IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 May 2022, accepted 21 June 2022, date of publication 23 June 2022, date of current version 29 June 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3185732

RESEARCH ARTICLE

PlaneLoc2: Indoor Global Localization Using Planar Segments and Passive Stereo Camera

JAN WIETRZYKOWSKI
Institute of Robotics and Machine Intelligence, Poznan University of Technology, 60-965 Poznan, Poland
e-mail: jan.wietrzykowski@put.poznan.pl

This work was supported by the Polish National Science Centre (NCN) under Contract UMO-2018/31/N/ST6/00941.

ABSTRACT This paper introduces PlaneLoc2 - a novel indoor global localization system designed to harness the potential of stereo cameras. A need for robust global localization that does not produce incorrect results (false positives) is present in almost every life-long autonomy task. We show that planar segments extracted from stereo vision data by a neural network enable such robust localization. Planar segments are easier to discriminate than keypoint features and provide easy-to-use geometric constraints. We propose an architecture that exploits a single deep neural network (DNN) to detect planar segments, produce appearance descriptors, and estimate segment geometry. Moreover, we introduce a novel view-based segment map and a novel pose retrieval procedure that considers the uncertainty of features to efficiently use the geometric constraints provided by them. We also show that the new learned descriptor provides better discrimination than the hand-crafted one. Finally, we present experimental results that show that our solution outperforms other state-of-the-art global localization methods and does not produce incorrect agent poses. For both test scenes it recognizes at least 15% more poses than the second best method without incorrect recognitions.

INDEX TERMS Simultaneous localization and mapping, artificial neural networks, stereo image processing.

I. INTRODUCTION

Accuracy of modern simultaneous localization and mapping (SLAM) systems over the last years has improved significantly, yet they are still not applicable to many real-world tasks. The main reason is that to work for a prolonged time these systems have to be able to recover from failures and have to correct localization drift that inevitably accumulates over time. When no external source of positioning is available, e.g. in indoor environments where there is no Global Positioning System (GPS) signal, global localization becomes essential. Global localization is a problem of localizing an agent with respect to a known map without knowledge of its previous poses [1]. In the case of metric global localization, the pose (translation and rotation) is expressed in a frame of reference of the map using appropriate representation, e.g. translation vector and rotation matrix. Metric global localization is a vital component of solutions to problems such as recovery after loosing pose tracking due to occlusion or other external factors, or loop closing when a robot arrives at a previously visited scene after traversing a long loop and drift has to be rectified. In order to compute the pose in this situation, it is necessary to match a selected type of features or objects between a local view and the global map. The more discriminative the features or objects, the better, because it is easier to avoid incorrect associations. However, a nontrivial problem is to reliably and repeatedly detect such objects and to exploit geometric constraints provided by their associations. One possibility is to use planar segments that are common in indoor environments. They are not so easily detected as keypoint features and geometric constraints are more complex than point-to-point constraints, nonetheless, they are more discriminative and there are usually fewer of them, which reduces the number of possible association combinations. Therefore, to build a global localization system that will benefit from planar segments, it is necessary to develop proper detection and pose retrieval algorithms. The detection of planar segments is usually done using RGB-D sensors because of the availability of depth information that helps to segment the scene and enables geometry estimation, i.e. plane equations supporting segments. Unfortunately, RGB-D sensors have limited effective range, and other sensors providing depth information, such as LiDARs (Light Detection and Ranging), are expensive. An interesting alternative is a pas-

The associate editor coordinating the review of this manuscript and approving it for publication was Thomas Canhao Xu.

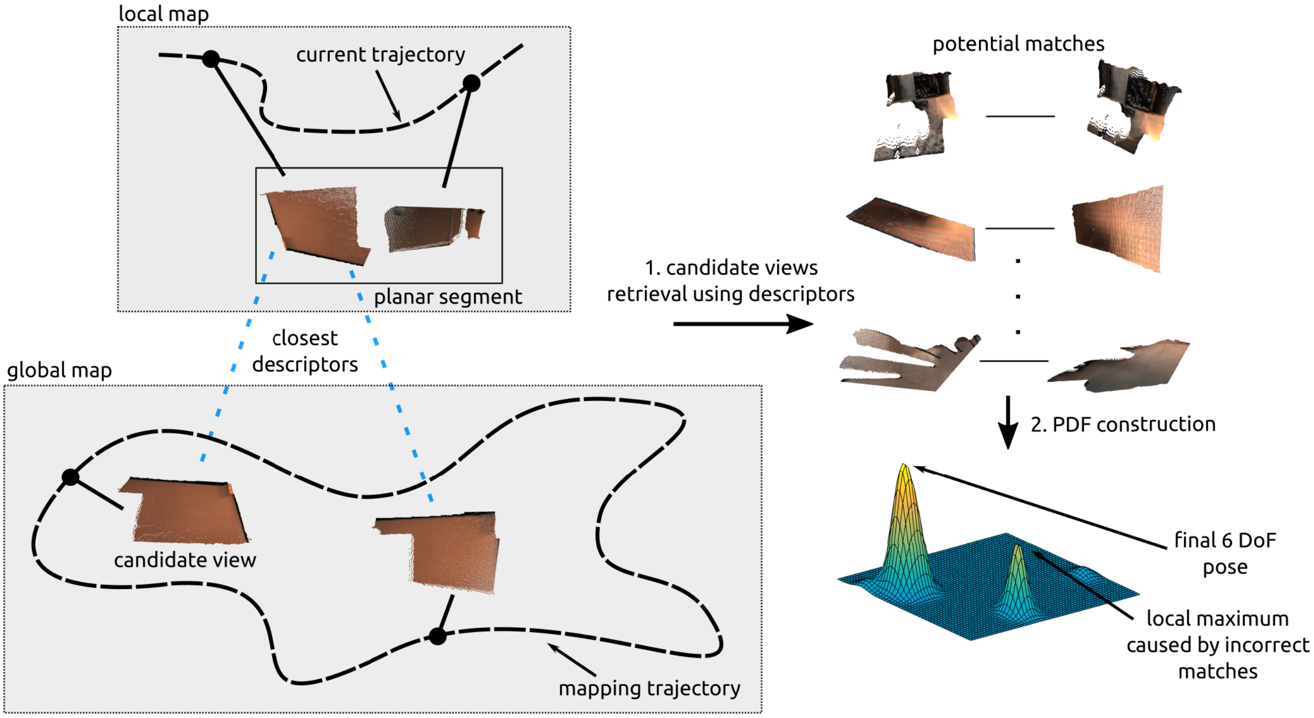

FIGURE 1. PlaneLoc2 retrieves candidate views using appearance descriptors and builds a PDF of pose using all potential matches. The final pose is a maximum of the PDF, verified by the fail-safe checks.

sive stereo camera that also facilitates unambiguous geometry recovery, but has a longer effective range than RGB-D sensors and is cheaper than LiDARs. However, to harness the full potential of stereo cameras, special care has to be taken because stereo estimated depth is not as accurate as the one from RGB-D sensors or LiDARs. Whereas multiple papers discuss planar segment detection without explicit depth information [2], localization using planar segments [3], and some systems allow localization using stereo sensors [4], no significant prior work exists that combines those topics to propose a robust global localization system. This paper closes this gap by introducing the PlaneLoc2 (Sec. III), depicted in Fig. 1. The goal of the presented research is to develop a system that delivers a metric pose of the agent with respect to a known map, using a passive stereo camera, and exploiting planar segments as reference objects. The contribution of the paper can be summarized as follows[1]:

- Extending Stereo Plane R-CNN planar segment detection network with a module to extract the geometry and uncertainty of geometry of planar segments. This enables application of this network architecture to the real-world problem of global localization (Sec. IV-A).
- Developing a planar segment appearance description method that is embedded in the segment detection network. The enhanced descriptor significantly limits the number of potential matches considered during localization (Sec. IV-B).
- Proposing a novel view-based map and a novel pose retrieval method that better suit the characteristic of passive stereo cameras (Sec. V).

The rest of the article is structured as follows.

In Sec. II we survey other papers and compare them with our approach. Sec. III is dedicated to the overview of the global localization pipeline. In Sec. IV we describe the planar segment extraction mechanism, while the view-based approach to global localization is presented in Sec. V. The proposed methods are extensively evaluated and compared to other state-of-the-art systems in Sec. VI. Finally, conclusions are drawn in Sec. VII.

This work builds on results from our previous articles. A planar segment detection DNN that enables accurate geometry retrieval was introduced in [5]. We use this network in the PlaneLoc2, but add a segment geometry extraction mechanism that can be used in global localization. The extracted information include the uncertainty that is a vital part of the description of geometry. The segment appearance description learning is inspired by our previous successful loop closing method [6], where descriptors of general (not necessarily planar) segments were computed from LiDAR data. The general idea of inference by building a probability density function (PDF) describing agent pose is borrowed from the PlaneLoc system that uses RGB-D data [7]. However, a completely new mapping approach and pose retrieval procedure are introduced in this article to handle a stereo sensor.

[1] Implementation and dataset are available at https://github.com/LRMPUT/plane_loc_2

II. RELATED WORK

In this section we describe other papers related with our work. The description is divided into three subsections concerning different aspects of global localization: sensors, features, and methods in general.

A. SENSORS

Rapid development of RGB-D sensors that followed the introduction of Kinect, brought a variety of sensors that use different measurement techniques, such as structured light, time of flight (ToF), and active stereo. However, all those solutions have a limited effective range of 4-6 m [8], even Kinect v2 that is especially vulnerable to reflective surfaces [8]. Therefore, modern RGB-D sensors often resort to passive stereo for larger distances, which increases the effective range [9]. The limited range poses problems for many real-world applications and makes a stereo camera the preferable sensor. The applications include, but are not limited to, tracking human motion [10], SLAM [11], and scene reconstruction [12]. Moreover, depth information is sometimes used to simulate a view-based stereo measurement to achieve better results [4]. Also, when significant scene sizes are considered, stereo is the only viable option [13] with monocular cameras struggling with scale ambiguity [14]. Aware of those results in related areas, we resort to a passive stereo camera to increase the effective range of perception of planar segments with respect to our earlier PlaneLoc system from [7].

B. FEATURES IN GLOBAL LOCALIZATION

One of the key aspects of global localization is a choice of features to be matched. Algorithms like DBoW2 [15], used in ORB-SLAM3, resort to classical, non-learned keypoint features, such as BRIEF (Binary Robust Independent Elementary Features) or ORB (ORiented FAST and Rotated BRIEF). A more recent approach is to use a trained keypoint detector and descriptor, as in [16] where finding dense pixel-wise correspondences between two images is enabled by a pyramid of coarse-to-fine features. Learned features are oftentimes combined with learned matching methods, such as SuperPoint detector and descriptor [17] and SuperGlue [18] matcher that uses a graph neural network to aggregate global context. A more localization-oriented feature learning was proposed in [19], where supervision at the level of pose was applied to train a multiscale feature generator. However, the pose estimation is left to a principled algorithm and the method requires a coarse initialization of pose, therefore being not suitable for global localization. In our work, we adopt a different approach and instead of resorting to a complex description and matching methods, we use planar segments that are easier to describe and match.

Planar segments are not as commonly used as reference objects, compared to keypoint features, mainly because difficulties with their detection, and with exploiting the geometric constraints they provide. Nonetheless, there are SLAM systems that use planar segments, such as the one presented in [20], where planar segments enabled loop closures in a LiDAR-based system. LiDAR measurements facilitates accurate estimation of planar segments' geometry, therefore the solution cannot be directly applied to a camera-based system, such as ours. In camera-based SLAM, planar segments were used in [3], [21], however, planar constraints were used only during incremental localization and loop closing was based on keypoint features. Contrarily, PlaneLoc2 uses planar segments to recover global pose, which is a part of loop closing procedure. A demonstration of global registration of camera pose with planar segments was presented in [22], but no quantitative localization results were provided. Global localization was also considered in [23], where graphs of incidence of planar segments were used to compare their sets. However, the method was tested only in a small environment, where objects were close to a sensor and their geometry could be accurately estimated using RGB-D data. In opposition, in this paper, we quantitatively evaluate the proposed solution in a workshop-sized environment to enable a fair comparison with other systems.

C. GLOBAL LOCALIZATION METHODS

Most of the global localization methods use associations between keypoint features to recover a pose. Loop closing and relocalization mechanisms in ORB-SLAM3 [4], based on DBoW2 [15], use sparse ORB features and hierarchical tree to quickly retrieve candidate images to match against. The pose is computed by point-to-point correspondences and later verified by tracking a local map. A solution using learned descriptors is presented in [24], where candidate images are found using NetVLAD [25] descriptor, followed by dense matching and pose verification using view synthesis. Unfortunately, view synthesis requires the database images to contain dense depth maps, which can be troublesome to obtain. A conceptually similar approach was described by Sarlin *et. al* [26], where localization is done in two steps: global candidate images retrieval, followed by local feature matching. Our solution follows a different strategy than those algorithms, matching directly objects of reference and including context description in the appearance descriptor of those objects. A data-driven approach could alleviate the need to choose a specific strategy and combine benefits of both solution. However, despite the enormous capabilities of DNNs, they have been applied mainly to feature generation and incremental localization [27], whereas global pose retrieval is done using principled algorithms, as in the aforementioned papers.

Uncertainty in global localization is not easy to capture and has been discussed only in a few articles. In [28] a place recognition method was proposed that uses Bayesian filtering with simple motion and sensor models. The model is used in prediction and resampling steps of a particle filter, but the computed place gives only a coarse pose. Another example of Bayesian localization is presented in [29], where authors integrated LiDAR and camera measurements and proposed an efficient inference method with a decomposition of the

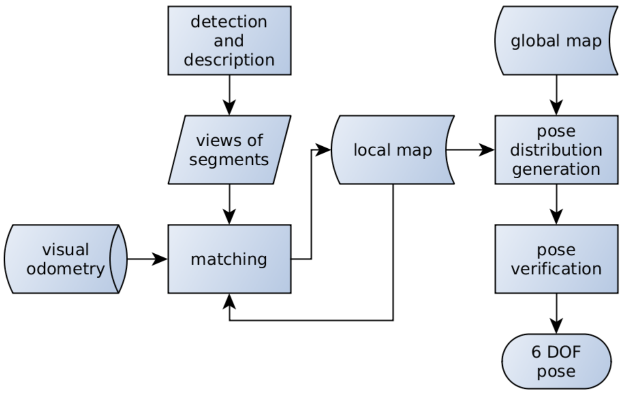

FIGURE 2. **Processing pipeline of PlaneLoc2.**

global map into local places. Those two methods maintain a probability distribution of poses and constrain transitions between locations using a motion model. Such an approach differs from the one presented in this article, because we assume that visual odometry in a short horizon is precise enough to neglect its uncertainty and represent the pose distribution using kernels.

III. GLOBAL LOCALIZATION USING PLANAR SEGMENTS

The RGB-D based PlaneLoc, despite achieving good results in terms of precision and recall, had a few issues that were identified during the research and hindered further development:

- Ignoring planar segments further than 4 m due to a limited effective range of RGB-D sensors. During global localization, using only the part of the image that is close to the sensor significantly limits the context and limits the number of geometrical constraints.
- Using poorly discriminating appearance descriptors based on color histograms. They were dependent on illumination and did not include context, therefore their comparison produced many spurious potential matches.
- Using pose retrieval optimization based on infinite planes. It did not include information about the boundaries of planar segments and produced implausible solutions that had to be additionally verified.

In the new approach, PlaneLoc2, the above-mentioned issues were addressed to improve robustness and recall. Nonetheless, the inference procedure is based on the previous version [7] in which all plausible pose hypotheses are generated and a PDF representing knowledge about the pose is built. In the PDF the maximum is sought and additional asserts are performed to ensure that the returned pose is correct. The same idea is applied here, although most of the other components had to be redesigned to benefit from a stereo sensor.

The processing pipeline (see Fig. 2) starts with planar segments detection and description using a DNN. To maximize computation sharing during this stage, we use a single DNN that extracts all information necessary for further processing, including segments' 3-D geometry. The geometry and visual odometry are used to match segments from the current frame to those present in the local map. Information from the current frame is then used to either update segments in the local map or to add new ones, depending on the matching results. Both maps, the local and the global one, do not merge segments explicitly to get a single representation but rather store information about views of the segments. After updating the local map, a localization procedure is performed, that associates segments between the local and the global map and builds the PDF. The procedure starts with the retrieval of candidate global map views using appearance descriptors. As a result of using deep learned descriptors that provide good discrimination, only 2 candidate views have to be retrieved to get a high probability of including a correct match. Using retrieved views, all plausible pose hypotheses are generated by examining triplets of matched segments and every hypothesis is inserted into the 6-D pose PDF as a kernel:

$$p(\mathbf{q}) = \frac{1}{Z}\tilde{p}(\mathbf{q}) = \frac{1}{Z}\sum_{a} w_a K_a(\mathbf{q}), \tag{1}$$

where $\mathbf{q}$ is a 6 element pose vector (a logarithm of the $SE(3)$ transformation matrix), Z is a normalizing factor, K_a is a kernel function for hypothesis a, and w_a is a weight of the kernel a. The weights are computed as follows:

$$w_a = \sum_{b} \alpha_b, \tag{2}$$

where b ranges over all local segment views used in the hypothesis a, and α_b is an area of the segment view b. A novel procedure to retrieve a pose hypothesis based on a set of matches is used to exploit view-based representation and provide as many geometric constraints as possible. The pose retrieval procedure is critical during the pose hypothesis generation and the final pose computation. When the PDF maximum is found and the final pose $\mathbf{q}^*$ is computed, three fail-safe checks are performed to ensure that the pose is correct:

- $\tilde{p}(\mathbf{q}^*) > \tau_p$ - the value of the unnormalized PDF $\tilde{p}(\mathbf{q}^*)$ for the final pose $\mathbf{q}^*$ has to be above a threshold τ_p to assert that enough positive evidence was collected.
- $\min\left(\frac{\alpha_m^l}{\alpha_t^l}, \frac{\alpha_m^g}{\alpha_t^g}\right) > \tau_r$ - the ratio of the area of segment views that were matched α_m to the total area of visible segment views α_t has to be above a threshold τ_r for, both, the local map (denoted by a superscript l) and the global map (denoted by a superscript g) to verify that there is no significant amount of negative evidence.
- $|\mathcal{M}| > \tau_d$ - the number of distinct matched pairs of segments has to be above a threshold to make sure that the positive evidence is diverse enough.

Our system has three main threads that can be executed concurrently. The first one is responsible for detecting planar segments and creating views – its processing takes 557 ms per frame on average on RTX 3090 GPU. The second one builds

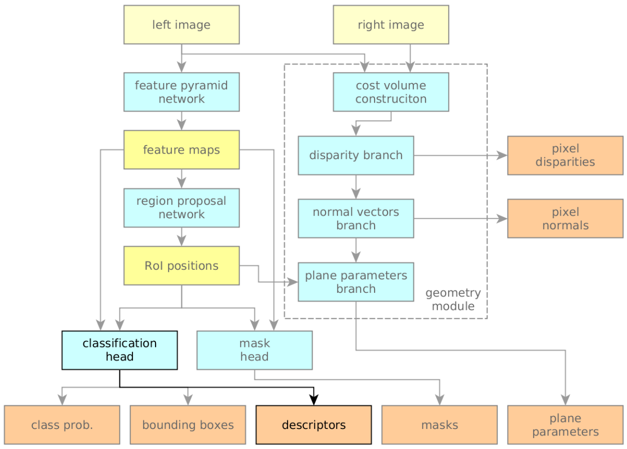

FIGURE 3. **An overview of DNN used to detect and describe planar segments. Gray blocks and connections were not modified.**

and manages the local map, using approximately 11 ms of i5-8250U CPU time for each frame. The last thread is a pose inference thread that returns results every 2883 ms on average using CPU only. The execution time allows to update the local map with a frequency of approximately 2 Hz, which is enough for global localization, since consecutive frames usually do not contain significant amount of new information. Although the local map can be updated with a frequency of 2 Hz, the agent pose cannot be retrieved after each update due to the longer processing time of the inference thread. Nonetheless, in the considered scenarios, information about the global pose yielded every 3 s is enough to recover from loosing pose tracking or to correct the drift.

IV. PLANAR SEGMENTS EXTRACTION

As mentioned in Sec. I, reliable and repeatable object detection is essential if they are to be used during localization. Drawing from development in the object detection field, where DNNs achieve the best results, outperforming classical methods by a large margin, we also use DNN to detect reference objects in the form of planar segments. The DNN, introduced in our recent work [5] (see Fig. 3), simultaneously produces image masks of individual planar segments, their appearance descriptors used to preliminarily match segments between the local map and the global map, and retrieves the 3-D geometry of the segments. It was trained on a photo-realistic synthetic `SceneNet Stereo` dataset containing approximately 35k images from 200 different scenes. The training was started from weights pretrained on the real-world `Coco` and `ScanNet` datasets, same as in [5]. We trained the network for 10 epochs using Adam optimizer with a learning rate equal to 10^{-5} and weight decay equal to 10^{-4}. Training examples were augmented using random color and sharpness manipulation, Gaussian noise, and random cropping. Despite using only a synthetic dataset for the final training, the network performs well on real-world data, as evaluated in Sec. VI.

A. DETECTION

To exploit more information about the scene by including also distant segments, we use a stereo camera instead of an RGB-D sensor. However, stereo estimated depth is not accurate enough to reliably segment an image into planar segments and to fit supporting 3-D planes for those segments. Nonetheless, a pair of stereo images is still a valuable source of information regarding the geometry of the scene and can be used without explicit depth reconstruction. In the PlaneLoc2 a DNN is used to segment image into planar segments and to estimate segments' supporting planes. The Stereo Plane R-CNN architecture detailed in [5] uses camera-agnostic geometry representation to provide robustness to camera parameters change and to enhance the results. To use this network for localization purposes, an export mechanism had to be added that handles the depth uncertainty. Besides a plane equation and a hull denoting the boundary of the segment, we also store a mean value and a covariance matrix of 3-D points forming this segment. The points are calculated using the estimated depth and the uncertainty of their estimation is extracted from the disparity estimation branch of the geometry module of the DNN. In this branch, a cost volume is created that holds the probability distribution over disparity values for each pixel. It is straightforward to compute the standard deviation of disparity σ_d from this distribution:

$$\sigma_d = \sqrt{\sum_d p(d)(d - \overline{d})}, \tag{3}$$

where $p(d)$ is a probability that d is a disparity for this pixel, and $\overline{d}$ is an expected value of the disparity. Then, a standard deviation of depth σ_z can be calculated using a camera model as follows:

$$\sigma_z = \sigma_d \frac{z}{f_x b}, \tag{4}$$

where z is a depth value, f_x is a focal length for X axis of the camera, and b is a baseline of the stereo setup. Finally, a covariance of 3-D point $\mathbf{x}_i$ in a camera frame of reference can be approximated as:

$$\mathbf{S}_i = \begin{pmatrix} 0.05^2 & & \\ & 0.05^2 & \\ & & \sigma_z^2 \end{pmatrix}. \tag{5}$$

A small, constant value of uncertainty of 0.05 m was used for the X and Y axes because uncertainty in those directions can be neglected compared to uncertainty in the Z axis. The uncertainties of individual points are aggregated to obtain a covariance matrix of the whole point cloud as follows:

$$\mathbf{S} = \sum_i \mathbf{S}_i + \mathbf{x}_i \mathbf{x}_i^T - \boldsymbol{\mu}\boldsymbol{\mu}^T, \tag{6}$$

where $\boldsymbol{\mu}$ is a centroid of the point cloud. This uncertainty is necessary to accommodate for inaccurate geometry estimation during the association check and the pose retrieval.

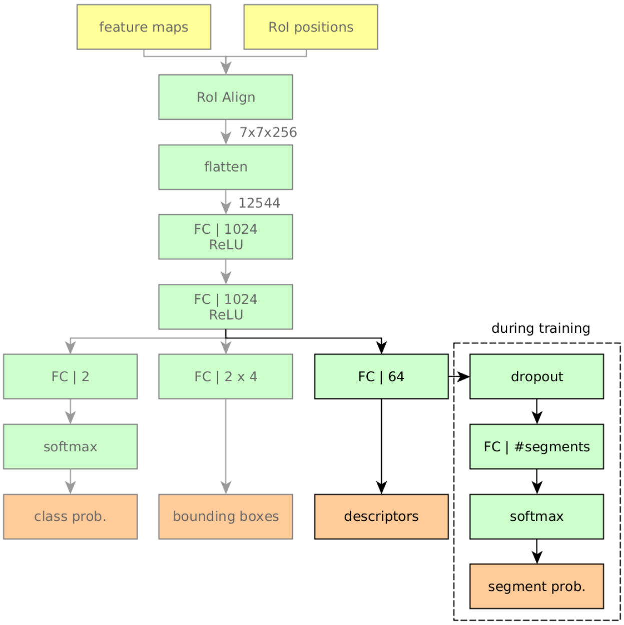

FIGURE 4. A modified classification head from stereo plane R-CNN that produces descriptors. Gray blocks and connections were not modified. Location of this classification head in the entire structure of the DNN is presented in Fig. 3.

B. DESCRIPTION

The DNN also helped resolve another issue of the previous version of PlaneLoc system, namely poorly discriminating appearance descriptors. We added additional layers in a classification head (location of this classification head in the entire structure of the DNN is presented in Fig. 3) of the DNN that produce descriptors as presented in Fig. 4. Features that are used to compute class probabilities and bounding box refinements are processed by a fully connected layer to output a descriptor. However, the most troublesome part of training a DNN that computes descriptors is the way of supervision. Inspired by [6], we also formulate this problem as a classification task. During training, every instance of a planar segment in the training dataset is a separate class and all observations of this segment should be classified as this class. To increase the robustness of the descriptor, a dropout layer is added between the descriptor and fully connected and softmax layers that output segment instance probabilities. The segment instance probabilities are used to compute a cross entropy loss by comparing with target annotations. Correspondences between observations and instances of planar segments that serve as the target annotations are computed using 3-D mesh models, eliminating the need for tedious manual labeling. Such a modification adds little overhead to the Stereo Plane R-CNN model from [5], while producing discriminative descriptors.

V. VIEW-BASED APPROACH TO GLOBAL LOCALIZATION

Distant planar segments, even if not useful to constrain how far the sensor is from the segment because of problems with accurate depth estimation, still provide good orientation constraints. To exploit those constraints, we proposed a novel, view-based map and a pose retrieval procedure that takes into consideration the uncertainty of depth estimation. Moreover, the new pose retrieval procedure treats planar segments as spatially bounded, providing more constraints as opposed to the previous approach that treated them as infinite planes.

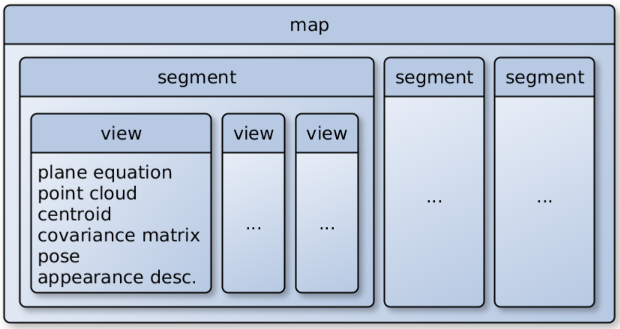

FIGURE 5. The map in PlaneLoc2 contains planar segments, whereas segments store information about their views.

A. PLANAR SEGMENT MAP

The new map structure, instead of explicitly merging different observations of the same planar segment to produce a single representation in the form of a point cloud, stores information about separate views of segments. The structure of the map is depicted in Fig. 5, showing the following information is stored for each view:

- Plane equation ($\boldsymbol{\pi}$) - estimated by the plane parameters branch of the geometry module.
- Point cloud - 3-D points constituting the segment. Points are reprojected using the stereo estimated depth from the disparity branch of the DNN. To limit storage requirements, they are downsampled using a voxel grid filter with a raster of 0.05 m.
- Centroid and covariance matrix ($\boldsymbol{\mu}, \mathbf{S}$) - computed from the point cloud.
- Pose - a visual odometry pose from which the segment was observed.
- Appearance descriptor - produced by the DNN and used to retrieve global map view candidates.

By avoiding merging, we circumvent the problem of, usually computationally costly, information merging and uncertainty propagation from different views. When a new frame is processed, a depth buffer is built to check which segments from the local map can be visible. During the buffer construction, for every segment we select a view with the observation pose $\mathbf{T}^v$ (expressed as a $SE(3)$ transformation matrix) closest to the current pose $\mathbf{T}^c$, according to the following metric, that is a weighted sum of translational and rotational differences:

$$d\left(\mathbf{T}^c, \mathbf{T}^v\right) = d_t\left(\left(\mathbf{T}^v\right)^{-1}\mathbf{T}_1\right) + w_r d_r\left(\left(\mathbf{T}^v\right)^{-1}\mathbf{T}_1\right), \quad (7)$$

where w_r is a weight of the rotational difference, $d_t(\cdot)$ is a function returning translation of the transformation, and $d_r(\cdot)$

is a function returning rotation of the transformation. The weight $w_r = 5$ is set to make an error of approximately 5° equal to an error of 0.5 m. The new views are matched against the potentially visible segments by a geometry test that employs the same error function as the pose retrieval procedure:

$$g\left(\mathcal{P}^c, \mathcal{N}(\boldsymbol{\mu}^v, \mathbf{S}^v)\right) = \sqrt{e_{c,v}\left(\mathbf{I}, \mathbf{0}\right)} < \tau_g, \tag{8}$$

where $\mathcal{P}^c$ is a set of points representing the currently considered new view, $\mathcal{N}(\boldsymbol{\mu}^v, \mathbf{S}^v)$ is a distribution representing the local map view, $e_{c,v}\left(\mathbf{I}, \mathbf{0}\right)$ is the error function defined in (9) for identity transformation, and τ_g is a threshold. Depending on the results of this test, views are either added to existing segments or create new ones. Additionally, we store an end-of-life (EOL) counter for every segment. It is initialized with a value of 4 and increased by 2 whenever a new view is added and decreased by 1 whenever the segment is potentially visible but no new view was added. Segments with EOL higher than 8 are treated as mature and their counter is not decreased anymore. When EOL drops to 0, the segment is considered an invalid observation and is removed from the map.

The local map has a limited time horizon of 2 seconds. Such a horizon prevents accumulation of the drift from the visual odometry, yet includes a broader context of a scene than a single frame. As a result of the view-based approach, older information can be easily removed by dropping information about outdated views.

B. POSE RETRIEVAL

The aim of pose retrieval is to compute a pose of the sensor with respect to the global map, given a set of matches between views of planar segments in the local map and ones in the global map. The novel pose retrieval used in this work does so by minimizing an error of fitting virtual points of the first planar segment to a distribution describing the second planar segment. Such formulation allows exploiting uncertainty of depth estimation while also providing a system of linear equations that can be quickly solved. Consider a planar segment from the local map (denoted by a superscript l) and a planar segment from the global map (denoted by a superscript g) described by their centroids $\boldsymbol{\mu}$, covariance matrices $\mathbf{S}$, and plane equations $\boldsymbol{\pi}$. To assess how N transformed points $\mathbf{R}\mathbf{x}^l_i + \mathbf{t}$ forming the local segment distribution fit the global segment distribution $\mathcal{N}(\boldsymbol{\mu}^g, \mathbf{S}^g)$, one can use a squared Mahalanobis distance:

$$\begin{aligned}
e_{l,g}\left(\mathbf{R}, \mathbf{t}\right) &= \\
&= \frac{1}{N}\sum_i (\mathbf{R}\mathbf{x}^l_i + \mathbf{t} - \boldsymbol{\mu}^g)^T (\mathbf{S}^g)^{-1} (\mathbf{R}\mathbf{x}^l_i + \mathbf{t} - \boldsymbol{\mu}^g) \\
&= \frac{1}{N}\sum_i (\mathbf{R}\mathbf{x}^l_i + \mathbf{t} - \boldsymbol{\mu}^g)^T (\mathbf{V}^g \boldsymbol{\Lambda}^g (\mathbf{V}^g)^T)^{-1} \\
&\quad \times (\mathbf{R}\mathbf{x}^l_i + \mathbf{t} - \boldsymbol{\mu}^g) \\
&= \frac{1}{N}\sum_i (\mathbf{R}\mathbf{x}^l_i + \mathbf{t} - \boldsymbol{\mu}^g)^T (\mathbf{V}^g \boldsymbol{\Lambda}^g_s (\boldsymbol{\Lambda}^g_s)^T (\mathbf{V}^g)^T) \\
&\quad \times (\mathbf{R}\mathbf{x}^l_i + \mathbf{t} - \boldsymbol{\mu}^g) \\
&= \frac{1}{N}\sum_i \sum_k \left((\mathbf{R}\mathbf{x}^l_i + \mathbf{t} - \boldsymbol{\mu}^g)^T \mathbf{v}^g_k \frac{1}{\sqrt{\lambda^g_k}} \right)^2,
\end{aligned} \tag{9}$$

where $\mathbf{V}^g \boldsymbol{\Lambda}^g (\mathbf{V}^g)^T$ is an eigen decomposition of the covariance matrix $\mathbf{S}^g$, $\boldsymbol{\Lambda}^g_s$ is a matrix with inverses of square roots of eigenvalues $\frac{1}{\sqrt{\lambda^g_k}}$ on the diagonal, and $\mathbf{v}^g_k$ are columns of the matrix $\mathbf{V}^g$ and eigenvectors of the covariance matrix $\mathbf{S}^g$. To minimize $e_{l,g}\left(\mathbf{R}, \mathbf{t}\right)$, a set of linear equations can be build in the form:

$$(\mathbf{R}\mathbf{x}^l_i + \mathbf{t} - \boldsymbol{\mu}^g)^T \mathbf{v}^g_k \frac{1}{\sqrt{\lambda^g_k}} = 0, \tag{10}$$

and then solved using SVD-based least squares algorithm. Unfortunately, this gives $3N$ equations when using all points from the local segment distribution. Hence, instead of all points, we use virtual points that subsume the distribution:

$$\mathbf{x}^l_{\pm k} = \boldsymbol{\mu}^l \pm K\sqrt{\lambda^l_k}\,\mathbf{v}^l_k, \tag{11}$$

where K is a number of dimensions used. We use 4 virtual points that correspond to two principal directions ($K = 2$) of the distribution $\mathcal{N}(\boldsymbol{\mu}^l, \mathbf{S}^l)$ projected onto plane $\boldsymbol{\pi}^l$. Those points lay on the plane and are in a distance of two standard deviations from the centroid. Using only points on the plane from the local segment distribution, instead of using 6 points that would represent the distribution before the projection, is of utmost importance to conserve the planar nature of those constraints. If all 6 points were used, and the uncertainty of estimation would be high in a direction of a normal vector of the local segment (i.e. due to poor depth estimation), the fitting error would be high if the global distribution was mainly planar (see Fig. 6). This high fitting error could cause minimization to favor undesired rotations. Moreover, using only 4 points further reduce the number of equations by exploiting the planarity of the segments. Additionally, the centroid and the covariance matrix are computed using stereo estimated depth and give a less accurate description of the geometry than the plane equation from the specialized branch of the DNN. Hence, by projecting distribution on the plane, the accuracy is increased. However, centroids and covariance matrices are still used because they are the only source of uncertainty measures.

After solving a system of 36 equations (3 pairs of matched segments, 3 dimensions, 4 virtual points) in the form of Eq. (10), we get values of the matrix $\mathbf{R}$ and the vector $\mathbf{t}$. Unfortunately, there are no constraints on the orthonormality of the values in $\mathbf{R}$, so it might not be a valid rotation matrix. To obtain a proper rotation matrix, we perform orthonormalization using the SVD decomposition:

$$\mathbf{R}' = \mathbf{U}\mathbf{V}^T, \tag{12}$$

where $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \mathbf{R}$ is the SVD decomposition. To refine the transformation, a Gauss Newton optimization in the Lie

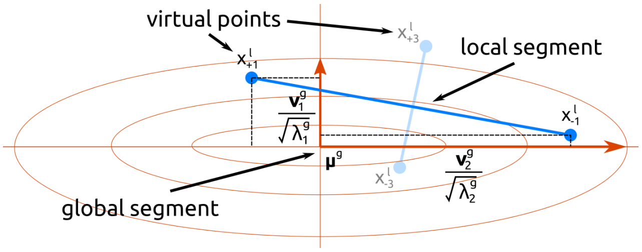

FIGURE 6. Schematic illustration of fitting error of local planar segment (blue) to global planar segment (orange) using 2-D section. Lengths of $\frac{\mathbf{v}^g}{\lambda^g}$ vectors correspond to a unit of error. The error is denoted using a dashed line. Using virtual points perpendicular to the plane $\mathbf{x}^l_{-3}$ and $\mathbf{x}^l_{+3}$ could yield high errors and undesired behavior during optimization (see text).

algebra is performed by minimizing a sum of squares of the following residuals:

$$r_{i,k} = (\mathbf{R}\exp(\boldsymbol{\omega})\mathbf{x}^l_i + \mathbf{t} - \boldsymbol{\mu}^g)^T \mathbf{v}^g_k \frac{1}{\sqrt{\lambda^g_k}}, \qquad (13)$$

where $\boldsymbol{\omega}$ is a rotation increment. The Jacobians of the residuals are as follows:

$$\frac{\partial r_{i,k}}{\partial \mathbf{t}} = (\mathbf{v}^g_k)^T \frac{1}{\sqrt{\lambda^g_k}} \qquad (14)$$

$$\left.\frac{\partial r_{i,k}}{\partial \boldsymbol{\omega}}\right|_{\boldsymbol{\omega}=\mathbf{0}} = (\mathbf{v}^g_k)^T \frac{1}{\sqrt{\lambda^g_k}} \mathbf{R} \left[\mathbf{x}^l_i\right]_\times, \qquad (15)$$

where $\left[\mathbf{x}^l_i\right]_\times$ is skew symmetric matrix formed from elements of $\mathbf{x}^l_i$. We can assume that $\boldsymbol{\omega}$ is close to $\mathbf{0}$ because it is an increment. By empirical examination, the number of iterations was set to a constant value of 5. In a vast majority of cases, further iterations do not alter the transformation, whereas using a constant value bounds the execution time. The result is a transformation $(\mathbf{R}, \mathbf{t})$ that stems from the geometric constraints imposed by a set of matched planar segments and is used later to build the PDF of the agent pose. The same procedure is also used to compute the final pose, after the maximum of the PDF was found and all matches were established.

VI. EXPERIMENTAL VERIFICATION

We use a real-world TERRINet dataset[2] to evaluate the proposed solution. The dataset contains trajectories from 3 different scenes with reference poses from Qualisys motion capture system. We recorded stereo images along with Velodyne VLP-16 LiDAR scans that were later used to generate ground truth depth maps for every image. The ground truth depth maps enabled the computation of correspondences between planar segments detected in different image frames.

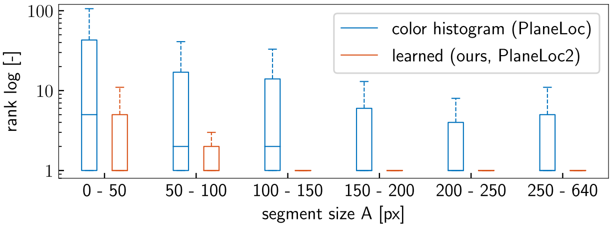

FIGURE 7. Statistics on ranks as a function of the square root of segment size *A*. Values on the box plot further than 1.5 inter-quartile range from the box were treated as outliers and removed.

A. DESCRIPTION

The aim of the first experiment is to show the effectiveness of our new learned descriptors. We compare them with descriptors based on color histograms used in the previous version of PlaneLoc. For every detected planar segment we compute its rank, i.e. the number of nearest neighbors necessary to fetch from the database of all descriptors to include a correct match. We exclude segments from the same trajectory, as images containing them could be very similar to the image of the query segment. To give more insight on the characteristic of descriptors, we present the rank as a function of the size of detected segments. We divided segments based on the square root of their area in pixels, denoted as A, into 6 bins (see Fig. 7). It is clearly visible that the learned descriptors outperform the histogram-based ones by a large margin for all sizes. It is also worth noting that from A equal to 100, the first neighbor is almost always the correct one (values on the box plot further than 1.5 inter-quartile range from the box were treated as outliers and removed).

B. LOCALIZATION

The second experiment compares the proposed solution with other state-of-the-art global localization systems. As avoiding an incorrect loop closure or relocalization is of utmost importance to the precision of most SLAM systems, we report a percentage of correct and incorrect localization acts (called recognitions hereinafter) and their precision. We compare the pose computed by a considered method with the reference pose and compute the translational and the rotational error. The threshold for assuming a recognition correct is 0.5 m and 10° as an error within such bounds usually enables resuming tracking in SLAM systems [7]. If a method returns no result, we do not compute the errors and treat such outcome as an unknown pose. For each scene, we use one trajectory to build a map and a different one to evaluate localization with a known map. The map is built using the reference poses for all tested solutions to exclude the factor of map precision. We tested the following solutions:

- `OS3/r` - relocalization mechanism from ORB-SLAM3. The system was forced to relocalize every frame and pose after local map tracking was evaluated if the relocalization was successful. Localization is performed every frame.

[2] This dataset was collected during the author's visit to LAAS-CNRS in Touluse, within the TERRINet project funded by EU H2020 under GA No.730994.

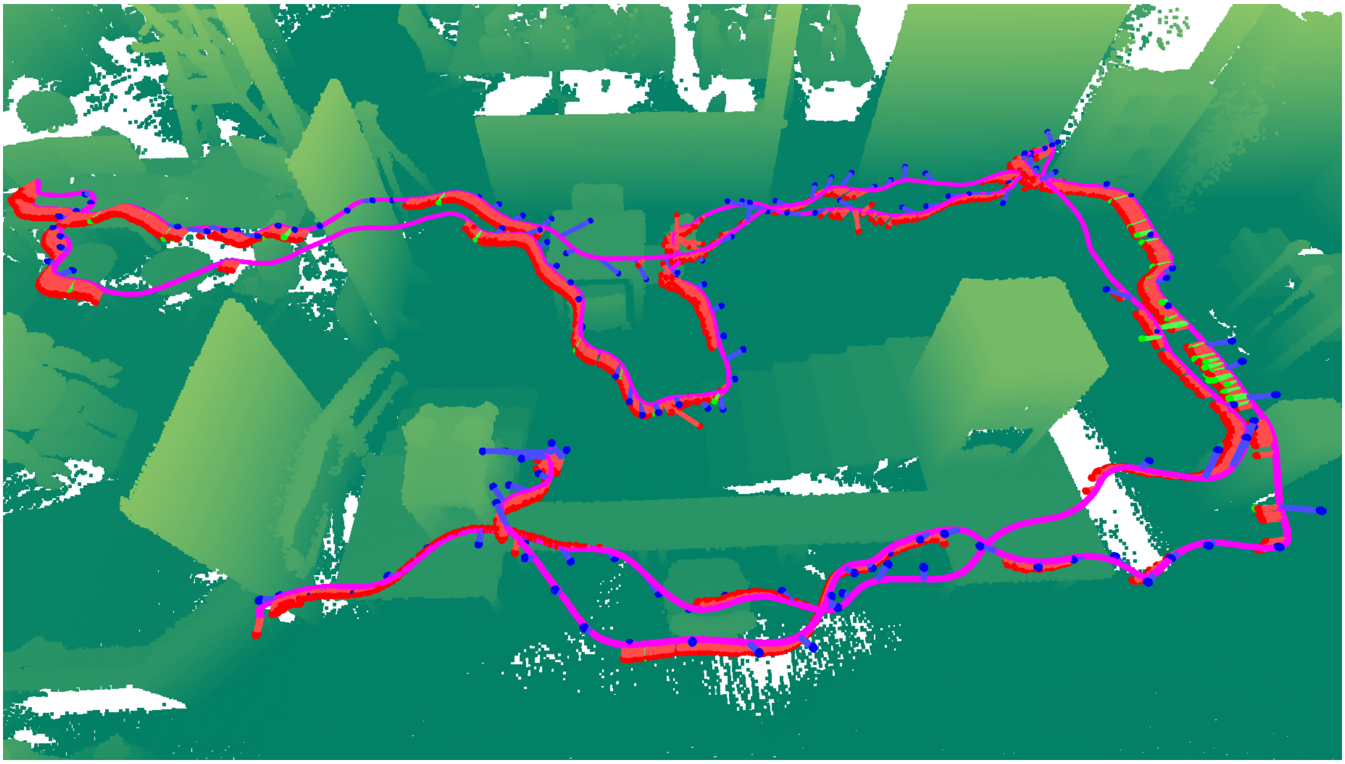

FIGURE 8. **Visualization of results with reference trajectory (magenta line), ORB-SLAM3 relocalization poses (red points), ORB-SLAM3 map merge poses (green points), and PlaneLoc2 poses (blue points). Lines of corresponding colors connect reference poses with computed poses. Points in the point cloud are colored according to their height above the ground. Results for NV+SP were omitted for clarity.**

- `OS3/m` - map merging mechanism from ORB-SLAM3. A new map was being build for the test trajectory and a transformation between the current map and the prebuilt map was evaluated if a merge was successful. Localization is performed every time a keyframe is inserted into the map.
- `NV+SP` - hierarchical localization [26] with SuperGlue [18] and NetVLAD [25] was evaluated with a global map constructed using COLMAP software.[3] Localization is performed every frame.
- `PL2` (ours) - the solution presented in this paper. The local map is updated every 15 frames because consecutive frames are similar to each other and do not provide diverse views, therefore localization is performed every 15 frames.

Setting proper values of parameters is a troublesome task, especially in complex systems. To facilitate this task in the PlaneLoc2, we follow a data-driven paradigm and use the first scene to perform statistical analysis and compute the values of parameters:

- τ_d - a maximum distance between descriptors that is considered during candidate segment views retrieval. It is set to include 90% of all correct matches.
- $\tau_{svd,t}$ and $\tau_{svd,r}$ - a minimum value of a singular value for translational and rotational part Jacobians in the gradient descent optimization of the pose to assume that the pose is constrained in all dimensions. It is set to include 90% of all correct triplets.
- τ_e - a maximum value of residual error to consider a fitting of planar segments as correct during the pose retrieval. It is set to include 75% of all correct triplets. Value of 75% was used instead of 90% to limit the number of considered triplets and to reduce the computational burden.
- τ_p, τ_r, and τ_d - thresholds that are used during the final safe-checks. They are set to maximize the number of correct matches, while keeping the number of incorrect matches equal to 0. Multiplied by a factor of 1.2, inspired by the Lowe's ratio test [30], to add a safety margin.
- τ_g - threshold used to determine whether two segment observations should be merged (see Eq. (8)) in a map. Empirically set to a value of 2 that prevents most of the incorrect data associations.

To enable a fair comparison, for ORB-SLAM3 we used the parameter setting designed by the authors and used in the EuRoC indoor experiments [4]. Likewise, for NV+SP we used parameters set for the InLoc dataset [26] that is similar in characteristic to the TERRINet dataset.

Quantitative results are gathered in Tab. 1, while visualization of results for scene 02 are presented in Fig. 8. Both ORB-SLAM3 mechanisms, relocalization and map merging, recognize a lower percentage of poses than our solution.

[3] https://colmap.github.io

TABLE 1. Results of global localization on TERRINet dataset. Cases with incorrect recognitions are colored red. The best correct recognitions rates for cases without incorrect recognitions are emboldened.

scene	measure	method			
		OS3/r	OS3/m	NV+SP	PL2
02	correct [%]	58.1	40.3	95.0	**73.8**
	incorrect [%]	0.0	0.0	5.0	0.0
	unknown [%]	41.9	59.6	0.0	26.2
	mean error lin. [m]	0.08	0.09	0.10	0.09
	mean error ang. [°]	0.7	0.7	1.9	1.1
	max. error lin. [m]	0.27	0.17	18.35	0.37
	max error ang. [°]	5.3	1.7	156.3	3.2
03	correct [%]	49.4	18.2	94.3	**47.9**
	incorrect [%]	0.2	0.0	5.7	0.0
	unknown [%]	50.4	81.8	0.0	52.1
	mean error lin. [m]	0.06	0.04	0.20	0.10
	mean error ang. [°]	1.5	0.7	2.9	1.1
	max. error lin. [m]	0.81	0.09	12.88	0.38
	max error ang. [°]	10.5	1.2	174.8	3.2

An exception is scene 02, where relocalization recognized slightly more poses, but also yielded incorrect ones. The NV+SP recognized a higher percentage of poses but also produced many incorrect ones, some of which were distant more than 18 m from the reference pose. Such behavior can be attributed to a lack of fail-safe checks that inevitably reject some of the correct recognitions, but also prevent incorrect ones. Thus, our system recognized the highest percentage of poses among cases where no incorrect results were produced. Moreover, our system did not produce any incorrect recognitions in all test cases.

The accuracy of all tested methods is similar, with mean error values varying slightly on different scenes. Maximum errors depend mainly on incorrect recognitions and are the lowest for the ORB-SLAM3 map merging mechanism, while being below 0.4 m and 3.5° for the PlaneLoc2.

VII. CONCLUSION

In this article, we present the PlaneLoc2 global localization method that utilizes a passive stereo camera to detect planar segments and compute a PDF of the 6-D pose. The method uses a DNN that jointly detects planar segments, describes their appearance, and estimates their geometry. The detected segments are used to build view-based local and global maps, that are easily manageable and store information about the uncertainty of geometry of planar segments. The uncertainty is exploited in a novel pose retrieval procedure that is designed with stereo sensors in mind. In the experimental section, we show that the new learned appearance descriptor outperforms the classic, based on color histograms one. We also tested the global localization performance of our system and show that it achieves the best percentage of recognized poses, when cases without incorrect recognitions are considered (15.7% more poses in the first scene than the second best solution and 29.7% more poses in the second scene). Moreover, the PlaneLoc2 did not produce incorrect recognitions in all cases, which is of pivotal importance in navigation and SLAM systems, proving its suitability as a global localization system.

The most important changes, with respect to the previous version of PlaneLoc, that helped achieve good results include the new appearance descriptor. Results in Sec. VI-A suggest that it significantly limits the number of incorrect potential matches. Additionally, considering geometric constraints from distant segments enabled correct pose retrieval in higher percentage of situations. The new pose retrieval procedure that accommodates the spatial boundaries of planar segments further increases the number of geometric constraints available. All those factors facilitate a high correct recognition rate without incorrect recognitions.

As a part of the future work, we plan to expand the system with other types of geometric features, such as edges. Edges could provide additional constraints that are unused in this version of the system.

REFERENCES

[1] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics* (Intelligent Robotics and Autonomous Agents). Cambridge, U.K.: MIT Press, 2006.

[2] C. Liu, K. Kim, J. Gu, Y. Furukawa, and J. Kautz, "PlaneRCNN: 3D plane detection and reconstruction from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4445–4454.

[3] S. Yang and S. Scherer, "Monocular object and plane SLAM in structured environments," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 3145–3152, Oct. 2019.

[4] C. Campos, R. Elvira, J. J. G. Rodriguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

[5] J. Wietrzykowski and D. Belter, "Stereo plane R-CNN: Accurate scene geometry reconstruction using planar segments and camera-agnostic representation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4345–4352, Apr. 2022.

[6] J. Wietrzykowski and P. Skrzypczynski, "On the descriptive power of LiDAR intensity images for segment-based loop closing in 3-D SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 79–85.

[7] J. Wietrzykowski and P. Skrzypczyński, "PlaneLoc: Probabilistic global localization in 3-D using local planar features," *Robot. Auto. Syst.*, vol. 113, pp. 160–173, Mar. 2019.

[8] G. Halmetschlager-Funek, M. Suchi, M. Kampel, and M. Vincze, "An empirical evaluation of ten depth cameras: Bias, precision, lateral noise, different lighting conditions and materials, and multiple sensor setups in indoor environments," *IEEE Robot. Autom. Mag.*, vol. 26, no. 1, pp. 67–77, Mar. 2019.

[9] I. C. F. S. Condotta, T. M. Brown-Brandl, S. K. Pitla, J. P. Stinn, and K. O. Silva-Miranda, "Evaluation of low-cost depth cameras for agricultural applications," *Comput. Electron. Agricult.*, vol. 173, Jun. 2020, Art. no. 105394.

[10] P. Hausamann, C. B. Sinnott, M. Daumer, and P. R. MacNeilage, "Evaluation of the Intel RealSense T265 for tracking natural human head motion," *Sci. Rep.*, vol. 11, no. 1, p. 12486, Dec. 2021.

[11] K. Chappellet, G. Caron, F. Kanehiro, K. Sakurada, and A. Kheddar, "Benchmarking cameras for open VSLAM indoors," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 4857–4864.

[12] M. Senthilvel, R. K. Soman, and K. Varghese, "Comparison of handheld devices for 3D reconstruction in construction," in *Proc. Int. Symp. Autom. Robot. Construction (IAARC)*, Taipei, Taiwan, Jul. 2017, pp. 698–705.

[13] I. Cvisic, J. Cesis, I. Markovic, and I. Petrovic, "SOFT-SLAM: Computationally efficient stereo visual simultaneous localization and mapping for autonomous unmanned aerial vehicles," *J. Field Robot.*, vol. 35, no. 4, pp. 578–595, Jun. 2018.

[14] N. Smolyanskiy, A. Kamenev, and S. Birchfield, "On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 1007–1015.

[15] D. Galvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1188–1197, Oct. 2012.

[16] X. Li, K. Han, S. Li, and V. Prisacariu, "Dual-resolution correspondence networks," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 17346–17357.

[17] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 224–236.

[18] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 4937–4946.

[19] P.-E. Sarlin, A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl, and T. Sattler, "Back to the feature: Learning robust camera localization from pixels to pose," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3246–3256.

[20] K. Ćwian, M. R. Nowicki, J. Wietrzykowski, and P. Skrzypczyński, "Large-scale LiDAR SLAM with factor graph optimization on high-level geometric features," *Sensors*, vol. 21, no. 10, p. 3445, May 2021.

[21] X. Zhang, W. Wang, X. Qi, Z. Liao, and R. Wei, "Point-plane SLAM using supposed planes for indoor environments," *Sensors*, vol. 19, no. 17, p. 3795, Sep. 2019.

[22] Y. Taguchi, Y.-D. Jian, S. Ramalingam, and C. Feng, "Point-plane SLAM for hand-held 3D sensors," in *Proc. IEEE Int. Conf. Robot. Autom.*, Karlsruhe, Germany, May 2013, pp. 5182–5189.

[23] E. Fernández-Moral, W. Mayol-Cuevas, V. Arévalo, and J. González-Jiménez, "Fast place recognition with plane-based maps," in *Proc. Int. Conf. Robot. Autom.*, Karlsruhe, Germany, May 2013, pp. 2719–2724.

[24] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "InLoc: Indoor visual localization with dense matching and view synthesis," Apr. 2018, *arXiv:1803.10368*.

[25] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1437–1451, Jun. 2018.

[26] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 12708–12717.

[27] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[28] M. Xu, N. Snderhauf, and M. Milford, "Probabilistic visual place recognition for hierarchical localization," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 311–318, Apr. 2021.

[29] R. Steiner, M. Cox, P. V. K. Borges, L. Bernreiter, and J. Nieto, "Certainty aware global localisation using 3D point correspondences," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 8710–8717, Oct. 2021.

[30] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Jan. 2004.

JAN WIETRZYKOWSKI received the B.Sc. and M.Sc. degrees in automatic control and robotics from the Poznan University of Technology, in 2014 and 2015, respectively. He is currently pursuing the Ph.D. degree with the Faculty of Control, Robotics, and Electrical Engineering. Since 2016, he has been a Research Assistant with the Institute of Robotics and Machine Intelligence. He is the author or coauthor of multiple technical papers in the area of robotics and machine learning, including ICRA and IROS conference papers, and RAS and RAL journal articles. His current research interests include robotic global localization, machine learning, and simultaneous localization and mapping.