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

TriLoc-NetVLAD: Enhancing Long-term Place Recognition in Orchards with a Novel LiDAR-Based Approach

Na Sun[1,3], Zhengqiang Fan[2], Quan Qiu[2]*, Tao Li[1]*, Qingchun Feng[1], Chao Ji[4], Chunjiang Zhao[1,3]

***Abstract*—Accurate long-term place recognition is crucial for agricultural robots operating in unstructured environments. However, in the challenging scene of orchard with high-frequency repetitive features, traditional LiDAR-based localization methods relying on geometric features prove to be inadequate. To address this challenge, we propose TriLoc-NetVLAD, a novel LiDAR-based long-term place recognition approach designed to handle the repetitive and ambiguous features of orchards. This approach initially fuses the point cloud density, height and spatial information to encode unordered 3D point clouds into a spatial context descriptor. Then，channel selection strategy based on descriptor's sublayer similarity between query and its corresponding positive and negative samples is proposed to amplify the differences in environmental features. Finally, we use a Triplet Network to extract local features, encompassing both high-dimensional and low-dimensional information. These local features are then cascaded through NetVLAD layer to form a global descriptor. Furthermore, we have built a cross-seasonal orchard dataset to evaluate the performance of our place recognition method. The experiment results demonstrate the advantageous localization performances of the proposed place recognition algorithm over the existing methods.**

I. Introduction

Localization is an indispensable part of many robotic systems. Within localization algorithms, place recognition enables robots to identify loop closure candidates for Simultaneous Localization and Mapping (SLAM)[1]. This means that robots can determine their current position in pre-recorded maps based on the features observed from sensor data, such as visual, geometric, or semantic information. Owing to its robustness against variations in lighting and weather conditions, LiDAR-based localization techniques have been widely employed in large-scale urban outdoor environments for Advanced Driver Assistance Systems (ADAS) and robot navigation domains[2-4]. However, when the application scenario shifts to a dwarf and high-density apple orchard, which contrasts sharply with urban scenes, it presents unique challenges for the LiDAR-based place recognition approach.

Orchards differ from urban scenes because they lack structured elements like vehicles and buildings, which poses a challenge for place recognition approach. Early LiDAR-based place recognition methods relied on hand-crafted descriptors extracted from points, lines, and planes, which proved to be ineffective in orchard settings[5,6]. This ineffectiveness stems from the fact that these features, while abundant in urban areas, are notably sparse in orchards.

Furthermore, the uniform arrangement of trees and consistent row spacing create repetitive features, leading to significant geometric similarities in scans. Additionally, the time-varying characteristics of orchard environments further increase the difficulty of place recognition based on LiDAR. Seasonal variations can affect canopy density, resulting in sparse scans and poor descriptive features. These challenges reduce the reliability and efficiency of long-term LiDAR-based place recognition methods.

Recently, notable advancements in deep learning have led a growing number of researchers to explore the use of deep neural networks for long-term place recognition based on LiDAR. Various methods, such as PointNetVLAD[7], PTNet-PLT[3], and LOGG3D-Net[8], have been developed to directly learn global features descriptors from point clouds to achieve robot localization. These methods have achieved impressive results in this domain. However, such methods require substantial computational resources. The challenge of real-time operation limits their practicality in real-world applications. Additionally, ambient temperatures in orchard may limit the performance of computing devices, resulting in issues with real-time processing. To address this, several studies have projected 3D point clouds into regular formats 2D images[9-13], subsequently applying existing or adapted neural networks to learn global descriptors. While this approach has shown considerable promise, the localization accuracy significantly drops in orchard environments when using images encoded by point cloud because the unique characters of orchards mentioned above.

In response, we propose TriLoc-NetVLAD, a novel LiDAR-based place recognition framework. Additionally, the scarcity of datasets encompassing cross-seasonal orchard scenarios hampers the development of data-driven place recognition algorithms for such contexts. To bridge this gap, we also build a comprehensive cross-seasonal orchard dataset to evaluate the performance of our proposed place recognition algorithm. The contributions of this paper can be summarized as follow:

*This work was funded by Beijing Nova Program (20220484023), Youth Research Foundation of Beijing Academy of Agriculture and Forestry Sciences (QNJJ202318), Science and Technology Cooperation Project of Xinjiang Production and Construction Crops (2022BC007).

[1]N. Sun, T. Li, Q. Feng and C. Zhao are with Intelligent Equipment Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China. Email: sunaswu@email.edu.cn (N. Sun)

[2]Z. Fan and Q. Qiu are with College of Intelligent Science and Engineering, Beijing University of Agriculture, Beijing, 102206, China.

[3]N. Sun and C. Zhao are with College of Engineering and Technology, Southwest University, Chongqing 400715, China.

[4]C. Ji is with Xinjiang Academy of Agricultural and Reclamation Science, Shihezi, 832000, China.

*T. Li and Q. Qiu are the corresponding authors. Email: lit@nercita.org.cn (T. Li), qiuquan0110@ustc.edu (Q. Qiu).

1) We propose a novel 3D LiDAR-based long-term place recognition approach specifically designed for tree environments such as orchard.

2) A novel point cloud descriptor called Multi-layer Spatial Context based on Point Density (MSCD) is proposed, which integrates geometric, density and spatial characteristics.

3) A novel channel selection strategy of point cloud descriptor is proposed to extract the most effective sublayer, thereby enhancing the uniqueness of environmental feature information.

II. Related Work

LiDAR-based place recognition approaches are divided into two main categories 3D point cloud-based and projected image-based.

Point-cloud-based approach enables the direct extraction of features, encompassing both geometric and semantic attributes, from 3D point cloud data. Recent advances in deep learning have significantly contributed to the development of point-cloud-based place recognition methods. PointNetVLAD [7] combines PointNet[14] with NetVLAD[15] for global descriptor extraction in place recognition task. LoGG3D-Net [8] enhances optimization through sparse convolution and integrates both local consistency loss and scene-level loss, whereas MinkLoc3D[16] addresses disordered datasets using sparse voxelization representation and convolutional techniques. PTC-Net[17] fuses sparse convolution and Point-wise Transformer methods to address potential information loss caused by sparse voxelization in point clouds. KPPR[18] leverages pre-trained encoder networks and stem architecture for improved network performance without negative sample calculations during training.

The projected image-based approach converts a 3D point cloud frame into a 2D image by means of polar or columnar coordinate projection. These images can be classified Range Image Views (RIV) or Bird's Eye Views (BEV) according to the projection direction method. OverlapNet[19] and its improved version[20], for instance, utilizes depth images, to estimate the overlap and relative yaw angles between query and reference scans. BEVPlace[21] uses grouped convolution in its rotation-invariant network for place recognition. MMCS-Net[22], mimicking human eyes, processes projected images with a cascading Siamese convolutional neural network to generate versatile and distinctive feature descriptors.

Some BEV-based methods trade fine-depth details for consistent height information, benefiting to reduce the robot response time. Scan Context (SC)[23] introduces a two-dimensional descriptor reliant on maximum bin height, with variants like Intensity Scan Context (ISC)[24], Semantic Scan Context (SSC)[25] and Scan Context++[26]. However, the aforementioned BEV-based methods transform point clouds into a single-layer 2D matrix representation through dimension reduction and compression, resulting in significant information loss, particularly in the vertical dimension. DiSCO[12] combines multi-layer SC descriptors with spectral analysis for global, rotation-invariant descriptors. Spatial Binary Pattern (SBP)[5] divides 3D point clouds into eight layers, using point density instead of height for a binary descriptor of global spatial information. Contour Context[27] uses BEV slice image connection points from various layer heights to form contours, aiding in similarity calculations using metrics like pixel count and center position.

III. Methodology

This section presents a comprehensive exposition of our algorithm, a deep learning framework for LiDAR-based long-term place recognition in orchards. Fig.1 illustrates the architecture of our proposed framework, named TriLoc-NetVLAD. It comprises of four primary modules: multi-layer descriptor generation, effective sublayer extraction, feature extraction and database retrieval. Following the methodology employed in prior literatures[12,13,23], the point cloud is initially encoded to generate a multi-layer spatial context descriptor that fuses the point cloud's geometry, density and spatial information. At the same time, effective sublayers that can amplify the environmental features are screened using an image similarity index. Subsequently, local features are computed through convolutional neural network (CNN) operations, and NetVLAD is utilized to aggregate the global descriptor. Ultimately, the robot's relocation is determined by querying adjacent keyframes of the current point cloud in the database by KD-Tree and verifying feature descriptor similarity through Euclidean distance comparison.

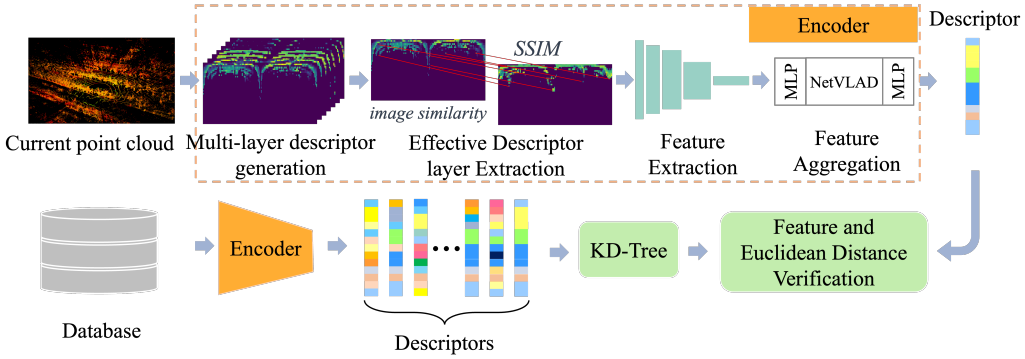


Figure 1. I Illustration of the overall architecture of the TriLoc-NetVLAD model. Database is the orchard map generated by KISS-ICP[28].

A. Multi-Layer Spatial Context Descriptors

LiDAR has been extensively utilized in the acquisition of canopy information for fruit trees, including measurements such as canopy volume and size, which demonstrates its feasibility in orchards[29,30]. Considering the robustness of LiDAR to environmental conditions like changes in illumination, we aim to leverage its point cloud data for long-term place recognition in orchards. However, the orchard which robot works differs from previous literatures[5,6]. The fruit trees in the orchard have smaller crowns and exhibit similar morphology. Consequently, traditional methods that rely solely on canopy information cannot be used for place recognition by LiDAR. And the orchard undergoes over time, such as the gentle movement of leaves caused by the breeze and the growth of leaves and fruits. This change may introduce potential errors between data frames. Therefore, we propose a point cloud descriptor based on spatial density context information to capture the distinctive characteristics of the orchard environment effectively.

As described in[5,12], we also partition 3D point cloud into 3D bins based on azimuthal, radial, and layer directions, as illustrated in Fig.2. To optimize information usage, we integrate height, density and spatial context information from the point cloud. In simpler terms, a bin is considered occupied when it contains more than 10 points; Otherwise, it is labeled as unoccupied. The 3D bins assignment function is defined as:

$$\delta(P)=\begin{cases}1, if\ P\ is\ occupied\\ 0, otherwise\end{cases} \tag{1}$$

where $\delta(P)$ is the bin status, assigning '0' for non-occupation, and '1' for occupation. Thus, the point cloud can be encoded into a descriptor in a $N_l \times N_r \times N_s$ subspace, called Multi-layer Spatial Context based on point cloud Density (MSCD). It is defined as:

$$SCD(k,i,j)=z_{mean}\delta(P) \tag{2}$$

where MSCD represents a multi-level global feature capturing the geometric shape and density distribution of the point cloud, providing a comprehensive spatial information reflection form a bird's-eye view.

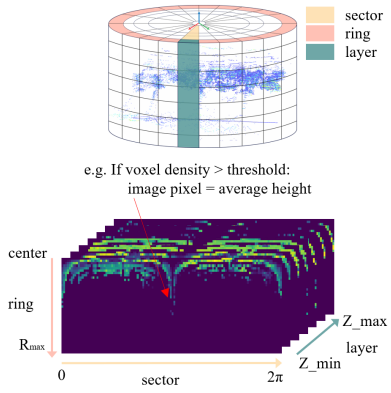


Figure 2. Schematic of our proposed MSCD descriptor, drawing inspiration from methodologies described in reference [5]. Our method innovatively adopts the concept of "layers" to more precisely convey spatial information. Furthermore, we enhance the selectivity of data in each bin, implementing a stringent filtering based on point density to refine accuracy and relevance.

B. Effective Descriptor Layer Extraction

In orchard environments with high-frequency repetitive features, we notice that some point cloud descriptors [12,23,24] based on BEV representation exhibit high feature similarities between positive samples (same position) and negative samples (different position) within the same query. In certain cases, the query descriptor shows greater similarity to negative samples than to positive ones. Similarly, our proposed point cloud descriptor, MSCD, also shows this issue in certain layers, significantly impacting robot localization robustness in the orchard. To track this challenge, we propose a channel selection strategy based on relies on comparing the descriptors similarities. This approach aims to reduce the negative impact of high-similarity feature layers on robot place recognition in environments with high-frequency repetitive features. Since the point cloud descriptors mainly capture the spatial occupancy details and height information of the orchards. We evaluate the similarity of descriptors using the Structural Similarity Index Measurement (SSIM)[31,32].

SSIM provides a nuanced approach to image comparison, emphasizing perceptual differences over simple pixel discrepancies. By analyzing changes in structure, luminance, and contrast, it aligns its assessment of image similarity with the complexities of human visual perception. The SSIM index, which scales from 0 to 1, measures the similarity between images. Higher values indicate greater similarity, while lower values indicate less likeness. The formula for calculating SSIM is detailed below:

$$SSIM(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\delta_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\delta_x^2+\delta_y^2+C_2)} \tag{3}$$

where, μ_x and μ_y are the average pixel values of image x and y, respectively. δ_x and δ_y are the standard deviations of x and y, respectively. δ_{xy} is the covariance between x and y. C_1 and C_2 are constants introduced to stabilize the division with weak denominators. In this paper, we calculate the SSIM index between the query and its corresponding positive and negative sample for every layer or channel, respectively. After that, the effective descriptor layer is then evaluated by calculating the SSIM difference dif_{SSIM} between positive and negative samples. when dif_{SSIM} is positive, it indicates that the current query is similar to its positive samples. Conversely, it indicates that the current query is more similar to its negative samples.

We perform the same process for all samples, calculating the SSIM variance value between the query and its corresponding positive and negative samples. These results are then statistically analyzed to calculate their distribution within each interval. Based on the above results, we choose the effective sublayer of MSCD with a small percentage of negative samples and a large percentage of dif_{SSIM} values in [0.1 inf).

C. Feature Extraction

The traditional Triplet Network often struggles to capture the subtle nuances in data that needs to be retrieved, especially for fine-grained image analysis. They fail to pay attention to intricate features that are essential for nuanced comprehension

and identification. High-frequency repetitive characteristics of the orchard resulted in the effective sublayers still having similarities. The local subtle differences between descriptors are important for determining the robot place recognition algorithm robustness. Thus, images encoded by point cloud can be seen as a fine-grained image retrieval problem in this paper. Inspired by[12,20,22], we introduce TriLoc-NetVLAD, a cascading network architecture that enhances long-term robot place recognition by combining the strengths of Triplet Netwrok and NetVLAD, as illustrated in Fig.3.

TriLoc-NetVLAD consists of two main parts: local feature extraction and feature aggregation. A weight-shared Triplet Network, called TriLocNet, is employed to extract local feature. TriLocNet has three branches for the anchor, positive, and negative samples. Each branch is equipped with nine convolutional layers that use asymmetric convolutional kernels. While this design could capture rich semantic information, it sacrifices resolution and detail perception, which contrasts with low-level features known for their high resolution and detail sensitivity. To address this issue, features from the third and seventh convolutional layers are integrated using a cascaded fusion technique, blending shallow and deep features to improve detail recognition.

In addition, our analysis in Section III.B reveals that the interaction of MSCD across different channels or height layers has a significant impact on robot localization accuracy. Because convolutional operations are linear and may neglect the spatial context of descriptors, we propose the use of a channel attention mechanism. This mechanism highlights important features for place recognition while diminishing less relevant ones.

The local feature descriptors generated by TriLocNet are fed into the NetVLAD layer, where they are aggregated into a global feature descriptor. It serves as a unique fingerprint in the database retrieval process for place recognition pipeline. By integrating this method, the TriLoc-NetVLAD architecture is highly effective in capturing detailed features. The strong performance in handling unstructured, dynamic, and repetitive environments showcases the potential of our approach in robotic place recognition.

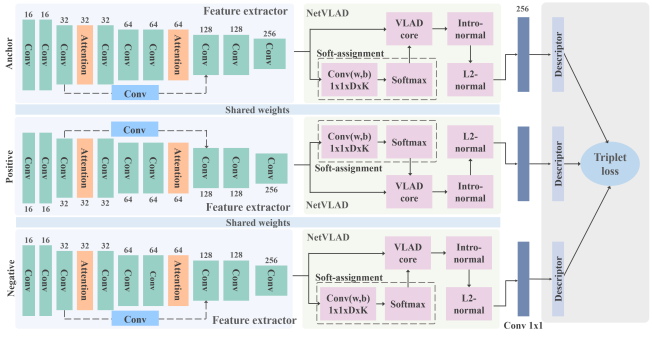


Figure 3. Architecture of the proposed TriLocNet.

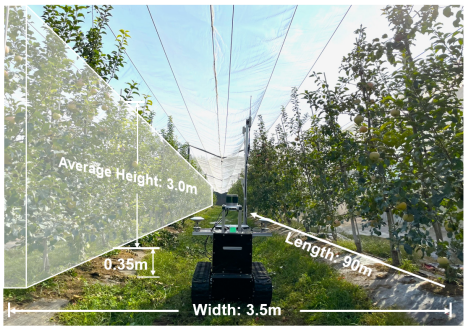
IV. Experiments

The experimental evaluation aims to show the performance of our approach and evaluate its capability in three key areas. i) Detecting loop closure candidates based solely on LiDAR data in large-scale scenes, without using any additional information. ii) Comparing to other point cloud descriptors that directly use point cloud intensity, height maximum or spatial context information, MSCD, fusing point cloud density, height and spatial context, can effectively improve the robustness of robot place recognition in unstructured environments. iii) Introducing a novel descriptor effective sublayer selection strategy to highlight task-relevant features and improve the accuracy of loop closure detection.

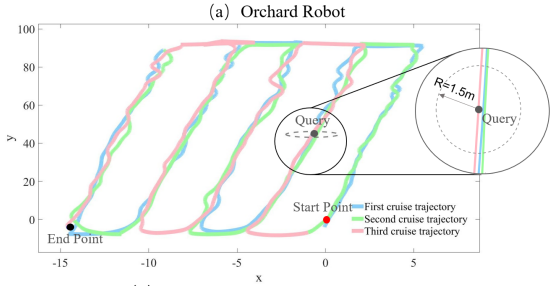
The performance of our approach will be evaluated using self-made orchard datasets. The experiments are conducted on a server equipped with an NVIDIA RTX 2080Ti 11GB GPU. Our model is implemented in Pytorch 1.10.2.

A. Experiment Settings

To evaluate the performance of our approach in orchard, we create a custom-made orchard dataset. This dataset comprises eight distinct sequences that cover the entire apple growth cycle. The sequences encompass a wide range of tree growth stages and diverse weather conditions experienced throughout the fruiting cycle. The dataset presents a significant challenge for robot place recognition efforts due to the dynamic changes in tree structure, weather patterns, and additional environmental factors. Fig.4 illustrates our robot and its corresponding trajectory. Tab.I provides details on the number of frames and loops in the dataset.



(a) Orchard Robot



(b) Travelling path of the robot during data acquisition tasks

Figure 4. Robot operating in orchard scenario and its travel paths. a) shows the robot platform and its actual working scenario. b) describes the robot's travel trajectory during each operation. To ensure the dataset containing both forward and reverse information, the robot traverses back and forth three times along a set routine. The pink, blue and green lines represent the robot's three different travel paths, respectively.

Our metric learning methodology follows the framework outlined in [7], incorporating a selective downsampling process for positive and negative samples. Within our custom-made orchard dataset, scan separated from the current query by 1.5m are categorized as positive sample pairs, while scan frames separated by 3m or more are deemed negative sample pairs, as shown in Fig.4(b). Since the number of negative samples is larger, we randomly select negative samples equal to the number of positive samples. Then, the

current query, positive samples and negative samples corresponding to the current scan are fed together into the network for training.

During the training stage, seq.00 is used, while seq.01, seq.02, seq.03, and seq.04 are methodically utilized for cross-validation. Additionally, in an effort to assess performance consistency within a single sequence and season, we carefully extracted data a single cruise in seq.03. As depicted in Fig.4, the three distinct trajectories within sequence 03 are sequentially represented as the first cruise trajectory (seq.03-1), second cruise trajectory (seq.03-2), and third cruise trajectory (seq.03-3). These trajectories distinctly marked with a pink line, green line, and blue line, respectively. In this paper, a position is considered successfully recognized if the retrieved point clouds are within a 1.5m proximity of the intended target location.

TABLE I. NUMBER OF FRAMES AND LOOPS IN THE DATASET.

Season	Seq	Time	Length	Loops
spring	00	2023-03	20788	13075
	01	2023-04	17474	8737
	02	2023-05	28814	17288
summer	03	2023-06	22795	13904
	04	2023-07	21147	12265
	05	2023-08	19464	12126
autumn	06	2023-09	19202	9631
	07	2023-10	17710	10626
	08	2023-10	22837	14387

B. *Effective Layer Extraction*

In view of the size of the custom-made dataset in question, we use random sampling method to statistically determine the SSIM index of the samples for descriptor effective layer selection. We choose 3000 groups of samples from all sequences in this paper. These samples include queries and their associated positive and negative samples. Following the image similarity comparison approach described in Section III.B, we meticulously calculate the difference between selected sequences and their corresponding positive and negative samples. We then display the results for 100 sets of data using heat maps, as shown in Fig.5.

It is evident form Fig.5 that there are inconsistent differences between positive sample and negative samples. In some cases, the similarity between the query and its negative samples surpasses that with positive samples. This phenomenon also provides insights into the limitations of existing open-source place recognition algorithms based on LiDAR in such environments. Therefore, it is possible to carry out an in-depth statistical analysis of the query in each layer by adopting our proposed channel selection strategy. This strategy allows us to identify representative sub-layers with significant feature differences to amplify the environmental features, so as to improve the place recognition accuracy of robot in orchard.

As shown in Fig.5b, we use mathematical statistics to analyze the SSIM variances of 3000 selected samples at different intervals. We observe that the point cloud descriptor, MSCD, does not consistently show positive and negative values for each layer in the custom-made orchard dataset. Specifically, the second to fourth layers contains relatively fewer negative similarity values and a larger proportion of samples fell within the [0.3, inf) interval compared to other layers. This suggests that these layers are more sensitive to variations in environments characterized by high-frequency repetitions. Thus, we can choose 2nd~4th as the effective layers. Furthermore, 7th and 8th layers of point cloud descriptor display a complete absence of negative percentage. This phenomenon may be attributed to the height of these layers surpassing that of the fruit tree, resulting in manifestation of zero-pixel values within this descriptor. Consequently, these layers are deemed ineffective for utilization as key layer.

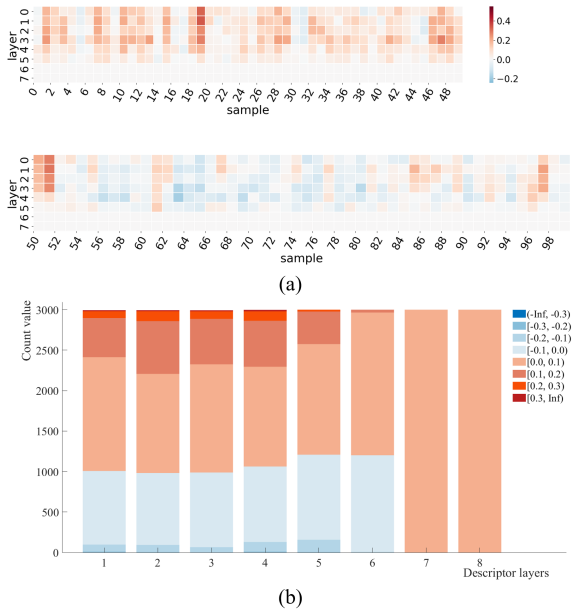


Figure 5. Visualization of effective layer selection strategy. a) is the heat map of SSIM variance of query with its positive and negative samples, respectively. We randomly select 100 groups for visualization. b) is the distribution of SSIM variances across the different intervals for 3000 group samples. In this figure, positive values are indicated by hot colors. Negative values are represented by cold colors.

C. *Performance on Orchard*

The experiment evaluation is grounded on several performance metrics, including top 1 recall (Recall@1), top 5 recall (Recall@5), and top 20 recall (Recall@20). The results, as outlined in Tab.II, indicate that TriLoc-NetVLAD can be applied to the orchard. As mentioned earlier, the prevalent issue in this environment is the high frequency of indistinguishable, repetitive features. To solve these challenges, this paper focuses on enhancing distinction capabilities through the strategic selection of effective descriptor layers. These layers empower descriptor to discern subtle variances. The network thereby can extract more meaningful feature sequences through these effective layers. In terms of Recall@1 metric, our approach achieves significant improvements of 11.9% over multi-level descriptors in the same season and 5.1% across different seasons. The results demonstrate our approach can enhance robot's place recognition performance by the effective descriptor layer extraction. However, our approach exhibits some instability performance when cross-seasonal experiment. This may be due to seasonal changes in the orchard as well as human intervention factors, such as branches pruning, flower

thinning and fruit thinning. Additionally, we also apply the effective descriptor layer extraction strategy to SC. The place recognition performance of robot improves 6.5% and 6.9% in the same season and across seasons, respectively.

Further, we also compared different voxel-property point cloud descriptors, such as SC, and SBP. Among them, SBP is encoded by binary to describe the point cloud spatial context information. To maintain a fair comparison, the layer is set to 8. The comparison results of different point cloud descriptors are shown in Tab.II. In terms of single-layer descriptor, SBP is superior, followed by MSCD, and SC performs the worst. Following integration with efficient layers extraction approach, the MSCD descriptor proposed is improved by 2.7% relative to SC in Recall@1 metric.

Fig.6 shows the performance of place recognition in the orchard using our proposed approach. We select four queries from the orchard dataset and rank the results from the top 1 to the top 5. Correct matches are highlighted with green bounding boxes, while incorrect ones are marked with red. Below each image, the Euclidean distance quantifies the spatial difference between the predicted and actual locations. The angle indicates the yaw angle difference between the predicted and actual positions. The first two rows of the figure depict the robot's relocation results in the headland of the orchard, while the last two demonstrate the localization outcomes between rows. The results suggest that the robot's localization is less accurate in the inter-row scenarios. In these cases, the predicted descriptors closely resemble the true descriptors, yet the distances between the predicted and actual positions show significant variation, ranging from close to far.

TABLE II. COMPARISON OF DIFFERENT DESCRIPTORS' RECOGNITION PERFORMANCES ON OUR ORCHARD DATASET

Method	Same Season			Cross Season		
	Recall@1	Recall@5	Recall@20	Recall@1	Recall@5	Recall@20
TriLoc-NetVLAD-SBP	0.643	0.718	0.880	0.468	0.557	0.627
TriLoc-NetVLAD-SC(single)	0.605	0.698	0.784	0.424	0.497	0.568
TriLoc-NetVLAD-SC(multiple)	0.655	0.747	0.869	0.459	0.576	0.631
TriLoc-NetVLAD-SC(effective)	0.720	0.769	0.873	0.528	0.595	0.683
TriLoc-NetVLAD-MSCD(single)	0.604	0.709	0.786	0.439	0.536	0.653
TriLoc-NetVLAD-MSCD(multiple)	0.628	0.754	0.848	0.499	0.578	0.680
TriLoc-NetVLAD-MSCD(effective)	0.747	0.821	0.875	0.550	0.653	0.697

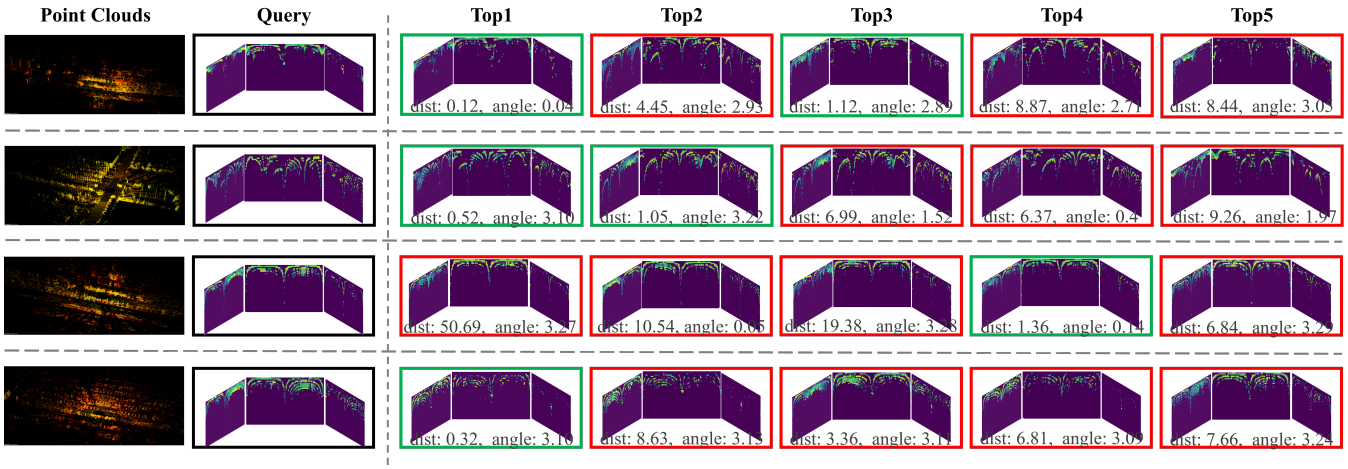


Figure 6. Top 5 database results for the provided query point clouds. A green border denotes a correct match while a red border is from a different location. Additionally, the Euclidean distance and yaw angle difference between the descriptors are provided.

V. CONCLUSION

Point cloud-based place recognition is an important part and challenging task in orchard with high-frequency repetitive features. To address this challenge, we propose a novel LiDAR-based place recognition framework, which is named TriLoc-NetVLAD, in this paper. It mainly consists of three parts: i) a new point cloud descriptor fuse point cloud density, height, and spatial information. ii) channel selection strategy chooses the effective descriptor layer by comparing the query and its corresponding samples similarity. iii) A Triplet Network is used to extract local features and NetVLAD is used to aggregate the local feature to global descriptor. To verify the performance of our algorithm, we conduct two experiments on our self-made orchard dataset. The results demonstrate that our algorithm meets the requirements for robot localization. In future work, we plan to explore the performance of inter-layer feature correlations of the descriptor in the context of robot place recognition.

REFERENCES

[1] H. Yin, X. Xu, S. Lu, X. Chen, R. Xiong, S. Shen, C. Stachniss, and Y. Wang, "A survey on global lidar localization: Challenges, advances and open problems," *arXiv preprint arXiv:2302.07433*, 2023.

[2] B. Peng, H. Xie, and W. Chen, "Roll: Long-term robust lidar-based localization with temporary mapping in changing environments," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp.2841-2847.

[3] C. Wen, H. Huang, Y. Liu, and Y. Fang, "Pyramid Learnable Tokens for 3D LiDAR Place Recognition," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp.4143-4149.

[4] Z. Wu, W. Wang, J. Zhang, Q. Lyu, H. Zhang, and D. Wang, "Global localization in repetive and ambiguous environments," in *2023 International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp.12374-12380.

[5] F. Ou, Y. Li, and Z. Miao, "Place recognition of large-scale un structured orchards with attention score maps," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 958–965, 2023.

[6] T. Barros, L. Garrote, P. Conde, M. Coombes, C. Liu, C. Premebida, and U. Nunes, "Orchnet: A robust global feature aggregation approach for 3d lidar-based place recognition in orchards," *arXiv preprint arXiv:2303.00477*, 2023.

[7] M. A. Uy and G. H. Lee, "Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4470–4479.

[8] 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.

[9] B. Li, T. Zhang, and T. Xia, "Vehicle detection from 3d lidar using fully convolutional network," *arXiv preprint arXiv:1608.07916*, 2016.

[10] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

[11] J. Beltr´ an, C. Guindel, F. M. Moreno, D. Cruzado, F. Garcia, and A. De La Escalera, "Birdnet: a 3d object detection framework from lidar information," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3517–3523.

[12] X. Xu, H. Yin, Z. Chen, Y. Li, Y. Wang, and R. Xiong, "Disco: Differentiable scan context with orientation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2791–2798, 2021.

[13] J. Ma, X. Chen, J. Xu, and G. Xiong, "Seqot: A spatial–temporal transformer network for place recognition using sequential lidar data," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 8, pp. 8225 8234, 2023.

[14] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[15] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Pro ceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[16] J. Komorowski, "Minkloc3d: Point cloud based large-scale place recognition," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1789–1798.

[17] L. Chen, H. Wang, H. Kong, W. Yang, and M. Ren, "Ptc-net: Point-wise transformer with sparse convolution network for place recognition," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3414–3421, 2023.

[18] L. Wiesmann, L. Nunes, J. Behley, and C. Stachniss, "Kppr: Exploiting momentum contrast for point cloud-based place recognition," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 592–599, 2023.

[19] X. Chen, T. L¨ abe, A. Milioto, T. R¨ ohling, J. Behley, and C. Stachniss, "OverlapNet: A siamese network for computing lidar scan similarity with applications to loop closing and localization," *Autonomous Robots*, vol. 46, pp. 61–81, 2021.

[20] J. Ma, J. Zhang, J. Xu, R. Ai, W. Gu, and X. Chen, "Overlaptrans former: 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.

[21] L. Lun, Z. Shuhang, L. Yixuan, F. Yongzhi, Y. Beinan, C. Siyuan, and S. Hui-Liang, "BEVPlace: Learning LiDAR-based place recognition using bird's eye view images," *arXiv preprint arXiv:2302.14325*, 2023.

[22] D. Kong, X. Li, Y. Cen, Q. Xu, and A. Wang, "Simultaneous viewpoint- and condition-invariant loop closure detection based on lidar descriptor for outdoor large-scale environments," *IEEE Transac tions on Industrial Electronics*, vol. 70, no. 2, pp. 2117–2127, 2023.

[23] G. Kim and A. Kim, "Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4802–4809.

[24] H. Wang, C. Wang, and L. Xie, "Intensity scan context: Coding intensity and geometry relations for loop closure detection," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 2095–2101.

[25] L. Li, X. Kong, X. Zhao, T. Huang, W. Li, F. Wen, H. Zhang, and Y. Liu, "Ssc: Semantic scan context for large-scale place recognition," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 2092–2099.

[26] G. Kim, S. Choi, and A. Kim, "Scan context++: Structural place recog -nition robust to rotation and lateral variations in urban environments," *IEEE Transactions on Robotics*, vol. 38, no. 3, pp. 1856–1874, 2022.

[27] B. Jiang and S. Shen, "Contour context: Abstract structural distribution for 3d lidar loop detection and metric pose estimation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 8386–8392.

[28] I. Vizzo, T. Guadagnino, B. Mersch, L. Wiesmann, J. Behley and C. Stachniss, "KISS-ICP: In Defense of Point-to-Point ICP – Simple, Accurate, and Robust Registration If Done the Right Way," IEEE Robotics and Automation Letters, vol. 8, no. 2, pp. 1029-1036, 2023.

[29] L. Du, Y. Pang, "Identifying regenerated saplings by stratifying forest overstory using airbone lidar data," Plant Phenomics, vol. 6, no. 145, pp.1-14, 2024.

[30] Y. Hao, F.R.A. Widagdo, X.Liu, Y. Liu, L. Dong, and F. Li, "A hierarchical region-merging algprithm for 3d segmentation of individual trees using uav-lidar point clouds," *IEEE Transactiona on Geoscience and Remote Sensing*, vol. 60, no. 5701416, pp.1-16, 2022.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image qualoty assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp.600-612, 2004.

[32] G. Lu, "Bird-view 3d reconstruction for crops with repeated textures," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp.4263-4270.