

Article

Motion-Adaptive Online Loop Closure Detection via Foundation Descriptors and Geometric Consistency

Duc Thinh Duong ¹  and Jong-Eun Ha ^{2,*} 

¹ Graduate School of Automotive Engineering, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea; ducthinh@seoultech.ac.kr

² School of Mechanical Systems Engineering-Automotive Engineering, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea

* Correspondence: jeha@seoultech.ac.kr

Abstract

This paper proposes a robust visual loop closure detection pipeline for outdoor environments, utilizing DINOv3-SALAD, a foundation Transformer-based model. Given a sequence of outdoor images, an adaptive temporal buffering mechanism dynamically rejects neighboring frames when feature similarity drops. This effectively prevents false positive loops caused by variable vehicle velocities, which result in inconsistent rates of visual change between consecutive frames. To obtain geometrically valid and locally optimal loop candidates, the visually retrieved pairs are further verified using trajectory-based distance ratios and filtered via 1D Non-Maximum Suppression (NMS). The resulting loop closures are then quantitatively and qualitatively compared with a traditional fixed-interval retrieval approach to evaluate the accuracy, robustness, and effectiveness of the proposed method in handling dynamic outdoor conditions.

Keywords: loop closure detection; simultaneous localization and mapping (SLAM); vision transformers

1. Introduction

With the rapid advancement of intelligent mobile systems, such as autonomous vehicles and robotic navigation, the importance of Simultaneous Localization and Mapping (SLAM) [1,2] has continued to increase for achieving long-term, reliable localization and mapping in complex environments. To compensate for the accumulated localization error that inevitably arises during long-term operation, a SLAM system requires robust Visual Loop Closure Detection (VLCD), which enables the robot to recognize previously visited places and establish loop constraints for trajectory correction. However, outdoor environments present significant challenges for place recognition due not only to drastic illumination changes but also to perceptual aliasing caused by repetitive structures, such as similar buildings and roadside trees. Furthermore, in real driving scenarios, vehicle speed continuously varies and vehicles frequently stop at intersections or traffic signals. These dynamic motion characteristics become a major factor degrading the reliability of loop closure prediction.

Traditional visual place recognition and loop closure detection methods have mainly relied on hand-crafted local features [3,4] or Bag-of-Words (BoW) representations [5]. Although effective under limited viewpoint variations, their recognition performance deteriorates significantly under large viewpoint and environmental changes. In addition, to suppress redundant consecutive frames acquired during stationary or low-speed driving,



Academic Editors: Pengcheng Shi and Qin Ye

Received: 2 July 2026

Revised: 27 July 2026

Accepted: 6 August 2026

Published: 7 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

conventional retrieval systems typically exclude a fixed number of neighboring frames from the search space using a fixed-interval exclusion strategy. However, this approach suffers from an inherent limitation because it does not account for the actual motion of the vehicle. When the vehicle moves slowly, a fixed exclusion interval may produce false loop closure predictions by allowing highly redundant frames into the search. Conversely, when the vehicle moves rapidly, valid loop closure candidates may be mistakenly discarded, resulting in missed loop detections.

Recently, state-of-the-art systems such as VGGT-Long [6] and DA3-Streaming [7] have adopted foundation-model-based global descriptors, particularly DINOv2-SALAD [8], for visual loop closure detection and have demonstrated noticeable performance improvements. Nevertheless, these methods still exhibit a structural limitation: even when two images correspond to the same physical location, their cosine similarity can decrease substantially under large viewpoint changes, causing valid loop closures to be missed. Simply lowering the similarity threshold is not an effective solution, since it dramatically increases the number of false loop closures, while genuine loop closures may still remain undetected if their descriptor similarity does not exceed the lowered threshold.

To fundamentally address these limitations, this paper proposes a novel visual loop closure detection framework. First, to overcome the shortcomings of fixed-interval exclusion, we introduce an Adaptive Temporal Buffering (ATB) mechanism that dynamically determines the temporal exclusion range by tracking the similarity decay between consecutive feature descriptors. Second, to prevent the increase in false loop closures caused by lowering the similarity threshold, we propose a local optimization strategy that combines geometric verification based on the traveled-distance ratio along the trajectory with one-dimensional Non-Maximum Suppression (1D NMS). Third, to maintain high descriptor similarity for genuine loop closures under large viewpoint changes, we replace the backbone of DINOv2-SALAD with the recently introduced DINOv3-SALAD, whose visual representations are substantially more robust to viewpoint variation. The proposed method is evaluated on a self-collected outdoor loop-closure dataset whose ground-truth trajectories are reconstructed using COLMAP [9]. Through comprehensive comparisons with the conventional FAISS-based [10] fixed-interval retrieval framework, we demonstrate that the proposed pipeline effectively suppresses false loop closures while maintaining robust detection performance under variable vehicle speeds and challenging outdoor environments.

The main contributions of this paper are summarized as follows:

- (1) We formulate temporal-neighbor rejection as a motion-adaptive search-space pruning problem rather than a fixed-frame exclusion problem.
- (2) We introduce a similarity-decay-based adaptive temporal buffer that suppresses redundant neighboring frames while preserving valid long-term loop closure candidates.
- (3) We combine descriptor-based retrieval with trajectory-consistency verification and 1D Non-Maximum Suppression (1D NMS) to effectively reduce false loop closures before pose-graph optimization.
- (4) We demonstrate that the proposed pipeline significantly improves loop closure reliability under stop-and-go vehicle motion and severe viewpoint variation using both KITTI sequences and a self-collected outdoor dataset.

2. Related Works

Early Visual Place Recognition (VPR) and Visual Loop Closure Detection (VLCD) systems primarily relied on the combination of hand-crafted feature extraction algorithms, such as SIFT [3], SURF [4], and ORB [11], with the Bag-of-Words (BoW) model [5]. Although VPR and VLCD are closely related and rely on similar image-retrieval techniques, they are not identical: VPR is a broader task that aims to recognize the location depicted

in a query image against a large geo-referenced database, whereas VLCD is a dedicated module within a SLAM system that must additionally enforce temporal and geometric consistency to establish reliable loop constraints for trajectory correction. The BoW framework constructs a visual vocabulary by clustering extracted local features and computes image similarity based on the occurrence frequency of visual words. However, because this traditional approach ignores the global spatial context of an image, its performance deteriorates significantly under challenging conditions such as illumination changes (e.g., shadows) and environmental variations. Moreover, to suppress redundant consecutive frames, most conventional systems employ a fixed-interval filtering strategy that excludes a predefined number of neighboring frames from the retrieval process. Since this strategy does not account for variations in vehicle speed, it often degrades the reliability of loop closure detection.

To overcome the limitations of hand-crafted features, recent VPR research has shifted from convolutional neural network (CNN)-based approaches toward Vision Transformer (ViT)-based foundation models. Through the self-attention mechanism, Transformers capture global interactions among image patches and learn robust visual representations that are less sensitive to illumination and environmental changes. In particular, state-of-the-art methods such as VGGT-Long [6] and DA3-Streaming adopt the DINOv2-SALAD [8] architecture to improve VLCD performance. This framework combines the self-supervised foundation model DINOv2 [12] with the SALAD spatial feature aggregation module to generate highly discriminative global descriptors. Compared with conventional BoW-based methods, DINOv2-SALAD better preserves both spatial and semantic context, leading to improved place recognition accuracy in outdoor environments. Nevertheless, although DINOv2 effectively extracts dense pixel-level patch features, substantial viewpoint or perspective changes at the same physical location can significantly reduce the cosine similarity between descriptors, causing valid loop closures to fall below the retrieval threshold and resulting in false-negative detections.

DINOv3 [13], released by Meta AI in August 2025, is the latest generation of the DINO family and serves as the successor to DINOv2. Compared with its predecessor, DINOv3 is trained on a dataset approximately twelve times larger and scales the model size to seven billion parameters (ViT-7B/16). More importantly, it introduces a novel Gram Anchoring mechanism to address the dense feature collapse problem observed during long-term self-supervised training. By regularizing the Gram matrix, which captures the second-order statistics among patch features, Gram Anchoring preserves patch-level locality and spatial consistency throughout training. Consequently, DINOv3 effectively mitigates the degradation of dense feature representations observed in DINOv2 and achieves superior performance on dense prediction tasks, including semantic segmentation, depth estimation, and object detection.

In this work, we replace the DINOv2 backbone of DINOv2-SALAD with DINOv3 to construct a DINOv3-SALAD pipeline. While preserving the original Optimal Transport-based feature aggregation module of SALAD [14], the proposed framework exploits the stronger dense feature representations and improved viewpoint robustness of DINOv3. As a result, it maintains consistently high cosine similarity between descriptors of the same location even under significant viewpoint variations, thereby improving the robustness of visual loop closure detection in challenging outdoor environments.

3. Method

3.1. Overview of DA3-Streaming

DA3-Streaming [7] is a streaming visual SLAM pipeline designed to enable Depth Anything 3 (DA3) [15] to efficiently process long video sequences and large-scale scenes

under limited CPU and GPU memory resources. The framework directly adopts the core ideas of VGGT-Long [6], including chunk-based sequence partitioning, inter-chunk state propagation, and global optimization through loop closure detection.

Following the chunk-based processing strategy of VGGT-Long, DA3-Streaming divides an input video into overlapping chunks and processes each chunk independently, thereby avoiding the need to load the entire sequence into memory simultaneously. The estimated state of each chunk is propagated to the subsequent chunk to preserve the continuity of camera trajectories and reconstructed point clouds. When a loop closure is detected between non-adjacent chunks, a lightweight loop closure optimization module performs global trajectory refinement to reduce accumulated drift while integrating the reconstructed 3D map with only a modest computational overhead.

Despite their similar overall framework, DA3-Streaming and VGGT-Long differ substantially in both their backbone architectures and the way geometric information is generated. VGGT-Long employs the VGGT [16] backbone to jointly estimate depth, camera parameters, and dense 3D points for each chunk. The predicted 3D points are subsequently used to identify overlapping regions between neighboring chunks, estimate the corresponding Sim(3) transformation, and align both the camera trajectories and reconstructed point clouds into a globally consistent map.

In contrast, DA3-Streaming adopts Depth Anything 3 as its backbone. Unlike VGGT, DA3 does not directly predict dense 3D points. Instead, it estimates dense depth maps together with camera parameters, from which the corresponding 3D point clouds are reconstructed through geometric back-projection. These reconstructed 3D points are then utilized in the same manner as VGGT-Long for Sim(3)-based alignment between adjacent chunks. Consequently, although both methods employ the same alignment strategy, DA3-Streaming differs fundamentally in that its geometric representation is reconstructed rather than directly predicted.

Replacing the VGGT backbone with Depth Anything 3 alone yields substantial performance improvements over VGGT-Long. On the TUM RGB-D benchmark, DA3-Streaming achieves an average Absolute Trajectory Error (ATE) of 0.087, representing approximately a 21% reduction compared with 0.110 obtained by VGGT-Long, while consistently outperforming the baseline across most evaluation sequences. Furthermore, on KITTI [17] Sequences 00, 05, and 08 (11,373 frames in total), DA3-Streaming achieves an inference speed of 8.51 FPS, which is nearly three times faster than the 2.91 FPS reported by VGGT-Long under identical evaluation settings (chunk size of 120 frames, overlap of 60 frames, NVIDIA A100 GPU, and without keyframe selection). These improvements are primarily attributed to the superior depth estimation and camera parameter prediction capabilities of Depth Anything 3. As a result, the reconstructed 3D points become more accurate, the Sim(3) alignment between chunks becomes more reliable, and the accumulated trajectory drift is significantly reduced.

Motivated by these advantages, this work adopts DA3-Streaming as the underlying SLAM framework and further enhances its visual loop closure module by incorporating DINOv3-SALAD. First, we introduce an Adaptive Temporal Buffering (ATB) mechanism that dynamically adjusts the temporal exclusion range according to the similarity decay between consecutive descriptors, allowing the system to adapt to varying vehicle speeds. Second, we perform local optimization by combining trajectory-based geometric verification with one-dimensional Non-Maximum Suppression (1D NMS) to effectively suppress false loop closures introduced by relaxed similarity thresholds. Finally, by integrating the viewpoint-robust DINOv3-SALAD descriptor into the retrieval pipeline, the proposed system maintains consistently high true loop closure detection performance even under severe viewpoint changes and complex outdoor environments.

3.2. DINOv3-SALAD

One of the key improvements of the proposed framework is the replacement of the DINOv2-SALAD [8] backbone with DINOv3 [13], resulting in the DINOv3-SALAD pipeline, whose overall architecture is illustrated in Figure 1. This modification is intended to improve robustness against viewpoint variations commonly encountered in outdoor environments while maintaining consistently high cosine similarity between descriptors of true loop pairs.

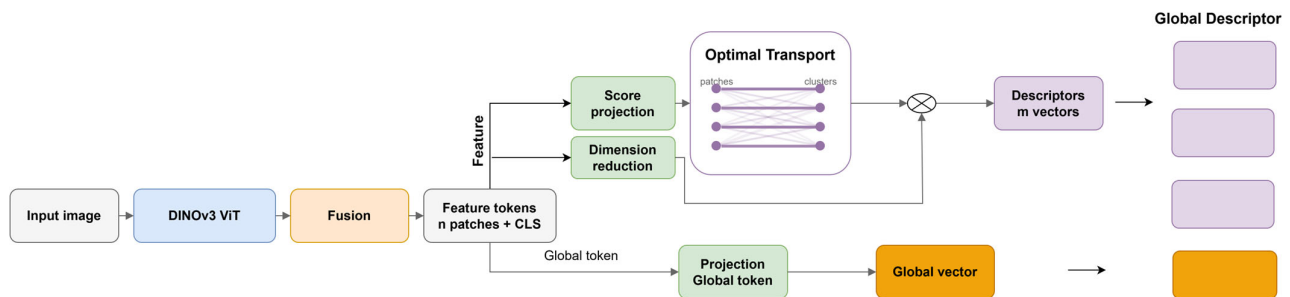


Figure 1. DINOv3-SALAD model architecture (the SALAD module is adopted from the original DINOv2-SALAD [8]).

DINOv2-SALAD employs DINOv2 [12] as its backbone to extract dense patch features and generates discriminative global descriptors through the Optimal Transport-based aggregation module of SALAD [14]. Although this architecture provides strong visual representations, the cosine similarity between descriptors often decreases substantially even under moderate viewpoint or perspective changes at the same physical location. Consequently, many genuine loop closures fall below the retrieval threshold, leading to false-negative detections.

To overcome this limitation, we retain the original Optimal Transport-based aggregation module of SALAD and replace only the backbone with DINOv3. Owing to its Gram Anchoring mechanism, DINOv3 effectively mitigates the dense feature collapse problem while preserving patch-level locality and spatial consistency throughout training. As a result, the extracted visual representations remain significantly more stable under severe viewpoint variations, allowing higher cosine similarity to be maintained for true loop pairs.

Specifically, the DINOv2 backbone is replaced with DINOv3 (ViT-B/16), which is fine-tuned on the GSV-Cities dataset. Following the original SALAD configuration, attention-based feature extraction is applied only to the last four transformer blocks. The SALAD aggregation module remains unchanged, including the original hyperparameter settings such as `num_clusters = 64`, `cluster_dim = 128`, and `token_dim = 256`.

The dense patch features extracted by DINOv3 are subsequently aggregated by SALAD to generate a compact global descriptor, and loop closure candidates are retrieved based on cosine similarity between descriptors. Unlike DINOv2-SALAD, which utilizes only the features from the final transformer layer, DINOv3 exploits representations from both intermediate and final transformer layers. These multi-level features are fused through a layer-fusion strategy to produce richer and more discriminative visual representations, thereby improving robustness to viewpoint changes and enhancing loop closure retrieval performance.

3.3. Adaptive Temporal Buffering

Loop closure is a fundamental component of visual SLAM that enables a robot or autonomous vehicle to recognize previously visited locations, optimize the pose graph, and recover the global consistency of the estimated trajectory and map. In outdoor envi-

ronments, however, reliable loop closure detection remains challenging due to viewpoint variation, illumination changes, and dynamic objects. Consequently, a practical loop closure system must simultaneously suppress false positives, which typically arise from temporally adjacent frames, while minimizing false negatives, in which genuine loop closures are missed. The primary objective of the proposed method is therefore to achieve accurate and robust loop closure detection for effectively correcting the accumulated drift encountered during long-term outdoor navigation.

Conventional retrieval frameworks, such as FAISS [10], identify loop closure candidates through approximate nearest-neighbor search over a global descriptor database. In contrast, the proposed approach exploits the sequential nature of video streams by comparing each query frame with all previously observed frames to identify potential loop closures. To eliminate temporally adjacent frames that provide little useful loop closure information, we introduce an Adaptive Temporal Buffering (ATB) mechanism.

The primary objective of ATB is to suppress temporal duplicates, which are a major source of false-positive detections caused by temporally neighboring frames, while preserving valid long-term loop closure candidates. Unlike conventional approaches based on fixed temporal windows or rigid distance thresholds, ATB dynamically analyzes the evolution of descriptor similarity in real time and automatically determines the effective retrieval range. Consequently, the proposed mechanism effectively removes long chains of redundant neighboring frames during stationary or low-speed motion while avoiding excessive suppression during high-speed driving. This adaptive behavior improves both the precision and recall of the overall loop closure detection system.

Let $\mathbf{d}_i$ denote the L2-normalized global descriptor extracted from frame i , the descriptor is defined as follows:

$$\mathbf{d}_i = \frac{f(I_i)}{\|f(I_i)\|_2} \quad (1)$$

where $f(I_i)$ denotes the feature descriptor extracted from image I_i .

The similarity between the query frame q and a previous frame p is defined as follows:

$$s_{q,p} = \mathbf{d}_q^T \mathbf{d}_p, 1 \leq p < q \quad (2)$$

The historical similarity sequence for the query frame q is defined as follows:

$$\mathbf{s}_q = [s_{q,1}, s_{q,2}, \dots, s_{q,q-1}] \quad (3)$$

The sequence $\mathbf{s}_q$ is a one-dimensional vector containing the cosine similarity scores between the current query frame q and all previously observed frames from frame 1 to frame $q - 1$.

Adaptive Neighbor Cut: Starting from $p = q - 1$, the similarity sequence is scanned in reverse order toward Frame 1 to determine the cut index k :

$$k = \max\{p | 1 \leq p < q \text{ and } s_{q,p} < \theta_n\} \quad (4)$$

Here, θ_n denotes the neighbor similarity threshold. The cut index k determined by Equation (4) corresponds to the nearest previous frame whose similarity to the query frame falls below θ_n . This index defines the adaptive search boundary by excluding only the temporally adjacent frames that remain highly similar to the query.

The loop closure search is then performed only over the interval from frame 1 to frame k . The frame with the highest similarity within this search range is selected as the loop

closure candidate. If its similarity score exceeds the threshold θ_s , the corresponding frame pair is accepted as the final loop closure candidate.

$$p^* = \arg \max_{1 \leq p \leq k} s_{q,p} \text{ if } s_{q,p} > \theta_s \quad (5)$$

3.4. Geometric Verification and 1d Non-Maximum Suppression

A loop closure fundamentally refers to the physical event in which a vehicle revisits a previously traversed location after traveling a considerable distance. Although the trajectory estimated by DA3-Streaming inevitably contains accumulated drift before loop closure optimization, it still exhibits a strong tendency to return to the vicinity of the previously visited location when a true loop closure occurs. This behavior is attributed to the robust depth estimation and relative motion prediction provided by DA3, which effectively suppress the accumulation of drift even over long video sequences without loop closure optimization.

Motivated by this observation, we introduce a simple geometric verification scheme that compares the Euclidean displacement between two camera poses with the corresponding traveled path length.

$$r_{q,p} = \frac{D}{S + \varepsilon}, \varepsilon = 10^{-6} \quad (6)$$

$$\text{Accept}(q, p^*) = \begin{cases} \text{True} & \text{if } r_{q,p} < \theta_r \\ \text{False} & \text{otherwise} \end{cases} \quad (7)$$

Here, D denotes the Euclidean displacement between the camera positions of Frames q and p , while S represents the total traveled path length from frame p to frame q . The quantity $r_{q,p}$ denotes the path ratio, and θ_r is the corresponding path-ratio threshold. Here, ε is a small positive constant (set to 10^{-6}) added to the denominator to ensure numerical stability by preventing division by zero when the traveled path length S is close to zero, for example, when the vehicle is stationary or moving very slowly.

For a genuine loop closure, the path ratio $r_{q,p}$ is expected to be smaller than the threshold because the two camera poses are spatially close despite being separated by a long temporal interval. This property accurately reflects the fundamental characteristic of a loop closure, namely that the vehicle revisits a previously explored location after traversing a long path. Conversely, candidate pairs satisfying $r_{q,p} > \theta_r$ are likely to correspond to false loop closures, even if they exhibit high cosine similarity, since their camera positions remain physically distant. Such candidates are therefore rejected during the geometric verification stage.

The loop closure candidates that pass geometric verification tend to form clusters within temporally adjacent regions. These redundant matches unnecessarily increase the computational burden of subsequent pose graph optimization. To address this issue, we adopt the one-dimensional Non-Maximum Suppression (1D NMS) algorithm introduced in the DA3-Streaming framework as the final filtering stage.

$$S \leftarrow S \cup \{x \mid |x - q| \leq \tau_{nms} \vee |x - p^*| \leq \tau_{nms}\} \quad (8)$$

Here, S denotes the suppression set, and τ_{nms} represents the temporal distance threshold used for one-dimensional Non-Maximum Suppression (1D NMS).

All candidate loop closure pairs, $\mathcal{C} = \{(q_i, p_i^*, \sigma_i^*)\}_{i=1}^M$ are first sorted in descending order according to their cosine similarity scores σ_i^* . The candidates are then processed sequentially from the highest score to the lowest. Whenever a candidate pair (q, p^*) is accepted into the final loop closure set $\mathcal{L}_{\text{final}}$, all temporally neighboring frames located

within a radius of τ_{nms} around either frame q or frame p^* are added to the suppression set S .

Any subsequent candidate involving an index already contained in S is immediately discarded without further evaluation. Consequently, the suppression set effectively eliminates redundant loop closure matches originating from the same temporal neighborhood, ensuring that only the highest-confidence correspondence is retained for each loop event. This strategy significantly reduces the computational cost of pose graph optimization while preserving loop closure accuracy.

4. Experimental Results

4.1. Experimental Setup

To comprehensively evaluate the proposed loop closure detection pipeline, experiments were conducted using both public benchmark datasets and a self-collected dataset. To verify the scalability and robustness of the proposed method in large-scale environments with multiple loop closures, KITTI [17] Sequences 00, 02, and 05 were selected for evaluation. These sequences cover extensive urban driving environments and contain multiple complex loop closure events, making them suitable for assessing the performance of large-scale visual SLAM systems.

To further evaluate the robustness of the proposed method under viewpoint variation, we constructed a dedicated outdoor dataset using an Intel RealSense D415 camera. The data were collected along a 72.3 m trajectory within a university campus under daytime conditions, and recorded at an average frame rate of 9.37 fps. The dataset consists of 1473 pairs of RGB images and depth maps with a resolution of 640×480 pixels. The camera focal lengths were ($f_x = f_y = 619$) pixels, and the depth scale was set to 1000.

To obtain reliable ground-truth trajectories for quantitative evaluation, the initial camera poses and sparse scene geometry were first reconstructed using COLMAP [9]. Since COLMAP suffers from scale ambiguity, depth measurements with high reliability (within 10 m) were extracted from the depth sensor and aligned with the point cloud reconstructed by COLMAP. The resulting scale adjustment factor was subsequently applied to the translation vectors, yielding an accurate ground-truth camera trajectory expressed in metric units. Notably, the self-collected dataset contains only a single loop closure event throughout the entire trajectory. To specifically evaluate robustness against viewpoint variation, the loop closure images were intentionally captured from noticeably different viewpoints despite representing the same physical location.

The proposed method was implemented on top of the DA3-Streaming framework. For memory-efficient streaming inference, the chunk size was fixed to 60 frames with an overlap of 30 frames between consecutive chunks. All experiments and model inference were performed on a single NVIDIA GeForce RTX 4090 GPU equipped with 24 GB of GPU memory.

4.2. Experimental Results and Analysis

To evaluate the robustness of the proposed descriptor under viewpoint variation, four representative ground-truth loop closure pairs extracted from the two datasets were selected, and the cosine similarity scores produced by DINOv2-SALAD and DINOv3-SALAD were compared. Figure 2 illustrates the four representative image pairs, while the quantitative results are summarized in Table 1.



Figure 2. Loop closure image pairs from the two datasets (a) pair1 (b) pair2 (c) pair3 (d) pair4.

Table 1. Comparison of cosine similarity scores between DINOv2-SALAD [7] and DINOv3-SALAD under different viewpoint conditions.

Model	Pair1 ↑	Pair2 ↑	Pair3 ↑	Pair4 ↑
DINOv2-SALAD	0.7375	0.6852	0.9538	0.8972
DINOv3-SALAD (ours)	0.8102	0.7658	0.9763	0.9277

Among the evaluated samples, Pairs 1 and 2 represent challenging loop closure cases involving significant viewpoint changes caused by variations in the vehicle’s traveling direction. In contrast, Pairs 3 and 4 correspond to ideal loop closure scenarios in which the viewpoints are nearly identical.

For the challenging cases (Pairs 1 and 2), DINOv3-SALAD consistently achieved substantially higher cosine similarity scores than DINOv2-SALAD, demonstrating stronger robustness to viewpoint variation. Even for the nearly identical-viewpoint cases (Pairs 3 and 4), DINOv3-SALAD still produced consistently higher similarity scores than its predecessor. These results indicate that replacing the DINOv2 backbone with DINOv3 enables the extraction of richer and more robust visual representations, thereby effectively mitigating the adverse effects of viewpoint variation and improving the reliability of visual loop closure detection.

To evaluate the performance of the proposed loop closure detection pipeline, two standard metrics, precision and recall, were adopted. In practical scenarios with severe viewpoint changes, the descriptor similarity of genuine loop closures often drops below 0.80; therefore, setting an excessively strict threshold (e.g., 0.85) would likely lead to the omission of valid loops. To address this, the cosine similarity threshold was intentionally reduced to 0.70. This reduction serves as a deliberate stress test to evaluate two key aspects: (1) the pipeline’s ability to retrieve genuine loop closures with low visual similarity, and (2) its capacity to suppress the inevitable influx of false-positive candidates under relaxed conditions. The quantitative results are summarized in Table 2.

Table 2. Comparison of loop detection performance between DA3-Streaming and the proposed method.

Dataset	DA3-Streaming [14]		Our Pipeline	
	Precision ↑	Recall ↑	Precision ↑	Recall ↑
KITTI sequence 00	95.24%	100%	100%	100%
KITTI sequence 02	100%	100%	100%	100%
KITTI sequence 05	83.3%	100%	100%	100%
Ours data	0%	0%	100%	100%

On KITTI Sequences 00, 02, and 05, both DA3-Streaming and the proposed method achieved 100% recall, indicating that no genuine loop closures were missed owing to the high visual similarity of the benchmark dataset. In particular, both methods achieved 100% precision on Sequence 02. However, clear differences were observed on Sequences 00 and 05. After lowering the cosine similarity threshold, DA3-Streaming produced one false-positive loop closure on each sequence, reducing the precision to 95.24% and 83.3%, respectively. In both cases, the false detections occurred when the vehicle slowed down or temporarily stopped before entering a curved road segment. Figure 3 presents the corresponding loop pairs detected on Sequence 00, including the false pair (pair 2) that was incorrectly recognized as a genuine loop closure.

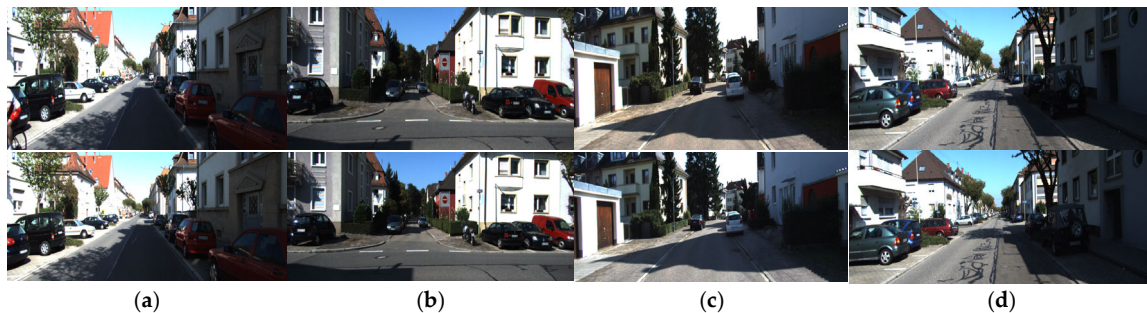


Figure 3. Loop pairs from the KITTI sequence 00 dataset, shown as panels (a) pair 1, (b) pair 2, (c) pair 3, and (d) pair 4. It includes true loop pairs: pair 1 (3627↔ 680), pair 3 (1595↔ 150), and pair 4 (4590↔ 65), as well as a false loop pair: pair 2 (561↔ 547). Here, pair 2 consists of two images that are spatially close to each other but were incorrectly recognized by the model as a true loop.

These false positives are mainly attributed to two limitations of DA3-Streaming. First, its fixed-interval temporal filtering strategy cannot adequately suppress highly redundant neighboring frames generated during low-speed or stationary motion. Second, it lacks a subsequent geometric verification stage capable of rejecting physically inconsistent loop closure candidates. Although these false loop closures introduce little distortion to the overall trajectory because the associated frames are spatially close, they nevertheless trigger unnecessary loop closure reasoning and pose graph optimization, resulting in wasted computational resources.

In contrast, the proposed method completely eliminated these false-positive detections through the proposed geometric verification module. Consequently, it maintained 100% precision on all KITTI sequences while simultaneously improving the computational efficiency of the loop closure pipeline.

The quantitative improvements are further illustrated by the qualitative trajectory visualizations shown in Figures 4 and 5. In Figure 4, panels (a) and (b) present the reconstructed 3D point clouds of DA3-Streaming and the proposed method, respectively, at three successive stages (before loop closing, after LM optimization, and after loop closing), showing how the global geometric consistency of the reconstruction is progressively restored once loop closures are applied, whereas panels (c) and (d) show the corresponding estimated trajectories together with the detected loop closures. In these figures, valid loop closures are represented by star-shaped markers, whereas false loop closures are indicated by circular markers. For DA3-Streaming, a false loop closure is clearly observed in the deceleration region of Sequences 00 and 05. In contrast, the proposed method successfully removes these erroneous detections, leaving only the genuine loop closure pairs.

As shown by the qualitative results in Table 3, both DA3-Streaming and the proposed pipeline successfully detected all genuine loop closures in KITTI Sequences 00, 02, and 05, thereby achieving 100% recall. However, as discussed previously, the false-positive

loop closures produced by DA3-Streaming during vehicle stops and deceleration unnecessarily increase computational overhead despite having only a minor influence on the final trajectory. By incorporating geometric verification, the proposed method completely suppresses these redundant loop closures. Furthermore, the quantitative trajectory evaluation presented in Table 4 demonstrates that the proposed approach achieves trajectory estimation accuracy comparable to DA3-Streaming while consistently outperforming both VGGT-Long and Pi-Long.

The most significant performance difference was observed on the self-collected outdoor dataset. Unlike KITTI, this dataset contains only a single loop closure event, while the revisited location is captured from substantially different viewpoints. Such severe viewpoint variation significantly reduces the cosine similarity between descriptors of genuine loop closure pairs.

Under this challenging condition, DA3-Streaming completely failed to detect the true loop closure, resulting in 0% precision and 0% recall, whereas the proposed method achieved 100% precision and 100% recall. Although the similarity threshold was relaxed and the true loop pair still exceeded the threshold, DA3-Streaming failed because of the inherent limitation of its retrieval mechanism. Specifically, DA3-Streaming combines the FAISS Top-(K) nearest-neighbor retrieval strategy with a fixed temporal exclusion window. Consequently, several temporally adjacent but non-loop frames remain within the candidate list due to their high visual similarity, occupying the limited Top-(K) retrieval slots and preventing the genuine loop closure from being selected.

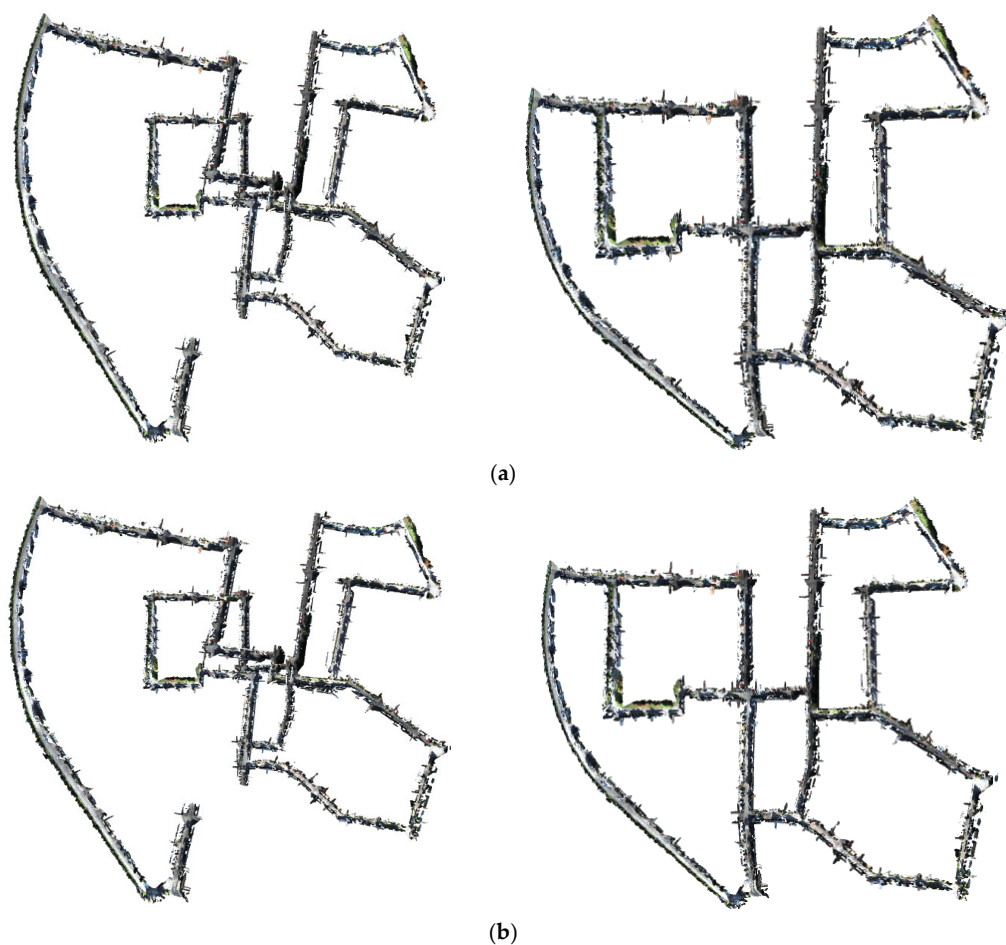


Figure 4. Cont.

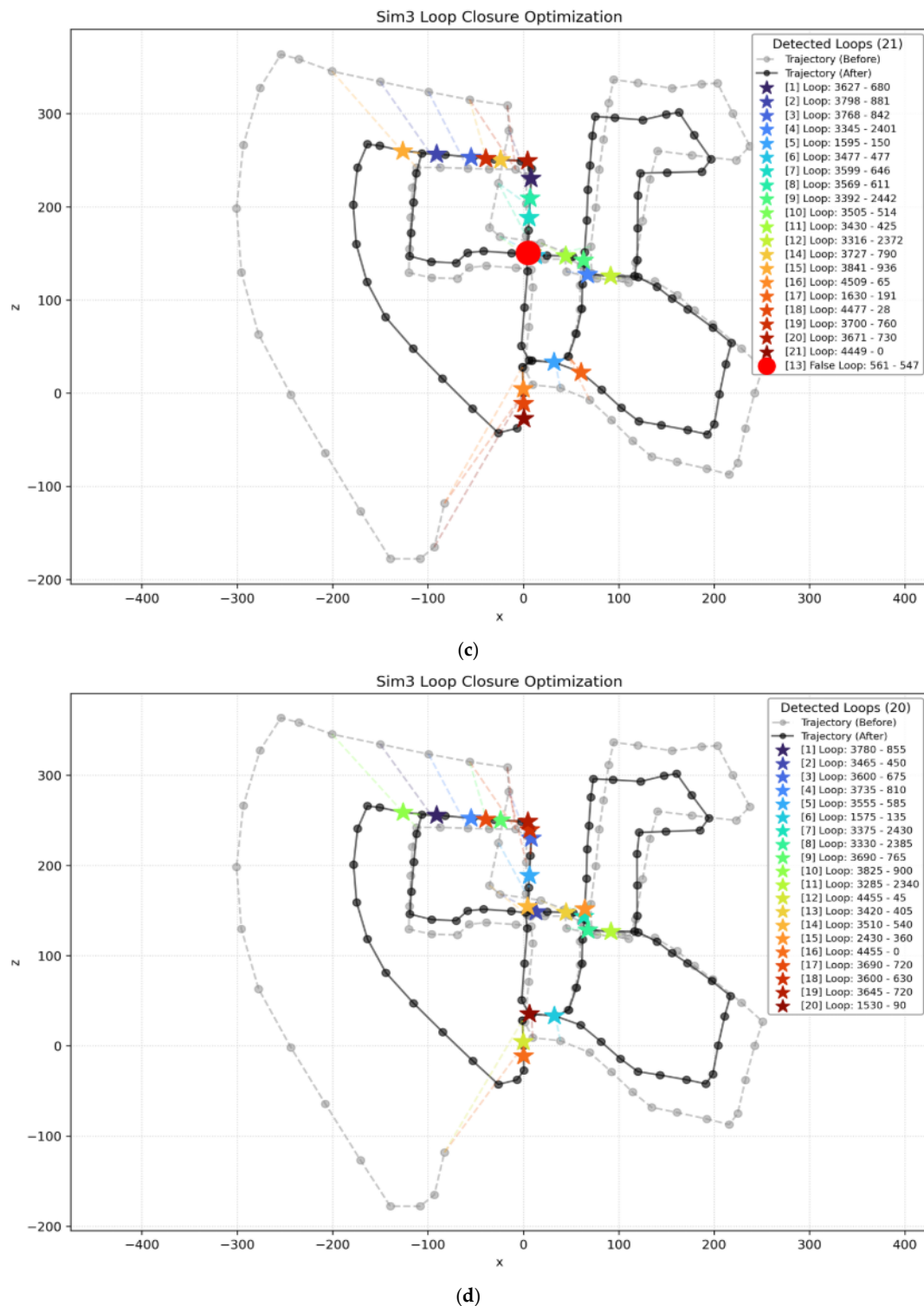


Figure 4. Loop closure detection on the KITTI sequence 00 dataset: reconstructed 3D point clouds of (a) DA3-Streaming and (b) the proposed method (columns, left to right: before loop closing, after LM optimization, after loop closing); estimated trajectories with detected loops for (c) DA3-Streaming and (d) the proposed method, where “Before”/“After” denote the trajectory before/after Sim(3) loop closure optimization. Coordinates are in the relative (up-to-scale) units of the DA3-Streaming reconstruction, not metric scale. Star and circle markers denote true and false loop closures; markers of the same color mark the two frames of one loop pair.

Table 3. Performance comparison between the DA3-Streaming method and our proposed pipeline on KITTI sequences 00, 02, and 05, based on APE, RRA, and RTA metrics (results reproduced on our local machine).

Model (Chunk Size = 60)	Seq.00			Seq.02			Seq.05		
	APE RMSE (m) ↓	RRA Mean (degree) ↓	RTA Mean (m) ↓	APE RMSE (m) ↓	RRA Mean (degree) ↓	RTA Mean (m) ↓	APE RMSE (m) ↓	RRA Mean (degree) ↓	RTA Mean (m) ↓
VGGT-Long	8.34	0.16	0.09	53.63	0.19	0.24	9.96	0.15	0.11
Pi-Long	9.73	0.15	0.11	42.72	0.14	0.18	5.62	0.13	0.08
DA3-Streaming	5.46	0.15	0.06	40.32	0.20	0.14	3.04	0.14	0.05
Ours pipeline	5.46	0.15	0.06	40.32	0.18	0.14	3.04	0.14	0.05

Table 4. Performance comparison between the DA3-Streaming method and Ours pipeline based on APE, RRA, and RTA metrics on our data.

Model Chunk Size = 60	APE RMSE (m) ↓	RRA Mean (degree) ↓	RTA Mean (m) ↓
VGGT-Long	2.71	2.69	0.04
Pi-Long	2.30	2.72	0.03
DA3-Streaming	2.39	2.71	0.03
Ours pipeline	1.53	2.72	0.02

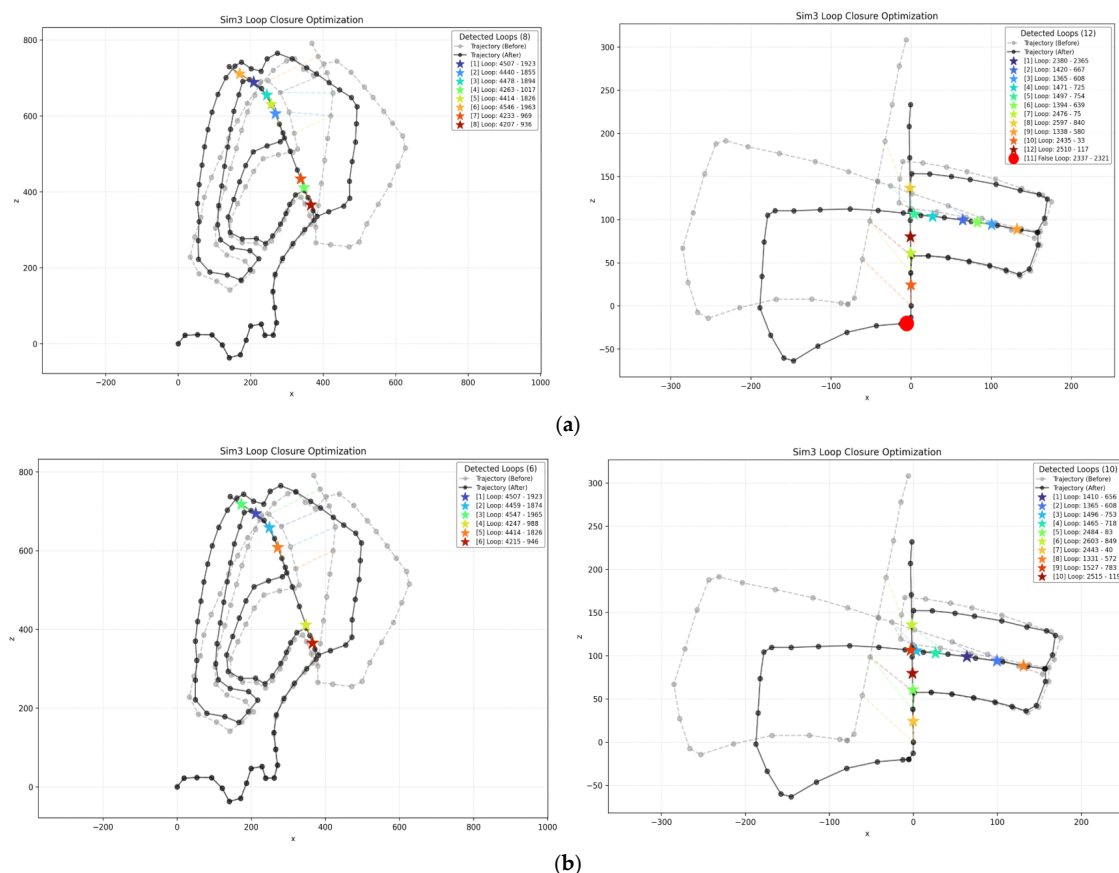


Figure 5. Detected loops and trajectories after loop optimization on KITTI sequences 00, 02, and 05 (see Figure 3 for the sequence 00 detected loops): (a) DA3-Streaming and (b) the proposed method, where “Before”/“After” denote the trajectory before/after Sim(3) loop closure optimization. Coordinates are in the relative (up-to-scale) units of the DA3-Streaming reconstruction, not metric scale. Star and circle markers denote true and false loop closures; markers of the same color mark the two frames of one loop pair.

Because the self-collected dataset contains only a single loop closure, this retrieval failure causes both precision and recall to drop to 0%. The qualitative comparison shown in Figure 6 clearly demonstrates this limitation. DA3-Streaming fails to detect the true loop closure and instead produces five false-positive loop closures (circular markers). In contrast, the proposed method successfully suppresses all false-positive detections and correctly identifies the only genuine loop closure (star-shaped marker). More importantly, the successful loop closure enables simultaneous correction of both the estimated camera poses and the reconstructed 3D point cloud, thereby restoring the global geometric consistency of the entire SLAM system.

The quantitative trajectory evaluation and the predicted trajectory visualization shown in Table 4 and Figure 7 further confirm that the proposed method achieves the best overall performance among all compared approaches. Successful loop closure detection substantially reduces the accumulated trajectory error. Specifically, the Absolute Pose Error (APE) is reduced from 2.39 m to 1.53 m, corresponding to an error reduction of approximately 36%. Similarly, the Relative Translational Error (RTA) decreases from 0.03 m to 0.02 m, demonstrating that accumulated drift is effectively compensated. In contrast, the Relative Rotational Error (RRA) increases only marginally from 2.71° for DA3-Streaming to 2.72° for the proposed method. This negligible increase is attributed to the pose graph optimization process, in which the camera orientations are slightly adjusted to achieve globally consistent trajectory correction after successful loop closure.

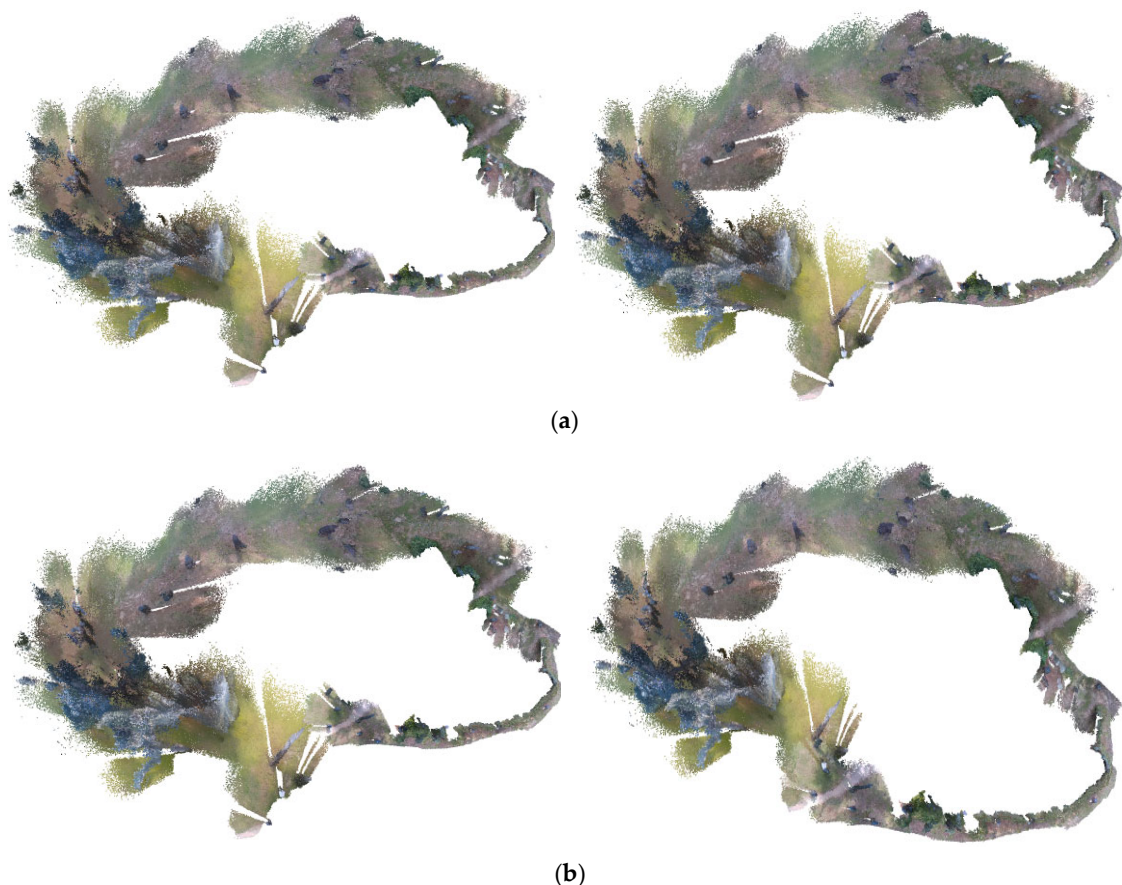


Figure 6. Cont.

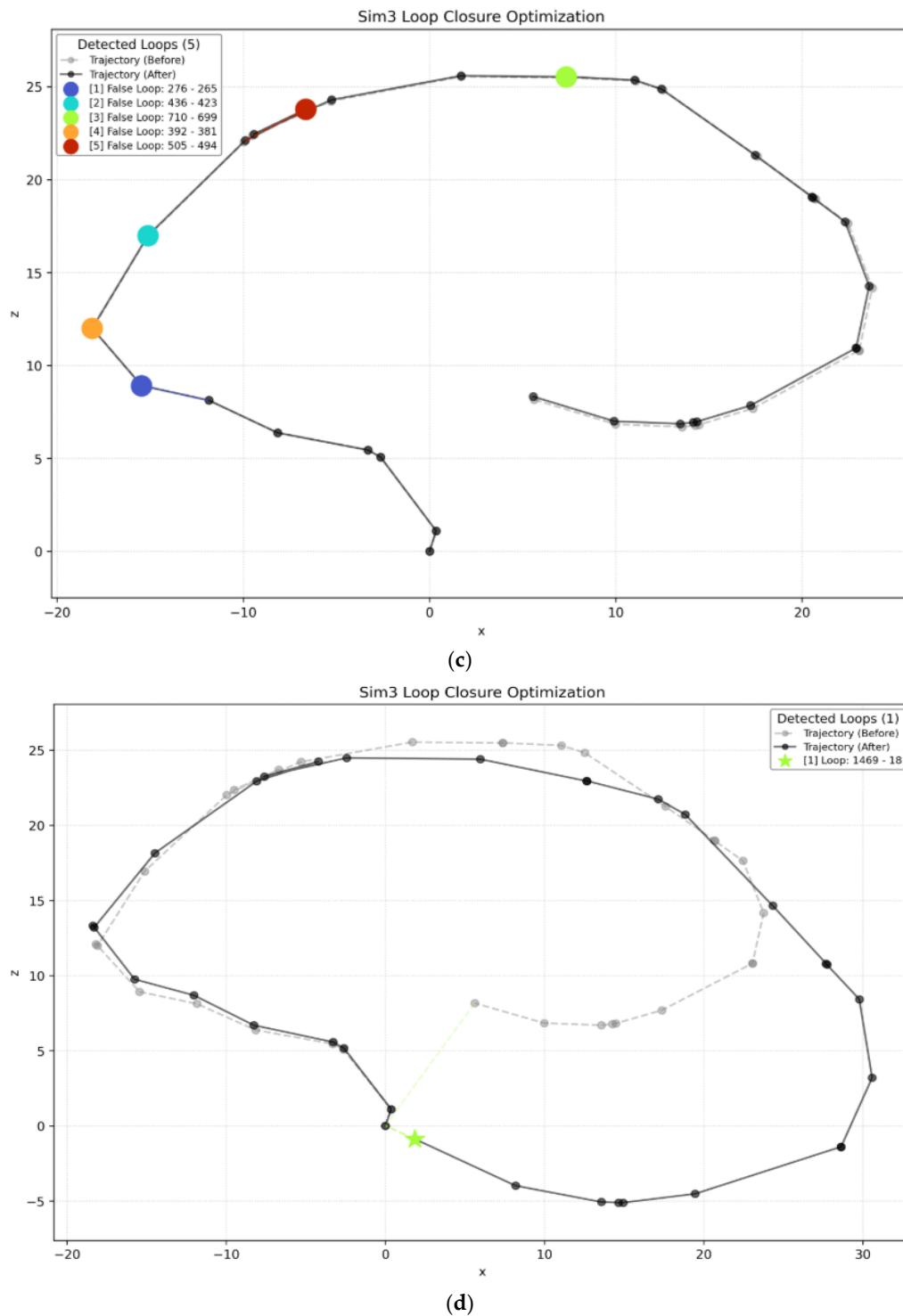


Figure 6. Loop closure detection on private data: reconstructed 3D point clouds of (a) DA3-Streaming and (b) the proposed method (columns, left to right: before loop closing, after LM optimization, after loop closing); estimated trajectories with detected loops for (c) DA3-Streaming and (d) the proposed method, where “Before”/“After” denote the trajectory before/after Sim(3) loop closure optimization. Coordinates are in the relative (up-to-scale) units of the DA3-Streaming reconstruction, not metric scale. Star and circle markers denote true and false loop closures; markers of the same color mark the two frames of one loop pair.

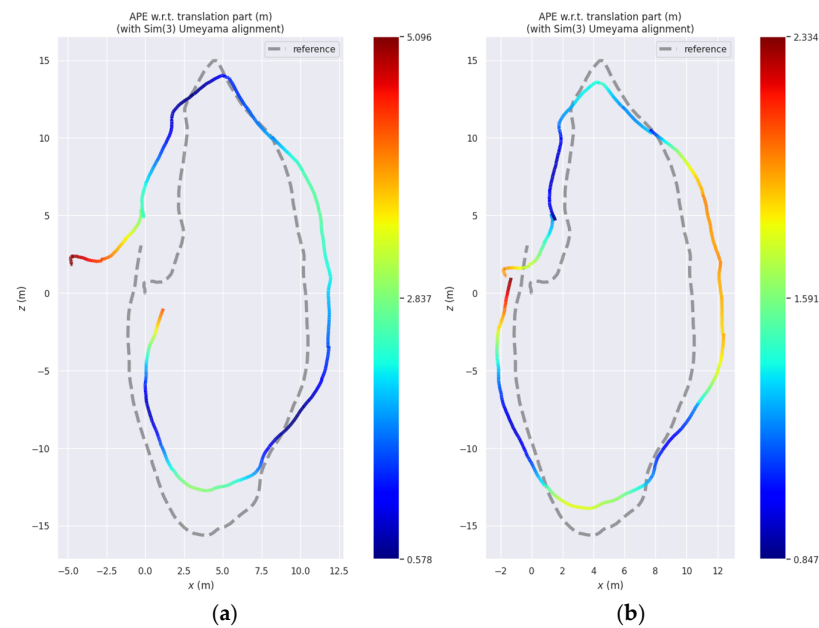


Figure 7. Trajectory comparison of DA3-Streaming and ours pipeline against the ground truth on ours data (a) DA3-Streaming (b) ours.

4.3. Runtime Analysis

To assess the real-time capability of the proposed method, we measured its runtime, inference latency, and memory usage on the full KITTI sequence 00 (4541 frames), using a chunk size of 60 with an overlap of 30 on a single RTX 4090 GPU; the resulting performance metrics are summarized in Table 5. The streaming front-end, consisting of chunk inference and inter-chunk alignment, constitutes the online cost of the pipeline. Although its effective end-to-end throughput is 7.19 FPS owing to the 50% chunk overlap, the raw per-frame inference runs at 15.67 FPS (a latency of 63.8 ms per frame), indicating that the overall pipeline sustains near-real-time streaming operation. The proposed loop closure module runs as a single global operation over the whole sequence and adds 127.2 s in total, corresponding to an amortized overhead of only 28 ms per frame. The peak GPU memory usage is 13.04 GB, well within the 24 GB budget of the RTX 4090. Point-cloud export I/O and one-time model loading are excluded, as they are not part of the online operation.

Table 5. Real-time performance summary on KITTI sequence 00 (single RTX 4090).

Metric	Value
Per-frame inference latency	63.8 ms (15.67 FPS)
Effective streaming throughput (inference + alignment)	7.19 FPS
Loop-closure runtime (one-time, full sequence)	127.2 s
Amortized loop-closure overhead	28 ms per frame
Peak GPU memory	13.04 GB

5. Conclusions

This paper proposed a novel Visual Loop Closure Detection (VLCD) pipeline that is robust to both viewpoint variation and variable-speed vehicle motion. Experimental results demonstrated that the proposed framework effectively overcomes several structural limitations of existing state-of-the-art methods. First, by adopting the DINOv3-SALAD backbone, the proposed method significantly improves robustness to viewpoint variation. Compared with the conventional DINOv2-based descriptor, DINOv3-SALAD maintains

consistently higher cosine similarity under challenging viewpoint changes, enabling more discriminative and reliable feature representations for loop closure detection. Second, the proposed Adaptive Temporal Buffering (ATB) mechanism replaces the conventional fixed-interval exclusion strategy, allowing the retrieval process to adapt dynamically to changes in vehicle speed. As a result, false-positive detections caused by temporally adjacent frames are effectively suppressed. Furthermore, the integration of one-dimensional Non-Maximum Suppression (1D NMS) prevents genuine loop closures from being displaced by redundant neighboring candidates in the retrieval list, thereby maximizing loop closure recall. Third, a rigorous Geometric Verification module completely eliminates false loop closures before pose graph optimization, resulting in 100% precision across all evaluated datasets. This improvement further enhances the global consistency of the estimated trajectory. On the self-collected dataset, the proposed framework reduced the Absolute Pose Error (APE) by approximately 36% compared with DA3-Streaming. In conclusion, the proposed pipeline provides reliable and robust loop closure detection under challenging outdoor environments characterized by severe viewpoint variation and variable vehicle dynamics. The proposed framework is expected to serve as an effective solution for improving long-term localization accuracy and global map consistency in future intelligent mobile systems and autonomous driving SLAM applications.

For future work, we plan to further improve the robustness and accuracy of the proposed framework by integrating multi-sensor information, including LiDAR and IMU measurements, with visual features. Such sensor fusion is expected to provide reliable loop closure detection under challenging conditions, such as low-light, foggy, and textureless environments. In addition, we will incorporate local Bundle Adjustment [18] (BA) within each streaming chunk to jointly refine camera poses and 3D point clouds, thereby reducing drift accumulation and improving the geometric consistency of long-term SLAM.

Author Contributions: Conceptualization, D.T.D. and J.-E.H.; methodology, D.T.D. and J.-E.H.; software, D.T.D.; validation, D.T.D.; formal analysis, D.T.D.; investigation, D.T.D.; resources, D.T.D.; data curation, D.T.D.; writing—original draft preparation, D.T.D.; writing—review and editing, J.-E.H.; visualization, D.T.D.; supervision, J.-E.H.; project administration, J.-E.H.; funding acquisition, J.-E.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) through Korean Government(MSIT) under Grant 2023R1A2C1005870. This research was supported by the ANCHOR through the Seoul ANCHOR Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government (2026-ANCHOR-01-014-02).

Data Availability Statement: The authors do not have permission to share data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Campos, C.; Elvira, R.; Rodríguez, J.J.G.; Montiel, J.M.M.; Tardós, J.D. ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multi-map SLAM. *IEEE Trans. Robot.* **2021**, *37*, 1874–1890. [CrossRef]
2. Qin, T.; Li, P.; Shen, S. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robot.* **2018**, *34*, 1004–1020. [CrossRef]
3. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
4. Bay, H.; Tuytelaars, T.; Van Gool, L. SURF: Speeded up robust features. In Proceedings of the European Conference on Computer Vision (ECCV), Graz, Austria, 7–13 May 2006; pp. 404–417.
5. Liu, J. Image retrieval based on Bag-of-Words model. *arXiv* **2013**, arXiv:1304.5168.
6. Deng, K.; Ti, Z.; Xu, J.; Yang, J.; Xie, J. VGGT-Long: Chunk it, loop it, align it—Pushing VGGT’s limits on kilometer-scale long RGB sequences. *arXiv* **2025**, arXiv:2507.16443.
7. Deng, K. DA3-Streaming: Memory-Efficient Inference for Videos via Chunk Streaming. Available online: https://github.com/ByteDance-Seed/Depth-Anything-3/tree/main/da3_streaming (accessed on 24 July 2026).

8. Izquierdo, S.; Civera, J. Optimal transport aggregation for visual place recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 17658–17668.
9. Schönberger, J.L.; Frahm, J.M. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4104–4113.
10. Johnson, J.; Douze, M.; Jégou, H. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data* **2021**, *7*, 535–547. [[CrossRef](#)]
11. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
12. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193.
13. Siméoni, O.; Vo, H.V.; Seitzer, M.; Baldassarre, F.; Oquab, M.; Jose, C.; Khalidov, V.; Szafraniec, M.; Yi, S.; Ramamonjisoa, M.; et al. DINOv3. *arXiv* **2025**, arXiv:2508.10104.
14. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Lake Tahoe, NV, USA, 5–10 December 2013; pp. 2292–2300.
15. Lin, H.; Chen, S.; Liew, J.; Chen, D.Y.; Li, Z.; Shi, G.; Feng, J.; Kang, B. Depth Anything 3: Recovering the visual space from any views. *arXiv* **2025**, arXiv:2511.10647.
16. Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Rupprecht, C.; Novotny, D. VGGT: Visual geometry grounded transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 5294–5306.
17. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
18. Triggs, B.; McLauchlan, P.F.; Hartley, R.I.; Fitzgibbon, A.W. Bundle adjustment—A modern synthesis. In Proceedings of the International Workshop on Vision Algorithms, Corfu, Greece, 21–22 September 1999; pp. 298–372.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.