2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SCDCE-3D: Soft-weighted Covariance and Dual-branch Channel Enhancement for 3D Place Recognition in Complex Orchard Environments

Yuping Tan[1],Chunjiang Zhao[1,2],Qin Zhao[3],Xinhong Hei[1],Xiaogang Song[1*]

Abstract— **Recent progress in 3D place recognition has delivered strong results in urban and indoor scenarios, but orchards remain largely unexplored. In these environments, unreliable or absent GNSS signals necessitate LiDAR-based place recognition for robust long-term localization, yet challenges such as ill-defined geometry, semi-transparent foliage, and severe inter-/intra-row overlaps cause high structural ambiguity. To address these challenges, we propose SCDCE-3D, a novel framework that integrates soft-weighted covariance representation with dual-branch channel enhancement. The soft-weighted covariance module adaptively down-weights noisy or overlapping points using a sigmoid-based weighting strategy, enabling robust second-order statistical representation that suppresses cross-row interference. In parallel, a dual-branch backbone extracts complementary global and local features, which drive a dynamic channel enhancement mechanism to emphasize discriminative feature channels while suppressing redundancy. Furthermore, multi-level triplet learning is applied not only to the final descriptor but also to intermediate statistical features, reinforcing robustness against structural ambiguity. Experiments on orchard-based LiDAR datasets demonstrate that SCDCE-3D significantly outperforms state-of-the-art methods in both recall and robustness, offering a reliable solution for long-term 3D place recognition in agricultural robotics. Code is available at https://github.com/typist2001/SCDCE-3D.**

I. INTRODUCTION

Autonomous robots are increasingly recognized as a key enabler for improving productivity and operational efficiency in the agricultural sector [1]. These robots are expected to perform complex tasks such as pruning [2], harvesting [3], and precision spraying [4] over large-scale farms, which requires accurate and reliable localization over extended periods. However, in many agricultural environments, GNSS signals are unreliable or entirely unavailable [5] due to dense canopy cover, terrain occlusion, or seasonal variations, making satellite-based positioning insufficient for long-term autonomous operation. To overcome these limitations, alternative localization strategies such as Simultaneous Localization and Mapping (SLAM) [6] and place recognition [7] have been widely adopted, providing robust global positioning and loop closure capabilities.

Among these approaches, 3D LiDAR-based place recognition [8] has emerged as a particularly effective solution, both as a core component within SLAM systems and as a standalone method, owing to its resilience to illumination changes, weather conditions, and seasonal variations. By capturing precise geometric measurements, LiDAR enables the extraction of high-fidelity spatial features that are less sensitive to appearance variations than visual sensors, which is crucial for outdoor agricultural deployment. While recent breakthroughs have largely stemmed from autonomous driving research tailored to structured urban settings, directly applying these methods to unstructured agricultural environments remains highly challenging. The ability to robustly operate under sparse, repetitive, and partially occluded point cloud conditions is therefore essential for enabling reliable, long-term autonomous robot operation in orchards, vineyards, and other complex agricultural scenarios.

Most breakthroughs in 3D LiDAR-based place recognition [8]–[10] have been optimized for structured urban environments, such as KITTI [11], KITTI-360 [12], NCLT [13], Oxford datasets [14] and Mulran [15], where scans capture abundant and well-defined geometric structures, including corners, planes, and edges. These stable features enable robust descriptor extraction and reliable retrieval across large distances. However, such features are extremely scarce in orchard environments, where trees, crops, and semi-transparent foliage dominate the scene. The resulting LiDAR scans are often sparse, partially occluded, and contain high structural similarity between adjacent rows, creating severe intra-row and inter-row descriptor ambiguity that significantly undermines retrieval accuracy. Moreover, repetitive row layouts exacerbate the risk of false positives during loop closure, posing a critical challenge for autonomous navigation and long-term localization.

Although several works have recently adapted place recognition to agricultural settings, orchard-specific challenges remain largely unresolved. SPVSoAP3D [16] integrates a voxel-based backbone with second-order pooling and descriptor enhancement to improve noise robustness. Yet voxelization introduces quantization errors and additional preprocessing latency, while global pooling struggles to suppress fine-grained cross-row overlaps in repetitive orchard layouts. TriLoc-NetVLAD [17] fuses geometric, density, and spatial cues into multi-layer descriptors through sub-layer selection

*This work was funded by the National Natural Science Foundation of China(52372418),the National Natural Science Foundation of China Joint Fund Key Project(U2368203),the National Key R&D Program of China(2022YFB2602203).

[1]Y.Tan, C.Zhao, X.Hei and X.Song are with School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, 710048, China.

[2]C.Zhao is with Intelligent Equipment Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing, 102206, China.

[3]Q.Zhao is with School of Civil Engineering and Architecture, Xi'an University of Technology, Xi'an, 710048, China.

*X.Song is the corresponding authors. Email:songxg@xaut.edu.cn.

and a CNN-NetVLAD pipeline, but its multi-stage design hinders real-time scalability under field conditions. PointNetPGAP [18] combines mean and covariance aggregators with a segment-level consistency module, offering high efficiency but relying on a static fusion strategy that cannot adapt to varying sparsity and semi-transparent foliage.

To address these challenges, we propose SCDCE-3D, a novel framework tailored for 3D place recognition in complex orchard environments. Our method integrates statistical feature aggregation with adaptive weighting and structural enhancement to overcome the issues of noisy overlaps and structural ambiguity. Specifically, it combines a soft-weighted covariance representation with a dual-branch channel enhancement mechanism and a multi-level learning strategy to ensure robustness in challenging agricultural conditions. Our contributions can be summarized as follows:

- We introduce a soft-weighted covariance module that adaptively assigns weights to points using local statistics, effectively suppressing noise and mitigating descriptor ambiguity.
- We design a dual-branch channel enhancement architecture that extracts global and auxiliary features and dynamically refines channel responses to highlight discriminative structures.
- SCDCE-3D adopts a multi-level triplet learning strategy applied to both final descriptors and intermediate features. Extensive experiments demonstrate that it consistently outperforms state-of-the-art baselines.

II. RELATED WORK

3D LiDAR-based place recognition has become a prominent research direction in robotics due to its robustness against illumination, weather, and seasonal variations [19]. Compared to vision-based approaches, LiDAR provides geometry-centric measurements that are less sensitive to appearance changes, making it particularly suitable for long-term outdoor deployment.

Traditional methods relied heavily on handcrafted feature engineering, leveraging prior knowledge of 3D geometric properties. Common early features include normal vectors, curvature, and point density [20]: normals capture local surface orientation by analyzing covariance eigenvectors, curvature quantifies surface bending to distinguish planes, edges, and corners, while point density estimates spatial sampling uniformity. Composite descriptors such as FPFH (Fast Point Feature Histograms) [21] and its simplified variant SPFH further incorporated angular and distance distributions within neighborhoods to encode local geometry in histogram form. While these features proved effective in early LiDAR-based localization, they degrade significantly in orchard environments, where geometric structures are sparse and repetitive.

Recent years have witnessed a paradigm shift toward deep learning (DL), where most frameworks consist of two key components: a local feature extraction module that embeds raw point clouds into high-dimensional feature space, and a feature aggregation module that compresses these local descriptors into compact global representations. For feature extraction, point-based approaches such as PointNetVLAD [8] combine PointNet [22] and NetVLAD [23] to enable end-to-end training for global descriptor learning, introducing triplet and quadruplet loss functions for enhanced discriminability. PointNetPGAP [18] integrates first-order and second-order pooling into a unified representation, reinforced by a piecewise consistency module for improved robustness under challenging conditions. Voxel-based methods, including MinkLoc3D [24], employ sparse voxelization and 3D sparse convolutions [25] to capture fine-grained local geometry, overcoming limitations of purely point-based models. LCDNet [26] extends PV-RCNN [27] for LiDAR loop closure detection by leveraging voxel-based backbones and pose estimation heads built on differentiable optimal transport. Similarly, LoGG3D-Net [10] incorporates sparse convolutions with local consistency loss to ensure repeatable feature learning across revisits, while TriLoc-NetVLAD [17] fuses density, height, and spatial cues into context-aware descriptors with channel selection strategies for improved distinctiveness. Other notable approaches include Overlap-Transformer [9], which leverages yaw-invariant range-image projections and hierarchical attention mechanisms, and SPV-SoAP3D [16], which combines voxel-based backbones with second-order pooling and descriptor enhancement for noise robustness. In this work, we adopt a point-based paradigm for its computational efficiency and scalability, but address its inherent limitations by introducing a Soft-weighted Covariance (SWC) mechanism that adaptively suppresses noisy or overlapping points, together with a dual-branch channel enhancement module that amplifies discriminative dimensions.

III. METHOD

This section provides a detailed description of the proposed SCDCE-3D model within our retrieval-based place recognition framework, as illustrated in Fig.1. In this approach, place recognition is formulated as a retrieval problem and evaluated at the segment level, where the orchard is partitioned into discrete segments, as depicted in Fig.2.

A. Problem Formulation

A typical retrieval-based framework for 3D LiDAR place recognition involves two main steps. First, a LiDAR scan consisting of n points, $P \in \mathbb{R}^{n\times 3}$, is transformed into a compact, high-dimensional descriptor $D \in \mathbb{R}^d$ via a mapping function $\Theta : \mathbb{R}^{n\times 3} \to \mathbb{R}^d$. In the subsequent stage, this descriptor is matched against a pre-constructed reference database using a similarity measure, and the top-k most similar entries are selected as candidate revisited locations. The quality of the mapping function Θ is therefore critical, as it directly determines the discriminative capability of the resulting descriptor and, consequently, the retrieval performance.

In this study, we aim to construct a robust mapping function Θ, which can be conceptually divided into two stages: $\Theta = \varphi \circ \psi$. The first stage, $\psi : \mathbb{R}^{n\times 3} \to \mathbb{R}^{n\times c}$, encodes the raw LiDAR scan P into point-level features $F \in \mathbb{R}^{n\times c}$ within

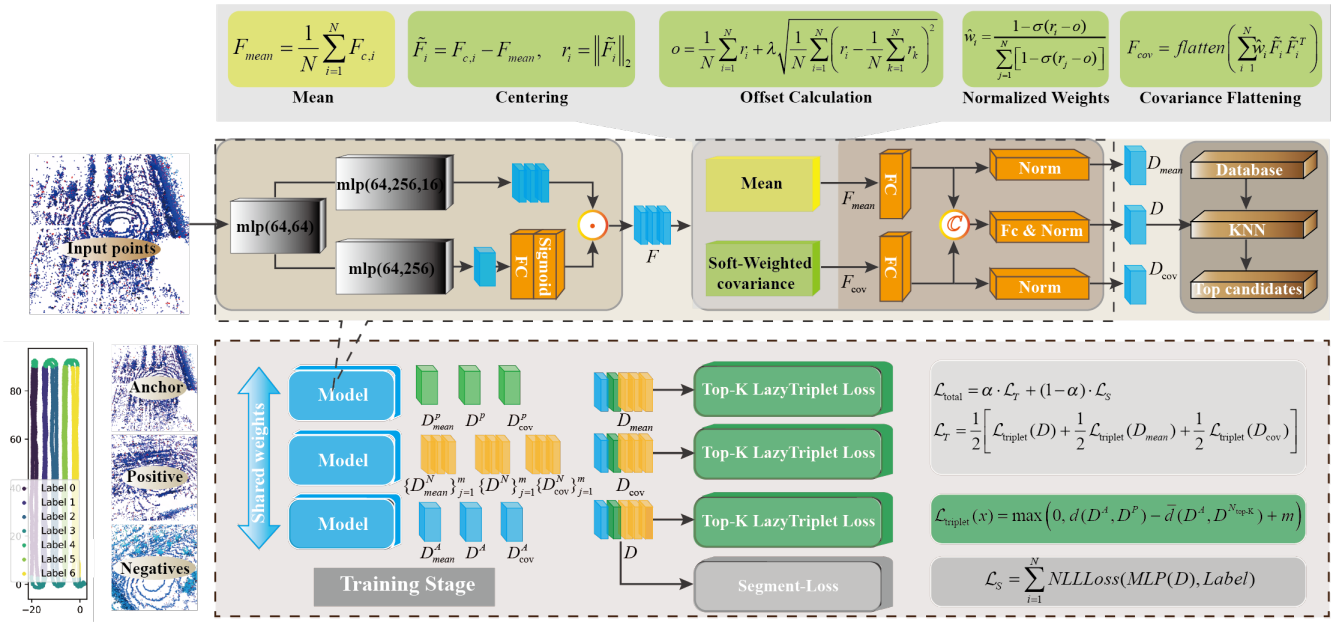

Fig. 1. The overall pipeline of the proposed SCDCE-3D framework. Raw point clouds are encoded into local features, which are refined through dual-branch channel enhancement and soft-weighted covariance modeling. The resulting multi-level statistics are aggregated into compact global descriptors. During training, both intermediate statistical features and final descriptors are jointly optimized via multi-level triplet learning.

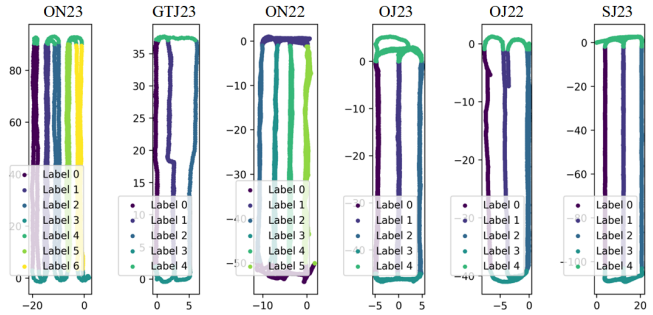

Fig. 2. 2D visualization of sequence paths with segmented regions.

a higher-dimensional space. Subsequently, $\varphi : \mathbb{R}^{n\times c} \rightarrow \mathbb{R}^{d}$ consolidates these local features into a compact global descriptor D, which serves as the basis for retrieval, following the paradigm in [28].

B. Local Feature Extraction

Given a raw LiDAR scan $P \in \mathbb{R}^{n\times 3}$, the goal of local feature extraction is to obtain discriminative point-wise embeddings that can later be aggregated into compact global descriptors. To this end, we adopt a point-based backbone, which applies shared multi-layer perceptrons (MLPs) to each point for hierarchical feature transformation. Specifically, the backbone outputs two levels of features: (i) high-dimensional point-wise features $x_{\text{feat}} \in \mathbb{R}^{B\times C\times N}$ capturing fine-grained geometric structures, and (ii) intermediate embeddings $x_{\text{mid}} \in \mathbb{R}^{B\times C}$ obtained through an auxiliary MLP branch that provides additional channel-level statistics.

To adaptively emphasize informative channels, we introduce a lightweight channel re-weighting mechanism. Specifically, a global average pooling is first applied to intermediate feature, producing a compact channel descriptor $g \in \mathbb{R}^{B\times C}$. This descriptor is then passed through a bottleneck fully connected layer, followed by a sigmoid activation, yielding channel-wise weights $s \in [0,1]^{B\times C}$. These weights are used to selectively amplify discriminative channels while suppressing redundant responses, enabling the network to dynamically focus on informative feature dimensions.The original point-wise features are then re-weighted as

$$F = x_{\text{feat}} \odot s, \tag{1}$$

where $\odot$ denotes channel-wise multiplication. This design allows the network to emphasize discriminative channels while reducing noise, combining point-wise embeddings with channel reweighting to produce robust features that support covariance modeling and global descriptor learning.

C. Feature Aggregation

After obtaining the channel-enhanced local features $F \in \mathbb{R}^{B\times C\times N}$, we aggregate them into compact global descriptors through a joint modeling of first-order and second-order statistics, as illustrated at the top of Fig. 1.

a) First-order statistics.: We first compute the mean feature vector across all points:

$$F_{\text{mean}} = \frac{1}{N}\sum_{i=1}^{N} F_{c,i}, \tag{2}$$

where $F_{c,i} \in \mathbb{R}^{C}$ denotes the feature vector of the i-th point. This captures the overall distribution of point-wise embeddings.

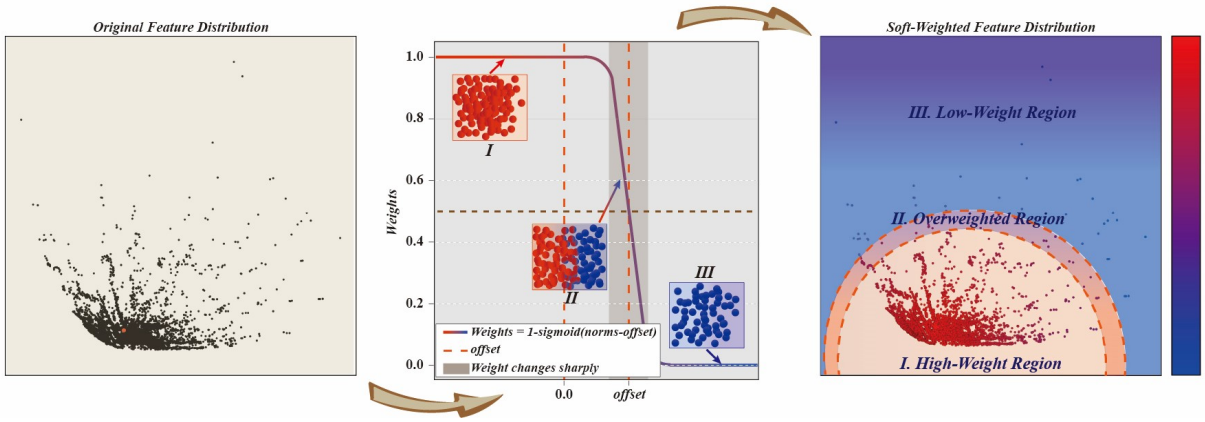

Fig. 3. PCA visualization of the Soft-weighted Covariance pipeline: raw feature distribution (left), weighting process (middle), and soft-weighted feature distribution (right). The learnable weighting suppresses outliers and enhances discriminative structure.

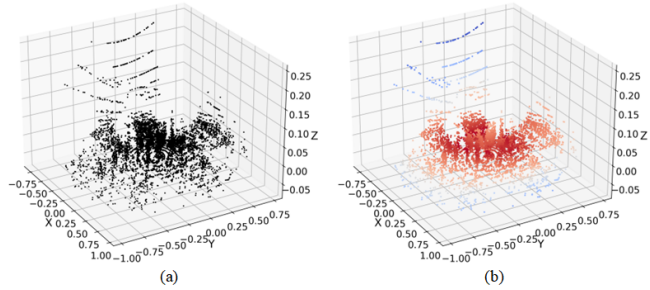

Fig. 4. (a) Original points in black. (b) Points colored by soft weights, red for high weights and blue for low weights.

b) Second-order statistics.: Then, each point feature is centered by subtracting the mean:

$$\tilde{F}_i = F_{c,i} - F_{\text{mean}}, \quad r_i = \|\tilde{F}_i\|_2. \tag{3}$$

Here r_i measures the deviation of each point from the global mean, which reflects its relative importance.

To suppress outliers and repetitive structures, we design a soft weighting function based on a dynamic offset:

$$o = \frac{1}{N}\sum_{i=1}^{N} r_i + \lambda \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(r_i - \frac{1}{N}\sum_{k=1}^{N} r_k\right)^2}, \tag{4}$$

where λ is a learnable scaling factor. This offset adaptively balances between the mean and the dispersion of residual norms. The normalized weights are then computed as

$$\hat{w}_i = \frac{1-\sigma(r_i - o)}{\sum_{j=1}^{N}[1-\sigma(r_j - o)]}, \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid function. This design assigns higher weights to informative points while smoothly suppressing less discriminative ones.

The weighted covariance matrix is obtained as

$$F_{\text{cov}} = \text{flatten}\left(\sum_{i=1}^{N} \hat{w}_i \tilde{F}_i \tilde{F}_i^{\top}\right). \tag{6}$$

This second-order representation captures the dispersion and correlations across channels, providing richer geometric information than mean pooling alone.

The pipeline of the Soft-weighted Covariance is illustrated in Fig. 3, where we visualize the feature distribution $\tilde{F}_i$ using PCA for dimensionality reduction. The left panel shows the raw point cloud feature distribution, the middle panel depicts the weighting process, and the right panel presents the resulting soft-weighted features. Directly computing covariance over raw features suffers from the influence of outliers, leading to unstable representations. To address this, we introduce a learnable offset parameter that shapes the weighting function: within the offset boundary, points receive nearly uniform weights; near the boundary, weights decay rapidly; and for distant outliers, weights approach zero.

This design is particularly effective in orchard environments, where semi-transparent foliage produces sparse, ill-defined geometry and high structural similarity across rows. Such conditions exacerbate intra-row and inter-row ambiguity, severely impairing descriptor discriminability. As shown in Fig. 4, when visualizing points colored by soft weights, most noisy or interfering points are suppressed, while regions with significant structural features are highlighted. By adaptively suppressing noisy and overlapping points while preserving structurally informative ones, the proposed soft-weighted covariance provides a more robust statistical representation, directly mitigating the ambiguity inherent in repetitive orchard layouts.

c) Feature fusion.: Finally, the first-order descriptor F_{mean} and the second-order descriptor F_{cov} are projected into a common space through independent linear layers, and then concatenated:

$$F_{\text{fusion}} = \text{MLP}\big([F_{\text{mean}}; F_{\text{cov}}]\big), \tag{7}$$

yielding a compact global descriptor of dimension d. This fusion strategy leverages the complementary strengths of first-order and second-order statistics, enhancing robustness against both noisy outliers and highly repetitive orchard structures.

D. Training

We adopt a hybrid training strategy following PointNetPGAP-SLC [18], which combines a retrieval-oriented triplet loss [8] with a segment-level consistency regularization. The overall objective is defined as

$$\mathscr{L}_{\text{total}} = \alpha \cdot \mathscr{L}_T + (1-\alpha) \cdot \mathscr{L}_S, \tag{8}$$

where α is fixed at 0.5 as suggested in PointNetPGAPP-SLC.

a) Triplet loss.: To encourage descriptors from the same location to remain close while pushing apart those from different locations, we adopt a Top-k LazyTripletLoss:

$$\mathscr{L}_{\text{tri}}(x) = \max\left(0, d(D^A, D^P) - \bar{d}(D^A, D^{N_{\text{top-k}}}) + m\right), \tag{9}$$

where D^A, D^P, D^N denote anchor, positive, and negative descriptors, respectively, $d(\cdot,\cdot)$ is the Euclidean distance, $D^{N_{\text{top-k}}}$ denotes the k hardest negative descriptors, aggregated by mean to compute $\bar{d}$, and m is a predefined margin. In our experiments, we set $k = 5$ to balance mining challenging negatives with training stability. Compared to standard lazy triplet loss, this Top-k variant considers multiple hard negatives rather than only the single closest negative, providing more informative gradient signals and improving convergence, while reducing sensitivity to outlier negatives.

Instead of relying solely on the fused global descriptor, we jointly supervise three complementary representations:

$$\mathscr{L}_T = \tfrac{1}{2}\left[\mathscr{L}_{\text{tri}}(D) + \tfrac{1}{2}\mathscr{L}_{\text{tri}}(D_{mean}) + \tfrac{1}{2}\mathscr{L}_{\text{tri}}(D_{cov})\right], \tag{10}$$

where D denotes the fused descriptor, while D_{mean} and D_{cov} correspond to the first-order and second-order representations, respectively. This multi-view supervision enforces discriminability across different levels of feature statistics.

b) Segment-level consistency loss.: To further regularize the descriptor space, we follow the segment-level consistency (SLC) training strategy [18]. Specifically, an auxiliary classifier takes the global descriptors D as input and predicts segment-level labels (as illustrated in Fig. 2). The corresponding loss is formulated as:

$$\mathscr{L}_S = \text{NLLLoss}\big(\text{MLP}(D), \text{Label}\big), \tag{11}$$

where $\text{MLP}(D)$ denotes a small multi-layer perceptron that maps the global descriptor D to a prediction over segment-level labels, and Label represents the ground-truth segment ID. The NLLLoss then penalizes discrepancies between the predicted and true segment labels.

In summary, the Top-k LazyTripletLoss enforces discriminative retrieval, while the segment-level consistency loss regularizes descriptors within the same spatial segment.

IV. Experiments

This section presents the evaluation protocols, outlines the implementation details, and reports both quantitative and qualitative retrieval results along with runtime analysis.

A. Dataset and Evaluation Protocols

We conduct experiments on the HORTO-3DLM+ dataset [16], [29], which contains six sequences collected from diverse horticultural environments. Following the evaluation protocol of PointNetPGAP [18], each sequence is divided into multiple segments for training and testing. During training, an anchor−positive pair is defined when the two scans are within a radius of 2 m, while anchors are sampled at least 0.5 m apart to reduce redundancy.

For evaluation, we adopt a cross-validation strategy: models are trained on five sequences and tested on the remaining one, and this process is repeated six times so that each sequence serves once as the test set. A retrieved candidate is considered a true positive if it is located within 10 m of the query and belongs to the same segment; otherwise, it is treated as a false positive. Performance metrics are reported using Recall for the top−k retrieved candidates (i.e., Recall@k).This protocol ensures a fair comparison across different methods and emphasizes generalization across distinct horticultural environments.

B. Implementation Details

To ensure fairness, all models were trained and evaluated under identical settings. Each model was optimized for 50 epochs with early stopping to select the best checkpoint on the validation set. The input scans were uniformly downsampled to 10k points, and randomly rotated around the z-axis. Training was conducted on an NVIDIA GeForce RTX 3090 GPU. For each anchor, the nearest positive and 20 negatives were sampled ($m = 20$). The margin in the contrastive loss was fixed to 0.5. Model parameters were optimized using the AdamW optimizer [30] with a learning rate of 1×10^{-4} and a weight decay of 5×10^{-4}.

C. Evaluation

1) Comparison to State-of-the-Art: In this section, we present and analyze the empirical results. We compare the proposed SCDCE-3D against several state-of-the-art (SOTA) methods, including PointNetVLAD [8], OverlapTransformer [9], LOGG3D-Net [10], SPVSoAP3D [16], and PointNetPGAP [18]. The quantitative results under a fixed retrieval radius of $r_{th} = 10$ m, summarized in Table I, demonstrate that SCDCE-3D consistently outperforms all state-of-the-art baselines under the adopted evaluation protocol. In particular, at the stringent threshold of $r_{th} = 10$ m, our framework achieves an average improvement of 5.6 percentage points in top-1 recall over the second-best baseline, together with a 2.8 percentage point gain in top-5 recall.

Furthermore, Fig. 5 illustrates the performance variation as the retrieval radius r_{th} increases from 0 to 50 m across all six evaluation sequences. The results reveal several consistent trends. First, as r_{th} grows, the recall naturally improves for all methods due to the relaxed matching criteria. More importantly, our method maintains a clear performance advantage across the entire range of thresholds, with Recall@1 consistently higher than all baselines. This superiority holds not only at stringent thresholds, where descriptor discriminability

TABLE I

PERFORMANCE AT A 10 M RETRIEVAL RADIUS, REPORTED AS RECALL@k WITH $k \in [1,5]$. THE BOTTOM THREE ROWS SHOW OUR ABLATIONS: 'D' DENOTES THE DUAL-BRANCH CHANNEL ENHANCEMENT MODULE, 'S' THE SPATIAL CONSISTENCY MODULE, AND 'M' THE MULTI-LEVEL LEARNING COMPONENT. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD, WHILE THE SECOND-BEST RESULTS ARE UNDERLINED.

				Recall@1							Recall@5						
Model				ON23	GTJ23	ON22	OJ23	OJ22	SJ23	MEAN	ON23	GTJ23	ON22	OJ23	OJ22	SJ23	MEAN
PointNetVLAD [8]				0.307	0.678	0.600	0.521	0.519	0.719	0.557	0.510	0.780	0.662	0.557	0.779	0.837	0.688
OverlapTransformer [9]				0.068	0.129	0.208	0.143	0.054	0.073	0.116	0.151	0.434	0.437	0.313	0.127	0.164	0.271
LOGG3D-Net [10]				0.124	0.431	0.496	0.395	0.356	0.308	0.352	0.344	0.742	0.701	0.703	0.696	0.532	0.620
SPVSoAP3D [16]				0.280	0.759	0.696	0.696	0.883	0.548	0.644	0.590	**0.908**	**0.826**	0.856	0.950	0.770	0.816
PointNetPGAP [18]				0.394	0.715	0.705	0.688	0.861	0.690	0.675	0.640	0.831	0.816	0.636	0.901	0.830	0.776
	D	S	M	ON23	GTJ23	ON22	OJ23	OJ22	SJ23	MEAN	ON23	GTJ23	ON22	OJ23	OJ22	SJ23	MEAN
Ours	✓			**0.476**	0.703	0.678	0.667	0.890	0.675	0.681	**0.722**	0.816	0.790	0.719	0.943	0.815	0.801
	✓	✓		0.447	0.722	**0.725**	0.785	0.914	0.720	0.719	0.676	0.843	0.818	0.849	**0.973**	0.861	0.836
	✓	✓	✓	0.460	**0.763**	0.716	**0.803**	**0.915**	**0.730**	**0.731**	0.702	0.871	0.822	**0.861**	0.941	**0.868**	**0.844**

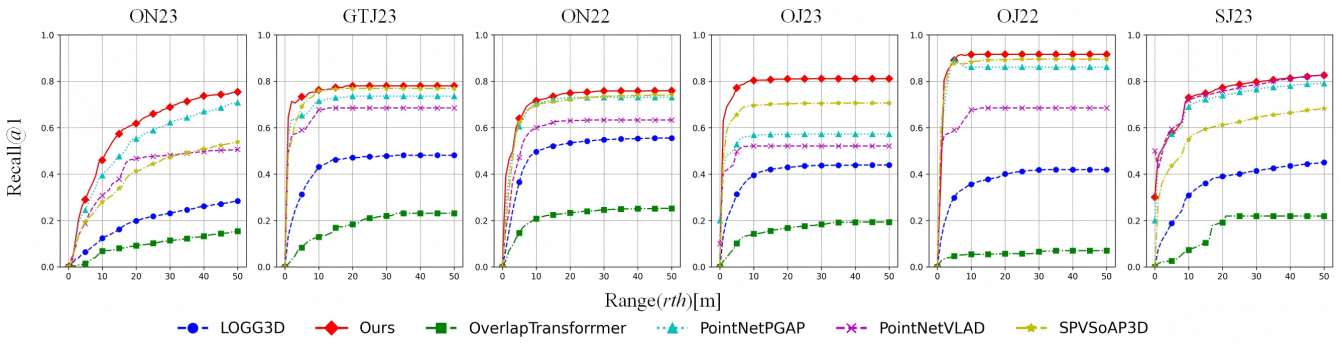

Fig. 5. Performance results along the segments, reported using the Recall@1 for rth [1, .., l] where l is the segment length.

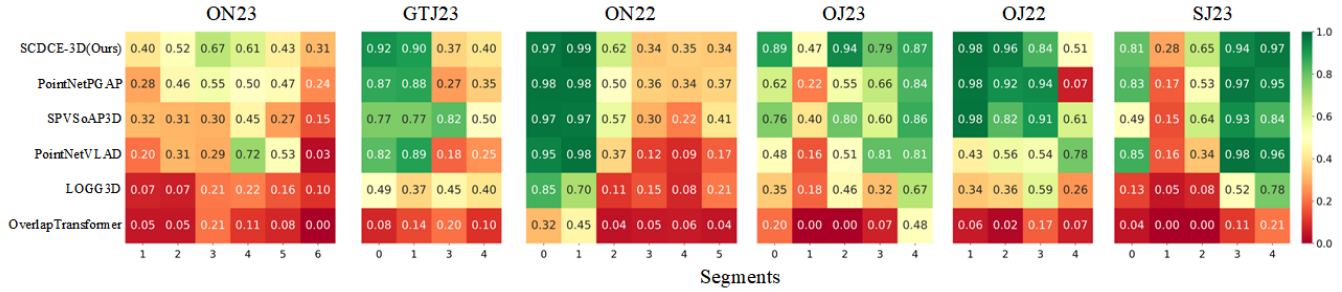

Fig. 6. Segment-level loop closure retrieval performance on six sequences. Each heatmap shows the Top-1 true positive rate, where cooler tones indicate higher reliability.

is critical, but also at more relaxed settings, demonstrating the robustness of our framework under different retrieval conditions. Notably, the performance margin of SCDCE-3D remains stable across the six sequences, highlighting its ability to generalize across diverse orchard layouts and scanning conditions.

Beyond overall accuracy, we further analyze performance on a per-sequence basis. As shown in Fig. 6, we present segment-level Recall@1 across six orchard sequences. Each heatmap depicts the true positive rate of the top-1 retrieved candidate, where warmer colors indicate failures and cooler tones denote reliable retrievals. Overall, SCDCE-3D consistently produces higher-confidence predictions at the segment level compared to competing methods.

In addition, we visualize the true positive predictions of the top-1 retrieved candidates. In Fig. 7, a qualitative comparison shows how each model performs along the paths for Top-1 candidate retrieval. In certain segments, some models either predict incorrect loops or fail to identify any loops correctly, whereas our method substantially reduces the error rate.

In addition to accuracy, computational efficiency and model size are crucial for real-world deployment. Table II reports the runtime and number of parameters for all compared methods. Runtime was measured on an NVIDIA GeForce RTX 3090 GPU, with the reported values (in milliseconds) representing the average time to process a batch of 20 scans across six sequences. Retrieval latency is not included. Model size is reported as the number of parameters (in millions, M).

The results show that SCDCE-3D balances efficiency and accuracy, with only 0.5M parameters and a runtime of 5.89ms per batch of 20 scans. Its compact design lowers memory and computation requirements, making it suitable

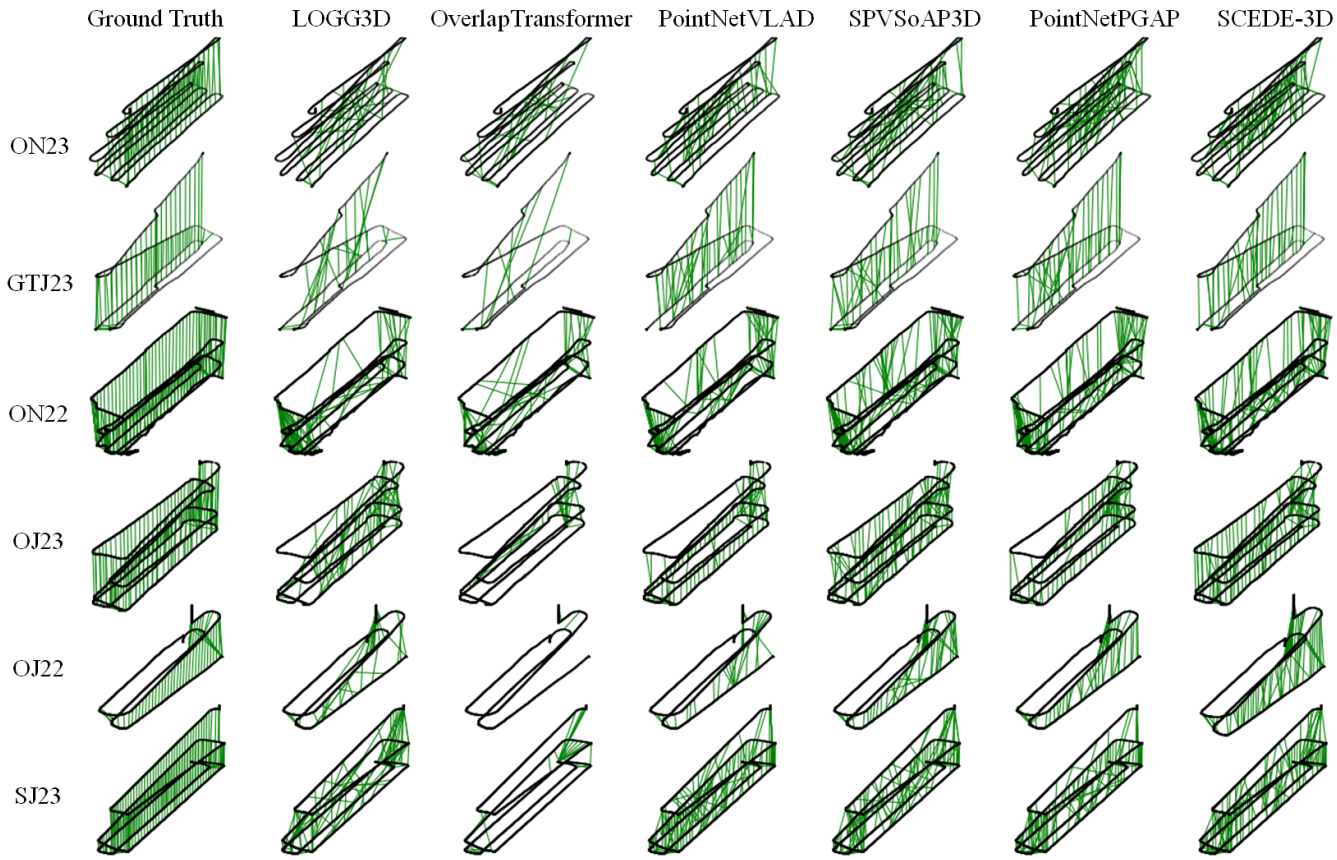

Fig. 7. True positive predictions of the top-1 retrieved candidates.

for real-time deployment on agricultural robots.

TABLE II

THE AVERAGE RUNTIME PER BATCH, EVALUATED WITH A BATCH SIZE OF 20 SCANS.

	Param. [M]	MEAN [ms]
PointNetVLAD [8]	19.8	44.51
OverlapTransformer [9]	48.2	12.29
LOGG3D [10]	8.8	256.07
SPVSoAP3D [16]	8.9	257.54
PointNetPGAP [18]	0.4	3.42
Ours	0.5	5.89

2) Ablation Study: To systematically evaluate the contribution of each component in SCDCE-3D, we conducted a series of ablation studies, with results summarized in the lower part of Table I. Specifically, 'D' denotes the dual-branch channel enhancement module, 'S' represents the soft-weighted covariance operator, which is replaced by standard covariance when disabled, and 'M' refers to the multi-level triplet learning strategy, where supervisory signals are applied not only to the final global descriptor but also to intermediate statistical features. When Multi-level Triplet Learning is removed, the loss function is restricted to the final descriptors alone. If all three modules are removed, the resulting architecture resembles conventional statistical descriptor methods such as PointNetPGAP. The ablation

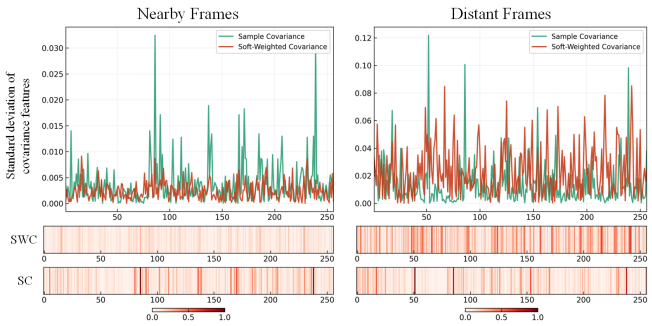

Fig. 8. Standard deviation of covariance features and inter-sample difference heatmaps. With Soft-weighted Covariance, intra-row variance is reduced while inter-row differences are enhanced.

results indicate that all three components play a crucial role in enhancing performance and that their effects are complementary, collectively contributing to the robustness of the framework.

To further investigate the mechanism of SWC, we provide a visualization in Fig. 8. By comparing the standard deviation distributions of F_{cov} across corresponding dimensions between adjacent and non-adjacent frames, as well as heatmaps of inter-sample differences, we observe that SWC substantially reduces intra-row variance while amplifying inter-row differences. This effect can be attributed to the sigmoid-

based weighting strategy, which adaptively down-weights noisy points and overlapping regions while assigning higher importance to structurally discriminative points. In essence, SWC acts as a joint mechanism of feature denoising and discriminative enhancement. Consequently, in adjacent frames, it suppresses the noise introduced by semi-transparent foliage and stabilizes statistical features, reducing random fluctuations. In contrast, across non-adjacent frames, it highlights structural differences between rows, mitigating descriptor ambiguity in repetitive orchard layouts. Ultimately, this dual effect directly translates into substantial gains in retrieval performance, demonstrating the effectiveness of SWC in complex agricultural environments.

V. Conclusions

We present SCDCE-3D, a tailored framework for 3D place recognition in complex orchard environments, designed to address structural repetitiveness and sensor noise inherent in agricultural settings. The approach integrates dual-branch channel enhancement to capture complementary geometric patterns, soft-weighted covariance aggregation to emphasize discriminative structures while suppressing noise, and multi-level triplet learning to enforce consistency across both global and statistical representations. Evaluations on the HORTO-3DLM+ benchmark demonstrate the robustness of our design, highlighting its potential for reliable long-term localization and navigation in agricultural robotics.

References

[1] Y. Liu, J. Ji, D. Pan, L. Zhao, and M. Li, "Localization method for agricultural robots based on fusion of lidar and imu," *Smart Agriculture*, vol. 6, no. 3, pp. 94–106, 2024.

[2] A. Navone, M. Martini, and M. Chiaberge, "Autonomous robotic pruning in orchards and vineyards: a review," *arXiv preprint arXiv:2505.07318*, 2025.

[3] M. Chen, Z. Chen, L. Luo, Y. Tang, J. Cheng, H. Wei, and J. Wang, "Dynamic visual servo control methods for continuous operation of a fruit harvesting robot working throughout an orchard," *Computers and electronics in agriculture*, vol. 219, p. 108774, 2024.

[4] S. Jiang, P. Qi, L. Han, L. Liu, Y. Li, Z. Huang, Y. Liu, and X. He, "Navigation system for orchard spraying robot based on 3d lidar slam with ndt_icp point cloud registration," *Computers and Electronics in Agriculture*, vol. 220, p. 108870, 2024.

[5] O. Abdi, J. Uusitalo, J. Pietarinen, and A. Lajunen, "Evaluation of forest features determining gnss positioning accuracy of a novel low-cost, mobile rtk system using lidar and treenet," *Remote Sensing*, vol. 14, no. 12, p. 2856, 2022.

[6] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *IEEE Robotics & Automation Magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[7] S. Lowry, N. Snderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[8] 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 (CVPR)*, June 2018.

[9] 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.

[10] 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)*. IEEE, 2022, pp. 2215–2221.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, "Mulran: Multimodal range dataset for urban place recognition," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 6246–6253.

[16] T. Barros, C. Premebida, S. Aravecchia, C. Pradalier, and U. J. Nunes, "Spvsoap3d: A second-order average pooling approach to enhance 3d place recognition in horticultural environments," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 9–15.

[17] N. Sun, Z. Fan, Q. Qiu, T. Li, Q. Feng, C. Ji, and C. Zhao, "Triloc-netvlad: Enhancing long-term place recognition in orchards with a novel lidar-based approach," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 16–22.

[18] T. Barros, L. Garrote, P. Conde, M. J. Coombes, C. Liu, C. Premebida, and U. J. Nunes, "Pointnetpgap-slc: A 3d lidar-based place recognition approach with segment-level consistency training for mobile robots in horticulture," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 10 471–10 478, 2024.

[19] Y. Zhang, P. Shi, and J. Li, "Lidar-based place recognition for autonomous driving: A survey," *ACM Computing Surveys*, vol. 57, no. 4, pp. 1–36, 2024.

[20] M. Pauly, R. Keiser, and M. Gross, "Multi-scale feature extraction on point-sampled surfaces," in *Computer graphics forum*, vol. 22, no. 3. Wiley Online Library, 2003, pp. 281–289.

[21] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *2009 IEEE international conference on robotics and automation*. IEEE, 2009, pp. 3212–3217.

[22] 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.

[23] 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*, 2016, pp. 5297–5307.

[24] J. Komorowski, "Minkloc3d: Point cloud based large-scale place recognition," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1790–1799.

[25] B. Graham, "Sparse 3d convolutional neural networks," *arXiv preprint arXiv:1505.02890*, 2015.

[26] 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.

[27] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 529–10 538.

[28] T. Barros, L. Garrote, R. Pereira, C. Premebida, and U. J. Nunes, "Attdlnet: Attention-based deep network for 3d lidar place recognition," in *Iberian Robotics conference*. Springer, 2022, pp. 309–320.

[29] 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.

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.