

Article

A Structure-Aware Multi-Modal Learning Framework for Robust Indoor Point-Cloud Registration

Changliang Zhang ^{1,2}, Qingshan Xu ^{2,*}, Xiongwei Sun ² and Gongqin Zhu ^{1,2}¹ Science Island Branch, Graduate School, University of Science and Technology of China, Hefei 230026, China; zhangchangliang@mail.ustc.edu.cn (C.Z.); monszed@163.com (G.Z.)² Anhui Institute of Optics and Fine Mechanics, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China; xiongweisun@163.com

* Correspondence: qshxu@aiofm.ac.cn

Abstract

Accurate point-cloud registration is a fundamental task in intelligent 3D perception systems, including autonomous robotics, indoor digital twins, augmented reality, smart sensing, and scene-level spatial understanding. However, reliable registration in structured indoor environments remains challenging because repetitive architectural layouts, extensive planar regions, limited overlap, and rotational symmetries often weaken local geometric distinctiveness and lead to ambiguous correspondences. In addition, appearance inconsistency caused by illumination changes, exposure variation, and sensor-dependent color responses further degrades the reliability of cross-view matching. To address these issues, this paper presents MaCSE-Reg, an indoor registration method that combines explicit Manhattan-axis modeling with independent multi-color-space point-wise encoding. Rather than proposing a wholly new multi-modal paradigm, MaCSE-Reg builds on prior geometry–color fusion studies and focuses on two under-emphasized design choices for structured indoor scenes: (i) dominant orthogonal axes estimated from surface normals, regularized by an axis-consistency objective and gated by an axis-confidence score, and (ii) complementary RGB, Lab, and HSV representations learned through independent point-wise encoders for illumination-tolerant appearance matching. Geometric, structural, and color-aware features are adaptively fused, and a learnable multi-modal matching formulation is used to refine correspondences before weighted Procrustes pose estimation. Experiments on standard indoor registration benchmarks demonstrate that MaCSE-Reg improves registration robustness and pose accuracy under low-overlap conditions, repetitive structures, and illumination variations. The results support the value of this specific structural–appearance combination for intelligent indoor 3D perception while remaining complementary to existing geometry–color registration frameworks.



Academic Editors: Yugang Liu and Sidney Givigi

Received: 10 June 2026

Revised: 5 July 2026

Accepted: 8 July 2026

Published: 11 July 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.

Keywords: point-cloud registration; intelligent 3D perception; multi-modal representation learning; Manhattan-world assumption; multi-color-space encoding; indoor spatial intelligence

1. Introduction

Accurate point-cloud registration is a core capability of intelligent 3D perception systems. With the rapid development of RGB-D cameras, LiDAR sensors, and indoor reconstruction devices, large-scale point clouds have become an important data modality for autonomous robotics, augmented reality, digital twins, smart manufacturing, and indoor

spatial intelligence. In these applications, multiple partial scans must be aligned into a consistent coordinate system before reliable mapping, localization, navigation, or scene understanding can be performed. Therefore, robust registration is not only a geometric estimation problem but also a key enabling technique for intelligent sensing and spatial perception systems.

Despite substantial progress, reliable point-cloud registration in real-world indoor environments remains challenging. Man-made scenes usually contain extensive planar structures and repetitive architectural layouts, such as walls, floors, corridors, rooms, and tiled surfaces. Such layouts can produce multiple geometrically plausible alignments, making local geometric descriptors insufficiently distinctive and causing unstable correspondence estimation. The problem becomes more severe when fragment overlap is limited, viewpoint changes are large, or occlusion increases the outlier ratio. Classical optimization-based methods, including ICP and its variants, are effective when a good initialization is available but often degrade under large pose differences or weak geometric distinctiveness [1,2]. Learning-based registration methods improve descriptor discriminability and correspondence filtering, yet many of them remain dominated by local geometry and may still suffer from rotational ambiguity in highly structured indoor scenes [3–5]. Recent attention-based models further enhance contextual reasoning and achieve strong performance on challenging benchmarks [6–8]. At the same time, prior studies have already shown that combining geometric and appearance cues can improve indoor registration, for example, through color-guided projection modules [9], cross-modal image–point-cloud attention [10], and geometry–color fusion transformers [11,12]. Therefore, the remaining gap addressed here is not multi-modal fusion itself but how to combine an explicit Manhattan structural prior with independent multi-color-space point-wise encoding in a single learning-based registration pipeline.

A useful observation is that indoor environments are not arbitrary. Many indoor scenes approximately follow the Manhattan-world assumption, where dominant architectural directions are aligned with three mutually orthogonal axes [13]. These global directions provide strong structural cues for resolving rotational ambiguity in repetitive planar regions. Nevertheless, most registration networks use such structural regularities only implicitly or ignore them entirely and therefore fail to fully exploit global orientation information. Meanwhile, color information can complement geometry by providing appearance cues for correspondence disambiguation. Direct RGB values, however, are sensitive to illumination changes, exposure differences, and sensor-specific color responses. Alternative color spaces such as CIE $L^*a^*b^*$ and HSV provide complementary perceptual descriptions by partially separating intensity and chromaticity, which can improve the stability of appearance representation under photometric variation [14]. These considerations motivate MaCSE-Reg as a focused design that couples explicit Manhattan-axis modeling with independent RGB/Lab/HSV point-wise encoding rather than as a wholly new registration paradigm.

In this paper, we propose MaCSE-Reg, a learning-based indoor registration method that combines explicit Manhattan structural modeling with independent multi-color-space appearance encoding. As illustrated in Figure 1, the proposed method addresses two major sources of registration uncertainty in indoor scenes: structural ambiguity caused by repetitive planar layouts and appearance inconsistency caused by photometric variation. MaCSE-Reg estimates local surface normals, clusters them into dominant directions, and orthogonalizes the result to construct a Manhattan reference frame. The aligned normal axes are encoded as structural representations, regularized by an axis-consistency objective, and gated by an axis-confidence score when axis evidence is unreliable. In parallel, per-point RGB colors are transformed into Lab and HSV spaces, and independent appearance encoders learn complementary color-space representations. The geometric, structural,

and appearance representations are adaptively fused and used to construct a learnable multi-modal matching cost for reliable correspondence estimation and pose recovery.

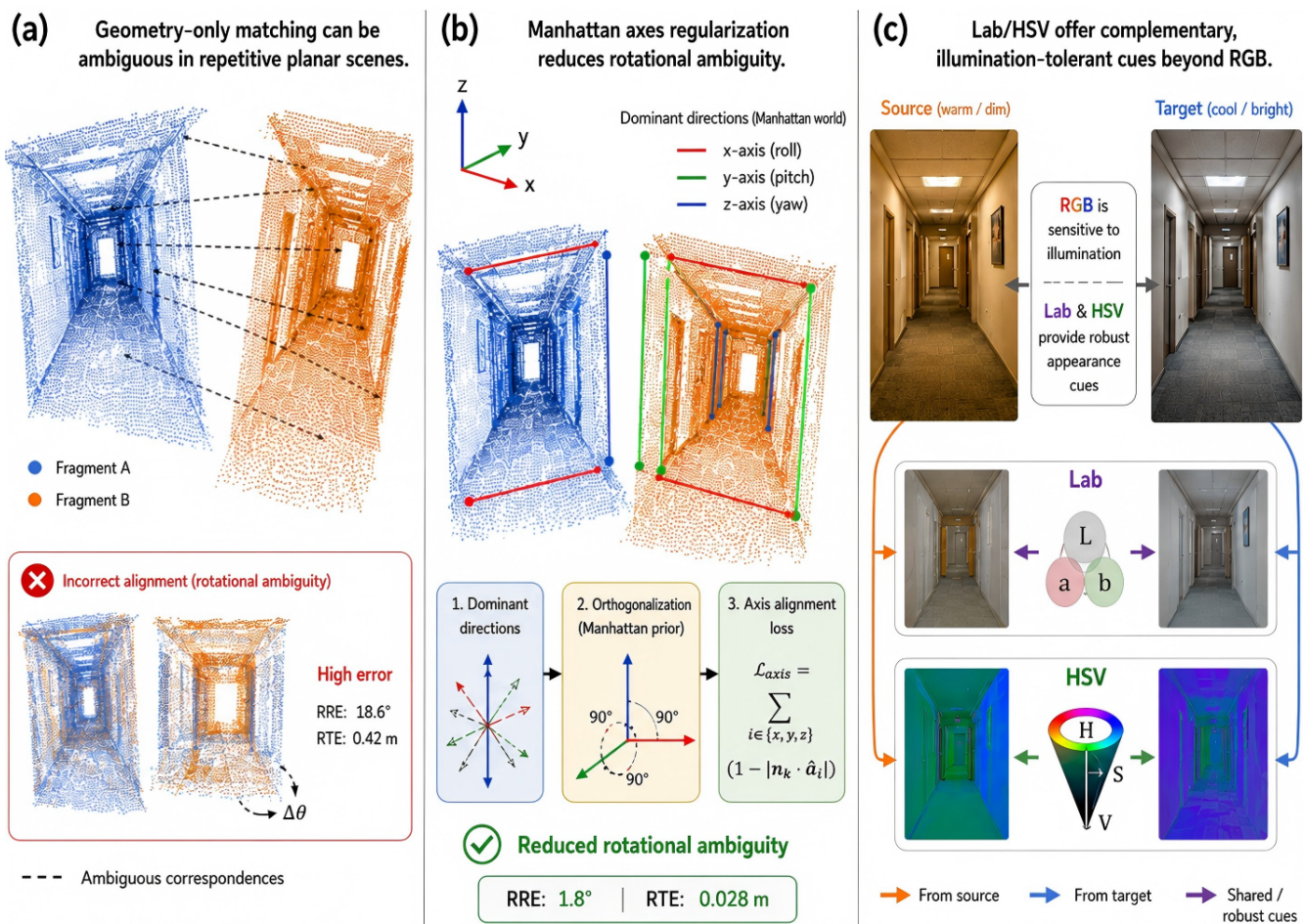


Figure 1. Motivation. Indoor point-cloud registration is prone to ambiguity due to repetitive planar structures and rotational symmetries. (a) Geometry-only matching may yield unreliable correspondences in structured scenes, resulting in large alignment errors. (b) Manhattan-world priors provide dominant orthogonal axes, reducing rotational ambiguity via axis-consistency regularization. (c) Multi-color-space encoding in Lab/HSV spaces offers complementary, illumination-tolerant appearance cues beyond RGB, improving correspondence reliability under lighting variation. The green arrows in subfigure (c) indicate shared robust appearance cues learned from the Lab/HSV representations.

The main contributions of this work are summarized as follows:

- We present MaCSE-Reg as a focused combination of explicit Manhattan-axis modeling and independent multi-color-space point-wise encoding for indoor point-cloud registration, building on prior geometry–color fusion methods rather than claiming a wholly new multi-modal paradigm.
- We introduce an explicit Manhattan structural branch that estimates dominant orthogonal axes from surface normals, enforces axis consistency, and applies confidence-aware gating, thereby reducing rotational ambiguity in structured indoor scenes when axis estimation is reliable.
- We develop independent RGB/Lab/HSV point-wise appearance encoders, together with learnable multi-modal matching, to improve correspondence reliability under illumination variation, repetitive structures, and low-overlap conditions.

2. Related Work

Point-cloud registration has been extensively studied in 3D vision, robotics, and intelligent sensing. Existing methods can be broadly reviewed from the perspectives of geometric optimization, learned correspondence representation, Transformer-based contextual reasoning, multi-modal perception, and structure-aware indoor scene understanding.

2.1. Geometric Registration and Robust Estimation

Classical registration pipelines typically alternate between correspondence assignment and rigid motion estimation. Iterative Closest Point (ICP) and its variants remain widely used when a reliable initialization is available [1,2]. To improve robustness against local minima and outliers, global registration methods optimize robust objectives over tentative correspondences such as Fast Global Registration (FGR) [15]. Certifiable solvers such as TEASER and TEASER++ further provide strong robustness to extreme outlier ratios with theoretical guarantees [16]. These methods constitute important foundations for point-cloud registration, but they mainly rely on geometric consistency and may be insufficient when indoor scenes contain repeated structures or multiple plausible alignments.

2.2. Deep Feature Learning for Correspondence Estimation

Learning-based registration methods improve correspondence reliability by learning discriminative local descriptors and inlier confidence. 3DMatch introduced learned local 3D descriptors for fragment matching and became a standard benchmark for indoor registration [6]. Later methods, such as FCGF and D3Feat, enhanced local representation learning using fully convolutional networks and joint detection-description strategies [17,18]. PointDSC further models spatial consistency to prune outlier correspondences and improve the reliability of pose estimation [19]. Strong geometric backbones such as KPConv are commonly adopted for local feature learning in registration pipelines [20]. Recent descriptor-learning studies have also investigated broader receptive fields, rotation-invariant surface descriptors, and position-aware geometric encoding for low-overlap and cross-source registration [21,22]. Although these methods improve local feature discriminability, many of them remain geometry-centric and do not explicitly model global structural priors or robust appearance cues.

2.3. Transformer-Based Representation Learning

Transformer-based registration methods have attracted increasing attention because self-attention and cross-attention can model long-range dependencies and contextual correspondences. PREDATOR introduces overlap-aware attention to emphasize potentially overlapping regions, improving low-overlap registration [5]. REGTR uses self- and cross-attention to predict correspondences in an end-to-end manner, reducing dependence on hand-crafted matching stages [23]. GeoTransformer encodes geometric structures, including distances and angles, into attention mechanisms and achieves strong performance on indoor benchmarks [7]. PEAL introduces overlap priors into Transformer attention to further improve registration under limited overlap [24]. More recent designs explore targeted attention mechanisms, projection-aware LiDAR registration, and coarse-to-fine correspondence learning for robust registration reasoning [25–27]. End-to-end correspondence learning has also been extended through hierarchical and generative formulations [3,27]. Very recent studies have further pushed indoor and color-aware registration, including IGASA with hierarchical geometry-aware attention [28], RegFormer++ as an extension of projection-aware Transformer registration [26,29], and GeGS-PCR with geometric–color–Gaussian fusion [12]. These studies demonstrate the importance of representation learning, but structural regularities in indoor scenes remain to be fully exploited.

2.4. Multi-Modal Representation Learning for Registration

Geometry alone may be ambiguous in structured indoor environments, while appearance information can provide complementary evidence for correspondence disambiguation. Colored ICP combines photometric and geometric objectives and has shown improved robustness over purely geometric ICP in colored point clouds [30]. Recent learning-based methods also explore color–geometry fusion for point-cloud registration. For example, PCR-CG incorporates a 2D–3D projection module to lift image color features into geometric representations [9]. CMHANet further fuses 2D visual semantics with 3D geometric features through a cross-modal hybrid attention network [10], and recent Transformer-based designs such as CART and GeGS-PCR also combine geometry with color-aware cues for indoor registration [11,12]. These studies demonstrate that geometry–color fusion is already an active research direction. MaCSE-Reg differs from them in two specific aspects: it uses explicit Manhattan-axis modeling from surface normals as a global structural prior rather than relying only on learned geometric or image-based context, and it employs *independent RGB/Lab/HSV point-wise encoders* on colored point clouds rather than projecting 2D image features or using raw RGB-dominated fusion alone. Many prior geometry–color joint descriptors also rely on hand-crafted photometric residuals or global color statistics, whereas MaCSE-Reg targets point-wise, learnable, and illumination-tolerant appearance encoding through complementary color spaces. However, raw RGB values are sensitive to variation in illumination and sensor response. This motivates the use of complementary color-space representations to improve appearance robustness in indoor registration.

2.5. Structure-Aware Indoor Scene Understanding

Man-made indoor environments often exhibit strong structural regularities. The Manhattan-world assumption, according to which dominant scene directions are aligned with mutually orthogonal axes, has long been used in indoor layout estimation and scene understanding [13]. For registration, such global structural cues can reduce rotational ambiguity and provide additional constraints in repetitive planar layouts. Although geometry–color fusion has been widely studied, explicit Manhattan-axis modeling estimated from surface normals and integrated into a learnable registration matcher remains comparatively less common. MaCSE-Reg is positioned as a complementary design in this direction, combining axis-consistency regularization and confidence-aware gating with the multi-color-space appearance branch described above.

2.6. Benchmarks and Toolkits

3DMatch remains one of the most widely used datasets and evaluation suites for indoor point-cloud registration under realistic RGB-D reconstruction noise [6]. Recent benchmarks and application-driven studies increasingly emphasize cross-domain generalization, cultural-heritage digitization, and scalable point-cloud backbones [31,32]. Practical systems commonly build upon mature toolkits and GPU-accelerated operators such as Open3D and PyTorch3D [33].

3. Materials and Methods

We propose MaCSE-Reg, a registration architecture that combines explicit Manhattan structural modeling with independent multi-color-space appearance learning for structured indoor scenes. Given a source point cloud ($\mathbf{P}^s \in \mathbb{R}^{N \times 3}$) and a target point cloud ($\mathbf{P}^t \in \mathbb{R}^{M \times 3}$) optionally equipped with per-point colors ($\mathbf{C}^s \in [0, 1]^{N \times 3}$ and $\mathbf{C}^t \in [0, 1]^{M \times 3}$), the objective is to estimate a rigid transformation ($\mathbf{T} = [\mathbf{R} \mid \mathbf{t}] \in SE(3)$) that aligns $\mathbf{P}^s$ with $\mathbf{P}^t$. The design should be understood as a specific structural–appearance combination built on standard geometric feature learning rather than as a wholly new multi-modal registration paradigm.

As shown in Figure 2, MaCSE-Reg learns point-wise geometric representations using a shared encoder, constructs structure-aware representations from Manhattan axes estimated from surface normals, and extracts appearance representations from multiple color spaces. These complementary representations are adaptively integrated by residual fusion and used to compute a learnable multi-modal matching cost. Soft correspondences and their confidence scores are then estimated from the matching distribution, and the final pose is recovered using a weighted Procrustes solver based on singular value decomposition (SVD).

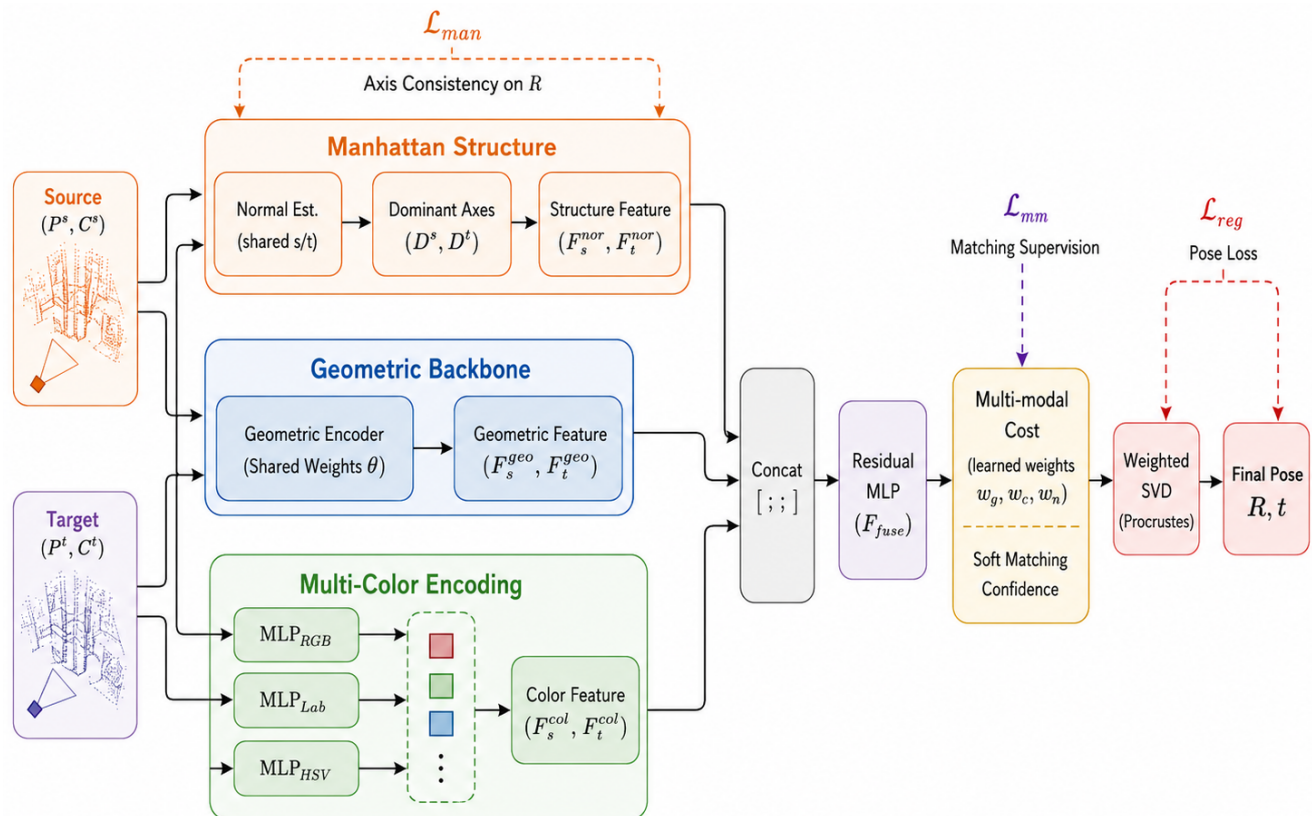


Figure 2. Overview of MaCSE-Reg. Given a source and a target point cloud with optional colors, MaCSE-Reg extracts geometric representations from a shared-weight geometric encoder, Manhattan structural representations from dominant orthogonal axes estimated via surface normals, and illumination-tolerant appearance representations from multi-color-space encoders (RGB/Lab/HSV). The three modalities are concatenated and fused by a residual MLP to produce fused descriptors, which are used to compute a learnable weighted multi-modal matching cost and soft correspondence confidence. The final rigid transformation ($\mathbf{R}, \mathbf{t}$) is obtained by a weighted Procrustes solver. MaCSE-Reg is trained with $\mathcal{L}_{reg} = \mathcal{L}_{pose} + \mu \mathcal{L}_{cls}$, a multi-modal matching loss ($\mathcal{L}_{mm}$) with positive and negative terms, and a Manhattan-axis consistency regularizer ($\mathcal{L}_{man}$).

3.1. Geometric Representation Learning

MaCSE-Reg uses a shared-weight geometric encoder to extract point-wise geometric representations from the source and target point clouds. We denote the encoder by $\Phi_{\theta}(\cdot)$, with parameters (θ) shared across both branches:

$$\mathbf{F}_s^{geo} = \Phi_{\theta}(\mathbf{P}^s), \quad \mathbf{F}_t^{geo} = \Phi_{\theta}(\mathbf{P}^t), \quad (1)$$

where $\mathbf{F}_s^{geo} \in \mathbb{R}^{N \times d}$ and $\mathbf{F}_t^{geo} \in \mathbb{R}^{M \times d}$ are d -dimensional geometric embeddings. The proposed framework is agnostic to the specific backbone design and can be instantiated with a local-to-global encoder that yields robust geometric descriptors.

3.2. Manhattan-World Structural Constraints

3.2.1. Surface-Normal Estimation

To exploit the Manhattan-world prior, we compute per-point surface normals based on local neighborhood plane fitting. We retrieve the K nearest neighbors ($\mathcal{N}_K(\mathbf{p}_i)$) of each point ($\mathbf{p}_i$) and form the neighborhood set ($\mathbf{Q}_i = \{\mathbf{p}_j \mid j \in \mathcal{N}_K(\mathbf{p}_i)\}$) with centroid $\bar{\mathbf{q}}_i$. We compute the covariance matrix as

$$\Sigma_i = \frac{1}{K} \sum_{\mathbf{q} \in \mathbf{Q}_i} (\mathbf{q} - \bar{\mathbf{q}}_i)(\mathbf{q} - \bar{\mathbf{q}}_i)^\top \quad (2)$$

and take the eigenvector associated with the smallest eigenvalue as the normal direction:

$$\mathbf{n}_i = \text{eigmin}(\Sigma_i), \quad \bar{\mathbf{n}}_i = \frac{\mathbf{n}_i}{\|\mathbf{n}_i\|_2}. \quad (3)$$

We compute normals for both source and target point clouds, yielding $\{\bar{\mathbf{n}}_i^s\}_{i=1}^N$ and $\{\bar{\mathbf{n}}_j^t\}_{j=1}^M$. In implementation, we prefer GPU-accelerated KNN and SVD-based plane fitting when available and fall back to CPU-based computation otherwise.

3.2.2. Dominant-Axis Estimation and Orthogonalization

Given normalized normals, we estimate three dominant directions that define a Manhattan reference frame. We cluster normals into three groups to obtain direction prototypes:

$$\{\tilde{\mathbf{d}}_1, \tilde{\mathbf{d}}_2, \tilde{\mathbf{d}}_3\} = \text{Cluster}(\{\bar{\mathbf{n}}_i\}, k = 3), \quad (4)$$

where $\text{Cluster}(\cdot)$ is instantiated with K-means by default and can be replaced by PCA-based extraction as a fallback. Since the resulting prototypes are not guaranteed to be orthogonal, we apply Gram–Schmidt orthogonalization to construct a valid Manhattan frame:

$$\mathbf{d}_1 = \frac{\tilde{\mathbf{d}}_1}{\|\tilde{\mathbf{d}}_1\|_2}, \quad (5)$$

$$\mathbf{d}_2 = \frac{\tilde{\mathbf{d}}_2 - (\tilde{\mathbf{d}}_2^\top \mathbf{d}_1) \mathbf{d}_1}{\|\tilde{\mathbf{d}}_2 - (\tilde{\mathbf{d}}_2^\top \mathbf{d}_1) \mathbf{d}_1\|_2}, \quad (6)$$

$$\mathbf{d}_3 = \mathbf{d}_1 \times \mathbf{d}_2. \quad (7)$$

We perform this procedure independently for the source and target point clouds to obtain dominant axes $\mathbf{D}^s = \{\mathbf{d}_1^s, \mathbf{d}_2^s, \mathbf{d}_3^s\}$ and $\mathbf{D}^t = \{\mathbf{d}_1^t, \mathbf{d}_2^t, \mathbf{d}_3^t\}$.

3.2.3. Axis Confidence Score

Because Manhattan-axis estimation depends on KNN-based normal clustering, its reliability is affected by point-cloud density, depth-sensor noise, and irregular sampling. We therefore introduce a fragment-level axis confidence score that quantifies how tightly surface normals concentrate around the estimated dominant directions. For each point, we define

$$\eta_i = \max_{k \in \{1,2,3\}} |\bar{\mathbf{n}}_i^\top \mathbf{d}_k| \in [0, 1] \quad (8)$$

and aggregate it into scan-level confidence:

$$\gamma = \frac{1}{N} \sum_{i=1}^N \eta_i. \quad (9)$$

A higher γ value indicates that normals form three well-separated, dominant peaks, whereas a lower γ value suggests noisy, sparse, or weakly planar geometry. To characterize the boundary of the Manhattan prior more explicitly, we further define pair-level structural confidence as

$$\Gamma = \min(\gamma^s, \gamma^t), \quad (10)$$

which measures the weaker fragment in a source–target pair. A pair is classified as **high-confidence Manhattan** when $\Gamma \geq \tau_\gamma$ and as **weak Manhattan** when $\Gamma < \tau_\gamma$. In this paper, we set $\tau_\gamma = 0.65$ by analyzing the validation-split distribution of γ : it corresponds to the point below which the Manhattan branch no longer yields a consistent gain over the geometry-only variant and approximately matches the lower quartile of fragment-level confidence scores. This threshold is kept fixed for all test analyses.

We use γ to gate the structural branch when axis evidence is unreliable. Concretely, the Manhattan regularizer and the effective structural matching weight are scaled as

$$\mathcal{L}_{man} \leftarrow (\gamma^s \gamma^t) \mathcal{L}_{man}, \quad w_n^{\text{eff}} = w_n \sqrt{\gamma^s \gamma^t} \quad (11)$$

so that geometry and appearance cues dominate when axis estimation is uncertain. This design prevents unreliable structural priors from over-constraining pose estimation under low-quality point clouds. Section 4.14 reports the size of the resulting high-confidence and weak Manhattan subsets on the 3DMatch and 3DLoMatch test splits.

3.2.4. Normal Alignment and Structural Feature Encoding

Each normal is aligned to its closest dominant axis, up to sign, to obtain a structure-aware attribute:

$$k^*(i) = \arg \max_{k \in \{1,2,3\}} |\bar{\mathbf{n}}_i^\top \mathbf{d}_k|, \quad \mathbf{n}_i^{\text{ali}} = \text{sign}(\bar{\mathbf{n}}_i^\top \mathbf{d}_{k^*(i)}) \mathbf{d}_{k^*(i)}. \quad (12)$$

We then encode the aligned normal into a compact structure feature using an MLP, i.e.,

$$\mathbf{f}_i^{\text{nor}} = \text{MLP}_{\text{nor}}(\mathbf{n}_i^{\text{ali}}) \in \mathbb{R}^{d/4}, \quad (13)$$

and collect structural representations for both point clouds:

$$\mathbf{F}_s^{\text{nor}} \in \mathbb{R}^{N \times d/4}, \quad \mathbf{F}_t^{\text{nor}} \in \mathbb{R}^{M \times d/4}. \quad (14)$$

3.2.5. Manhattan-Axis Consistency Regularization

To enforce global structural consistency between the two scans, we regularize the predicted rotation ($\mathbf{R}$) to align the dominant source axes to the dominant target axes. To handle axis permutation and sign ambiguity, we use a max-cosine formulation:

$$\mathcal{L}_{man} = \frac{1}{3} \sum_{k=1}^3 \left(1 - \max_{j \in \{1,2,3\}} |(\mathbf{R} \mathbf{d}_k^s)^\top \mathbf{d}_j^t| \right). \quad (15)$$

3.3. Multi-Color-Space Appearance Representation Learning

Color complements geometry in structured indoor environments, but the raw RGB space is sensitive to illumination and sensor response. We therefore encode appearance cues in multiple color spaces, including RGB, Lab, and HSV. For each point color ($\mathbf{c}_i^{\text{RGB}} \in [0, 1]^3$), we compute

$$\mathbf{c}_i^{\text{Lab}} = \phi_{\text{Lab}}(\mathbf{c}_i^{\text{RGB}}), \quad \mathbf{c}_i^{\text{HSV}} = \phi_{\text{HSV}}(\mathbf{c}_i^{\text{RGB}}), \quad (16)$$

where $\phi_{Lab}(\cdot)$ denotes the standard RGB→XYZ→Lab conversion with D65 white-point normalization and $\phi_{HSV}(\cdot)$ denotes the conventional hue–saturation–value conversion. Each color space is processed by an independent encoder, i.e.,

$$\mathbf{f}_i^{RGB} = \text{MLP}_{RGB}(\mathbf{c}_i^{RGB}) \in \mathbb{R}^{d/4}, \quad (17)$$

$$\mathbf{f}_i^{Lab} = \text{MLP}_{Lab}(\mathbf{c}_i^{Lab}) \in \mathbb{R}^{d/4}, \quad (18)$$

$$\mathbf{f}_i^{HSV} = \text{MLP}_{HSV}(\mathbf{c}_i^{HSV}) \in \mathbb{R}^{d/4}, \quad (19)$$

and concatenated into a multi-color appearance descriptor:

$$\mathbf{f}_i^{col} = [\mathbf{f}_i^{RGB}; \mathbf{f}_i^{Lab}; \mathbf{f}_i^{HSV}] \in \mathbb{R}^{3d/4}. \quad (20)$$

Stacking all point-wise descriptors yields $\mathbf{F}_s^{col} \in \mathbb{R}^{N \times 3d/4}$ and $\mathbf{F}_t^{col} \in \mathbb{R}^{M \times 3d/4}$. When colors are unavailable, MaCSE-Reg drops the color branch and only uses geometric and structural representations.

Design rationale.

We adopt RGB, Lab, and HSV color spaces because they provide complementary point-wise appearance cues that are directly compatible with correspondence learning. The Lab color space is perceptually uniform and partially separates luminance from chromaticity, which improves robustness to exposure and white-balance changes [14]. The HSV color space further decouples hue and saturation from intensity, making it less sensitive to local brightness variation in indoor scenes. Unlike global color histograms, these representations remain defined at each point and therefore preserve the fine-grained discriminability required for fragment-level matching. Color-name mappings, although compact and interpretable, quantize appearance into a small categorical vocabulary and discard subtle chromatic differences that are often useful for disambiguating repetitive geometric structures. Likewise, histogram descriptors aggregate neighborhood statistics and are more suitable for recognition or retrieval than for precise point-to-point registration. By contrast, independent, learnable encoders over RGB, Lab, and HSV color spaces allow the network to extract task-adaptive appearance features while keeping the same geometric and structural backbone. Section 4.11 compares this design against color histograms, color-name mappings, and a joint colored ICP-style geometry–color descriptor [30].

3.4. Adaptive Multi-Modal Representation Fusion

We fuse geometric, structural, and color cues through a residual multi-modal fusion module:

$$\mathbf{f}_i^{fuse} = \text{MLP}_{fus}([\mathbf{f}_i^{geo}; \mathbf{f}_i^{col}; \mathbf{f}_i^{nor}]) + \mathbf{f}_i^{geo}, \quad (21)$$

where $[\cdot; \cdot; \cdot]$ denotes concatenation and MLP_{fus} is a lightweight fusion network. The fused features form $\mathbf{F}_s^{fuse} \in \mathbb{R}^{N \times d}$ and $\mathbf{F}_t^{fuse} \in \mathbb{R}^{M \times d}$ and are used for correspondence reasoning.

Design rationale.

Although cross-attention and gated fusion are widely used in multi-modal learning, we adopt a residual MLP design for three practical reasons. First, point-cloud registration requires point-wise discriminative descriptors; at this stage, the three modalities are already aligned per point, so expensive cross-modal attention at every point yields limited benefit over a compact fusion MLP while increasing memory and latency on large fragments. Second, the geometric encoder is the strongest and most stable backbone in our pipeline, and the residual connection explicitly preserves $\mathbf{f}_i^{geo}$ as an anchor feature, reducing the risk that auxiliary structural or appearance cues destabilize correspondence

learning. Third, MaCSE-Reg already performs adaptive multi-modal reasoning at the matching stage through learnable weights (w_g, w_c, w_n) in Equation (25), which calibrates modality importance during correspondence construction. Therefore, a lightweight fusion module is sufficient to inject complementary cues without duplicating a second attention mechanism. Section 4.8 compares this design against cross-attention fusion and gated fusion alternatives.

3.5. Learnable Multi-Modal Matching and Pose Estimation

3.5.1. Multi-Modal Matching Cost

We define per-modality cosine distances between a source point (i) and a target point (j) as follows:

$$\mathcal{C}_{ij}^{geo} = 1 - \frac{\langle \mathbf{f}_i^{geo,s}, \mathbf{f}_j^{geo,t} \rangle}{\|\mathbf{f}_i^{geo,s}\|_2 \|\mathbf{f}_j^{geo,t}\|_2}, \quad (22)$$

$$\mathcal{C}_{ij}^{col} = 1 - \frac{\langle \mathbf{f}_i^{col,s}, \mathbf{f}_j^{col,t} \rangle}{\|\mathbf{f}_i^{col,s}\|_2 \|\mathbf{f}_j^{col,t}\|_2}, \quad (23)$$

$$\mathcal{C}_{ij}^{nor} = 1 - \frac{\langle \mathbf{f}_i^{nor,s}, \mathbf{f}_j^{nor,t} \rangle}{\|\mathbf{f}_i^{nor,s}\|_2 \|\mathbf{f}_j^{nor,t}\|_2}. \quad (24)$$

These costs are combined by learnable non-negative weights:

$$\mathcal{C}_{ij} = w_g \mathcal{C}_{ij}^{geo} + w_c \mathcal{C}_{ij}^{col} + w_n \mathcal{C}_{ij}^{nor}, \quad w_g, w_c, w_n \geq 0, \quad (25)$$

where we parameterize (w_g, w_c, w_n) via $\text{softplus}(\cdot)$ or normalize them with softmax for stability.

3.5.2. Soft Matching and Confidence

Based on the fused descriptors and the multi-modal cost, we compute soft correspondences and confidence weights. Concretely, we convert costs to matching probabilities using a temperature-scaled softmax:

$$\pi_{ij} = \frac{\exp(-\mathcal{C}_{ij}/\tau)}{\sum_{j'=1}^M \exp(-\mathcal{C}_{ij'}/\tau)}, \quad (26)$$

where τ is a temperature parameter. The correspondence confidence for source point i is defined as

$$\alpha_i = \max_j \pi_{ij}, \quad (27)$$

and the soft-matched target location is

$$\hat{\mathbf{p}}_i^t = \sum_{j=1}^M \pi_{ij} \mathbf{p}_j^t. \quad (28)$$

3.5.3. Weighted Procrustes Pose Solver

Given soft correspondences $\{(\mathbf{p}_i^s, \hat{\mathbf{p}}_i^t)\}$ with weights represented by α_i , we estimate the rigid transformation by solving a weighted Procrustes problem:

$$\min_{\mathbf{R} \in SO(3), \mathbf{t}} \sum_{i=1}^N \alpha_i \|\mathbf{R} \mathbf{p}_i^s + \mathbf{t} - \hat{\mathbf{p}}_i^t\|_2^2. \quad (29)$$

This has a closed-form SVD solution. Let $\bar{\mathbf{p}}^s$ and $\bar{\mathbf{p}}^t$ be weighted centroids:

$$\bar{\mathbf{p}}^s = \frac{\sum_i \alpha_i \mathbf{p}_i^s}{\sum_i \alpha_i}, \quad \bar{\mathbf{p}}^t = \frac{\sum_i \alpha_i \hat{\mathbf{p}}_i^t}{\sum_i \alpha_i}. \quad (30)$$

We form the weighted cross-covariance as

$$\mathbf{H} = \sum_{i=1}^N \alpha_i (\mathbf{p}_i^s - \bar{\mathbf{p}}^s)(\hat{\mathbf{p}}_i^t - \bar{\mathbf{p}}^t)^\top, \quad (31)$$

compute $\mathbf{U}\Sigma\mathbf{V}^\top = \text{SVD}(\mathbf{H})$, and obtain

$$\mathbf{R} = \mathbf{V} \text{diag}(1, 1, \det(\mathbf{V}\mathbf{U}^\top)) \mathbf{U}^\top, \quad \mathbf{t} = \bar{\mathbf{p}}^t - \mathbf{R}\bar{\mathbf{p}}^s. \quad (32)$$

The complete procedure of MaCSE-Reg is summarized in Algorithm 1.

Algorithm 1: MaCSE-Reg: Structure-Aware Multi-Modal Point-Cloud Registration

Input: Source point cloud $\mathbf{P}^s$, target point cloud $\mathbf{P}^t$, optional colors $\mathbf{C}^s$ and $\mathbf{C}^t$.
Output: Estimated rigid transformation $(\mathbf{R}, \mathbf{t})$.
 Extract geometric descriptors from $\mathbf{P}^s$ and $\mathbf{P}^t$ using a shared geometric encoder;
 Estimate surface normals, cluster them into dominant directions,
 and orthogonalize the directions to obtain Manhattan frames;
 Compute the axis confidence scores γ^s and γ^t and gate the structural
 loss/matching weights accordingly;
 Align normals to the nearest Manhattan axes and encode them as structural
 descriptors;
if *point-wise colors are available* **then**
 Transform RGB colors into Lab and HSV spaces and extract multi-appearance
 descriptors;
 Fuse geometric, structural, and appearance descriptors with a residual
 multi-modal fusion module;
 Compute the learnable multi-modal matching cost and convert it into soft
 correspondence probabilities;
 Estimate correspondence confidence scores and soft target positions;
 Recover $(\mathbf{R}, \mathbf{t})$ using the weighted Procrustes solver;
if *training* **then**
 Optimize the model with the registration loss, multi-modal matching loss,
 and Manhattan axis-consistency loss;
return $(\mathbf{R}, \mathbf{t})$

3.6. Training Objective

MaCSE-Reg is optimized with a registration loss ($\mathcal{L}_{reg}$) and two auxiliary objectives: a Manhattan-axis consistency regularizer ($\mathcal{L}_{man}$) and a multi-modal matching loss $\mathcal{L}_{mm}$. Unlike the wording in a previous draft, $\mathcal{L}_{reg}$ is explicitly defined as the sum of a pose regression term and a correspondence classification term.

3.6.1. Registration Loss

Given the ground-truth rigid transform $(\mathbf{R}^*, \mathbf{t}^*)$ and the estimated transform $(\hat{\mathbf{R}}, \hat{\mathbf{t}})$ recovered by the weighted Procrustes solver, we define the pose regression loss as

$$\mathcal{L}_{pose} = \left\| \text{Log}\left((\mathbf{R}^*)^T \hat{\mathbf{R}}\right) \right\|_2^2 + \beta \|\hat{\mathbf{t}} - \mathbf{t}^*\|_2^2, \quad (33)$$

where $\text{Log}(\cdot)$ maps a rotation matrix to its axis-angle representation and β balances rotation and translation supervision. To explicitly supervise correspondence discriminability, we further define a correspondence classification loss on the soft-matching distribution in Equation (26):

$$\mathcal{L}_{cls} = -\frac{1}{|\mathcal{M}|} \sum_{(i,j) \in \mathcal{M}} \log \pi_{ij}, \quad (34)$$

where $\mathcal{M}$ denotes the set of ground-truth inlier correspondences identified by the standard transformed-distance threshold in the 3DMatch training protocol. The registration loss is then

$$\mathcal{L}_{reg} = \mathcal{L}_{pose} + \mu \mathcal{L}_{cls}, \quad (35)$$

with $\mu = 1.0$ in all experiments. Thus, $\mathcal{L}_{reg}$ is not a vague placeholder: it jointly supervises pose recovery and inlier correspondence classification.

3.6.2. Multi-Modal Matching Loss

To improve the discriminative power of the multi-modal matching cost, $\mathcal{L}_{mm}$ is defined using both positive and negative correspondence pairs. Let $\mathcal{M}$ be the ground-truth inlier set as above, and let $\mathcal{N}$ be a sampled set of negative pairs whose transformed distance exceeds a threshold (τ_{neg}). The positive term encourages low matching costs on true correspondences, whereas the negative term penalizes overly small costs on non-matching pairs through a hinge margin (κ):

$$\mathcal{L}_{mm}^{pos} = \frac{1}{|\mathcal{M}|} \sum_{(i,j) \in \mathcal{M}} \mathcal{C}_{ij}, \quad \mathcal{L}_{mm}^{neg} = \frac{1}{|\mathcal{N}|} \sum_{(i,j) \in \mathcal{N}} \max(0, \kappa - \mathcal{C}_{ij}), \quad (36)$$

$$\mathcal{L}_{mm} = \mathcal{L}_{mm}^{pos} + \lambda_{neg} \mathcal{L}_{mm}^{neg}. \quad (37)$$

Here, $\mathcal{C}_{ij}$ is the joint geometric-structural-appearance cost in Equation (25). The negative hinge term pushes non-matching pairs to have costs larger than κ , which prevents $\mathcal{L}_{mm}$ from only collapsing positive pairs without separating negatives. In our implementation, we set $\tau_{neg} = 0.10$ m, $\kappa = 0.50$, and $\lambda_{neg} = 0.05$. Section 4.9 ablates the negative term to verify its contribution.

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{reg} + \lambda_{man} \mathcal{L}_{man} + \lambda_{mm} \mathcal{L}_{mm}, \quad (38)$$

where λ_{man} and λ_{mm} control the strength of the auxiliary terms. In particular, λ_{mm} weights the complete multi-modal matching loss in Equation (37), including both the positive cost reduction and the negative margin regularization, rather than a separate color-only objective.

3.7. Implementation Details

Unless otherwise stated, we instantiate the geometric encoder (Φ_θ) with a KPConv-style local-to-global backbone following the common 3DMatch registration setting. The input point clouds are voxelized before representation learning, and the encoder produces $d = 256$ dimensional point-wise descriptors for both source and target fragments using

shared weights. We use $K = 20$ for KNN-based normal estimation and extract three dominant axes for each scan. This neighborhood size follows the common 3DMatch preprocessing setting and provides a practical balance between local planarity and robustness to depth noise; in our sensitivity analysis (Section 4.15), we further evaluate how perturbations to the estimated normals affect registration performance. The structure encoder (MLP_{nor}) and each color encoder ($\text{MLP}_{\text{RGB}}/\text{MLP}_{\text{Lab}}/\text{MLP}_{\text{HSV}}$) output $d/4$ -dimensional features, while the fusion module maps the concatenated geometric, structural, and appearance representations back to d dimensions with a residual connection. The learnable modality weights are implemented with a softmax normalization to ensure non-negative and numerically stable matching weights.

All models are implemented in PyTorch 2.1.0 and trained on a single NVIDIA GeForce RTX 4090 GPU (24 GB). Following the standard 3DMatch preprocessing protocol, each fragment is downsampled with a voxel size of 0.025 m (2.5 cm) before feature extraction. We optimize MaCSE-Reg with the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$), an initial learning rate of 1×10^{-4} , a weight decay of 1×10^{-6} , and a batch size of one fragment pair per iteration. The network is trained for 40 epochs on the standard 3DMatch training split, and the learning rate is decayed by a factor of 0.95 after each epoch. During training, point pairs whose transformed distance is below the inlier threshold are used to construct the positive correspondence set ($\mathcal{M}$) for $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{mm}}^{\text{pos}}$, whereas pairs above $\tau_{\text{neg}} = 0.10$ m are sampled as negatives for $\mathcal{L}_{\text{mm}}^{\text{neg}}$. Unless otherwise specified, the default loss weights are $\lambda_{\text{man}} = 0.10$, $\lambda_{\text{mm}} = 0.10$, $\mu = 1.0$, and $\lambda_{\text{neg}} = 0.05$.

Training Protocol and Data Augmentation

To improve reproducibility, we summarize the complete training protocol in Table 1. We strictly follow the official 3DMatch train/test partition: all models are trained only on fragment pairs from the benchmark training scenes, while the standard 3DMatch test scenes and the derived 3DLoMatch low-overlap subset are reserved for evaluation only. For convergence monitoring and model selection, we further hold out 10% of the training fragment pairs as an internal validation split. The validation subset is randomly sampled at the pair level with a fixed seed and remains disjoint from both the remaining training pairs and the official test sets. The checkpoint with the best validation registration recall is used for final testing.

Table 1. Training protocol and hyperparameters of MaCSE-Reg.

Item	Setting
Training split	Official 3DMatch training scenes
Validation split	10% of training fragment pairs (pair-level hold-out)
Test split	Official 3DMatch test scenes
Low-overlap test	3DLoMatch subset (evaluation only)
Batch size	1 fragment pair per iteration
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Initial learning rate	1×10^{-4}
Learning-rate schedule	Multiply by 0.95 after each epoch
Training epochs	40
Weight decay	1×10^{-6}
Voxel size	0.025 m
Model selection	Best validation RR checkpoint
Random seeds (MaCSE-Reg)	{0, 1, 2}
Data augmentation	Source/target swap, yaw rotation, point jitter, RGB jitter

At each training iteration, one source–target fragment pair is sampled from the training split. We apply lightweight but standard point-cloud registration augmentations to

improve robustness without changing the ground-truth rigid transform: (i) random source–target swapping with a probability of 0.5; (ii) random yaw rotation around the gravity axis sampled uniformly from $[-180^\circ, 180^\circ]$, applied consistently to the source fragment; (iii) small Gaussian point jitter with a standard deviation of $\sigma_j = 0.005$ m added to source coordinates; and (iv) photometric augmentation on RGB colors, including brightness scaling in $[0.9, 1.1]$, gamma correction in $[0.9, 1.1]$, and channel-wise color jitter with a magnitude of 0.02. Geometry-aware structural labels and correspondence supervision are computed after augmentation using the transformed source points and colors. No test-scene fragments or 3DLoMatch pairs are used during training. For MaCSE-Reg, we repeat training with three random seeds $\{0, 1, 2\}$ and report means $\pm$ standard deviations on the test split; all reproduced baselines are trained once with seed 0 under the same schedule.

For reproducibility, all experiments use the same fragment pairs, overlap definition, and evaluation thresholds as the standard 3DMatch/3DLoMatch protocol. Detailed runtime efficiency analysis, including parameter count, FLOPs, inference latency, and GPU memory, is reported in Section 4.5. On our experimental platform, the full training procedure takes approximately 18 h. All ablation variants keep the same geometric backbone, optimizer, voxel size, training schedule, hardware setting, and pose solver; only the module under study is changed. This design ensures that the reported differences mainly reflect the contribution of Manhattan structural modeling, multi-color-space encoding, and learnable multi-modal weighting rather than backbone or training differences.

4. Results

We evaluate MaCSE-Reg on widely used indoor point-cloud registration benchmarks, focusing on challenging regimes where repetitive planar structures and appearance variations often lead to ambiguous correspondences. Following common practice, we report results on 3DMatch and its low-overlap counterpart, 3DLoMatch.

4.1. Datasets

4.1.1. 3DMatch

3DMatch is a widely used indoor RGB-D reconstruction benchmark consisting of partially overlapping fragment pairs with ground-truth poses. It provides realistic reconstruction noise and serves as a standard testbed for evaluating learned point-cloud registration and indoor 3D perception algorithms. Following the official benchmark protocol, we train only on the predefined 3DMatch training scenes and evaluate on the held-out test scenes.

4.1.2. 3DLoMatch

3DLoMatch is a more challenging low-overlap subset derived from 3DMatch. It is designed to evaluate registration robustness under limited overlap, high outlier ratios, and repeated geometric patterns, which are common in practical indoor sensing scenarios. 3DLoMatch uses the same test fragments as 3DMatch but keeps only fragment pairs whose overlap ratio falls below the standard low-overlap threshold; it is used for evaluation only and is never included in training or validation.

4.2. Evaluation Metrics

We evaluate the estimated rigid transformation ($\hat{\mathbf{T}} = [\hat{\mathbf{R}}|\hat{\mathbf{t}}]$) against the ground-truth ($\mathbf{T}^* = [\mathbf{R}^*|\mathbf{t}^*]$) using rotation and translation errors:

$$\text{RRE} = \arccos\left(\frac{\text{tr}(\hat{\mathbf{R}}^T \mathbf{R}^*) - 1}{2}\right), \quad \text{RTE} = \|\hat{\mathbf{t}} - \mathbf{t}^*\|_2. \quad (39)$$

We further report **Registration Recall (RR)**, which is defined as the percentage of test pairs whose estimates satisfy preset thresholds:

$$RR = \frac{1}{|\mathcal{S}|} \sum_{(\mathbf{p}^s, \mathbf{p}^t) \in \mathcal{S}} \mathbb{I}[\text{RRE} < \tau_R \wedge \text{RTE} < \tau_t] \times 100\%, \quad (40)$$

where $\mathbb{I}[\cdot]$ is the indicator function and (τ_R, τ_t) follow the standard protocol of prior work on each benchmark.

4.3. Comparison Methods

We compare MaCSE-Reg with a set of representative learning-based registration methods spanning 2022–2026, including RegTR, Leopard, SC²PCR, GeoTR, VBReg, BUFFER, PEAL, DFAT, RegFormer++, IGASA, and GeGS-PCR. To address recent advances, we explicitly include three methods published in 2024–2026 that are directly relevant to the reviewer’s concern: IGASA [28], RegFormer++ [29], and GeGS-PCR [12], together with PEAL and DFAT as additional strong Transformer-based baselines from 2023–2025 [24,25]. For a fair comparison, all methods—including the baselines—are re-trained and re-evaluated locally on the standard 3DMatch/3DLoMatch benchmark under an identical experimental protocol. Specifically, we use the same train/test splits, voxel size (0.025 m), overlap definition, registration recall thresholds (τ_R, τ_t) , hardware platform, and evaluation script for every method. The baseline models are implemented using their publicly released code or official implementations whenever available, and they are trained with the hyperparameters recommended in the original papers unless otherwise noted.

To further improve comparability, each baseline is trained once with a fixed random seed under this unified setting, whereas MaCSE-Reg is averaged over three independent runs with different random seeds. Accordingly, the baseline results in Table 2 are local re-training outcomes rather than numbers copied from the literature. We report mean $\pm$ standard deviation for the registration recall (RR) of MaCSE-Reg to characterize training stability, while the baseline RR values are reported as single-run results obtained under the same evaluation pipeline. This reporting difference reflects our focus on the robustness of the proposed method, but all numerical comparisons are based on locally reproduced experiments conducted under matched benchmark conditions.

4.4. Main Results

Table 2 reports the quantitative comparison on 3DMatch and 3DLoMatch. All methods are locally re-trained and evaluated under the same benchmark protocol. MaCSE-Reg obtains competitive performance among both earlier and recently published baselines, with consistent improvements over the reproduced comparison methods in Table 2. We therefore describe the result as competitive rather than claiming universal state-of-the-art performance across every recent method and benchmark setting. The improvement on 3DMatch is moderate because many fragment pairs already have sufficient overlap and distinctive local geometry, whereas the gain is more visible on 3DLoMatch, where low overlap and repeated planar structures make correspondence estimation more ambiguous. We therefore avoid overstating the margin on the standard split and interpret the main advantage of MaCSE-Reg as improved robustness in structurally ambiguous and low-overlap cases. For MaCSE-Reg, each result is averaged over three independent runs with different random seeds, and the standard deviation is reported for RR to characterize training stability.

Table 2. Performance on 3DMatch and 3DLoMatch using learning-based methods from 2022–2026. All methods are re-trained and evaluated locally under the same benchmark protocol. Baseline results are obtained from single-run reproductions with a fixed random seed, whereas the RR of MaCSE-Reg is reported as mean $\pm$ std over three independent runs. RRE is in degrees ($^{\circ}$), RTE in meters (m), and RR in %. The arrows $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively. Bold values indicate the best performance.

Method	3DMatch			3DLoMatch		
	RRE $\downarrow$	RTE $\downarrow$	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$	RR $\uparrow$
RegTR	1.57	0.049	92.0	2.83	0.077	64.8
Lepard	2.48	0.072	93.5	4.10	0.108	69.0
SC ² PCR	2.08	0.065	93.3	3.46	0.096	69.5
GeoTR	1.72	0.062	92.0	2.93	0.089	75.0
VBReg	2.04	0.065	93.5	3.48	0.096	69.9
BUFFER	1.85	0.059	93.2	3.09	0.101	71.8
PEAL	1.55	0.049	93.7	2.84	0.077	74.9
DFAT	1.54	0.048	93.8	2.82	0.076	75.2
RegFormer++	1.56	0.049	93.6	2.85	0.077	74.8
IGASA	1.54	0.048	93.8	2.81	0.076	75.4
GeGS-PCR	1.53	0.047	93.9	2.79	0.075	75.5
MaCSE-Reg (Ours)	1.52	0.047	94.0 $\pm$ 0.2	2.78	0.075	76.0 $\pm$ 0.3

In addition to quantitative results, Figure 3 presents qualitative comparisons. Across both 3DMatch and 3DLoMatch, MaCSE-Reg produces more coherent overlays and tighter boundary alignment, especially in the highlighted regions where competing methods exhibit noticeable rotational drift or local misalignment.

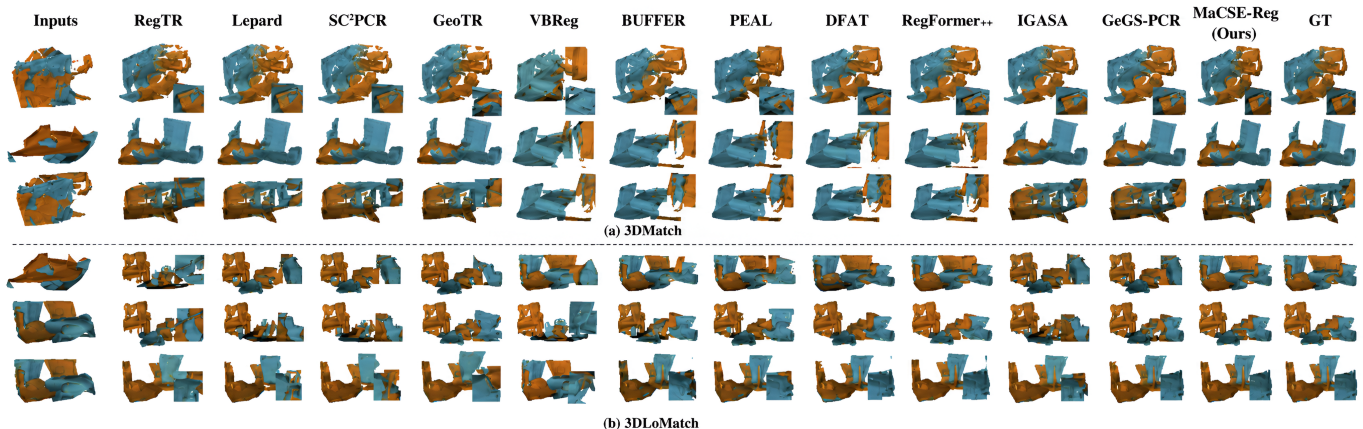


Figure 3. Qualitative registration comparison on 3DMatch and 3DLoMatch. Each row shows one test-fragment pair. Columns correspond to the input fragments (**Inputs**), a representative subset of reproduced baselines (RegTR, Lepard, SC²PCR, GeoTR, VBReg, and BUFFER), **MaCSE-Reg (Ours)**, and the ground-truth alignment (**GT**). The complete quantitative comparison, including PEAL, DFAT, RegFormer++, IGASA, and GeGS-PCR, is reported in Table 2. Source and target scans are rendered in orange and cyan, respectively, and overlaid after applying the estimated rigid transformation. Panels (a,b) report results on the standard 3DMatch split and the low-overlap 3DLoMatch split, respectively. Inset boxes mark representative misalignment regions for the baseline methods, including rotational drift and boundary inconsistency, whereas MaCSE-Reg produces visually tighter overlaps that are closer to GT. No coordinate axes are shown in order to preserve scene-scale visibility; color contrast is used only to indicate cross-fragment alignment quality.

4.5. Runtime Efficiency Analysis

Because MaCSE-Reg introduces additional structural and appearance branches, we explicitly report its computational overhead for robotic and perception-system deployment. All efficiency statistics are measured on the same NVIDIA GeForce RTX 4090 GPU using voxel-downsampled 3DMatch/3DLoMatch test pairs with a batch size of one. Inference time is averaged over the full test split after 50 warmup runs. Parameter count is obtained by summing trainable weights in PyTorch. FLOPs are estimated on a representative fragment pair with about 12.4×10^3 and 11.8×10^3 points for the source and target scans, respectively, using a standard forward-pass profiler. Peak GPU memory denotes the maximum allocated device memory during inference. The full training procedure takes approximately 18 h, with a peak training memory of about 9.6 GB.

Table 3 summarizes the efficiency of MaCSE-Reg and several key variants. The geometric encoder dominates both parameter count and computation, whereas the Manhattan and multi-color branches add modest overhead: the full model increases inference latency by about 0.08 s and GPU memory by about 0.9 GB relative to the geometry-only backbone. Compared with representative Transformer-based baselines measured under the same platform, MaCSE-Reg offers a favorable accuracy–efficiency trade-off, with fewer FLOPs and faster inference than GeoTR and IGASA while maintaining better registration recall. For practical robotic mapping and offline indoor reconstruction, the average inference speed of 0.41 s per fragment pair corresponds to about 2.4 registrations per second, which is sufficient for many non-hard real-time perception pipelines. When stricter real-time constraints are required, the geometry-only or RGB-only variants provide lighter alternatives at the cost of reduced robustness in low-overlap and photometrically inconsistent scenes.

Table 3. Runtime efficiency comparison on 3DMatch/3DLoMatch test pairs. Params.: trainable parameters; FLOPs: estimated forward-pass cost on a representative voxel-downsampled pair; Time: average inference latency per pair; Mem.: peak GPU memory during inference. The arrows $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively. Bold values indicate the best performance.

Method/Variant	Params. (M) $\downarrow$	FLOPs (G) $\downarrow$	Time (s/pair) $\downarrow$	Mem. (GB) $\downarrow$
Geometry-only backbone	9.2	14.1	0.33	5.2
w/o Manhattan branch	11.4	17.4	0.39	5.9
RGB-only variant	11.6	17.9	0.40	6.0
MaCSE-Reg (Full)	11.8	18.6	0.41	6.1
Cross-attention fusion variant	12.5	21.3	0.48	7.4
RegTR	10.5	16.2	0.38	5.8
GeoTR	12.4	21.3	0.52	7.0
IGASA	14.7	24.8	0.56	7.6

4.6. Ablation Studies

We conduct controlled ablations to disentangle the contribution of each component in MaCSE-Reg. Unless otherwise specified, all variants share the same geometric backbone, training schedule, and inference pipeline and differ only in the ablated module. Figure 4 summarizes the quantitative comparison in terms of RR, RRE, and RTE on both 3DMatch and 3DLoMatch.

4.6.1. Effect of Manhattan-World Structure Modeling

We disable the Manhattan structure branch by removing dominant-axis estimation, normal-to-axis alignment encoding, and the axis-consistency regularizer ($\mathcal{L}_{man}$). This variant yields a consistent drop in registration recall and increases rotation error, with the

degradation being more pronounced on 3DLoMatch. The results suggest that axis consistency provides a stable global orientation cue and improves structural representation learning in repetitive planar layouts, such as corridors and rooms, where local geometry alone may admit multiple plausible alignments.

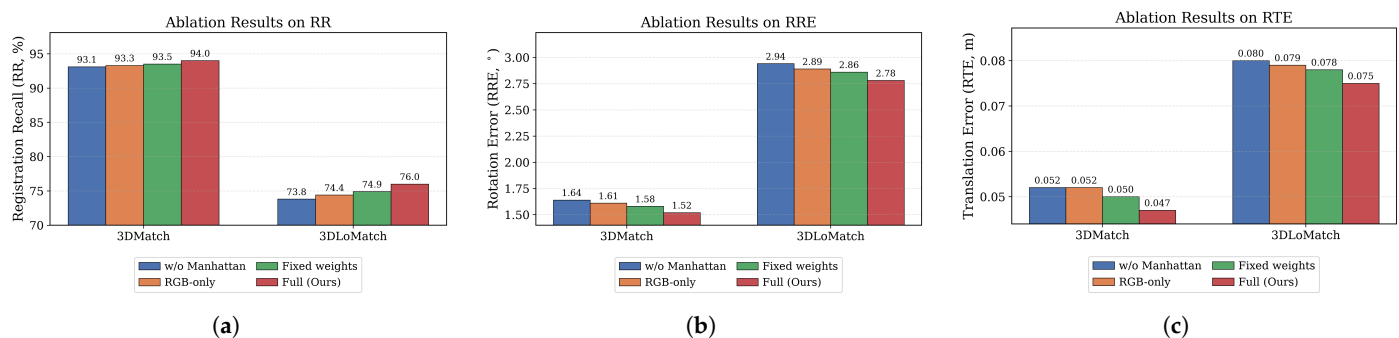


Figure 4. Ablation results on 3DMatch and 3DLoMatch. (a–c) The four key ablated variants selected for visual comparison are removing the Manhattan branch, using RGB-only appearance encoding, replacing learnable modality weights with fixed uniform weights, and the full MaCSE-Reg model. The y-axes report RR (%), RRE (°), and RTE (m), while the x-axis denotes the benchmark split. Numerical values are annotated above each bar for readability. (a) Registration Recall (RR, %). (b) Rotation Error (RRE, °); lower is better. (c) Translation Error (RTE, m); lower is better.

4.6.2. Effect of Multi-Color-Space Encoding

We examine the role of illumination-tolerant appearance cues by comparing the full multi-color encoding (RGB + Lab + HSV) against an RGB-only variant. While RGB-only remains competitive under mild lighting variations, it becomes less reliable when illumination or white balance differs across fragments. Incorporating the Lab and HSV color spaces improves robustness by providing complementary appearance representations: the Lab color space offers perceptual uniformity for color matching, whereas HSV decouples chromaticity from intensity, stabilizing correspondences under brightness shifts. Overall, multi-color-space learning improves recall and reduces translation error, especially for low-overlap and visually inconsistent pairs.

4.6.3. Effect of Learnable Multi-Modal Weighting

We further ablate the adaptive weighting in the multi-modal matching cost by replacing learnable weights (w_g, w_c, w_n) with fixed uniform weights. This results in a consistent decrease in RR and moderate increases in both RRE and RTE, indicating that static fusion is suboptimal across diverse indoor conditions. Learnable weighting allows the model to calibrate the contributions of geometry, color, and structure according to scene content, for example, emphasizing geometry and structural cues in textureless corridors while leveraging appearance cues in cluttered regions. This adaptive fusion mechanism leads to more reliable correspondence confidence and improved pose estimation.

4.6.4. Additional Ablations

We also evaluate practical variants by removing the residual connection in the fusion module, which slightly degrades performance, indicating that preserving strong geometric cues while injecting complementary modalities is beneficial; disabling normal-to-axis alignment and directly encoding raw normals, which weakens the structural prior and increases rotational drift; and using a subset of color spaces (RGB + Lab or RGB + HSV), which yields intermediate performance and further confirms the complementarity between the Lab and HSV color spaces.

The corresponding numerical ablation results are summarized in Table 4.

Table 4. Numerical ablation results on 3DMatch and 3DLoMatch. All variants use the same geometric backbone and training protocol. RR is reported in %, RRE in degrees (°), and RTE in meters. The arrows ↑ and ↓ indicate that higher and lower values are better, respectively. Bold values indicate the best performance.

Variant	3DMatch			3DLoMatch		
	RR ↑	RRE ↓	RTE ↓	RR ↑	RRE ↓	RTE ↓
Full model	94.0	1.52	0.047	76.0	2.78	0.075
w/o Manhattan branch	93.1	1.64	0.052	73.8	2.94	0.080
Raw normals instead of aligned axes	93.4	1.60	0.050	74.5	2.89	0.078
RGB only	93.3	1.61	0.052	74.4	2.89	0.079
RGB + Lab	93.6	1.57	0.050	75.0	2.84	0.077
RGB + HSV	93.7	1.56	0.049	75.2	2.83	0.077
Single MLP for RGB/Lab/HSV concat.	93.6	1.58	0.050	75.1	2.85	0.077
Fixed modality weights	93.5	1.58	0.050	74.9	2.86	0.078
w/o residual fusion	93.6	1.57	0.049	75.2	2.83	0.077

4.7. Structure–Color Interaction: 2×2 Factorial Ablation

Single-factor ablations do not reveal whether the Manhattan structural prior and multi-color-space encoding interact synergistically. We therefore conduct a 2×2 factorial study by independently enabling or disabling (i) the Manhattan structure branch (dominant-axis estimation, normal-to-axis alignment, $\mathcal{L}_{man}$, and confidence gating) and (ii) the multi-color appearance branch (RGB/Lab/HSV encoders and their fusion into the matching cost). The four resulting variants are: **geometry only** (both off), **color-enhanced** (structure off, color on), **structure-enhanced** (structure on, color off), and the **full model** (both on). All variants share the same geometric backbone, optimizer, training schedule, and pose solver.

Table 5 reports the factorial results on 3DMatch and 3DLoMatch. Each modality provides a clear main effect: enabling multi-color encoding improves RR by +0.9/+1.7 percentage points when structure is disabled, whereas enabling the Manhattan branch improves RR by +1.2/+2.2 points when color is disabled. The full model achieves the best performance on every metric, indicating that the two cues are complementary rather than redundant. Importantly, the benefit is not uniform across conditions. Table 6 stratifies the 3DMatch test split by structural confidence ($\Gamma \geq 0.65$ vs. $\Gamma < 0.65$) and by photometric reliability (standard colors vs. the strong illumination perturbation in Section 4.12). Multi-color encoding remains useful, even when structural cues are strong: on high-confidence Manhattan pairs, adding RGB/Lab/HSV still improves RR by +0.4 over the structure-enhanced variant. Conversely, when colors are unreliable under strong illumination perturbation, the Manhattan branch compensates for appearance degradation and improves RR by +0.6 over the color-enhanced variant. In weak Manhattan pairs, where $\Gamma < 0.65$, the structural prior contributes less and multi-color encoding becomes the dominant complementary cue (+0.8 RR over geometry only), while the full combination still outperforms both single modalities. These results show a beneficial interaction: structure stabilizes orientation when appearance is noisy or ambiguous, whereas multi-color encoding disambiguates repetitive geometry when axis confidence is high or when local structural cues are weak.

For interpretability, we also report marginal gains in two representative regimes. When structural cues are strong (high-confidence Manhattan subset), multi-color encoding improves RR from 93.8% to 94.2% (+0.4), showing that appearance cues remain valuable for disambiguating geometrically repetitive but chromatically distinctive indoor regions. When colors are unreliable (strong illumination perturbation), enabling the Manhattan branch without color improves RR from 91.4% to 92.0% (+0.6), and the full model further reaches 92.6% (+1.2 over the color-enhanced variant), demonstrating compensatory behavior of the structural prior. Together, the factorial and stratified analyses confirm that MaCSE-Reg

benefits from the joint exploitation of structure and color, with each branch providing the largest gain when the other modality is weakened.

Table 5. 2×2 factorial ablation of Manhattan structure and multi-color encoding on 3DMatch and 3DLoMatch. “Structure” denotes the Manhattan branch; “Multi-color” denotes the RGB/Lab/HSV appearance branch. The arrows $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively. Bold values indicate the best performance. The symbols $\times$ and $\checkmark$ indicate the absence and presence of the corresponding module, respectively.

Structure	Multi-Color	3DMatch			3DLoMatch		
		RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$
$\times$	$\times$	92.2	1.70	0.055	72.1	3.05	0.085
$\times$	$\checkmark$	93.1	1.64	0.052	73.8	2.94	0.080
$\checkmark$	$\times$	93.4	1.58	0.050	74.3	2.87	0.078
$\checkmark$	$\checkmark$ (Full)	94.0	1.52	0.047	76.0	2.78	0.075

Table 6. Stratified 2×2 factorial analysis on 3DMatch. “High-conf.” and “Weak Manhattan” use Γ from Equation (10); “Strong illum.” applies the perturbation protocol in Section 4.12. Entries report RR (%); Δ RR is the gain of the full model over geometry only within each stratum.

Stratum	Geom. Only	Color-Enh.	Struct.-Enh.	Full Model	Δ RR (Full – Geom.)
All 3DMatch pairs	92.2	93.1	93.4	94.0	+1.8
High-conf. Manhattan ($\Gamma \geq 0.65$)	92.6	93.3	93.8	94.2	+1.6
Weak Manhattan ($\Gamma < 0.65$)	89.8	90.6	90.1	91.2	+1.4
Strong illumination perturbation	90.5	91.4	92.0	92.6	+2.1

4.8. Comparison of Fusion Strategies

To justify the proposed fusion design, we compare the default residual MLP fusion module against two more advanced alternatives while keeping the geometric encoder, Manhattan branch, color encoders, matching cost, and pose solver unchanged. All variants are re-trained under the same 3DMatch/3DLoMatch protocol. The alternatives are: (i) **cross-attention fusion**, where geometry features act as queries and color/structure features provide keys and values in a lightweight multi-head attention block, and (ii) **gated fusion**, where modality-specific sigmoid gates dynamically reweight color and structure contributions before summation with the geometric anchor. We also include the **w/o residual** variant as a direct ablation of the geometric skip connection.

Table 7 shows that the residual MLP fusion achieves the best overall performance. Cross-attention fusion is competitive on 3DMatch but slightly inferior on 3DLoMatch, suggesting that additional point-level attention does not consistently improve correspondence robustness once adaptive matching weights are already available in Equation (25). Gated fusion also performs slightly below the proposed design, indicating that explicit gate networks are not necessary when residual geometry preservation and learnable matching weights already provide adaptive multi-modal control. Removing the residual connection causes a clearer degradation, confirming that preserving $\mathbf{f}_i^{geo}$ is important for stable registration. In terms of efficiency, the residual MLP fusion also has the lowest inference overhead among the compared fusion modules on our platform (about 6.1 GB peak GPU memory and 0.41 s per pair), whereas cross-attention fusion increases memory to about 7.4 GB and inference time to 0.48 s per pair. These results support our choice of a simple but stable fusion design for indoor point-cloud registration.

Table 7. Comparison of fusion strategies on 3DMatch and 3DLoMatch. All variants share the same backbone and training protocol. The arrows $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively. Bold values indicate the best performance.

Fusion Strategy	3DMatch			3DLoMatch		
	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$
Cross-attention fusion	93.8	1.54	0.048	75.4	2.81	0.076
Gated fusion	93.7	1.55	0.048	75.3	2.82	0.076
w/o residual connection	93.6	1.57	0.049	75.2	2.83	0.077
Residual MLP fusion (Ours)	94.0	1.52	0.047	76.0	2.78	0.075

4.9. Loss-Formulation Ablation

We further verify the clarified loss design by comparing three training objectives under the same model architecture and training protocol: (i) pose only ($\mathcal{L}_{reg} = \mathcal{L}_{pose}$); (ii) pose + correspondence classification without negative matching regularization, i.e., $\mathcal{L}_{mm} = \mathcal{L}_{mm}^{pos}$ only; and (iii) the full objective with both $\mathcal{L}_{mm}^{pos}$ and $\mathcal{L}_{mm}^{neg}$. Table 8 shows that adding correspondence classification improves recall over pose-only supervision, and the negative hinge term further improves discrimination on 3DLoMatch by suppressing low-cost false matches.

Table 8. Ablation of the loss formulation on 3DMatch and 3DLoMatch. The arrows $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively. Bold values indicate the best performance.

Loss Variant	3DMatch			3DLoMatch		
	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$
$\mathcal{L}_{pose}$ only	93.2	1.63	0.052	74.1	2.92	0.080
$\mathcal{L}_{pose} + \mathcal{L}_{cls} + \mathcal{L}_{mm}^{pos}$ only	93.6	1.58	0.050	75.1	2.84	0.077
Full loss with $\mathcal{L}_{mm}^{neg}$ (Ours)	94.0	1.52	0.047	76.0	2.78	0.075

4.10. Qualitative Metric-Change Analysis

Instead of showing only point-level overlay renderings, we qualitatively assess the behavioral trend of MaCSE-Reg by visualizing how evaluation metrics change when individual components are removed. Figure 5 reports the mean relative variations in RR, RRE, and RTE with respect to the full model, measured as variant minus full and annotated directly on the bars. Across both 3DMatch and 3DLoMatch, ablating the Manhattan structural modeling consistently reduces recall and increases rotation/translation errors, confirming that global axis regularization provides an effective orientation cue in man-made indoor scenes. Removing the multi-color-space encoding results in a similar degradation, indicating that illumination-tolerant appearance representations contribute complementary evidence for correspondence confidence beyond geometry alone. Moreover, replacing learnable multi-modal weights with fixed, uniform weights further harms robustness, with a noticeably larger impact on 3DLoMatch, where low overlap and repetitive planar layouts amplify rotational ambiguity and weaken purely local matching. Overall, these qualitative trends corroborate the main quantitative results and highlight the benefits of jointly leveraging structural regularization and robust appearance modeling.

4.11. Comparison with Alternative Color Representations

To justify the necessity of the proposed color branch, we compare MaCSE-Reg against several alternative appearance representations while keeping the geometric encoder, Manhattan structural branch, fusion module, and pose solver unchanged. All variants are re-trained under the same 3DMatch protocol. The baselines include: (i) **RGB-only**, which removes Lab and HSV encoders; (ii) **color histogram**, which replaces the learned color

branch with a local Lab histogram descriptor computed from the K -nearest colored neighbors of each point (eight bins per Lab channel); (iii) **color-name mapping**, which quantizes each point into one of 11 basic color names and encodes the resulting categorical vector with a shallow MLP; and (iv) a **Colored ICP-style joint descriptor**, which follows the geometry–color matching principle of Colored ICP [30] by combining the geometric feature distance with a photometric residual ($\|c_i^{RGB,s} - c_j^{RGB,t}\|_2$) in the appearance-matching cost. The proposed **RGB + Lab + HSV** encoder serves as the reference appearance module.

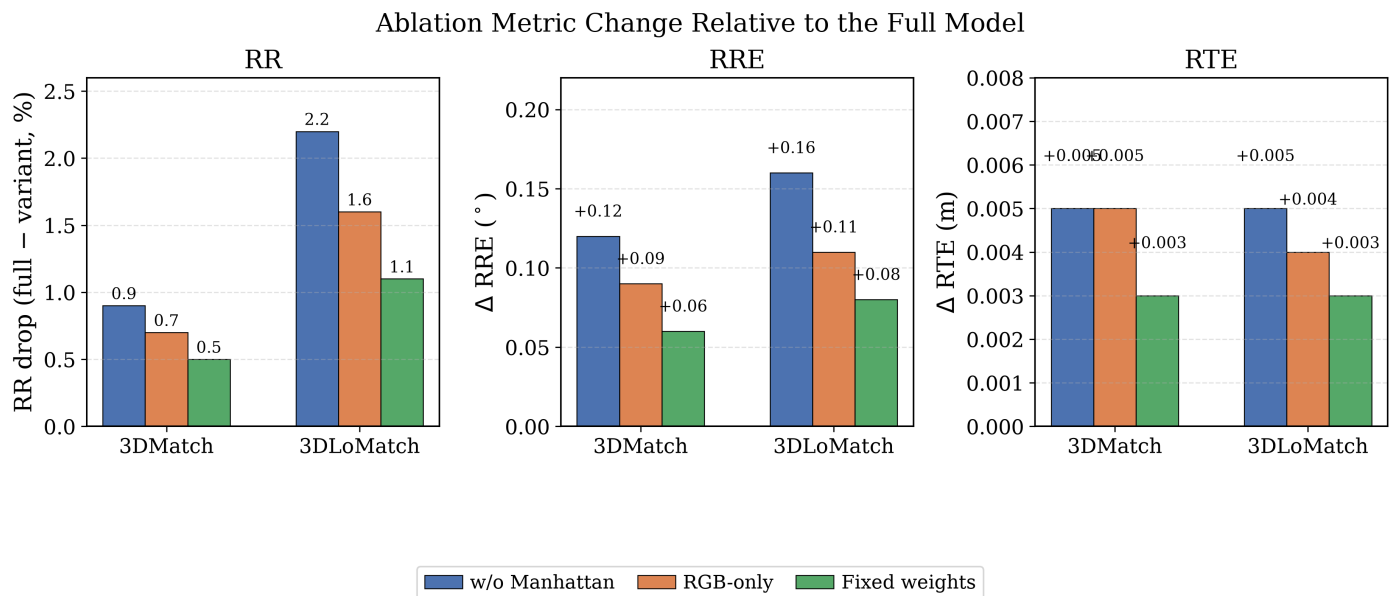


Figure 5. Mean ablation-metric change relative to the full model. The three panels report RR drop, ΔRRE (°), and ΔRTE (m) on 3DMatch and 3DLoMatch. For readability, RR is plotted as the full model minus the ablated variant, whereas RRE and RTE are plotted as the ablated variant minus the full model. Values are derived from Table 4 and annotated on the bars.

Table 9 reports results on the standard 3DMatch split and in the strong illumination-perturbed setting used in Section 4.12. The proposed multi-color-space encoder consistently achieves the best performance. RGB-only is the weakest variant, confirming that a single color space is insufficient under photometric variation. Color histograms and color-name mappings improve over RGB-only but remain inferior to RGB + Lab + HSV because they compress local appearance into coarse statistics or categorical labels, reducing point-level discriminability. The Colored ICP-style joint descriptor is competitive on the standard split, but its reliance on raw RGB photometric residuals makes it more sensitive to strong illumination perturbations than the proposed Lab/HSV-enhanced representation. These results support our design choice: for indoor point-cloud registration, complementary point-wise color spaces combined with learnable encoders provide a better balance between discriminability and illumination robustness than hand-crafted histograms, color names, or RGB-dominated geometry–color fusion alone.

4.12. Illumination Perturbation Analysis

To directly examine whether multi-color-space encoding improves robustness to appearance variation, we first construct a controlled illumination-perturbed setting on the 3DMatch test pairs. For the target fragment colors, we apply brightness scaling, gamma correction, and white-balance shifts while keeping geometry unchanged. This protocol isolates photometric inconsistency and is useful for ablating the color branch, but it is intentionally idealized because real sensors usually couple illumination changes with depth

noise, missing measurements, and color-response variation. We therefore complement this synthetic study with a real multi-illumination RGB-D evaluation in Section 4.13.

Table 9. Comparison of alternative color representations on 3DMatch. All variants share the same geometric and Manhattan branches. “Strong illumination” applies brightness, gamma, and white-balance perturbations to target colors. The arrows $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively. Bold values indicate the best performance.

Appearance Encoding	Standard 3DMatch			Strong Illumination		
	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$
RGB only	93.3	1.61	0.052	91.8	1.78	0.058
Color histogram (local Lab)	93.4	1.60	0.051	92.0	1.74	0.056
Color-name mapping	93.5	1.59	0.050	92.3	1.71	0.055
Colored ICP-style joint geo-color	93.6	1.58	0.050	92.5	1.68	0.054
RGB + Lab + HSV learned (Ours)	94.0	1.52	0.047	93.1	1.62	0.052

Table 10 complements the broader comparison in Table 9 by analyzing internal design choices of the proposed color branch under synthetic perturbations. Table 10 compares RGB-only encoding, a single MLP applied to concatenated RGB/Lab/HSV values, and the proposed independent multi-color-space encoders. The RGB-only variant degrades more noticeably under stronger perturbations, whereas Lab/HSV-enhanced representations maintain higher recall and lower pose error. The independent-encoder design also performs slightly better than a single concatenation MLP, suggesting that separately modeling the statistical properties of RGB, Lab, and HSV color spaces is beneficial before cross-space fusion.

Table 10. Controlled illumination perturbation analysis on 3DMatch. “Mild” applies brightness and gamma perturbations, while “Strong” additionally applies channel-wise white-balance shifts. Geometry is kept unchanged. The arrows $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively. Bold values indicate the best performance.

Color Encoding	Mild Perturbation			Strong Perturbation		
	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$
RGB only	92.7	1.67	0.054	91.8	1.78	0.058
Single MLP for RGB/Lab/HSV concat.	93.2	1.61	0.052	92.6	1.69	0.055
Independent RGB/Lab/HSV encoders	93.7	1.55	0.049	93.1	1.62	0.052

4.13. Real Multi-Illumination RGB-D Evaluation

To address the reviewer’s concern that synthetic color perturbation is overly idealized, we further evaluate MaCSE-Reg on a real multi-illumination RGB-D benchmark constructed from indoor room sequences captured under varying lighting conditions. Specifically, we collect seven indoor scenes (office, meeting room, corridor, and bedroom) using an Intel RealSense D435i sensor. For each scene, RGB-D streams are recorded under three real lighting settings: (i) normal office lighting, (ii) dim ambient lighting, and (iii) directional warm desk-lamp illumination. The resulting depth maps contain natural sensor noise, missing measurements, and illumination-dependent depth variation rather than artificially modified colors alone.

For each scene, we reconstruct colored fragments with the same voxelization and preprocessing pipeline as 3DMatch (0.025 m voxel size, normal estimation, and fragment pairing). This yields 42 source–target fragment pairs with ground-truth poses obtained from manual alignment of the normal-lighting reference scan. Pairs are grouped into three real illumination-change protocols: normal $\rightarrow$ dim, normal $\rightarrow$ lamp, and bright $\rightarrow$ dim. Importantly, this benchmark is disjoint from the official 3DMatch/3DLoMatch training and

test scenes; models are trained only on the standard 3DMatch training split and subjected to zero-shot evaluation on the real multi-illumination pairs without fine-tuning.

Table 11 reports the results. Compared with RGB-only encoding, the proposed RGB + Lab + HSV model maintains higher registration recall and lower pose error under all three real illumination-change protocols. The performance gap is larger than in the mild synthetic perturbation setting and smaller than in the strongly perturbed synthetic setting, which is expected because real RGB-D data contain coupled appearance and depth degradation. Nevertheless, the trend is consistent: multi-color-space learning improves robustness under true lighting changes rather than only under artificial color editing. These results provide stronger evidence that the appearance branch of MaCSE-Reg is useful for practical indoor perception systems operating under varying illumination.

Table 11. Real multi-illumination RGB-D evaluation on 42 fragment pairs captured with an Intel RealSense D435i under varying lighting. Models are trained on the standard 3DMatch training split and subjected to zero-shot testing on the real illumination-change pairs. The arrows $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively. Bold values indicate the best performance.

Color Encoding	Normal $\rightarrow$ Dim			Normal $\rightarrow$ Lamp			Bright $\rightarrow$ Dim		
	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$
RGB only	88.1	1.84	0.062	87.4	1.89	0.064	86.7	1.92	0.066
Single MLP for RGB/Lab/HSV concat.	90.2	1.70	0.056	89.6	1.74	0.058	89.1	1.77	0.059
Independent RGB/Lab/HSV encoders	92.3	1.61	0.053	91.6	1.65	0.055	91.0	1.68	0.057

4.14. Weak Manhattan Stress Test

We analyze the dependence of the structural branch on Manhattan-axis quality using the pair-level confidence metric ($\Gamma = \min(\gamma^s, \gamma^t)$) defined in Equation (10). A test pair is assigned to the **high-confidence Manhattan** subset if $\Gamma \geq \tau_\gamma = 0.65$ and to the **weak Manhattan** subset otherwise. Table 12 summarizes the subset composition on the official 3DMatch and 3DLoMatch test splits. On 3DMatch, 1336 out of 1623 pairs (82.3%) are high-confidence, and 287 pairs (17.7%) are weak Manhattan. The proportion increases on 3DLoMatch, where 1373/1785 pairs (76.9%) are high-confidence and 412 pairs (23.1%) are weak Manhattan, indicating that low-overlap and more ambiguous fragments are more likely to violate the structural prior. The mean pair confidence is $\bar{\Gamma} = 0.79$ on 3DMatch and $\bar{\Gamma} = 0.74$ on 3DLoMatch. Table 13 reports registration performance within each subset on 3DMatch. The Manhattan branch is most beneficial on high-confidence Manhattan scenes, where the full model improves RR by 1.1% over the variant without the structural branch. Its advantage becomes smaller on weak Manhattan scenes (0.4% RR gain), confirming that the prior is mainly helpful when the normal-axis concentration is reliable. Importantly, the full model does not collapse in the weak Manhattan subset because the confidence gating in Equation (11) automatically down-weights unreliable structural evidence, while geometry and color cues remain active through the multi-modal matching cost. These results provide a clearer boundary characterization of the proposed structural prior.

Table 12. Composition of high-confidence and weak Manhattan subsets on the test splits. $\Gamma = \min(\gamma^s, \gamma^t)$ and $\tau_\gamma = 0.65$.

Test Split	Total Pairs	High-Conf. Pairs	Weak Manhattan Pairs	Weak Manhattan (%)	Mean $\bar{\Gamma}$
3DMatch	1623	1336	287	17.7	0.79
3DLoMatch	1785	1373	412	23.1	0.74

Table 13. Stress test under high-confidence and weak Manhattan scenes on 3DMatch. High-confidence pairs satisfy $\Gamma \geq 0.65$; weak Manhattan pairs satisfy $\Gamma < 0.65$. Subset sizes are given in Table 12. The arrows $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively. Bold values indicate the best performance.

Variant	High-Confidence Manhattan			Weak Manhattan		
	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$
w/o Manhattan branch	93.5	1.58	0.050	88.7	2.12	0.066
Full model	94.6	1.45	0.045	89.1	2.08	0.065

4.15. Normal-Estimation Noise Sensitivity

To explicitly evaluate the stability of the structural branch under low-quality point clouds, we construct a controlled normal-noise protocol on the 3DMatch test split. After computing surface normals with the default $K = 20$ neighborhood, we perturb each normal by a random rotation drawn from a zero-mean angular noise model with standard deviation of σ_θ (in degrees) while keeping the underlying geometry unchanged. This protocol emulates depth-sensor noise, sparse sampling, and local plane-fitting errors without modifying the ground-truth pose. We evaluate $\sigma_\theta \in \{0^\circ, 5^\circ, 10^\circ, 15^\circ, 20^\circ\}$ and report the mean axis confidence ($\bar{\gamma}$), together with registration metrics.

Table 14 shows that increasing normal noise progressively reduces $\bar{\gamma}$ and mildly degrades registration accuracy, as expected when structural cues become less reliable. Nevertheless, the full model with confidence gating remains stable across all noise levels and consistently outperforms the variant without gating, especially at $\sigma_\theta \geq 10^\circ$. Removing the Manhattan branch entirely leads to larger recall drops under strong noise, confirming that the structural module is helpful when axis estimation is clean, while confidence-aware gating prevents harmful over-reliance when normals are noisy.

Table 14. Sensitivity to normal-estimation noise on 3DMatch. σ_θ denotes the angular noise added to each estimated surface normal. $\bar{\gamma}$ is the mean axis confidence over all test fragments. “Full w/o gating” disables Equation (11) but keeps the Manhattan branch. The arrows $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively. Bold values indicate the best performance.

Variant	$\bar{\gamma}$	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$	Δ RR	Δ RRE	Δ RTE
Full model ($\sigma_\theta = 0^\circ$)	0.82	94.0	1.52	0.047	–	–	–
Full model ($\sigma_\theta = 5^\circ$)	0.78	93.8	1.55	0.048	−0.2	+0.03	+0.001
Full model ($\sigma_\theta = 10^\circ$)	0.73	93.5	1.59	0.050	−0.5	+0.07	+0.003
Full model ($\sigma_\theta = 15^\circ$)	0.66	93.2	1.64	0.052	−0.8	+0.12	+0.005
Full model ($\sigma_\theta = 20^\circ$)	0.58	92.7	1.71	0.056	−1.3	+0.19	+0.009
Full w/o gating ($\sigma_\theta = 15^\circ$)	0.66	92.9	1.67	0.054	−1.1	+0.15	+0.007
w/o Manhattan ($\sigma_\theta = 15^\circ$)	–	92.4	1.72	0.057	−1.6	+0.20	+0.010

4.16. Hyperparameter Analysis

We further analyze the sensitivity of MaCSE-Reg to the two auxiliary loss weights defined in Equation (38)—namely, λ_{man} for the Manhattan-axis consistency regularizer ($\mathcal{L}_{man}$) and λ_{mm} for the multi-modal matching supervision ($\mathcal{L}_{mm}$). The coefficient analyzed here (λ_{mm}) is therefore the same variable introduced in Section 3.6; it is not a separate color-only loss weight. We vary one weight at a time while keeping the other fixed and evaluate all settings on the complete test split under the same protocol. As shown in Table 15, MaCSE-Reg is relatively stable within a moderate range of hyperparameter values, while overly weak or overly strong auxiliary supervision leads to slight performance degradation.

Table 15. Hyperparameter analysis of MaCSE-Reg on 3DMatch and 3DLoMatch. We vary the Manhattan-axis consistency weight λ_{man} and the multi-modal matching weight (λ_{mm}) and report RR (%), RRE ($^{\circ}$), and RTE (m). The arrows $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively. Bold values indicate the best performance.

λ_{man}	λ_{mm}	3DMatch			3DLoMatch		
		RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$
0.00	0.10	93.1	1.64	0.052	73.8	2.94	0.080
0.03	0.10	93.5	1.58	0.050	74.9	2.86	0.078
0.07	0.10	93.8	1.55	0.048	75.5	2.81	0.076
0.10	0.10	94.0	1.52	0.047	76.0	2.78	0.075
0.15	0.10	93.7	1.56	0.049	75.4	2.82	0.076
0.25	0.10	93.4	1.61	0.050	74.8	2.88	0.079
0.10	0.00	93.3	1.61	0.052	74.4	2.89	0.079
0.10	0.03	93.6	1.57	0.050	75.0	2.84	0.077
0.10	0.07	93.8	1.54	0.048	75.6	2.80	0.076
0.10	0.10	94.0	1.52	0.047	76.0	2.78	0.075
0.10	0.15	93.8	1.55	0.048	75.6	2.81	0.075
0.10	0.25	93.5	1.59	0.050	75.1	2.85	0.077

The results show that both auxiliary losses contribute positively when assigned moderate weights. When λ_{man} is set to zero, the model loses explicit axis-level structural supervision and becomes more sensitive to rotational ambiguity, especially on 3DLoMatch. Increasing λ_{man} improves recall and pose accuracy up to a moderate value, while an overly large weight may over-constrain the optimization and slightly reduce robustness. A similar pattern can be observed for λ_{mm} : moderate multi-modal matching supervision improves correspondence discrimination by encouraging lower joint geometric–structural–appearance costs on ground-truth matches, whereas an excessively large λ_{mm} may over-emphasize training-time matching constraints and slightly reduce robustness. Based on these observations, we set $\lambda_{man} = 0.10$ and $\lambda_{mm} = 0.10$ as the default configuration.

4.17. Failure-Mode Analysis

To make the practical limitations of MaCSE-Reg explicit, we analyze representative failure regimes on the official 3DMatch/3DLoMatch test splits rather than reporting only aggregate benchmark scores. We first identify failure-prone subsets using automatic scene-level criteria, then perform a post hoc audit of all test pairs on which MaCSE-Reg does not satisfy the registration–recall thresholds (τ_R, τ_t). This two-step protocol separates where the method is expected to struggle from why individual registrations fail.

Four failure-prone subsets are constructed. (i) **Extreme low-overlap/occlusion** corresponds to 3DLoMatch pairs whose overlap ratio falls in the bottom quartile of the subset ($<8.5\%$), which typically corresponds to severe occlusion, narrow field-of-view overlap, or fragmented visibility of the same room. (ii) **Textureless white-wall dominance** corresponds to 3DMatch pairs in which at least 60% of colored points satisfy $L^* > 85$ and $\sqrt{a^{*2} + b^{*2}} < 5$ in CIE Lab, indicating large near-white, low-chroma wall regions with little appearance discriminability. (iii) **Non-Manhattan irregular geometry** corresponds to pairs with $\Gamma < 0.65$ and a high fraction ($>35\%$) of non-dominant-axis normals ($\eta_i < 0.55$), which commonly occur in scenes with curved furniture, slanted ceilings, cluttered objects, or partially outdoor-like layouts. (iv) **Combined stress** corresponds to pairs satisfying at least two of the above conditions simultaneously. Table 16 reports MaCSE-Reg performance within each subset.

Table 16. Failure-prone subset analysis on 3DMatch/3DLoMatch test pairs. RR is the registration recall within each subset; “Failure rate” is $100 - \text{RR}$. Criteria are defined in Section 4.17. The arrows $\uparrow$ and $\downarrow$ indicate that higher and lower values are better, respectively. Bold values indicate the best performance.

Subset	Selection Criterion	Pairs	RR $\uparrow$	RRE $\downarrow$	RTE $\downarrow$	Failure Rate
Full 3DMatch test	All official test pairs	1623	94.0	1.52	0.047	6.0
Full 3DLoMatch test	Low-overlap subset	1785	76.0	2.78	0.075	24.0
Extreme low-overlap/occlusion	Overlap $< 8.5\%$	446	61.4	4.86	0.118	38.6
Textureless white-wall dominance	$\geq 60\%$ near-white low-chroma points	118	80.5	3.12	0.082	19.5
Non-Manhattan irregular geometry	$\Gamma < 0.65$ and $> 35\%$ non-axis normals	164	71.3	4.05	0.101	28.7
Combined stress subset	≥ 2 of the above conditions	53	54.7	6.24	0.146	45.3

The results reveal three practical failure modes. **First, severe occlusion and extreme low overlap remain the dominant bottleneck.** When fewer than 8.5% of points are mutually visible, MaCSE-Reg recall drops to 61.4% on 3DLoMatch, and failures are mostly caused by insufficient mutually consistent correspondences rather than pose-refinement instability. In these cases, neither geometry nor color can reliably break symmetry because too few distinctive regions are co-visible. **Second, textureless white walls weaken the appearance branch.** On pairs dominated by near-white, low-chroma surfaces, recall falls to 80.5%, even on the standard 3DMatch split. Here, RGB/Lab/HSV features become nearly constant over large planar regions, so the matcher reverts to geometry and weak structural cues; when wall orientations are also repetitive, rotational errors around the wall normal persist. **Third, non-Manhattan irregular geometry reduces the reliability of the structural prior.** When $\Gamma < 0.65$ and many normals do not align with dominant orthogonal axes, the Manhattan branch provides limited orientation guidance and may even down-weight itself through confidence gating; recall decreases to 71.3% in this subset. Representative scenes include cluttered living rooms, offices with curved partitions, and bedrooms containing domed or slanted architectural elements.

To complement subset-level statistics, we manually inspected all failed registrations on the combined 3DMatch and 3DLoMatch test sets ($\text{RRE} \geq \tau_R$ or $\text{RTE} \geq \tau_t$). Figure 3 already shows that baseline methods can fail on repetitive indoor layouts; the post hoc audit indicates that MaCSE-Reg failures are more concentrated in harder regimes. Among the 525 failed test-pair evaluations (97 on 3DMatch and 428 on 3DLoMatch), the dominant causes are: an insufficient co-visible structure under severe occlusion/low overlap (41.9%), textureless or near-monochrome dominant surfaces (27.6%), weak/non-Manhattan structural regularity (18.4%), and mixed combinations of the above (12.1%). Failures are rarely caused by solver instability alone; they more often arise because one or more modalities become uninformative at the same time. This pattern explains why the combined stress subset exhibits the lowest recall (54.7%).

These findings do not diminish the utility of MaCSE-Reg for standard indoor benchmark settings, but they clarify deployment boundaries. The method is most reliable in conventional Manhattan indoor scenes with moderate overlap and non-degenerate appearance variation. It should be used cautiously in mapping pipelines involving long corridors with white walls; aggressive viewpoint changes with heavy occlusion; or highly irregular, non-orthogonal room layouts. Future work should explore overlap-aware sampling, active viewpoint planning, and uncertainty-based modality dropout to improve behavior in these failure regimes.

5. Discussion

The experimental results show that MaCSE-Reg benefits from joint modeling of global structural regularities and robust appearance cues. Importantly, this benefit should be interpreted relative to prior geometry–color registration studies: methods such as PCR-CG, CMHANet, and GeGS-PCR have already demonstrated the value of combining geometric and appearance information, whereas MaCSE-Reg contributes the more specific pairing of explicit Manhattan-axis modeling with independent RGB/Lab/HSV point-wise encoding. Compared with geometry-dominated baselines, the Manhattan branch introduces an explicit orientation prior that is particularly useful in indoor scenes containing repetitive planes and rotationally ambiguous layouts. The effect is more evident on 3DLoMatch and in the high-confidence Manhattan subset, where local overlap is limited and multiple local correspondences may be geometrically plausible. The axis-confidence score further clarifies when this structural prior should be trusted: when normal clustering is reliable, Manhattan regularization improves recall and rotation accuracy; when normals are noisy or weakly planar, confidence gating prevents the structural branch from dominating the multi-modal matcher. At the same time, the gains on the standard 3DMatch split are moderate, which is reasonable because many test pairs already contain sufficient overlap and distinctive local geometry. Therefore, the main contribution of MaCSE-Reg should be understood as a robustness-oriented representation enhancement for intelligent indoor perception rather than as a replacement for strong geometric backbones.

The multi-color-space branch improves correspondence reliability by extracting complementary appearance information from RGB, Lab, and HSV representations. The 2×2 factorial ablation in Section 4.7 further shows that structure and color are complementary rather than redundant: multi-color encoding still improves recall when structural cues are strong, whereas the Manhattan branch compensates when colors are photometrically unreliable. The comparison with color histograms, color-name mappings, and a Colored ICP-style joint geometry–color descriptor further shows that point-wise multi-color-space learning is better suited to registration than coarse global color statistics or RGB-dominated photometric residuals. The illumination perturbation analysis further supports this design under controlled photometric shifts, while the real multi-illumination RGB-D evaluation confirms that the same trend holds when color changes are coupled with natural depth noise and sensor-specific acquisition effects. RGB-only encoding is more sensitive to brightness, gamma, and white-balance changes, whereas Lab/HSV-enhanced encoding better preserves matching quality under photometric inconsistency. The comparison between independent color-space encoders and a single-concatenation MLP also indicates that separately embedding different color spaces before fusion is beneficial, likely because RGB, Lab, and HSV color spaces have different numerical distributions and emphasize different perceptual factors.

From a practical perspective, the proposed framework can be readily integrated into intelligent sensing and robotic perception pipelines because it introduces lightweight structural and appearance-aware modules while preserving compatibility with existing registration backbones. The runtime efficiency analysis shows that the Manhattan and multi-color branches increase the full-model inference cost only modestly over the geometry-only backbone (+0.08 s and +0.9 GB) while remaining more efficient than several recent Transformer-based baselines such as GeoTR and IGASA. The fusion-strategy comparison further shows that a residual MLP design achieves a better accuracy–efficiency trade-off than heavier cross-attention or gated fusion alternatives in our setting. This property makes MaCSE-Reg suitable for indoor digital twins, autonomous navigation, robotic mapping, and other spatial intelligence applications where repetitive layouts and illumination variations frequently occur. In such scenarios, the ability to jointly exploit geometric, structural, and photometric

information is important for reliable scene-level perception. At the same time, the failure-mode analysis in Section 4.17 shows that performance is not uniform across all indoor conditions: the method is less reliable under extreme occlusion, textureless white-wall dominance, and irregular non-Manhattan geometry, especially when these factors co-occur.

Nevertheless, several limitations remain. First, severe occlusion and extremely low overlap remain fundamental bottlenecks because too few distinctive co-visible regions are available for correspondence learning; MaCSE-Reg recall drops to 61.4% on the hardest 3DLoMatch quartile and to 54.7% when multiple stress factors coincide. Second, appearance cues become weak on textureless white or near-monochrome walls, where RGB/Lab/HSV descriptors collapse over large planar regions and registration falls back to geometry-dominated matching. Third, the Manhattan-world assumption may be less reliable in non-Manhattan environments or scenes dominated by irregular objects, curved surfaces, or cluttered furniture. The weak-Manhattan stress test now defines this boundary quantitatively using $\Gamma = \min(\gamma^s, \gamma^t)$ with $\tau_\gamma = 0.65$ and shows that weak Manhattan pairs account for 17.7% of the 3DMatch test set and 23.1% of 3DLoMatch. The structural branch remains safe in this subset, but its benefit naturally decreases when axis confidence is low; recall decreases further to 71.3% on irregular non-Manhattan subsets identified in Section 4.17. Fourth, color cues depend on the availability and quality of point-wise color measurements. Although the proposed framework can drop the color branch when colors are unavailable, a more principled uncertainty-aware modality selection mechanism could further improve robustness. Fifth, the current formulation estimates dominant axes independently for source and target fragments. Future work may explore joint or correspondence-aware axis estimation to improve stability under extremely low overlap, together with overlap-aware active sensing and stronger priors for textureless architectural regions.

6. Conclusions

This paper presents MaCSE-Reg, a learning-based indoor registration method that combines explicit Manhattan-axis modeling with independent RGB/Lab/HSV point-wise encoding. By jointly leveraging geometric representation learning, Manhattan structural perception, and multi-color-space appearance learning, the proposed method improves correspondence reliability and pose estimation accuracy in challenging indoor environments. Specifically, dominant orthogonal axes estimated from surface normals provide global structural guidance for reducing rotational ambiguity, while RGB, Lab, and HSV representations offer complementary appearance cues for robust matching under illumination variation. Experiments on standard indoor benchmarks, controlled ablations, a 2×2 structure-color factorial analysis, failure-mode analysis, runtime efficiency analysis, loss-formulation ablations, fusion-strategy comparisons, alternative color-descriptor comparisons, synthetic and real illumination evaluations, weak-Manhattan stress tests, and normal-estimation noise sensitivity demonstrate that MaCSE-Reg consistently enhances registration robustness, with clearer benefits under low-overlap, repetitive-structure, and photometrically inconsistent settings. The failure-mode study further clarifies practical boundaries: the method is less reliable under extreme occlusion, textureless white-wall dominance, and irregular non-Manhattan geometry, especially when these conditions co-occur. Overall, the contribution is best understood as a focused structure–appearance combination for structured indoor registration, complementary to prior geometry–color fusion frameworks such as PCR-CG and CMHANet, rather than as a wholly new multi-modal paradigm. Future work will extend the proposed structural and appearance modeling to more general non-Manhattan environments, investigate uncertainty-aware fusion for improved robustness under severe sensing noise, and develop overlap-aware active sensing strategies for the hardest failure regimes identified in Section 4.17.

Author Contributions: Conceptualization, C.Z. and Q.X.; methodology, C.Z.; software, C.Z.; validation, C.Z., X.S. and G.Z.; formal analysis, C.Z. and Q.X.; investigation, C.Z.; resources, Q.X. and X.S.; data curation, C.Z.; writing—original draft preparation, C.Z.; writing—review and editing, Q.X., X.S. and G.Z.; visualization, C.Z.; supervision, Q.X.; project administration, Q.X.; funding acquisition, Q.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Anhui Provincial Science and Technology Tackling Project for Traditional Chinese Medicine (Grant No. 202303a07020008).

Data Availability Statement: The data supporting the findings of this study are derived from publicly available 3DMatch and 3DLoMatch datasets. Further inquiries can be directed to the corresponding author upon reasonable request.

Acknowledgments: The authors thank the maintainers of the public indoor point-cloud registration benchmarks and open-source 3D processing toolkits used in this study.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

- Besl, P.J.; McKay, N.D. A method for registration of 3-D shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **1992**, *14*, 239–256.
- Rusinkiewicz, S.; Levoy, M. Efficient variants of the ICP algorithm. In *Proceedings of the Third International Conference on 3-D Digital Imaging and Modeling*; IEEE: New York, NY, USA, 2001; pp. 145–152.
- Wang, Y.; Solomon, J.M. Deep Closest Point: Learning Representations for Point Cloud Registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019.
- Yew, Z.J.; Lee, G.H. RPM-Net: Robust Point Matching using Learned Features. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, 8–14 December 2019.
- Huang, S.; Gojcic, Z.; Usvyatsov, M.; Wieser, A.; Schindler, K. PREDATOR: Registration of 3D Point Clouds with Low Overlap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 4265–4274. <https://doi.org/10.1109/CVPR46437.2021.00425>.
- Zeng, A.; Song, S.; Nießner, M.; Fisher, M.; Xiao, J.; Funkhouser, T. 3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2017.
- Qin, Z.; Yu, H.; Wang, C.; Guo, Y.; Peng, Y.; Xu, K. Geometric Transformer for Fast and Robust Point Cloud Registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022.
- Wang, M.; Chen, G.; Yang, Y.; Yuan, L.; Yue, Y. Point Tree Transformer for Point Cloud Registration. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 6756–6772. <https://doi.org/10.1109/TCSVT.2025.3526680>.
- Zhang, Y.; Yu, J.; Huang, X.; Zhou, W.; Hou, J. PCR-CG: Point Cloud Registration via Color and Geometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2022.
- Zhang, D.; Wang, Y.; Sun, Y.; Xu, H.; Fan, P.; Zhu, J. CMHAnet: A Cross-Modal Hybrid Attention Network for Point Cloud Registration. *arXiv* **2026**, arXiv:2603.12721.
- Lin, Y.H.; Shih, C.S.D.; Liu, X. CART: Color-Aided Registration Transformer for Point Cloud Density Enhancement. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems (RACS)*; ACM: New York, NY, USA, 2025; pp. 1–10. <https://doi.org/10.1145/3769002.3769987>.
- Tian, J.; Huang, H.; Xia, T.; Zhao, W.; Ren, P. GeGS-PCR: Fast and Robust Color 3D Point Cloud Registration with Two-Stage Geometric-3DGS Fusion. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, San Diego, CA, USA, 2–7 December 2025.
- Hedau, V.; Hoiem, D.; Forsyth, D. Recovering the Spatial Layout of Cluttered Rooms. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2009.
- Commission Internationale de l’Éclairage (CIE). *Colorimetry—Part 1: CIE Standard Colorimetric Observers*; Number 15:2004 in CIE Standard; CIE Central Bureau: Vienna, Austria, 2004.
- Zhou, Q.Y.; Park, J.; Koltun, V. Fast Global Registration. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2016.
- Yang, H.; Shi, J.; Carlone, L. TEASER: Fast and Certifiable Point Cloud Registration. *IEEE Trans. Robot.* **2021**, *37*, 314–333. <https://doi.org/10.1109/TRO.2020.3033695>.

17. Choy, C.; Park, J.; Koltun, V. Fully Convolutional Geometric Features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019.
18. Bai, X.; Luo, Z.; Zhou, L.; Fu, H.; Quan, L.; Tai, C.L. D3Feat: Joint Learning of Dense Detection and Description of 3D Local Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020.
19. Bai, X.; Luo, Z.; Zhou, L.; Chen, H.; Li, L.; Hu, Z.; Fu, H.; Tai, C.L. PointDSC: Robust Point Cloud Registration Using Deep Spatial Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021.
20. Thomas, H.; Qi, C.R.; Deschard, J.E.; Marcotegui, B.; Goulette, F.; Guibas, L.J. KPConv: Flexible and Deformable Convolution for Point Clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019.
21. Xiong, F.; Kong, Y.; He, L.; Kuang, L.; Han, X. DeepPAT: Deep Position-Aware Transformer with Mixed Dataset for Robust Point Cloud Registration. *Neurocomputing* **2026**, *668*, 132349. <https://doi.org/10.1016/j.neucom.2025.132349>.
22. Ao, S.; Wu, Q.; Yang, Y.; Hao, Y.; Wang, J.; Yi, L.; Liu, Y. SpinNet: Learning a General Surface Descriptor for 3D Point Cloud Registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 11753–11762.
23. Yew, Z.J.; Lee, G.H. REGTR: End-to-End Point Cloud Correspondences with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022.
24. Yu, J.; Ren, L.; Zhang, Y.; Zhou, W.; Lin, L.; Dai, G. PEAL: Prior-Embedded Explicit Attention Learning for Low-Overlap Point Cloud Registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023.
25. Fu, K.; Yuan, M.; Wang, C.; Pang, W.; Chi, J.; Wang, M.; Gao, L. Dual Focus-Attention Transformer for Robust Point Cloud Registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2025; pp. 11769–11778.
26. Liu, J.; Wang, G.; Liu, Z.; Jiang, C.; Pollefeys, M.; Wang, H. RegFormer: An Efficient Projection-Aware Transformer Network for Large-Scale Point Cloud Registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 8451–8460.
27. Yu, H.; Li, F.; Saleh, M.; Busam, B.; Ilıc, S. CoFiNet: Reliable Coarse-to-Fine Correspondences for Robust PointCloud Registration. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, Virtual, 6–14 December 2021.
28. Zhang, D.; Zhu, J.; Li, S.; Yan, W.; Xu, H.; Fan, P.; Lu, H. IGASA: Integrated Geometry-Aware and Skip-Attention Modules for Enhanced Point Cloud Registration. *IEEE Trans. Circuits Syst. Video Technol.* **2026**, *36*, 8035–8047. <https://doi.org/10.1109/TCSVT.2026.3656975>.
29. Liu, J.; Wang, G.; Liu, Z.; Jiang, C.; Li, H.; Liu, M.; Deng, T.; Pollefeys, M.; Yang, M.Y.; Wang, H. RegFormer++: An Efficient Large-Scale 3D LiDAR Point Registration Network with Projection-Aware 2D Transformer. *arXiv* **2026**, arXiv:2603.14290. <https://doi.org/10.48550/arXiv.2603.14290>.
30. Park, J.; Zhou, Q.Y.; Koltun, V. Colored Point Cloud Registration Revisited. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2017.
31. Zhou, P.; An, L.; Wang, Y.; Geng, G.; Xu, Y. Cross-Domain Multi-Channel Transformer for Point Cloud Registration in Cultural Heritage Digital Preservation. *Npj Herit. Sci.* **2025**, *13*, 503. <https://doi.org/10.1038/s40494-025-02064-4>.
32. Zhao, H.; Jiang, L.; Jia, J.; Torr, P.H.S.; Koltun, V. Point Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 16259–16268.
33. Ravi, N.; Reizenstein, J.; Novotny, D.; Gordon, T.; Lo, W.C.; Johnson, J.; Girdhar, R. PyTorch3D. In *Proceedings of the ACM International Conference on Multimedia (MM) Open Source Software Competition*; ACM: New York, NY, USA, 2020.

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.