ISSN: 2664-9640

Deep Learning for Point Cloud Registration: A Review of Core Architectures and Feature Extraction Strategies

Zhiqiu Zhang[1], Ye Chen[2]

[1] University of Shanghai for Science and Technology, Shanghai 200093, China

[2] University of Shanghai for Science and Technology, Shanghai 200093, China

Abstract

Point cloud registration aims to estimate the spatial transformation that aligns a source point cloud with a target point cloud, and constitutes a fundamental problem in three-dimensional reconstruction, robotic perception, autonomous driving, industrial metrology, and medical digitization. In recent years, deep learning has substantially reshaped the research paradigm of point cloud registration. The field has progressively shifted from traditional pipelines based on handcrafted geometric descriptors and local optimization toward data-driven frameworks that jointly learn feature representations, correspondence patterns, and rigid transformations. Focusing on the research question of which core architectures and feature extraction strategies are adopted by deep learning methods for point cloud registration, this review systematically analyzes the representative methodological routes developed in recent years. Specifically, we examine global feature regression architectures, correspondence learning architectures, coarse-to-fine hierarchical frameworks, and Transformer-based geometric attention models, and further discuss the evolution of feature learning from point-wise encoding to local geometric description, global context aggregation, multi-scale fusion, and cross-cloud interaction. On this basis, the review summarizes the intrinsic coupling among feature extraction, correspondence construction, and transformation estimation, compares the performance of representative models under a unified benchmark protocol, and outlines the main challenges and future directions in terms of generalization, efficiency, and geometric interpretability. This review is expected to provide a structured reference for subsequent studies on deep-learning-based point cloud registration.

Keywords

Point Cloud Registration; Deep Learning; Feature Extraction; Correspondence Learning; Transformer; Coarse-to-Fine Matching.

1. Introduction

Point cloud registration refers to the process of estimating the spatial transformation between two or more three-dimensional point sets so that the source point cloud can be geometrically aligned with the target point cloud in a common coordinate system[1]. This problem is central to tasks such as three-dimensional reconstruction, simultaneous localization and mapping, autonomous driving perception, industrial inspection, and medical digital modeling. Consequently, its accuracy, robustness, and efficiency directly affect the performance of upper-level systems.

Traditional point cloud registration methods typically rely on handcrafted local descriptors, nearest-neighbor matching, and iterative optimization, such as ICP and its variants[4,5], or global coarse registration based on local feature matching combined with RANSAC. These methods are geometrically interpretable, yet they often suffer performance degradation in

ISSN: 2664-9640

scenarios characterized by low overlap, large pose discrepancies, repetitive structures, noise, and high outlier ratios.

The introduction of deep learning has changed this situation. Compared with conventional methods, learning-based approaches can not only automatically extract discriminative geometric representations, but also integrate feature learning, correspondence construction, and transformation estimation into a unified optimization framework[1-3]. Accordingly, the central issue addressed in this review is not merely whether deep learning can improve point cloud registration, but rather through which architectural mechanisms and feature extraction strategies deep learning has driven this improvement.

2. Problem Definition and Analytical Framework

Let the source point cloud be denoted by $X=\{x_i\}$ and the target point cloud by $Y=\{y_j\}$, where $x_i, y_j \in R^3$. The objective of pairwise rigid point cloud registration is to estimate a rotation matrix R and a translation vector t such that the transformed source point cloud $R X + t$ is geometrically consistent with the target point cloud Y. This review primarily focuses on pairwise rigid registration, which remains the most extensively studied setting for deep-learning-based approaches[1-3].

From a methodological perspective, most deep-learning-based point cloud registration models can be decomposed into three interrelated stages: feature extraction, correspondence construction, and rigid transformation estimation. Under this unified pipeline, the major differences among methods lie in how features are encoded, when and how information is exchanged between the source and target point clouds, and how matching results are converted into a solvable geometric transformation.

A second important analytical dimension concerns the direction of information flow between the two point clouds. Early learning-based methods typically encoded the source and target point clouds independently and performed matching only in feature space. By contrast, more recent methods increasingly introduce cross-cloud interaction during encoding, so that each point representation is constrained not only by local geometry within its own cloud, but also by the contextual information from the other cloud. This shift indicates that feature learning itself has become an integral part of the matching mechanism.

The rigid registration objective can be formally expressed as $(R^*, t^*) = \arg\min_{R\in SO(3), t\in R^3} \Sigma_i ||R x_i + t - y_{\pi(i)}||^2$,

where $\pi(i)$ denotes the ideal correspondence mapping. Under the deep learning paradigm, the core problem is therefore transformed into the following question: how can a network learn intermediate representations that are both geometrically discriminative and capable of supporting stable correspondence inference and rigid pose estimation? The subsequent analysis thus focuses on the intrinsic relationship between architectural design and feature extraction strategy.

3. Core Architectures for Deep Learning-Based Point Cloud Registration

3.1. Global Feature Regression and Feature-Space Alignment

A representative line of research formulates point cloud registration as an alignment problem in a global feature space. Within this framework, the network first encodes the source and target point clouds into global representations, and then estimates the rigid transformation through feature differences, optimization in feature space, or end-to-end regression. Owing to their compact structures and relatively direct training procedures, such methods played an important role in the early development of deep-learning-based point cloud registration[8,9].

ISSN: 2664-9640

PointNetLK is a representative example of this category. It treats PointNet as a learnable feature mapping and borrows the Lucas–Kanade alignment strategy to iteratively update pose parameters in feature space[8]. DCP further extends the global representation framework by incorporating point embeddings, attention-based interaction, and differentiable SVD, thereby making the estimation process more closely coupled with geometric constraints[9].

Despite their structural elegance, global regression methods also exhibit clear limitations. Because transformation estimation mainly relies on global representations or weakly explicit correspondences, these architectures are more susceptible to local ambiguities in low-overlap scenarios, repetitive local structures, or large-scale scenes. They are therefore more suitable for registration tasks involving relatively high overlap and limited geometric variation.

3.2. Correspondence Learning Architectures

A second major line of research returns to the geometric essence of registration: establishing reliable correspondences before solving for the pose. Compared with direct regression methods, explicit correspondence learning is more consistent with the classical registration pipeline and offers stronger interpretability at intermediate stages. When registration fails, one can often trace the error back to concrete issues such as feature ambiguity, inaccurate overlap estimation, or insufficient correspondence filtering[10,13].

RPM-Net is an important representative of this direction. It learns hybrid features from spatial coordinates and local geometry, and constructs matching relationships through a Sinkhorn-based soft assignment matrix[10]. REGTR further advances this idea under a Transformer framework by directly predicting correspondences in an end-to-end manner[13].

Overall, correspondence learning architectures provide a more natural mechanism for converting similarity in feature space into a geometrically solvable relationship in Euclidean space, and have therefore become one of the dominant methodological routes in recent years. Their upper performance bound depends primarily on the quality of the inferred correspondences rather than simply on network depth.

3.3. Coarse-to-Fine Hierarchical Architectures

A third class of methods addresses the challenges of low overlap, severe occlusion, and excessively large search spaces in real-world scenes by adopting coarse-to-fine hierarchical registration strategies[11,12]. The key idea is to first establish relatively reliable sparse correspondences at a coarse scale, thereby reducing the search space, and then progressively refine local geometric alignment at finer scales.

PREDATOR is a representative method for low-overlap registration. Through its overlap-attention mechanism, the two point clouds exchange information at an early encoding stage, while overlap scores and matchability scores are explicitly predicted to emphasize potentially overlapping regions and suppress non-overlapping areas[11].

CoFiNet, in contrast, extracts coarse correspondences and fine correspondences in a hierarchical manner and completes local refinement through multi-level propagation[12]. The contribution of coarse-to-fine architectures lies not merely in multi-resolution engineering, but in their ability to decompose an otherwise difficult global matching problem into a sequence of easier subproblems under hierarchical constraints.

3.4. Transformer and Geometry-Aware Attention Architectures

The rapid development of Transformer-based methods for point cloud registration reflects a shift in research emphasis from merely enhancing local descriptors to modeling richer global context and cross-cloud interactions. Self-attention captures long-range dependencies within a point cloud, whereas cross-attention directly models relationships between the source and

ISSN: 2664-9640

target point clouds. These properties make Transformers particularly suitable for dealing with repetitive structures, long-distance dependencies, and low-overlap configurations[13,14].

REGTR employs Transformer layers to generate correspondences directly[13]. GeoTransformer further argues that generic attention alone is insufficient to stably capture geometric constraints, and therefore explicitly incorporates rigid-transformation-invariant geometric quantities, such as pairwise distances and triplet angles, into the attention mechanism. This design substantially improves matching robustness under low-overlap conditions[14].

It should be noted that Transformer architectures are not without cost. Attention mechanisms are typically associated with higher computational and memory overhead, and many methods therefore combine them with downsampling, superpoint modeling, or local-window strategies to control complexity. Nevertheless, geometry-aware attention has become one of the most important directions in current high-performance point cloud registration.

Table 1. Comparison of core architectural paradigms in deep-learning-based point cloud registration.

Architectural paradigm	Representative methods	Core mechanism	Main strengths	Typical limitations
Global feature regression	PointNetLK; DCP	Alignment of global latent representations or optimization in feature space	Compact pipeline; amenable to end-to-end training; elegant formulation	Sensitive to low overlap and local ambiguity
Correspondence learning	RPM-Net; REGTR	Soft/hard correspondences are first predicted, followed by pose estimation	Strong interpretability; closer to the classical geometric pipeline	Performance heavily depends on correspondence quality
Coarse-to-fine hierarchy	PREDATOR; CoFiNet	Reliable coarse correspondences are established first and then progressively refined	Suitable for low-overlap scenes; balances robustness and precision	More complex multi-stage design
Transformer / geometry-aware attention	REGTR; GeoTransformer	Self-/cross-attention with geometric priors for cross-cloud interaction	Strong contextual modeling; high performance under low overlap	Relatively high computational and memory cost

Overall, the four architectural paradigms reviewed above reveal a clear methodological evolution. Early work primarily investigated whether neural networks could replace or enhance handcrafted features; intermediate studies gradually incorporated correspondences into the learning framework explicitly; and more recent methods increasingly emphasize robustness under low overlap, cross-cloud interaction, and geometry-aware priors. Hence, performance gains in point cloud registration are driven not simply by deeper networks, but by more effective decomposition and joint optimization of the feature–correspondence–pose pipeline.

ISSN: 2664-9640

4. Feature Extraction Strategies

4.1. From Point-Wise Encoding to Local Geometric Description

Feature extraction is the component most directly influenced by deep learning in point cloud registration. Early learning-based methods usually relied on point coordinates and obtained point-wise features through shared MLPs or lightweight mappings, followed by global aggregation or simple interaction for registration[8,9]. The advantage of point-wise encoding lies in its preservation of raw geometric detail; however, without neighborhood context, it remains sensitive to noise, density variation, and local structural ambiguity.

Subsequently, research attention shifted toward local geometric structure modeling. By learning local volumetric descriptors from RGB-D reconstructions, 3DMatch advanced the application of learned local features to point cloud registration[6]; FCGF later used a fully convolutional structure to generate dense local geometric features efficiently in large-scale scenes, thereby greatly improving both discriminability and computational efficiency[7].

The importance of local geometric features stems from the fact that rigid registration is fundamentally a problem of relational modeling. An isolated point is often ambiguous, whereas its neighborhood shape, relative orientation, and local surface structure provide a far more stable basis for matching. Consequently, modern registration methods rarely remain at the level of single-point encoding, and instead generally regard local geometry as the core foundation of feature learning.

4.2. Global Context Aggregation

Although local descriptors form the basis of registration, they are often insufficient to resolve matching ambiguity in the presence of repetitive structures, self-similar surfaces, or incomplete local evidence. As a result, an increasing number of methods incorporate global contextual information to help the model judge whether a local feature is trustworthy, whether it lies within the overlap region, and whether it satisfies broader structural consistency with other candidates.

It should be emphasized that global context is not a substitute for local geometry, but rather a higher-level constraint imposed on local representations. Its role is to reduce the ambiguity of local features and enable the model to understand structural relations over a larger spatial extent. This idea is reflected in the attention interaction of DCP[9] and the overlap-aware information exchange of PREDATOR[11].

4.3. Multi-Scale Fusion

Multi-scale fusion has become a hallmark of high-performance point cloud registration systems, because geometric evidence is distributed across different spatial scales. Coarse scales are more suitable for representing overall structure and long-range dependencies, whereas fine scales preserve local boundaries and surface details. If only a single scale is used, it is difficult for a model to achieve both robustness and precision simultaneously.

CoFiNet is a typical representative of the multi-scale paradigm. It first extracts reliable correspondences at a coarse scale and then propagates information to finer scales for local refinement[12]. Similarly, GeoTransformer also embodies a hierarchical propagation process from superpoints to denser correspondences[14].

Conceptually, multi-scale fusion can be understood as a trade-off between robustness and resolution: coarse representations reduce noise interference and shrink the search space, while fine representations recover geometric detail. The frequent adoption of multi-scale design in modern registration architectures is therefore a direct consequence of the need to simultaneously maintain global stability and local precision.

ISSN: 2664-9640

4.4. Cross-Cloud Interaction as a Feature Learning Mechanism

A major recent shift is the realization that feature extraction in registration should not be performed independently on the source and target point clouds. If the two point clouds can share information during encoding, each point representation is influenced not only by its local structure, but also by the context of potential corresponding regions in the other cloud.

PREDATOR's overlap-attention is an early representative of this idea, enabling the network to emphasize points that are both locally discriminative and likely to lie in overlapping regions[11]. REGTR and GeoTransformer implement cross-cloud interaction more systematically under Transformer frameworks, thereby making feature learning itself directly oriented toward matching[13,14].

This shift is highly relevant to the research question addressed in this review. The most effective feature extraction strategies are no longer those that merely learn stronger descriptors, but those that learn representations capable of maintaining geometric alignability throughout cross-cloud interaction. In other words, the boundary between feature extraction and matching is increasingly blurred.

5. Correspondence Construction, Transformation Estimation, and Evaluation

5.1. Matching Mechanisms

The practical value of feature extraction only emerges when features are actually used to construct correspondences. Current matching mechanisms in deep-learning-based point cloud registration can be broadly divided into three categories: hard matching based on nearest-neighbor retrieval, soft matching based on assignment matrices or optimal transport, and direct prediction of correspondence locations through attention or regression[10,13].

RPM-Net is a typical representative of soft matching mechanisms. Its Sinkhorn-based matching layer preserves multiple hypotheses during the early stage of optimization, thereby better handling the uncertainty introduced by noise, outliers, and partial overlap[10]. Methods such as REGTR attempt to simplify the pipeline further by directly predicting correspondences and overlap probabilities through the network, reducing the dependence on traditional matching and filtering stages[13].

These advances suggest that correspondence construction has become a major point of competition in point cloud registration research. Performance improvements increasingly arise from higher-quality matching representations rather than from more complex pose regressors alone.

5.2. Rigid Transformation Solvers

Once correspondences or matching probabilities have been obtained, the system must estimate the final rigid transformation. Current methods generally favor explicit geometric solvers over completely black-box parameter regression. Typical approaches include solving for rotation and translation through SVD or Procrustes analysis, or performing iterative refinement on the basis of an initial estimate.

The advantage of explicit rigid solvers lies in their preservation of the algebraic structure of SE(3), which improves numerical stability and alleviates the burden on the network to learn a direct mapping in pose space. The differentiable SVD in DCP, the iterative updates in PointNetLK, and the weighted rigid alignment in RPM-Net all demonstrate the effectiveness of integrating analytic geometry modules with neural networks[8-10].

ISSN: 2664-9640

5.3. Benchmark Datasets and Evaluation Metrics

Evaluation protocols, in turn, influence architectural design. Indoor benchmarks such as 3DMatch and 3DLoMatch emphasize registration capability under RGB-D reconstruction noise and partial overlap, whereas outdoor benchmarks such as KITTI place more emphasis on large-scale scenes, sparse sampling, and ego-motion[6,15]. Different benchmarks thus place different demands on low-overlap robustness, contextual modeling, and computational efficiency.

Common evaluation metrics include registration recall, inlier ratio, rotation error, translation error, and root-mean-square alignment error. For a review study, understanding the relationship between evaluation metrics and dataset settings is essential, because some architectural advantages become fully visible only under specific evaluation protocols.

Table 2. Performance comparison of different deep learning models on the 3DMatch test set (all evaluated under the unified RANSAC-50k protocol with 1000 sampled correspondences).

Model	IR (%)	FMR (%)	RR (%)
FCGF	48.7	97.0	83.3
D3Feat	40.4	94.5	83.4
SpinNet	39.8	97.6	84.1
PREDATOR	58.0	96.6	89.0
CoFiNet	61.3	98.1	89.3
GeoTransformer	71.8	97.9	91.8
YOHO	64.0	98.2	89.6
RIGA	67.6	98.0	90.8
OIF-PCR	72.3	97.9	91.8
RoITr	75.3	97.1	91.8
HST	74.0	98.5	92.6

Note: The numerical results in the table are taken from publicly reported comparisons under the unified 3DMatch protocol, using RANSAC-50k and 1000 sampled correspondences.

5.4. Comparative Analysis under a Unified Benchmark Protocol

As shown in Table 2, when registration recall (RR) is used as the principal criterion, HST achieves the highest score of 92.6% on 3DMatch, while GeoTransformer, OIF-PCR, and RoITr all reach 91.8%, forming the second-best tier. In terms of inlier ratio (IR), RoITr attains 75.3%, slightly exceeding HST (74.0%) and OIF-PCR (72.3%), which suggests that different methods do not necessarily dominate all metrics simultaneously[16].

A further comparison of feature match recall (FMR) shows that many recent methods are already approaching saturation on this metric. HST reaches 98.5%, only marginally higher than CoFiNet at 98.1% and OIF-PCR and GeoTransformer at 97.9%. This indicates that FMR alone is no longer sufficient to clearly distinguish high-performing models. In contrast, RR remains more discriminative because it reflects the effectiveness of the entire pipeline, including correspondence quality, geometric verification, and transformation estimation[17].

5.5. Open Challenges

Despite the substantial progress achieved in this field, several unresolved problems remain. First, cross-domain generalization is still limited: a model trained on indoor RGB-D data may not transfer robustly to LiDAR, industrial scanning, or medical point clouds. Second, under extremely low overlap, high outlier ratios, and severe occlusion, even state-of-the-art methods may still become unstable.

Another issue worth noting concerns the realism of benchmark datasets. Public datasets provide a unified platform for comparison, yet real-world applications often involve more

ISSN: 2664-9640

complex sampling patterns, noise distributions, scale variation, and semantic structure. Therefore, performance on standard benchmarks alone is insufficient to fully reflect the engineering usability of a method.

5.6. Future Research Directions

Future research is likely to proceed along four interrelated directions. First, geometric priors and equivariant learning may be integrated more tightly into neural architectures to strengthen the explicit modeling of rigid structure. Second, stronger self-supervised or weakly supervised mechanisms are needed to reduce dependence on precise pose annotations. Third, further progress in lightweight design and real-time inference is required to satisfy the demands of robotics, autonomous driving, and online industrial inspection.

From the perspective of feature extraction, future development should not focus solely on building more complex encoders, but rather on learning representations that remain geometrically comparable and stable for matching across different sensors, overlap ratios, and noise conditions. Consequently, the coordinated optimization of feature learning, correspondence construction, and pose estimation will continue to be a central theme in subsequent research.

6. Conclusion

This review has examined the core architectures and feature extraction strategies of deep-learning-based point cloud registration. Overall, the field has evolved from early global feature regression toward explicit correspondence learning, coarse-to-fine hierarchical matching, and Transformer-based joint modeling enhanced by geometric priors.

From the viewpoint of representation learning, the dominant trajectory has progressed from point-wise encoding to local geometric modeling, and then to global context aggregation, multi-scale fusion, and cross-cloud interactive feature learning. Effective features are not merely discriminative; more importantly, they must support correspondence construction and rigid transformation estimation in a stable manner. Future research must therefore continue to improve generalization, efficiency, and geometric interpretability in order to move point cloud registration closer to robust real-world deployment.

References

[1] Z. Zhang, Y. Dai and J. Sun: Deep learning based point cloud registration: an overview, Virtual Reality & Intelligent Hardware, Vol. 2 (2020) No. 3, p. 222-246.

[2] M. Lyu, J. Yang, Z. Qi, R. Xu, J. Liu and Y. Liu: Rigid pairwise 3D point cloud registration: A survey, Pattern Recognition, Vol. 151 (2024), 110408.

[3] Y.-X. Zhang, J. Gui, B. Yu, X. Cong, X. Gong, W. Tao and D. Tao: Deep Learning-Based Point Cloud Registration: A Comprehensive Survey and Taxonomy, arXiv:2404.13830 (2024).

[4] P. J. Besl and N. D. McKay: A method for registration of 3-D shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 14 (1992) No. 2, p. 239-256.

[5] S. Rusinkiewicz and M. Levoy: Efficient variants of the ICP algorithm, Proc. Third International Conference on 3-D Digital Imaging and Modeling (Quebec City, Canada, May 28-June 1, 2001), p. 145-152.

[6] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao and T. Funkhouser: 3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions, Proc. IEEE Conference on Computer Vision and Pattern Recognition (Honolulu, USA, July 21-26, 2017), p. 199-208.

[7] C. Choy, J. Park and V. Koltun: Fully Convolutional Geometric Features, Proc. IEEE/CVF International Conference on Computer Vision (Seoul, Korea, October 27-November 2, 2019), p. 8958-8966.

ISSN: 2664-9640

[8] Y. Aoki, H. Goforth, R. A. Srivatsan and S. Lucey: PointNetLK: Robust & Efficient Point Cloud Registration Using PointNet, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (Long Beach, USA, June 16-20, 2019), p. 7163-7172.

[9] Y. Wang and J. M. Solomon: Deep Closest Point: Learning Representations for Point Cloud Registration, Proc. IEEE/CVF International Conference on Computer Vision (Seoul, Korea, October 27-November 2, 2019), p. 3523-3532.

[10] Z. J. Yew and G. H. Lee: RPM-Net: Robust Point Matching Using Learned Features, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (Seattle, USA, June 13-19, 2020), p. 11824-11833.

[11] S. Huang, Z. Gojcic, M. Usvyatsov, A. Wieser and K. Schindler: Predator: Registration of 3D Point Clouds with Low Overlap, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (virtual, June 19-25, 2021), p. 4267-4276.

[12] H. Yu, F. Li, M. Saleh, B. Busam and S. Ilic: CoFiNet: Reliable Coarse-to-Fine Correspondences for Robust Point Cloud Registration, Advances in Neural Information Processing Systems, Vol. 34 (2021), p. 23872-23884.

[13] Z. J. Yew and G. H. Lee: REGTR: End-to-End Point Cloud Correspondences with Transformers, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (New Orleans, USA, June 18-24, 2022), p. 6677-6686.

[14] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng and S. Ilic: Geometric Transformer for Fast and Robust Point Cloud Registration, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (New Orleans, USA, June 18-24, 2022), p. 11143-11152.

[15] A. Geiger, P. Lenz and R. Urtasun: Are we ready for autonomous driving? The KITTI vision benchmark suite, Proc. IEEE Conference on Computer Vision and Pattern Recognition (Providence, USA, June 16-21, 2012), p. 3354-3361.

[16] J. Qian and D. Tao: RRGA-Net: Robust Point Cloud Registration Based on Graph Convolutional Attention, Sensors, 2023, 23(24): 9710.

[17] C. Ren, Y. Feng, W. Zhang, X.-P. Zhang and Y. Gao: Multi-scale Consistency for Robust 3D Registration via Hierarchical Sinkhorn Tree, Advances in Neural Information Processing Systems, Vol. 37 (2024), pp. 63991-64020.