisession mapping, visual localization, wheeled robots

1 | INTRODUCTION

Visual localization systems are able to provide centimeter-accurate pose estimations of mobile robots with a low-cost sensor setup. This renders visual localization an attractive alternative to light detection and ranging (LiDAR)-based localization which today still requires mechanically complex and thus expensive hardware. However, and in contrast to aforementioned LiDAR localization, visual localization systems targeting long-term usage suffer from variations in appearance conditions which render matching between currently observed visual cues and landmarks stored in the map difficult. A promising approach to address this problem has been proposed in the form of multisession maps (Churchill & Newman, 2013; Mühlfellner et al., 2016; Paton, Mactavish, Warren, & Barfoot,

2016) that incorporate visual cues from more than one appearance condition. The resulting maps, however, quickly grow in size and become impractical to handle on the mobile robotic platform. To mitigate this problem, the map can be stored on a cloud-based backend and made available to the robots in operation over a mobile data network. Apart from relieving the mobile platforms from storing large maps, such a shared-map scenario offers further advantages such as the reduction of redundant data, more efficient map maintenance, and an increased potential for collaboration between the robots. However, it also requires map data to be exchanged between the map backend and the robots in operation over bandwidth constrained mobile data networks. This renders it important to only exchange map data that can be used for localization at the particular time and place of operation. For this purpose, it may be sufficient to only transmit a fraction of map data available in the multisession map, since the latter must cover all possible appearance conditions, while the robots in operation are exposed to only one condition at a certain time and place. It is the aim of this study to exploit this potential and select landmarks for localization based on the current appearance condition. This serves the following two purposes: (a) Keep data exchange between the map backend and the mobile platform, and therewith the bandwidth consumption on a mobile network, as low as possible, and (b) lower the computational resource demands on the mobile platform, increasing the real-time capability of visual localization. At the same time, a localization performance as good as if all landmarks are used ought to be maintained. Additionally, the appearance-based landmark selection enables decoupling of the localization performance from map management. While the multisession map at the backend may be large, and resource intensive to maintain, localization on the vehicles remains as efficient as if only one map session of the current appearance condition was available (Figure 1).

In summary, we present a complete visual localization system yielding six-degree-of-freedom (6DoF) pose estimates at each timestep with the capability to perform efficient online data association through appearance-based landmark selection.

The main contributions of this paper are as follows:

- We derive, analyze, and compare a ranking function for appearance-based landmark selection based on appearance equivalence classes, which can be shown to maximize the number of observed landmarks with respect to the current appearance condition.
- We investigate in detail the impact of the incorporation of *observation sessions*, a lightweight extension to the visual maps boosting the landmark selection performance.
- In an extensive evaluation involving three collections of outdoor data sets, one of them publicly available, we thoroughly investigate the performance of the appearance-based landmark selection in real-world conditions, and compare against related popular ranking schemes from information retrieval.
- An analysis of the computational performance demonstrates the real-time capability of the appearance-based landmark selection and reveals its potential to reduce the computational load on the vehicle platforms.

This paper builds upon our previous work on appearance-based landmark selection presented in Bürki et al. (2018) and Bürki, Gilitschenski, Stumm, Siegwart, and Nieto (2016) and extends it in several aspects: We derive several appearance-based ranking functions, relate them to popular ranking schemes from information retrieval, and evaluate the expected performance of our proposed solution on a related state-of-the-art simultaneous localization and mapping (SLAM) framework which keeps separate maps for different appearance conditions. In addition to that, we present an extensive evaluation on the publicly available *University of Michigan North Campus long-term vision and LIDAR data sets* (*NCLT*) collection, including an assessment of the localization *accuracy* with respect to ground truth. The evaluation on the *NCLT* data set collection further demonstrates the applicability of our proposed appearance-based landmark selection on a second robotic platform in highly challenging long-term outdoor conditions, and with a considerably different camera system than the one on the vehicle used in the *parking-lot* and *city environment*. A detailed investigation of the computational

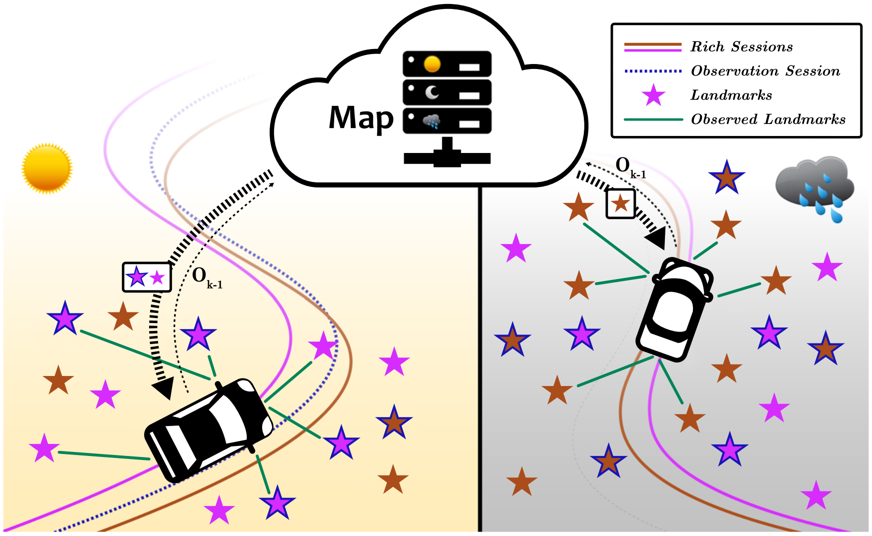


FIGURE 1 Shared-map scenario motivating our work. One large map containing landmarks from multiple *rich-* and *observation sessions* is stored and maintained on a cloud-based map backend. Vehicles en route under different appearance conditions retrieve selected landmarks matching their operation conditions (thick dashed arrow), use those landmarks for visual localization (turquoise lines), and report back a set of recently observed landmark identifiers (thin dashed arrow) [Color figure can be viewed at wileyonlinelibrary.com]

performance further not only shows the real-time capability of the localization pipeline, but also reveals lower computational resource demands as a second benefit of our proposed appearance-based landmark selection apart from reduced bandwidth consumption.

2 | RELATED WORK

Outdoor environments are subject to appearance change, such as change in illumination, as well as change in weather and seasonal conditions. This has a severe impact on long-term operations of outdoor visual localization systems, as in many environments, change in appearance is much more pronounced than structural change, and already with relatively small time offsets of only several hours between mapping and localization it may become difficult to match currently observed visual cues against a visual map. The approaches to overcome this can in general be distinguished into two categories: (a) initiatives to overcome the appearance dependency, and (b) attempts to collect and organize appearance-dependent visual features from differing conditions. We first present an overview over relevant work associated with category (a), before investigating approach (b) in detail in the remainder of this section.

Lategahn, Beck, and Stiller (2014) propose a local feature descriptor named DIRD which exhibits illumination invariance superior to other popular local features such as SURF (Bay, Tuytelaars, & Van Gool, 2006) or BRIEF (Calonder et al., 2012). Nevertheless, the ability to cover appearance change is ultimately still limited in situations with such strong differences in illumination that let already the location of keypoints be different. In another approach to reduce the appearance change in images, Maddern et al. make use of the spectral properties of color cameras to apply an illumination invariant gray-scale transformation to images, effectively removing shadows and reducing the appearance variation due to sunlight (Clement, Kelly, & Barfoot, 2017; Maddern et al., 2014; McManus, Churchill, Maddern, Stewart, & Newman, 2014; Paton, MacTavish, Ostafew, & Barfoot, 2015). This on the one hand requires a photometrically calibrated color camera, and on the other hand is only able to reduce the appearance change due to sunlight. Any other source of appearance change, such as seasonal change, or daytime versus nighttime, are not tackled. McManus, Upcroft, and Newman (2014) propose to learn location-dependent detectors that retrieve large patches in images deemed descriptive for the respective place. While this shows promising redetection performance across vastly different appearance conditions, it is not able to allow as precise a metric localization compared to using local corner-based features.

As mentioned above, an alternative approach to tackle the challenge of appearance change lies in the attempt to enrich a visual map with features from varying conditions to extend its appearance coverage and allow localization across a wide range of differing conditions. Konolige and Bowman (2009) present a visual mapping algorithm that is able to aggregate visual cues from different states of the environment into so-called "views," which are managed over long time spans. Their system, however, mainly targets structural changes in dynamic indoor environments.

In a similar vein, Milford and Prasser have extended RatSLAM (Milford, Wyeth, & Prasser, 2004) in Prasser, Milford, and Wyeth (2006) and Milford, Prasser, and Wyeth (2005) to include "local view cells" and abstract "experience maps" which allow associating previously visited places under varying appearance with the same physical location on the one hand, and the creation and maintenance of a spatially consistent map representations across different environmental states on the other hand. However, the ability to yield a precise metric pose estimate of the robot in a Euclidean coordinate system is limited. In contrast to that, Churchill and Newman (2013) propose a visual mapping framework called "experience-based mapping" which explicitly creates and maintains separate and detached visual maps for varying outdoor environmental conditions. While this allows precise metric localization under essentially any appearance condition, the visual pose estimate can only be expressed with respect to a Euclidean coordinate system that is unique to each experience. Any interpretation in a common coordinate frame requires links between experiences based on additional sensor modalities, such as (differential) global positioning system (GPS), which may considerably deteriorate the accuracy of the resulting pose estimate. For this reason, attempts have been made to represent visual features—or landmarks respectively—from different appearance conditions in a single Euclidean coordinate frame. Paton et al. (2016) present a visual mapping framework able to incorporate and correlate landmarks from different appearance conditions in outdoor environments with respect to a manually taught reference path. This enables a mobile robot to autonomously repeat the reference route in vastly different appearance conditions. The principle behind the multisession mapping framework proposed by Mühlfellner, Furgale, Derendarz, and Philippsen (2015) and Mühlfellner et al. (2016) is similar. However, there is no notion of a privileged path, or session, respectively, in the map. Instead, the resulting map offers accurate metric localization under any appearance condition represented by the map sessions with respect to a single coordinate frame.

While incorporating landmarks from varying environmental states into a single map can successfully enable visual localization in vastly different appearance conditions, the resulting maps quickly grow in size and become impractical to maintain. Therefore, considerable efforts have been made to optimize map representations such that keeping redundant landmarks is avoided and only a minimal set of landmarks that allow localization across different appearance conditions is maintained. In Dayoub, Cielniak, and Duckett (2011), a long-term short-term memory model is proposed to dynamically distinguish useful from outdated landmarks. Such a model of change is especially suited to environments that exhibit some fraction of features stable in appearance (e.g., corners on the ceiling), but does not have the ability to represent multiple environment states at the same time. In contrast to that, Hochdorfer and Schlegel (2009), Konolige and Bowman (2009), and Milford and Wyeth (2010) employ clustering of images, or landmark, respectively, to keep the number of visual cues bounded. While Konolige and Bowman (2009) use a similarity measure between local feature clusters to discard redundant "views," Hochdorfer and Schlegel (2009) remove visual data on the landmark level by assessing the usefulness of individual landmarks inside a local feature cluster based on position uncertainty. Milford and Wyeth (2010), on the other hand, simply discard landmarks randomly to keep the data density within a cluster bounded. More recent and advanced approaches to bounding the map size for metric visual localization systems are presented in Dymczyk, Lynen, Bosse, and Siegwart (2015) and Mühlfellner et al. (2016). Mühlfellner et al. (2016) compare a number of different algorithms to prune landmarks in a multisession map, demonstrating selection criteria involving the number of observed sessions, and the total number of observations of a landmark to yield good metric localization performance over long time spans while keeping the map size limited. Along a similar vein, Dymczyk et al. (2015) propose to solve an integer linear problem with cost terms favoring landmarks with a large number of observations on the one hand, and guaranteeing a minimal number of landmarks observed from every keyframe on the other hand.

In contrast to metric localization, efficient map representations and landmark selection has also been studied in the context of place recognition. In Fayin and Košecká (2006) and Schindler, Brown, and Szeliski (2007), only the SIFT (Lowe, 1999) features contributing the most to the distinctiveness of places are retained in the map. Similarly, in Cummins and Newman (2011), Johns and Yang (2013), Li, Snavely, and Huttenlocher (2010), Stumm, Mei, Lacroix, and Chli (2015), and Johns and Yang (2014), covisibility of features is used to efficiently and effectively solve the place recognition problem.

While all of these works describe successful approaches to mitigate the problem of ever-growing visual maps, they only address offline map maintenance with the goal of computing as small a map representation as possible while at the same time maintaining the appearance coverage over different conditions. However, as mentioned in Section 1, in long-term operations in outdoor environments, the map must cover a far wider range of appearance conditions than what the robots in operation require at a given point in time. This offers a potential to further optimize data usage and minimize computational demands on the robot platforms by distinguishing currently useful data based on the observed appearance conditions in an online fashion. In this regard, Linegar, Churchill, and Newman (2015) have presented an algorithm for the experience-based mapping framework which adaptively selects the best matching "experience" in an online fashion. While their work addresses a similar motivation as ours, there are substantial differences as a consequence of the different underlying map representation and mapping framework. For instance, the different appearance conditions are represented as individual maps, and therefore their selection of useful map data occurs on the level of "experiences." In contrast to that, and due to the fact that our landmarks in the map from the different appearance conditions are all expressed with respect to a single coordinate frame, we are able to select map data matching the current appearance conditions on the level of individual landmarks. In addition to that, we may also select landmarks from more than one session in the map at a time, allowing to benefit from potentially overlapping appearance conditions. In a similar vein, Mactavish, Paton, and Barfoot (2017) propose an online selection of useful map data for their Visual Teach & Repeat framework. Analogous to Linegar et al. (2015) and in contrast to our work, they perform the selection on the level of "experiences," are, however, able to simultaneously use more than one "experience" for localization. Their work differs further to ours in the methodology at the basis of the selection algorithm. While they compute and compare current images to their map images employing a visual bag-of-words representation, we evaluate the current appearance conformity on the basis of co-observability of recently observed landmarks. This relieves us from having to train and rely on a vocabulary.

3 | BACKGROUND

In this section we briefly introduce the components of our localization and mapping system. This overview supports and facilitates the understanding of subsequent sections in this paper. We first describe the mapping process and the resulting map structure, before presenting our visual localization module in detail.

3.1 | Mapping

Mapping is performed in an offline process. We track FREAK (Alahi, Ortiz, & Vandergheynst, 2012) features[1] from one camera frame to the next, and triangulate the position of these landmarks using wheel odometry. With this, a map is generated with a graph of the vehicle's poses (position and orientation) at image acquisition times, as well as the landmark positions in 3D space. If necessary, loops are closed using the matching-based loop-closure algorithm (Sattler, Leibe, & Kobbelt, 2011). Finally, both the poses of the vehicle and the positions of the landmarks are jointly optimized in a bundle-adjustment routine.

Further mapping sessions are added by first localizing the new data set in an offline process against the pre-existing map. This generates both initial pose estimates for the vehicle in the new data set and associations between features from the camera images of the new data set and landmarks of the pre-existing map. In addition, new landmarks are spawned from features of the new data set that failed to find a matching map landmark. Finally, the resulting multisession map is optimized again with bundle adjustment. Note that all information, that is, both the landmark positions and vehicle poses, of all map sessions, is expressed in the same metric three-dimensional coordinate frame of reference, denoted by $\mathcal{F}_W$.

3.2 | Localization

The aim of the localization module is to estimate the vehicle's 6DoF pose with respect to the map coordinate frame of reference $\mathcal{F}_W$, given one or more camera images acquired at a specific point in time,

[1] As we demonstrate in Section A.2, our apperance-based landmark selection algorithm is agnostic to the type of local feature descriptor used. However, in practice, not every descriptor may be equivalently well suited for building multisession maps, and the choice of descriptor can further be restricted by computational constraints.

and some rough prior knowledge about the current vehicle's location. We refer to this localization paradigm as local iterative localization, in contrast to global localization or loop closure where no a priori knowledge of the vehicle's pose is available.

Let $map \coloneqq \{V, L, E\}$ denote the map containing a set of vertices V (robot's poses), a set of landmarks' positions L, and a set of edges E capturing the observation relation between vertices and landmarks. Let further $\mathcal{F}_B$ denote the vehicle body coordinate frame. Image acquisitions occur repeatedly along a traversal through the mapped area at a given frequency. Instead of referring to the time of image acquisition, we enumerate them with index k, and refer to the set of images recorded at the kth acquisition with I_k. With this, we can formulate our local iterative localization problem as follows:

$$\bar{\mathcal{T}}_{W\ B_k} = localize(I_k, \hat{\mathcal{T}}_{W\ B_k}, map), . \quad (1)$$

with $\bar{\mathcal{T}}_{W\ B_k}$ denoting the estimate of the vehicle's pose expressed in the map coordinate frame of reference. Analogously, $\hat{\mathcal{T}}_{W\ B_k}$ denotes the rough prior guess of the same quantity. Using $\hat{\mathcal{T}}_{W\ B_k}$, landmarks are retrieved from the map that have been observed from near-by, and their respective 3D points are backprojected into the camera image plane, where they are matched against the feature descriptors extracted on the query images based on pixel and descriptor distance. The refined pose estimate $\bar{\mathcal{T}}_{W\ B_k}$ is calculated from solving a nonlinear least-squares optimization problem involving an image-plane projection error constraint with a robust cost function for every keypoint-landmark match. Observations under a predefined backprojection error are considered inliers of the localization iteration k, and the respective landmarks form the set of observed landmarks $\mathcal{O}_k$. The prior guess of the pose for the subsequent localization at iteration $k + 1$ is readily obtained from forward propagating the previous pose estimate with the use of wheel odometry:

$$\hat{\mathcal{T}}_{W\ B_{k+1}} \coloneqq \bar{\mathcal{T}}_{W\ B_k} \mathcal{T}^{\text{odo}}_{B_k B_{k+1}}. \quad (2)$$

The main steps of the localization module are summarized in Algorithm 1 in Section 4.

Note that the matching in image space between 2D features and 3D landmarks requires an association of one feature descriptor for every landmark in the map. For our experiments, we group all observations associated with the same 3D point based on their association with the respective *rich session* (see Section 3.3). For every group, we then evaluate the one observation with the smallest accumulated descriptor distance to all other descriptors of the same group, have the descriptor of this observation, and, together with respective 3D point, form a landmark used for selection and matching.

3.3 | Rich and observation sessions

In Section 3.1, we have described how a map can be enriched with landmarks from multiple sessions by localizing a data set against the map in an offline process. We refer to a data set added to a map in this fashion as a *rich session*. Adding a *rich session* to a map extends the appearance coverage of the map with the conditions present in the respective data set. At the same time, however, the size of the map, and the complexity and runtime of the optimization with bundle adjustment are considerably increased.

In contrast to that, a data set can also be added to the map without the addition of new landmarks. For this, the data set is localized against the map, and the vertex poses along the trajectory are added to the pose graph of the map, analogous to adding a *rich session*. Instead of tracking and triangulating new landmarks, however, only the relation between keypoints from the new data set and observed pre-existing map landmarks is registered. This barely increases the size of the map and does not have an impact on the complexity of bundle adjustment. Although this does not extend the appearance coverage either, it increases the landmark co-observation statistics, which can be beneficial for the performance of appearance-based landmark selection. A data set added to the map in this fashion is referred to as an *observation session*.

4 | APPEARANCE-BASED LANDMARK SELECTION

In this section, the selection of landmarks for localization based on appearance is described in detail. After formally presenting the problem at hand, we introduce a landmark ranking function used to prioritize relevant landmarks for the selection process. We conclude this section by relating our problem of appearance-based landmark selection to popular ranking schemes from information retrieval.

4.1 | Problem formulation

The goal of appearance-based landmark selection is to decide which of the landmarks in the map are likely to be seen under the present appearance condition. In a generalized manner, this problem can be formulated as follows:

$$S_k = selectLandmarks(f, C_k, n, \mathcal{A}), \quad \text{with} \quad S_k \subseteq C_k, \quad (3)$$

where C_k denotes the set of geometrically visible candidate landmarks, S_k denotes the set of selected landmarks, f refers to the landmark ranking function, n to the number landmarks to select, and the current appearance condition is expressed as $\mathcal{A}$. The ranking function f maps a landmark l to a score, that is,

$$f: l \to [0, 1] \quad \forall\, l \in C_k, \quad (4)$$

whereas a landmark is defined as three tuples:

$$l \coloneqq (p_l, d_l, Z_l), \quad \text{with} \quad Z_l \subseteq Z, \quad \forall\, l \in L$$

with p_{l_i} denoting the 3D point expressed in the frame of reference $\mathcal{F}_W$, d_l denoting the descriptor associated with landmark l, and Z_l denoting the set of map sessions in which the landmark was observed. The set of all map sessions is denoted by Z.

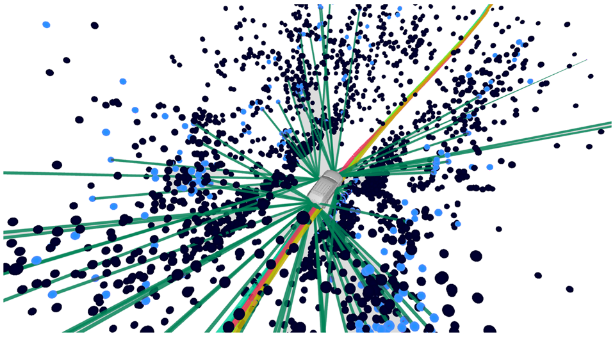


FIGURE 2 Snapshot visualization of our landmark selection. The thick colored lines depict the pose graph of the map, while the candidate landmarks C are shown as black spheres, and selected landmarks S as blue spheres. The turquoise lines indicate inlier observations between the four cameras and some of the selected landmarks after the pose refinement step [Color figure can be viewed at wileyonlinelibrary.com]

The set of selected landmarks S_k is formed by applying the ranking function f to every landmark $l \in C_k$, before selecting n top-ranked landmarks. In this study, we choose n to be relative to the number of candidate landmarks available at iteration k, formally expressed as follows:

$$\begin{aligned} U_k &:= \{l \in C_k | f(l) > 0\}, \\ n &:= \min(\alpha |C_k|, |U_k|), \quad \text{with} \quad \alpha \in [0, 1]. \end{aligned} \tag{5}$$

Preselecting the candidate landmarks based on the condition $f(l) > 0$ allows the ranking function to exclude certain landmarks from being selected. This property is used by the ranking function f_{MRS} as described in Section 5.3. A visualization of this landmark selection paradigm can be found in Figure 2.

In the following section, we elaborate in detail on how to find a tangible expression for $\mathcal{A}$ and propose a formulation for a ranking function.

4.2 | The ranking function

The ranking function ought to reflect the probability of successfully forming a match between a map landmark and a feature extracted from the current set of images under the current appearance condition $\mathcal{A}$.

To motivate the formulation for our proposed appearance-based landmark ranking function, we introduce it from a probabilistic perspective. We are thus interested in evaluating the following quantity: $P(l \in \mathcal{O} | \mathcal{A})$. This denotes the probability of observing landmark l under the current appearance condition $\mathcal{A}$. By ranking all candidate landmarks according to this probability, and selecting some number of top-ranked landmarks, we achieve our goal of maximizing the number of observed landmarks.

4.2.1 | Ranking landmarks based on appearance equivalence classes

Unfortunately, $\mathcal{A}$ is an abstract, intangible entity and not directly observable. However, as every traversal through the environment is related to the particular appearance condition present during that time, all available information regarding the probability of observing landmark l under some appearance condition $\mathcal{A}$ is encoded in the map session observation relation of landmarks. That is, if l_i and l_j were observed in the same sessions, that is, $Z_{l_i} = Z_{l_j}$, it can be assumed that

$$P(l_i | \mathcal{A}) = P(l_j | \mathcal{A}). \tag{6}$$

This allows approximation by substituting the current appearance condition $\mathcal{A}$ by the respective set of map sessions a landmark has been observed in, that is,

$$P(l \in \mathcal{O} | \mathcal{A}) \approx P(l \in \mathcal{O} | Z_l). \tag{7}$$

This renders the conditioning on the appearance condition tangible, as the observing map session relations of landmarks are well-defined and countable. Note that we employ a common abuse of notation by interpreting the expression $P(l \in \mathcal{O} | Z_l)$ as the probability of observing landmark l, given it has been observed in the past in the map sessions Z_l. We can thus group all landmarks into distinct equivalence classes, and model the observation likelihood with a simple Bernoulli distribution, that is,

$$P(l \in \mathcal{O} | Z_l) \sim Ber(\theta^{[l]}), \quad \text{with} [l] := \{l_j \in L | Z_{l_j} = Z_l\}. \tag{8}$$

It remains to estimate the appearance-dependent parameters $\theta^{[l]}$. For this, we employ the principle of local temporal stability of appearance conditions: Whenever the mapped area is traversed, the appearance conditions are expected to change along the route in the same manner as they have in previous traversals. Following this principle, we thus expect to again observe the same landmarks together with those that have already in the past been co-observed. This allows to compute a maximum likelihood estimate for $\theta^{[l]}$ using recently selected and observed landmarks from previous localization iterations. For this, we add subscript k to refer to localization iteration k, as described in Section 3.2:

$$\theta_k^{[l]} = P(l_o \in \mathcal{O}_k | l_o \in [l]) = \frac{P(l_o \in \mathcal{O}_k, l_o \in [l])}{P(l_o \in [l])} \approx \frac{|\mathcal{O}_{k-1}^{[l]}|}{|S_{k-1}^{[l]}|}, \tag{9}$$

with

$$\mathcal{O}_{k-1}^{[l]} := \{l_o \in \mathcal{O}_{k-1} | l_o \in [l]\}, \tag{10}$$

$$S_{k-1}^{[l]} := \{l_s \in S_{k-1} | l_s \in [l]\}. \tag{11}$$

We can interpret this quantity as the estimated relevance of appearance equivalence class $[l]$, based on recently collected statistical samples. With a limited budget of landmarks to select, prioritizing the selection according to this ranking function maximizes the number of expected observed landmarks under the current appearance condition. Note, however, that this statement of optimality only refers to the selection of landmarks based on appearance. There are further nonappearance related effects (e.g., geometry, occlusion, etc.) having an impact on whether a landmark is observed or not.

For our experiments, we use a temporal smoothing of $\theta^{[l]}$ over the $N = 50$ most recent iterations and define our ranking function accordingly:

$$f_{AEC}(l) := \frac{1}{N}\sum_{w=0}^{N-1}\theta_{k-w}^{[l]}. \tag{12}$$

4.3 | Relation to information retrieval

In this section, we relate our proposed appearance-based landmark ranking approach to common concepts in the field of information retrieval. With this, we aim at providing further theoretical context and facilitating the understanding and interpretation of the ranking function described in Equation (12).

The principles of information retrieval are usually stated in a linguistic context, where the overall goal is to retrieve a set of text documents most relevant to a given search query consisting of a set of query words (Salton & Buckley, 1988). Analogous to appearance-based landmark selection for visual localization, a ranking function is required, which assigns a relevance score to each document in the collection, according to how well the document matches the query words. It has thereby proven to be most successful to take two distinct aspects of relevance into consideration when assessing the relevance of a query word to a document. The *term frequency* aspect reflects how well a given query term represents the given document, while the *inverse document frequency* aspect attempts to reflect the overall discriminatory power of a word with respect to the entire document collection. These two aspects form two separate terms, whose product is assigned as the relevance weight of a query word with respect to a document. The overall ranking score can readily be computed either by summing over all relevance weights, or by representing the relevance weights in vector form and employing cosine similarity (Salton & Buckley, 1988). The result is the well-known *tf-idf* ranking scheme. Drawing the analogy with appearance-based landmark selection, we can interpret recently observed landmarks as the query. This allows expressing the appearance-based ranking function f_{AEC} described in Equation (12) as follows:

$$tf(l_o, l) := \begin{cases} 1 & \text{if } [l_o] = [l], \\ 0 & \text{otherwise,} \end{cases} \quad idf(l, S_{k-1}) := \frac{1}{|S_{k-1}^{[l]}|}, \tag{13}$$

$$f_{AEC}(l) = \sum_{l_o} tf(l_o, l)\, idf(l_o, S_{k-1}). \tag{14}$$

A unary *term frequency* only considers query landmark relevant if they belong to the same appearance equivalence class. The *inverse document frequency* term downweights contributions of landmarks if a large quantity of landmarks from the same appearance equivalence class have recently been selected. We note, however, that this interpretation of the *idf* term deviates from the text-book definition. This is because in the context of appearance-based landmark selection, we are rather interested in weighting the query words in relation to the set of recently selected landmarks, as opposed to the set of candidate landmarks. We further note that there are countless variations in how to formulate *tf* and *idf* terms to achieve optimal retrieval performance in a given application (Aizawa, 2003; Salton & Buckley, 1988). In Section 5.3, we introduce further sensible formulations that we compare against in our experiments.

Algorithm 1 Iterative local localization. The retrieval of nearby vertices from the pose-graph employs a distance δ and yaw angle discrepancy ϕ around the pose guess $\hat{\mathcal{T}}_{W\,B_k}$.

```
function LOCALIZE($I_k$, $\hat{\mathcal{T}}_{W\,B_k}$, $map$, $\mathcal{O}_{k-1}$)
    $K$← extractFeatures($I_k$)
    $V_k$← retrieveNearbyVertices($\hat{\mathcal{T}}_{W\,B_k}$, $\delta$, $\phi$, $map$)
    $C_k$← getLandmarksObservedFromVertices($V_k$, $map$)
    $S_k$← selectLandmarks($C_k$, $\mathcal{O}_{k-1}$, $f$)
    $M$← match2D3D($K$, $S_k$)
    $\bar{\mathcal{T}}_{W\,B_k}$, $\mathcal{O}_k$← estimatePose($M$, $\hat{\mathcal{T}}_{W\,B_k}$)
end function
```

An overview of the localization with appearance-based landmark selection in pseudocode can be seen in Algorithm 1.

5 | EVALUATION

In this section, we present the results of our evaluation, focusing on (a) demonstrating the effectiveness of selecting landmarks using the appearance-based ranking function presented in Section 4.2 in multiple challenging long-term outdoor environments, (b) comparing our proposed ranking function with related popular ranking schemes, (c) reporting on the resulting localization precision and accuracy, and (d) analyzing the computational performance of the respective localization algorithm.

To facilitate the navigation within and reading of this section, we first present a concise summary of the conducted experiments. Subsequently, the data set collections, respective sensor configurations, and evaluation metrics are introduced, before the various experiments are presented in detail. A paragraph containing our key findings concludes the evaluation section.

Please note that a direct comparison of our appearance-based landmark selection performance with the most related works (Linegar et al., 2015; Mactavish et al., 2017) is inherently difficult, as the underlying mapping framework and visual feature representations are fundamentally different, and the selection of relevant data on the level of individual landmarks constitutes a unique feature of our method. With the ranking function f_{MRS}, as introduced in Section 5.3, we aim at comparing our method with the performance that is to be expected with an "experience-based" mapping framework, which creates and maintains separate maps for each map session. In addition to that, comparisons of the localization performance with selecting landmarks randomly, and with the localization performance using all landmarks, serve as lower and upper bounds for properly assessing the effectiveness of our proposed landmark selection on the one hand, and the extent of saving mobile network bandwidth on the other hand. We further assess and compare the selection performance with various ranking schemes inspired by the *tf-idf* concept in information retrieval.

To keep the evaluation section as concise as possible, we prefer to present metrics aggregated over all data sets of the respective data set collection. However, the interested reader is kindly invited to study the graphs showing the performance on each data set separately in Section A.1.

5.1 | Experiments overview

Our experiments can be divided into four groups as follows.

5.1.1 | Rich sessions only

We first investigate the effectiveness of the proposed appearance-based landmark selection and the resulting localization precision with maps containing only *rich sessions*. This allows us to restrict the landmark selection to select from at most one *rich session* at any localization iteration along the trajectory with the ranking function f_{MRS}, as described in Section 5.3. It corresponds to the localization performance attainable with mapping frameworks that keep separate maps for every session, such as the Experience-Based mapping framework by Churchill and Newman (2013). In reverse, it shows the benefit in localization precision achievable in a multisession mapping framework as the one used for this study, which expresses all landmarks from all sessions in a common reference coordinate frame and thus allows selecting landmarks from more than one *rich session* at the same time. The respective experiments can be found in Section 5.5 (Figures 4 and 5).

5.1.2 | Observation sessions

With the presence of *observation sessions*, the selection performance of different appearance-based landmark ranking functions exhibit more pronounced variance. Therefore, the experiments in this section aim at analyzing these differences in performance and relate them to the varying environmental conditions. Note that since *observation sessions* span across multiple *rich sessions*, the ranking function f_{MRS} is no longer properly defined and is thus not included in these experiments. The experiments can be found in Section 5.6 (Figures 6–8).

5.1.3 | Localization accuracy

The *NCLT* data set collection provides ground-truth pose estimates. This allows us to evaluate the localization accuracy along the trajectories of all *NCLT* data sets. Apart from yielding an absolute estimate of the localization accuracy achieved by the different selection policies and landmark ranking functions, we can further investigate and validate the relation between the localization accuracy and other performance influencing metrics such as the distance from the map trajectories, or the number of observed landmarks. The respective experiments can be found in Section 5.7, and Table 1. Furthermore, two special phenomena are analyzed in detail in two case studies in Figures 9–11.

5.1.4 | Computational performance analysis

The potential to significantly reduce the computational requirements on the vehicle side constitutes—apart from a reduction in mobile network bandwidth consumption—a second strong incentive to employ the proposed appearance-based landmark selection. To support this claim, we have measured and analyzed the computational costs involved for the different components of our visual localization pipeline, both with, and without appearance-based landmark selection. The respective results are presented in Section 5.8 and Figure 12.

5.2 | Data set collections

The selection of data sets for evaluating the performance of the proposed appearance-based landmark selection has been driven by the following main criteria: (a) The data set collection ought to cover a wide range of varying appearance conditions, with still sufficient appearance overlap allowing to build a multisession map. (b) The sensor setup must include an odometry sensor, which we require for the forward propagation of the pose states in our iterative localization pipeline. (c) Ideally, the data set collection offers ground-truth poses, which enable an evaluation of the localization accuracy. Many popular publicly available data set collections fail to meet these criteria. With the NCLT data sets, however, there exists a data set collections offering all features relevant for us. Furthermore, the appearance conditions covered by the NCLT data sets are diverse and very challenging, with changing weather conditions, often a setting sun, or strong shadows in the field of view. They thus provide an ideal settings for putting the different appearance-based landmark ranking functions through their paces.

We extend the evaluation with two self-collected data sets, named *parking-lot* and *city environment*. Similar to the *NCLT* data sets, the *parking-lot* data sets cover long-term appearance change during daytime. The respective sensor setup and platform dynamics differ, however, which adds further variation to the evaluation scenarios. In contrast to the *NCLT* and *parking-lot* data sets, the *city environment* data sets cover a very specific scenario of appearance change, namely that of the change from daytime to nighttime.

5.2.1 | NCLT

In *The University of Michigan North Campus long-term vision and LIDAR data sets* (Carlevaris-Bianco, Ushani, & Eustice, 2016), a *Ladybug 3* camera is used, together with wheel odometry from the Segway platform. All images are undistorted and downscaled to dimensions of 808 px × 616 px to be comparable in resolution to the images recorded in the *parking-lot* and *city environment* collection respectively. The 27 data sets from the *NCLT* collection were recorded between January 2012 and April 2013 on the north campus of the Michigan University in Ann-Arbor. The route and direction of traversal followed during the individual recordings, however, varies considerably between the different data sets. For the purpose of this evaluation, we have extracted an approximately 750 m long segment of the routes that has been traversed in all data sets, except the one recorded on January 10, 2013. Furthermore, the data set from December 1, 2012 has been excluded from the evaluation since it comprises the only nighttime recording. Due to a lack of any recordings from transitioning conditions at dusk or dawn, it is not possible to extend the appearance coverage of the map to an extent that would allow proper localization at nighttime. The traversing direction of all the remaining data sets is the same, except for the recordings from February 4, 2012, November 4, 2012, and February 23, 2013 which traverse the mapped area in opposite direction. These data sets can be successfully localized, even though the respective precision and accuracy are worse compared to the other data sets.

5.2.2 | Parking-lot

The *parking-lot* data sets cover a circular traversals of a car on a open space parking-lot. A total of 28 data sets recorded between August 2013 and July 2014 cover a vast variety of different weather and seasonal conditions during daytime. Among others, they include low-standing sun, rain and wet snow, as well as scattered clouds and clear skies.

5.2.3 | City environment

To cover the extreme change in appearance from daytime to nighttime, 23 drivings in a *city environment* have been recorded during the course of a day, starting around noon, and ending around 6 p.m. in the evening. While the weather condition across these data sets remains static, illumination undergoes drastic change from diffuse daylight to nighttime with artificial street lighting.

The sensor setup used in the *parking-lot* and *city environment* data sets consists of four fish-eye cameras mounted on a car (facing front, left, rear and right), and wheel odometry. Images are recorded at 12.5 Hz in gray scale at a resolution of 640 px × 400 px.

An overview over the weather conditions, the usage of each data set in the corresponding multisession maps, as well as example images for all three data set collections can be found in Tables A1–A3. More sample images of the *parking-lot* data sets can be found in Mühlfellner et al. (2016) and in Carlevaris-Bianco et al. (2016) for the *NCLT* data sets.

5.3 | Ranking functions

Before presenting the metrics and experimental results, we introduce additional ranking functions used for comparison and as baselines in the evaluation.

We employ localization with the following pseudoranking function and selection fraction $\alpha = 1.0$ as a baseline to evaluate the performance of our proposed appearance-based ranking functions:

$$f_0(l) := 1 \ \forall \ l \in C, \quad \alpha = 1.0. \tag{15}$$

This corresponds to using all landmarks in the candidate set C for localization, and in general serves as an upper bound for the performance of any other ranking function with $\alpha < 1.0$.

As a lower bound for the performance of landmark selection, we further compare against selecting landmarks randomly:

$$f_{\text{random}}(l) := v, \quad v \sim \mathcal{U}[0, 1]. \tag{16}$$

In addition to that, we also compare the performance of the appearance-based ranking functions introduced in Section 4.2 to the performance of the following ranking function:

$$f_{MRS}(l_i) := \begin{cases} 1 & \text{if } \ p([l_i]|\mathcal{A}) = \max\limits_{[l]}(p([l]|\mathcal{A})), \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

This ranking function selects at most $n = \alpha|C|$ landmarks observed from the *rich session* with currently the best conformity with the encountered appearance condition. While switching the selection of landmarks from one *rich session* to another is allowed along the traversal, selecting landmarks from more than one *rich session* for a specific localization iteration is prohibited. It thus demonstrates the localization performance attainable with separate maps from each *rich session*, in contrast to having all landmarks and observer vertices expressed in one common coordinate frame of reference.

We further include the following appearance-based ranking functions introduced in Bürki et al. (2016) in our comparison:

$$f_{NCV}(l) := \frac{1}{|Z_l|} \sum_{z \in Z_l} |\mathcal{O}_{k-1}^{z}|, \tag{18}$$

with

$$\mathcal{O}_{k-1}^{z} := \{l \in \mathcal{O}_{k-1} | z \in Z_l\}. \tag{19}$$

It corresponds to a normalized voting-based ranking. Every landmark observed in the previous localization iteration casts a vote for each of its observing sessions. To prevent landmarks observed from multiple map sessions to always dominate over landmarks observed from fewer or only one map session, the accumulated votes are normalized by the number of map observer sessions.

In addition to that, we compare our proposed appearance-based ranking functions with different variations of the *tf-idf* ranking scheme used in information retrieval.

$v_l := [x_i] \in \mathbb{R}^{\lvert Z \rvert} \text{ with } x_i = \begin{cases} 1, & \text{if } z_i \in Z_l \\ 0, & \text{otherwise} \end{cases}$ $v_q := \sum_{l \in \mathcal{O}_{k-1}} v_l$ $f_{AV}(l) := cosine(v_l, v_q)\ \forall l \in C$	
$tf(l, z)$	$\begin{cases} 1, & \text{if } z \in Z_l \\ 0, & \text{otherwise} \end{cases}$
$idf()$	1.0

The ranking function f_{AV} uses a vector space representation for landmarks with a binary *tf* term representing the observing map session relation. A *cosine* similarity metric is used as the ranking score.

$f_{TfIdfA}(l) := \sum_{l_o \in \mathcal{O}_{k-1}} \sum_{z \in Z_{l_o}} tf(z, l) idf(z, S_{k-1})$	
$tf(l, z)$	$\begin{cases} 1, & \text{if } z \in Z_l \\ 0, & \text{otherwise} \end{cases}$
$idf(z, C)$	$log(\frac{\lvert C \rvert}{\lvert C^z \rvert})$, with $C^z := \{l \in C \mid z \in Z_l\}$

Ranking function f_{TfIdfA} follows an analogy with text document retrieval where landmarks are interpreted as documents containing words in the form of observing map sessions. The multiset of query words is built from all observing map sessions from the set of recently observed landmarks. Further, a standard *inverse document frequency* term is used, downweighting the contribution of map sessions frequently present in the observing map sessions of the candidate landmarks.

$r(z) := \sum_{l \in \mathcal{O}_{k-1}} tf(l, z) idf(l, Z)$ $f_{TfIdfB}(l) := \sum_{z \in Z_l} r(z)$	
$tf(l, z)$	$\begin{cases} 1 & \text{if } z \in Z_l \\ 0, & \text{otherwise} \end{cases}$
$idf(l, Z)$	$log(\frac{\lvert Z \rvert}{\lvert Z_l \rvert})$

In contrast to f_{TfIdfA}, the ranking function f_{TfIdfB} first attempts to rank map sessions, instead of directly ranking landmarks. For this, roles are switched, and map sessions are interpreted as documents, containing words in the form of landmarks observed in the respective session. The set of recently observed landmarks $\mathcal{O}_{k-1}$ forms the set of query words, upon which the map sessions are ranked, following a standard *tf-idf* scheme. The ranking score for a candidate landmark is eventually formed as the sum of the respective observing session relevances.

$r(z) := \sum_{l \in \mathcal{O}_{k-1}} tf(l, z) idf(z, S_{k-1})$ $f_{WRS}(l) := \sum_{z \in Z_l} r(z)$	
$tf(l, z)$	$\begin{cases} 1 & \text{if } z \in Z_l \\ 0, & \text{otherwise} \end{cases}$
$idf(z, S_{k-1})$	$\frac{1}{\lvert S^z_{k-1} \rvert}$ with $S^z_{k-1} := \{l \in S_{k-1} \mid z \in Z_l\}$

Ranking function f_{WRS} is defined similar to f_{AEC}, evaluates, however, the relevances of individual map sessions $r(z)$, instead of appearance equivalence classes. Analogous to f_{TfIdfB}, a sum over all observing session relevances is used as the final ranking score of a candidate landmark.

5.4 | Metrics

An informative measure for the quality of the ranking function is the comparison between the number of observed landmarks using only some percentage of selected landmarks, and the number of observed landmarks using all landmarks for localization at a given iteration k. This ratio is denoted r_k^{obs} and formally defined as follows:

$$r_k^{\text{obs}} := \frac{\lvert \mathcal{O}_k^{f,\alpha} \rvert}{\lvert \mathcal{O}_k^{f_0,\alpha=1.0} \rvert}. \tag{20}$$

An ideal ranking function f achieves an observation ratio r_k^{obs} close to 1.0 with a selection fraction α as low as possible. This would indicate that only landmarks currently observable receive a high score and are selected, whereas unobservable landmarks receive a low score and are discarded.

The *NCLT* data set collection provides ground-truth poses based on fused and globally optimized pose estimates computed from the 3D LiDAR scans and the differential GPS sensor measurements. We make use of this to evaluate the accuracy of the localized poses within a local neighborhood of the map (Burgard et al., 2009). For every localization iteration along a traversal of a data set, we compare the transformation between the pose resulting from solving our visual localization optimization problem, $\bar{\mathcal{T}}_{W\,B_k}$, and the pose of the nearest vertex in the map, $\mathcal{T}_{W\,nnV_k}$, with the transformation between the ground-truth pose for the current image, $\mathcal{T}_{GB_k}$, and the ground-truth pose of the same nearest vertex in the map, $\mathcal{T}_{GnnV_k}$. This results in the following formulation for the local-error transformation:

$$\mathcal{T}^{W}_{nnVB_k} = \mathcal{T}^{-1}_{W\,nnV_k} \bar{\mathcal{T}}_{W\,B_k}, \tag{21}$$

$$\mathcal{T}^{G}_{nnVB_k} = \mathcal{T}^{-1}_{GnnV_k}\mathcal{T}_{GB_k}, \tag{22}$$

$$\mathcal{T}_{LEGT_k} := \mathcal{T}^{W-1}_{nnVBk}\mathcal{T}^{G}_{nnVBk}. \tag{23}$$

All involved transformations are schematically depicted in Figure 3.

Apart from the inaccuracy of the visual localization, there are further sources of errors affecting $\mathcal{T}_{LEGT}$, such as (a) inherent inaccuracies of the ground-truth transformations, (b) time synchronization, (c) sensor intrinsics and extrinsics calibration, (d) scale and space distortions between the two involved coordinate systems $\mathcal{F}_W$ and $\mathcal{F}_G$, and (e) inconsistencies in the pose graph of the visual map. The effect of the distortion between the involved coordinate systems is almost entirely mitigated by employing local errors as described above. To eliminate any errors due to inconsistencies in the pose graph of the visual map, we optimize the poses of the *NCLT* maps with an additional prior constraint linked to the ground-truth transformation closest in time. The inherent inaccuracies of the ground-truth solution are expected to be considerably lower than the localization accuracies from the visual localization system, as the former is computed from a globally optimized SLAM solution using the 3D LiDAR scans and differential GPS, with all data sets cross-registered, and a manual removal of wrong loop-closure constraints (Carlevaris-Bianco et al., 2016). As a consequence, we expect $\mathcal{T}_{LEGT}$ to reflect primarily the (in-)accuracy of the visual localization.

The local-error transformation $\mathcal{T}_{LEGT_k}$ is further decomposed into the corresponding three-dimensional translation and rotation vector, denoted by p_{LEGT_k}, and a_{LEGT_k} respectively.

In the cases of the *parking-lot* and *city environment* data sets, no ground-truth solution is available. Since in each localization iteration, a visual-only pose optimization problem is solved (see Section 3.2) we can still assess how well the resulting pose estimate is constrained along a data set by computing statistics on the transformations between the pose estimates, and the respective pose guess of the same iteration:

$$\mathcal{T}_{LEO_k} := \hat{\mathcal{T}}^{-1}_{W\ B_k}\bar{\mathcal{T}}_{W\ B_k}. \tag{24}$$

We refer to this as localization precision, as opposed to localization accuracy as described in Equation (23). In addition to the error induced by the visual localization itself on the current and previous pose estimate, the local-error transformation $\mathcal{T}_{LEO_k}$ also contains the local drift of the wheel odometry in between. However, the latter is expected to be at least one magnitude smaller, leaving the magnitude of $\mathcal{T}_{LEO_k}$ to be dominated by the visual localization errors.

5.5 | Rich sessions only

We first present the ratios of observed and selected landmarks, as well as the precision results, for all three data set collections, whereas the presented values are aggregated over all data sets of the respective collection.

In Figure 4, the relation of observed versus selected landmarks is shown for selection fractions between 10% and 40%. Since there is a significant discrepancy in the observation percentage during daytime as opposed to at night, we further show the observation percentage in the *city environment* aggregated over daytime data sets, that is, up and including 17:30, and over the remaining nighttime data sets, separately. In addition to that, Figure 5 shows a comparison of the localization precision.

Note that in this scenario with a map containing only *rich sessions*, it is straightforward to see that the ranking score with the appearance-based ranking functions f_{AEC}, f_{WRS}, f_{AV}, and f_{TfIdfB} is identical. We therefore only show the results for the ranking function f_{AEC}. In all three environments, ranking landmarks with to f_{AEC} yields a consistently high observation percentage, and localization precision close to the one achieved using all landmarks. In contrast to that, ranking landmarks using f_{TfIdfA} fails, as it yields a consistently lower observation percentage than random selection, and precision values considerably worse than the other ranking

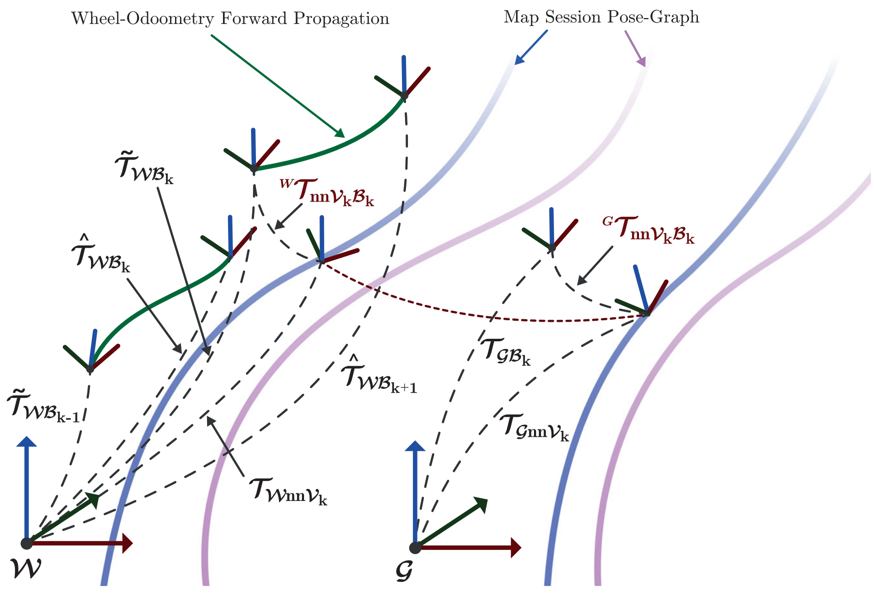

Wheel-Odoometry Forward Propagation
Map Session Pose-Graph
$\tilde{\mathcal{T}}_{\mathcal{W}\mathcal{B}_k}$
$^{W}\mathcal{T}_{nn\mathcal{V}_k\mathcal{B}_k}$
$\hat{\mathcal{T}}_{\mathcal{W}\mathcal{B}_k}$
$^{G}\mathcal{T}_{nn\mathcal{V}_k\mathcal{B}_k}$
$\mathcal{T}_{\mathcal{G}\mathcal{B}_k}$
$\tilde{\mathcal{T}}_{\mathcal{W}\mathcal{B}_{k-1}}$
$\hat{\mathcal{T}}_{\mathcal{W}\mathcal{B}_{k+1}}$
$\mathcal{T}_{\mathcal{G}nn\mathcal{V}_k}$
$\mathcal{T}_{\mathcal{W}nn\mathcal{V}_k}$
$\mathcal{W}$
$\mathcal{G}$

FIGURE 3 The two coordinate systems $\mathcal{F}_W$ and $\mathcal{F}_G$ and all relevant transformations used for the calculation of the local localization precision with respect to the wheel odometry, and the local localization accuracy with respect to the ground-truth solution [Color figure can be viewed at wileyonlinelibrary.com]

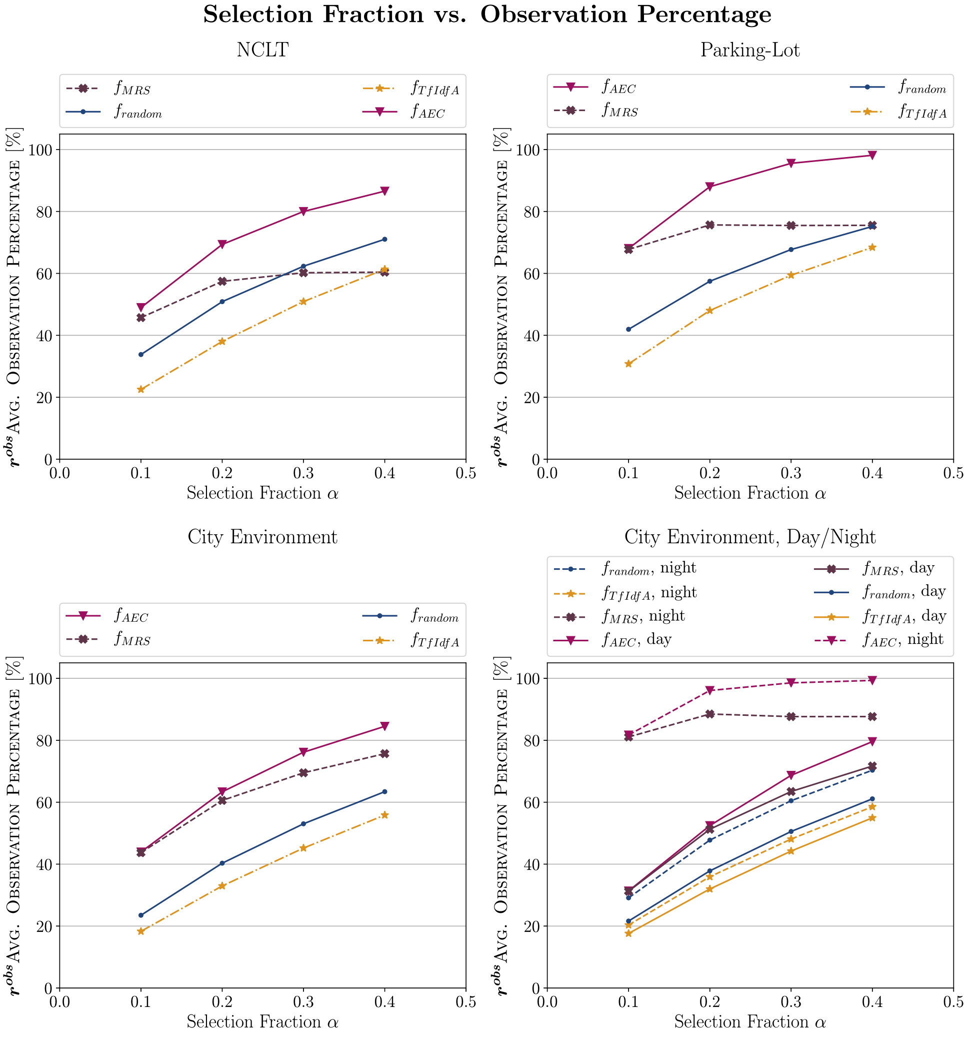


FIGURE 4 The average observation percentage r^{obs} in relation to the selection fraction α for different choices of ranking functions, and for all three data set collections against maps containing only *rich sessions*. In the *city environment*, data sets are further split up into daytime data sets (up until 17:30), and nighttime data sets [Color figure can be viewed at wileyonlinelibrary.com]

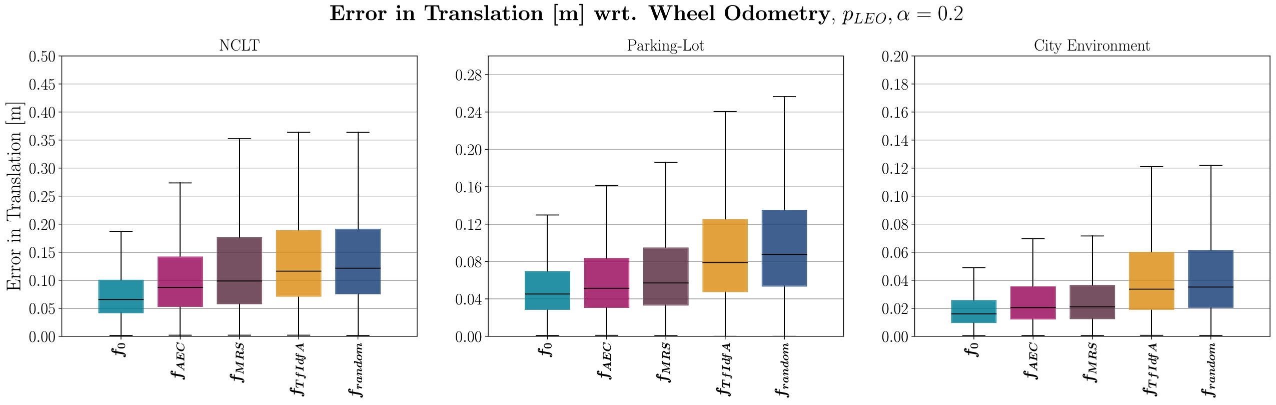


FIGURE 5 The aggregated localization translation precision for all three data set collection against the map containing only *rich sessions*. The following ranking functions are shown: localization using all landmarks, f_0, $\alpha = 1.0$, appearance-based landmark selection with f_{AEC}, f_{TfIdfA}, f_{MRS}, and random selection with f_{random}, all with a selection fraction of $\alpha = 0.2$ [Color figure can be viewed at wileyonlinelibrary.com]

functions. The *idf* term of f_{TfIdfA} follows the text-book definition of *inverse document frequency*, thus downweighting the influence of map sessions if there are many candidate landmarks observed in the respective session. As described in Section 4.3, this criteria does not well reflect the appearance conformity of a landmark, and instead tends to favor map sessions with only few landmarks.

We further note the performance limitations of f_{MRS}. With low selection fractions, the attained observation percentage is on the same level as other well-performing ranking functions, such as f_{AEC}, f_{AV}, and f_{TfIdfB}. However, the restriction to only select from one *rich session* results in performance saturation for larger selection fractions. The respective loss in precision is clearly visible in case on the *NCLT* and *parking-lot* data sets at a selection fraction of 20%, and demonstrates one of the benefits of having all landmarks, even from multiple *rich sessions*, registered in one common coordinate frame of reference. Note that this loss in precision is less pronounced on the *city environment* data sets, as in this case, there are only few different appearance conditions represented in the map, with a clear best-matching *rich session* at any time.

It can further be observed that the overall observation percentage and localization precision in the *NCLT* environment is lower as compared to the *parking-lot*, despite both environments reflecting long-term daytime conditions. This discrepancy suggests that there is a larger difference in encountered appearances over the year in relation to the number of *rich sessions* in the map in the *NCLT* scenario as compared to the *parking-lot* scenario. Precision in the *NCLT* environment further pays tribute to the fact that the trajectories in the *NCLT* data sets often do not follow the exact same route and exhibit lateral offsets of up to 12 m. This renders the visual localization considerably more challenging as opposed to the *parking-lot* scenario, where there is a quite precisely repeated driving pattern on the parking-lot.

In the *city environment*, the comparatively low observation percentage is attributed to the fact that there are fewer diverse appearance conditions covered, resulting in a lower number of *rich sessions* present in the map. During daytime, where there are considerably more landmarks than at dusk and nighttime, even a selection fraction of up to 40% may not be sufficient to select all landmarks matching the current appearance condition. This effect is supported by the graph in Figure 4 showing the observation percentage aggregated separately over daytime and nighttime data sets. All ranking functions, including random selection, exhibit a significantly higher observation percentage at night as opposed to during daytime. The increase at night is, however, most pronounced for the appearance-based ranking functions f_{AEC} and f_{MRS}.

5.6 | Observation sessions

Adding *observation sessions* to the map can further improve the performance of the appearance-based landmark selection at a negligible additional map storage or computational burden.

It can be seen in Figures 6 and 7 that for low selection fractions, the best observation percentage and localization precision is achieved using the proposed f_{AEC} ranking function. In addition to that, ranking functions f_{AV} and f_{TfIdfB} also achieve favorable performance for low selection fractions, and even achieve the best observation percentage on the *NCLT* and *parking-lot* data sets for higher selection fractions. We attribute this phenomena to saturation effects with higher selection fractions on the one hand, and to data set specific artifacts, such as the *lock-in* effect discussed in Section 5.7 on the other hand. Both may undermine the theoretical optimality of f_{AEC} under ideal conditions.

In addition to that, the ranking function f_{NCV} exhibits the highest variance in performance. As can be seen in Figures 6, and 8, this ranking function fails during daytime, yielding an observation percentage worse than that of random selection. It further performs poorly in general for low selection fractions on the other two data set collections too. However, for high selection fractions, the opposite is the case, and appearance-based landmark selection with f_{NCV} achieves the best performance on the *NCLT* and *parking-lot* data sets.

We further observe ranking functions f_{AV} and f_{TfIdfB} perform very similarly in all three environments. This is remarkable, as the respective expressions of the ranking functions are considerably different.

Similar as with maps containing only *rich sessions*, the ranking function f_{WRS} outperforms random selection, but falls short of any of the before mentioned appearance-based ranking functions. The presence of *observation sessions* further is not able to improve the poor performance of f_{TfIdfA}.

The benefit of ranking based on appearance equivalence classes is most pronounced in the *city environment* at dusk around 17:25, as can be seen in Figure 8. Despite this being only one data set, it is exemplary for a general phenomena: The heavy bias in the number of landmarks towards daytime *rich sessions* lets most other ranking functions preferably select landmarks from daytime. However, nighttime landmarks would, although fewer in absolute numbers, already yield more inlier observations relative to the number of selected nighttime landmarks. Only the two ranking functions f_{AEC} and f_{WRS} are able to exploit this and achieve almost 20% more landmark observations in this case. The ranking function f_{WRS}, however, suffers from suboptimal performance during daytime, leaving the ranking function based on appearance equivalence classes as the only one with high observation percentage all the time.

Before elaborating on the localization accuracy evaluation in the subsequent section, we summarize the key findings of the different ranking function's performance. On all three data set collections, a significant boost in observation percentage by the use of *observation sessions* is well visible. In addition to that, the ranking function f_{AEC} exhibits the best performance at low selection fractions, while the performance of f_{NCV} is most susceptible to the selection fraction, performing poorly for low selection fractions, but even outperforming all other ranking functions by a small margin at a selection fraction of 40%.

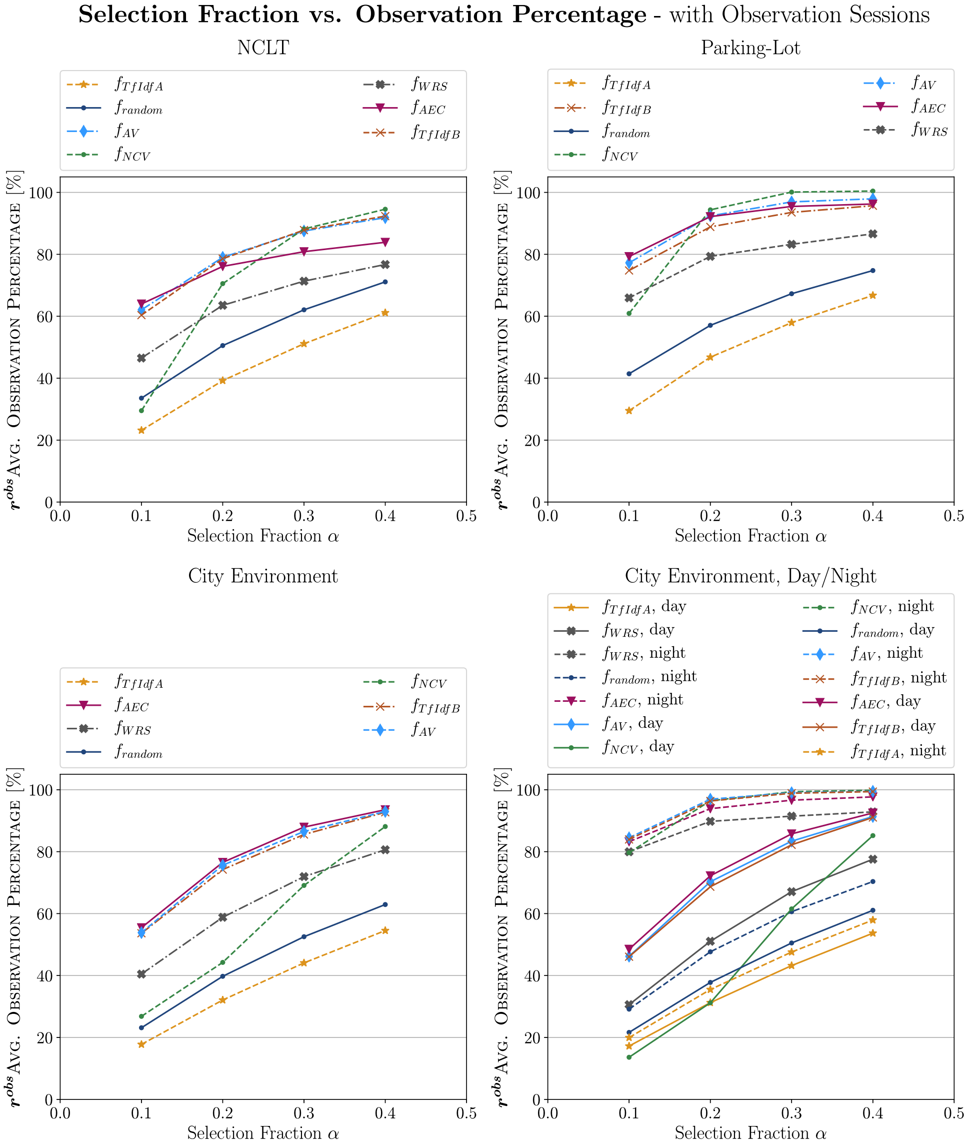


FIGURE 6 The average observation percentage r^{obs} in relation to the selection fraction α for different choices of ranking functions, and for all three data set collections against maps containing *observation sessions*. In the *city environment*, data sets are further split up into daytime data sets (up until 17:30), and nighttime data sets [Color figure can be viewed at wileyonlinelibrary.com]

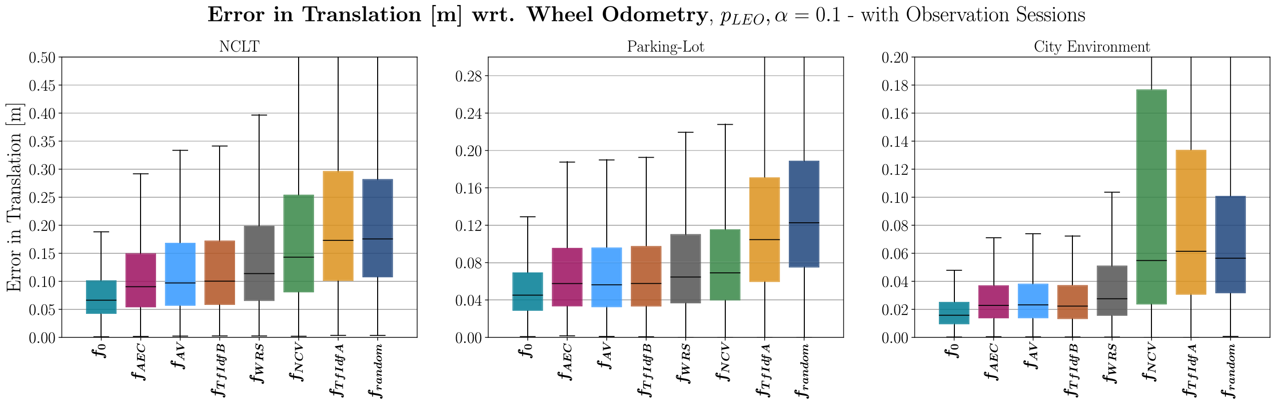


FIGURE 7 The aggregated localization translation precision for all three data set collection against the map containing *observation sessions*. The following ranking functions are shown: localization using all landmarks, $f_0, \alpha = 1.0$, appearance-based landmark selection with $f_{AEC}, f_{AV}, f_{TfIdfB}, f_{NCV}, f_{TfIdfA}$, and random selection with f_{random}, all with a selection fraction of $\alpha = 0.1$ [Color figure can be viewed at wileyonlinelibrary.com]

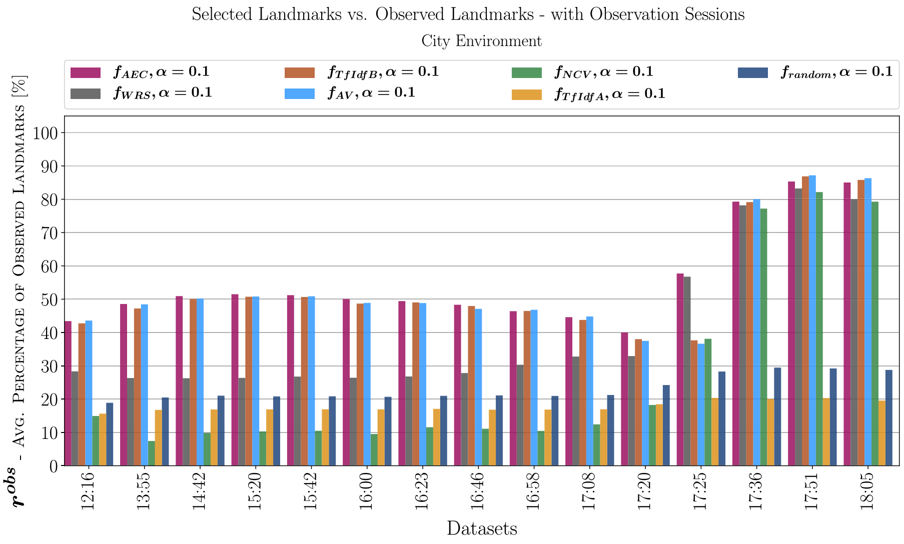


FIGURE 8 The average observation percentage r^{obs} for different choices of ranking functions and a selection fraction $\alpha = 0.1$, for all data sets of the *city environment* against the map containing *observation sessions* [Color figure can be viewed at wileyonlinelibrary.com]

5.7 | Localization accuracy

In this section, we present the visual localization accuracy results of the *NCLT* data sets using the ground-truth solution based on 3D LiDAR and differential GPS as a reference.

The median errors in translation of $\mathcal{T}_{LEGT}$, denoted by p_{LEGT}, are listed in Table 1 for localization using all landmarks, f_0, as well as appearance-based selection with the ranking functions $f_{AEC}, f_{NCV}, f_{MRS}$, and random selection, f_{random}; all with a selection fraction of 10%. Furthermore, the translation accuracy using the ranking function f_{AEC} is listed both when localizing against the map containing only *rich sessions*, as well as when localizing against the map containing both *rich sessions* and *observation sessions*. For the latter, the ranking function is denoted by "f_{AEC} os."

TABLE 1 The translation localization accuracy for the *NCLT* data sets, using the ground-truth poses as a reference

	Median p_{LEGT} [m], $\alpha = 0.1$					
Date	**$f_0(\alpha = 1.0)$**	**f_{AEC}**	**f_{AEC}, os.**	**f_{NCV}**	**f_{MRS}**	**f_{random}**
January 8, 2012	0.155	0.2	**0.181**	0.2	0.206	0.247
January 15, 2012	0.215	**0.317**		0.35	0.363	0.329
February 2, 2012	0.14	0.196	**0.145**	0.222	0.203	0.187
February 12, 2012	0.179	0.263	**0.211**	0.268	0.266	0.3
February 18, 2012	0.18	0.247	**0.182**	0.28	0.287	0.268
February 19, 2012	0.129	**0.191**		0.215	0.211	**0.191**
March 17, 2012	0.129	0.196	**0.155**	0.235	0.204	0.212
March 25, 2012	0.265	0.304	**0.271**	0.301	0.294	0.329
March 31, 2012	0.112	0.189		0.197	0.196	**0.163**
April 29, 2012	0.141	0.196	**0.161**	0.202	0.203	0.216
May 26, 2012	0.121	0.151	**0.135**	0.154	0.161	0.181
August 4, 2012	0.139	0.159	**0.157**	0.16	0.158	0.233
September 28, 2012	0.124	0.18	**0.149**	0.259	0.194	0.202
October 28, 2012	0.138	**0.201**		0.251	0.236	0.219
November 4, 2012	0.308	0.396	**0.363**	0.458	2.282	0.41
November 17, 2012	0.174	0.215	**0.185**	0.219	0.24	0.253
February 23, 2013	0.445	0.485		0.544	0.524	**0.477**
April 5, 2013	0.168	0.225	**0.188**	0.218	0.237	0.253

Note. The columns show the translation median error in meters for the following six ranking functions: localization using all landmark (f_0, $\alpha = 1.0$), appearance-based landmark selection with f_{AEC}, f_{NCV}, and f_{MRS}, and random selection with f_{random}. For appearance-based and random selection, a selection fraction $\alpha = 0.1$ is used. All ranking function localize against the map containing only *rich sessions*, except for f_{AEC}, *os.* which localizes against the map containing both *rich sessions* and *observation sessions*.

The bold values mark the minimum in each row, which corresponds to the ranking function achieving the lowest median translation error.

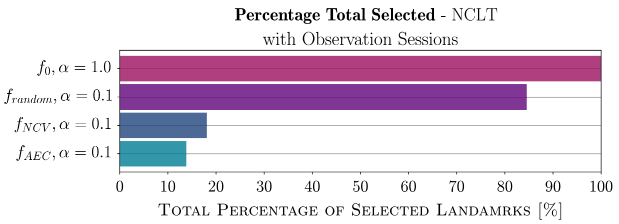


FIGURE 9 The total percentage of unique landmarks selected over the coarse of the entire trajectory of a data set. This percentage directly conforms to the overall savings in data transmissions between a map backend and a mobile robot in a shared-map scenario. While appearance-based selection only uses a percentage of landmarks approximately equivalent to the respective selection fraction in each iteration, random selection makes use of almost all landmarks at least once along the trajectory [Color figure can be viewed at wileyonlinelibrary.com]

We first note that the median translation accuracy of the reference localization using all landmarks exhibits a rather large span, ranging from 11 up to 44 cm. This is on the one hand due to the varying trajectories of the respective data sets. On the other hand, not every appearance condition encountered in the data sets used for evaluation is equally well covered by the *rich sessions* in the map, resulting in differing visual localization performance. The most important factor for deteriorated localization performance, however, is the direction of traversal, resulting in the data sets from November 4, 2012 and February 23, 2013 to perform considerably worse than any other data set.

We further observe that the accuracy using the appearance-based ranking function f_{AEC} on the map containing only *rich sessions* slightly outperforms selecting landmarks based on f_{NCV}, but both perform significantly better than using f_{MRS} for landmark selection. This again demonstrates the gain in performance due to the ability to select landmarks from more than one *rich session* at a time.

In addition to that, there is a clearly pronounced boost in localization accuracy when using the map with additional *observation sessions* for localization, with landmark selection based on f_{AEC} even achieving accuracy values close to those of the reference localization using all landmarks for certain data sets.

It is noticeable, however, that random selection often achieves accuracy values close to those of appearance-based selection, at least in the case of using the map with no *observation sessions*. In this regard, we notice that the random selection of landmarks occurs for every localization iteration along the trajectory. Despite only selecting 10% at each iteration, even after short traversals of a few meters, almost all landmarks available in the vicinity of the respective map segment have been selected at least once by f_{random}. This effect is well visible in Figure 9 which displays the total number of selected landmarks for each of the different ranking functions and selection policies. While all appearance-based ranking functions only select a fraction of all landmarks across the entire data set trajectory approximately equal to the selection fraction at each iteration of 10%, random selection selects up 85% of all landmarks in the map. For this reason, localization using random selection may be considerably less precise, but its accuracy is not in the same extent worse compared to both appearance-based landmark selection and localization using all landmarks.

In addition to that, the challenging route selection of the *NCLT* data sets lead to varying localization performance along a trajectory of a specific data set. Even though the influence of outliers in the localization performance on the overall accuracy is limited by our choice of the median error, the magnitude of the latter is often still heavily influenced by short segments of the trajectory with very poor localization performance. To render these effects more tangible, we investigate the localization accuracy in relation to the number of observed landmarks and the distance to the nearest vertex in the map in detail for the two data sets of January 8, 2012 and February 2, 2012.

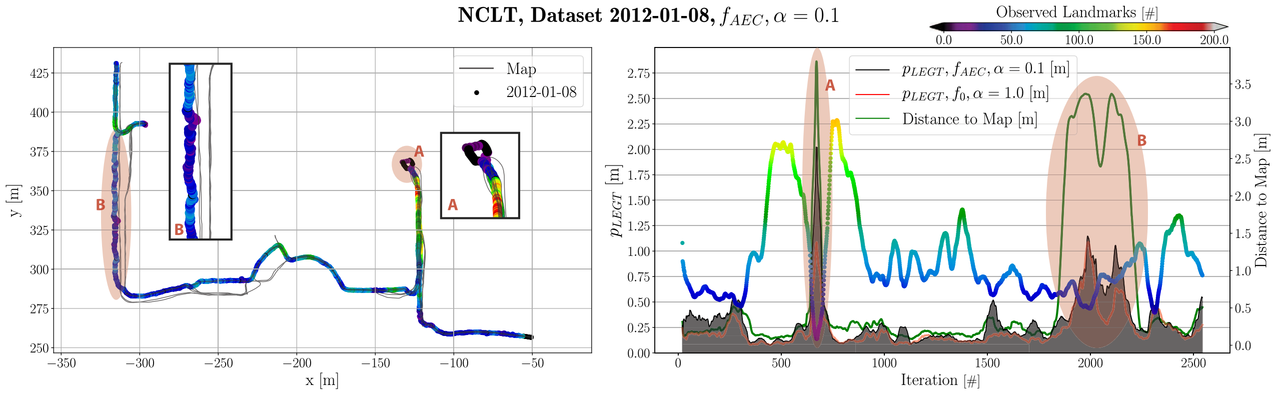


FIGURE 10 Bird's eye perspective onto the mapped segment of the *NCLT* scenario, together with a temporal analysis of the number of observed landmarks, the translation localization accuracy, and the distance to the nearest vertex in the map, referred to as "distance to map." On the left-hand side, the trajectories of the *richsessions* in the map are drawn in light gray, while the localized poses of the data set from January 8, 2012 is drawn in color indicating the number of observed landmarks along the route. Two particular situations along the trajectory of this data set are marked by capital letters "A" and "B" and analyzed in detail in Section 5.7 [Color figure can be viewed at wileyonlinelibrary.com]

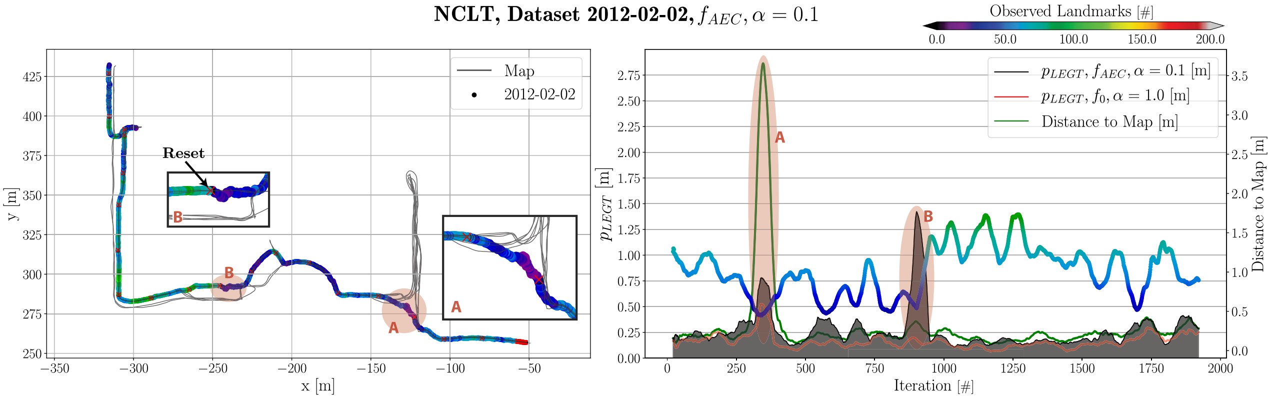


FIGURE 11 Bird's eye perspective onto the mapped segment of the *NCLT* scenario, together with a temporal analysis of the number of observed landmarks, the translation localization accuracy, and the distance to the nearest vertex in the map, referred to as "distance to map." On the left-hand side, the trajectories of the *richsessions* in the map are drawn in light gray, while the localized poses of the data set from February 2, 2012 is drawn in color indicating the number of observed landmarks along the route. Two particular situations along the trajectory of this data set are marked by capital letters "A" and "B" and analyzed in detail in Section 5.7 [Color figure can be viewed at wileyonlinelibrary.com]

The left-hand side of Figure 10 shows a birds-eye perspective of the mapped segment used in the *NCLT* scenario. The trajectories of the *rich sessions* present in the map are drawn in gray, whereas the trajectory of the data set being localized from January 8, 2012 is drawn in color indicating the number of landmark observations along the trajectory. In contrast to that, the bottom half depicts the relation between time—or iteration index respectively—on the one hand, and the number of observed landmarks, the localization accuracy, and the distance to the nearest vertex in the map on the other hand.

The trajectories from all *rich sessions* in the map follow up and down a long aisle between iteration 300 and 900. While the trajectory from January 8 in general follows the same pattern, the turning point at the back of the aisle occurs a few meters farther into the aisle compared to the trajectories present in the map. This situation is marked with the letter "A" in Figure 10. While the distance to the nearest vertex in the map suddenly increases from approximately 30 cm up to almost 3 m, the number of observed landmarks drops to almost zero. At the same time, the localization accuracy is greatly reduced, both in case of the reference localization with f_0, as well as and considerably more severely in the case of appearance-based localization. In this regard, situation "A" also serves as a good example for the strong correlation between the number of observed landmarks and the localization precision and accuracy respectively.

In situation "B," the number of observed landmarks is only slightly lower than in the preceding trajectory segment. The distance to the map, however, is considerably increased, since the trajectory of the data set travels along the street instead of on the parallel sidewalk as in all the *rich sessions* in the map. This again results in a peak degradation of the localization accuracy and demonstrates the correlation between the distance to the map, and the localization performance.

The data set from February 2 exhibits even slightly lower localization accuracy with the map containing only *rich sessions* for appearance-based landmark selection with f_{AEC} compared to performing random selection with f_{random}. As can be seen in Figure 11, the errors in localization are mainly attributed to two peaks, again marked with a capital letter "A" and "B."

In situation "A," the trajectory from February 2 directly crosses the street, while all the data sets used for building the map take a right turn up and down the aisle. This leads to a sudden increase in the distance to the nearest vertex in the map, and as a result of that a simultaneous drop in the number of observed landmarks and the respective localization accuracy.

In contrast to that, the peak drop in localization accuracy in situation "B" originates from a *lock-in* effect inherent to the presented appearance-based landmark selection. To understand the cause of this peak drop, we point out that the relevance of the different appearance equivalence classes is evaluated based on recently observed landmarks. The latter themselves are a subset of recently selected landmarks. This does not constitute a problem as long as the availability of landmarks from all *rich sessions* in the map along the trajectory is maintained, and the appearance conditions encountered do not show any abrupt change that is not also reflected in the respective *rich* or *observation sessions*. In the *NCLT* scenario, however, the trajectory segment between iteration 600 and 1100 is characterized by the different data sets taking varying routes. Thus, the *rich session* with the best appearance conformity may at once exhibit a large lateral offset, or not be available at all temporarily, resulting in the number of observed landmarks to decrease and the localization accuracy to drop. To recover from this *lock-in* situation, the appearance-based landmark selection can be "reset." For this, all candidate landmark are used for localization of a single iteration, allowing to properly re-evaluate the relevance of all available appearance equivalence classes. For the presented experiments, such a "reset" is set to occur at every 100th iteration, and its effect is clearly visible in situation "B": After the "reset," the number of inliers swiftly increases from approximately 20 up to 100, and the respective localization accuracy recovers. In practice, it is advisable to link the

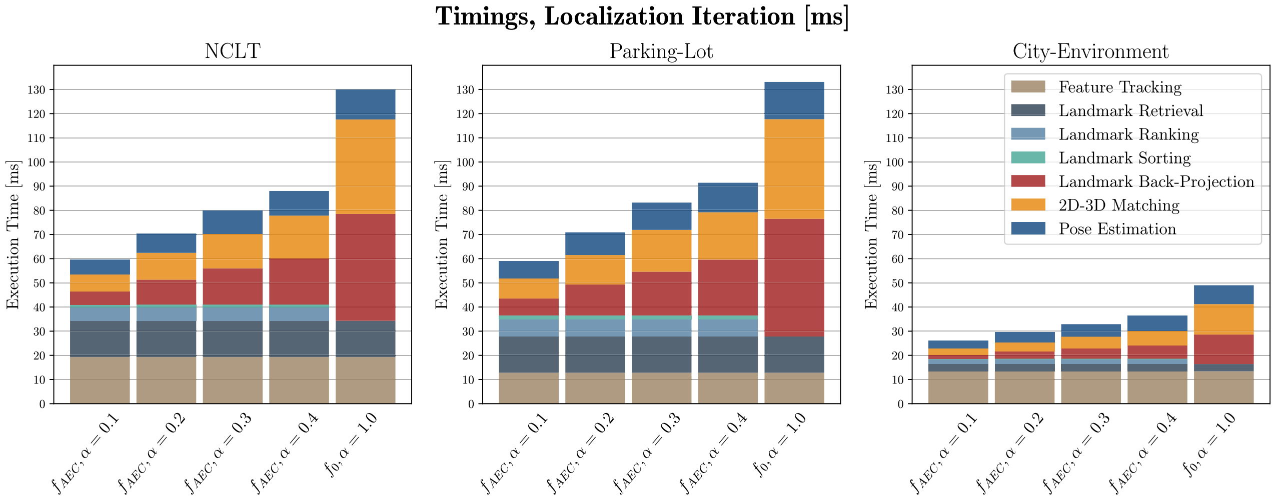


FIGURE 12 The execution times of the individual building blocks of our localization pipeline, including the modules specific to appearance-based landmark selection. While the overall execution times vary in the different scenarios, in all cases localization with appearance-based landmark selection is able to perform significantly faster compared to localization without landmark selection. This is due to the fact that the time invested into landmark ranking and sorting is more than compensated by the time saved in back-projecting and matching of the landmarks in the subsequent modules of the localization pipeline [Color figure can be viewed at wileyonlinelibrary.com]

triggering of "resets" to a metric reflecting the condition of poor localization performance in situations where localization is expected to perform reasonably well (e.g., when the assumed location is close to the mapped trajectories). Such a metric is, however, application and use-case specific.

Apart from the exemplary situations mentioned and discussed in detail above causing the localization accuracy to degrade, Figures 10 and 11 also indicate that under normal circumstances, that is, with the localized trajectory and all map *sessions* following the same route, a localization accuracy of around 10 cm is achieved, which is in accordance with the results found in Mühlfellner et al., (2016).

5.8 | Computational performance analysis

We conclude the evaluation section by analyzing the computational time spent on the major blocks of our localization pipeline. All computations have been carried out on a Lenovo W530 with an Intel i7 CPU, and without the use a graphics processing unit. In addition to the incentive of lower data bandwidth usage, the computational performance analysis reveals a second benefit of appearance-based landmark selection in the form of reduced computational demands on the mobile platform side.

Figure 12 shows the execution times of the following building blocks:

- Feature tracking: The time to extract keypoints and compute FREAK descriptors on all involved cameras.
- Landmark retrieval: The time to retrieve all near-by pose-graph vertices, and their observed landmarks.
- Landmark ranking: The time to apply f on all candidate landmarks.
- Landmark selection: The time to select n top-ranked candidate landmarks, yielding S_k.
- Landmark backprojection: This step involves the look-up of the landmark descriptor for each selected landmark, and the back-projection of the landmark's 3D point into the camera image planes, using the pose guess $\hat{\mathcal{T}}_{W\ B_k}$.
- 2D–3D matching: The formation of associations between the FREAK features in the current camera images, and the back-projected map landmarks.
- Pose estimation: Refinement of the pose estimation employing a nonlinear least-squares optimization problem, yielding $\bar{\mathcal{T}}_{W\ B_k}$.

Apart from the feature tracking, the overall execution times in the *city environment* are by more than a Factor 2 lower. This is due to the fact that there are only four *rich sessions* in the map, with only one from bright daylight. Thus, the resulting number of candidate landmarks being retrieved in each iteration is considerably lower as compared to the *NCLT* data sets.

We further note that the feature tracking, the landmark retrieval, and the pose estimation step all have to be carried out both in case of localization with, as well as without appearance-based landmark selection, and that their running time is mostly independent of the number of selected landmarks. Nevertheless, these blocks are included in Figure 12 to give a comprehensive overview over the running time and real-time capability of the localization pipeline.

The computational differences between localization with and without appearance-based landmark selection can be summarized as follows: In contrast to localization without landmark selection, localization with landmark selection has to invest time in ranking and sorting the candidate landmarks. In exchange, however, considerably fewer landmarks have to be backprojected and matched against the features in the image, reducing the runtime of these modules in relation to the respective selection fraction. As can be seen in Figure 12, in both scenarios, even with a selection fraction of

40%, the total runtime of appearance-based selection is considerably lower than without landmark selection. The computational load on the mobile platform side is even further reduced in a shared-map scenario as motivated in Section 1, where ranking and sorting of candidate landmarks is carried out on the backend side.

From the accumulated running times in the *NCLT* scenario, it can be deduced that localization with appearance-based landmark selection is able to run at 10–15 Hz, while localization without landmark selection may not be able to exceed 8 Hz. Only accounting for the modules running on the mobile platform in case of a shared-map scenario, namely feature-tracking, landmark backprojection, matching, and pose estimation, the resulting difference in runtime performance is increased to 15–30 Hz with landmark selection, as opposed to only 10 Hz without landmark selection.

6 | CONCLUSIONS

In this section, we summarize our key findings and draw conclusions for the use of appearance-based landmark selection in practice.

At first, we note that substantial differences in the camera setup in the *NCLT* data sets, such as the lack of fish-eye distortion, does not have a significant effect on the performance of the appearance-based landmark selection. Similar as with the *city environment* and the *parking-lot* data sets, an appearance-based selection of 20–30% of the available landmarks allows achieving a localization performance similar to using all landmarks.

Furthermore, we have analyzed in detail the performance of several appearance-based landmark ranking function in combination with maps with, and without *observation sessions*. Selecting landmarks using the proposed f_{AEC} ranking function yields the best performance, especially for low selection fractions. However, other formulations for the ranking functions, most notably f_{AV} and f_{TfIdfB}, achieve favorable performance too. This observation, together with the independence with respect to the distribution of landmarks in map sessions, let f_{AEC} be the ranking function of choice in general.

With the *lock-in* effect observed on the data set example depicted in Figure 11, we have analyzed and described a potential pitfall inherent to the use of appearance-based landmark selection. In practice, an application and use-case specific monitoring of the observed localization performance in relation to what performance is to be expected is pivotal to swiftly detect a *lock-in* situation and initiate a “reset” of the appearance-based landmark selection. Easily trackable metrics, such as the number of observed landmarks, and the distance from the nearest vertex in the map, may serve as potent indicators to distinguish *lock-in* situation from poor localization due to too large divergence from the mapped territory. This suggestion is supported by the strong correlations between the aforementioned metrics and the localization performance, as shown in Figures 10 and 11.

The localization accuracy achieved in the *NCLT* scenario is in general in accordance with the respective precision, although the magnitude of the former is slightly higher. This is attributed to the fact that there are more sources of error involved, such as the error of the ground-truth solution itself, and inaccuracies of the intrinsic and extrinsic sensor calibrations. It is in this regard important to note again that the pose estimated in each iteration is computed from solving a nonlinear least-squares optimization problem only containing constraints between the image keypoints and the matched 3D landmarks from the map. In particular, there is no temporal smoothing or sensor fusion, which would prevent immediate degradation of accuracy in many situations where temporarily only few landmark are observed.

In a detailed computational performance analysis, we have shown that our localization pipeline with appearance-based landmark selection is able to run in real time. Furthermore, the use of appearance-based landmark selection significantly lowers the computational demand on the mobile platform, as only a fraction of landmarks have to be processed in each localization iteration.

ACKNOWLEDGMENTS

This project has received funding from the EU H2020 research project “UP-Drive” under grant no. 688652 and from the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 15.0284.

ORCID

Mathias Bürki http://orcid.org/0000-0002-6988-9990

REFERENCES

Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing and Management*, *39*(1), 45–65.

Alahi, A., Ortiz, R., & Vandergheynst, P. (2012). Freak: Fast retina keypoint. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Rhode Island, RI, pp. 510–517.

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. *European Conference on Computer Vision*, Graz, Austria, pp. 404–417.

Burgard, W., Stachniss, C., Grisetti, G., Steder, B., Kümmerle, R., Dornhege, C., … Tardós, J. (2009). A comparison of SLAM algorithms based on a graph of relations. 2009 *IEEE/RSJ International Conference on Intelligent Robots and Systems*, St. Louis, MO, pp. 2089–2095.

Bürki, M., Dymczyk, M., Gilitschenski, I., Cadena, C., Siegwart, R., & Nieto, J. (2018). Map management for efficient long-term visual localization in outdoor environments. 2018 IEEE Intelligent Vehicles Symposium, Daejeon, South Korea, pp. 682–688.

Bürki, M., Gilitschenski, I., Stumm, E., Siegwart, R., & Nieto, J. (2016). Appearance-based landmark selection for efficient long-term visual localization. 2016 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, *IEEE*, Daejeon, South Korea, pp. 4137-4143.

Calonder, M., Lepetit, V., Özuysal, M., Trzcinski, T., Strecha, C., & Fua, P. (2012). BRIEF: Computing a local binary descriptor very fast. *Transactions on Pattern Analysis and Machine Intelligence*, *34*(7), 1281–1298.

Carlevaris-Bianco, N., Ushani, A., & Eustice, R. (2016). University of Michigan North Campus long-term vision and lidar datase. *The International Journal of Robotics Research*, *35*(9), 1023–1035.

Churchill, W., & Newman, P. (2013). Experience-based navigation for long-term localisation. *The International Journal of Robotics Research*, *32*(14), 1645–1661.

Clement, L., Kelly, J., & Barfoot, T. (2017). Robust monocular visual teach and repeat aided by local ground planarity and color-constant imagery. *Journal of Field Robotics*, *34*(1), 74–97.

Cummins, M., & Newman, P. (2011). Appearance-only SLAM at large scale with FAB-MAP 2.0. *The International Journal of Robotics Research*, *30*(9), 1100–1123.

Dayoub, F., Cielniak, G., & Duckett, T. (2011). Long-term experiments with an adaptive spherical view representation for navigation in changing environments. *Robotics and Autonomous Systems*, *59*(5), 285–295.

Dymczyk, M., Lynen, S., Bosse, M., & Siegwart, R. (2015). Keep it brief: Scalable creation of compressed localization maps. 2015 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Hamburg, Germany, pp. 2536–2542.

Fayin, L., & Košecká, J. (2006). Probabilistic location recognition using reduced feature set. *Proceedings of* 2006 *IEEE International Conference on Robotics and Automation*, Orlando, FL, vol. 2006, pp. 3405–3410.

Hochdorfer, S., & Schlegel, C. (2009). Towards a robust visual SLAM approach: Addressing the challenge of life-long operation. 2009 *International Conference on Advanced Robotics*, Munich, Germany, pp. 1–6.

Johns, E., & Yang, G. (2013). Feature co-occurrence maps: Appearance-based localisation throughout the day. 2013 *IEEE International Conference on Robotics and Automation*, St. Louis, MO, pp. 3212–3218.

Johns, E., & Yang, G. (2014). Generative methods for long-term place recognition in dynamic scenes. *International Journal of Computer Vision*, *106*(3), 297–314.

Konolige, K., & Bowman, J. (2009). Towards lifelong visual maps. *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 115–1163.

Lategahn, H., Beck, J., & Stiller, C. (2014). DIRD is an illumination robust descriptor. 2014 *IEEE Intelligent Vehicles Symposium Proceedings*, Ypsilanti, MI, pp. 756–761.

Li, Y., Snavely, N., & Huttenlocher, D. P. (2010). Location recognition using prioritized feature matching. *ECCV'10 Proceedings of the 11th ECCV, Hersonissos*, Greece, pp. 791–804.

Linegar, C., Churchill, W., & Newman, P. (2015). Work smart, not hard: Recalling relevant experiences for vast-scale but time-constrained localisation. 2015 *IEEE International Conference on Robotics and Automation* (*ICRA*), Seattle, WA, pp. 90-97.

Lowe, D. (1999). Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Corfu, Greece, vol 2, pp. 1150–1157.

Mactavish, K., Paton, M., & Barfoot, T. (2017). Visual triage: A bag-of-words experience selector for long-term visual route following. 2017 *IEEE International Conference on Robotics and Automation* (*ICRA*), Singapore, pp. 2065–2072.

Maddern, W., Stewart, A., McManus, C., Upcroft, B., Churchill, W., & Newman, P. (2014). Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles. *IEEE International Conference on Robotics and Automation*, Hong Kong, China, p. 5.

McManus, C., Churchill, W., Maddern, W., Stewart, A., & Newman, P. (2014). Shady dealings: Robust, long-term visual localisation using illumination invariance. 2014 *IEEE International Conference on Robotics and Automation*, Hong Kong, China, pp. 901–906.

McManus, C., Upcroft, B., & Newman, P. (2014). Scene signatures: Localised and point-less features for localisation. RSS. Rome, Italy.

Milford, M., Prasser, D., & Wyeth, G. (2005). Experience mapping: Producing spatially continuous environment representations using RatSLAM. *Proceedings of Australasian Conference on Robotics and Automation 2005*, pp. 1–10.

Milford, M., & Wyeth, G. (2010). Persistent navigation and mapping using a biologically inspired slam system. *The International Journal of Robotics Research*, *29*(9), 1131–1153.

Milford, M., Wyeth, G., & Prasser, D. (2004). RatSLAM: A hippocampal model for simultaneous localization and mapping. *IEEE International Conference on Robotics and Automation*, New Orleans, LA, vol 1, IEEE, pp. 403–408.

Mühlfellner, P., Bürki, M., Bosse, M., Derendarz, W., Philippsen, R., & Furgale, P. (2016). Summary maps for lifelong visual localization. *Journal of Field Robotics*, *33*(5), 561–590.

Mühlfellner, P., Furgale, P., Derendarz, W., & Philippsen, R. (2015). *Designing a relational database for long-term visual mapping*. http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A811166&dswid=-565

Paton, M., MacTavish, K., Ostafew, C., & Barfoot, T. (2015). Itas not easy seeing green: Lighting-resistant stereo Visual Teach & Repeat using color-constant images. 2015 *IEEE International Conference on Robotics and Automation* (*ICRA*), Seattle, WA, pp. 1519–1526.

Paton, M., Mactavish, K., Warren, M., & Barfoot, T. D. (2016). Bridging the appearance gap: Multi-experience localization for long-term visual teach and repeat. 2016 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Daejeon, South Korea, pp. 1918–1925.

Prasser, D., Milford, M., & Wyeth, G. (2006). Outdoor simultaneous localisation and mapping using RatSLAM. *Field and Service Robotics*, *25*, 143–154.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, *24*(5), 513–523.

Sattler, T., Leibe, B., & Kobbelt, L. (2011). Fast image-based localization using direct 2D-to-3D matching. 2011 *International Conference on Computer Vision*, Barcelona, Spain, pp. 667–674.

Schindler, G., Brown, M., & Szeliski, R. (2007). City-scale location recognition. 2007 *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, pp. 1–7.

Stumm, E., Mei, C., Lacroix, S., & Chli, M. (2015). Location graphs for visual place recognition. 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, pp. 5475-5480.

How to cite this article: Bürki M, Cadena C, Gilitschenski I, Siegwart R, Nieto J. Appearance-based landmark selection for visual localization. *J Field Robotics*. 2019;36:1041–1073. https://doi.org/10.1002/rob.21870

APPENDIX A

We present the observation percentage and localization precision separately for each data set of the three data set collections in Section A.1. This relates to Sections 5.5, and 5.6, where the same metrics are shown in aggregated form.

Furthermore, we compare the localization performance with different choices of feature descriptors on the *parking-lot* data sets. The respective results can be found in Section A.2.

We conclude the appendix with a list of all data sets used in this evaluation, the respective weather conditions, and some sample images in Section A.3.

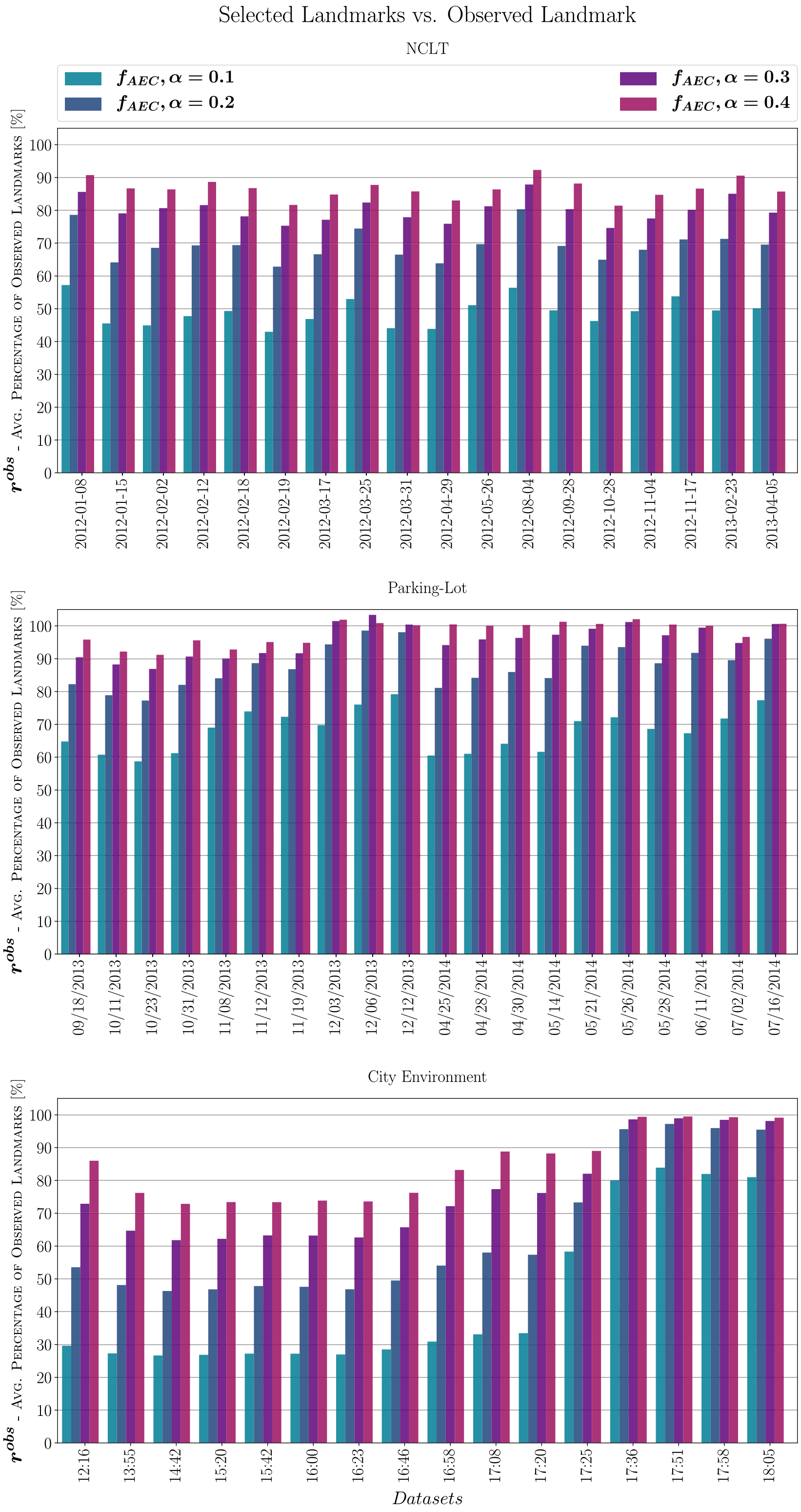
A.1 | Individual data set performance analysis

In Figure A1, the observation percentage is shown with ranking function f_{AEC} for selection fractions of 10–40% with maps containing only *rich sessions*. It can be observed that for certain data sets of the *parking-lot* collection, the average number of observed landmarks with a 30% or

40% selection fraction can even exceed the average number of observed landmarks when using all candidate landmarks. This exhibits a saturation effect, resulting in occasionally achieving a higher number of observed landmarks with only a subset of selected landmarks, as opposed to using all candidate landmarks. While counterintuitive at first, this is due to the fact that including more candidate landmarks increases the chance of forming wrong 2D–3D matches. After the subsequent pose estimation step, these wrong matches are then classified as outliers, resulting in a potentially lower number of observed landmarks.

Furthermore, the different observation percentage characteristics during daytime as opposed to at night are clearly visible in the *city environment*. During the day, a selection of 40% of the landmarks

FIGURE A1 The average observation percentage r^{obs} for selection fractions between 10% and 40%, for every data set of the *NCLT* (top), *parking-lot* (middle), and *city environment* data set collection against maps containing only *rich sessions* [Color figure can be viewed at wileyonlinelibrary.com]



is not sufficient for an observation percentage of more than 90%, while at nighttime, even 20% of selected landmarks achieve almost an observation percentage of 100%.

The localization precision using different ranking functions and with a selection fraction of 20% are shown for each data set of all three collections in Figure A2–A4, respectively. The results reflect the patterns visible in Figure A1, and in Section 5.5. The best performance is achieved using f_{AEC}, f_{AV}, and f_{TfIdfB} for ranking landmarks, with precision values often close to that of using all landmarks for localization instead. While the precision using f_{MRS} can vary considerably between different data sets, ranking functions f_{TfIdfA} fails, resulting in occasionally even worse precision than selecting landmarks randomly.

Enriching the maps with *observation sessions* results in a higher variance of performance between different ranking functions, as can be seen in Figure A5 for the *NCLT* and *parking-lot* collection, and in Section 5.6 in Figure 8 for the *city environment*. The respective localization precision results are shown in Figures A6–A8. Most notable is the failure of the ranking function f_{NCV} during daytime in the *city environment*. As discussed in Section 5.6, ranking function f_{AEC} is the only one achieving consistently high localization precision in the *city environment* both during daytime, at dusk, as well as at nighttime.

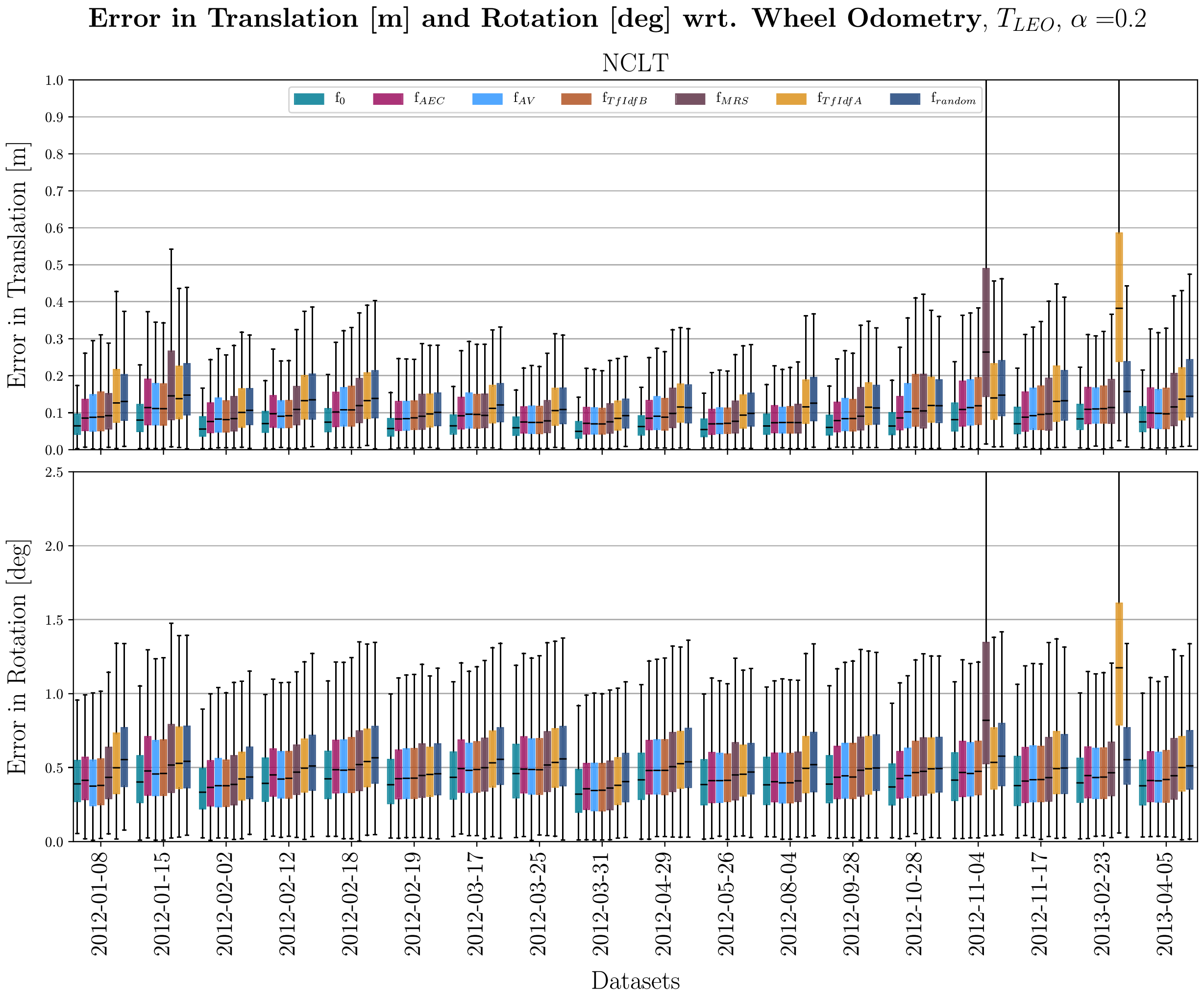


FIGURE A2 The localization precision for a selection fraction of 20%, for every data set of the *NCLT* collection against the map containing only *rich sessions* [Color figure can be viewed at wileyonlinelibrary.com]

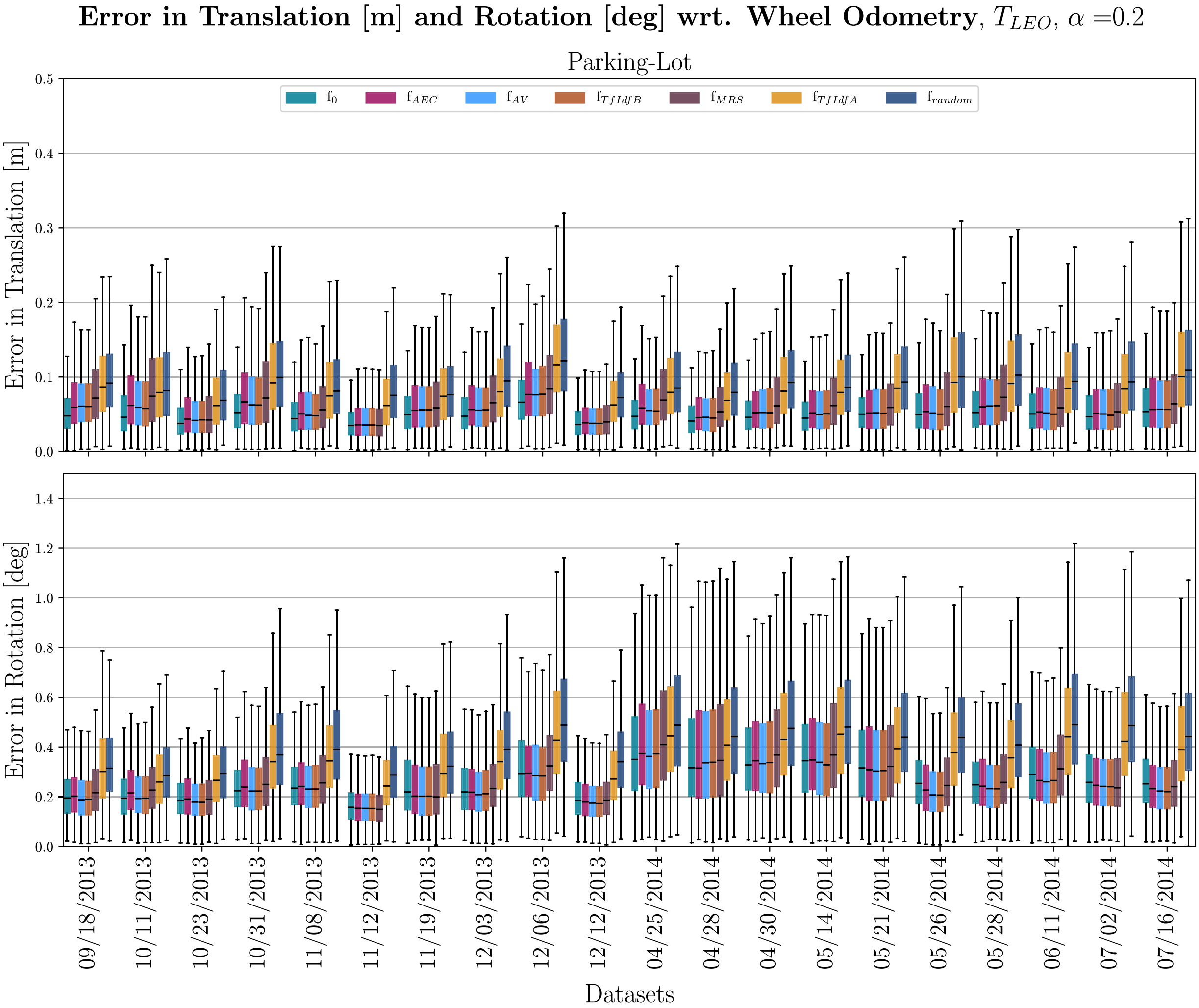


FIGURE A3 The localization precision for a selection fraction of 20%, for every data set of the *parking-lot* collection against the map containing only *rich sessions* [Color figure can be viewed at wileyonlinelibrary.com]

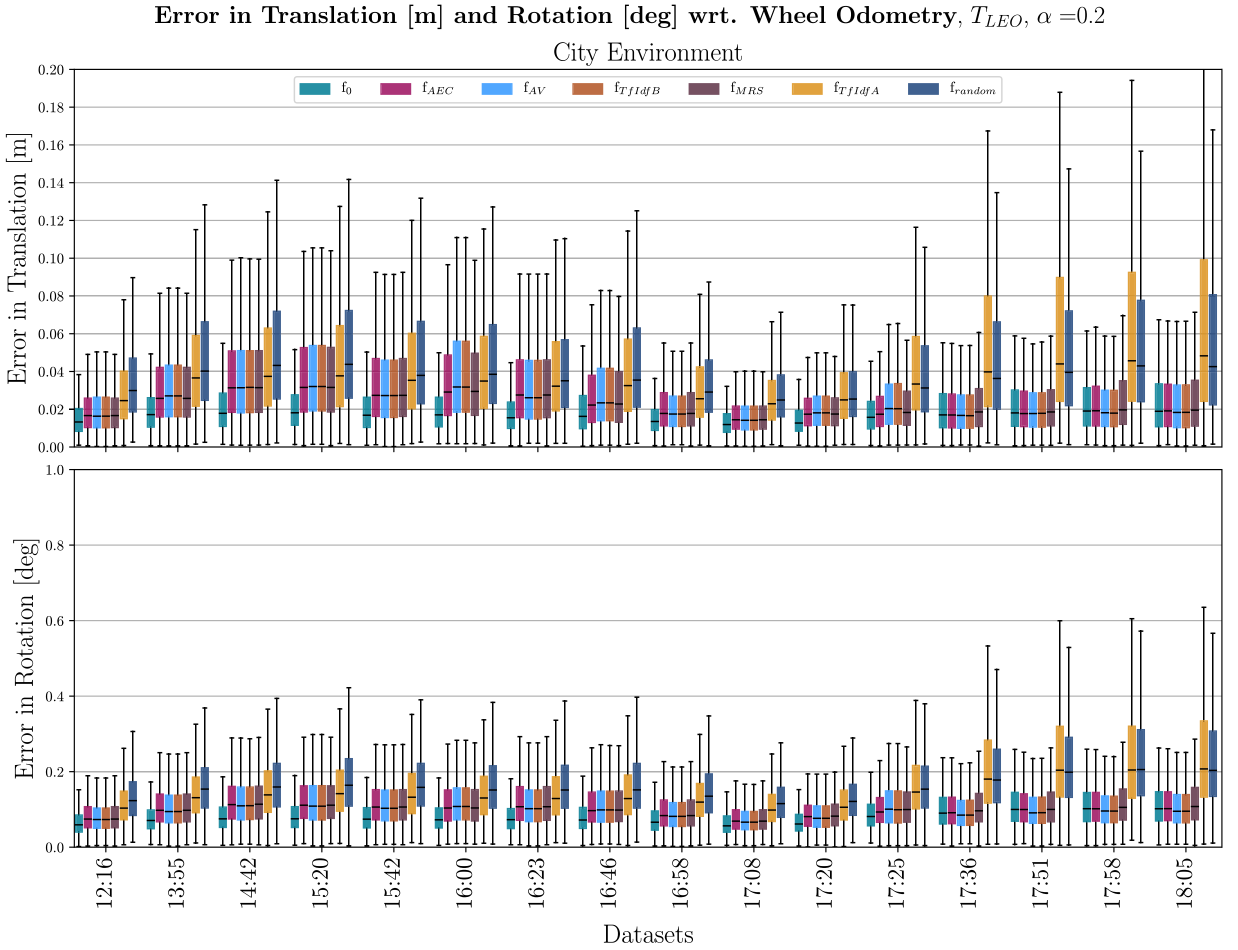


FIGURE A4 The localization precision for a selection fraction of 20%, for every data set of the *city environment* collection against the map containing only *rich sessions* [Color figure can be viewed at wileyonlinelibrary.com]

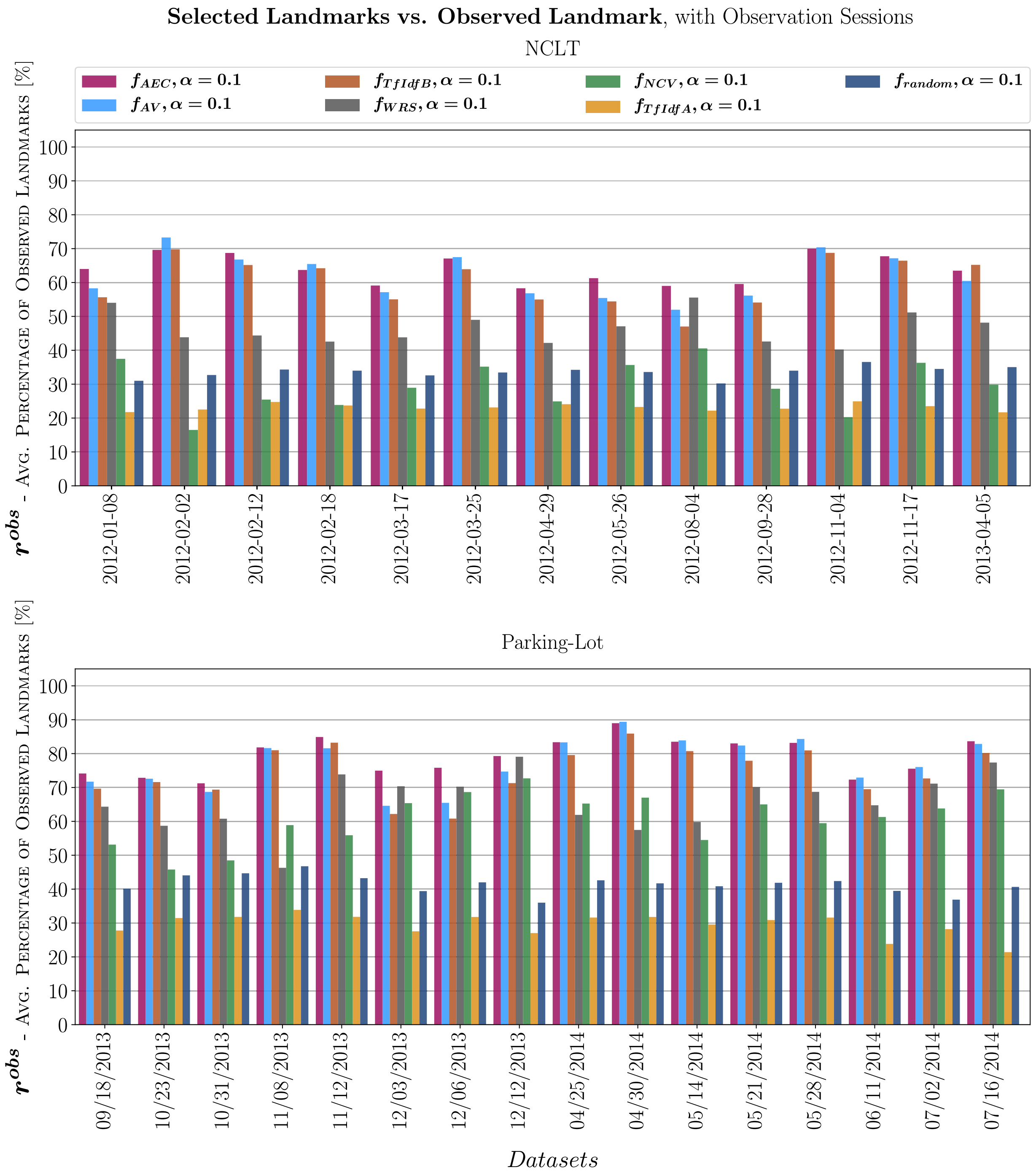


FIGURE A5 The average observation percentage r^{obs} for a selection fraction of 10%, for every data set of the *NCLT* and *parking-lot* collection against the map with *observation sessions* [Color figure can be viewed at wileyonlinelibrary.com]

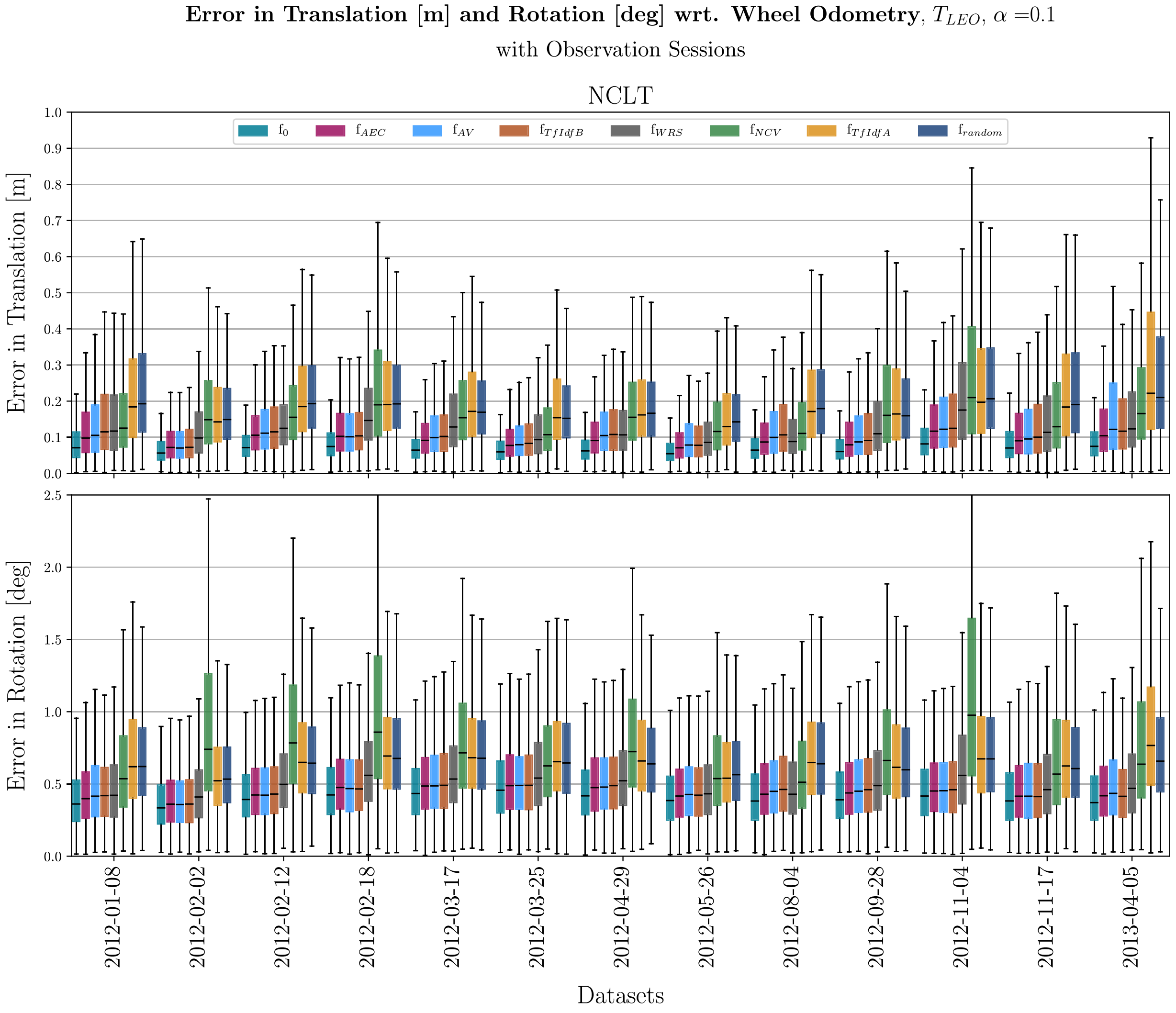


FIGURE A6 The localization precision for a selection fraction of 10%, for every data set of the *NCLT* collection against the map with *observation sessions* [Color figure can be viewed at wileyonlinelibrary.com]

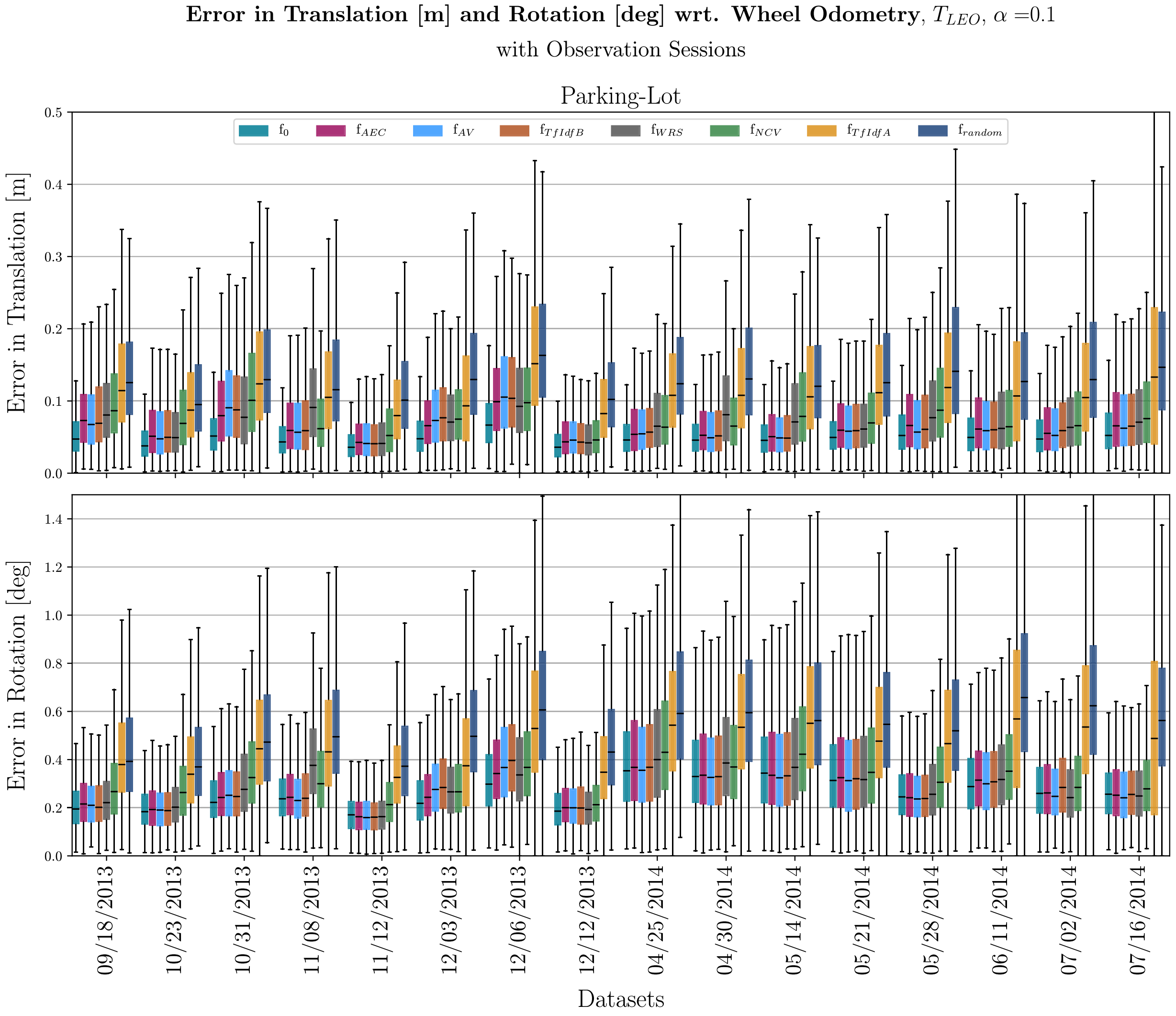


FIGURE A7 The localization precision for a selection fraction of 10%, for every data set of the *parking-lot* collection against the map with *observation sessions* [Color figure can be viewed at wileyonlinelibrary.com]

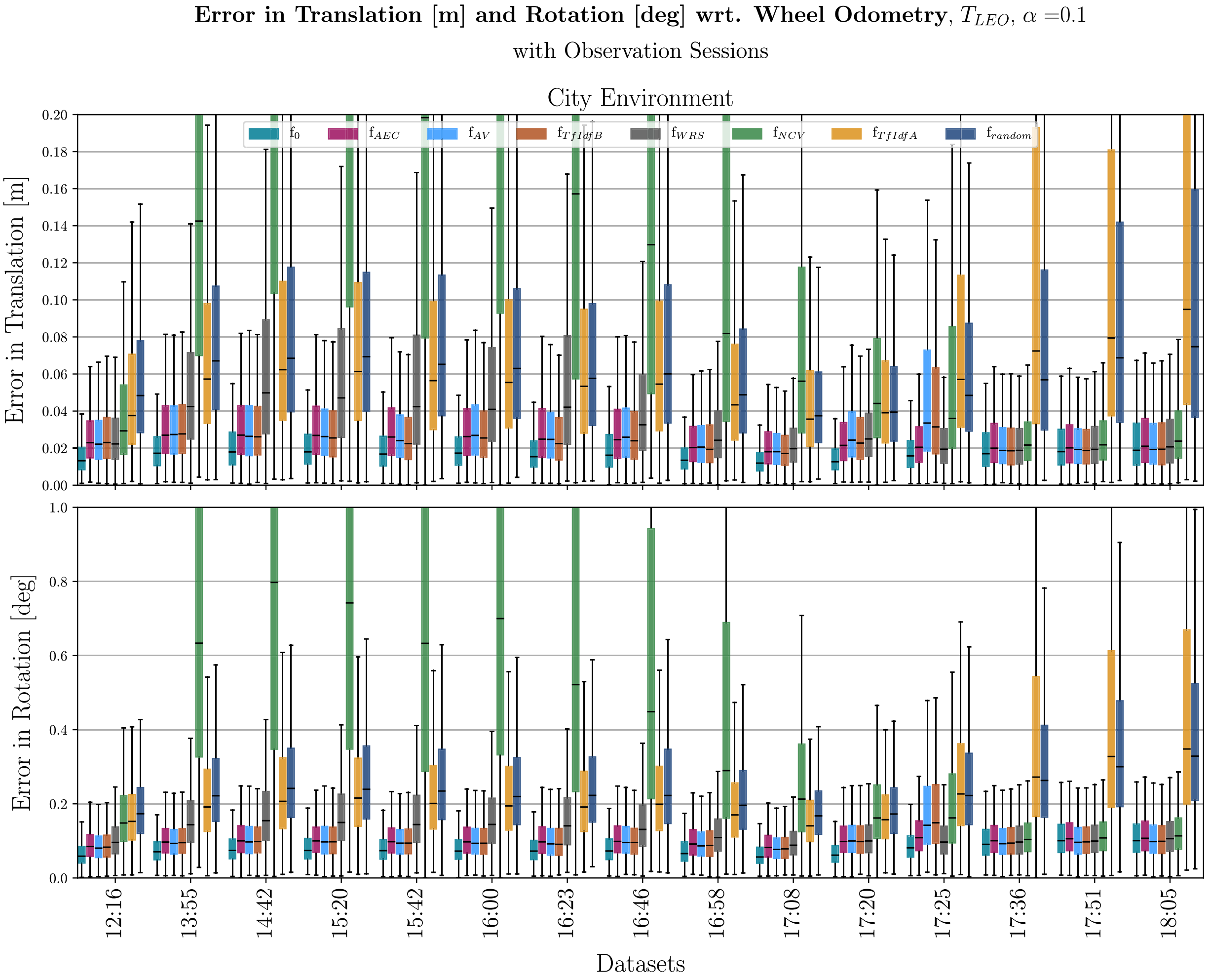


FIGURE A8 The localization precision for a selection fraction of 10%, for every data set of the *city environment* collection against the map with *observation sessions* [Color figure can be viewed at wileyonlinelibrary.com]

A.2 | Feature descriptor comparison

Our proposed appearance-based landmark ranking functions are per construction independent of the local feature descriptor used for mapping and localization, as they only take the co-observability patterns of landmarks into account. Nevertheless, the feature descriptor is an integral part of the localization pipeline, and thus the resulting performance of the localization with appearance-based landmark selection may not be identical with every choice of local feature descriptor. We have therefore evaluated the localization performance with popular choices of different local feature descriptors on the *parking-lot* data set collection. The results are shown in Figures A9–A11. As expected, the results are similar regardless of the choice of descriptor.

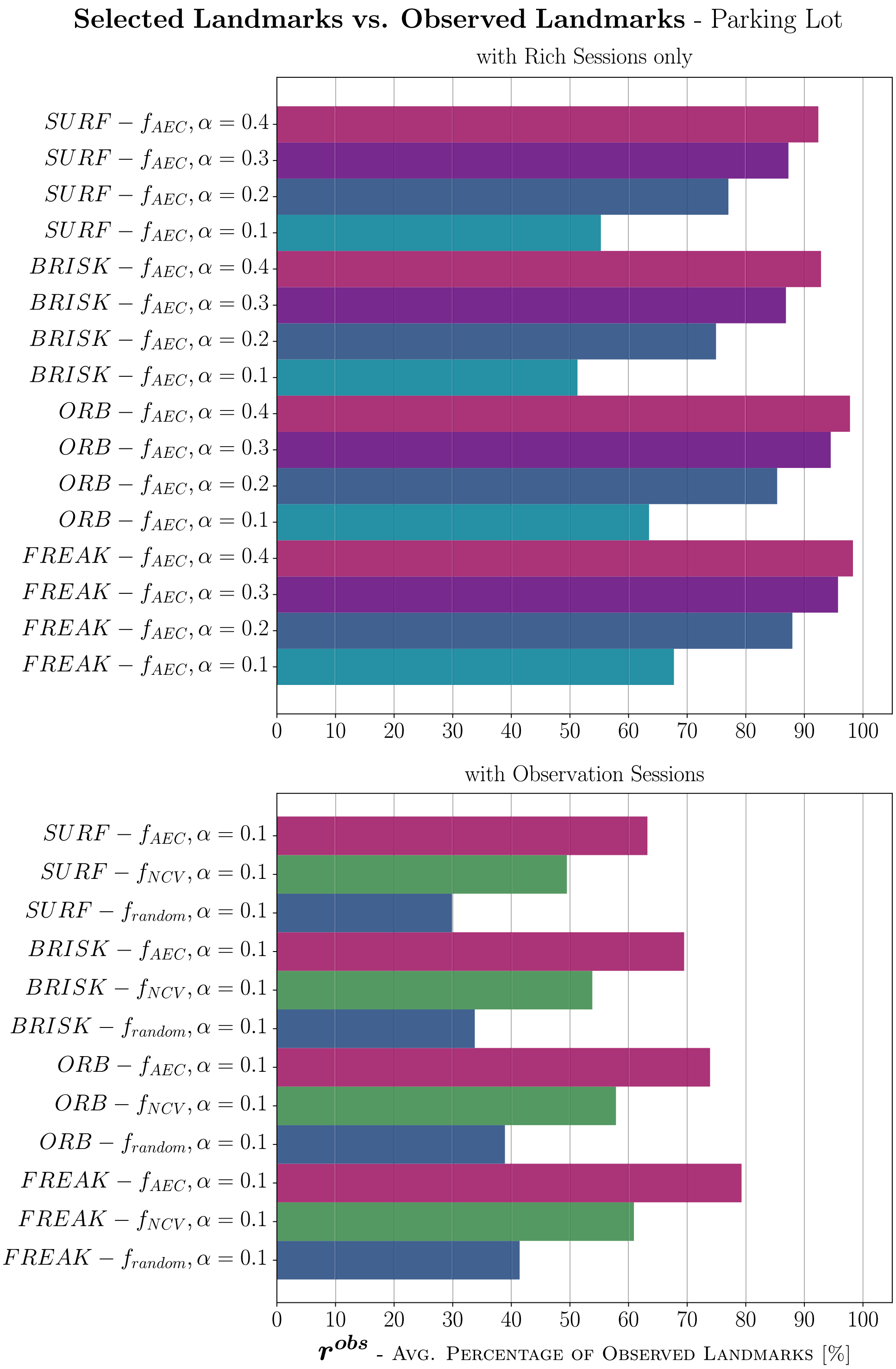


FIGURE A9 Observation percentage for different choices of feature descriptors, with a selection fraction of 20% against the map with only *rich sessions*, aggregated over all data sets of the *parking-lot* collection [Color figure can be viewed at wileyonlinelibrary.com]

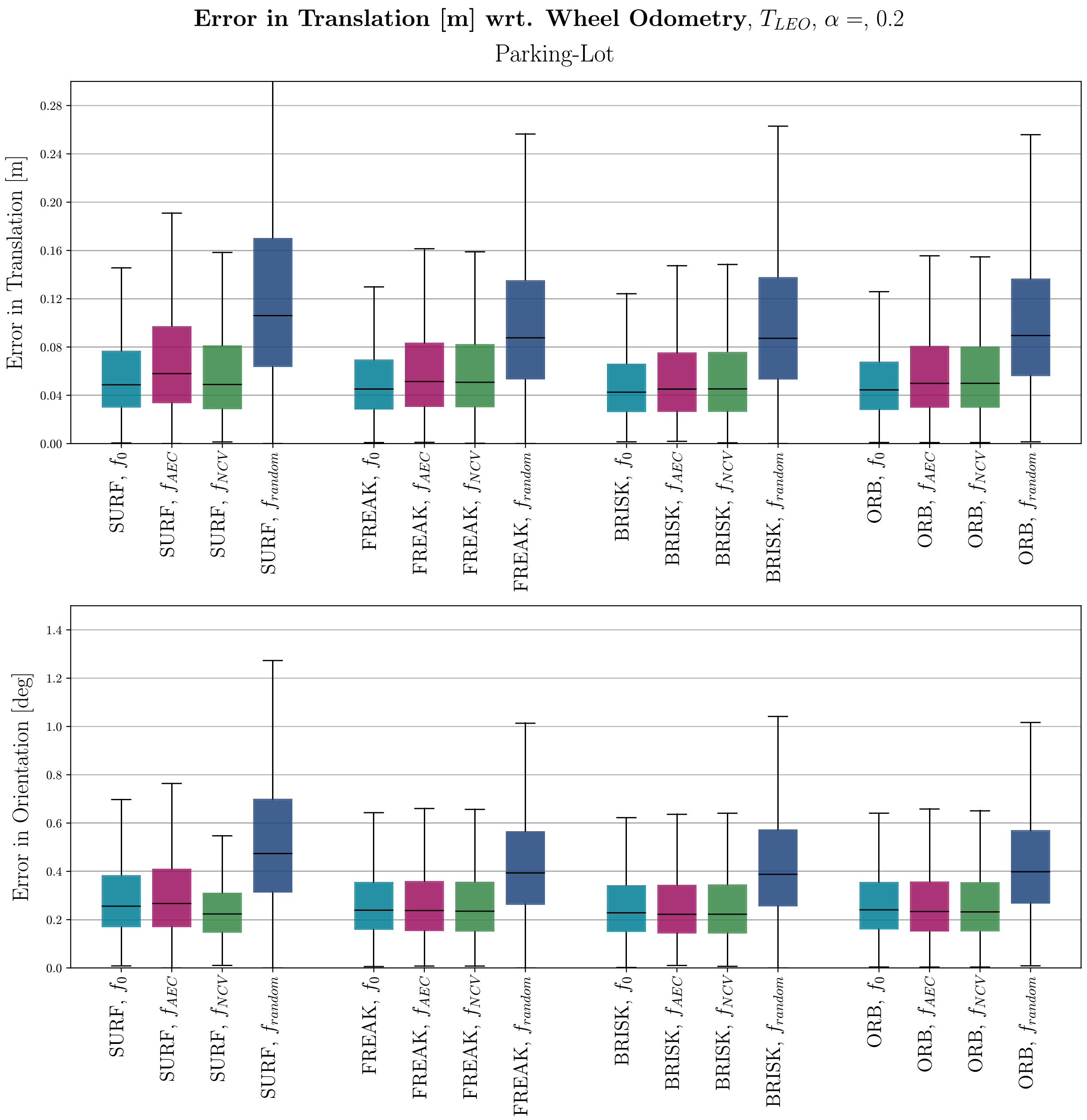


FIGURE A10 Localization precision with different choices of feature descriptors, a selection fraction of 20% against the map with only *rich sessions*, aggregated over all data sets of the *parking-lot* collection [Color figure can be viewed at wileyonlinelibrary.com]

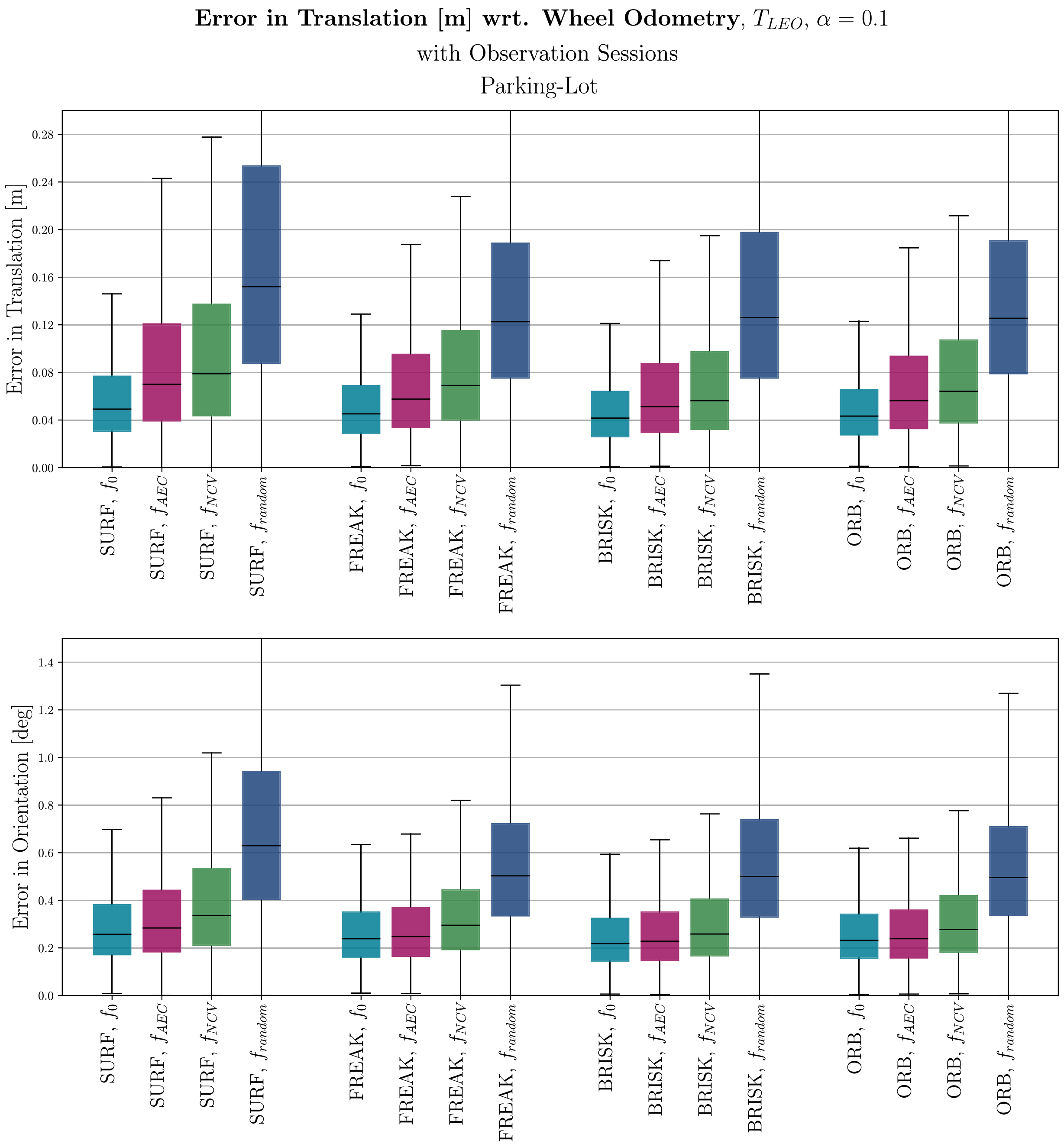


FIGURE A11 Localization precision with different choices of feature descriptors, a selection fraction of 10% against the map with *observation sessions*, aggregated over all data sets of the *parking-lot* collection [Color figure can be viewed at wileyonlinelibrary.com]

A.3 | Sample images

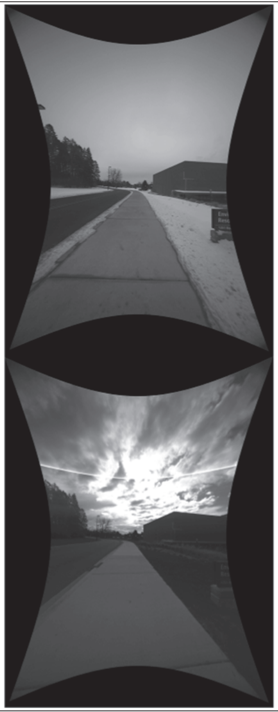
TABLE A1 List of the *parking-lot* data sets with their respective weather condition and usage in the maps. The lower-case "*r*" and "*o*" indicate that the data set has been added to the map as *rich* and *observation sessions*, respectively

Date	W.	U.	Example Images
20th August, 2013	☁	*r*	
17th September, 2013	☼		
18th September, 2013	🌧		
11th October, 2013	☁	*o*	
16th October, 2013	⛅	*r*	
23rd October, 2013	☁		
31st October, 2013	☼		
7th November, 2013	☁	*r*	
8th November, 2013	☼		
12th November, 2013	☁		
19th November, 2013	☁	*o*	
3rd December, 2013	☼		
6th December, 2013	🌨		
10th December, 2013	☁	*r*	
12th December, 2013	☁		
14th January, 2014	☼	*r*	
25th April, 2014	⛅		
28th April, 2014	⛅	*o*	
30th April, 2014	⛅		
5th May, 2014	☼	*r*	
14th May, 2014	⛅		
21st May, 2014	☼		
26th May, 2014	⛅	*o*	
28th May, 2014	🌧		
11th June, 2014	⛅		
30th June, 2014	⛅	*r*	
2nd July, 2014	⛅		
16th July, 2014	☼		

TABLE A2 List of the *city environment* data sets with their respective weather condition and usage in the maps. The lower-case "*r*" and "*o*" indicate that the data set has been added to the map as *rich* and *observation sessions*, respectively

Date	W.	U.	Example Images
11:49	☼	*r*	
12:16	☼		
13:55	☼		
14:31	☼	*o*	
14:42	☼		
15:20	☼		
15:42	☼		
15:56	☼	*o*	
16:00	☼		
16:23	☼		
16:46	☼		
16:58	◒		
17:03	◒	*o*	
17:08	◒		
17:15	◒	*r*	
17:20	◒		
17:25	◒		
17:30	◒	*r*	
17:36	●		
17:43	●	*r*	
17:51	●		
17:58	●	*o*	
18:05	●		

TABLE A3 List of the *NCLT* data sets with their respective weather condition and usage in the maps. The lower-case "*r*" and "*o*" indicate that the data set has been added to the map as a *rich* and *observation sessions*, respectively

Date	W.	U.	Example Images
8th January, 2012	☼		
15th January, 2012	☼	*o*	
22nd January, 2012	●	*r*	
2nd February, 2012	⛅		
4th February, 2012	☼	*r*	
5th February, 2012	☼	*r*	
12th February, 2012	☼		
18th February, 2012	☼		
19th February, 2012	⛅	*o*	
17th March, 2012	☼		
25th March, 2012	☼		
31st March, 2012	☁	*o*	
29th April, 2012	⛅		
11th May, 2012	☼	*r*	
26th May, 2012	☼		
16th June, 2012	☼	*r*	
4th August, 2012	☼		
20th August, 2012	☼	*r*	
28th September, 2014	⛅		
28th October, 2014	⛅	*o*	
4th November, 2014	☁		
16th November, 2014	☼	*r*	
17th November, 2014	☼		
23rd February, 2013	●	*o*	
5th April, 2013	☼		