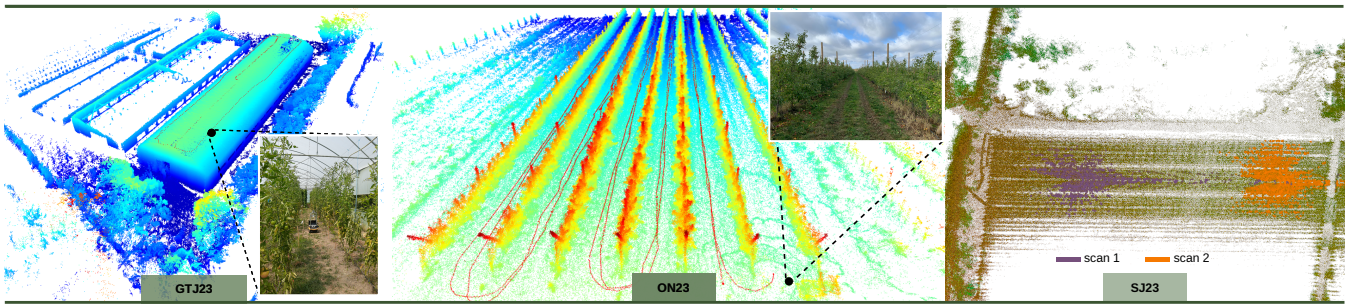
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

SPVSoAP3D: A Second-order Average Pooling Approach to enhance 3D Place Recognition in Horticultural Environments

T. Barros[1], C. Premebida[1], S. Aravecchia[2], C. Pradalier[2], U.J. Nunes[1]



(a) (b) (c)

Fig. 1: 3D maps of the two new sequences from horticultural environments proposed in this work are as follows: (a) GTJ23, recorded in a greenhouse tomato production facility in Coimbra, Portugal, and (b) ON23, recorded in an orchard in Metz, France. Additionally, (c) illustrates a 3D map of an HORTO-3DLM sequence (SJ23) with two scans, showcasing the permeable nature of these environments to LiDAR scans.

***Abstract*— 3D LiDAR-based place recognition has been extensively researched in urban environments, yet it remains underexplored in agricultural settings. Unlike urban contexts, horticultural environments, characterized by their permeability to laser beams, result in sparse and overlapping LiDAR scans with suboptimal geometries. This phenomenon leads to intra- and inter-row descriptor ambiguity. In this work, we address this challenge by introducing SPVSoAP3D, a novel modeling approach that combines a voxel-based feature extraction network with an aggregation technique based on a second-order average pooling operator, complemented by a descriptor enhancement stage. Furthermore, we augment the existing HORTO-3DLM dataset by introducing two new sequences derived from horticultural environments. We evaluate the performance of SPVSoAP3D against state-of-the-art (SOTA) models, including OverlapTransformer, PointNetVLAD, and LOGG3D-Net, utilizing a cross-validation protocol on both the newly introduced sequences and the existing HORTO-3DLM dataset. The findings indicate that the average operator is more suitable for horticultural environments compared to the max operator and other first-order pooling techniques. Additionally, the results highlight the improvements brought by the descriptor enhancement stage. The code is publicly available at https://github.com/Cybonic/SPVSoAP3D.git**

I. INTRODUCTION

Autonomous robots are pivotal in enhancing productivity within the agricultural sector [1], [2]. To operate reliably, these robots require precise and accurate localization systems. In GNSS-denied environments, perception-based methodologies such as Simultaneous Localization and Mapping (SLAM) and place recognition offer viable solutions. Among these, 3D LiDAR-based place recognition has gained prominence both as a component of SLAM and as a standalone method, providing an efficient approach for global localization and loop detection [3].

3D LiDAR-based place recognition has primarily been advanced by the autonomous driving community, which has introduced novel methodologies specifically designed for urban-like environments [4]. Recent efforts have focused on adapting these methods to natural and agricultural environments [5], [6], which pose fundamentally different challenges. Unlike urban settings that return scans with well-defined geometries (e.g., corners, planes, lines), agricultural, and particularly horticultural environments, produce LiDAR scans with poor geometries. The permeable nature of these environments to laser beams allows the beams to traverse various row layers, resulting in scans that cover multiple rows (illustrated in Fig 1.c). Consequently, this generates sparse scans with poor geometries and highly overlapped scans from neighboring rows, leading to intra- and inter-row descriptor ambiguity.

This work addresses the challenge of 3D LiDAR place recognition, specifically focusing on the inter-row descriptor ambiguity within horticultural environments, by making two key contributions. Firstly, we introduce two new curated sequences comprising 3D LiDAR and GNSS/localization data, thereby enhancing the existing HORTO-3DLM dataset*. These sequences add diversity and representativeness to the dataset; one sequence was recorded in an apple orchard

[1] T. Barros, C. Premebida and U.J.Nunes are with the University of Coimbra, Institute of Systems and Robotics, Department of Electrical and Computer Engineering, Portugal. E-mails: {tiagobarros, cpremebida, urbano}@isr.uc.pt

[2] S. Aravecchia and C. Pradalier are with the IRL2958 GeorgiaTech-CNRS, Metz 57070, France. E-mail: {stephanie.aravecchia, cedric.pradalier}@georgiatech-metz.fr

*https://github.com/Cybonic/HORTO-3DLM.git

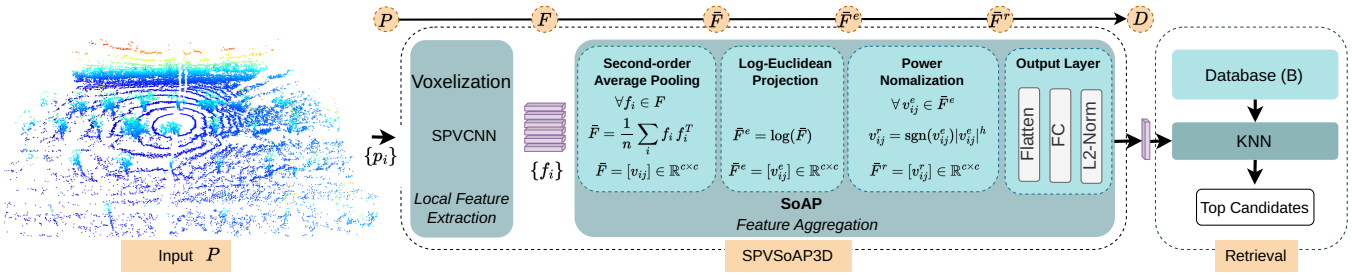


Fig. 2: SPVSoAP3D in a retrieval-based framework. SPVSoAP3D has 5 main stages. In **stage 1**, the input 3D LiDAR scan P is voxelized and fed to the backbone SPVCNN, which returns local features F. In **stage 2**, these local features are aggregated using second-order average pooling, resulting in $\bar{F}$. In **stage 3**, the aggregated features $\bar{F}$ are projected to a tangent space using Log-Euclidean projection, yielding $\bar{F}^e$. In **stage 4**, the $\bar{F}^e$ features are rescaled using power normalization. In **stage 5**, the rescaled features $\bar{F}^r$ are flattened and fed to a fully connected layer followed by $L2$-norm. Finally, the model outputs a descriptor D, which is used to query the database for the top-k most similar descriptors.

in Metz, France, using a 16-beam 3D LiDAR mounted on a mobile platform, and the other in a greenhouse in Coimbra, Portugal, employing a 64-beam 3D LiDAR on a different mobile platform. Figure 1 illustrates 3D maps of each sequence.

Secondly, we propose a novel 3D LiDAR-based place recognition method named SPVSoAP3D. SPVSoAP3D uses a sparse point-voxel-based backbone to extract local features from 3D LiDAR scans and utilizes a second-order average pooling approach to aggregate these features into a descriptor. Our contribution lies particularly in the adoption of the average operator as an aggregation approach for 3D LiDAR-based frameworks and a descriptor enhancement approach based on Log-Euclidean projection and power normalization. The proposed approach is illustrated in Fig. 2.

The proposed approach was evaluated using the HORTO-3DLM dataset, along with two additional sequences, benchmarking SPVSoAP3D against state-of-the-art (SOTA) models such as OverlapTransformer [7], PointNetVLAD [8], and LOGG3D-Net [9] using a cross-validation protocol, specifically the *leave-one-out* approach. The empirical results show that average pooling outperforms the second-order max pooling approach used in LOGG3D-Net [9], as well as other first-order pooling methods employed in SOTA 3D LiDAR place recognition models. Additionally, the enhancement stage further improves performance.

II. RELATED WORK

Over the past few years, the majority of advancements in 3D LiDAR place recognition have been proposed for urban environments [10], [11]. These advancements have been mainly driven by Deep Neural Networks (DNNs), which have enabled the learning of features directly from point clouds in an end-to-end fashion [4]. A less studied field is place recognition in agricultural or natural environments, where efforts have recently been directed towards adapting urban-based approaches for these semi-structured environments. Despite these efforts, such endeavors face significant challenges due to the limited availability of datasets, given the substantial amounts of data required to train DL-based approaches.

A. Datasets

The autonomous driving community has extensively utilized well-established urban environment datasets to advance 3D LiDAR-based place recognition. Datasets such as KITTI [12], KITTI360 [13], NCLT [14], Oxford [15], and MulRan [16] are just a few examples of popular benchmarks used in both SLAM and place recognition tasks. The diversity in these datasets in terms of environment and sensor characteristics (e.g., resolution, range, or precision) make them well-suited for training and evaluating DL-based place recognition approaches.

In contrast, natural and agricultural environments possess different characteristics compared to on-road urban environments. In recent years, new LiDAR datasets from these environments have emerged to bridge this gap. Datasets such as Wild-places [17], RELLIS-3D [18], ORFD [19], TreeScope [6], and VineLiDAR [20] exemplify these efforts to provide multi-modality real-world sensory data of natural and agricultural environments for localization, mapping, and particularly for place recognition tasks.

Despite these efforts, there remains a lack of representative data from environments such as horticulture. Therefore, one of the objectives of this work is to provide real-world sensory data from such environments, specifically from two distinct settings: an apple orchard and a tomato plantation within a greenhouse. These two sequences will complement an existing horticulture dataset called HORTO-3DLM, which includes 3D LiDAR sensory data from orchards of apples and strawberries.

B. 3D LiDAR-based Place Recognition

Although hand-crafted descriptors are still used, Deep Learning (DL) techniques have become increasingly popular for place modeling [4]. Among these techniques, most 3D LiDAR-based place recognition approaches rely on DNNs, typically comprising two main modules: a backbone for extracting local features from input point clouds, and a feature

aggregation module for combining these local features into a global descriptor.

In this work, we focus on the feature aggregation module, an area extensively explored in vision perception tasks such as classification [21], [22] and segmentation [23], where first-order and second-order statistics are commonly compared. These methods are less studied in 3D LiDAR-based place recognition. Many works in the field resort to first-order methods such as NetVLAD [8], [24], [25], or generalized mean approaches [26], while higher-order pooling receive less attention. Recent works such as LOGG3D-Net [9] and Locus [27] explore second-order pooling utilizing the max operator to aggregate features.

Our approach builds upon the concept of second-order pooling, proposing the use of the average operator instead of the max operator. We show that this operator serves as a more suitable feature aggregator for second-order features in horticultural environments than the max pooling operator. Our findings align with those in vision-related research [21], [22], which have shown the superiority of second-order average pooling, for instance, in image segmentation tasks [23]. Furthermore, inspired by the work in [23], we additionally project the average representation into a Log-Euclidean tangent space and apply power normalization. Our results indicate that these additional stages further enhance performance.

III. PROPOSED APPROACH

This section details the proposed SPVSoAP3D model within a retrieval-based place recognition framework as shown in Fig. 2. In this work, we frame the place recognition problem as a retrieval tasks, evaluated at the segment level, meaning that the plantations are divided into segments. For instance, each row constitutes a segment, and the extremities of the plantations (perpendicular to the rows) are also considered segments. This segmentation approach enables the evaluation of performance at the segment level.

A. Problem Formulation

A retrieval-based 3D LiDAR place recognition framework comprises two stages: first, a LiDAR scan $P \in \mathbb{R}^{n\times 3}$ with n points, which is mapped to a descriptor $D \in \mathbb{R}^d$, using a function $\Theta : \mathbb{R}^{n\times 3} \rightarrow \mathbb{R}^d$; second, the descriptor D queries a database for the top-k candidates based on a similarity metric. The retrieved candidates represent potential revisited places. The modeling capacity of Θ directly impacts the retrieval performance.

This work focuses on the mapping function Θ, which can be decomposed into two functions, $\Theta \equiv \varphi \circ \phi$, where $\phi : \mathbb{R}^{n\times 3} \rightarrow \mathbb{R}^{n\times c}$ maps an input scan P (with 3 dimensions) to a hyper-dimensional feature vector $F = \{f_i\} \in \mathbb{R}^{n\times c}$ with c dimensions; and $\varphi : \mathbb{R}^{n\times c} \rightarrow \mathbb{R}^d$ aggregates the features F into a global descriptor D with d dimensions.

B. SPVSoAP3D

The proposed place modeling approach, SPVSoAP3D, inspired by LOGG3D-Net [9], can be summarized in five main stages: (1) local features extraction; (2) second-order average pooling; (3) Log-Euclidean tangent space projection; (4) power normalization; and (5) linear projection using a fully connected layer (FC) and descriptor normalization. The model is trained using a contrastive learning approach employing a triplet loss for parameter optimization.

1) Local Feature Extraction: Local features are extracted with the Sparse Point-Voxel (**SPV**) CNN [28], which serves as a backbone. Initially, the input scan $P = \{(x, y, z)_i\} \in \mathbb{R}^{n\times 3}$ is converted to a voxel representation $V \in \mathbb{R}^{n'\times 3}$ with n' occupied voxels where $n' < n$. This voxel representation is then processed by the SPVCNN backbone to extract the local features $F = \{f_i\} \in \mathbb{R}^{n'\times c}$, consisting of n' features, and each with c dimensions. Details on the voxelization process and on SPVCNN can be found in [28].

2) Second-order Average Pooling: Local features are aggregated by averaging over second-order interactions (*e.g.*, outer products) [23]. Thus, given a set of local features F, the second-order average pooling operation is defined as follows:

$$\bar{F} = \frac{1}{n'} \sum_{f_i \in F} f_i \, f_i^T, \tag{1}$$

where $\bar{F} \in \mathbb{R}^{c\times c}$ is the aggregated average representation.

3) Log-Euclidean Projection: In this work, the descriptors (output of the model) are compared using the L_2-norm, which is an Euclidean metric. However, $\bar{F}$ leads to a symmetric positive definite (SPD) matrix, forming a Riemannian manifold *i.e.*, a non-Euclidean space [23], [29], [30]. The geometric relationships in the manifold can be preserved by mapping the SPDs to a Euclidean tangent space. This mapping can be achieved using a principal matrix logarithm operation [23] as follows,

$$\bar{F}^e = \log(\bar{F}) \in \mathbb{R}^{c\times c}. \tag{2}$$

The matrix logarithm $\log(\cdot)$ is implemented as described in [31], [32], it is obtained by computing the logarithm of the eigenvalues *s.t.* $\log(\bar{F}) = U \log(\Sigma) V^T$ with Σ being the diagonal matrix containing the eigenvalues of $\bar{F}$. The eigenvalues are computed using the SVD decomposition.

4) Power Normalization: The descriptiveness of the features can be improved reducing the sparsity [23]. One approach to reducing sparsity is through power normalization, which rescales feature values by amplifying smaller values, while saturating larger ones [33]. Given $\bar{F}^e = [v_{ij}] \in \mathbb{R}^{c\times c}$, the rescaling is computed using element-wise power normalization as follows:

$$v_{ij}^r = \text{sgn}(v_{ij})|v_{ij}|^h, \; \forall \, v_{ij} \in \bar{F}^e \tag{3}$$

where v_{ij}^r is the rescaled feature element. The rescaled representation is $\bar{F}^r = [v_{ij}^r] \in \mathbb{R}^{c\times c}$, and $h \in [0, 1]$ is the power parameter. In contrast to previous works such as [23], [33], where h is fixed, here h is a trainable parameter.

5) Output Layer: Before obtaining the descriptor $D \in \mathbb{R}^d$ with d dimensions, $\bar{F}^r \in \mathbb{R}^{c\times c}$ is flattened and fed to a fully-connected layer $FC : \mathbb{R}^{c^2} \rightarrow \mathbb{R}^d$, and normalized using the L_2-norm.

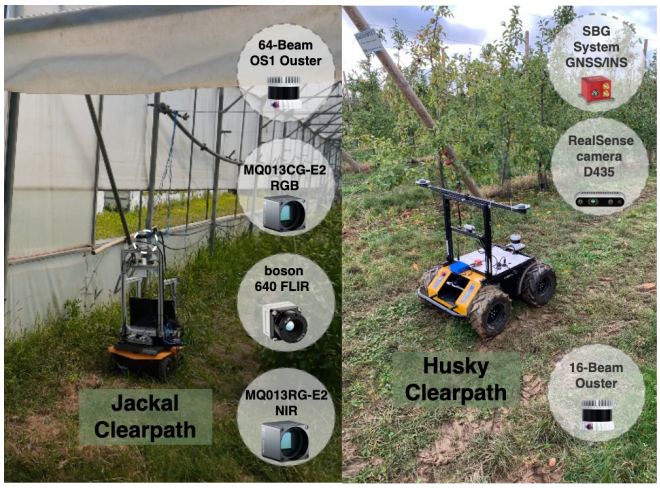


Fig. 3: Mobile platforms and sensors used for recording the sequences. The Jackal platform was used to record the GTJ23, while the Husky platform was used to collect data for the ON23 dataset.

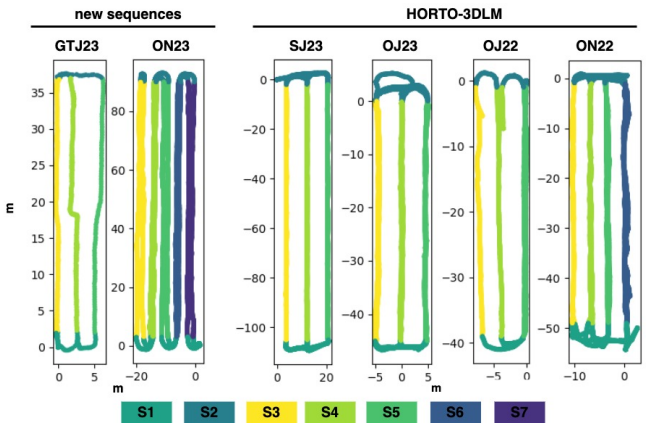


Fig. 4: Paths with segments identified by colored S1,...,S7. GTJ3, SJ23, OJ23 and OJ22 have 5 segments, ON22 has 6 segments, while ON23 has 7 segments.

C. Network Training

SPVSoAP3D is trained using the LazyTriplet loss as proposed in [8], which employs tuples of scans. Given a tuple $\tau_i = (P_a, P_p, \{P_{n_i}\})$, P_a and P_p form an anchor-positive pair, and $\{P_{n_i}\}_{i=1}^{m}$ is a set of negative scans.

As in traditional place recognition, the anchor-positive pair must come from the same place. However, in this work, we define "same place" differently by adding a third constraint. Thus, the anchor-positive pair selection is based on the following three constraints (where the third is our additional constraint): (1) both scans must be within a threshold range r_{th} (in meters); (2) both scans must be from different revisits, meaning that the positive scan cannot be the immediate previous scan of the anchor; (3) both scans must be from the same segment (see Fig. 4). All scans that do not satisfy these three constraints are considered negatives.

To compute the loss, $\tau_i = (P_a, P_p, \{P_{n_i}\})$ is mapped to the respective descriptors $\hat{\tau_i} = (D_a, D_p, \{D_{n_i}\})$, which are employed in the loss as follows:

$$\mathcal{L}_T = \max(d_{AP} - d_{AN} + m, 0), \tag{4}$$

where $d_{AP} = \|D_a - D_p\|_2$ is the Euclidean distance between the anchor and the positive, and $d_{AN} = \min_i(\|D_a - \{D_{n_i}\}\|_2)$ is the Euclidean distance between the anchor and the hardest negative, *i.e.*, the negative that is closest in the descriptor space. Finally, m designates a margin value.

IV. EXPERIMENTAL EVALUATION

The proposed approach was evaluated on data from horticultural environments, utilizing two new sequences, as well as the HORTO-3DLM dataset.

A. HORTO-3DLM+

The HORTO-3DLM dataset comprises four sequences from distinct horticultural environments: three from orchards (two from apples and one from cherries), and another sequence from strawberries within polytunnels with a table-top growing system. These sequences were recorded in England, UK, using a Clearpath Husky mobile robot equipped with a Velodyne VLP32 3D LiDAR (10Hz) and a ZED-F9P RTK-GPS (5Hz). The objective of this dataset is to provide real-world sensory data from horticultural environments to support research, particularly in localization, mapping, and place recognition.

In this work, we introduce two additional sequences to enhance the original HORTO-3DLM: one sequence recorded in an apple orchard in Metz, France (referred to as ON23); and another recorded in a greenhouse tomato production in Coimbra, Portugal (referred to as GTJ23). Each sequence was recorded with a different setup, illustrated in Fig. 3.

Sequence ON23 was captured in November of 2023 with an 16-beam Ouster 3D LiDAR and an SBG GNSS/INS system (without RTK) mounted on a Clearpath Husky mobile platform. To address the low LiDAR resolution, the original scans were merged to increase point density, resulting in sub-maps with approximately 100k points per sub-map. This operation reduced the original sequence from 25836 scans to 3086 sub-maps in total. Both SLAM and GNSS positioning data are provided, and the synchronization between the sub-maps and the positioning data was achieved by associating the nearest neighbor timestamps in each modality, using the first scan in the sub-maps as the timestamp reference.

On the other hand, sequence GTJ23 was recorded in June of 2023 using a 64-beam Ouster 3D LiDAR mounted on a Clearpath Jackal mobile platform. In this sequence the the ground-truth positions were computed using a SLAM approach due to signal interference caused by the greenhouse structure.

Figure 1 outlines the 3D maps of each sequence and their respective environments, while Fig. 3 illustrates the two mobile platforms and the sensors used in the recordings. The details of the augmented dataset HORTO-3DLM+, which comprises the original four sequences and the two new

TABLE I: HORTO-3DLM+: summary of the sequences used in the experiments. The upper four rows represent Seqs. from the HORTO-3DLM dataset, while the bottom two rows represent the two new Seqs. The 'Seq.' column contains the sequence names. The 'M', 'Y', and 'C' columns refer to the month, year, and country of recording, respectively. The total distance ('Dist') of each sequence is measured in meters, while the scan size is given by the number of points in each scan.

Seq.	M	Y	C	Nº Scans	Nº Rows	Dist. [m]	Scan Size	Plantation Type
ON22	Nov.	2022	UK	7974	4	514	48k	Apple (open)
OJ22	July	2022	UK	4361	3	206	50k	Apple (open)
OJ23	June	2023	UK	7229	3	459	46k	Cherry (open)
SJ23	June	2023	UK	6389	3	742	48k	Strawberry (polytunnels)
ON23	Nov.	2023	FR	3086	5	966	105k	Apple (open)
GTJ23	June	2023	PT	661	3	202	60k	Tomato (greenhouse)

TABLE II: Dataset split used for the evaluation. In leave-one-out cross-validation, a model that is evaluated on a sequence, is trained on the remaining sequences. For instance, to assess a model's performance on sequence OJ22, the same model is trained on OJ23, ON22, SJ23, ON23, and GTJ23.

		Cross-validation	
Sequence	Training anchors in the Seq.	Test anchors in the Seq.	Remaining training anchors combined
OJ22	100	1797	2302
OJ23	512	4801	1890
ON22	495	7367	1907
SJ23	598	3509	1804
ON23	580	1657	1822
GTJ23	117	295	2285

sequences, are presented in Table I, while Fig. 4 outlines the paths of all six sequences.

B. Dataset and Evaluation Protocol

In this work, we divide the path into various segments (S1, S2, ..., S7), as illustrated in Fig. 4. In addition to the rows, each corresponding to a segment, we also consider the extremities (i.e., S1 and S7) as separate segments. This segmentation allows for the imposition of the third constraint defined in Section III-C.

The training set is generated based on the three constraints defined in Section III-C, using a search radius of $2\,m$ (*i.e.*, $r_{th} = 2\,m$) for the anchor-positive pairs. Among all anchor-positive pairs, only the anchors that are at least 0.5 m apart are selected to reduce repetitive scans. For each anchor-positive pair, a set of negative scans is randomly selected from the pool of scans that meet two criteria: (1) they originate from different segments, and (2) they are located outside a 10-meter radius relative to the anchor.

For the evaluation, we employ a cross-validation strategy, specifically the *leave-one-out* protocol, and report the results from two distinct experiments. In the first experiment, we evaluate the model's performance using a fixed radius $r_{th} = 10\,m$ and report the recall@k with $k \in [1, 1\%]$. In the second experiment, we analyze the performance along the segments, *i.e.*, $r_{th} \in [1, ..., l]$, where l represents the segment length, and report recall@k with $k \in [1, 10]$. Table II presents the training and evaluation data generated with the aforementioned conditions.

C. Implementation and Training Details

All models are trained and evaluated under identical conditions. Specifically, all scans are downsampled to 10000 points without cropping. For the proposed model, the downsampled scans are voxelized with a grid size of $0.1m$. The backbone is configured to return features with 16 dimensions (i.e., $c = 16$), and h is initialized at 0.75, while the descriptor size is set to 256 dimensions (i.e., $d = 256$).

Regarding training, all models are trained for 50 epochs, using training tuples consisting of an anchor, the closest positive, and 20 negatives. The margin of the loss is set to 0.5 ($m = 0.5$), and model parameters are optimized using the AdamW optimizer with a learning rate of 0.0001 and a weight decay of 0.0005. Additionally, all state-of-the-art models are implemented based on the original code.

The models are executed on a machine equipped with an AMD Ryzen 9 5900X 12-Core CPU, 64 GB RAM, and a NVIDIA GeForce RTX 3090 GPU. The code is implemented in Python 3.9, utilizing PyTorch with CUDA 11.7.

D. Results & Discussion

In this section, we present and discuss the empirical results. As outlined in Section IV-B, we conducted two experiments comparing SPVSoAP to SOTA models, namely PointNetVLAD, OverlapTransformer, and LOGG3D-Net.

The results of the experiment with a fixed radius ($r_{th} = 10\,m$) are reported in Table III. This table also includes the results of an ablation study, which highlights the contribution of each stage in the aggregation module to the overall performance. The results of the experiment with r_{th} varying along the segments are reported in Fig. 5.

1) Comparison to SOTA: In the experiment with $r_{th} = 10\,m$, the results indicate that SPVSoAP3D, on average, outperforms the SOTA models. It achieves 7.8 percentage points (pp) higher performance than the next best model in top-1 retrieval (PointNetVLAD), and 5 pp in top-1% (LOGG3D-Net). Particularly, when compared to LOGG3D-Net, which shares the same backbone and uses second-order max pooling, the proposed approach performs 40 pp higher in top-1 and 5 pp in top-1%, suggesting that the average operator is a more suitable aggregation approach for horticultural environments than the max operator.

Compared to PointNetVLAD and OverlapTransformer, which rely on first-order statistics, SPVSoAP3D exhibits greater robustness, achieving higher performance on average in both top-1 and top-1% retrieval. SPVSoAP3D only ranks second in the SJ23 sequence to PointNetVLAD in top-1

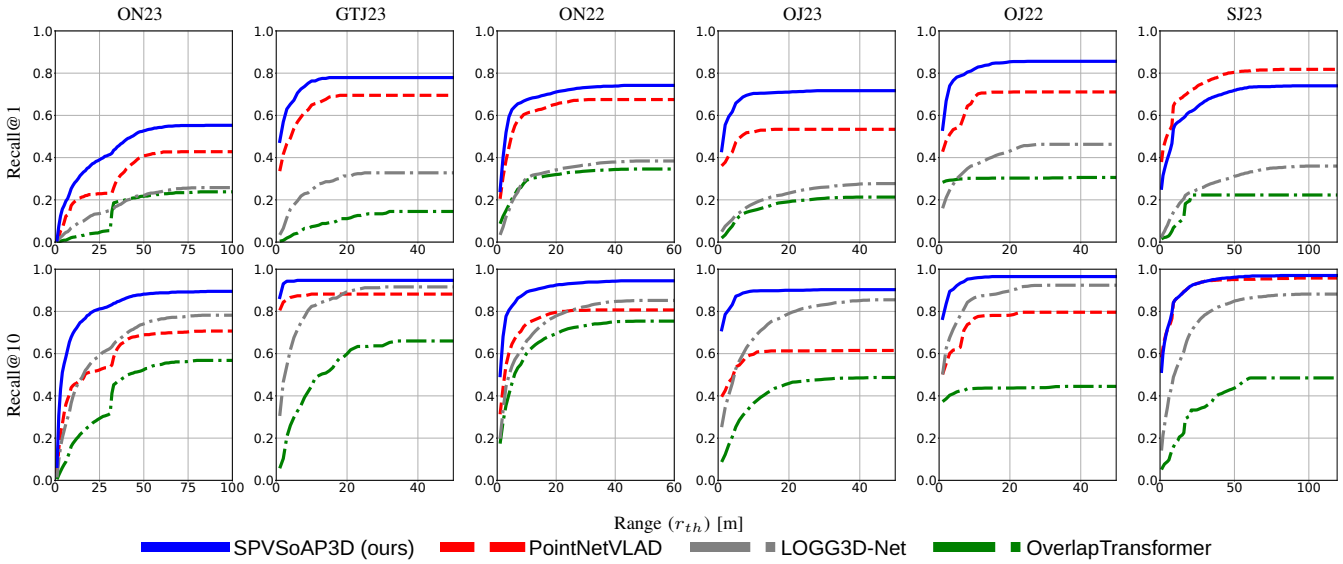


Fig. 5: Performance results along the segments, reported using the Recall@k with $k \in [1, 10]$ for $r_{th} \in [1, .., l]$ where l is the segment length.

TABLE III: Performance results for $r_{th} = 10\,m$ are reported using Recall@k with $k \in [1, 1\%]$. The bottom four rows refer to results from the ablation study, where the row without any marks represents SPVSoAP3D only with second-order average pooling, *i.e.*, until stage 2 (see Fig. 2). 'LOG' refers to stage 3, 'PN' refers to stage 4, while 'FC' refers to stage 5.

				Recall@1							Recall@1%						
				ON23	GTJ23	ON22	OJ23	OJ22	SJ23	MEAN	ON23	GTJ23	ON22	OJ23	OJ22	SJ23	MEAN
PointNetVLAD[8]				0.184	0.649	0.610	0.521	0.686	**0.659**	0.551	0.652	0.920	0.926	0.754	0.873	0.965	0.848
OverlapTransformer[7]				0.023	0.073	0.293	0.150	0.302	0.075	0.153	0.300	0.679	0.854	0.487	0.476	0.203	0.500
LOGG3D-Net[9]				0.074	0.248	0.303	0.182	0.364	0.153	0.221	0.693	0.947	0.975	0.950	0.989	0.928	0.914
SPVSoAP3D	FC	LOG	PN	ON23	GTJ23	ON22	OJ23	OJ22	SJ23	MEAN	ON23	GTJ23	ON22	OJ23	OJ22	SJ23	MEAN
				0.200	0.692	0.467	0.649	**0.862**	0.501	0.562	0.826	**0.997**	0.896	0.978	0.960	0.983	0.940
	✓			0.209	0.740	0.676	0.663	0.817	0.484	0.598	0.829	0.981	0.981	0.978	0.935	0.977	0.947
	✓	✓		0.189	0.721	**0.707**	0.677	0.820	0.541	0.609	0.844	0.966	**0.989**	**0.982**	0.952	0.982	0.953
	✓	✓	✓	**0.268**	**0.763**	0.674	**0.704**	0.820	0.554	**0.630**	**0.887**	0.966	0.988	0.963	**0.994**	**0.986**	**0.964**

retrieval. These results hold true when evaluating along the segments, with all models retrieving more true loops as the range increases along the segments. Upon analyzing the results of recall@10 (second row) in Fig. 5, SPVSoAP3D surpasses PointNetVLAD in SJ23, showcasing the superior modeling capacity of the proposed approach.

2) Ablation Study: This study highlights the contribution of each stage in the aggregation process to the overall performance. The results of this study are presented in the lower section of Table III, where FC refers to stage 5 (*i.e.*, Output Layer) in Fig 2. Regardless of whether FC is utilized, descriptor normalization is always computed. LOG refers to stage 3 (Log-Euclidean projection), while PN refers to stage 4 (power normalization). The first row of the ablation study section in Table III (without any marks) refers to SPVSoAP3D using only second-order average pooling (up to stage 2, included).

The results suggest that the average operator is sufficient to outperform SOTA models. However, when all stages are included, retrieval performance increases by 6.8 pp for top-1 and 2.4 pp for top-1%. These findings highlight the effectiveness of the descriptor enhancement strategy proposed in this work.

V. CONCLUSIONS

We have introduced SPVSoAP3D, a novel 3D LiDAR-based place recognition model suitable for agricultural robotics, and extend the HORTO-3DLM dataset with two new sequences from horticultural environments: one recorded in an orchard in Metz, France, and another in a greenhouse in Coimbra, Portugal. This expansion enriches the dataset with more diversified data, enhancing its utility for agricultural robotics research.

Our proposed SPVSoAP3D model utilizes a second-order average pooling approach and an additional enhancement stage for local feature aggregation. Through experimentation, and compared to SOTA approaches, the results show the superiority of the average operator over the max operator on second-order statistics in horticultural environments. Moreover, SPVSoAP3D outperforms models relying on first-order approaches. Additionally, the results highlight the effectiveness of the enhancement stage, yielding performance

improvements compared to SPVSoAP3D using second-order average pooling alone.

ACKNOWLEDGMENTS

This work has been supported by the project GreenBotics (ref. PTDC/EEI-ROB/2459/2021), funded by Fundação para a Ciência e a Tecnologia (FCT), Portugal. It was also partially supported by FCT through grant UIDB/00048/2020 and under the PhD grant with reference 2021.06492.BD. Furthermore, we extend our sincere thanks to Cueillette de Peltre for providing access to their orchards.

REFERENCES

[1] R. Sparrow and M. Howard, "Robots in agriculture: prospects, impacts, ethics, and policy," *precision agriculture*, vol. 22, pp. 818–833, 2021.

[2] G. Rivera, R. Porras, R. Florencia, and J. P. Sánchez-Solís, "Lidar applications in precision agriculture for cultivating crops: a review of recent advances," *Computers and Electronics in Agriculture*, vol. 207, p. 107737, 2023.

[3] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[4] T. Barros, R. Pereira, L. Garrote, C. Premebida, and U. J. Nunes, "Place recognition survey: An update on deep learning approaches," *arXiv preprint arXiv:2106.10458*, 2021.

[5] 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.

[6] D. Cheng, F. C. Ojeda, A. Prabhu, X. Liu, A. Zhu, P. C. Green, R. Ehsani, P. Chaudhari, and V. Kumar, "Treescope: An agricultural robotics dataset for lidar-based mapping of trees in forests and orchards," *arXiv preprint arXiv:2310.02162*, 2023.

[7] J. Ma, J. Zhang, J. Xu, R. Ai, W. Gu, and X. Chen, "OverlapTransformer: An Efficient and Yaw-Angle-Invariant Transformer Network for LiDAR-Based Place Recognition," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6958–6965, 2022.

[8] M. Angelina Uy and G. Hee Lee, "PointNetVLAD Deep Point Cloud Based Retrieval for Large-Scale Place Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4470–4479.

[9] K. Vidanapathirana, M. Ramezani, P. Moghadam, S. Sridharan, and C. Fookes, "LoGG3D-Net: Locally Guided Global Descriptor Learning for 3D Place Recognition," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 2215–2221.

[10] T. Barros, L. Garrote, R. Pereira, C. Premebida, and U. J. Nunes, "AttDLNet: Attention-Based Deep Network for 3D LiDAR Place Recognition," in *ROBOT2022: Fifth Iberian Robotics Conference: Advances in Robotics, Volume 1*. Springer, 2022, pp. 309–320.

[11] S. Siva, Z. Nahman, and H. Zhang, "Voxel-Based Representation Learning for Place Recognition Based on 3D Point Clouds," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8351–8357.

[12] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[13] Y. Liao, J. Xie, and A. Geiger, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3292–3310, 2022.

[14] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of michigan north campus long-term vision and lidar dataset," *The International Journal of Robotics Research*, vol. 35, no. 9, pp. 1023–1035, 2016.

[15] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[16] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, "Complex urban dataset with multi-level sensors from highly diverse urban environments," *The International Journal of Robotics Research*, vol. 38, no. 6, pp. 642–657, 2019.

[17] J. Knights, K. Vidanapathirana, M. Ramezani, S. Sridharan, C. Fookes, and P. Moghadam, "Wild-Places: A Large-Scale Dataset for Lidar Place Recognition in Unstructured Natural Environments," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 322–11 328.

[18] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "RELLIS-3D Dataset: Data, Benchmarks and Analysis," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 1110–1116.

[19] C. Min, W. Jiang, D. Zhao, J. Xu, L. Xiao, Y. Nie, and B. Dai, "ORFD: A Dataset and Benchmark for Off-Road Freespace Detection," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2532–2538.

[20] A. Prabhu, X. Liu, I. Spasojevic, Y. Wu, Y. Shao, D. Ong, J. Lei, P. C. Green, P. Chaudhari, and V. Kumar, "UAVs for forestry: Metric-semantic mapping and diameter estimation with autonomous aerial robots," *Mechanical Systems and Signal Processing*, vol. 208, p. 111050, 2024.

[21] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global Second-order Pooling Convolutional Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3024–3033.

[22] K. Yu and M. Salzmann, "Statistically-motivated second-order pooling," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 600–616.

[23] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12*. Springer, 2012, pp. 430–443.

[24] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5297–5307.

[25] D. Cattaneo, M. Vaghi, and A. Valada, "LCDNet: Deep Loop Closure Detection and Point Cloud Registration for LiDAR SLAM," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2074–2093, 2022.

[26] J. Komorowski, "MinkLoc3D: Point Cloud Based Large-Scale Place Recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1790–1799.

[27] K. Vidanapathirana, P. Moghadam, B. Harwood, M. Zhao, S. Sridharan, and C. Fookes, "Locus: LiDAR-based Place Recognition using Spatiotemporal Higher-Order Pooling," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5075–5081.

[28] H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han, "Searching efficient 3d architectures with sparse point-voxel convolution," in *European Conference on Computer Vision*, 2020.

[29] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM journal on matrix analysis and applications*, vol. 29, no. 1, pp. 328–347, 2007.

[30] R. Bahtia, "Positive definite matrices, Princeton Series in Applied Mathematics," 2007.

[31] Z. Huang and L. Van Gool, "A Riemannian Network for SPD Matrix Learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[32] Z. Gao, Y. Wu, Y. Jia, and M. Harandi, "Learning to optimize on spd manifolds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7700–7709.

[33] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*. Springer, 2010, pp. 143–156.