ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XI-2-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission II, 4–11 July 2026, Toronto, Canada

Differentiable deep consistency for point cloud registration

Tian Zhang*, Sagi Filin

Mapping and Geo-Information Engineering, Technion – Israel Institute of Technology, Haifa, Israel
(tianz, filin)@technion.ac.il

Keywords: Point cloud registration, Laser scanners, Deep learning, Correspondence matching

Abstract

Point cloud registration is a key facilitator for mapping, autonomous driving, and robotic applications. Current neural-based pipelines focus on learning view-consistent descriptors for correspondence matching, typically followed by geometric verification to assess distance/angular preservation and aid transformation estimation. Though beneficial, pairwise correspondence verification scales quadratically, creating a computational bottleneck. Moreover, since matching and verification are optimized separately, the latter cannot guide descriptor learning or foster geometric awareness. To address both limitations, we introduce an end-to-end neural registration framework that blends correspondence learning and verification into a single differentiable formulation. We propose a consistency-driven cross-attention module that dynamically correlates cross-scan neighborhoods to suppress inconsistent matches and reinforce inter-scan feature coherence, generating robust, discriminative descriptors without the quadratic cost of explicit pairwise verification. Our formulation integrates seamlessly into state-of-the-art architectures, GeoTransformer and RoITr, without additional supervision or post-processing. Results demonstrate superiority in challenging setups, where competing methods either produce few correct correspondences or fail entirely. Our method consistently achieves superior inlier ratios and the lowest registration errors on 3DMatch, 3DLoMatch, and KITTI benchmarks, improving registration recall by up to 2.6%, directly addressing setups where state-of-the-art frameworks fail. Beyond accuracy, our model converges faster during training and achieves the quickest inference among state-of-the-art methods, reflecting the value and soundness of our differentiable formulation.

1. Introduction

Point cloud registration concerns estimating a relative rigid transformation whose aim is to align two scans acquired from different views. It acts as a fundamental task in 3D mapping, autonomous driving and robotics, as only a few examples (e.g., Yin et al., 2024; Hu et al., 2025). Registration schemes tend to involve the detection of salient repeatable points, generation of descriptive information to establish correspondences across scans, and finally the estimation of rigid transformation parameters (Theiler et al., 2015; Dong et al., 2018; Li et al., 2020; Huang et al., 2021; Yu et al., 2021, 2023b; Zhao et al., 2025). A refinement phase that usually ensues, employs the iterative closest point algorithm, or its variants, for optimal alignment (e.g., Besl and McKay, 1992; Chen and Medioni, 1992).

Recent point registration methods are neural-driven, attempting to learn point descriptors that remain consistent across views (Zeng et al., 2017; Choy et al., 2019; Gojcic et al., 2019; Bai et al., 2020; Qin et al., 2022; Ao et al., 2023; Zhao et al., 2025). The dominant paradigm is correspondence-based, where a coarse-to-fine matching mechanism is utilized to yield a dense set of candidates. It down-samples the original scan into a sparse set of points, refines the descriptors with transformer-based modules, and performs coarse matching and correspondences propagation to the point cloud (Qin et al., 2022; Yu et al., 2023a). Despite their success, the established correspondences contain numerous outliers, which challenge the transformation estimation (Fig. 1a). To mitigate their effect, an additional verification stage ensues, in which point pairs are tested for their distance preservation, normal consistency, or feature similarity under a rigid transformation (Chen et al., 2022; Zhang et al., 2023; Bai et al., 2021; Jiang et al., 2023). Nonetheless, existing frameworks treat the correspondence matching and verification as two disjoint stages, optimized by separate networks, limiting the valuable feedback from the latter be utilized to guide feature learning in the former. This decoupling impairs the geometric awareness of the feature descriptors and inlier ratios (Zhao et al., 2025). Furthermore, as the verification phase merely filters correspondences without generating new ones, the transformation estimation may become challenging, especially in low-overlap scans.

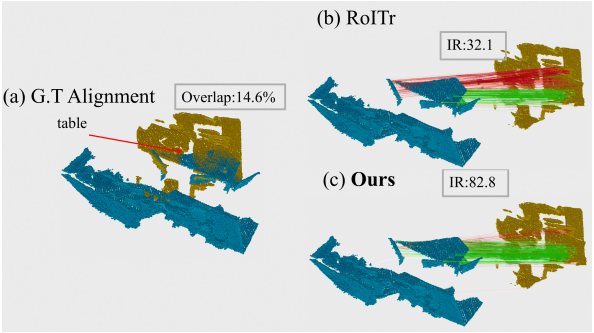


Figure 1. Registration of two low-overlap scans of a kitchen scene (a), the application of the state-of-the-art method RoITr yields numerous outliers leading to failure (b), while our method retrieves a high inlier rate with successful alignment (c).

To address these challenges, this paper introduces a novel end-to-end point cloud registration framework that integrates correspondence matching and verification within a single differentiable formulation. Our model learns expressive, unique and spatially coherent representations for each point via integrated self- and cross-attention layers and employs a new geometric differentiable consistency module that explicitly enforces descriptor consistency through a novel cross-attention unit. We amplify

* Corresponding author

attention in regions exhibiting locally consistent spatial structures while enhancing features in inconsistent ones to promote coherence. Consequently, our framework learns geometry-preserving features that produce reliable matches even under challenging low-overlap conditions. Hence, our contributions are the following: *i*) a differentiable feature consistency module that unifies matching and verification within a new attention mechanism, promoting distinct and cross-view consistent features; *ii*) an efficient and theoretically grounded formulation that ensures rapid convergence during training and accelerated inference; *iii*) seamless integration into state-of-the-art architectures, such as GeoTransformer and RoITr, without requiring additional supervision or post-processing; and *iv*) superior robustness in challenging scenarios, where competing methods produce few correct correspondences or fail entirely. Extensive experiments demonstrate fast convergence during training and fastest inference at test time. In challenging low-overlap scenarios, our proposed model attains the highest inlier ratio and lowest registration error among compared methods, successfully retrieving correspondences even in such hard cases where other methods fail.

2. Related Work

The dominant neural point cloud registration paradigm attempts learning consistent point descriptors across views. Based on their representation and processing strategies, they can be divided into two groups: patch-based methods that generate descriptors separately for a selected set of anchor points from the original point cloud (Zeng et al., 2017; Deng et al., 2018; Ao et al., 2021, 2023; Wang et al., 2023), and fragment-based methods that process all points simultaneously (Choy et al., 2019; Bai et al., 2020; Huang et al., 2021; Yu et al., 2021; Qin et al., 2022; Yu et al., 2023a; Chen et al., 2024). The former was pioneered by Zeng et al. (2017), where local neighborhoods per point were converted into occupancy grids and processed by convolutional networks. The obtained descriptors were trained with contrastive loss to minimize the metric distance between matching points while maximizing it for non-matched ones. Addressing limitations of data quantization in occupancy grid, Deng et al. (2018) utilized the raw 3D coordinates and applied the PointNet architecture to yield per point descriptor. Ao et al. (2021, 2023) reoriented the local neighborhood with principal directions, and employed spherical convolutions to achieve rotation invariance. While effective for local descriptor learning, patch-based methods rely solely on local context and do not explicitly reason about pairwise geometric compatibility between correspondences. As a result, they often produce ambiguous matches, require post hoc filtering, and are slow to perform. To improve computational efficiency, fragment-based method employed a coarse-to-fine strategy and utilized fully convolutional architectures for dense matching (Choy et al., 2019; Thomas et al., 2019). For instance, building on KPConv, Yu et al. (2021) and Qin et al. (2022) utilized alternating self- and cross-attention layers, to capture intra-scan contextual relationships and enforce inter-scan feature consistency, respectively. Sparse matches obtained from the network bottleneck were then propagated via interpolation to the full point cloud level. To achieve view independence, Yu et al. (2023a) further improved the positional encoding in self- and cross-attention, using point pair-features such as normal dot product and distance between point pairs. Despite their architectural advances and computational efficiency, current frameworks still solely rely on feature similarity to establish correspondences, without evaluating their mutual geometric agreement (Zhao et al., 2025). Consequently, their application remains susceptible to outliers and is typically coupled with an additional verification stage.

Correspondence verification aims to increase the inlier ratio by examining the geometric consistency of matched pairs and removing incompatible ones (Chen et al., 2022; Zhang et al., 2023; Zhao et al., 2025). Classical geometry-based approaches, constructed compatibility graphs to encode spatial consistency (e.g., pairwise distances or normal dot products) and selected subsets of mutually consistent correspondences by maximizing the overall compatibility scores (e.g., Chen et al., 2022; Zhang et al., 2023). These methods relied on dataset-specific thresholds and often struggled with noisy or partial scans. Learning-based verification methods (e.g., Bai et al., 2021; Jiang et al., 2023) incorporated neural models to predict inlier probabilities from correspondence features. Bai et al. (2021) used feature similarity and geometric distances as attention weights in a self-attention module to update correspondence features. A neural spectral matching module was then employed to predict inlier probabilities, which was trained with a binary cross-entropy loss. Jiang et al. (2023) processed pairs of correspondence features through a variational Bayesian framework to estimate inlier probabilities, capturing uncertainty and improving outlier rejection. Zhao et al. (2025) proposed a progressive iterative framework that regenerated correspondences using multi-point geometric consistency. In each iteration, correspondences satisfying local and global distance compatibility constraints were used to estimate the transformation and update the raw matches, yielding a larger correspondence set than standard filtering. Although these verification techniques improved robustness, they were applied post hoc, treating verification as an independent stage. This post hoc design breaks the feedback loop between matching and verification, preventing the geometric consistency signals used during verification, from influencing the feature learning. As a result, descriptors learned during matching lack spatial consistency, and the verification mainly filters correspondences rather than refining them, usually leading to a reduced set of inliers and unstable transformation estimation.

These limitations motivate us to develop a joint formulation, which integrates correspondence matching and verification into a single differentiable framework. By embedding feature consistency reasoning directly into the feature learning process, our model jointly learns to identify, refine, and validate correspondences in a geometry-aware manner. We show how our generated correspondence secures high consistency that reduces the need for further post-processing. With our new formulation, our network utilizes the same number of parameters as state-of-the-art method, but achieves high convergence speed and reduces the training time needed.

3. Methodology

Given two partially overlapping scans, $\mathcal{P} = \{\mathbf{p}_i \in \mathbb{R}^3 | i = 1, \dots, N\}$ and $\mathcal{Q} = \{\mathbf{q}_i \in \mathbb{R}^3 | i = 1, \dots, M\}$, our goal is to estimate a rigid transformation, $\mathbf{T} = \{\mathbf{R}, \mathbf{t}\}$, that aligns the two scans, where $\mathbf{R} \in SO(3)$ is a 3D rotation, and $\mathbf{t} \in \mathbb{R}^3$ is a 3D translation. The transformation is solved by,

$$(\mathbf{R}^*, \mathbf{t}^*) = \underset{\{\mathbf{R},\mathbf{t}\} \subset \mathrm{SE}(3)}{\arg\min} \sum_{(\mathbf{p}_{x_i}, \mathbf{q}_{y_i}) \in \mathcal{C}^*} \|\mathbf{R}\mathbf{p}_{x_i} + \mathbf{t} - \mathbf{q}_{y_i}\|_2^2 , \tag{1}$$

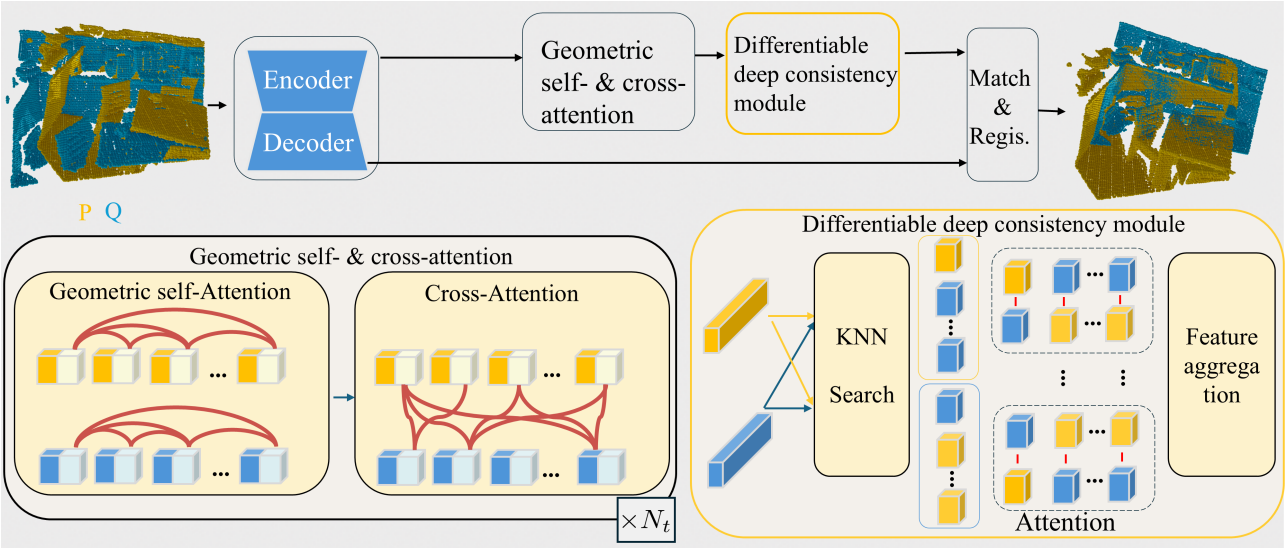


Figure 2. Overview of our proposed framework. The differentiable deep consistency module lies at the core of our design.

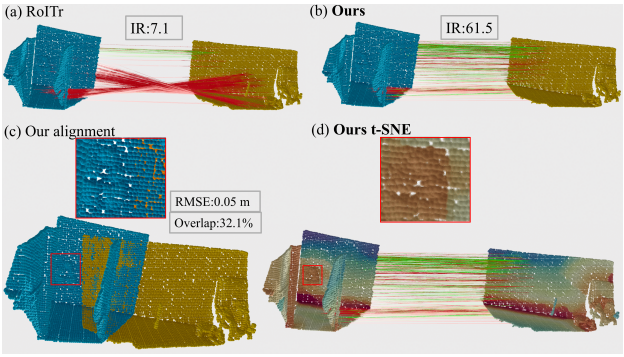


Figure 3. Matching over a nearly flat scene. Our model highlights subtle patterns (t-SNE plot in red box) to facilitate consistent matches, evident by the registration RMSE.

where $\mathcal{C}^*$ is the correspondences set between between $\mathcal{P}$ and $\mathcal{Q}$.

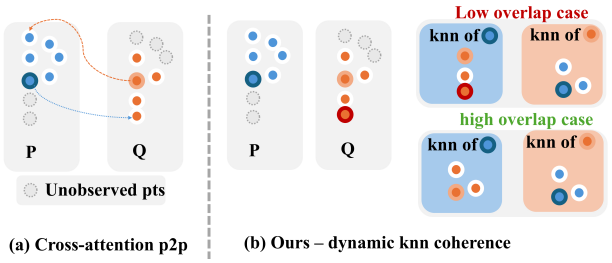
Our framework follows the coarse-to-fine matching paradigm (Qin et al., 2022; Yu et al., 2023a) to extract correspondences. We use their feature extraction backbones as they are to extract multi-level descriptors (Fig. 2). Given the two scans, $\mathcal{P}$ and $\mathcal{Q}$, the encoder outputs the downsampled coarse-level superpoints, $\hat{\mathcal{P}}$ and $\hat{\mathcal{Q}}$, with their associated learned descriptors, $\mathbf{F}_{\hat{\mathcal{P}}} \in \mathbb{R}^{\hat{n}\times\hat{d}}$ and $\mathbf{F}_{\hat{\mathcal{Q}}} \in \mathbb{R}^{\hat{m}\times\hat{d}}$, respectively. The decoder outputs the fine-level points, $\tilde{\mathcal{P}}$ and $\tilde{\mathcal{Q}}$, with the associated descriptors, $\mathbf{F}_{\tilde{\mathcal{P}}} \in \mathbb{R}^{\tilde{n}\times\tilde{d}}$ and $\mathbf{F}_{\tilde{\mathcal{Q}}} \in \mathbb{R}^{\tilde{m}\times\tilde{d}}$, respectively. The super points, $\hat{\mathcal{P}}$ and $\hat{\mathcal{Q}}$, with the associated learned descriptors, $\mathbf{F}_{\hat{\mathcal{P}}}$ and $\mathbf{F}_{\hat{\mathcal{Q}}}$, are then fed into geometric self- and cross-attention layers. The output is introduced into our newly proposed differentiable deep consistency module to generate highly consistent features across views. We employ superpoint matching as in Yu et al. (2023a) to yield the coarse set correspondence, $\hat{\mathcal{C}} = \{(\hat{\mathbf{p}}_{x_i}, \hat{\mathbf{q}}_{y_i}) | \hat{\mathbf{p}}_{x_i} \in \hat{\mathcal{P}}, \hat{\mathbf{q}}_{y_i} \in \hat{\mathcal{Q}}\}$, and then perform point level matching to propagate the coarse set correspondence into dense point correspondences, $\tilde{\mathcal{C}} = \{(\tilde{\mathbf{p}}_{x_i}, \tilde{\mathbf{q}}_{y_i}) | \tilde{\mathbf{p}}_{x_i} \in \tilde{\mathcal{P}}, \tilde{\mathbf{q}}_{y_i} \in \tilde{\mathcal{Q}}\}$.

3.1 Differentiable deep consistency module

Self- and cross-attention mechanisms are commonly employed to model intra-scan geometric context and inter-scan feature consistency, yielding refined superpoint features, $\mathbf{H}_{\hat{\mathcal{P}}}$ and $\mathbf{H}_{\hat{\mathcal{Q}}}$. Despite their effectiveness, the coarse correspondences derived from these features often include incorrect matches caused by locally inconsistent representations across scans. Left unchecked these incorrect matches propagate to fine-level point matching (Fig. 3a), degrading the registration accuracy. We address this by introducing a differentiable consistency-driven module that actively suppresses inconsistent matches and promotes mutually coherent correspondences. A direct way to assess correspondence compatibility is to measure the pairwise geometric, or feature-level consistency between all correspondence pairs. However, this operation is of quadratic complexity in the number of superpoints and is computationally intractable during training.

Our formulation evolves from the intuition that true correspondences are mutually supported by their local neighborhood structures, whereas false matches exhibit incoherent or unstable neighborhood relationships. In contrast to static neighborhood definitions, such as Euclidean proximity or fixed feature similarity thresholds, we employ a dynamic cross-scan neighbor search performed in feature space. All superpoint features are first normalized onto a unit hypersphere, and for each one in $\hat{\mathcal{P}}$, we search for the ordered k nearest neighbors in the feature space of $\hat{\mathcal{Q}}$ using $\mathbf{H}_{\hat{\mathcal{P}}}$ and $\mathbf{H}_{\hat{\mathcal{Q}}}$, forming the cross-scan neighbor feature matrix, $\mathbf{Z}_{\hat{\mathcal{P}}} \in \mathbb{R}^{\hat{n}\times k\times d}$. Similarly, we obtain the feature matrix, $\mathbf{Z}_{\hat{\mathcal{Q}}} \in \mathbb{R}^{\hat{m}\times k\times d}$, for $\hat{\mathcal{Q}}$. As these neighbor sets are re-evaluated, the notion of proximity becomes feature-adaptive, and does not need to coincide with geometric proximity in the raw input space.

To exploit these dynamic neighborhoods for information aggregation, we design a consistency cross-attention mechanism that evaluates the mutual compatibility of potential correspondences. Rather than comparing neighbor similarities across scans as in standard cross-attention, we compute ranked correlations that highlight neighborhood pairs exhibiting coherent local structure while dynamically refining features to encourage coherence among initially inconsistent true pairs (Fig. 4b). The attention score for each superpoint pair is defined as cumulative correlations between their local feature neighborhoods, $\mathbf{z}^m_{\hat{\mathcal{P}}_i} \in \mathbf{Z}_{\hat{\mathcal{P}}}$ & $\mathbf{z}^m_{\hat{\mathcal{Q}}_j} \in \mathbf{Z}_{\hat{\mathcal{Q}}}$, via the learnable projection matrices,



(a) Cross-attention p2p (b) Ours – dynamic knn coherence

Figure 4. (a) Cross-attention scores candidate correspondences using only point-to-point (P2P) similarity. (b) In contrast, our deep consistency module is driven by dynamic neighborhood consistency, where the attention score is determined by the overlap of mutual neighbors across two partially overlapping scans. The low- and high-overlap examples demonstrate that this mechanism better captures local structural agreement.

$\mathbf{W}^Q$ and $\mathbf{W}^K \in \mathbb{R}^{d\times d}$, where d is the feature dimension:

$$e_{\hat{\mathcal{P}}_i\hat{\mathcal{Q}}_j} = \sum_{m=1}^{k} \frac{(\mathbf{z}^m_{\hat{\mathcal{P}}_i}\mathbf{W}^Q)(\mathbf{z}^m_{\hat{\mathcal{Q}}_j}\mathbf{W}^K)^T}{\sqrt{d}}. \tag{2}$$

The resulting attention scores form the attention matrix $\mathbf{E} \in \mathbb{R}^{|\hat{\mathcal{P}}|\times|\hat{\mathcal{Q}}|}$, where each element $e_{\hat{\mathcal{P}}_i\hat{\mathcal{Q}}_j}$ quantifies the consistency confidence between points $\hat{\mathcal{P}}_i$ and $\hat{\mathcal{Q}}_j$, implicitly encoding a local pairwise compatibility test between their corresponding neighborhoods. We then apply a row-wise softmax normalization and update features through a learnable value projection $\mathbf{W}^V \in \mathbb{R}^{d\times d}$:

$$\mathbf{U}_{\hat{\mathcal{P}}} = \mathrm{Softmax}\big(\mathbf{E}\big)(\mathbf{H}_{\hat{\mathcal{Q}}}\mathbf{W}^V). \tag{3}$$

This aggregation integrates correspondence validation directly into the feature update process, effectively embedding consistency check within the attention mechanism. By dynamically correlating cross-scan neighborhoods in feature space, the module we introduce implicitly captures local relational dependencies without incurring the quadratic cost of explicit pairwise verification. As a result, it reinforces inter-scan feature coherence and produces refined, mutually consistent correspondences, as illustrated in (Fig. 3b).

3.2 Differentiability and backpropagation

We analyze the properties of our formulation in the backpropagation phase. The overall loss follows Qin et al. (2022) and combines the superpoint matching loss L_c and the fine-level point matching loss L_f, $L = L_c + L_f$. The coarse loss encourages corresponding superpoints $(\hat{p}_i, \hat{q}_j) \in \hat{\mathcal{C}}$ to have similar features, i.e.,

$$\mathbf{U}_{\hat{\mathcal{P}}_i} \approx \mathbf{U}_{\hat{\mathcal{Q}}_j}. \tag{4}$$

Each superpoint feature is updated as in Eq. (3) where $\mathbf{W}^V$ is shared. This implies $\mathbf{H}_{\hat{\mathcal{P}}_i} \approx \mathbf{H}_{\hat{\mathcal{Q}}_j}$, and the attention scores between corresponding neighbor pairs are maximized. According to Eq. (2), this is achieved when the neighbors of $\hat{\mathcal{P}}_i$ and $\hat{\mathcal{Q}}_j$ exhibit maximal overlap and feature similarity, as we now derive.

Specifically, when true correspondences exist, the neighboring points of $\hat{\mathcal{P}}_i$ and $\hat{\mathcal{Q}}_j$ exhibit similar feature patterns, leading their associated feature embedding to be well aligned. Consequently, the majority of the r matched neighbors contribute positively and coherently to the total attention score, while the remaining $(k-r)$ unmatched neighbors contribute weaker values. Formally, the total attention score can be written as $e_{\hat{\mathcal{P}}_i\hat{\mathcal{Q}}_j} = \sum_{m=1}^{r}\mathbf{x}_m\cdot\mathbf{y}_m + \sum_{m=r+1}^{k}\mathbf{x}_m\cdot\mathbf{y}_m$, where $\mathbf{x}_m = \mathbf{z}^m_{\hat{\mathcal{P}}_i}\mathbf{W}^Q/d^{1/4}$, and $\mathbf{y}_m = \mathbf{z}^m_{\hat{\mathcal{Q}}_j}\mathbf{W}^K/d^{1/4}$, denote normalized local features. For the matched pairs $\mathbf{x}_m$, the norm is bounded due to feature normalization, and the weight regularization bounds the spectral norm of $\mathbf{W}^Q$, so that $\|\mathbf{x}_m\| \le B$, where B denotes the product of the spectral norm and the scaling factor $d^{-1/4}$. Moreover, the contrastive loss $\mathcal{L}_c$ encourages feature alignment between matched pairs, enforcing proximity $\|\mathbf{x}_m - \mathbf{y}_m\| \le \varepsilon$, where ε is a contrastive-loss-determined constant. Then, by the Cauchy–Schwarz inequality, feature consistency for the matched pairs implies

$$\begin{aligned}
\mathbf{x}_m\cdot\mathbf{y}_m &= \mathbf{x}_m\cdot\big(\mathbf{x}_m - (\mathbf{x}_m - \mathbf{y}_m)\big) && (5)\\
&= \|\mathbf{x}_m\|^2 - \mathbf{x}_m\cdot(\mathbf{x}_m - \mathbf{y}_m) && (6)\\
&\ge \|\mathbf{x}_m\|^2 - \|\mathbf{x}_m\|\,\|\mathbf{x}_m - \mathbf{y}_m\| && (7)\\
&\ge B^2 - B\cdot\varepsilon. && (8)
\end{aligned}$$

In contrast, for non-matching pairs, $\mathbf{x}_m\cdot\mathbf{y}_m$ is treated as a smaller value where the average we denote as μ_0, where $\mu_0 \ll B$. Therefore, the total score satisfies $e_{\hat{\mathcal{P}}_i\hat{\mathcal{Q}}_j} \ge r\,(B^2 - B\cdot\varepsilon) + (k-r)\mu_0$.

Interpretation. When true correspondences exist, the network dynamically adjusts the features such that $r > (k-r)$, allowing the coherent signal to dominate the noisy component and producing a distinct peak in the attention response. The optimal state corresponds to a full-overlap condition (Fig. 4b). Conversely, when no correspondence is present, the neighborhood correlations are unstructured, leading to uniformly low attention scores across candidates. This built-in mechanism effectively replaces explicit correspondence verification by modeling *relational feature consistency* among local neighborhoods, analogous to validating rigid point-pair relations, but in a fully differentiable and learnable form. In contrast to the standard cross-attention form, our formulation explicitly enforces correspondence consistency through joint neighbor alignment. Furthermore, we demonstrate that when our consistency module added, it leads to faster convergence compared to the vanilla self-cross attention, as shown in Fig. (5).

3.3 Matching and registration

We employ coarse-to-fine matching strategy to find correspondences (Qin et al., 2022; Wang et al., 2025). For each normalized feature pair $\mathbf{U}_{\hat{\mathcal{P}}_i}$ and $\mathbf{U}_{\hat{\mathcal{Q}}_j}$, the Gaussian correlation matrix is computed as $\mathbf{S}(i,j) = \exp\left(-\|\mathbf{U}_{\hat{\mathcal{P}}_i} - \mathbf{U}_{\hat{\mathcal{Q}}_j}\|^2\right)$. After dual normalization of $\mathbf{S}$ to capture global correlations, we select the top-$\hat{k}$ entries to form the coarse correspondence set $\hat{\mathcal{C}} = \{\,(\hat{p}_i,\hat{q}_j)\,|\,\hat{p}_i\in\hat{\mathcal{P}},\ \hat{q}_j\in\hat{\mathcal{Q}}\,\}$. We then employ a point-to-node strategy to propose fine-level correspondences, where each dense point in $\tilde{\mathcal{P}}$ and $\tilde{\mathcal{Q}}$ is associated with its nearest superpoint $\hat{p}_i$ and $\hat{q}_j$, respectively (Qin et al., 2022; Yu et al., 2023a). The group of points assigned to each superpoint $\hat{p}_i$ and $\hat{q}_j$ is denoted as, $\mathcal{G}_{\hat{P}_i}\subseteq\tilde{\mathcal{P}}$ and $\mathcal{G}_{\hat{Q}_j}\subseteq\tilde{\mathcal{Q}}$, with their associated features defined as, $\mathbf{F}_{\mathcal{G}_{\hat{P}_i}}\subseteq\mathbf{F}_{\tilde{\mathcal{P}}}$ and $\mathbf{F}_{\mathcal{G}_{\hat{Q}_j}}\subseteq\mathbf{F}_{\tilde{\mathcal{Q}}}$, respectively. The similarity between feature groups, $\mathbf{F}_{\mathcal{G}_{\hat{P}_i}}$ and $\mathbf{F}_{\mathcal{G}_{\hat{Q}_j}}$, is computed as, $\mathbf{S}_g = \mathbf{F}_{\mathcal{G}_{\hat{P}_i}}\mathbf{F}^{\top}_{\mathcal{G}_{\hat{Q}_j}}/\sqrt{\tilde{d}}$. The resulting similarity matrix $\mathbf{S}_g$ is then processed by the Sinkhorn algorithm (Sarlin et

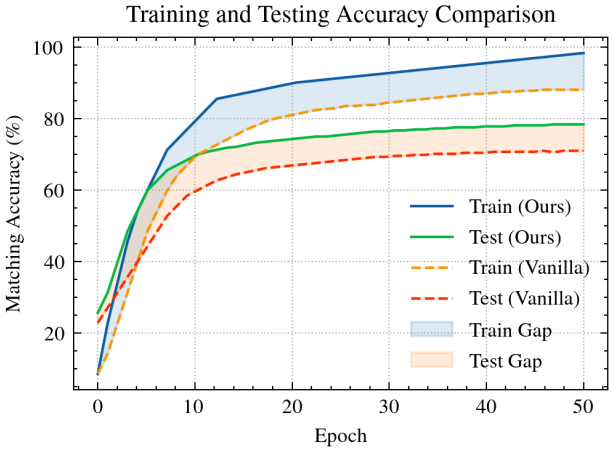


Figure 5. Comparison of training and testing matching accuracy between our proposed differentiable deep consistency module (Ours), and the vanilla self-cross attention baseline. Our method exhibits faster convergence and achieves consistently higher training and testing accuracy.

al., 2020) to produce the fine-level correspondence set, $\tilde{\mathcal{C}}_g$. The final correspondence set is obtained by aggregating across all neighborhood pairs $\tilde{\mathcal{C}} = \bigcup_{g=1}^{|\tilde{\mathcal{C}}|} \tilde{\mathcal{C}}_g$. A robust estimator such as RANSAC or LGR is used to compute the rigid transformation and register the point clouds (Qin et al., 2022). Notably, our network is trained using the loss function defined in Sec. (3.2), and is therefore not elaborated here.

4. Results

Datasets We assess our proposed model on the 3DMatch, 3DLoMatch, and the KITTI benchmarks (Geiger et al., 2012; Zeng et al., 2017; Huang et al., 2021), following standard evaluation protocols (Qin et al., 2022; Yu et al., 2023a). The 3DMatch dataset is a large-scale indoor RGB-D collection comprising 62 diverse scenes, of which 46 are used for training, eight for validation, and eight for testing. The 3DLoMatch benchmark is a more challenging variant of 3DMatch that focuses on scan pairs with limited overlap, typically ranging from 10% to 30%. Our model is trained exclusively on 3DMatch and evaluated directly on 3DLoMatch, without any additional fine-tuning. The third dataset, KITTI, is an outdoor mobile LiDAR dataset, which features different scanning pattern and data characteristics. We use sequences 0–10 for the analysis and the same training and evaluation as in Qin et al. (2022).

Metrics We evaluate the registration performance using three widely adopted metrics (e.g., Huang et al., 2021; Qin et al., 2022; Wang et al., 2023): *i*) Inlier Ratio (IR) – the proportion of correspondences whose residuals, computed under the ground-truth transformation, fall below a predefined threshold (0.1 m); *ii*) Feature Matching Recall (FMR) – the percentage of scan pairs whose estimated transformation achieves an error within a relative bound of 5%; and *iii*) Registration Recall (RR) – the percentage of scan pairs whose registration result attains a root mean square error (RMSE) smaller than 0.1 m. A higher value in any of these metrics reflects stronger registration quality. In addition, we quantify the accuracy of the recovered transformation using the relative rotation error (RRE), the angular deviation between predicted and ground-truth rotations, and the relative translation error (RTE), the Euclidean distance between their translation vectors.

Experiment setting and baselines To validate the effectiveness and generality of our proposed model, we integrate it into two state-of-the-art registration frameworks: GeoTrans and RoITr. The former employs a KPConv-based backbone, while the latter adopts a PointTransformer-based design, both with an alternated sequence of self- and cross-attention layers in the network bottleneck. In both cases, we replace the last cross-attention layer with our differentiable deep consistency module. The resulting variants are denoted as **Ours**+GeoTrans and **Ours**+RoITr, respectively. We compare against seven state-of-the-art methods: GeoTrans (Qin et al., 2022), RoITr (Yu et al., 2023a), SpinNet (Ao et al., 2021), Predator (Huang et al., 2021), CoFiNet (Yu et al., 2021), YOHO (Wang et al., 2022), and RIGA (Yu et al., 2024).

4.1 Quantitive registration evaluation

To evaluate the model performance under different level of correspondence, we perform experiments on 3DMatch and 3DLoMatch under a varying number of sampled counts, following the common convention (Qin et al., 2022; Wang et al., 2023). Regarding IR, **Ours**+RoITr records the best inlier ratio, outperforming all others on both datasets, with the exception when sampling correspondence at 250 (Table 1). For example, on 3DMatch, the IR increased to 83.3–83.5% across all correspondence densities, while on 3DLoMatch, the IR reached 55.1–56.0%. Regarding the FMR, we lead the chart, with 98.3% on 3DMatch, and 89.8% on 3DLoMatch. These results highlights the ability to extract more correct correspondences on challenging cases (Fig. 3). In terms of registration recall, we achieve the best registration recall compared to all others. With **Ours**+RoITr, RR reaches 94.2% on 3DMatch, and 75.9% on 3DLoMatch, substantially higher than all baselines. We achieve consistent improvement over RoITr, and GeoTrans with our new differentiable consistency module. On challenging cases, when the structure is almost flat and low overlap (Fig. 7), our model manages to estimate close to ground-truth transformations, compared to the complete failure of the RoITr. This improvement is a direct consequence of the neighborhood-to-neighborhood attention: by embedding inter-scan geometric coherence into feature learning, our new module produces highly consistent superpoint correspondences and generates more correspondence proposals than the standard verification, which enables more accurate rigid transformations and final registration. On the KITTI dataset, **Ours**+GeoTrans improves the IR by 1.7% comparing to GeoTrans, and achieves RR and FMR rates of 99.8%, which are on par with those of GeoTrans. We attribute these high rates to the relatively small baselines between scan pairs.

Hyperparameter sensitivity The sensitivity of the hyperparameter, k, in the differentiable consistency module is evaluated on 3DLoMatch. Fig. 6 shows that for $k < 6$, the registration recall consistently improves, and then reaches a steady value. Hence, we set $k = 6$, remains fixed for all experiments.

Registration error analysis To evaluate the registration quality, we compare the estimated transformations against the ground truth using RRE and RTE. As shown in Table 2, our model achieves the highest registration recall on both 3DMatch (94.5%) and 3DLoMatch (75.9%), demonstrating effectiveness in both standard and low-overlap scenarios. Notably, the rotation error shows a clear reduction, indicating improved pose

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-179-2026 | © Author(s) 2026. CC BY 4.0 License.

Table 1. Quantitative results on 3DMatch & 3DLoMatch with a varying number of points/correspondences. Bold and underscore, here and following, indicate the best and second-best results, respectively.

	3DMatch					3DLoMatch				
# Samples	5000	2500	1000	500	250	5000	2500	1000	500	250
	Feature Matching Recall (%) ↑									
SpinNet (Ao et al., 2021)	97.4	97.0	96.4	96.7	94.8	75.5	75.1	74.2	69.0	62.7
Predator (Huang et al., 2021)	96.6	96.6	96.5	96.3	96.5	78.6	77.4	76.3	75.7	75.3
CoFiNet (Yu et al., 2021)	98.1	**98.3**	98.1	98.2	**98.3**	83.1	83.5	83.3	83.1	82.6
YOHO (Wang et al., 2023)	**98.2**	97.6	97.5	97.7	96.0	79.4	78.1	76.3	73.8	69.1
RIGA (Yu et al., 2024)	97.9	97.8	97.7	97.7	97.6	85.1	85.0	85.1	84.3	85.1
GeoTrans (Qin et al., 2022)	97.9	97.9	97.9	97.9	97.6	88.3	88.6	88.8	88.6	88.3
Ours + GeoTrans	98.1	98.0	98.1	98.2	98.2	88.5	88.9	88.9	88.7	88.5
RoITr (Yu et al., 2023a)	98.0	98.0	97.9	98.0	97.9	89.6	89.6	89.5	89.4	89.3
Ours + RoITr	**98.2**	**98.3**	**98.2**	**98.3**	98.2	**89.7**	**89.7**	**89.8**	**89.7**	**89.8**
	Inlier Ratio (%) ↑									
SpinNet (Ao et al., 2021)	48.5	46.2	40.8	35.1	29.0	25.7	23.7	20.6	18.2	13.1
Predator (Huang et al., 2021)	58.0	58.4	57.1	54.1	49.3	26.7	28.1	28.3	27.5	25.8
CoFiNet (Yu et al., 2021)	49.8	51.2	51.9	52.2	52.2	24.4	25.9	26.7	26.8	26.9
YOHO (Wang et al., 2023)	64.4	60.7	55.7	46.4	41.2	25.9	23.3	22.6	18.2	15.0
RIGA (Yu et al., 2024)	68.4	69.7	70.6	70.9	71.0	32.1	33.4	34.3	34.5	34.6
GeoTrans (Qin et al., 2022)	71.9	75.2	76.0	82.2	85.1	43.5	45.3	46.2	52.9	57.7
Ours + GeoTrans	72.3	75.8	76.7	82.8	**85.2**	43.8	45.6	46.5	53.2	**58.1**
RoITr (Yu et al., 2023a)	82.6	82.8	83.0	83.0	83.0	54.3	54.6	55.1	55.2	55.3
Ours + RoITr	**83.3**	**83.4**	**83.5**	**83.5**	83.5	**55.1**	**55.3**	**55.8**	**55.8**	56.0
	Registration Recall (%) ↑									
SpinNet (Ao et al., 2021)	88.8	88.0	84.5	79.0	69.2	58.2	56.7	49.8	41.0	26.7
Predator (Huang et al., 2021)	89.0	89.9	90.6	88.5	86.6	59.8	61.2	62.4	60.8	58.1
CoFiNet (Yu et al., 2021)	89.3	88.9	88.4	87.4	87.0	67.5	66.2	64.2	63.1	61.0
YOHO (Wang et al., 2023)	90.8	90.3	89.1	88.6	84.5	65.2	65.5	63.2	56.5	48.0
RIGA (Yu et al., 2024)	89.3	88.4	89.1	89.0	87.7	65.1	64.7	64.5	64.1	61.8
GeoTrans (Qin et al., 2022)	92.0	91.8	91.8	91.4	91.2	75.0	74.8	74.2	74.1	73.5
Ours + GeoTrans	92.0	91.9	92.0	91.7	91.5	75.3	74.9	74.4	74.4	74.0
RoITr (Yu et al., 2023a)	91.9	91.7	91.8	91.4	91.0	74.7	74.8	74.8	74.2	73.6
Ours + RoITr	**94.5**	**94.0**	**93.8**	**93.5**	**93.6**	**75.9**	**75.7**	**75.7**	**75.7**	**75.8**

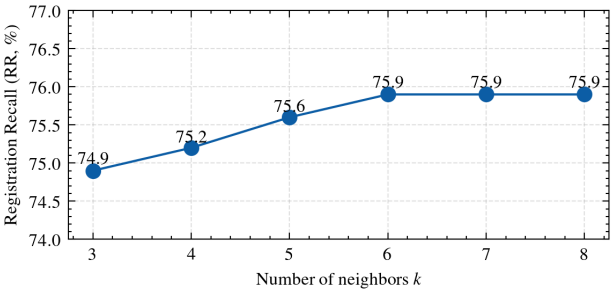


Figure 6. Sensitivity analysis of the hyperparameter k in the differentiable consistency module on 3DLoMatch.

estimation accuracy. Specifically, compared to RoITr (Yu et al., 2023a), we reduce the rotation error from 1.81° to 1.73° on 3DMatch, and from 2.92° to 2.66° on 3DLoMatch. Similar consistent improvements are observed when applying our module with GeoTransformer. The most significant gains in registration recall are achieved with the RoITr backbone, where the rotation-invariant features likely provide a more stable initialization for our consistency-guided module. On the KITTI dataset, **Ours**+GeoTrans reduces the rotation error from 0.25° to 0.19° and the translation error from 6.5 cm to 6.2 cm compared to GeoTrans. Notably, the orientation differences between consecutive scans are relatively small from the outset compared to the indoor datasets, which explains both the high overall accuracy and the modest improvement. These results confirm that our formulation produces more reliable correspondences, leading to more accurate and robust transformation estimation. Visualizations evaluated using RMSE, highlight improvements on 3DMatch and 3DLoMatch (Fig. 7). In challenging cases, such as two scans with low overlap of an incomplete chair (middle row) or an incomplete sofa (bottom row) where RoITr failed, our model achieved precise alignment close to the ground truth. We achieve RMSE on the order of several centimeters, improving by one to two orders of magnitude over competing approaches. These challenging cases further demonstrate the value of our formulation in handling incomplete and low-overlap scenarios.

4.2 Runtime and complexity analysis

To evaluate our model performance, we compare the inference-time runtime at test time against state-of-the-art methods (Table 3). The total runtime comprises two components: the *model time*, corresponding to feature extraction, and the *pose time*, corresponding to transformation estimation. As shown in Table (3), **Ours**+RoITr achieves the lowest total runtime when performing pairwise registrations. From the detailed runtime breakdown, **Ours**+RoITr and **Ours**+GeoTrans maintained a comparable *model time* to their respective backbone counterparts, exhibiting only a slight increase due to the additional neighbor-search operation introduced for feature association.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-179-2026 | © Author(s) 2026. CC BY 4.0 License.

Table 2. Registration error comparison on 3DMatch and 3DLoMatch benchmarks.

Methods	3DMatch			3DLoMatch		
	RR(%)↑	RRE(°)↓	RTE(m)↓	RR(%)↑	RRE(°)↓	RTE(m)↓
Predator Huang et al. (2021)	90.6	2.029	0.064	61.2	3.048	0.093
CoFiNet Yu et al. (2021)	88.4	2.011	0.061	64.2	3.280	0.094
GeoTrans (Qin et al., 2022)	91.2	1.758	0.063	73.5	2.716	0.097
Ours + GeoTrans	91.5 (+0.3)	1.748 (-0.010)	0.063 (–)	74.0 (+0.5)	2.690 (-0.026)	0.091 (-0.006)
RoITr (Yu et al., 2023a)	91.8	1.810	0.061	73.6	2.917	0.087
Ours + RoITr	**94.0** (+2.2)	**1.734** (-0.076)	**0.058** (-0.003)	**75.8** (+2.2)	**2.666** (-0.251)	**0.086** (-0.001)

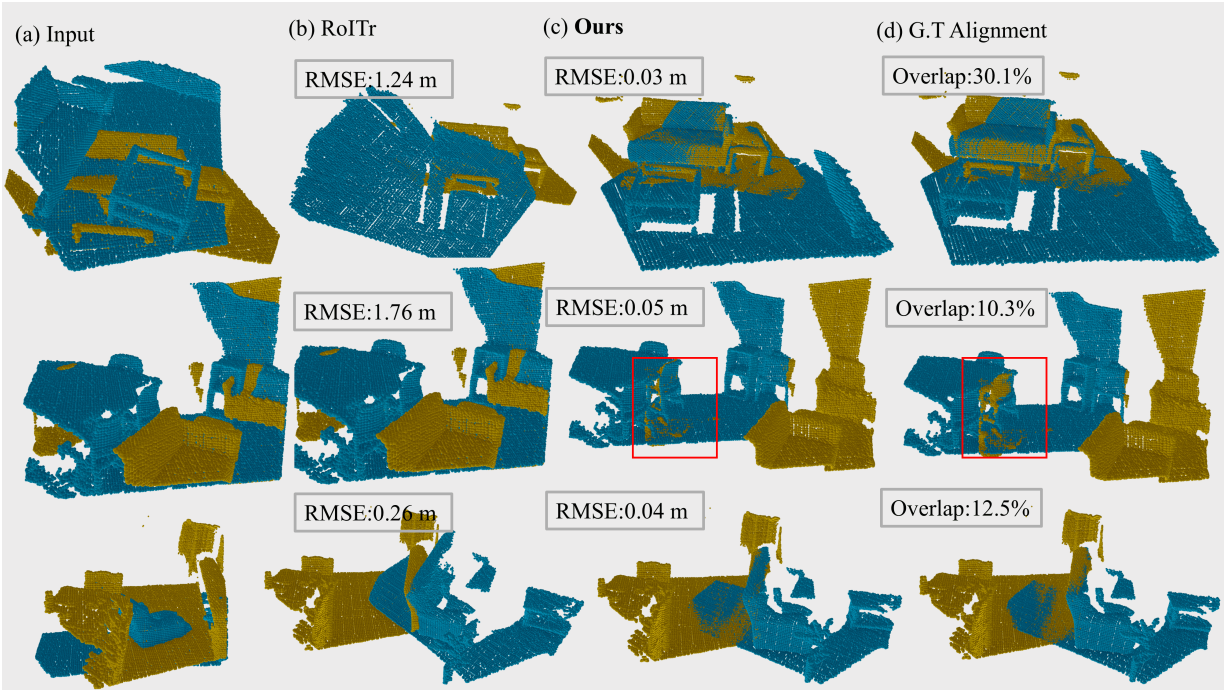


Figure 7. Registration comparison on 3DMatch and 3DLoMatch. Red squares indicate registered scan pairs with low overlap.

Table 3. Runtime comparison on on 3DMatch.

Methods	Model (s)↓	Pose (s)↓	Total (s)↓
Predator (Huang et al., 2021)	0.032	5.120	5.152
CoFiNet (Yu et al., 2021)	0.115	1.807	1.922
GeoTrans (Qin et al., 2022)	0.075	1.558	1.633
Ours + GeoTrans	**0.078**	1.547	1.625
RoITr (Yu et al., 2023a)	0.124	1.479	1.621
Ours + RoITr	0.127	**1.435**	**1.562**

Our proposed differentiable deep consistency module only adds 3 ms of latency to GeoTransformer and RoITr. This negligible overhead is compensated by the reduced *pose time* (1.43 s for **Ours**+RoITr vs. 1.48 s for RoITr), as the improved feature consistency leads to more stable and efficient transformation estimation. In terms of model capacity, **Ours**+RoITr contains approximately 10M parameters, equivalent to the best-performing RoITr configuration. With the same number of parameters, our model not only improves training convergence, as illustrated in Fig. (5), but also achieves consistent gains in inference efficiency.

5. Conclusions

Neural point cloud registration has demonstrated strong performance in registering scans for a wide range of applications in mapping and robotics. In this paper, we presented a novel end-to-end framework that integrates correspondence matching and verification within a single differentiable formulation. Through our proposed differentiable consistency module, the network evaluates the mutual compatibility of potential correspondences by ranking correlations, thereby highlighting neighborhood pairs with coherent local structures. This mechanism enables dynamic feature refinement in inconsistent regions to promote coherence. By embedding consistency verification directly into the attention mechanism, our network enforced geometric agreement during feature aggregation, learned expressive, distinctive, and spatially coherent point representations, and produced more reliable correspondences. We have shown the flexibility and generality of our formulation by seamlessly integrating it with state-of-the-art backbones such as GeoTransformer and RoITr. It has consistently outperformed state-of-the-art methods on both the 3DMatch, 3DLoMatch, and Kitti benchmarks, while maintaining comparable parameter complexity. In challenging low-overlap scenarios, our model has achieved the highest inlier ratio and the lowest registration error, successfully recovering correspondences where other methods fail. Furthermore, our method has demonstrated faster convergence during training and operates efficiently at inference. Overall, these results demonstrate the effectiveness, robustness, and generality of our consistency-driven formulation for accurate and efficient point cloud registration. Future work would extend our framework to more complex environments, e.g., forested or large-scale outdoor ones, and data streams acquired by

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-179-2026 | © Author(s) 2026. CC BY 4.0 License.

unmanned airborne scanners, where variable sampling density and cross-modal differences may be prevalent.

References

Ao, S., Hu, Q., Wang, H., Xu, K., Guo, Y., 2023. BUFFER: Balancing accuracy, efficiency, and generalizability in point cloud registration. *Proc. of CVPR*.

Ao, S., Hu, Q., Yang, B., Markham, A., Guo, Y., 2021. SpinNet: Learning a general surface descriptor for 3d point cloud registration. *Proc. of CVPR*.

Bai, X., Luo, Z., Zhou, L., Chen, H., Li, L., Hu, Z., Fu, H., Tai, C.-L., 2021. Pointdsc: Robust point cloud registration using deep spatial consistency. *Proc. of CVPR*, 15859–15869.

Bai, X., Luo, Z., Zhou, L., Fu, H., Quan, L., Tai, C.-L., 2020. D3Feat: Joint learning of dense detection and description of 3d local features. *Proc. of CVPR*.

Besl, P., McKay, N. D., 1992. A method for registration of 3-D shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*.

Chen, H., Yan, P., Xiang, S., Tan, Y., 2024. Dynamic cues-assisted transformer for robust point cloud registration. *Proc. of CVPR*, 21698–21707.

Chen, Y., Medioni, G., 1992. Object modelling by registration of multiple range images. *Image Vis. Comput.*, 10(3), 145–155.

Chen, Z., Sun, K., Yang, F., Tao, W., 2022. Sc2-pcr: A second order spatial compatibility for efficient and robust point cloud registration. *Proc. of CVPR*, 13221–13231.

Choy, C., Park, J., Koltun, V., 2019. Fully convolutional geometric features. *Proc. of ICCV*.

Deng, H., Birdal, T., Ilic, S., 2018. PPF-FoldNet: Unsupervised learning of rotation invariant 3d local descriptors. *Proc. of ECCV*, 602–618.

Dong, Z., Yang, B., Liang, F., Huang, R., Scherer, S., 2018. Hierarchical registration of unordered TLS point clouds based on binary shape context descriptor. *ISPRS J. Photogramm. Remote Sens.*, 144, 61–79.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. *Proc. of CVPR*.

Gojcic, Z., Zhou, C., Wegner, J. D., Andreas, W., 2019. The perfect match: 3d point cloud matching with smoothed densities. *Proc. of CVPR*.

Hu, X., Wu, J., Jia, M., Yan, H., Jiang, Y., Jiang, B., Zhang, W., He, W., Tan, P., 2025. Mapeval: towards unified, robust and efficient slam map evaluation framework. *IEEE Robot. Autom. Lett.*.

Huang, S., Gojcic, Z., Usvyatsov, M., Andreas Wieser, K. S., 2021. PREDATOR: Registration of 3d point clouds with low overlap. *Proc. of CVPR*.

Jiang, H., Dang, Z., Wei, Z., Xie, J., Yang, J., Salzmann, M., 2023. Robust outlier rejection for 3d registration with variational bayes. *Proc. of CVPR*, 1148–1157.

Li, J., Hu, Q., Ai, M., 2020. GESAC: Robust graph enhanced sample consensus for point cloud registration. *ISPRS J. Photogramm. Remote Sens.*, 167, 363–374.

Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Xu, K., 2022. Geometric transformer for fast and robust point cloud registration. *Proc. of CVPR*.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. SuperGlue: Learning feature matching with graph neural networks. *Proc. of CVPR*.

Theiler, P. W., Wegner, J. D., Schindler, K., 2015. Globally consistent registration of terrestrial laser scans via graph optimization. *ISPRS J. Photogramm. Remote Sens.*, 109, 126–138.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L. J., 2019. KPConv: Flexible and deformable convolution for point clouds. *Proc. of CVPR*.

Wang, H., Liu, Y., Dong, Z., Wang, W., 2022. You only hypothesize once: Point cloud registration with rotation-equivariant descriptors. *Proc. of ACM MM*, 1630–1641.

Wang, H., Liu, Y., Hu, Q., Wang, B., Chen, J., Dong, Z., Guo, Y., Wang, W., Yang, B., 2023. RoReg: Pairwise Point Cloud Registration with Oriented Descriptors and Local Rotations. *IEEE Trans. Pattern Anal. Mach. Intell.*.

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

Yin, H., Xu, X., Lu, S., Chen, X., Xiong, R., Shen, S., Stachniss, C., Wang, Y., 2024. A survey on global lidar localization: Challenges, advances and open problems. *Int. J. Comput. Vis.*, 132(8), 3139–3171.

Yu, H., Hou, J., Qin, Z., Saleh, M., Shugurov, I., Wang, K., Busam, B., Ilic, S., 2024. Riga: Rotation-invariant and globally-aware descriptors for point cloud registration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(5), 3796–3812.

Yu, H., Li, F., Saleh, M., Busam, B., Ilic, S., 2021. CoFiNet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *Proc. of NeurIPS*.

Yu, H., Qin, Z., Hou, J., Saleh, M., Li, D., Busam, B., Ilic, S., 2023a. Rotation-invariant transformer for point cloud matching. *Proc. of CVPR*.

Yu, J., Ren, L., Zhang, Y., Zhou, W., Lin, L., Dai, G., 2023b. PEAL: Prior-embedded explicit attention learning for low-overlap point cloud registration. *Proc. of CVPR*.

Zeng, A., Song, S., Nießner, M., Fisher, M., Xiao, J., Funkhouser, T., 2017. 3DMatch: Learning local geometric descriptors from rgb-d reconstructions. *Proc. of CVPR*.

Zhang, X., Yang, J., Zhang, S., Zhang, Y., 2023. 3d registration with maximal cliques. *Proc. of CVPR*, 17745–17754.

Zhao, G., Ao, S., Zhang, Y., Xu, K., Guo, Y., 2025. Progressive correspondence regenerator for robust 3d registration. *Proc. of CVPR*, 1210–1219.