Journal Title
XX(X):1–??

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Sequential Monte Carlo Localization in Topometric Appearance Maps

Alberto Jaenal, Francisco-Angel Moreno and Javier Gonzalez-Jimenez

Abstract

Representing the scene appearance by a global image descriptor (BoW, NetVLAD, etc.) is a widely adopted choice to address Visual Place Recognition (VPR). The main reasons are that appearance descriptors can be effectively provided with radiometric and perspective invariances as well as they can deal with large environments because of their compactness. However, addressing metric localization with such descriptors (a problem called Appearance-based Localization, or AbL) achieves much poorer accuracy than those techniques exploiting the observation of 3D landmarks, which represent the standard for Visual Localization. In this paper, we propose ALLOM (**A**ppearance-based **L**ocalization with **L**ocal **O**bservation **M**odels) which addresses AbL by leveraging the topological location of a robot within a map to achieve accurate metric estimations. This topology-assisted metric localization is implemented with a sequential Monte Carlo Bayesian filter that applies a specific observation model for each different place of the environment, thus taking advantage of the local correlation between the pose and the appearance descriptor within each region. ALLOM also benefits from the topological structure of the map to detect eventual robot loss-of-tracking and to effectively cope with its relocalization by applying VPR. Our proposal demonstrates superior metric localization capability compared to different state-of-the-art *AbL* methods under a wide range of situations.

Keywords

Visual Localization, Topometric Maps, Appearance-based Localization

Introduction

The development of emerging technologies like autonomous vehicles (robots, cars or UAVs) or Augmented Reality devices demands fast and reliable Visual Localization (VL) methods to determine the pose of a camera given a pre-built model of the environment (Piasco et al. 2018; Toft et al. 2020). Current state-of-the-art (SOTA) VL methods are commonly addressed from a 3D perspective, by relying on a model that comprises geometric entities (landmarks), mostly 3D points (Mur-Artal et al. 2015), and sometimes segments as well (Gomez-Ojeda et al. 2019). With this model, the camera pose is estimated by minimizing a cost function that accounts for the errors between the landmark projections and their corresponding image observations. This approach has proved to perform very accurately in a wide variety of scenarios, being nowadays adopted as the *de facto* standard for VL (Lynen et al. 2020).

However, relying on such 3D map for VL also comes with a number of limitations and drawbacks that arise when: (i) global localization is required (e.g. for relocalization, wake-up, and kidnapping problems); (ii) few and/or poorly distributed features are detected in the images; (iii) the lighting conditions of the scene vary substantially compared to those in the map (e.g. day/night, different seasons, etc.); or (iv) the map becomes very large, which demands further processing and memory resources.

In contrast, Appearance-based Localization (AbL) offers an entirely different perspective for VL, as it avoids modeling the 3D geometry of the world. Instead, the image content is encoded into a compact descriptor, typically through a Convolutional Neural Network (CNN) (Arandjelovic et al. 2016; Lopez-Antequera et al. 2017a), and the environment appearance is represented through a database of image descriptors annotated with their locations, often known as an *Appearance Map* (AM). These AMs have demonstrated to be particularly suitable for Visual Place Recognition (VPR) (Lowry et al. 2015), both in very large environments and when facing strong lighting changes. Unfortunately, their advantage for topological localization comes with a price: poor performance for metric localization.

Such shortcoming becomes particularly pronounced when using VPR to obtain accurate AbL results (Sattler et al. 2018), as the outcome is the pose of the element in the AM with the highest visual resemblance. A commonly adopted solution is to assume that the camera follows a previously traversed path (Maddern et al. 2012; Thoma et al. 2019), which leads to a substantial improvement in accuracy but with limited applicability. Aiming for better generalization, some authors opt to represent the environment by a dense AM (Ham et al. 2005; Lopez-Antequera et al. 2016), but this leads to maps containing unnecessary and redundant

Machine Perception and Intelligent Robotics (MAPIR) Group, Malaga Institute for Mechatronics Engineering and Cyber-Physical Systems (IMECH.UMA), University of Malaga, Spain.

Corresponding author:
Francisco-Angel Moreno, System Engineering and Automation Department, University of Malaga, Campus de Teatinos, 29071 Malaga, Spain.
Email: famoreno@uma.es

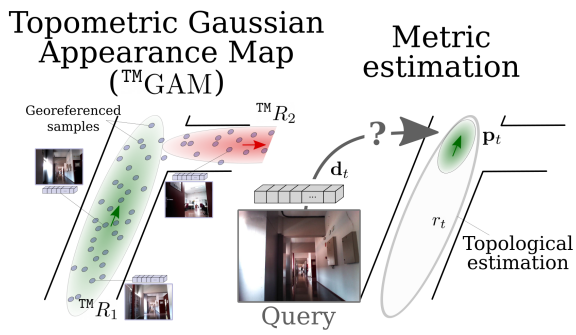


Figure 1. Our proposal models the correlation between pose and appearance at each region R_j with a Gaussian distribution. Metric pose can be estimated from this correlation which is specific for each region.

samples, resulting in substantial computational burden and memory expense.

As a solution to these dense, redundant AMs, in a previous work (Jaenal et al. 2022) we proposed to cluster samples that are close in both pose and appearance into places represented by multivariate Gaussian distributions. We call this abstracted representation of the AM a *Gaussian Appearance Map* (GAM). However, in order to estimate the Gaussian distributions in the high-dimensional joint space of descriptor and poses, that work assumed conditional independence between these variables. This consideration restricted the GAM to topological location only.

In this paper, we build on top of these topological GAMs to model the relationship between poses and appearance on a local basis, now defining correlated distributions over these two elements for each region (see Fig. 1). Specifically, we consider planar poses for the robot and a reduced (projected) version of the appearance descriptor. The resulting conditional probabilities serve as local observation models that are integrated into a complete probabilistic framework based on a sequential Montecarlo filter, resulting in a system called ALLOM (**A**ppearance-based **L**ocalization with **L**ocal **O**bservation **M**odels). This solution estimates the robot pose with higher precision than existing AbL techniques while keeping a low computational burden. Moreover, the system is able to address relocalization by triggering VPR when needed. This feature allows for fast initialization and the capability to deal with the kidnapped robot problem. This claim is supported by evaluating and contrasting ALLOM against other SOTA methods in three different indoor datasets: two gathered in real scenarios and another one created with synthetic images. Additionally, we provide a demonstration video* and have released the code†.

Related works

The most simplistic approach for addressing AbL is Visual Place Recognition (Lowry et al. 2015), a technique that, instead of estimating the query pose, assigns it the pose of the most similar element in the AM. Several works have proposed to exploit the spatio-temporal consistency conferred by image sequences to improve the accuracy of VPR (Milford and Wyeth 2012; Vysotska and Stachniss 2015, 2017). For example, the authors of Xu et al. (2020, 2021) rely on Bayesian filtering to achieve high performance even under challenging lighting conditions. Similarly, in Doan et al. (2020) the authors propose to incrementally build an abstracted database through clusterization, achieving state-of-the-art performance with a Bayesian filter upon optimal computational cost. However, VPR approaches are of a topological nature and cannot be employed for metric estimation, requiring to be assessed by some metric threshold (Sattler et al. 2017; Xu et al. 2021) that uniquely determines whether a query has been correctly localized.

Some authors have enabled metric localization by enhancing VPR with techniques for metric estimation such as Gaussian Processes (GPs) regression (Huhle et al. 2010; Schairer et al. 2011) or relative pose transform estimation through CNNs (Laskar et al. 2017; Balntas et al. 2018; Ding et al. 2019). On the other hand, CNN architectures have been proposed for absolute pose regression from whole images (Kendall et al. 2015; Brahmbhatt et al. 2018), although they still present several limitations as the high computational cost of learning a complete environment and a restricted generalization ability (Sattler et al. 2019). However, all these works are dedicated to estimating the pose of individual images, which makes localization more prone to inconsistency problems (e.g. Perceptual Aliasing) in challenging conditions.

In recent years, though, some authors have proposed more elaborate procedures for AbL. A common approach is to employ AMs created from a single trajectory, so the query pose can be determined through one-dimensional interpolation between map samples. This way, CAT-SLAM (Maddern et al. 2012) associates increments on pose and appearance in order to address AbL on the discrete maps proposed by FAB-MAP (Cummins and Newman 2008). Other authors have employed flow networks on single sequence maps (Naseer et al. 2014, 2018; Thoma et al. 2019), exploiting temporal and spatial correlation to efficiently address camera localization or mapping. In a similar manner, the AM proposed in (Jaenal et al. 2021) consists of a set of adjacent elements that describe areas where pose and appearance are assumed to be smooth, addressing AbL through a PF combining GP‡, regression and odometry. Obviously, using a pre-fixed sequence as AM is a very restricted solution that lacks applicability to realistic conditions where the robot moves freely in the environment. In contrast to all these approaches, our work grounds on maps expressed as a set of Multivariate Gaussians, which, instead of relying on a pre-established path, enables further generalization over the whole environment.

Another potential approach are generating dense, unordered AMs for the environment, where techniques such as Gaussian Process Particle Filters (GPPFs) (Lopez-Antequera et al. 2016, 2017b) can carry out AbL, using GPs as observation model and odometry for particle propagation. In short, the GPPF method first finds the nearest samples in the map in terms of pose for each particle and subsequently

*https://youtu.be/4vkuK4_RfVQ

†https://github.com/AlbertoJaenal/AppearanceSeqMCL

‡The Gaussian Process employed by these works uses the pose as input and the descriptor as output. However, the authors simplify the regression by assuming independence between the components of the descriptors.

weights the observation at that location through a GP. Another work operating in dense AMs is the one in (Jaenal et al. 2020), which proposes a position graph that combines VPR over urban-scale AMs with accurate Visual Odometry from local features, allowing the correction of the drift and also to geo-alignment of the sequences. However, these approaches require a high amount of samples to achieve reasonable accuracy (most likely with redundant data), at the cost of increasing both the computational time and the map size. The topometric abstraction employed in this work summarizes the information contained in dense AMs, achieving a consistent and precise localization while substantially improving its computational cost.

Finally, it is worth noting that, unlike the approaches cited above, which assume independence between descriptor components, ours is, to the best of our knowledge, the first work for AbL that exploits the correlation between the pose and appearance vector components (in this case, a reduced/projected version).

Topometric Gaussian Appearance Maps

Our localization system builds on top of the concept of Gaussian Appearance Map (GAM), proposed in our previous work (Jaenal et al. 2022). In a nutshell, a GAM is a probabilistic model consisting of a Mixture of Multivariate Gaussian distributions over the joint space of pose and appearance descriptor. In (Jaenal et al. 2022), a GAM provides a topological-only representation of an environment, suitable for efficient VPR but not for metric AbL. To make this topological nature explicit, from now on we will refer to it as ${}^{\mathtt{T}}$GAM. Here, we introduce a metric version of a GAM, denoted ${}^{\mathtt{M}}$GAM, that now takes into account the local correlation between pose and appearance in order to facilitate metric AbL.

In the following, we start by formally defining an Appearance Map (AM) as the source for building the GAMs; then we review the basics concepts of the original ${}^{\mathtt{T}}$GAM and address the synthesis of the new metric ${}^{\mathtt{M}}$GAM. Finally, we define the aggregation leading to the topometric ${}^{\mathtt{TM}}$GAM, which will allow for topology-assisted metric robot localization. A summary of the notation used throughout the paper is provided in Table 1, for quick reference.

Appearance Map

We define an Appearance Map ($\mathcal{AM}$) as an unordered set of geo-tagged image descriptors, typically collected during several robot navigations (Gálvez-López and Tardós 2012; Torii et al. 2015; Arandjelovic et al. 2016). Formally:

$$\mathcal{AM} = \{\mathbf{x}_i\}_{i=1}^{N}, \tag{1}$$

where N is the number of pairs $\mathbf{x}_i = [\mathbf{q}_i, \mathbf{d}_i]^\intercal$, formed by a pose vector $\mathbf{q}_i \in \mathrm{SE}(2)$ and a D-dimensional appearance descriptor $\mathbf{d}_i \in \mathbb{R}^D$.

In the case of indoors, where many parts of the environment are likely to be revisited multiple times, $\mathcal{AM}$ will include repeated (or very similar) views that do not add any meaningful information to the map, while increasing its size unnecessarily. Reducing this irrelevant data is the main motivation behind the abstraction process that leads to the ${}^{\mathtt{T}}$GAM described next.

Table 1. Summary of the employed notation

SYMBOL	MEANING
$\mathbf{x}_i, \mathbf{q}_i, \mathbf{d}_i$	i^{th} element in the database, its metric pose and its appearance descriptor, respectively
${}^{\mathtt{T}}\boldsymbol{\mathcal{R}}$, ${}^{\mathtt{T}}R_j$	**Topological** GAM, and its j^{th} region
${}^{\mathtt{M}}\boldsymbol{\mathcal{R}}$, ${}^{\mathtt{M}}R_j$	**Metric** GAM, and its j^{th} region
${}^{\mathtt{TM}}\boldsymbol{\mathcal{R}}$	**Topometric** GAM
${}^{\mathtt{T}}\mu_j^{\mathbf{q}}$, ${}^{\mathtt{T}}\Sigma_j^{\mathbf{q}}$	Mean and covariance of the pose Gaussian distribution for ${}^{\mathtt{T}}R_j$
${}^{\mathtt{T}}\mu_j^{\mathbf{d}}$, ${}^{\mathtt{T}}\sigma_j^2\mathbf{I}^D$	Mean and covariance of the descriptor Gaussian distribution for ${}^{\mathtt{T}}R_j$
$\delta_{j,i}$	PCA-reduced i^{th} descriptor for the j^{th} region
${}^{\mathtt{M}}\mu_j^{\mathbf{q}}$, ${}^{\mathtt{M}}\Sigma_j^{\mathbf{qq}}$	Mean and covariance of the pose Gaussian distribution for ${}^{\mathtt{M}}R_j$
${}^{\mathtt{M}}\mu_j^{\delta}$, ${}^{\mathtt{M}}\Sigma_j^{\delta\delta}$	Mean and covariance of the descriptor Gaussian distribution for ${}^{\mathtt{M}}R_j$
${}^{\mathtt{M}}\Sigma_j^{\mathbf{q}\delta}$	Pose-descriptor cross-covariance matrix for ${}^{\mathtt{M}}R_j$
$\mathbf{p}_t, r_t, \mathbf{u}_t$	Metric pose, topological region (in the GAM) and odometry reading of the robot at time t
$\tilde{\mathbf{p}}_t, \tilde{r}_t, \tilde{w}_t$	Metric pose, topological region and weight of a certain particle at time t

Topological GAMs: ${}^{\mathtt{T}}$GAM

The topological GAM consists of a Multivariate Gaussian Mixture Model that covers the mapped environment with a set of M regions ${}^{\mathtt{T}}\boldsymbol{\mathcal{R}} = \{{}^{\mathtt{T}}R_j\}_{j=1}^{M}$, each representing a local area with similar appearance. This similarity in both appearance and poses defines a *place* of the environment. The number of regions M chosen to cluster the ${}^{\mathtt{T}}$GAM is selected according to the Davies-Bouldin (DB) Index (Davies and Bouldin 1979), which estimates the optimal value from the existing poses of the map. Further information about this index can be found in our previous work.

According to this model, the elements ${}^{\mathtt{T}}\mathbf{x}$ within each region ${}^{\mathtt{T}}R_j$ are considered to follow a certain Gaussian distribution over the joint space of pose and appearance:

$${}^{\mathtt{T}}\mathbf{x} \sim \mathcal{N}\left(\begin{bmatrix} {}^{\mathtt{T}}\mu_j^{\mathbf{q}} \\ {}^{\mathtt{T}}\mu_j^{\mathbf{d}} \end{bmatrix}, \begin{bmatrix} {}^{\mathtt{T}}\Sigma_j^{\mathbf{q}} & \mathbf{0} \\ \mathbf{0} & {}^{\mathtt{T}}\sigma_j^2\mathbf{I}^D \end{bmatrix}\right) \tag{2}$$

A full explanation of the estimation of the distribution parameters can be found elsewhere (Jaenal et al. 2022). From now on, we will use ${}^{\mathtt{T}}R_j$ to refer indistinctly to both the topological region and its associated Gaussian distribution.

As seen in Eq. 2, two simplifications are adopted in the construction of this model due to computational reasons: (i) the independence between poses and descriptors, and (ii) the modeling of the appearance descriptor $\mathbf{d}$ as an isotropic normal distribution. In the first one, the lack of correlation between poses and descriptors prevents inferring information about the camera pose given the image descriptor and vice versa. Also, the assumption of an isotropic distribution of the appearance of the region is too simplistic to properly characterize the appearance variability of a place. Although both simplifications have proved to be assumable for reliable VPR, as demonstrated in (Jaenal et al. 2022), they are not

suitable for metric localization, where a given observed image descriptor must provide information about the pose of the camera.

Metric GAMs: $^{\mathtt{M}}$GAM

In this paper, we present the metric $^{\mathtt{M}}$GAM, which overcomes the above-mentioned limitations of the $^{\mathtt{T}}$GAM by modeling the local correlation between pose and appearance within each region $^{\mathtt{M}}R_j$.

The main hindrance to the calculation of such correlation is the high dimensionality of the appearance descriptor (typically, $D \geq 2^{10}$), which demands an unfeasibly huge amount of map samples to estimate the covariance between both variables (pose and appearance). Therefore, we project the descriptor to a subspace with a more manageable dimension ($D' \sim 2^6$). Particularly, we apply Principal Component Analysis (PCA) (Pearson 1901), since it preserves maximum variance between the descriptor components. However, employing a single PCA-based reduction for the entire map is inadequate because the AM forms a highly non-linear manifold embedded in the descriptor space, which can not be linearly mapped without losing meaningful appearance information. This becomes evident if we think that similar descriptors from different locations can be projected to the same simplified descriptor (i.e. Perceptual Aliasing), hence losing any trustable correlation between the projected descriptors and their corresponding poses. To avoid this, we apply PCA on a local basis, that is, reducing the dimensionality of the appearance by taking into account only those descriptors located within each region of the map. Formally, for a given appearance descriptor $\mathbf{d}_i$ associated to a sample located within the j^{th} region, we generate its projected descriptor using a PCA model specifically trained for that region:

$$\delta_{j,i} = \text{PCA}_j(\mathbf{d}_i) \in \mathbb{R}^{D'} \mid (D' \ll D). \tag{3}$$

This idea is key for the formulation of the localization process described next, and shares the same principles as other hybrid localization methods that use structure-based maps (Blanco et al. 2008; Mazuran et al. 2018).

Dimensionality Reduction training. The PCA model for the j^{th} region of the map is trained with a subset $\widehat{\mathcal{AM}}_j$ that collects those samples that fulfill two requirements: (i) they belong to the original dense $\mathcal{AM}$, and (ii) their likelihood of falling in the j-th region is the maximum among all regions. Formally:

$$\widehat{\mathcal{AM}}_j = \left\{ \mathbf{x}_i \mid (\mathbf{x}_i \in \mathcal{AM}) \wedge \left(\arg\max_k \mathcal{L}\left(\mathbf{x}_i \mid {}^{\mathtt{T}}R_k\right) = j \right) \right\}_{i=1}^{N_j}, \tag{4}$$

where N_j is the number of samples of the subset and $\mathcal{L}\left(\mathbf{x}_i \mid {}^{\mathtt{T}}R_k\right)$ represents the likelihood of the i^{th} sample within the Gaussian distribution associated to the k^{th} topological region (defined in Eq. 2). Note that, although this likelihood is evaluated in the space of concatenated pose and appearance, only the descriptors will be used to train the model PCA_j for the region. Then, this model is employed to project the descriptors $\mathbf{d}_i$ in $\widehat{\mathcal{AM}}_j$, leading to a new set:

$${}^{\mathtt{M}}\widehat{\mathcal{AM}}_j = \{{}^{\mathtt{M}}\mathbf{x}_i\}_{i=1}^{N_j} \mid {}^{\mathtt{M}}\mathbf{x}_i = [\mathbf{q}_i, \delta_{j,i}]^{\intercal}, \tag{5}$$

now including the projected descriptors $\delta_{j,i}$ instead of the original $\mathbf{d}_i$.

Creation of the metric regions. The final step consists of determining the parameters ${}^{\mathtt{M}}\mu_j$ and ${}^{\mathtt{M}}\Sigma_j$ of the Gaussian distribution associated to each metric region ${}^{\mathtt{M}}R_j$, given the samples ${}^{\mathtt{M}}\mathbf{x}_k \in {}^{\mathtt{M}}\widehat{\mathcal{AM}}_j$:

$${}^{\mathtt{M}}\mathbf{x}_k \sim \mathcal{N}({}^{\mathtt{M}}\mu_j, {}^{\mathtt{M}}\Sigma_j) \equiv \mathcal{N}\left(\begin{bmatrix} {}^{\mathtt{M}}\mu_j^{\mathbf{q}} \\ {}^{\mathtt{M}}\mu_j^{\delta} \end{bmatrix}, \begin{bmatrix} {}^{\mathtt{M}}\Sigma_j^{\mathbf{qq}} & {}^{\mathtt{M}}\Sigma_j^{\mathbf{q}\delta} \\ {}^{\mathtt{M}}\Sigma_j^{\delta\delta} & {}^{\mathtt{M}}\Sigma_j^{\delta\delta} \end{bmatrix} \right). \tag{6}$$

where ${}^{\mathtt{M}}\Sigma_j^{\mathbf{q}\delta} = \left({}^{\mathtt{M}}\Sigma_j^{\delta\mathbf{q}}\right)^{\mathtt{T}}$.

These parameters are computed by applying a Maximum Likelihood Estimation process:

$${}^{\mathtt{M}}\mu_j = \frac{\sum_k w_k \, {}^{\mathtt{M}}\mathbf{x}_k}{\sum_k w_k}, \tag{7}$$

$${}^{\mathtt{M}}\Sigma_j = \frac{\sum_k w_k \, ({}^{\mathtt{M}}\mathbf{x}_k \boxminus {}^{\mathtt{M}}\mu_j)({}^{\mathtt{M}}\mathbf{x}_k \boxminus {}^{\mathtt{M}}\mu_j)^{\intercal}}{\sum_k w_k}, \tag{8}$$

with k iterating over all elements in the subset and where $w_k = \mathcal{L}\left({}^{\mathtt{M}}\mathbf{x}_k \mid {}^{\mathtt{T}}R_j\right)$ is the likelihood of the sample given the topological region.

Here, the operator $\boxminus$ represents the *subtraction* between two samples in the concatenated $[\mathbf{q}, \delta]$ space, that is: ${}^{\mathtt{M}}\mathbf{x}_1 \boxminus {}^{\mathtt{M}}\mathbf{x}_2 = [\mathbf{q}_1 \ominus \mathbf{q}_2, \, \delta_{j,1} - \delta_{j,2}]^{\intercal}$, where the operator $\ominus$ stands for the pose inverse composition (Fernández-Madrigal and Blanco-Claraco 2012).

Topometric GAMs: $^{\mathtt{TM}}$GAM

At this point, we have defined two different Gaussian Appearance Maps for an environment: $^{\mathtt{T}}$GAM describes it from a topological perspective, while $^{\mathtt{M}}$GAM models the local correlation between the pose and the appearance, being suitable for metric localization.

For convenience of notation, we define a topometric map $^{\mathtt{TM}}$GAM as a dual Multivariate Gaussian Mixture Model composed of the following aggregation:

$${}^{\mathtt{TM}}\boldsymbol{\mathcal{R}} = \left\{ \left({}^{\mathtt{T}}R_j, \, {}^{\mathtt{M}}R_j\right) \right\}_{j=1}^{M}. \tag{9}$$

As a result of maintaining both GAMs, we can leverage the topological knowledge of the environment to improve metric localization, as described next.

Sequential Montecarlo Localization

The proposed system ALLOM (Appearance-based Localization with Local Observation Models) takes advantage of both GAMs to sequentially estimate the metric pose of the robot $\mathbf{p}_t \in SE(2)$ while keeping consistency with its topological location $r_t \in [1...M]$ in the $^{\mathtt{T}}$GAM. We call this: topology-assisted metric Appearance-based Localization. In addition, ALLOM provides methods for initialization and relocalization after loss-of-tracking that also relies on topological information.

The sequential localization, whose graphical model is shown in Fig. 3, is formulated as the estimation of the posterior distribution:

$$p\left(\mathbf{p}_t \mid \delta_t, \mathbf{u}_t, \mathbf{p}_{t-1}, \, {}^{\mathtt{TM}}\boldsymbol{\mathcal{R}}\right). \tag{10}$$

In this formulation, the probability of the current pose of the robot $\mathbf{p}_t$ depends on: (i) its previous pose $\mathbf{p}_{t-1}$;

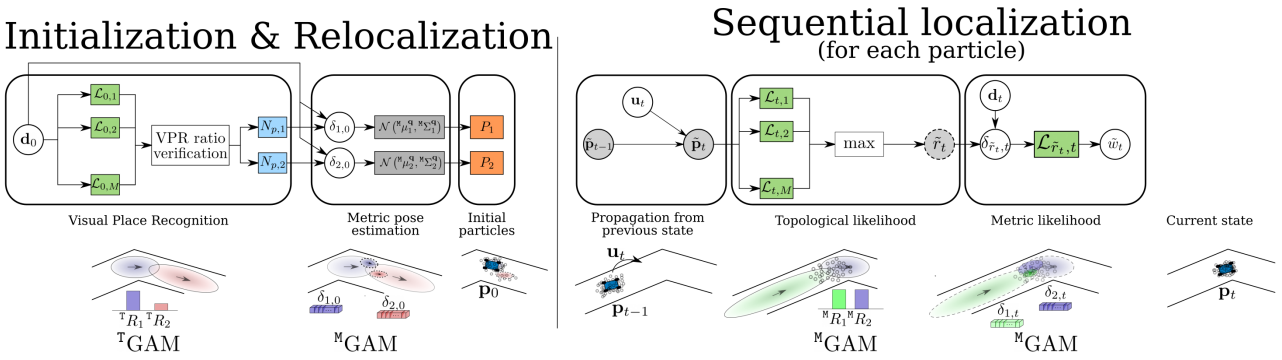


Figure 2. Operation of ALLOM for Appearance-based Localization over a Gaussian Appearance Map ($^{\mathtt{TM}}$GAM). The left part of the figure describes the process of initialization and relocalization after a robot loss-of-tracking: first, it applies VPR to detect the most probable regions in the $^{\mathtt{T}}$GAM and then uses the specific descriptor for each region δ_r to determine the initial pose of each particle. The right part describes the sequential localization: after propagating the particle pose with the odometry, each particle is assigned to a region of the map. The local observation model of this region is used to compute the likelihood that weights the particle.

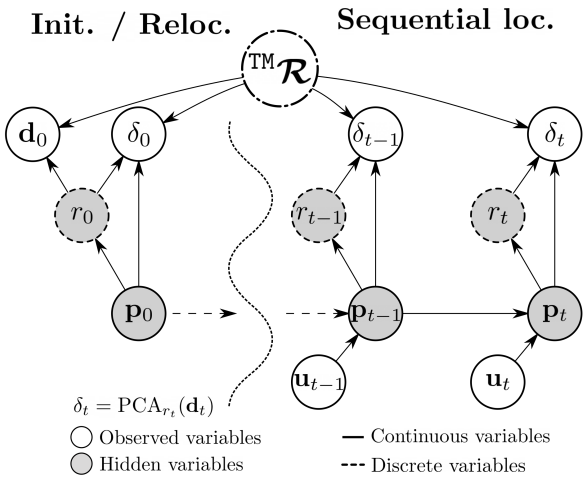


Figure 3. Probabilistic graphical model for ALLOM. The two unknown (hidden) variables at each time instant are the region r in the $^{\mathtt{TM}}$GAM that the robot is traversing and the global pose p.

(ii) the projection of the observed appearance descriptor $\delta_t = \mathrm{PCA}_{r_t}(\mathbf{d}_t)$, given the topological location of the robot; (iii) the odometry of the robot $\mathbf{u}_t$; and (iv) the $^{\mathtt{TM}}$GAM. For convenience, from now on, we will omit specifying the map $^{\mathtt{TM}}\boldsymbol{\mathcal{R}}$ in the formulation, since it is involved in all the derivations.

We estimate Eq. 10 through Bayesian filtering:

$$p\left(\mathbf{p}_t \mid \delta_t, \mathbf{u}_t, \mathbf{p}_{t-1}\right) \propto \underbrace{p\left(\delta_t \mid \mathbf{p}_t\right)}_{\text{Local obs. model}} \underbrace{p\left(\mathbf{p}_t \mid \mathbf{u}_t, \mathbf{p}_{t-1}\right)}_{\text{Transition model}}, \tag{11}$$

where the rightmost term is the transition model and the middle term represents a contribution of this work: a local observation model specifically fitted for the region r_t in which the robot is located. We solve this localization problem through a Particle Filter (PF), so that a set of N_p weighted particles approximates the posterior belief of the current robot pose in Eq. 10.

In practice, the localization process modeled by Eq. 11 is carried out as summarized in Fig. 2 (right): we first propagate the previous metric pose of the particles, we then use the resulting poses to determine their topological location, and, finally, we update their weights, as explained next. For the sake of clarity, in the following we will refer to the current topological state, metric pose and weight of a particle by $\tilde{r}_t$, $\tilde{\mathbf{p}}_t$ and $\tilde{w}_t$, respectively.

Transition model

The transition model in Eq. 11 updates the pose of the particles applying the odometry reading $\mathbf{u}_t$:

$$\tilde{\mathbf{p}}_t = \tilde{\mathbf{p}}_{t-1} \oplus \mathbf{u}_t \oplus \epsilon, \tag{12}$$

with $\epsilon \sim \mathcal{N}(0, \Sigma^{\mathbf{u}})$ being Gaussian noise, and $\oplus$ the pose composition operator (Fernández-Madrigal and Blanco-Claraco 2012).

Local observation model

In turn, the local observation model in Eq. 11 modifies the importance of each particle according to the likelihood of the current appearance observation δ_t.

For that, the first step is to determine the topological region $\tilde{r}_t$ where each particle lies in, defined as the region of the $^{\mathtt{M}}$GAM where the particle has the maximum likelihood from its pose:

$$\tilde{r}_t = \arg\max_j \mathcal{L}\left(\tilde{\mathbf{p}}_t \mid {}^{\mathtt{M}}\mu_j^{\mathbf{q}}, {}^{\mathtt{M}}\Sigma_j^{\mathbf{qq}}\right). \tag{13}$$

Once the particle region is known, its associated PCA projection function is applied to the observed image descriptor $\mathbf{d}_t$ to obtain the reduced descriptor $\delta_{\tilde{r}_t,t}$. Then, the weight $\tilde{w}_t$ of each particle is updated with the likelihood:

$$p\left(\delta_t \mid \mathbf{p}_t\right) = \mathcal{L}\left(\delta_{\tilde{r}_t,t} \mid {}^{\mathtt{M}}\mu^{\delta}_{\tilde{r}_t,t}, {}^{\mathtt{M}}\Sigma^{\delta}_{\tilde{r}_t,t}\right), \tag{14}$$

where $^{\mathtt{M}}\mu^{\delta}_{\tilde{r},t}$ and $^{\mathtt{M}}\Sigma^{\delta}_{\tilde{r},t}$ are the mean and covariance of the conditional distribution over the space of the projected descriptors given the current pose of the particle $\tilde{\mathbf{p}}_t$:

$$\begin{aligned} {}^{\mathtt{M}}\mu^{\delta}_{\tilde{r}_t,t} &= {}^{\mathtt{M}}\mu^{\delta}_{\tilde{r}_t} + {}^{\mathtt{M}}\Sigma^{\delta\mathbf{q}}_{\tilde{r}_t} {}^{\mathtt{M}}\Sigma^{\mathbf{qq}^{-1}}_{\tilde{r}_t} \left(\tilde{\mathbf{p}}_t \ominus {}^{\mathtt{M}}\mu^{\mathbf{q}}_{\tilde{r}_t}\right) \\ {}^{\mathtt{M}}\Sigma^{\delta}_{\tilde{r}_t,t} &= {}^{\mathtt{M}}\Sigma^{\delta\delta}_{\tilde{r}_t} + {}^{\mathtt{M}}\Sigma^{\delta\mathbf{q}}_{\tilde{r}_t} {}^{\mathtt{M}}\Sigma^{\mathbf{qq}^{-1}}_{\tilde{r}_t} {}^{\mathtt{M}}\Sigma^{\mathbf{q}\delta}_{\tilde{r}_t}. \end{aligned} \tag{15}$$

Note that, in this expression, all matrix multiplications can be calculated off-line, since they do not depend on the actual pose of the particle, but only on the parameters of the metric region defined in Eq. 6. Just the inverse pose composition and the subsequent product must be performed at each step.

Finally, in order to avoid the numerical divergences derived from evaluating high-dimensional multivariate likelihoods in the projected descriptor space, we weight the particles through the log-likelihood:

$$\tilde{w}_t \propto \tilde{w}_{t-1} \ \log \mathcal{L}\left(\delta_{\tilde{r}_t,t} \mid {}^{\mathtt{M}}\mu^{\delta}_{\tilde{r}_t,t}, {}^{\mathtt{M}}\Sigma^{\delta}_{\tilde{r}_t,t}\right). \quad (16)$$

As a common practice, when the Effective Sample Size (ESS) of the PF falls below $\tau_{ESS} = \frac{N_P}{2}$, we apply Sampling Importance Resampling (Rubin 1988).

Initialization and Relocalization

Once the procedure for sequential localization has been stated, we discuss here the detection of the robot loss-of-tracking and how the PF addresses initialization and particle relocalization.

Loss-of-tracking detection. Due to the fact that local observation models are defined specifically for each region, a correct topological localization must be guaranteed in order to obtain reliable metric estimations. This implies determining whether the robot tracking is consistent or, on the contrary, has been lost. Hence, we identify a *loss-of-tracking* when the metric pose estimation $\mathbf{p}_t$ of the robot is significantly far from the center of its topological location in the map r_t during a certain time window w_{loss}. This is measured by means of the set of Mahalanobis distances in pose $\mathbf{\Delta}_{\mathbf{M}}(\mathbf{p}_t) = \left\{\Delta_M\left(\mathbf{p}_{t'}, {}^{\mathtt{T}}R_{r_{t'}}\right)\right\}^{t}_{t'=t-w_{loss}}$ within the time window, with:

$$\Delta_M\left(\mathbf{p}_t, {}^{\mathtt{T}}R_{r_t}\right) = \sqrt{\left(\mathbf{p}_t \ominus {}^{\mathtt{T}}\mu^{\mathbf{q}}_{r_t}\right)^T \left({}^{\mathtt{T}}\Sigma^{\mathbf{q}}_{r_t}\right)^{-1} \left(\mathbf{p}_t \ominus {}^{\mathtt{T}}\mu^{\mathbf{q}}_{r_t}\right)}. \quad (17)$$

Inspired by (Xu et al. 2021), we employ a chi-squared cumulative distribution with three degrees of freedom over such distances:

$$P(\mathbf{\Delta}_{\mathbf{M}}(\mathbf{p}_t) < \tau_\chi) = \chi^2_3(\tau_\chi), \quad (18)$$

to measure the probability mass of the Mahalanobis distance being under the scalar threshold τ_χ. This probability describes how likely the tracking of the robot has been lost at time step t, in order to trigger a relocalization procedure.

Particle initialization and relocalization. Particle initialization in ALLOM is carried out according to the following procedure, illustrated in Fig. 2 (left):

- First, the initial topological location of the particles $\tilde{r}_0$ is obtained through probabilistic VPR. This involves evaluating the likelihood of the observation $\mathbf{d}_0$ at the appearance term of all the ${}^{\mathtt{T}}$GAM regions (see Eq. 2):

$$\mathcal{L}_{0,j} = \mathcal{L}\left(\mathbf{d}_0 \mid {}^{\mathtt{T}}\mu^{\mathbf{d}}_j, {}^{\mathtt{T}}\sigma^2_j \mathbf{I}^D\right). \quad (19)$$

- We select all the regions whose likelihood is greater than 80% of the most probable one (VPR ratio verification in Fig. 2 (left)). For each of them, a number of particles proportional to their likelihood are deployed: $N_{p,j} \propto \mathcal{L}_{0,j}$.
- We then compute the projected descriptor $\delta_{\tilde{r}_0,0}$ for each selected region and obtain a conditional Gaussian distribution in pose $\mathcal{N}\left({}^{\mathtt{M}}\mu^{\mathbf{q}}_{\tilde{r}_0,0}, {}^{\mathtt{M}}\Sigma^{\mathbf{q}}_{\tilde{r}_0,0}\right)$ from the metric regions in the ${}^{\mathtt{M}}$GAM, given the value of $\delta_{\tilde{r}_0,0}$:

$$\begin{aligned} {}^{\mathtt{M}}\mu^{\mathbf{q}}_{\tilde{r}_0,0} &= {}^{\mathtt{M}}\mu^{\mathbf{q}}_{\tilde{r}_0} + {}^{\mathtt{M}}\Sigma^{\mathbf{q}\delta}_{\tilde{r}_0} {}^{\mathtt{M}}\Sigma^{\delta\delta^{-1}}_{\tilde{r}_0} \left(\delta_{\tilde{r}_0,0} - {}^{\mathtt{M}}\mu^{\delta}_{\tilde{r}_0}\right) \\ {}^{\mathtt{M}}\Sigma^{\mathbf{q}}_{\tilde{r}_0,0} &= {}^{\mathtt{M}}\Sigma^{\mathbf{qq}}_{\tilde{r}_0} + {}^{\mathtt{M}}\Sigma^{\mathbf{q}\delta}_{\tilde{r}_0} {}^{\mathtt{M}}\Sigma^{\delta\delta^{-1}}_{\tilde{r}_0} {}^{\mathtt{M}}\Sigma^{\delta\mathbf{q}}_{\tilde{r}_0}. \end{aligned} \quad (20)$$

- The initial pose $\tilde{\mathbf{p}}_0$ of the particles for each region $\tilde{r}_0$ are determined by drawing samples from such conditional distribution.
- Finally, the initial weights of the particles are uniformly distributed $\tilde{w}_0 = \frac{1}{N_p}$.

In the case of loss-of-tracking, relocalization follows an analogous procedure, but starting from the observation $\mathbf{d}_t$ instead.

Experimental setup

Before describing the experiments, we first introduce the error measures chosen for the evaluation of our proposal and, subsequently, the datasets and parameter settings employed for the tests.

Evaluation error

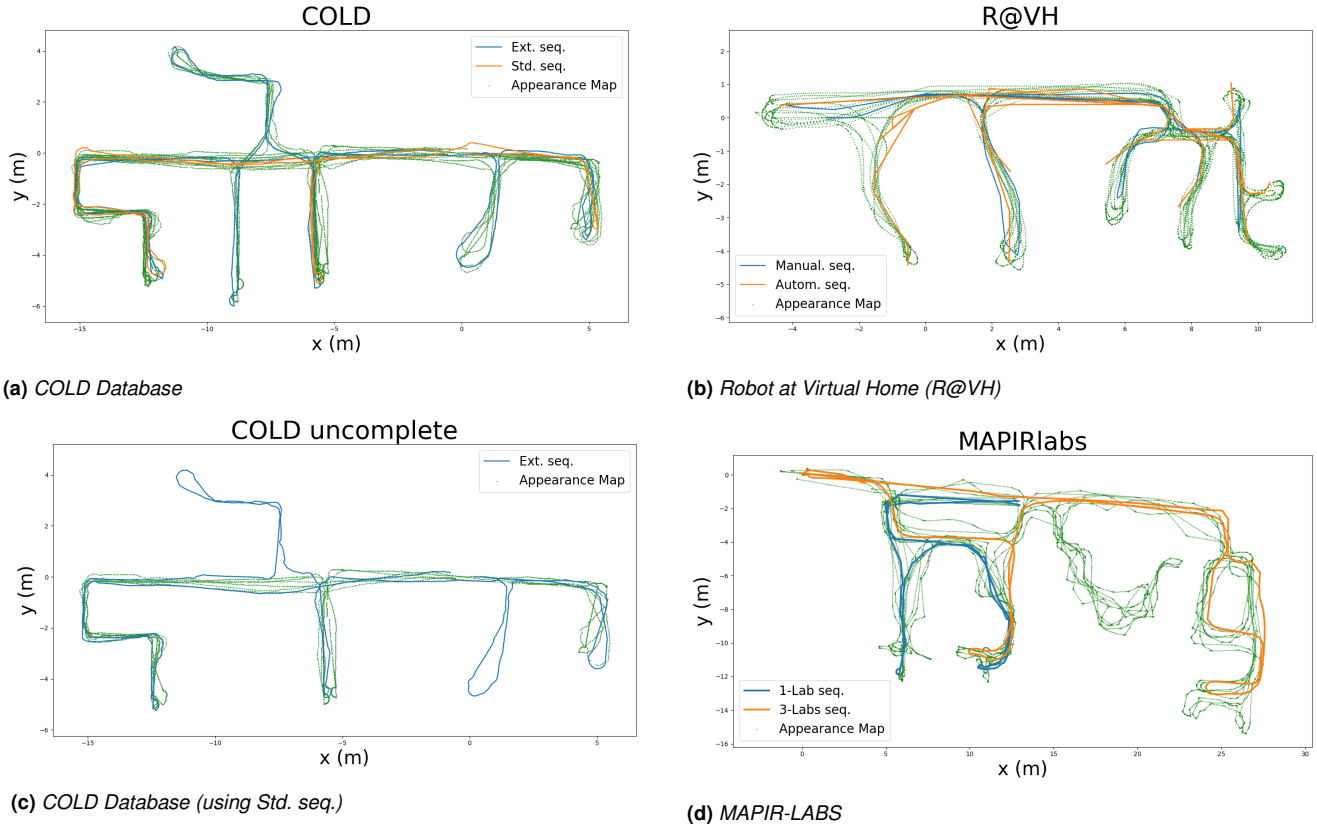
Given that ALLOM is a topology-assisted AbL method, we evaluate its performance in both metric and topological terms, as detailed next.

Metric localization accuracy. Two different criteria have been chosen to measure the metric accuracy: (i) the well-known *Absolute Trajectory Error* (ATE), widely employed in SLAM methods (Mur-Artal and Tardós 2017; Gomez-Ojeda et al. 2019), and (ii) the *Median Translation Error* (MTE) Sattler et al. (2018); Toft et al. (2020), since considering only ATE for methods that can incur in loss-of-tracking and perform relocalization yields biased results due to the deviation that outliers induce in ATE. In such situations, MTE brings more realistic results.

Topological localization accuracy. Regarding the topological aspect, a grounded criterion to decide whether the estimate at a certain time step is correct, consists of checking how likely is that the ground-truth robot pose $\mathbf{p}^*_t$ lies in the estimated topological region r_t. This measure is typically used in VPR-related works (Lowry et al. 2015; Xu et al. 2021; Jaenal et al. 2022). To this end, we consider that the estimated region is correct if the Mahalanobis distance between the ground-truth and the pose distribution of the region $\Delta_M\left(\mathbf{p}^*_t, {}^{\mathtt{T}}R_{r_t}\right)$ falls below a value of 4, since it represents the 99% confidence region for a three-dimensional distribution, such as the pose in $SE(2)$. From this detection process, we evaluate the topological accuracy with the *Area Under the Curve* (AUC) of the Precision-Recall (PR) curves (Xu et al. 2020), varying a region importance threshold. The importance of a region is computed as the sum of the weights of the particles that have been topologically assigned to it.

Dataset description

We evaluate our proposal against three datasets containing images of planar robot motion in indoor scenes: two publicly



(a) *COLD Database*

(b) *Robot at Virtual Home (R@VH)*

(c) *COLD Database (using Std. seq.)*

(d) *MAPIR-LABS*

Figure 4. Poses in the Appearance Maps (AMs) are plotted as green dots and the evaluation trajectories for the three employed datasets are plotted as solid lines. The poses of the AM are a collection of unordered elements obtained from different robot trajectories, while the evaluation dataset consists of sequences that do not totally overlap with the AM. The figure in (c) shows a partially mapped version of the COLD database with evaluation trajectories traversing uncharted areas.

available (named **COLD** and **R@VH**) and an additional one collected by ourselves (named **MAPIR-LABS**). All of them contain several loops and revisitations of parts of the environment, and include the challenges described below:

- *The COsy Localization Database* (**COLD**) (Pronobis and Caputo 2009) (concretely the sub-dataset *Freiburg-partA*) includes images and odometry readings from a robot navigating a real-world office at 5Hz, where sequences roughly follow two different routes (depicted in Fig. 4a): a standard one (*std.*), consisting of a smaller set of rooms; and an extended one (*ext.*), which visits the whole environment. Each type of route always follows the same order in which the rooms were visited. The AM was created from four of the sequences recorded under *cloudy* conditions (two standard and two extended), containing a total of $\sim$ 10k images of the environment under similar lighting conditions, accounting for slight viewpoint variations (see Fig. 4a). The evaluation for this dataset uses a standard and an extended sequence under all possible appearance settings, namely: *cloudy* (different from those sequences employed to create the map), *night* and *sunny*, presenting challenging appearance variations w.r.t. the map.
- The synthetic *Robot at Virtual Home* (**R@VH**) (Fernandez-Chaves et al. 2022) provides realistic simulations of robot navigations within $\sim$ 30 different houses. We simulated the *House21* subdataset, which covers 16m $\times$ 8m and has 8 different rooms connected by a corridor, as shown in Fig. 4b. We gathered images and poses with a simulated robot under the same lighting conditions at 33Hz, and we artificially generated the odometry for this dataset, adding zero-mean Gaussian noise with $\sigma = (0.05\text{m},\ 2.50^\circ)$ to the ground truth poses. In this case, the AM consisted of a sequence where the simulated robot visited each room more than once without following a particular order, resulting in $>$ 50k images. Two short evaluation sequences were provided for this dataset: the first following a path similar to that used for the map (*Sim.*), and another one where the virtual robot was driven manually within the environment, following a different order (*Diff.*).
- The third dataset, called **MAPIR-LABS**, was collected by ourselves, teleoperating a Giraff robot (Gonzalez-Jimenez et al. 2012) in three different laboratories and a corridor that connects them in the School of Computer Science and Engineering of the University of Malaga. The ground-truth for the sequences was obtained with a Graph-SLAM implementation (Grisetti et al. 2010) over the recorded laser scans, while the odometry readings were obtained using RF2O (Jaimez et al. 2016). The map consisted of three sequences visiting a 30m $\times$ 15m environment following different paths (as depicted in Fig. 4d), containing more than 150k images captured at 33Hz. Two evaluation sequences were also recorded: one visiting

the whole environment (*3-labs*) under the same radiometric conditions as the map but following a different path, and another one visiting only a room (*1-lab*) which presents dynamic changes (moved furniture, new objects, etc.).

The GAMs for each dataset were generated off-line, before addressing the localization process. According to the DB Index, the number of regions was set to $M = \{35, 40, 65\}$ for the **COLD**, **R@VH** and **MAPIR-LABS** datasets, respectively.

Parameter setup

For the experiments, we have chosen the following parameter settings:

- The GAMs were built employing samples from sequences captured under the same lighting conditions, which leads to underestimating the variance of the projected descriptors when captured in different circumstances. Consequently, we model the possible impact caused by such different lighting conditions as noise, modifying the covariance matrix of the estimated Gaussian distribution of the projected descriptor ${}^{\mathrm{M}}\Sigma_j^{\delta\delta}$ in Eq. 6. In particular, we multiply it with a matrix whose diagonal entries are $\sigma_{ii} = 1.5$ and whose off-diagonal entries are $\sigma_{ij|i\neq j} = 0.5$.
- The covariance matrix of the odometry noise in Eq. 12 has been set to $diag(\Sigma^{\mathbf{u}}) = \left[0.025^2, 0.01^2, 0.05^2\right]$.
- The time window for detection is $w_{loss} = 3$s, and the loss-of-tracking detection threshold is $\tau_\chi = 99\%$.
- The computational time for each analyzed method was measured in an Intel Core i7-6700K computer with 16-GB RAM, using Python and the NumPy library.
- For every experiment we depict the average results after 5 runs.

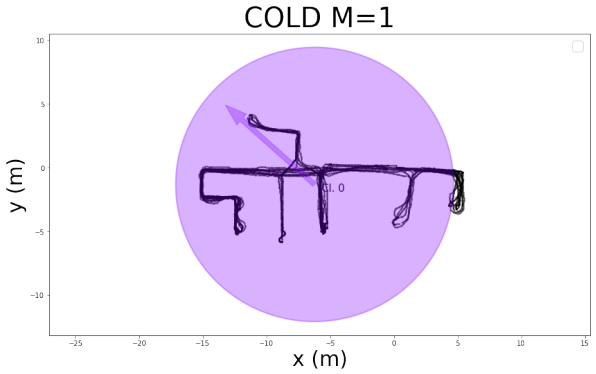
Experimental results

Here we assess the performance of our proposed system and compare it against other SOTA methods.

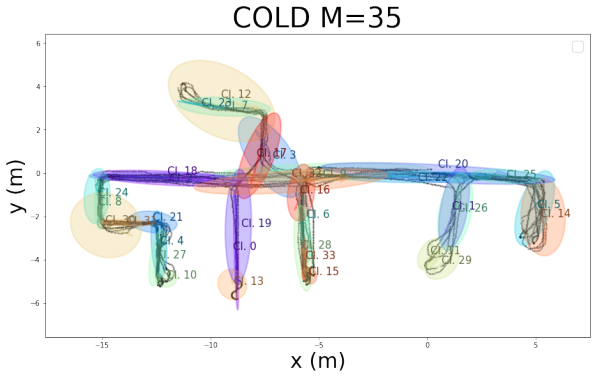
We first conduct a proof of concept to evaluate the benefits of modeling the local correlation between pose and appearance, a capital assumption in this work. Secondly, we explore two essential aspects of our method: the influence of the holistic descriptor chosen to represent the appearance, and the advantages of using local observation models against a global one. Then, we evaluate ALLOM's topological and metric accuracy, as well as its processing time, in comparison to other SOTA approaches for AbL. Finally, we analyze the capability of our proposal to achieve fast and precise relocalization.

A proof-of-concept on local vs. global pose-appearance correlation

This work is based on exploiting the correlation between the appearance descriptor and the pose. We claim that such correlation must be modeled locally, as a global correlation may not exist given the complexity of the appearance descriptor manifold. Intuitively, we can not assume a reliable correlation that relates any scene appearance descriptor to a pose. Thinking locally, however, one can expect that increments of the camera pose bring increments of appearance. This is the principle of our local observation model. To support this intuition, we have designed a simple proof of concept using two maps reflecting extreme scenarios: (i) a ${}^{\mathrm{TM}}$GAM with a single region $M = 1$ (see Fig. 5a) that models the pose-appearance correlation globally; and (ii) a ${}^{\mathrm{TM}}$GAM with $M = 35$ regions (see Fig. 5b).



(a) GAM with $M = 1$ region



(b) GAM with $M = 35$ regions

Figure 5. Two GAMs built from the *COLD Database* (see AM in Fig. 4a) with different number of regions.

Table 2. Results of ALLOM for maps with different number of regions on the **COLD** dataset.

M	AUC (%)	ATE (m)	MTE (m)
1	100.00	6.3493	5.5827
35	91.70	0.4974	0.2015

The results shown in Table 2 reflect the topological and metric accuracy of ALLOM with a number of particles $N_p = 100$ and using $D' = 128$ as the dimension of the projected descriptors, over the *Std. Cloudy* sequence on each GAM. The values of these parameters have been chosen empirically, as they provide an appropriate balance between accuracy and computational burden.

As can be seen, for the one-region map, even though the robot is always topologically localized within the environment (obviously, as there is just one region), the computed correlation between pose and appearance is completely uninformative, resulting in a very poor metric localization. In contrast, using multiple regions allows us

to establish local correlation models between pose and appearance, ultimately leading to a representation of the environment more suited for metric localization.

Influence of the appearance descriptor

One of the capital elements of AbL is the choice of the holistic descriptor employed to represent appearance, as it heavily influences its performance. Since ALLOM is not attached to any particular appearance representation and can be run with any of the existing descriptors, here we carry out some experiments using three of the most popular holistic descriptors and analyze its performance: i) NetVLAD (Arandjelovic et al. 2016), a 4096-sized Deep Learning descriptor, ii) ImRet (Radenović et al. 2019), the 2048-sized Resnet-101 Generalized Mean (GeM) descriptor, and iii) and an additional Bag of Words (BoW) descriptor (Gálvez-López and Tardós 2012) built from ORB features, with the vocabulary trained in (Mur-Artal and Tardós 2017) accumulated into 1024 bins.

Table 3. Accuracy of our proposal for different descriptors on the **COLD** dataset.

	AUC (%)	ATE (m)	MTE (m)
ALLOM + ORB-BoW	32.68	9.1335	1.0264
ALLOM + NetVLAD	36.01	4.8728	0.7550
ALLOM + ImRet	**91.70**	**0.4974**	**0.2015**

The results depicted in Table 3 were also obtained with $N_p = 100$ and using $D' = 128$ projected dimensions, and clearly indicate that ImRet is a descriptor more suited for our approach than NetVLAD and BoW. The results illustrate the failure of the latter descriptors to meet one of the main assumptions of this work: smooth descriptor variation with respect to the pose. In fact, NetVLAD and BoW present excessive variability even for close samples compared to ImRet. Besides, this variability lacks any proportionality to the pose, which becomes problematic when assuming smoothness across a spatial region such as the GAMs. This is an interesting issue that deserves to be addressed in more detail in future work, but, for the rest of the experiments and taking into account the empirical results, ImRet has been chosen as the appearance descriptor.

Local vs. global observation model

Our work, which advocates defining different local observation models for each region, postulates against those methods that build a single observation model for the entire environment.

In the following, we evaluate our proposal in comparison to a global-observation-model-based method in terms of metric accuracy as well as memory and computational cost. For that, we have chosen the Gaussian Process Particle Filter (GPPF) proposed in the series of works (Lopez-Antequera et al. 2016, 2017b), as they define a single Gaussian Process that operates indistinctly in all parts of the map. The main reason behind selecting this approach is that, as the $^{\text{TM}}$GAM in which our proposal is based, it carries out AbL on unordered sets of data points, which makes the two methods directly comparable.

Fig. 6 shows the localization error (using the ATE) and the computational time (seconds per step) for different number of particles incurred by the GPPF technique operating on the dense AM, and ALLOM over $^{\text{TM}}$GAM obtained from the same AM. Error bands representing the standard deviation are also shown for each method. As can be seen, our proposal reaches the accuracy of the GPPF for descriptor dimensions greater than 100 while requiring much less computational time. Notice that the ATE does not decrease significantly for $D' > 100$, indicating that, by applying PCA for each region, we obtain a very compact descriptor that reflects most of the pose information for that region. According to these results, ALLOM obtains a similar ATE than GPPF but requires one order of magnitude less time, thanks to the abstraction provided by the GAMs.

In addition, Table 4 shows that our proposal requires more lightweight maps, demanding at least $50\times$ less memory space. The size of the dense AM is calculated as the sum of the sizes of each appearance descriptor as $O(ND)$, with N being the number of samples and D the dimension of each descriptor. In turn, the size of the $^{\text{TM}}$GAM is independent of the original number of AM samples (reflected by the size of $\mathcal{AM}$), but depends on the number of regions M and the size of the projected descriptor ($D' = 128$ in this case), as $\sim O(MD + MD' + MD'^2)$, being M the number of regions. Note that, in both cases, the space occupied by the poses is disregarded, as it is negligible in comparison.

Table 4. Size of the Topometric Appearance Map $^{\text{TM}}$GAM (in MB) for different size of the ImRet descriptor.

	COLD	**R@VH**	**MAPIR-LABS**
Dense AM, $\mathcal{AM}$	195.31	259.26	1156.80
$^{\text{TM}}$GAM $D' = 10$	0.58	0.68	1.11
$^{\text{TM}}$GAM $D' = 50$	1.26	1.50	2.44
$^{\text{TM}}$GAM $D' = 100$	3.21	3.89	6.33
$^{\text{TM}}$GAM $D' = 128$	4.60	5.90	9.59
$^{\text{TM}}$GAM $D' = 200$	10.16	13.11	21.55

Topological localization

Given the topology-assisted nature of our proposal, it is of interest to evaluate the correctness of the estimated topological path, that is, the sequence of traversed regions, as it has a direct impact on the metric results.

Table 5 shows a comparison between the topological localization accuracy of three different approaches: (i) pure VPR over the $^{\text{T}}$GAM (i.e. to retrieve the topological region that maximizes the Gaussian likelihood of a query descriptor in appearance terms, as formulated in Eq. 19), using the best and the three best matches (shown as Top-1 and Top-3 in the table, respectively), (ii) the topological Bayesian filter proposed in our previous work (Jaenal et al. 2022), which makes use of a simple transition model between regions based on the pose distance between them (for more information, please refer to the paper), and (iii) ALLOM, again with $N_p = 100$ particles, and a projected descriptor with $D' = 128$ dimensions.

The results show that our method clearly outperforms the topological accuracy of the other approaches in most

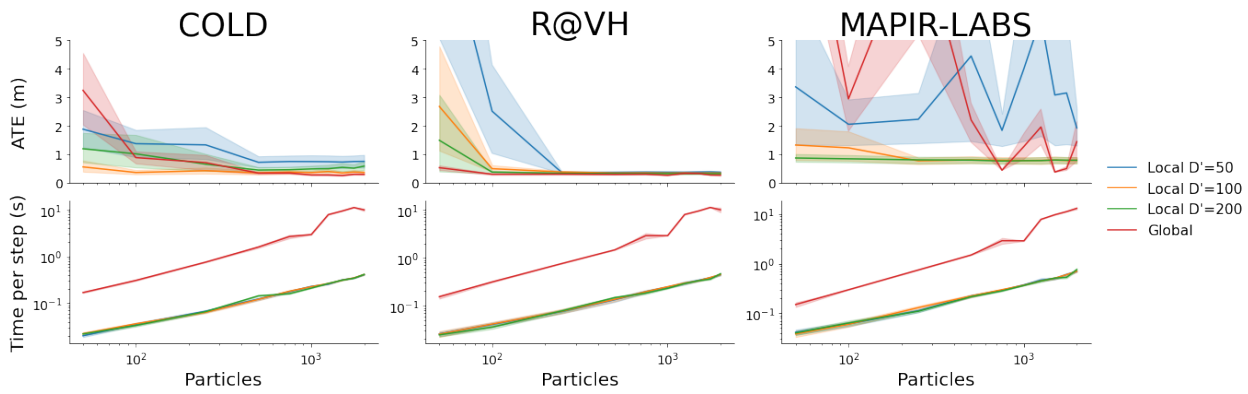


Figure 6. Absolute Trajectory Error (ATE) and computational time using the ImRet descriptor for GPPF (Lopez-Antequera et al. 2017b) (Global Observation Model) and our proposal (Local) for different number of particles.

Table 5. Precision of the topological localization over the Gaussian Appearance Map ™GAM measured through the AUC in %. Three different methods are compared: pure VPR, a topological sequential VPR filter and the topological part of ALLOM.

		VPR (Top-1)	VPR (Top-3)	Topol. Jaenal et al. (2022)	ALLOM
COLD	*Std. Cloudy*	86.20	94.78	64.98	**99.79**
	Ext. Cloudy	78.15	86.43	56.42	**91.16**
	Std. Night	75.75	84.27	51.23	**99.43**
	Ext. Night	72.99	81.33	43.79	**94.77**
	Std. Sunny	72.80	80.81	45.11	**90.21**
	Ext. Sunny	63.41	**73.55**	34.06	72.80
R@VH	*Sim.*	43.90	64.17	40.12	**81.98**
	Diff.	44.07	57.86	27.13	**67.19**
MAPIR -Labs	*1-lab*	82.96	92.54	43.05	**92.42**
	3-labs	74.44	86.41	79.50	**87.37**

scenarios, although pure VPR also obtains very reliable estimations. This seems to demonstrate that the abstracted maps provide a description of the environment appearance sufficiently adequate for VPR. The poor performance of the topological filter for VPR is explained by the limitations of the topological transition model in which it relies. The topological nature of this recursive approach makes it unable to incorporate metric information such as robot odometry, which makes it sensitive to observation noise and/or Perceptual Aliasing. In addition, it is also incapable of recovering after a loss-of-tracking, consequently obtaining reduced tracking performance in the long-term. These shortcomings were one of the main motivations for the development of this proposal.

Metric localization

To complete the localization evaluation, we now analyze the results between ALLOM and other SOTA AbL methods in terms of both topological and metric accuracy (through AUC, and ATE-MTE, respectively), and processing time (seconds per step).

We have compared three different techniques:

- *VPR-Reloc.* (Vysotska and Stachniss 2017), a graph-based VPR tool that matches sequences following different routes with partial overlap.
- *Image-Nav.* (Thoma et al. 2019), a flow network for localization.
- *MCL* (Xu et al. 2020), a Monte Carlo-based Localization method that employs odometry readings.

It should be noted in advance that all these approaches are designed to address localization on AMs based on a single sequence, usually subsampled for efficiency. Recall that, in contrast, ALLOM works on a probabilistic ™GAM that results from abstracting a dense, unordered set of georeferenced images, built from individual samples without any sequential information. Thus, to allow comparison, we have run these SOTA methods using as a database each of the sequences employed to create the ™GAM, and then averaging their final results. Furthermore, as they require subsampling to improve efficiency, we measured their performance using maps with an amount of samples similar to the evaluation conditions they follow in their respective proposals. The parameters of each method have been left as the authors set by default.

Similarly to the measure defined in equation Eq. 17 for our approach, the criteria to determine if a query image has been correctly localized in topological terms by these techniques turns into assessing whether the distance in pose between the query estimation and the ground truth falls below a certain threshold $\tau = (1\text{m},\ 10°)$. This threshold has been chosen by adapting to indoors the $(5\text{m},\ 10°)$ outdoors threshold proposed in (Sattler et al. 2018).

The localization results for the four techniques can be seen in Table 6, with ALLOM and *MCL* outperforming the other two approaches in all cases. Note that *VPR-Reloc* is a sequence-to-sequence matching proposal, so it is unable to provide metric poses estimations (hence shown as N/E in the table). *Image-Nav.*, on the other hand, performed poorly in most cases and also required excessively long processing times.

In the **COLD** dataset, our proposal demonstrates accuracy comparable to *MCL* on maps with considerably higher sampling rate. It is important to note that, in this dataset, the path of the robot between the evaluation sequences and those employed to build the map follow similar trajectories, a situation which is beneficial to single-sequence approaches. On the contrary, since the evaluation trajectories from **MAPIR-LABS** and **R@VH** datasets cover the environment following different paths than their mapping sequences,

Table 6. Compared ATE and MTE in meters, topological performance (through AUC) and time per step (p/s) in seconds of different methods: *VPR-Reloc* Vysotska and Stachniss (2017), *Image-Nav.* Thoma et al. (2019) and *MCL* Xu et al. (2020). N/E means not estimated.

	COLD					R@VH					MAPIR-LABS				
	M	AUC (%)	ATE (m)	MTE (m)	Time p/s	M	AUC (%)	ATE (m)	MTE (m)	Time p/s	M	AUC (%)	ATE (m)	MTE (m)	Time p/s
VPR-Reloc. + ImRet	250	34.15	N/E	N/E	0.0036	400	32.51	N/E	N/E	0.0066	500	21.48	N/E	N/E	0.0071
VPR-Reloc. + NetVLAD	250	35.64	N/E	N/E	0.0036	400	41.39	N/E	N/E	0.0065	500	36.86	N/E	N/E	0.0075
Image-Nav. + ImRet	250	38.46	3.8744	1.3149	3.7974	400	9.23	4.1191	1.6252	7.6912	500	14.53	5.9410	2.1751	15.9024
Image-Nav. + NetVLAD	250	42.72	2.7467	0.9574	3.7314	400	15.13	3.7591	1.2636	8.6912	500	22.02	4.6941	1.8753	16.0089
MCL + ImRet	250	87.29	0.5283	0.2273	0.0506	400	76.04	2.0171	1.1342	0.0642	500	39.81	8.7259	6.4464	0.0642
MCL + NetVLAD	250	87.06	0.5055	0.2070	0.0506	400	84.67	0.8018	0.4449	0.0516	500	58.84	4.2320	1.4241	0.0532
ALLOM + ImRet	35	91.70	0.4974	0.2015	0.0325	40	74.58	0.8797	0.3515	0.0369	65	89.89	0.5952	0.4737	0.0534
ALLOM + NetVLAD	35	42.02	4.3749	0.6418	0.0422	40	46.84	5.3729	0.8179	0.0401	65	27.96	6.4199	1.0457	0.0572

ALLOM significantly outperforms the metric accuracy of *MCL* as this is unable to generalize the environment beyond the mapping sequence. In this case, ALLOM obtains superior performance, demonstrating that the topological knowledge based on the spatial structure of the environment in which ™GAM grounds provides better understanding than the prefixed route from single-sequence maps. This improves the metric accuracy for Appearance-based Localization.

Relocalization

Finally, to test the relocalization capability of our proposal, we have designed a set of challenging scenarios in which the robot loses its track, namely:

- *Camera failure:* The robot keeps moving while the camera does not capture images (specifically, the image turns black). This situation lasts 5-10 seconds.
- *Slippery surface:* The robot remains static during 5-10 seconds while providing unrealistic odometry reading and receiving a still image from the camera.
- *Kidnapped robot:* We introduce a jump at some point in the robot's trajectory and place it at an arbitrary location, without the odometry reflecting such jump.
- *Unseen places:* The robot aims to localize in a map that only covers the scene partially (see Fig. 4c) , while the evaluation sequence visits rooms that are not covered by it.

To evaluate all these scenarios, we used the *Ext. Cloudy* from the **COLD** dataset. For the first three cases, we randomly generated 25 different sequences of ~ 10m length, which begins with a ~ 5m stretch where the robot must initialize and keep track of the path; then, the specific challenge takes place, and finally the robot continues following a path of the same length in a normal manner. In the *Unseen Places* case, the evaluation consists of sequence excerpts beginning in the map, then leaving and re-entering it, traversing uncharted rooms for more than 45 seconds.

After losing track, we consider the filter to be relocalized when more than 50% of the particles are topologically localized again, i.e. on the correct region. We measure the relocalization capability of ALLOM through three parameters: (i) whether relocalization was successful, by checking if the correct region could be detected after the event, (ii) the elapsed time between the occurrence of the event and the moment when ALLOM detects the loss-of-tracking, and (iii) the elapsed time between the occurrence of the event and the relocalization. Table 7 shows the results of the experiment yielded by our proposal with $N_p = 100$ and $D' = 128$ for each of the four situations.

Table 7. Relocalization performance of ALLOM in terms of success ratio (%), time to detect the loss (seconds, with mean and std. deviation) and relocalization time (seconds, with mean and std. deviation)

	Success (%)	Loss detection (s)	Reloc. time (s)
Camera failure	94.79	2.63 ± 1.20	2.98 ± 1.64
Slippery surface	88.89	2.74 ± 1.08	5.72 ± 4.63
Kidnapped robot	70.96	2.44 ± 1.49	10.75 ± 7.42
Unseen places	100.00	3.16 ± 3.01	4.16 ± 5.84

As can be seen, ALLOM can relocalize rapidly in the two first cases. In the camera failure scenario, relocalization is almost immediate upon the loss-of-tracking detection, indicating that the particles are quickly re-associated back to the correct region. In turn, the second scenario introduces more inconsistency between the observation and the odometry, so relocalization requires an average additional time of ~ 3 seconds. The kidnapped robot scenario, in turn, represents a bigger problem that poses more difficult challenges. However, ALLOM is still able to relocalize in more than 70% of the cases, although requiring considerable extra time to reach a successful relocalization. Finally, in the case of traversing places that are not covered by the map, ALLOM demonstrates its ability to promptly recognize that the robot is leaving a mapped area. Then, while traversing these off-map areas, ALLOM continuously tries to perform relocalization without success but, finally, it is able to regain tracking when the robot re-enters a mapped zone in ~ 4s.

Conclusions and future work

We have presented ALLOM, a method for Appearance-based robot Localization that provides consistent and reliable topology-assisted metric pose estimation at indoors where dense Appearance Maps are available.

First, we have introduced an extension for the topological abstraction provided by the Gaussian Appearance Maps (GAMs) proposed in previous work. Such extension relies on modeling the probabilistic correlation between the pose and a reduced version of the appearance descriptor in local regions of the environment. As a result, we produce enhanced GAMs (™GAM) that enable accurate robot metric localization by leveraging the information about its topological location. This has been addressed through a Particle Filter that resorts to conditional distributions derived

from the TMGAM, which, in turn, serve as observation models that are specifically fitted for each region. ALLOM also resorts to the topological structure of the map in order to detect any robot loss-of-tracking and to apply a VPR-based initialization and relocalization.

Our proposal has demonstrated comparable performance to a method based on global observation models running over the original dense, unordered AM, but with a significant increase in efficiency. Besides, ALLOM is also able to achieve better metric accuracy than current state-of-the-art techniques even under appearance challenges, since the topological knowledge provided by our maps allows us to handle situations where other techniques fail.

In our experiments, we have found several limitations regarding the poor performance of two foundational VPR descriptors such as NetVLAD and ORB-BoW. In this sense, an important subject of study for future works is to design specific appearance descriptors for AbL that are not invariant to point-of-view, so they can provide reliable information for localization. In addition, a promising research line for future work is to study different Dimensionality Reduction techniques, other than PCA, generating reduced descriptors that result in better correlation with the pose.

It is also of interest to extend our method to work outdoors with SE(3) poses, as well as to consider more sparse maps, adapting abstracted AMs to work with non-Gaussian regions.

Funding

This work was supported by the Government of Spain in part under grant FPU17/04512, in part under the *ARPEGGIO* (PID2020-117057GB-I00) research project, and also by the Andalucian Regional Government under the *Houndbot* (PY20_01302) research project.

Acknowledgements

The authors thank the Supercomputing and Bioinnovation Center (SCBI) of the University of Malaga for their provision of computational resources and technical support (www.scbi.uma.es/site); and the support of NVIDIA Corporation for the donation of the Titan X Pascal used in this work.

Finally, the authors want to thank the Department of Computer Architecture of the University of Malaga for the cession of their facilities for gathering data, especially to Francisco M. Castro.

References

Arandjelovic R, Gronat P, Torii A, Pajdla T and Sivic J (2016) NetVLAD: CNN architecture for weakly supervised place recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 5297–5307.

Balntas V, Li S and Prisacariu V (2018) Relocnet: Continuous metric learning relocalisation using neural nets. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 751–767.

Blanco JL, Fernandez-Madrigal JA and Gonzalez J (2008) Toward a unified bayesian approach to hybrid metric–topological slam. *IEEE Transactions on Robotics* 24(2): 259–270.

Brahmbhatt S, Gu J, Kim K, Hays J and Kautz J (2018) Geometry-aware learning of maps for camera localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2616–2625.

Cummins M and Newman P (2008) FAB-MAP: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research* 27(6): 647–665.

Davies DL and Bouldin DW (1979) A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Ding M, Wang Z, Sun J, Shi J and Luo P (2019) Camnet: Coarse-to-fine retrieval for camera re-localization. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 2871–2880.

Doan AD, Latif Y, Chin TJ and Reid I (2020) Hmˆ4: Hidden markov model with memory management for visual place recognition. *IEEE Robotics and Automation Letters* 6(1): 167–174.

Fernandez-Chaves D, Ruiz-Sarmiento JR, Jaenal A, Petkov N and Gonzalez-Jimenez J (2022) Robot@virtualhome, an ecosystem of virtual environments and tools for realistic indoor robotic simulation. *Expert Systems with Applications* : 117970.

Fernández-Madrigal JA and Blanco-Claraco JL (2012) *Simultaneous Localization and Mapping for Mobile Robots: Introduction and Methods: Introduction and Methods*. IGI global.

Gálvez-López D and Tardós JD (2012) Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics* 28(5): 1188–1197.

Gomez-Ojeda R, Moreno FA, Zuñiga-Noël D, Scaramuzza D and Gonzalez-Jimenez J (2019) PL-SLAM: A stereo SLAM system through the combination of points and line segments. *IEEE Transactions on Robotics* 35(3): 734–746.

Gonzalez-Jimenez J, Galindo C and Ruiz-Sarmiento J (2012) Technical improvements of the giraff telepresence robot based on users' evaluation. In: *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, pp. 827–832.

Grisetti G, Kümmerle R, Stachniss C and Burgard W (2010) A tutorial on graph-based slam. *IEEE Intelligent Transportation Systems Magazine* 2(4): 31–43.

Ham J, Lin Y and Lee DD (2005) Learning nonlinear appearance manifolds for robot localization. In: *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, pp. 2971–2976.

Huhle B, Schairer T, Schilling A and Straßer W (2010) Learning to localize with gaussian process regression on omnidirectional image data. In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, pp. 5208–5213.

Jaenal A, Moreno FA and Gonzalez-Jimenez J (2021) Appearance-based sequential robot localization using a patchwise approximation of a descriptor manifold. *Sensors* 21(7): 2483.

Jaenal A, Moreno FA and Gonzalez-Jimenez J (2022) Unsupervised appearance map abstraction for indoor visual place recognition with mobile robots. *IEEE Robotics and Automation Letters* : 1–7.

Jaenal A, Zuñiga-Nöel D, Gomez-Ojeda R and Gonzalez-Jimenez J (2020) Improving visual slam in car-navigated urban environments with appearance maps. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 4679–4685.

Jaimez M, Monroy JG and Gonzalez-Jimenez J (2016) Planar odometry from a radial laser scanner. a range flow-based approach. In: *2016 IEEE International Conference on Robotics*

and Automation (ICRA). IEEE, pp. 4479–4485.

Kendall A, Grimes M and Cipolla R (2015) Posenet: A convolutional network for real-time 6-dof camera relocalization. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2938–2946.

Laskar Z, Melekhov I, Kalia S and Kannala J (2017) Camera relocalization by computing pairwise relative poses using convolutional neural network. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. pp. 929–938.

Lopez-Antequera M, Gomez-Ojeda R, Petkov N and Gonzalez-Jimenez J (2017a) Appearance-invariant place recognition by discriminatively training a convolutional neural network. *Pattern Recognition Letters* 92: 89–95.

Lopez-Antequera M, Petkov N and Gonzalez-Jimenez J (2016) Image-based localization using gaussian processes. In: *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, pp. 1–7.

Lopez-Antequera M, Petkov N and Gonzalez-Jimenez J (2017b) City-scale continuous visual localization. In: *2017 European Conference on Mobile Robots (ECMR)*. IEEE, pp. 1–6.

Lowry S, Sünderhauf N, Newman P, Leonard JJ, Cox D, Corke P and Milford MJ (2015) Visual place recognition: A survey. *IEEE Transactions on Robotics* 32(1): 1–19.

Lynen S, Zeisl B, Aiger D, Bosse M, Hesch J, Pollefeys M, Siegwart R and Sattler T (2020) Large-scale, real-time visual–inertial localization revisited. *The International Journal of Robotics Research* 39(9): 1061–1084.

Maddern W, Milford M and Wyeth G (2012) Cat-slam: probabilistic localisation and mapping using a continuous appearance-based trajectory. *The International Journal of Robotics Research* 31(4): 429–451.

Mazuran M, Boniardi F, Burgard W and Tipaldi GD (2018) Relative topometric localization in globally inconsistent maps. In: *Robotics Research*. Springer, pp. 435–451.

Milford MJ and Wyeth GF (2012) Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In: *International Conference on Robotics and Automation*. pp. 1643–1649.

Mur-Artal R, Montiel JMM and Tardos JD (2015) Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics* 31(5): 1147–1163.

Mur-Artal R and Tardós JD (2017) ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *IEEE Transactions on Robotics* 33(5): 1255–1262.

Naseer T, Burgard W and Stachniss C (2018) Robust visual localization across seasons. *IEEE Transactions on Robotics* 34(2): 289–302.

Naseer T, Spinello L, Burgard W and Stachniss C (2014) Robust visual robot localization across seasons using network flows. In: *AAAI*. pp. 2564–2570.

Pearson K (1901) Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2(11): 559–572.

Piasco N, Sidibe D, Demonceaux C and Gouet-Brunet V (2018) A survey on visual-based localization: On the benefit of heterogeneous data. *Pattern Recognition* 74: 90–109.

Pronobis A and Caputo B (2009) Cold: The cosy localization database. *IJRR* 28(5): 588–594.

Radenović F, Tolias G and Chum O (2019) Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(7).

Rubin DB (1988) Using the sir algorithm to simulate posterior distributions. *Bayesian statistics* 3: 395–402.

Sattler T, Maddern W, Toft C, Torii A, Hammarstrand L, Stenborg E, Safari D, Okutomi M, Pollefeys M, Sivic J et al. (2018) Benchmarking 6dof outdoor visual localization in changing conditions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 8601–8610.

Sattler T, Torii A, Sivic J, Pollefeys M, Taira H, Okutomi M and Pajdla T (2017) Are large-scale 3D models really necessary for accurate visual localization? In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1637–1646.

Sattler T, Zhou Q, Pollefeys M and Leal-Taixe L (2019) Understanding the limitations of cnn-based absolute camera pose regression. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3302–3312.

Schairer T, Huhle B, Vorst P, Schilling A and Straßer W (2011) Visual mapping with uncertainty for correspondence-free localization using gaussian process regression. In: *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, pp. 4229–4235.

Thoma J, Paudel DP, Chhatkuli A, Probst T and Gool LV (2019) Mapping, localization and path planning for image-based navigation using visual features and map. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 7383–7391.

Toft C, Maddern W, Torii A, Hammarstrand L, Stenborg E, Safari D, Okutomi M, Pollefeys M, Sivic J, Pajdla T et al. (2020) Long-term visual localization revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44(4): 2074–2088.

Torii A, Arandjelovic R, Sivic J, Okutomi M and Pajdla T (2015) 24/7 Place Recognition by View Synthesis. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1808–1817.

Vysotska O and Stachniss C (2015) Lazy data association for image sequences matching under substantial appearance changes. *IEEE Robotics and Automation Letters* 1(1): 213–220.

Vysotska O and Stachniss C (2017) Relocalization under substantial appearance changes using hashing. In: *Proceedings of the IROS Workshop on Planning, Perception and Navigation for Intelligent Vehicles, Vancouver, BC, Canada*, volume 24.

Xu M, Fischer T, Sünderhauf N and Milford M (2021) Probabilistic appearance-invariant topometric localization with new place awareness. *IEEE Robotics and Automation Letters* 6(4): 6985–6992.

Xu M, Sünderhauf N and Milford M (2020) Probabilistic visual place recognition for hierarchical localization. *IEEE Robotics and Automation Letters* 6(2): 311–318.