Robust Cross-Modal Matching between LiDAR Point Clouds and Multi-Camera Images in Tunnel Environments via Surface Parameterization

Ying Jiang[1], Feng Liu[2,*],Han Hu[1,3], Yulin Ding[1,3], Chong Wang[3,4], Ping Wen[3,4], Qing Zhu[1,3]

[1] Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu 611756, China
[2] CRSC Communication & Information Group Co., Ltd., Beijing 100080, China - 13910533237@139.com
[3] Yunnan Engineering Research Center of 3D Real Scene, Kunming 650500, China
[4] Kunming Engineering Corporation Limited, Kunming 650500, China

Keywords: Cross-modal matching, Surface parameterization, Multi-camera stitching, Low-texture environment, Data fusion.

Abstract

Tunnel lining defect detection increasingly depends on LiDAR and multi-camera collaboration for high-precision 3D modeling. However, sparse textures, repetitive structures, and insufficient lighting limit camera overlap, compromising target-free cross-modal matching. We propose leveraging inherent tunnel structural regularity for robust matching by projecting LiDAR intensity data and stitched images onto a shared parameterized surface. Leveraging geometric priors, LiDAR point clouds are fitted to cross-sections and unrolled into a 2D intensity map within a longitudinal-angular domain. Simultaneously, eight images undergo photometric normalization and stitching into a panoramic image sampled in the same domain. Feature extraction and matching occur in this unified 2D space, with correspondences mapped back to 3D-2D relationships via invertible bi-directional mapping. This approach facilitates geometric constraints and consistency checks to filter unreliable matches. Experiments on the Daliang and Zhaobishan railway tunnels demonstrate that our method generates significantly more geometrically verified correspondences compared to direct image matching. Matching stability is enhanced, with high alignment precision at structural edges, effectively supporting subsequent fusion and analysis tasks.

1. Introduction

Railway tunnel inspection increasingly depends on integrating dense 3D geometry with high-resolution visual texture for lining defect detection, structural assessment, and asset documentation (Hawley and Gräbe, 2022). Single-modality sensing is often insufficient: LiDAR provides accurate metric geometry and intensity cues but limited semantic detail, whereas cameras capture rich appearance without depth. Consequently, mobile mapping systems commonly combine co-located LiDAR and multi-camera rigs to obtain complementary observations (Xiong et al., 2023). The effectiveness of such fusion ultimately hinges on reliable cross-modal registration, which typically requires robust 3D-2D homonymous correspondences to support accurate LiDAR-camera extrinsic calibration. This requirement is critical in tunnels, where geometric regularity and texture sparsity make matching difficult and robust registration becomes a prerequisite for downstream analysis (Xiong et al., 2025).

Existing cross-modal registration methods mainly fall into target-based and targetless paradigms. Target-based methods use artificial markers (e.g., checkerboards or spheres) to impose strong geometric constraints and can reach high accuracy in controlled settings, but they rely on field deployability (Zahs et al., 2023). Targetless methods avoid external aids by exploiting scene cues. These include keypoint matching such as SIFT (Lowe, 2004) and ORB (Rublee et al., 2011), as well as geometric primitives. Additionally, appearance-driven alignment optimizes objectives like mutual information or gradient consistency between images and LiDAR renderings (Sun et al., 2026). Recently, learning-based pipelines and panoramic representations have been adopted to enhance descriptor robustness and reduce viewpoint discrepancies in enclosed scenes (Lei et al., 2023). However, most approaches assume distinctive features and stable radiometry, and thus remain fragile in tunnels with repetitive structures and severe photometric inconsistencies (Yan et al., 2023).

Figure 1. Harsh tunnel environment (low light, low texture, dynamic flash) and the integrated LiDAR–camera mapping system mounted on a rail-based platform.

Tunnel environments, however, pose systematic challenges to these paradigms. First, tunnel linings are often texture-poor and exhibit repetitive patterns, making local features weak and ambiguous. Second, adverse illumination and auxiliary lighting introduce severe photometric inconsistencies, vignetting, and local over-/under-exposure, causing the same structure to appear substantially different across modalities and viewpoints (Su et al., 2025). Third, due to cost and installation constraints, many tunnel multi-camera systems provide only limited over-

lap between adjacent views, which reduces structural context and further degrades feature discriminability (Figure 1). More importantly, engineering deployments commonly adopt modular assembly: large devices are frequently disassembled and reassembled, leading to non-negligible extrinsic drift across acquisition sessions; meanwhile, deploying calibration targets is often impractical because of traffic control, safety, and access restrictions (Wang et al., 2025). As a result, establishing stable and reliable cross-modal homonymous points under low texture, limited overlap, strong photometric variations, and targetless operation remains an open practical problem (Shim et al., 2023).

This paper addresses the following question: how to reliably mine cross-modal homonymous points between LiDAR scans and multi-camera images in tunnels without calibration targets, under limited inter-camera overlap and significant photometric variations, and how to evaluate correspondence quality and spatial distribution in a reproducible manner. Our core idea is to exploit the strong geometric prior of tunnel interiors by introducing a shared parameterized surface that approximates the lining. Specifically, we map both a LiDAR intensity unwrapping and a stitched multi-camera ring image into a common longitudinal–angular parameter domain, transforming cross-modal matching into a unified 2D–2D problem. The mapping is designed to be invertible, so that matched keypoints in the parameter domain can be mapped back bidirectionally to LiDAR 3D points and image pixels, yielding 3D–2D correspondences suitable for downstream self-calibration and fusion. Beyond providing a shared representation, the parameterization also enables geometric constraints and bidirectional consistency checks for reliability filtering (Ma et al., 2025).

The main contributions of this paper are summarized as follows: A reliable homonymous point mining and filtering method tailored for low-texture tunnels is proposed. Specifically: First, a prior-guided dual-projection framework is designed to project LiDAR intensity and stitched multi-camera imagery into a shared parameter domain, establishing an invertible 3D–2D mapping interface. Second, a bidirectional consistency checking mechanism is introduced to effectively reject unreliable matches. Finally, comparative experiments on real railway tunnel datasets demonstrate significant improvements in correspondence quantity, accuracy, and coverage. The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 presents the proposed method, Section 4 reports experiments, and Section 5 concludes.

2. Related work

This section reviews existing literature on cross-modal correspondence extraction between LiDAR and cameras, categorized into marker-assisted and targetless paradigms.

2.1 Marker-assisted Correspondence Extraction

Conventional methods for cross-modal alignment rely on the strategic integration of auxiliary structures—such as checkerboards, retro-reflective spheres, or high-contrast coded targets—to establish mathematically rigorous 3D–2D correspondences. These engineered designs offer superior observability and geometric repeatability, enabling high-confidence matching and precision calibration in stable, controlled environments (Jiang et al., 2023). By providing distinct, easily identifiable features across both LiDAR and image modalities, marker-assisted approaches effectively bypass the ambiguities inherent in raw sensor data. However, the practical application of these methods in operational tunnel environments is severely hindered by stringent traffic safety regulations and the extremely narrow maintenance windows typical of modern infrastructure (Li et al., 2025). The logistical burden of deploying, measuring, and retrieving physical markers across several kilometers of repetitive tunnel sections is often prohibitive. Furthermore, the requirement for frequent, high-speed inspections makes repeated marker placement highly inefficient, while persistent challenges such as dynamic occlusions and poor localized lighting further degrade the reliability of physical targets. These cumulative constraints necessitate the development of robust, targetless alternatives capable of exploiting the tunnel’s inherent structural regularity and geometric priors for autonomous, large-scale calibration (Lin et al., 2024).

2.2 Targetless Correspondence Extraction

Targetless approaches aim to establish homonymous points by exploiting geometric, appearance, or feature-based cues shared across modalities. Geometry and appearance-driven techniques focus on structural primitives such as edges, planes, and surface normals, or convert LiDAR data into image-like representations to optimize similarity measures like mutual information (Huang et al., 2024). However, in tunnel environments, repetitive lining patterns and longitudinal seams create significant geometric ambiguity, while non-uniform auxiliary lighting induces strong photometric distortions and specular highlights that violate the stability assumptions of statistical optimization (Zhu et al., 2023). Similarly, feature-based workflows, which utilize either handcrafted descriptors or learning-based matchers to establish correspondences, frequently fail in these low-texture scenarios (Li et al., 2024). The high density of repetitive structures leads to severe mismatches, and the limited field of view of individual cameras often results in insufficient structural context. Consequently, initial matches in tunnel scenes often contain a high percentage of outliers, failing to provide the reliable spatial coverage required for robust calibration (Cheng et al., 2025).

Despite their flexibility, these methods frequently struggle in tunnel environments. Repetitive lining patterns and longitudinal seams create significant geometric ambiguity, while non-uniform auxiliary lighting leads to strong photometric distortions and unstable descriptors (Ma et al., 2025). Consequently, existing targetless frameworks often fail to produce sufficient reliable correspondences in low-texture, repetitive tunnel sections. Our work addresses this gap by introducing a shared parameterized surface that leverages the tunnel’s structural regularity to stabilize cross-modal matching.

3. Methodology

3.1 Overview

To address low texture, strong photometric variations, and repetitive structures in tunnels, we propose a cross-modal homonymous point mining framework based on a shared surface parameterization. As shown in Figure 2, the inputs include LiDAR scans with intensity, eight synchronized tunnel images, POS trajectory, and the intrinsic parameters and relative poses of the multi-camera rig. Cross-modal homonymous point extraction in tunnels is formulated as a coarse-to-fine matching

pipeline. First, a global timestamp-to-mileage gating is used to restrict the valid longitudinal coverage of each station, thereby suppressing ambiguities caused by repetitive tunnel patterns. Second, both LiDAR intensity and multi-camera imagery are embedded into a shared cylindrical surface grid, yielding two aligned 2D representations. Third, local keypoints and correspondences are established on the shared grid using SuperPoint and SuperGlue. Finally, reliability filtering and geometric pruning are performed to remove mismatches and retain stable 3D–2D homonymous points.

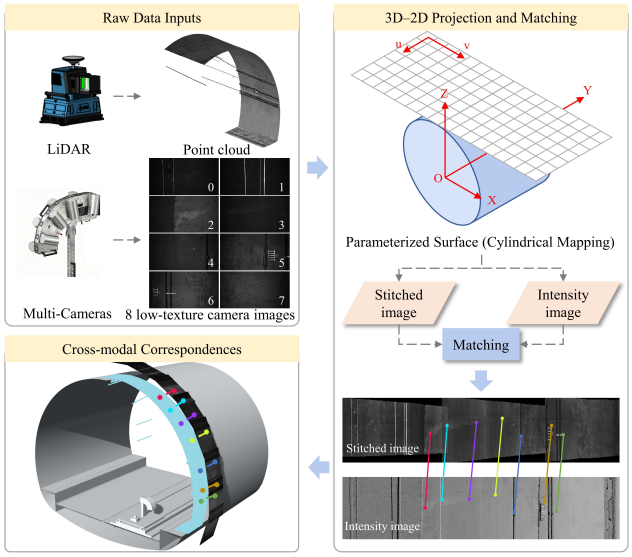


Figure 2. Overview of the proposed cross-modal matching framework in a shared cylindrical domain.

3.2 Global Spatio-Temporal Gating

One of the most challenging issues in tunnel environments is the highly repetitive geometry and texture: cables, lining joints, and grooves often recur with an approximately fixed period. As a consequence, purely local-feature-based matching can easily produce long-range visual confusion, generating a large number of seemingly plausible but spatially incorrect candidate correspondences. This effect becomes particularly severe in long and monotonous underground corridors; without effective suppression, the subsequent fine matching stage faces both a high outlier risk and unnecessary computational overhead. To mitigate this problem, a global spatio-temporal prior provided by the acquisition system is exploited for coarse localization and search-space restriction (Figure 3).

For each acquisition station (frame) f, the synchronized timestamp t_f is recorded and used to query the POS trajectory, yielding the cumulative mileage along the tunnel axis:

$$s_f = s(t_f). \tag{1}$$

Accordingly, all subsequent matching operations are restricted to a local longitudinal window centered at s_f:

$$\mathcal{W}_f = [s_f - \Delta s,\ s_f + \Delta s], \tag{2}$$

where the half-width Δs is determined by the physical coverage of a single station. The corresponding longitudinal coverage is $L = 2\Delta s$, which is typically on the order of 0.8 m in

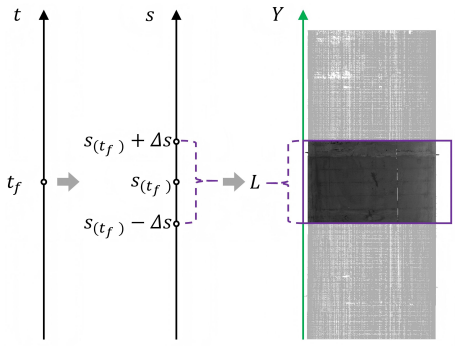


Figure 3. Global spatio-temporal gating for point-cloud cropping.

the proposed system. In implementation, the mileage window $\mathcal{W}_f$ is mapped to an equivalent axis-aligned cropping interval in the world frame. Since the platform motion is aligned with the tunnel axis, this interval is applied along the world coordinate Y direction to crop the LiDAR point cloud, and only points within the valid matching region are retained. All points outside the window are discarded before surface projection and feature matching.

This gating strategy provides two practical benefits. First, long-range mismatches induced by repetitive tunnel patterns are effectively suppressed: even if visually similar structures exist elsewhere, candidates with meter- to tens-of-meter offsets are prevented from entering the correspondence pool, eliminating “looks similar but far away” confusions at the source. Second, computational cost is reduced by shrinking the effective processing region from a long tunnel segment to a station-level window, thereby decreasing the number of valid points and the descriptor search space. Notably, this constraint does not rely on visual distinctiveness; instead, it leverages an external trajectory prior and remains robust under severe illumination changes or partial occlusions, as long as the POS drift stays within a bounded range. The cropped point cloud segment is then used for cylindrical unwrapping and ring-image construction.

3.3 Shared Surface Parameterization and Dual Projection

3.3.1 Cylindrical surface discretization As illustrated in Figure 4, the tunnel lining is approximated by a virtual cylinder with design radius R_0 and discretized into a 2D (u, v) grid. This rasterization establishes a one-to-one relationship between each grid cell and a 3D surface point, so that both LiDAR intensity and multi-camera imagery can be accumulated on the same grid for cross-modal matching. The circumferential extent is set to a full circumference and the longitudinal extent is set to a local window along the tunnel axis. Specifically, the circumferential length is $L_u = 2\pi R_0$, and the longitudinal window length is set to $L_v = 2\Delta s$ as defined by the mileage window $\mathcal{W}_f$ (typically ≈ 0.8 m per station). With circumferential and longitudinal resolutions (Δ_u, Δ_v) (in meters), the grid size is

$$N_u = \left\lceil \frac{L_u}{\Delta_u} \right\rceil, N_v = \left\lceil \frac{L_v}{\Delta_v} \right\rceil. \tag{3}$$

Let (u, v) denote integer grid indices with $u \in \{0, \ldots, N_u - 1\}$ and $v \in \{0, \ldots, N_v - 1\}$. Each cell is associated with its

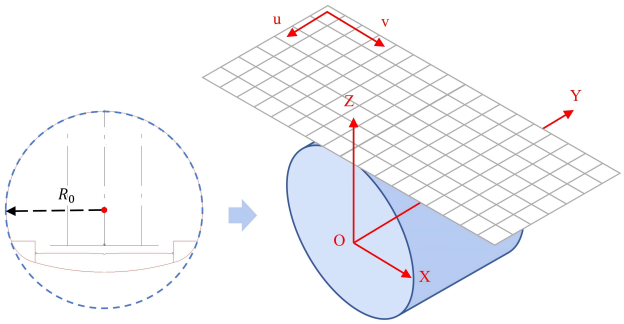


Figure 4. Surface parameterization and discretization. The tunnel profile is unwrapped from a virtual cylinder (radius R_0) into a 2D (u, v) grid representing circumferential and longitudinal directions.

physical coordinates on the cylinder by

$$\begin{aligned} \ell_u(u) &= \left(u + \tfrac{1}{2}\right) \Delta_u, \\ s_v(v) &= s_{\min} + \left(v + \tfrac{1}{2}\right) \Delta_v, \end{aligned} \tag{4}$$

$$\begin{aligned} &\varphi(u) = \frac{\ell_u(u)}{R_0} \in [0, 2\pi), \\ &\mathbf{X}_{u,v} = \mathbf{c}(s_v) + R_0 \Big(\cos\varphi(u)\, \mathbf{n}(s_v) + \sin\varphi(u)\, \mathbf{b}(s_v) \Big). \end{aligned} \tag{5}$$

where $\mathbf{c}(s)$ denotes the tunnel centerline and $\{\mathbf{n}(s), \mathbf{b}(s)\}$ span the cross-section plane orthogonal to the axis direction. Within a local window (~ 0.8 m), $\mathbf{c}(s)$ and the local frame are approximated as constant and obtained from the POS trajectory.

3.3.2 LiDAR intensity unwrapping Given a LiDAR scan $\{(\mathbf{X}_k, I_k)\}_{k=1}^{N}$, intensities are rasterized onto the shared cylindrical grid to obtain an unwrapped intensity map $I_L(u, v)$. Let $\mathcal{Q}(\cdot)$ denote the quantization from a 3D point to grid indices:

$$(u_k, v_k) = \mathcal{Q}(\mathbf{X}_k). \tag{6}$$

Grid values are assigned by an aggregation operator

$$I_L(u, v) \leftarrow \mathrm{Agg}\Big(\{ I_k \mid (u_k, v_k) = (u, v) \} \Big), \tag{7}$$

where $\mathrm{Agg}(\cdot)$ can be the average or maximum intensity. In addition, an index map $\Omega_L(u, v)$ is stored to record the contributing 3D point index (or a small set of indices) for inverse lookup, enabling back-mapping from a grid location to a LiDAR 3D point.

3.3.3 Geometry-driven ring image construction Prior to surface projection, each camera image is photometrically normalized to improve feature repeatability under strong illumination variations. Illumination equalization, vignetting correction, and contrast enhancement are applied, followed by optional denoising for low-light frames. The normalized images are then used for geometry-driven warping and blending to form the ring image $I_I(u, v)$. Traditional image stitching relies on sufficient overlap and feature matches between adjacent views, which is unreliable in low-texture tunnels with small overlaps. Instead, a stitched ring image is constructed by projecting each camera image onto the shared cylindrical surface using the known rig geometry.

For camera i, a pixel $\tilde{\mathbf{p}} = [x, y, 1]^\top$ defines a ray direction in the camera frame:

$$\mathbf{d}_i = \frac{\mathbf{K}_i^{-1} \tilde{\mathbf{p}}}{\| \mathbf{K}_i^{-1} \tilde{\mathbf{p}} \|}. \tag{8}$$

Let the camera-to-tunnel transformation be $\mathbf{T}_{\mathcal{T} \leftarrow \mathcal{C}_i} = \begin{bmatrix} \mathbf{R}_i & \mathbf{t}_i \\ \mathbf{0}^\top & 1 \end{bmatrix}$. Here, $\mathbf{a}$ denotes the unit direction of the tunnel axis and $\mathbf{c}_0$ is a reference point on the axis (approximated from POS within the local window). The ray in the tunnel frame is expressed as

$$\begin{aligned} \mathbf{x}(\lambda) &= \mathbf{o}_i + \lambda\, \mathbf{v}_i, \quad \lambda > 0, \\ \mathbf{o}_i &= \mathbf{t}_i, \\ \mathbf{v}_i &= \mathbf{R}_i\, \mathbf{d}_i. \end{aligned} \tag{9}$$

The ray is intersected with the cylinder. Let $\mathbf{m} = \mathbf{o}_i - \mathbf{c}_0$, and define the components orthogonal to the axis:

$$\begin{aligned} \mathbf{m}_\perp &= \mathbf{m} - (\mathbf{a}^\top \mathbf{m})\mathbf{a}, \\ \mathbf{v}_\perp &= \mathbf{v}_i - (\mathbf{a}^\top \mathbf{v}_i)\mathbf{a}. \end{aligned} \tag{10}$$

The intersection is obtained by solving

$$\| \mathbf{m}_\perp + \lambda \mathbf{v}_\perp \|^2 = R_0^2, \tag{11}$$

which yields a quadratic equation in λ. The smallest positive solution $\lambda^\star$ is selected and the surface point is computed as $\mathbf{X}^\star = \mathbf{x}(\lambda^\star)$. Subsequently, $\mathbf{X}^\star$ is mapped to the shared grid (u, v), and its radiometrically normalized intensity is splatted onto the ring image $I_I(u, v)$ with blending. A provenance map $\Omega_I(u, v)$ is stored to record the contributing camera ID and the original pixel coordinate (x, y), enabling inverse mapping from the grid back to the original images.

This dual projection produces two co-registered surface maps, namely the LiDAR intensity map $I_L(u, v)$ and the stitched ring image $I_I(u, v)$, defined on the same (u, v) grid together with inverse lookup tables (Ω_L, Ω_I). This representation enables direct 2D–2D feature matching and subsequent back-mapping to 3D–2D homonymous points.

3.4 Local Feature Matching on the Shared Domain

Tunnel environments typically contain large homogeneous surfaces and repetitive patterns, resulting in weak texture, low contrast, and poor or uneven illumination. Under such conditions, hand-crafted feature detectors (e.g., SIFT and ORB) often yield sparse and unstable keypoints, which in turn leads to unreliable cross-modal correspondences. To improve repeatability, local feature detection and matching are performed directly in the shared 2D domain defined by the projected representations $I_L(u, v)$ (LiDAR intensity map) and $I_I(u, v)$ (stitched ring image). Since both modalities are resampled onto the same cylindrical grid, keypoints can be localized in a consistent coordinate system despite the fundamentally different sensing mechanisms.

A learning-based pipeline, SuperPoint (DeTone et al., 2018) followed by SuperGlue (Sarlin et al., 2020), is adopted for its robustness in low-texture and low-light scenarios. SuperPoint provides repeatable keypoints and discriminative descriptors,

while SuperGlue leverages contextual reasoning through an attentional graph neural network and solves the assignment problem via differentiable optimal transport, which helps handle partial visibility, occlusions, and non-repeatable points.

Concretely, SuperPoint is applied independently to $I_L(u, v)$ and $I_I(u, v)$ to extract keypoints and descriptors. SuperGlue then predicts a set of tentative matches:

$$\mathcal{M} = \{(\mathbf{z}_m, \mathbf{z}'_m, \gamma_m)\}_{m=1}^{M}, \tag{12}$$

where $\mathbf{z}_m = (u_m, v_m)$ and $\mathbf{z}'_m = (u'_m, v'_m)$ denote the grid coordinates of the m-th matched keypoint pair in the shared domain, and $\gamma_m \in [0, 1]$ is the confidence score produced by SuperGlue.

Each tentative match is then back-mapped to the original 3D–2D space using the inverse lookup tables: $\Omega_L(\mathbf{z}_m)$ retrieves the associated LiDAR point $\mathbf{X}_m \in \mathbb{R}^3$, and $\Omega_I(\mathbf{z}'_m)$ retrieves the corresponding pixel coordinate $\mathbf{u}_m$ on the ring image and the original camera pixel. This yields an initial set of candidate homonymous points:

$$\mathcal{H}_0 = \{(\mathbf{X}_m, \mathbf{u}_m, \gamma_m)\}_{m=1}^{M}. \tag{13}$$

Performing learning-based matching on modality-aligned projections rather than raw inputs significantly improves both correspondence density and stability in visually challenging tunnel scenes. The resulting 3D–2D candidates $\mathcal{H}_0$ are subsequently passed to geometric verification and reliability filtering to suppress mismatches induced by repetitive structures and photometric inconsistencies.

3.5 Reliability Filtering and Geometric Pruning

Although SuperPoint+SuperGlue improves correspondence robustness in the shared projected domain, tunnel scenes remain prone to false matches due to highly repetitive structures (e.g., periodic segment joints, cable trays) and low-discriminative textures. Such outliers may still receive high confidence from the matcher because of local descriptor similarity, yet they violate global geometric consistency and can severely degrade subsequent pose estimation. Therefore, a pruning stage is introduced to convert dense tentative correspondences into a sparse but reliable 3D–2D set.

3.5.1 Bidirectional reprojection consistency For each tentative correspondence $(\mathbf{X}_m, \mathbf{u}_m^{\text{ring}}, \gamma_m) \in \mathcal{H}_0$, geometric plausibility is evaluated using a reprojection consistency error. Let $\Pi_{\text{ring}}(\mathbf{X})$ denote the forward projection that maps a 3D LiDAR point to the ring-image pixel through the shared grid and the known sensor geometry. The bidirectional error is defined as

$$e_m^{\text{bi}} = \left\| \Pi_{\text{ring}}(\mathbf{X}_m) - \mathbf{u}_m^{\text{ring}} \right\|_2. \tag{14}$$

A correspondence is retained only if

$$e_m^{\text{bi}} \le \tau_{\text{bi}} \quad \text{and} \quad \gamma_m \ge \tau_\gamma, \tag{15}$$

where τ_{bi} (e.g., 2–4 pixels) accounts for grid discretization, residual calibration errors, and minor acquisition perturbations, and τ_γ removes low-confidence assignments. This check effectively suppresses matches that are locally plausible in the shared domain but inconsistent with the underlying projection geometry.

3.5.2 Uniqueness constraint Repetitive tunnel patterns often induce many-to-one assignments, where multiple keypoints in one modality match to similar locations in the other. To enforce a one-to-one correspondence set, a mutual nearest-neighbor (MNN) constraint (cross-check) is applied in descriptor space: a match $(\mathbf{z}_m, \mathbf{z}'_m)$ is kept only if $\mathbf{z}_m$ is the nearest neighbor of $\mathbf{z}'_m$ and simultaneously $\mathbf{z}'_m$ is the nearest neighbor of $\mathbf{z}_m$ among all candidates. After applying both checks, each keypoint participates in at most one correspondence. The resulting filtered set is denoted by

$$\mathcal{H} = \{(\mathbf{X}_k, \mathbf{u}_k, \gamma_k)\}. \tag{16}$$

Overall, this pruning stage improves the inlier ratio and correspondence coverage on the tunnel surface, providing reliable inputs for downstream robust estimation in repetitive underground environments. As shown in Figure 5, the tentative correspondence set $\mathcal{H}_0$ contains a noticeable number of ambiguous matches induced by repetitive tunnel patterns, which appear plausible locally but are inconsistent geometrically. After applying bidirectional reprojection consistency and the uniqueness constraint, a large portion of such outliers is removed. The remaining set $\mathcal{H}$ becomes sparser yet more reliable, yielding improved geometric consistency and a cleaner correspondence distribution for subsequent robust estimation.

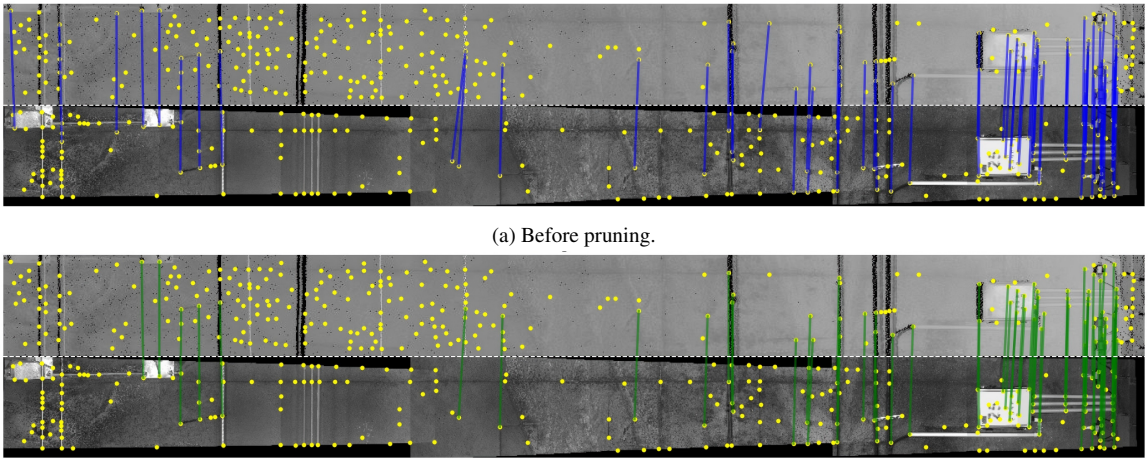
4. Experimental evaluations and analyses

4.1 Datasets Description

Experiments were conducted on two real-world railway tunnel datasets: the Daliang Tunnel and the Zhaobi Mountain Tunnel. Both datasets were collected using a tunnel mapping platform equipped with one LiDAR and an eight-camera rig. At each acquisition station, one LiDAR scan and eight synchronized images were captured under low-texture lining surfaces and uneven illumination, assisted by onboard lighting. The platform advanced along the tunnel axis with an approximately constant spacing (about 0.8 m per station), and odometry measurements were recorded for each station to provide the along-track mileage reference used in global gating.

4.2 Comparison with Classical Feature Matchers

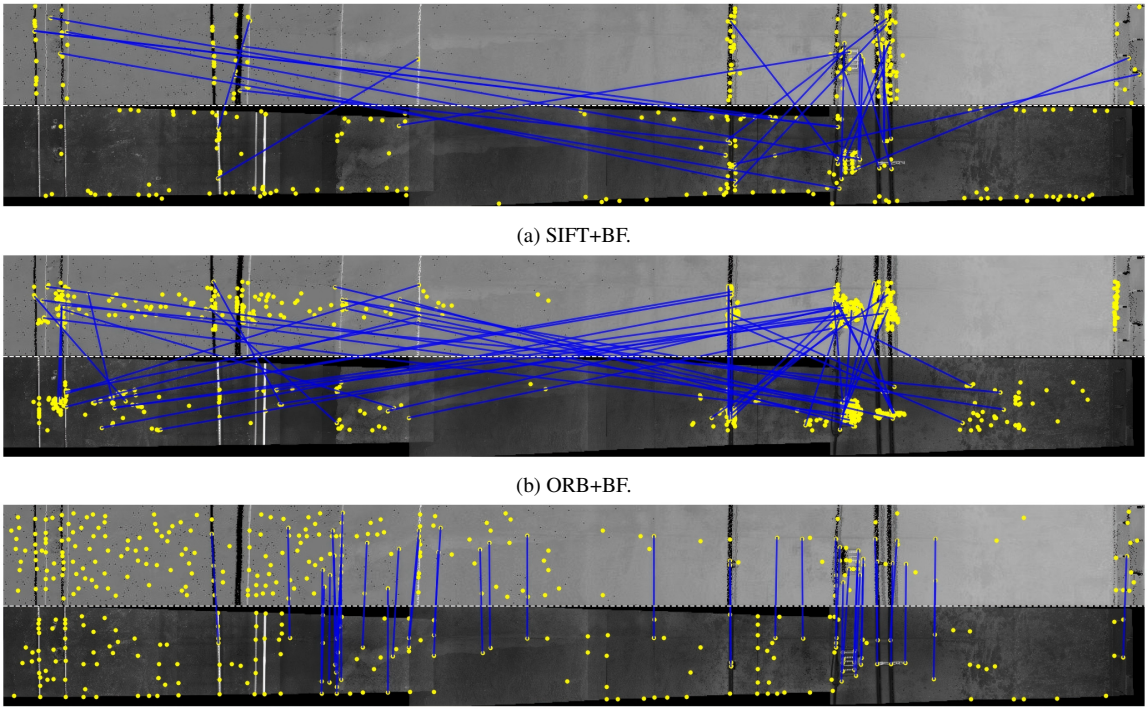
This experiment qualitatively compares conventional hand-crafted feature matchers with a learning-based matcher in the proposed shared projected domain. The objective is to examine whether classical local features can provide stable cross-modal correspondences in low-texture tunnels, and to identify typical failure modes caused by repetitive structures. All matchers operate on the same pair of projected representations, namely the LiDAR intensity map and the stitched ring image, both defined on the shared cylindrical grid. SIFT is matched using a brute-force matcher with L2 distance and mutual cross-check, while ORB is matched using Hamming distance with cross-check. SuperPoint and SuperGlue are applied directly in the same domain. For visualization, detected keypoints are shown as yellow dots and tentative matches are drawn as blue segments.



(a) Before pruning.

(b) After pruning.

Figure 5. Correspondences before and after reliability pruning (top: intensity map; bottom: stitched ring image).



(a) SIFT+BF.

(b) ORB+BF.

(c) SuperPoint+SuperGlue.

Figure 6. Qualitative comparison of cross-modal matching in the shared projected domain.

Figure 6 shows that SIFT+BF and ORB+BF produce a noticeable number of long-range and structurally inconsistent correspondences: many match segments span large distances and frequently connect visually similar but spatially unrelated repetitive patterns (e.g., segment joints and vertical seams). Although these matches appear plausible locally, their global layout indicates strong ambiguity in repetitive tunnel textures, suggesting that descriptor similarity alone is insufficient in this setting. In contrast, SuperPoint+SuperGlue yields correspondences that are more coherent with the shared-domain geometry, exhibiting a more consistent match structure and fewer long-range crossings. Matches tend to concentrate on stable structural cues and remain locally aligned, indicating improved repeatability under weak texture and uneven illumination.

These observations imply that, even after mapping both modalities into a common representation, hand-crafted features remain vulnerable to perceptual aliasing in tunnels, resulting in many-to-one ambiguities and spatially inconsistent correspondences. Learning-based matching with contextual reasoning provides a more discriminative assignment, producing a correspondence set that is better suited for subsequent geometric verification and reliability filtering. Therefore, SuperPoint+SuperGlue is adopted as the primary matching module in the proposed pipeline, while additional pruning is still necessary to further suppress residual outliers in highly repetitive segments.

4.3 Stitching-enhanced Cross-modal Matching

Figure 7 compares correspondence mining before and after stitching. Before stitching, limited context and cross-modal appearance gaps result in few valid correspondences surviving reliab-

ility filtering. After stitching, the expanded receptive field and consistent long-range structural cues reduce perceptual aliasing, yielding significantly denser and more reliable 3D–2D matches. This demonstrates that stitching-induced context aggregation is essential for robust cross-modal matching in low-texture tunnel environments. To quantify this effect, five station groups are uniformly sampled along the mileage of each tunnel dataset, and the number of retained homonymous points after reliability filtering, $|\mathcal{H}|$, is counted for both settings (before stitching vs. after stitching). The statistical results are summarized in Fig. 8, where stitching consistently yields a substantially larger number of valid correspondences across all sampled groups, with typical improvements on the order of 3–$5\times$ in low-texture segments. These results indicate that stitching-induced context expansion is critical for suppressing perceptual aliasing and stabilizing cross-modal matching in repetitive tunnel environments.

5. Conclusions

This paper investigated cross-modal homonymous point extraction between LiDAR point clouds and multi-camera images in low-texture tunnel environments. By introducing a shared cylindrical surface representation and a dual-projection strategy, both LiDAR intensity and stitched ring imagery were mapped into a common domain where robust feature matching could be performed. A global mileage-based gating and reliability filtering were further applied to suppress long-range perceptual aliasing and remove geometrically inconsistent correspondences caused by repetitive tunnel patterns. Experiments on two real railway tunnels demonstrated that image stitching and the proposed matching pipeline substantially increase the number and stability of valid correspondences, providing a reliable basis for subsequent multi-sensor fusion and tunnel inspection tasks. Future work will extend the parameterization to more general cross-sections and evaluate the impact on downstream localization and defect analysis.

Acknowledgments

This study was supported in part by the National Natural Science Foundation of China (Project No. 42230102) and the Natural Science Foundation of Sichuan Province under Grant 2026NSF-SCZY0054, Key Technology Research and Application Demonstration of Real-3D Scene Modeling for Green Energy Engineering Projects (Project No.: Yunnan Engineering Research Center Innovation Capacity Building and Enhancement 202403).

References

Cheng, L., Guo, L., Zhang, T., Bang, T., Harris, A., Hajij, M., Sartipi, M., Cao, S., 2025. CalibRefine: Deep Learning-Based Online Automatic Targetless LiDAR–Camera Calibration with Iterative and Attention-Driven Post-Refinement. *IEEE Transactions on Instrumentation and Measurement*.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 224–236.

Hawley, C. J., Gräbe, P., 2022. Water leakage mapping in concrete railway tunnels using LiDAR generated point clouds. *Construction and Building Materials*, 361, 129644.

Huang, Z., Zhang, Y., Chen, Q., Fan, R., 2024. Online, target-free LiDAR-camera extrinsic calibration via cross-modal mask matching. *IEEE Transactions on Intelligent Vehicles*, 10(5), 3531–3542.

Jiang, Y., Hu, H., Si, S., Zhang, Y., Chen, J., Guo, X., Ding, Y., Zhong, R., Zhu, Q., 2023. A flexible calibration method with multi-stage optimization for the axial error of mobile mapping systems. *International Journal of Applied Earth Observation and Geoinformation*, 118, 103240.

Lei, Y., Tian, T., Jiang, B., Qi, F., Jia, F., Qu, Q., 2023. Research and application of the obstacle avoidance system for high-speed railway tunnel lining inspection train based on integrated 3D LiDAR and 2D camera machine vision technology. *Applied Sciences*, 13(13), 7689.

Li, J., Wang, X., Liu, Y., Zeng, Z., 2024. CFN-ESA: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition. *IEEE Transactions on Affective Computing*, 15(4), 1919–1933.

Li, L., Han, L., Ye, Y., Xiang, Y., Zhang, T., 2025. Deep learning in remote sensing image matching: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 225, 88–112.

Lin, X., Yang, X., Yao, W., Wang, X., Ma, X., Ma, B., 2024. Graph-based adaptive weighted fusion SLAM using multimodal data in complex underground spaces. *ISPRS Journal of Photogrammetry and Remote Sensing*, 217, 101–119.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91–110.

Ma, W., Huang, Y., Tang, S., Zheng, X., Dong, Z., Ge, L., Pan, J., Li, Q., Wang, B., 2025. Cross-modal 2D-3D feature matching: simultaneous local feature description and detection across images and point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 229, 155–169.

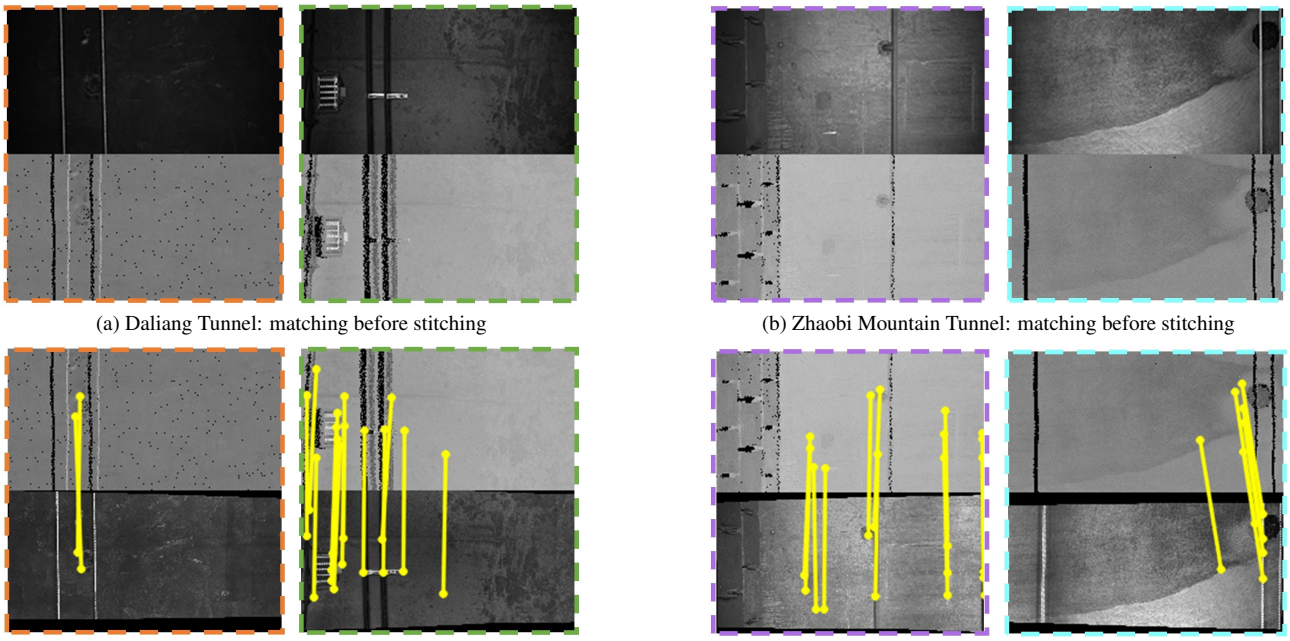
Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. Orb: An efficient alternative to sift or surf. *2011 International conference on computer vision*, Ieee, 2564–2571.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4938–4947.

Shim, S., Lee, S.-W., Cho, G.-C., Kim, J., Kang, S.-M., 2023. Remote robotic system for 3D measurement of concrete damage in tunnel with ground vehicle and manipulator. *Computer-Aided Civil and Infrastructure Engineering*, 38(15), 2180–2201.

Su, B., Zhang, J., Tang, Y., Yu, J., Wang, E., Vashishtha, G., Li, Z., 2025. Geological disaster detection in underground mining tunnels using a new electromagnetic method: Theoretical modeling and experimental evaluation. *Journal of Applied Geophysics*, 233, 105641.

Sun, Y., Dai, C., Li, W., Liu, X., Sun, Y., Zhang, Y., Guan, W., Zhang, Y., Guo, Y., Wang, H., 2026. RTPSeg: A multi-modality dataset for LiDAR point cloud semantic segmentation assisted with RGB-thermal images in autonomous driving. *ISPRS Journal of Photogrammetry and Remote Sensing*, 233, 25–38.



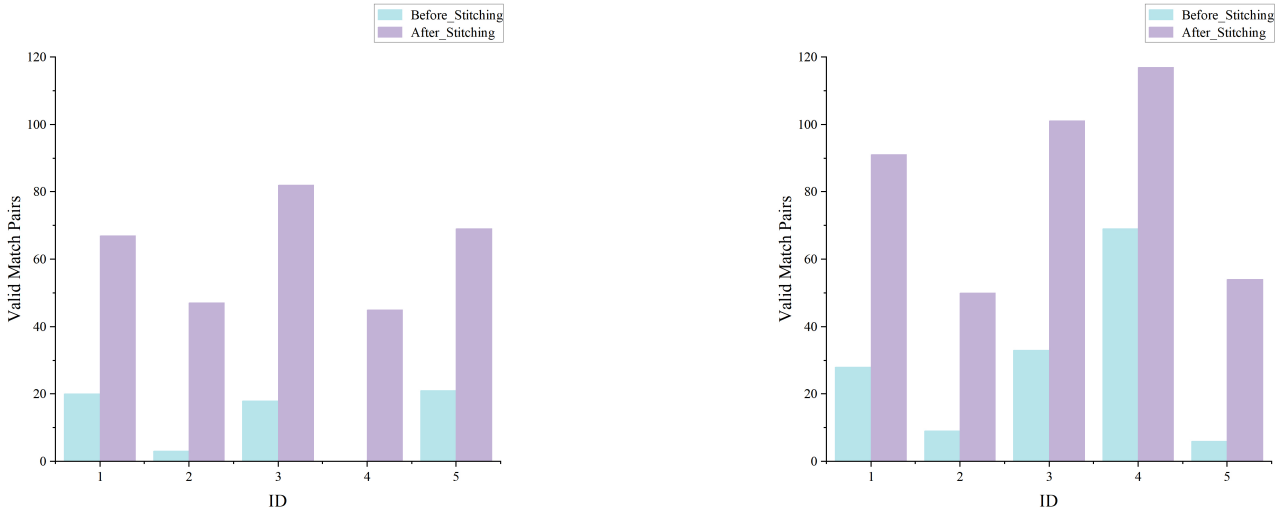
(a) Daliang Tunnel: matching before stitching

(b) Zhaobi Mountain Tunnel: matching before stitching

(c) Daliang Tunnel: matching after stitching

(d) Zhaobi Mountain Tunnel: matching after stitching

Figure 7. Cross-modal matching before and after image stitching. (a-b) Raw images fail to establish reliable correspondences due to limited field-of-view and structural context. (c-d) Stitched ring images from Daliang and Zhaobi Mountain Tunnels enable robust matching, demonstrating that expanding the receptive field is essential for cross-modal alignment in low-texture tunnels.



(a) Daliang Tunnel.

(b) Zhaobi Mountain Tunnel.

Figure 8. Quantitative comparison of valid cross-modal correspondences before and after stitching in two real tunnel datasets.

Wang, Z., Huang, S., Butt, J. A., Cai, Y., Varga, M., Wieser, A., 2025. Cross-modal feature fusion for robust point cloud registration with ambiguous geometry. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227, 31–47.

Xiong, H., Li, J., Liu, T., Zhang, K., Li, Z., 2025. Catenary clutter elimination network for railway tunnel ground penetrating radar data. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 1–14.

Xiong, H., Li, J., Su, G., Li, Z., Zhang, Z., 2023. Automatic defect detection in operational high-speed railway tunnels guided by train-mounted ground penetrating radar data. *Journal of Applied Geophysics*, 219, 105219.

Yan, Z.-g., Li, J.-t., Shen, Y., Xiao, Z.-q., Ai, Q., Zhu, H.-h., 2023. Damage identification method on shield tunnel based on PLSR and equivalent damage analysis. *Tunnelling and Underground Space Technology*, 137, 105127.

Zahs, V., Anders, K., Kohns, J., Stark, A., Höfle, B., 2023. Classification of structural building damage grades from multi-temporal photogrammetric point clouds using a machine learning model trained on virtual laser scanning data. *International Journal of Applied Earth Observation and Geoinformation*, 122, 103406.

Zhu, J., Li, X., Xu, Q., Sun, Z., 2023. Robust online calibration of LiDAR and camera based on cross-modal graph neural network. *IEEE Transactions on Instrumentation and Measurement*, 72, 1–14.