2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Spherical Multi-Modal Place Recognition for Heterogeneous Sensor Systems

Lukas Bernreiter, Lionel Ott, Juan Nieto, Roland Siegwart and Cesar Cadena

***Abstract*— In this paper, we propose a robust end-to-end multi-modal pipeline for place recognition where the sensor systems can differ from the map building to the query. Our approach operates directly on images and LiDAR scans without requiring any local feature extraction modules. By projecting the sensor data onto the unit sphere, we learn a multi-modal descriptor of partially overlapping scenes using a spherical convolutional neural network. The employed spherical projection model enables the support of arbitrary LiDAR and camera systems readily without losing information. Loop closure candidates are found using a nearest-neighbor lookup in the embedding space. We tackle the problem of correctly identifying the closest place by correlating the candidates' power spectra, obtaining a confidence value per prospect. Our estimate for the correct place corresponds then to the candidate with the highest confidence. We evaluate our proposal w.r.t. state-of-the-art approaches in place recognition using real-world data acquired using different sensors. Our approach can achieve a recall that is up to $10\,\%$ and $5\,\%$ higher than for a LiDAR- and vision-based system, respectively, when the sensor setup differs between model training and deployment. Additionally, our place selection can correctly identify up to $95\,\%$ matches from the candidate set.**

I. Introduction

Place recognition in mobile robotics with heterogeneous sensory systems are particularly challenging for current vision and laser-based place recognition systems. Many of the existing approaches are often tailored to a specific LiDAR or camera and often do not generalize well to another system with different resolutions, variations in densities, or camera lenses. Furthermore, the problem becomes significantly more challenging for learning-based methods when the employed sensors are switched between training and testing or building and querying a map.

Such scenarios are especially crucial for autonomous vehicles as it enables building a generic model using high-performance sensors and test on manufacturer-specific, low-cost sensory systems. Consequently, allowing manufacturers to freely choose the employed sensors and decrease their production costs using lower fidelity sensory systems. Additionally, heterogeneous models would remain applicable even when the hardware is upgraded and would not be specialized for a specific year or vehicle generation. However, it is common for state-of-the-art place recognition systems to simplify the problem by assuming the same sensory system for training and testing as it is further, in many cases, inherently impossible to change sensor configurations without degradations once a model is learned. Moreover, a robust and reliable place recognition is often not achieved through a single sensory system, e.g., GNSS-based localization is insufficiently accurate in various scenarios due to multipath effects. Image-based place recognition systems are prone to degradation when the data contains viewpoint and illumination changes and, further, commonly do not support multi-camera systems very well. Similarly, several LiDAR-based place recognition systems require computationally expensive preprocessing steps, e.g., ground and noise removal, and typically suffer from degraded performance with rotated scans from the same scene. Despite these challenges, the current state-of-the-art either addresses the vision- or laser-based place recognition problem and do not take advantage of their complementary nature. Visual sensors provide descriptive appearance information about the environment, whereas LiDAR sensors are useful to measure the range accurately and, when combined, can increase the robustness of localization or place recognition systems. Regardless of the employed sensors, most place recognition algorithms can only reliably retrieve the correct place when considering a high number of potential candidates. Thus, making outlier rejection algorithms or additional filter steps a critical requirement. While most of the currently proposed solutions rely on a single modality per network, our approach combines multiple modalities in a single network and forward pass. The employed spherical projection model seamlessly allows supporting high field-of-view systems without introducing distortions. In more detail, we use a spherical Convolutional Neural Network (CNN) [1]–[3] to learn an embedding optimized for place recognition taking the projected sensor data

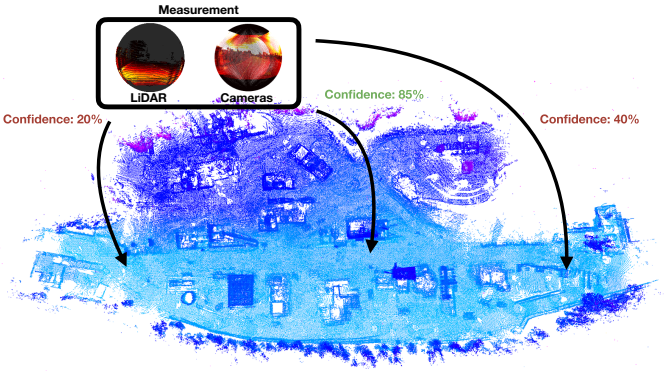

Fig. 1. We propose a place recognition pipeline that combines, but is not limited to, the output of a LiDAR scan and multiple cameras. Furthermore, the outcome of our pipeline yields a confidence value for the place matching.

This work was supported by the National Center of Competence in Research (NCCR) Robotics through the Swiss National Science Foundation.

All authors are with the Autonomous Systems Lab, ETH Zurich, Zurich 8092, Switzerland, {berlukas, liott, nietoj, rsiegwart, cesarc}@ethz.ch.

as input (cf. Figure 1). We retrieve loop closure candidates using a nearest neighbor search in the embedding space. Furthermore, we estimate each candidate's power spectrum and perform a correlation in the spherical harmonic domain to find the best match between all retrieved candidates. In summary, our contributions are

- An end-to-end pipeline for place recognition using a spherical projection of cameras and LiDARs.
- A probabilistic voting framework for finding the correct match given a set of potential candidates.

We conduct a detailed evaluation of several heterogeneous place recognition scenarios and sensor setups.

II. Related Work

This section reviews the current state-of-the-art vision- and LiDAR-based localization and place recognition systems related to our pipeline, where we begin with individual solutions and then conclude with multi-modal approaches.

A. Visual-based Approaches

Many traditional visual place recognition systems follow the traditional bag-of-words paradigm, where aggregated local descriptors (e.g. SIFT and SURF) represent an image or place [4], [5]. However, learning-based solutions have significantly boosted the field [6], [7] in recent years. Sünderhauf et al. [8] and Chen et al. [9] describe multiple spatial regions using CNN features for place recognition. The work of Hausler et al. [10] proposes a novel fusion of image processing methods for increased robustness and performance. PoseNet [11] is an end-to-end learning system for global localization, whereas NetVLAD [7] learns end-to-end a global descriptor for place recognition using a VLAD layer.

Furthermore, since many robotic scenarios inherently contain different appearances between map and query images, a significant amount of work tackles seasonal and viewpoint discrepancies to avoid degradation [12], [13].

Our proposed approach also learns a representation in an end-to-end manner but does not solely rely on visual images but also takes LiDAR scans as an additional input.

B. LiDAR-based Approaches

Although visual localization approaches are more mature than their LiDAR-based counterparts, the latter can outperform visual systems due to their illumination-invariance and 360-degree field-of-view. Early advances in LiDAR-based localization and place recognition explore point histograms over the whole cloud [14], [15]. However, histograms are inefficient under partially overlapping pointclouds, where they differ.

Here as well, learning-based approaches have gained tremendous success in recent years. Dube et al. [16] propose that with the accumulation of multiple LiDAR scans to individual segments it can learn a more distinctive description of the environment. The use of attention networks for place recognition was proposed by Zhang et al. [17] to reweight local feature points. The work of Du et al. [18] proposes a relocalization pipeline to extract local descriptors, a score map, and a global descriptor in a single forward pass.

Several other approaches utilize a projection to the 2D Euclidean domain and consequently rely on 2D CNNs [19]–[21]. LocNet [19] learns a rotation-invariant global descriptor of pointclouds using histograms per scan line. Uy et al. [22] propose a NetVLAD layer for description on top of PointNet [23]. Recently, Chen et al. [20] propose an end-to-end learning approach, which additionally evaluates the matching based on the overlap. Some other global LiDAR-based localization approaches additionally utilize the returned signal's strength (intensities) and have shown improvements to the localization quality [20], [24], [25].

While LiDAR-based approaches typically perform well in confined environments, their performance often suffers in wide open areas, due to the limited amount of returned beams. Consequently, the combination of multiple modalities on a fundamental level is paramount to design a robust and reliable place recognition system for arbitrary environments. Thus, our approach aims at combining the visual and LiDAR-based place recognition into a single pipeline.

C. Multi- and Cross-modal Approaches

Recently, there is great interest in robotics to perform multi- or cross-modal place recognition and localization. In the work of Ratz et al. [26], a multi-modal descriptor was learned by fusing the embedding of a NetVLAD and of LiDAR segments [16] using fully connected layers. Xie. et al. [27] presented an end-to-end approach with an image and pointcloud fusion to learn a multi-modal descriptor.

The work of Cattaneo et al. [28] and Caselitz et al. [29] deal with the task of localizing visual data on LiDAR maps. Feng et al. [30] deals with the extraction of descriptors of images and LiDAR patches in a metric learning approach.

Inspired by these findings, we propose an end-to-end place recognition pipeline that combines vision and LiDAR sensors. Both modalities are projected onto the unit sphere and are given as an input to a spherical CNN [1]–[3]. As a result, our network learns a global descriptor directly from the input modalities and does not require any local feature extraction modules.

III. Method

This section describes the core components of our place recognition framework (cf. Figure 2). First, we discuss the spherical projection model and how the modalities are combined. Second, we explain the spherical CNN architecture that takes the spherical signals as input and learns a global descriptor. These two parts form the map building stage of our proposed methodology. Finally, we describe our spectral analysis-based candidate selection and matching process.

A. Data Representation

Overall, our approach utilizes multiple camera images and LiDAR scans to learn a global multi-modal descriptor of all modalities. Although we only use two modalities (camera and LiDAR) in this work, our approach is not limited to these

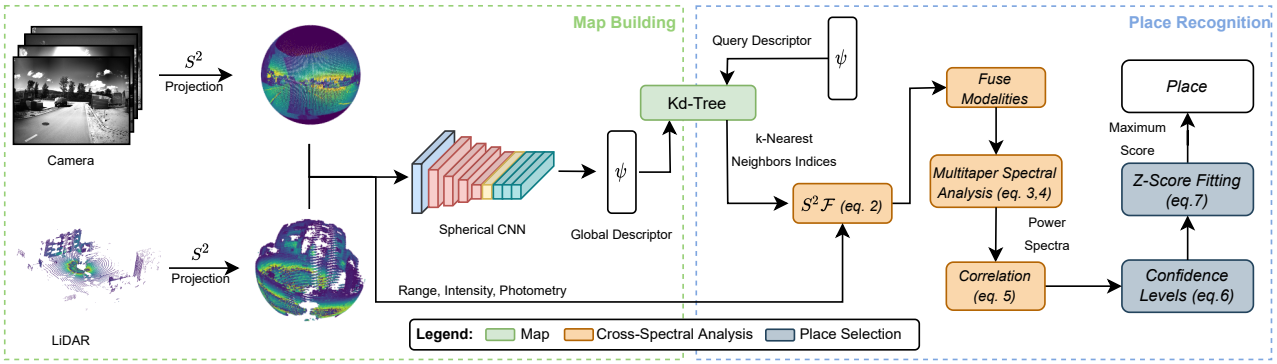

Fig. 2. Overview of the proposed place recognition framework. Spherical projections of multiple images and a LiDAR scan serve as input for a spherical CNN. The network learns to differentiate places through a global representation of the input modalities. The nearest neighbors are transformed using a spherical Fourier transform ($S^2\mathcal{F}$), fused, and then correlated. The correct place corresponds to the neighbor with the highest confidence in the correlation to the query. The related equations referenced by the components are discussed in Section III.

and can readily be extended to include more. We start by combining all modalities into one single joint representation to create an input feature vector for our network.

Since our pipeline's core component is a spherical CNN that takes arbitrary square-integrable functions on the two-dimensional hypersphere S^2 as input, we project all modalities onto the sphere. In addition, the spherical CNN requires discrete equiangular samples and generally the input has to comply with Discroll and Healy's (DH) sampling theorem [31]. Hence, we uniformly sample the projected modalities using an equiangular DH grid by performing a k-nearest neighbor lookup of the sampling points.

Generally, the DH grid is defined as a $2\tilde{B} \times 2\tilde{B}$ sampling grid, where $\tilde{B}$ is the spherical bandwidth which controls how dense the sphere is sampled and the discretization of the spherical spectrum. Using $\tilde{B} = 100$, we define a DH grid in a common base frame enabling us to combine all the individual modalities. The constructed equiangular sampling grid in the base frame is then projected into each sensor's local frame to sample the modality using each projected point that falls within the sensor's field-of-view. For camera images, this corresponds to the photometric pixel value, and for LiDAR scans, we sample the range and intensity values. Multiple sensors per modality (e.g. multiple cameras) use the same sampling grid such that overlapping regions can be averaged when sampled. Regions that do not have a corresponding measurement are set to zero. Finally, in the base frame, we combine all samples in our input feature vector $C \in \mathbb{R}^{3\times200\times200}$ comprising photometry, range, and beam intensity and forward it to the spherical CNN.

B. Spherical Convolutional Neural Network Architecture

We utilize a spherical CNN [1]–[3] to learn a unique embedding of the multi-modal input data. As the work of Cohen et al. [1] and Esteves et al. [2] have shown, spherical CNNs work very well for several classification tasks, especially when dealing with rotated data. Notably, for LiDAR-based systems, loop closure candidates can have arbitrary rotations between them, making spherical CNNs a good fit. Furthermore, several modern LiDAR systems leverage a high field-of-view resulting in high distortions for

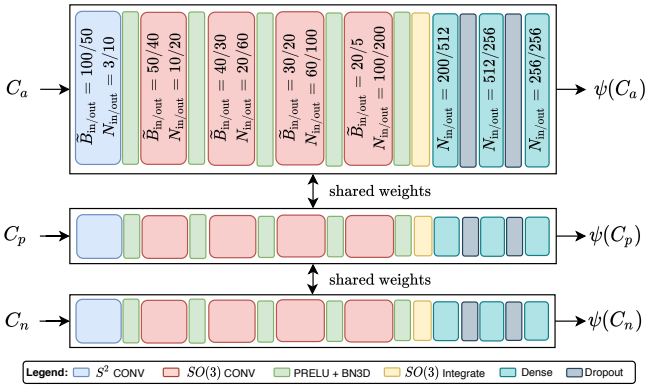

Fig. 3. Detailed network architecture used to train our place recognition pipeline. Each training triplet C_a, C_p and C_n is passed through a duplicated network with shared weights.

planar projection models. In contrast, spherical projections ideally model the nature of a rotating LiDAR and do not introduce any distortions into the projected pointcloud.

Our network architecture consists of $900k$ parameters, five convolutional and three fully-connected layers (cf. Figure 3). The first layer performs a convolution over S^2, whereas the remaining four convolutional layers act on $SO(3)$ to preserve the convolution's equivariance property [32]. Moreover, after each convolution, we apply a PReLu [33] activation function followed by a three-dimensional batch normalization (BN3D). After all convolutions, we integrate over $SO(3)$ and feed the result into our network's final three fully connected layers. Between the fully connected layers, we employ dropout layers with a probability of $40\,\%$. Finally, the outcome of the network will be a 256-dimensional descriptor.

We approximate a function ψ that maps our input C to a unique embedding $\psi(C)$ optimized for place recognition. Thereby, we employ a metric learning approach to learn a compact descriptor using a triplet network setup. Consequently, each training sample comprises an anchor C_a, a positive C_p and a negative C_n sample. A positive sample represents a spatially close area to the anchor, whereas a negative sample is a place with no or minimal overlap w.r.t. the anchor. The loss is defined to pull positive matches closer and negative matches further away in the embedding space [34], i.e.

$$\mathcal{L} = \frac{1}{N}\sum^{N}\left[\|\psi(C_a)-\psi(C_p)\|_2 - \|\psi(C_a)-\psi(C_n)\|_2 + \tau_1\right]_+ + \frac{1}{N}\sum^{N}\left[\|\psi(C_a)-\psi(C_p)\|_2 - \tau_2\right]_+, \quad (1)$$

where $[z]_+ = max(0, z)$, N the number of triplet samples and $\tau_1 = 2, \tau_2 = 0.2$ denote the margins in the embedding space. We employ a stochastic gradient descent and reduced the initial learning rate to 0.0046 to ensure stable conversion. Furthermore, we use a batch size of 13 samples for the optimization as this is our maximum GPU allowance.

During the map building step, a KD-tree with an L_2 norm of the learned descriptors serves as map representation. The initial step for the place recognition is to perform a k-nearest neighbor search given a query descriptor. Next, we retrieve the corresponding equiangular feature vectors and perform a cross-spectral analysis between each of the nearest neighbors and the query to find the correct match.

C. Spectral Analysis and Voting

In this step, we seek to identify the correct place from a set of k-nearest neighbors and reject all outliers. Each neighbor is transformed to the spherical harmonics domain where we fuse the modalities, correlate the fused spectra with the query, and finally evaluate the correlation.

In more detail, this process performs a cross-spectral analysis given the spherical harmonic coefficients of each neighbor in the spherical harmonic domain. Therefore, we first perform a spherical Fourier transform of each neighbor and each modality using the equiangular feature vectors.

Generally, any arbitrary function $\tilde{f} \in L^2(S^2)$ can be expanded in the base of the spherical harmonics, i.e.

$$\tilde{f}(\boldsymbol{\omega}) = \sum_{l\geq 0}\sum_{m\leq l}\tilde{F}_{lm}Y_{lm}(\boldsymbol{\omega}), \quad (2)$$

where $\tilde{F}$ is the spherical Fourier transformed signal and Y_{lm} are the so called spherical harmonics of degree $l \in \mathbb{N}_0$, order $m \in [0, l] \in \mathbb{N}_0$ and form an orthonormal basis over $L^2(S^2)$. For details on the spherical Fourier transform, we refer to the seminal work of Kostelec et al. [35].

Before performing the spectral analysis, we create a fused spectrum by combining each of the transformed modalities. Thereby, we compute the spectrum of each modality and select the modality's coefficient for the fused spectrum with the highest local power per degree [36]. The fused spectrum serves then as an input to the remaining part of the pipeline.

The correlation of two functions yields a measure of how strongly two functions are related and forms our basis for finding the correct match. Therefore, at this point we estimate the power spectrum for each candidate and query. Generally, the power spectrum $S_{\tilde{f}\tilde{f}}$ of a function $\tilde{f}$ is defined as the integral of $\tilde{f}$ squared over the spherical space. Concretely, for a given degree l, the power spectrum of $\tilde{f}$ is given by

$$S_{\tilde{f}\tilde{f}}(l) = \sum_{m=0}^{l}\tilde{F}_{lm}\cdot\tilde{F}^*_{lm} = \sum_{m=0}^{l}\mathrm{Re}(\tilde{F}_{lm})^2 + \mathrm{Im}(\tilde{F}_{lm})^2, \quad (3)$$

where $\tilde{F}^*$ is the complex conjugate of $\tilde{F}$. Similar, the cross power spectrum of two arbitrary functions $\tilde{f}$ and $\tilde{g}$ is defined by

$$S_{\tilde{f}\tilde{g}}(l) = \sum_{m=0}^{l}\tilde{F}_{lm}\cdot\tilde{G}^*_{lm}. \quad (4)$$

Given the global power $S_{\tilde{f}\tilde{f}}$, $S_{\tilde{g}\tilde{g}}$ and the cross-power $S_{\tilde{f}\tilde{g}}$ spectra, we define the correlation Q of two functions $\tilde{f}$ and $\tilde{g}$ for degree l as

$$Q(l) = \frac{S_{\tilde{f}\tilde{g}}(l)}{\sqrt{S_{\tilde{f}\tilde{f}}(l)\cdot S_{\tilde{g}\tilde{g}}(l)}}. \quad (5)$$

The correlation in eq. 5 forms the fundamental theory of a global spectral analysis in the spherical harmonic domain.

Generally, a traditional, global Fourier cross-spectral analysis is often biased in terms of larger variances due to leakage. Thomson's pioneering work [37] proposes an approach for alleviating biases by using multiple localized windows (tapers) to relate to the global spectrum. The main idea is to average a set of direct spectrum estimators using pairwise orthogonal tapers resulting in a less biased estimate of the power spectrum. Generalizing the theory of the Cartesian multitaper analysis to the sphere [38], [39] enables our approach to examine the place candidates based on their cross-spectral energy. For a detailed derivation of the spherical tapers, we refer the interested reader to the work of Wieczorek et al. [38], [40].

Initially, we create a set of tapers $h_1, ..., h_n$ to use during the complete candidate voting process. We first calculate the spectra $S_{\tilde{f}\tilde{f}}$, $S_{\tilde{g}\tilde{g}}$ and $S_{\tilde{f}\tilde{g}}$ using eq. 3 and eq. 4, respectively. Next, for each taper h_i, we calculate the respective windowed version of the global power spectra and cross-spectrum [40]. Subsequently, using the windowed spectra, we compute the correlation Q as in eq. 5 for each h_i. The average over all correlations yields our measure on how related the query and the candidate are.

As a final step, we infer a confidence value based on the averaged correlation. Generally, the candidate with the highest total confidence is selected as the final estimate for the robot's place. The probability that two functions correlate is expressed as bivariate normal distribution [41] using the correlation coefficient $Q(l)$, i.e.

$$G_1(Q, 1) = Q(1)$$
$$G_l(Q, l) = Q_{l-1} + Q(l)(1 - Q(l)^2)^{l-1}\prod_{i=1}^{l-1}\frac{2i-1}{2i}. \quad (6)$$

Analogous to a standard Fourier transform, most of the transformed signal's energy is concentrated in the lower bands. Hence, we solely utilize the first 15 spherical harmonic degrees to evaluate a correlation coefficient in eq. 5 and consequently, the confidence using eq. 6. The resulting confidence values are converted to z-score values using the inverse of the Cumulative Distribution Function (CDF) of a normal distribution. In more detail, given a confidence value

$g \in [0, 1]$ we infer the z-score s_g using

$$s_g = \Phi^{-1}\left(\frac{1-(1-g)}{2}\right), \tag{7}$$

where Φ is the CDF of $\mathcal{N}(0, 1)$. The z-score down-weighs correlations with less than $50\,\%$ and up-weighs correlations with more than $50\,\%$ confidence. We accumulate the z-scores for each degree l and the maximum of the accumulated values represents the place with the highest confidence based on the correlation and consequently our best match.

Thus, the spectral analysis and place voting constitutes our proposed place recognition pipeline.

IV. Experiments

This section evaluates our proposed spherical place recognition pipeline denoted as *S²Loc* using different sensory systems and environments. First, we discuss our used sensor setup and the data generation process. Then, the experimental evaluation of the learned descriptor and our proposed spectral place voting are presented.

We compare our approach to two current state-of-the-art solutions: NetVLAD [7] for image-based place recognition and OverlapNet [20] for laser-based place recognition. NetVLAD was set to treat each image independently for multi-camera datasets, and we configured OverlapNet to get the same information as our network, i.e., to use range and intensity information solely.

A. Sensor Setup and Data Generation

Generally, for each test environment, we created a global multi-session map from multiple runs and jointly optimized it using constraints from visual landmarks, LiDAR scan-to-scan matches, and RTK GPS. Our high-fidelity (HF) sensor setup comprises an Ouster OS-0 with 128 beams (131072 points per scan) and four $0.4\,\mathrm{MP}$ global shutter grayscale cameras (two forward, one to each side). The low-fidelity (LF) setup contains an Ouster OS-1 with 64 beams (65535 points per scan) and a single forward-facing $0.4\,\mathrm{MP}$ global shutter grayscale camera.

As discussed in Sec. III-B, our triplet network approach requires three samples per training input, an anchor, a positive and a negative sample. Positive candidates for training are extracted using a proximity search on the individual poses. Precisely, we extract positive intra- and inter-mission poses for each pose if their Euclidean distances are less than $5\,\mathrm{m}$. Similar, negative training candidates are extracted with a Euclidean distance of 6-20 m. Furthermore, we avoid clusters in the training and test data by constraining each sample to be at least $10\,\mathrm{cm}$ away from each other. OverlapNet and NetVLAD were given the same training data as our network.

Moreover, the data used for training and testing comprises an outdoor and indoor environment. Both environments were recorded with both sensor configurations using a handheld device [42]. The recordings of the LF and HF setup are roughly one year apart for both datasets.

Outdoor environment. We utilized a search and rescue testing facility (cf. Figure 1) that includes several multi-floor buildings, urban-like streets, and collapsed structures. The training, test split was $8\,\mathrm{km}/2\,\mathrm{km}$ and $2\,\mathrm{km}/1\,\mathrm{km}$ for the HF and LF map, respectively. Additionally, we ensured that the training and test data was not recorded on the same day.

Indoor environment. In total, the data covers $1\,\mathrm{km}$ of recordings in a building and is solely used for cross-modality tests, i.e., a HF map and LF queries. This data is never used for training.

Evaluation metric. We use the *recall@n* metric as the basis of our evaluation and comparison to the other approaches. Here, *n* refers to the number of nearest neighbors (database candidates) retrieved during the lookup, and *recall* refers to the correctly identified places w.r.t. the whole map. For OverlapNet, we created a pointcloud map and retrieved the top-n candidates in terms of their highest overlap.

B. Descriptor Lookup

This section investigates the accuracy and precision of our learned descriptor. Specifically, for a given descriptor map, we perform a nearest neighbor lookup and consider it as successful if one of the neighbors is within $5\,\mathrm{m}$ of the query sample.

Descriptor matching. This experiment aims to validate our descriptor's effectiveness when employing it on the same and different hardware from the outdoor environment (cf. Figure 4). In concrete, we compare different train and test configurations to validate the effectiveness of our multi-modal descriptor and show its versatile applicability. The

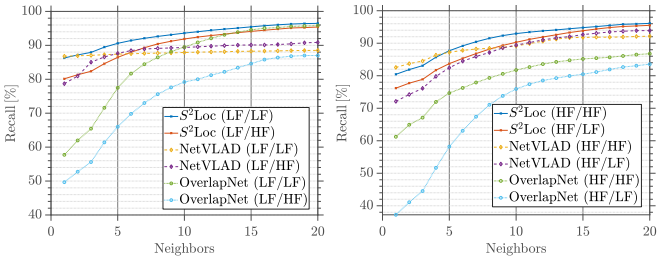

Fig. 4. Recall of the descriptor lookup in the outdoor environment. Each test set comprises 3000 places. The parentheses denote the used sensory system for training and testing, i.e. (train/test).

test data is randomly sampled and contains arbitrary large rotational differences between map and query, making the LF data particularly hard for NetVLAD as there was only one camera used. As a result, NetVLAD's recall is competitive for less nearest neighbors but does not improve drastically for a higher number of retrievals. Furthermore, the data's diversity results in many samples exceeding our success threshold but are still within high overlap to several queries. Consequently, OverlapNet's recall results in a steady improvement with the number of retrieved neighbors. Our approach benefits from both modalities and, especially for more neighbors, generalizes well on the different sensory systems. The visual data improves the retrieval when the LiDAR does not have reasonable good beam returns. Similarly, the LiDAR supports the retrieval when the visual counterpart suffers greatly from viewpoint and illumination changes. Moreover, our approach gains additional efficiency from

spherical CNN's rotational robustness, which we validate next along with the fact that our approach also works well with a single modality.

Descriptor matching on rotated samples. Next, we will investigate the descriptor matching when the data is corrupted with rotations. Generally, in place recognition, LiDAR scans used to build and query a map can be arbitrary rotated to each other. This experiment confirms that our approach is resilient against arbitrary rotations by corrupting the map with rotations around yaw. Figure 5 illustrates the evaluation of rotational shifts from 0° to 180°. These results indicate

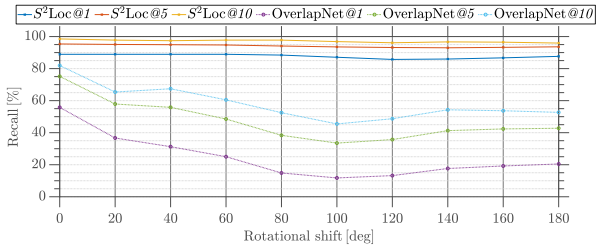

Fig. 5. Evaluation of the descriptor lookup for the LiDAR-only case. The map was built with artificially rotated pointclouds and queries remained unchanged. Results are shown over a 1000 places with the *recall@n* metric. The map was built and queried with the HF setup and both networks were trained on HF data.

our spherical projection and the spherical CNN generalize well on the rotated data and essentially account for almost no degradation in the recall. In contrast, OverlapNet significantly decreases with increasing yaw perturbations since rotations around yaw result in shifts for planar projection models. Next, we test the place selection part of our proposed pipeline.

C. Place Recognition

This section solely evaluates our proposed cross-spectral place selection algorithm that takes place after the descriptor lookup. Like the previous experiments, we only consider successful matches if the selected place from the z-score voting is within 5 m of the query location. However, we limited the input to cases where at least one retrieved neighbor is a correct match to evaluate solely the place selection. Figure 6 shows the percentage of selected places greater than 5 m w.r.t. the number of retrieved candidates. Our approach

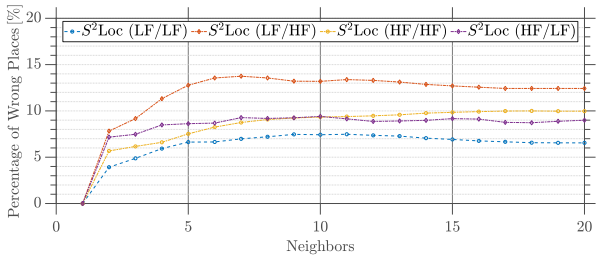

Fig. 6. Percentage of wrongly selected places per database retrievals using our cross-spectral place selection described in Section III-C. The results are shown for 2000 samples from the outdoor dataset.

considerably benefits from the descriptive information in the visual data to distinguish the correct from the wrong places.

Next, we evaluate our place selection using the cross-modal indoor environment where the sensors are switched between building (HF) and querying (LF) the map. Figure 7 illustrates the percentage of the selected places grouped by their distance to the query. The multi-modality and the

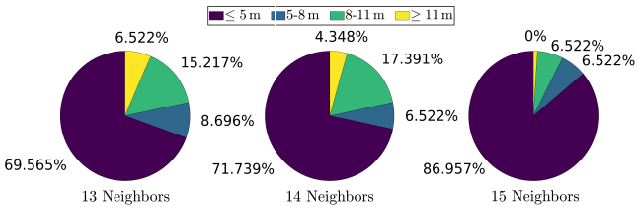

Fig. 7. Distribution of the distances between the selected map places and their respective queries using 13, 14 and 15 neighbors. The map was built with HF and queried with LF sensors from the indoor dataset, whereas the model was trained on the HF outdoor dataset.

correlation with less-biased spectrum estimates allow our approach to successfully distinguish close places as we are not computing correspondences but directly measure how similar the candidate with the query is. Furthermore, with only 15 neighbors, no place further than 11 m away was selected.

D. Performance Evaluation

We run our proposed pipeline on an Intel Xeon E5-2640v3 and an NVIDIA Titan RTX. Figure 8 shows the execution time per individual component. In total, a single sample takes ~300 ms processing time for $\tilde{B} = 100$ when considering 1.06 ms for the KD-tree lookup and disregarding the time needed to build the map.

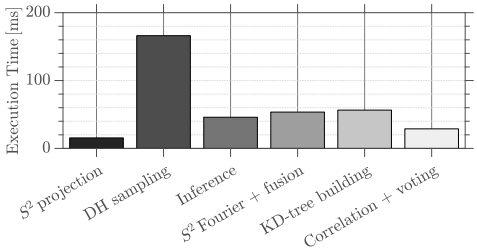

Fig. 8. Execution time in ms partitioned per component. All values are averaged over a 1000 samples.

V. Conclusion and Future Work

This paper presented a learning approach to multi-modal place recognition using vision and LiDAR sensors. Our approach operates end-to-end by projecting each modality onto the hypersphere and using a spherical CNN.

We showed that our multi-modal descriptor improve the state-of-the-art in place recognition, and more important, it generalizes to different sensor systems, in terms of training and deployment, and map building and querying. Additionally, our method benefits from spectral analysis to efficiently distinguish the correct place from the retrieved database candidates.

We will continue our research in two directions. First, the integration of semantic information (e.g. computed from the image) as an extra modality to improve in the place recognition task. Second, we will investigate a spherical decoder to exploit our multi-modal spherical parametrization for the actual task of semantic segmentation under heterogeneous sensor coverages and sensor setups.

REFERENCES

[1] T. S. Cohen, M. Geiger, J. Koehler, and M. Welling, "Spherical CNNs," *Proceedings of the International Conference on Learning Representations*, no. 3, pp. 1–15, 1 2018.

[2] C. Esteves, C. Allen-Blanchette, A. Makadia, and K. Daniilidis, "Learning SO(3) Equivariant Representations with Spherical CNNs," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[3] N. Perraudin, M. Defferrard, T. Kacprzak, and R. Sgier, "DeepSphere: Efficient spherical convolutional neural network with HEALPix sampling for cosmological applications," *Astronomy and Computing*, vol. 27, pp. 130–146, 2019.

[4] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.

[5] D. Galvez-López and J. D. Tardos, "Bags of Binary Words for Fast Place Recognition in Image Sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 10 2012.

[6] L. Zheng, Y. Yang, and Q. Tian, "SIFT Meets CNN: A Decade Survey of Instance Retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1224–1244, 5 2018.

[7] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1437–1451, 2018.

[8] N. Suenderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free," in *Robotics: Science and Systems XI*, vol. 11. Robotics: Science and Systems Foundation, 7 2015.

[9] Z. Chen, A. Jacobson, N. Sunderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5 2017, pp. 3223–3230.

[10] S. Hausler, A. Jacobson, and M. Milford, "Multi-Process Fusion: Visual Place Recognition Using Multiple Image Processing Methods," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1924–1931, 4 2019.

[11] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization," in *2015 IEEE International Conference on Computer Vision (ICCV)*, vol. 2015 Inter. IEEE, 12 2015, pp. 2938–2946.

[12] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 Place Recognition by View Synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 257–271, 2 2018.

[13] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "End-to-End Learning of Deep Visual Representations for Image Retrieval," *International Journal of Computer Vision*, vol. 124, no. 2, pp. 237–254, 9 2017.

[14] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3D recognition and pose using the viewpoint feature histogram," *IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings*, pp. 2155–2162, 2010.

[15] T. Rohling, J. Mack, and D. Schulz, "A fast histogram-based similarity measure for detecting loop closures in 3-D LIDAR data," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2015-Decem, pp. 736–741, 2015.

[16] R. Dubé, A. Cramariuc, D. Dugas, H. Sommer, M. Dymczyk, J. Nieto, R. Siegwart, and C. Cadena, "SegMap: Segment-based mapping and localization using data-driven descriptors," *International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 339–355, 2020.

[17] W. Zhang and C. Xiao, "PCAN: 3D Attention Map Learning Using Contextual Information for Point Cloud Based Retrieval," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2019-June. IEEE, 6 2019, pp. 12 428–12 437.

[18] J. Du, R. Wang, and D. Cremers, "DH3D: Deep Hierarchical 3D Descriptors for Robust Large-Scale 6DoF Relocalization," 2020.

[19] H. Yin, L. Tang, X. DIng, Y. Wang, and R. Xiong, "LocNet: Global Localization in 3D Point Clouds for Mobile Vehicles," *IEEE Intelligent Vehicles Symposium, Proceedings*, vol. 2018-June, no. Iv, pp. 728–733, 2018.

[20] X. Chen, T. Läbe, A. Milioto, T. Röhling, O. Vysotska, A. Haag, J. Behley, and C. Stachniss, "OverlapNet: Loop Closing for LiDAR-based SLAM," *Robotics: Science and Systems (RSS)*, no. July, 2020.

[21] L. Schaupp, M. Burki, R. Dube, R. Siegwart, and C. Cadena, "OREOS: Oriented Recognition of 3D Point Clouds in Outdoor Scenarios," *IEEE International Conference on Intelligent Robots and Systems*, pp. 3255–3261, 2019.

[22] M. A. Uy and G. H. Lee, "PointNetVLAD: Deep Point Cloud Based Retrieval for Large-Scale Place Recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, no. c. IEEE, 6 2018, pp. 4470–4479.

[23] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12 2017.

[24] J. Guo, P. V. Borges, C. Park, and A. Gawel, "Local Descriptor for Robust Place Recognition Using LiDAR Intensity," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1470–1477, 2019.

[25] K. P. Cop, P. V. Borges, and R. Dube, "Delight: An Efficient Descriptor for Global Localisation Using LiDAR Intensities," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 3653–3660, 2018.

[26] S. Ratz, M. Dymczyk, R. Siegwart, and R. Dube, "OneShot Global Localization: Instant LiDAR-Visual Pose Estimation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5 2020, pp. 5415–5421.

[27] S. Xie, C. Pan, Y. Peng, K. Liu, and S. Ying, "Large-scale place recognition based on camera-lidar fused descriptor," *Sensors (Switzerland)*, vol. 20, no. 10, pp. 1–21, 2020.

[28] D. Cattaneo, M. Vaghi, S. Fontana, A. L. Ballardini, and D. G. Sorrenti, "Global visual localization in LiDAR-maps through shared 2D-3D embedding space," 2019.

[29] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard, "Monocular camera localization in 3D LiDAR maps," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2016-Novem, pp. 1926–1931, 2016.

[30] M. Feng, S. Hu, M. H. Ang, and G. H. Lee, "2D3D-Matchnet: Learning To Match Keypoints Across 2D Image And 3D Point Cloud," in *2019 International Conference on Robotics and Automation (ICRA)*, vol. 2019-May. IEEE, 5 2019, pp. 4790–4796.

[31] J. R. Driscoll and D. M. Healy, "Computing fourier transforms and convolutions on the 2-sphere," pp. 202–250, 1994.

[32] T. Cohen, M. Geiger, J. Köhler, and M. Welling, "Convolutional Networks for Spherical Signals," *Principled Approaches to Deep Learning Workshop, ICML*, 9 2017.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 1026–1034.

[34] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 1335–1344, 2016.

[35] P. J. Kostelec and D. N. Rockmore, "FFTs on the Rotation Group," *Journal of Fourier Analysis and Applications*, vol. 14, no. 2, pp. 145–179, 4 2008.

[36] H. Falk, "Prolog To A Categorization Of Multiscale-decomposition-based Image Fusion Schemes With A Performance Study For A Digital Camera Application," *Proceedings of the IEEE*, vol. 87, no. 8, pp. 1313–1314, 8 1999.

[37] D. J. Thomson, "Spectrum Estimation and Harmonic Analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, 1982.

[38] M. A. Wieczorek and F. J. Simons, "Minimum-variance multitaper spectral estimation on the sphere," *Journal of Fourier Analysis and Applications*, vol. 13, no. 6, pp. 665–692, 2007.

[39] F. J. Simons, F. A. Dahlen, and M. A. Wieczorek, "Spatiospectral concentration on a sphere," *SIAM Review*, vol. 48, no. 3, pp. 504–536, 2006.

[40] M. A. Wieczorek and F. J. Simons, "Localized spectral analysis on the sphere," *Geophysical Journal International*, vol. 162, no. 3, pp. 655–675, 2005.
[41] M. Pauer, K. Fleming, and O. Čadek, "Modeling the dynamic component of the geoid and topography of Venus," *Journal of Geophysical Research E: Planets*, vol. 111, no. 11, pp. 1–18, 2006.
[42] F. Tschopp, M. Riner, M. Fehr, L. Bernreiter, F. Furrer, T. Novkovic, A. Pfrunder, C. Cadena, R. Siegwart, and J. Nieto, "VersaVIS—An Open Versatile Multi-Camera Visual-Inertial Sensor Suite," *Sensors*, vol. 20, no. 5, p. 1439, 3 2020.