ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0002

Alignment Study of Oral Cavity Point Cloud based on Improved Deep-interaction Transformer

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

In digital oral treatment, the accuracy of point cloud registration has a significant impact on treatment outcomes. To address the limitations of traditional point cloud registration methods when dealing with sparse and partially missing point clouds, an improvement based on the Deep Interaction Transformer (DIT) model is proposed. This study introduces new loss functions and an oral point cloud self-attention module to enhance the registration accuracy of single-frame oral point clouds.The proposed loss functions include symmetry loss of dental curves, smoothness loss of dental surfaces, and contour consistency loss, each optimized to cater to the geometric characteristics of oral point clouds. Additionally, the designed self-attention module, named Oral Point Cloud Self-Attention (OPCSA), enables more refined information interaction through point-wise calculation of queries, keys, and values, thereby enhancing feature extraction capabilities and robustness.Experimental results demonstrate that the improved DIT model significantly outperforms the baseline model in tasks of single-frame point cloud registration for oral and dental models. Not only does this research provide a new solution for high-precision registration of oral point clouds, but it also shows promising potential in terms of methodological effectiveness and application feasibility. The approach sets a foundation for future advancements in digital dental treatment and point cloud processing technologies.

Keywords

Oral Medicine; Point Cloud Registration; Deep Learning; Attention Mechanism.

1. Introduction

In recent years, oral health has received increasing attention, and the implementation of digital dental treatment plans relies on accurate dental impressions[1]. Traditional methods require placing a tray coated with impression material into the patient's mouth, waiting for a period of time, then removing it and using plaster to cast a hard mold that matches the original shape of the teeth[2]. Undoubtedly, this method is not suitable for all patients, and its accuracy is limited by the materials and the dentist's technique. Therefore, in order to improve the efficiency and precision of dental treatment, three-dimensional scanning technology has become the mainstream method for obtaining dental impressions in recent years[3]. High-precision modeling of a patient's oral condition must rely on modern point cloud registration algorithms.

Currently, 3D oral scanning technology divides the point cloud registration process into two steps. First, coarse registration is performed using Fast Point Feature Histograms (FPFH) and Random Sample Consensus (RANSAC). Specifically, the FPFH algorithm extracts a 33-dimensional feature descriptor for each point in both the target and source point clouds[4]. The Euclidean distances between the FPFH descriptors of the target and source point clouds are calculated to select similar

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0002

feature pairs. RANSAC then randomly selects point pairs from the matched features and estimates an initial rigid transformation matrix. This estimated matrix is applied to the source point cloud to compute its overlap with the target point cloud. By repeatedly executing this process, the optimal transformation matrix is obtained[5]. Next, fine registration is performed using the Iterative Closest Point (ICP) algorithm, which uses the optimal transformation matrix obtained in the first step as the initial value. ICP finds the nearest neighbor points between the two clouds and applies the least squares method to solve for the optimal transformation matrix. These steps iterate continuously until a termination condition is met, such as the distance between nearest neighbor pairs falling below a threshold or the change in the rotation and translation matrices between two iterations being sufficiently small[6]. Although this method is robust, resistant to outliers and noise, and relatively efficient, its performance is limited by the incomplete feature extraction of FPFH and ICP's dependence on a high-quality initial transformation matrix. In recent years, the success of deep learning in point cloud processing has prompted researchers to explore its application in point cloud registration. PointNetLK extracts global features of point clouds using the PointNet architecture and incorporates the traditional Lucas-Kanade algorithm to achieve end-to-end registration. However, due to the lack of information interaction between the source and target point clouds, it performs poorly on point clouds with partial visible features. Robust Point Matching Network (RPM-Net) improves robustness to noise and outliers by introducing learned feature distances in place of traditional spatial distances. It also employs a Sinkhorn layer and an annealing mechanism to perform point cloud registration[7]. However, the annealing mechanism results in high computational complexity and limited generalization. Inspired by the success of attention mechanisms in natural language processing and computer vision, researchers have begun incorporating them into point cloud registration. In addition, several emerging methods use attention mechanisms to aggregate contextual information, aiming to improve feature extraction and matching accuracy. GeoTransformer encodes the geometric structure of point clouds to learn features and adopts a superpoint strategy, reducing reliance on keypoint detection, discarding RANSAC, and improving registration efficiency. DIT simulates dependencies across the entire point cloud using attention mechanisms and introduces correspondence confidence estimation based on geometric matching to filter out high-quality correspondences, thereby improving registration accuracy and success rates[8]. However, due to the high computational complexity and large data requirements of Transformers, the training process remains challenging.

To address the limitations of the traditional two-stage registration approach—namely incomplete feature extraction, limited success rate, and accuracy—we propose a novel deep learning-based point cloud registration algorithm (HIT) in this paper. HIT enhances the relative positional representation between the source and target point clouds through a position encoding network[9]. It also employs a hybrid interaction attention mechanism that combines point convolution (PConv) and relative attention (Rel Attention) to achieve global modeling of geometric features and spatial information within the point clouds. This design overcomes the training inefficiency of Transformer architectures in small-sample scenarios. In addition, a loss function tailored to the specific geometric structure of oral point clouds is introduced. Furthermore, we construct a point cloud dataset obtained via intraoral 3D scanning devices, including real intraoral point clouds and dental model point clouds made of different materials, thus filling the current gap in point cloud datasets within the domain of oral healthcare.

2. Point Cloud Registration based on Hybrid Interaction Attention Section Headings

2.1 Overall Model Architecture

The overall architecture of the proposed HIT model is illustrated in Figure 1. It is designed as an end-to-end deep learning framework that takes a pair of unordered 3D point clouds-typically representing source and target dental structures-and estimates the optimal rigid transformation $\{R, t\}$ aligning

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0002

them. The model is modular, with each component tailored to address specific challenges in oral point cloud registration, including sparsity, anatomical asymmetry, and spatial ambiguity.

The first stage involves feature encoding, where local geometric features are extracted from both the source and target point clouds. This is achieved through a dual-path encoding scheme combining Local Feature Description (LFD) and Multi-head Self Attention. The LFD module leverages K-nearest neighbor aggregation and shared-weight MLPs to capture fine-grained spatial patterns within local neighborhoods. In parallel, self-attention is employed to model global contextual dependencies across the point cloud, enhancing the representation of complex anatomical structures such as occlusal surfaces and dental ridges[10].

Once local and global features are extracted, the model applies hybrid interaction attention, a key innovation of our approach. This mechanism integrates point convolution and relative attention operations in a point-wise manner. Point convolution captures low-level geometric interactions by aggregating features from spatially adjacent points, while relative attention introduces a spatially-aware bias into the attention weights, allowing the network to reason about long-range positional relationships. This hybrid strategy ensures that both local geometry and global spatial cues are preserved and effectively encoded.

To further refine the learned features, a Squeeze-and-Excitation (SE) module is introduced. This component performs channel-wise recalibration by adaptively adjusting the contribution of each feature dimension based on its global statistical relevance. In doing so, it suppresses noise and enhances salient channels, enabling the model to focus on structurally informative regions of the point cloud, such as cusp tips, crown margins, and dental arch centers.

Following feature extraction and enhancement, the model enters the correspondence estimation stage. Here, the similarity between each point in the source and target point clouds is computed based on their respective feature embeddings. These similarity scores are used to establish soft correspondences between the two clouds, avoiding hard point matching and allowing for better handling of partial occlusions and non-uniform sampling.

Finally, the matched points and their similarity scores are fed into a weighted Procrustes alignment module, which estimates the optimal rigid-body transformation $\{R, t\}$ that aligns the source point cloud to the target. The weight matrix ensures that more reliable correspondences contribute more heavily to the estimation, improving robustness to noise and outliers. The Procrustes solution is computed in closed form via singular value decomposition (SVD) of the cross-covariance matrix constructed from matched point pairs.

Overall, the architecture integrates local feature encoding, global attention, feature calibration, and geometric alignment into a cohesive pipeline optimized for oral point cloud registration. Its design is guided by the anatomical complexity and clinical precision requirements of intraoral data, enabling accurate and robust alignment in both model-based and real-world scenarios.

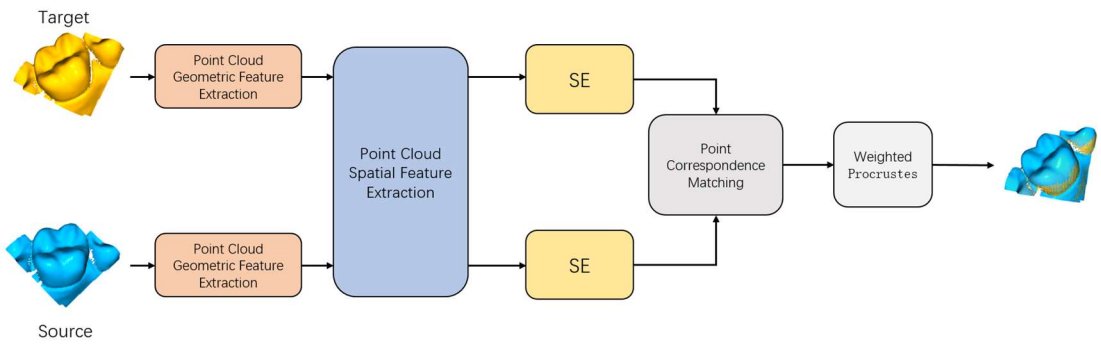

Figure 1. Overall Network Architecture

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0002

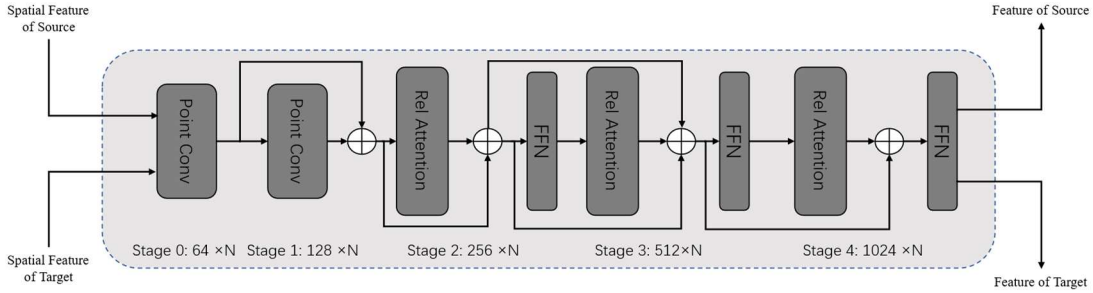

Figure 2. Hybrid Interaction Network Architecture for Point Clouds

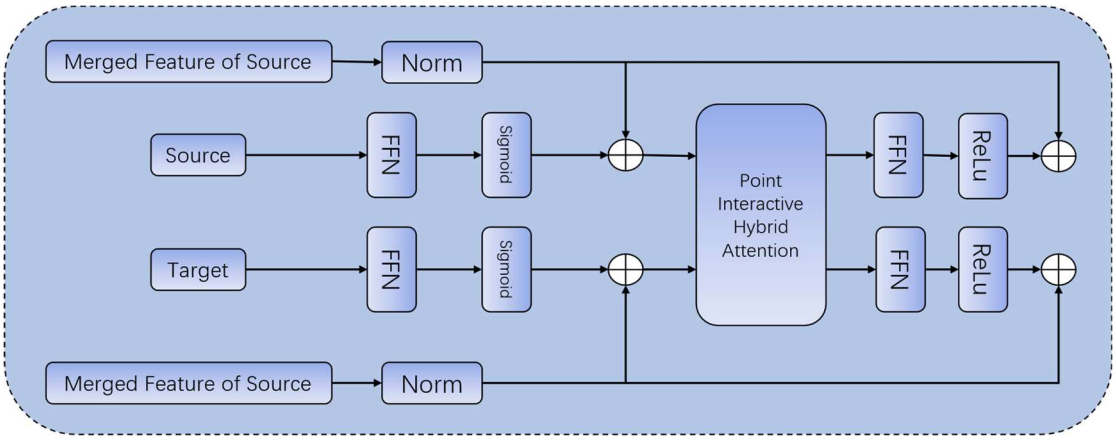

Figure 3. Spatial Feature Extraction Module for Point Clouds

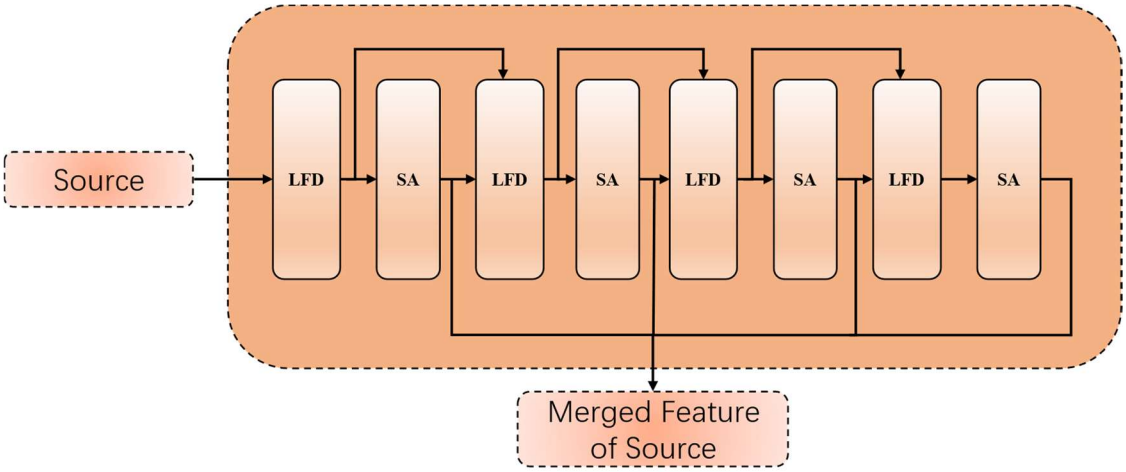

Figure 4. Geometric Feature Extraction Module for Point Clouds

2.2 Geometric Feature Extraction of Pointcloud

The geometric feature extraction module for point clouds serves as a crucial foundation for robust and accurate registration. In the proposed model, this module is composed of two key components that work in tandem to capture both local and global structural information from sparse, unordered point cloud data.

(1) Local Feature Description (LFD): The first stage focuses on the extraction of local geometric features that preserve spatial structures within a neighborhood. Specifically, we employ the K-Nearest Neighbors (KNN) algorithm in Euclidean space to construct local neighborhoods for each point. For a given query point, the algorithm selects the K closest points based on spatial proximity, forming a compact geometric context. These neighborhoods are then indexed to facilitate the construction of localized feature representations. To capture the intricate structural variations within these

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0002

neighborhoods, we compute relative positional offsets between each neighbor and the central point. These offsets serve as the input to shared-weight Multi-Layer Perceptrons (MLPs), enabling the network to learn nonlinear mappings from raw spatial coordinates to higher-dimensional feature embeddings. This procedure not only preserves local topological relationships but also introduces a convolutional inductive bias that improves the generalizability and data efficiency of the network. Unlike handcrafted descriptors such as FPFH, which lack adaptability, the learned local descriptors are dynamically optimized during training, leading to more discriminative features tailored to the specific geometry of oral point clouds. Furthermore, by leveraging the permutation invariance property inherent in point clouds, we apply symmetric functions (e.g., max pooling or average pooling) to aggregate the features within each neighborhood. This process effectively reduces sensitivity to point ordering while maintaining essential geometric information. The combination of KNN-based sampling, relative coordinate encoding, and shared-weight learning forms a lightweight yet expressive module for local structural modeling.

(2) Multi-head Attention: Once the local features are computed, we proceed to enhance global contextual understanding through a multi-head self-attention mechanism. First, a 1×1 convolution is applied to the local descriptors to produce query (Q), key (K), and value (V) matrices for each point. These matrices encode the learned representations in a way that enables pairwise interaction across the entire point cloud.

The attention weights are calculated using scaled dot-product attention, where the similarity between Q and K is computed as $\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, with d_k denoting the dimensionality of the key vectors. The resulting attention scores are used to weight the value vectors, and the outputs from multiple attention heads are concatenated and linearly projected to form the final representation. By employing multiple heads, the network captures diverse types of relationships among points - some heads may focus on spatial proximity, while others may attend to points with similar curvature or density characteristics. This multi-perspective interaction allows the network to model long-range dependencies and global spatial correlations, which are particularly important in oral point clouds where dental arches and occlusal surfaces span large spatial extents. Moreover, self-attention enhances the model's robustness to noise and missing data by allowing each point to gather contextually relevant information from the entire cloud, rather than relying solely on fixed-radius neighborhoods. This is especially valuable in intraoral scans where partial occlusions or sparse sampling are common. The ability to dynamically reweight feature contributions based on learned similarities endows the model with adaptive feature refinement capabilities, leading to improved discriminative power and better alignment performance. In summary, the integration of LFD and multi-head attention results in a hierarchical representation that captures fine-grained local structures and higher-order global relationships simultaneously. This dual-scale feature encoding forms a critical backbone for accurate point cloud correspondence estimation and registration in the downstream modules.

2.3 Hybrid Interaction Attention for Pointclouds

The hybrid interaction attention mechanism combines point convolution layers and attention layers, leveraging the strengths of both to construct a model that remains efficient and high-performing across varying data scales. Inspired by the PointNet model's approach to handling the sparsity and unordered nature of point cloud data, point convolution captures low-level features using shared-weight MLPs and symmetric functions (e.g., max pooling). To address the high computational complexity and poor generalization of traditional attention mechanisms, relative attention is introduced to extract high-level features from the source and target point clouds. A relative bias matrix P is added to the standard attention computation, storing the positional offset Pij between each pair of points iii and jjj, defined as:

$$P_{i,j} = (x_i - x_j, y_i - y_j, z_i - z_j)$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0002

The relative attention is computed as:

$$Attention(Q,K,V) = Softmax(\frac{QK^T + P_{i,j}}{\sqrt{d_k}})V$$

This adjustment ensures that attention weights incorporate both the dot-product similarity of Q and K and their relative spatial bias, allowing the model to better consider the spatial relationship between source and target point clouds. Moreover, a multi-stage design is adopted, in which different types of network blocks are used at each stage to progressively increase feature dimensionality, enabling the model to transition from low-level to high-level feature extraction.

2.4 Feature Enhancement

To further improve the discriminative capacity of the extracted features, a feature enhancement module is integrated into the architecture to dynamically recalibrate channel-wise feature responses. This module is designed to emphasize informative features while suppressing less relevant or noisy components, thereby improving the robustness and expressiveness of the network under challenging scenarios such as sparse sampling and partial occlusions commonly observed in oral point clouds.

The enhancement strategy is implemented using a channel-wise attention mechanism, inspired by the principles of Squeeze-and-Excitation (SE) networks. The core idea is to model the interdependencies among feature channels and use this learned information to adaptively reweight the feature maps.

The process begins with global average pooling applied independently to each feature channel. This operation compresses the spatial dimension, transforming each channel into a scalar that summarizes its overall activation. The resulting vector captures global contextual information and serves as a compact descriptor of the relative importance of each channel. Notably, in the context of 3D point clouds, this global pooling effectively captures shape-level statistics, which are critical for recognizing symmetric structures such as dental arches or identifying localized anomalies such as caries or implants.

Subsequently, the aggregated channel descriptors are passed through a fully connected (FC) layer, often accompanied by a non-linear activation function (e.g., ReLU or Sigmoid), to learn a set of weights that quantify the importance of each channel with respect to the given input. These learned weights encode high-level semantic relevance and are conditioned on the specific geometry and spatial distribution of the point cloud instance.

The final step involves applying the learned weights to the original feature maps via channel-wise multiplication. Channels with higher learned weights are selectively amplified, allowing the network to attend more strongly to structurally salient or diagnostically significant features. Conversely, channels associated with weak or irrelevant features are suppressed, reducing the influence of background noise or redundant information.

This adaptive feature recalibration not only improves the signal-to-noise ratio within the learned representations but also facilitates better gradient flow during backpropagation, leading to improved convergence behavior. In the case of oral point cloud registration, where minor geometric deviations (e.g., crown tilts or occlusal asymmetries) can critically affect alignment accuracy, the ability to prioritize informative channels enhances the model's sensitivity to clinically relevant structural cues.

Moreover, the lightweight nature of the channel attention mechanism ensures that the overall computational overhead remains minimal, making it suitable for real-time or resource-constrained dental applications. The integration of this feature enhancement module thus contributes to both the interpretability and performance of the proposed HIT model by enabling it to learn not only what to attend to, but also how strongly to attend to each feature dimension.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0002

2.5 Correspondence Estimation

Before computing the transformation {R,t}, accurate correspondences between the source and target point clouds must be established. The previously obtained feature embeddings of the source and target are averaged along the second dimension, and their absolute difference is computed to yield the feature residual[11]. This residual is passed through a linear layer, layer normalization, and a LeakyReLU activation function to obtain a similarity measure between the source and target point clouds. This similarity controls the contribution of each point pair to the model: if the embeddings are similar, the similarity measure tends to be small, indicating easier alignment[12]; if they differ greatly, the similarity measure increases, signaling the model to pay more attention to such pairs during training[13]. The matching scores are obtained via the dot product between the source and target embeddings, and are modulated by the similarity measure to control the sharpness of the matching weights: higher similarity values produce smoother weights, while lower ones result in sharper matching. Softmax is then applied to yield normalized matching scores between the source and target point clouds. The maximum values and their indices are selected from the score matrix, representing the most likely corresponding point in the target point cloud for each point in the source.

2.6 Estimation of the Transformation Matrix

First, the weight matrix is expanded to ensure that each point has an assigned value in 3D space, and normalization is applied to balance the influence of different point pairs in subsequent computations. Then, the weighted centroids of the source and target point clouds are calculated to eliminate errors caused by translation. Based on these centroids, a covariance matrix is constructed to describe the spatial correlation between the two point clouds, capturing the transformation relationships between points. By performing singular value decomposition (SVD) on the covariance matrix, the left singular vectors, singular values, and right singular vectors are obtained[14]. SVD simplifies the complex transformation between point clouds into a series of rotations and scalings, thus identifying the optimal linear transformation to align the two point clouds.

2.7 Loss Function Design

The overall loss function for oral point cloud registration includes: Contour consistency loss (L_{ct}), Cycle consistency loss (L_c), Discrimination loss (L_d) and Dental arch symmetry loss (L_{sy})[15]. Smoothing coefficients are introduced to adjust the contribution of each loss component. The total loss is defined as:

$$L = L_{ct} + \alpha L_c + \beta L_d + \theta L_{sy}$$

1) Contour consistency loss L_{ct}: Minimizes the distance between the transformed source point cloud and the target point cloud, promoting end-to-end learning. The model simultaneously learns transformation parameters and registration alignment, while averaging local errors to improve robustness against small noise and sampling variation. The formula is:

$$L_{ct} = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{N}\sum_{j=1}^{N}\left\|src_{b,j} - tgt_{b,j}\right\|_2\right)$$

Where B is the batch size, D is the dimension of each point (usually 3), N is the number of randomly sampled points, and $src_{b,j}$ is the coordinate of the *j-th* point in the *b-th* sample.

2) Cycle consistency loss L_c: Measures the consistency between the predicted motion from $X \rightarrow Y$ (R_{XY}, t_{XY}) and from $Y \rightarrow X$ (R_{YX}, t_{YX}). The goal is to ensure that:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0002

$$L_c = \left\| R_{XY} R_{YX} - I \right\|^2 + \left\| t_{XY} + t_{YX} \right\|^2$$

3) Discrimination loss L_d: Evaluates the discriminative power of extracted features and the correctness of established correspondences:

$$L_d = -\frac{1}{\|M\|} \sum_{(x_i, y_j) \in M} \times \begin{bmatrix} C(x_i, y_j) \times lnS(x_i, y_j) + \\ (1 - C(x_i, y_j) \times ln(1 - S(x_i, y_j)) \end{bmatrix}$$

Here, if the correspondence $\{x_i, y_j\}$ is correct, then $C(x_i, y_j) = 1$; otherwise, $C(x_i, y_j) = 0$。 $S(x_i, y_j)$ denotes the similarity between feature vectors ϕ_{xi} and ϕ_{yj}, and M is the set of index pairs corresponding to matched points.

4) Dental arch symmetry loss L_{sy}: Provides a prior constraint that reduces the search space during optimization. This allows the model to avoid searching within asymmetric solution spaces, accelerating convergence and reducing training time. It also helps preserve the symmetrical structure of the dental arch:

$$L_{sy} = v \cdot \gamma \sum_{j=1}^{[N/2]-1} \frac{1}{B} \sum_{b=1}^{B} \left(\frac{1}{D} \sum_{d=1}^{D} \left| src_{b,d,j} - src_{b,d,N-j-1} \right| \right)$$

Where $src_{b,d,j}$ is the coordinate of the j-th point in the d-th dimension of the b-th sample, and γ is a smoothness weight controlling the contribution of this term to the total loss. γ acts as a discount factor that gradually decreases with iteration index i.

3. Experimental Results Analysis

3.1 Experimental Setup

The dataset employed in this study comprises single-frame 3D point clouds acquired from a range of intraoral scanning sources, including yellow plaster dental models, resin-based models, implant-supported structures, plastic dental casts, and real human oral cavities. This diversity ensures that the dataset captures a wide spectrum of clinical scenarios, ranging from idealized physical models to complex anatomical variations present in vivo. All point clouds were obtained using structured light and laser-based intraoral scanning systems, and were subsequently preprocessed to remove outliers and normalize scale.

For each point cloud instance, we computed three key geometric descriptors: the total number of points, average point density (inversely proportional to the square of neighborhood radius), and average mean curvature. These statistics provide a quantitative basis for understanding the structural variability within the dataset and for evaluating model robustness across different scanning conditions. The distributions of these descriptors are visualized in Figure 5.

Subfigure (a) illustrates the distribution of point counts across samples, which range approximately from 3,000 to 11,000 points. Most samples cluster around 7,000–9,000 points, indicating a moderately dense sampling typical of intraoral scanners. Subfigure (b) shows the average density per sample, with the majority concentrated between 150 and 190 $1/\text{radius}^2$, reflecting localized sampling regularity. Subfigure (c) presents the mean curvature values, which capture the degree of local surface variation and anatomical complexity. The curvature distribution peaks around 0.35, suggesting that most regions are moderately curved-consistent with the smooth but non-planar nature of dental surfaces.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0002

The complete dataset was partitioned using an 80/10/10 split for training, validation, and testing, respectively. Stratified sampling was applied to ensure balanced representation of model types and anatomical variability across all subsets. The training set was used to optimize the HIT model parameters, the validation set guided early stopping and hyperparameter tuning, and the test set was reserved for final performance evaluation.

To ensure reproducibility and computational consistency, all experiments were conducted on a high-performance computing environment equipped with an Intel Core i9-14900 CPU and an NVIDIA RTX 3090 GPU. This setup provided sufficient memory and parallel processing capacity to handle large-scale point cloud data and the intensive computations associated with attention-based deep learning models.

This experimental protocol ensures that the evaluation of HIT is conducted under realistic, clinically relevant conditions, with sufficient diversity in both synthetic and real intraoral scenarios to validate generalizability. The inclusion of real patient data further strengthens the ecological validity of the proposed approach.

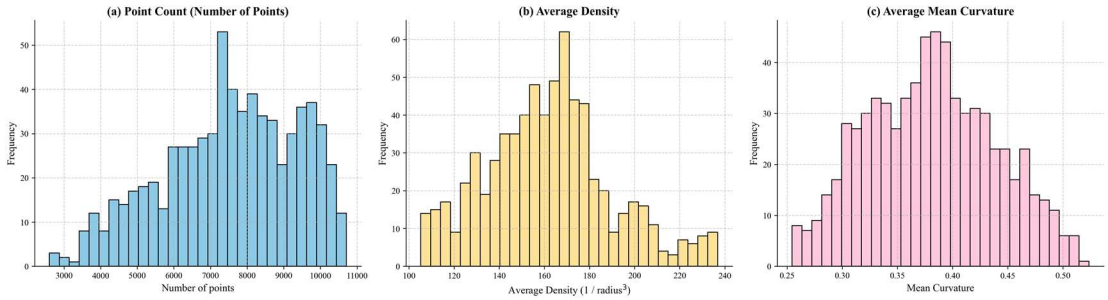

Figure 5. Distribution of Sample Statistics in the Oral Point Cloud Dataset

3.2 Evaluation Metrics

In this study, we adopt a set of quantitative evaluation metrics commonly used in point cloud registration tasks, particularly those aligned with benchmarks such as ModelNet40. These include Rotation Mean Absolute Error (RMAE), Rotation Root Mean Square Error (RRMSE), Translation Mean Absolute Error (TMAE), and Translation Root Mean Square Error (TRMSE). Collectively, these four metrics offer a comprehensive assessment of both the rotational and translational alignment quality between the predicted transformation and the ground truth.

For rotation-related errors, we first convert both the predicted rotation matrix and the ground truth rotation matrix into Euler angles following the z–y–x convention. This decomposition yields interpretable angle differences along three principal axes. The RMAE is then computed as the average of the absolute differences in Euler angles across all test samples, reflecting the typical misalignment in angular space. In contrast, RRMSE penalizes larger deviations more severely by squaring the angular errors before averaging and taking the square root, thereby offering a more sensitive measure for outlier rotations or poorly aligned instances.

For translation, errors are directly calculated in Euclidean space by comparing the predicted translation vector t_{pred} and the ground truth translation vector t_{gt}. Specifically, TMAE measures the average magnitude of the element-wise absolute differences, providing an intuitive sense of the expected translational offset. TRMSE, similar to RRMSE, emphasizes large deviations by using squared differences, and is particularly useful for capturing registration instability in noisy or complex scenes.

By evaluating both mean absolute and root mean square forms, we capture two complementary aspects of performance: typical-case accuracy (via RMAE and TMAE) and sensitivity to failure or inconsistency (via RRMSE and TRMSE). This dual perspective is especially important in the context of oral point clouds, where accurate registration depends not only on minimizing average error but also on avoiding localized alignment failures that may compromise clinical usability. The use of these

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0002

four metrics ensures a robust and interpretable framework for comparing different models and validating the performance of the proposed HIT approach across both synthetic and real-world datasets.

3.3 Comparative Experiments

The source point clouds were uniformly downsampled, and a fixed number of points were randomly sampled[16]. Transformation matrices were randomly generated from the following ranges: Rotation: [0°, 45°] along the x, y, and z axes; Translation: [-0.5, 0.5] along each axis. These transformations were applied to the downsampled source point clouds.

We compare our model with both traditional and learning-based methods, including: Traditional method: Generalized-ICP (implemented via the Open3D library); Learning-based methods: UGMMREG, IDAM, RPM-Net, and DIT[17]. For the learning-based baselines, we reproduced the results using open-source code released by the original papers and adjusted the settings and hyperparameters to suit our dataset[18]. We conducted tests on both dental model point clouds and real intraoral scans. Yellow plaster dental models were selected as the representative dental model type. Qualitative results are shown in Figure 6, and quantitative results are summarized in Table 1. Our model demonstrates a significant advantage in registration accuracy, validating the effectiveness of the hybrid interaction attention mechanism in recognizing and extracting geometric structures in oral point clouds. Moreover, the tailored loss function proves to be more suitable for oral point cloud registration tasks. Compared with traditional methods, HIT achieves substantial improvements in accuracy. Relative to other learning-based approaches, HIT exhibits faster training and better convergence behavior[19].

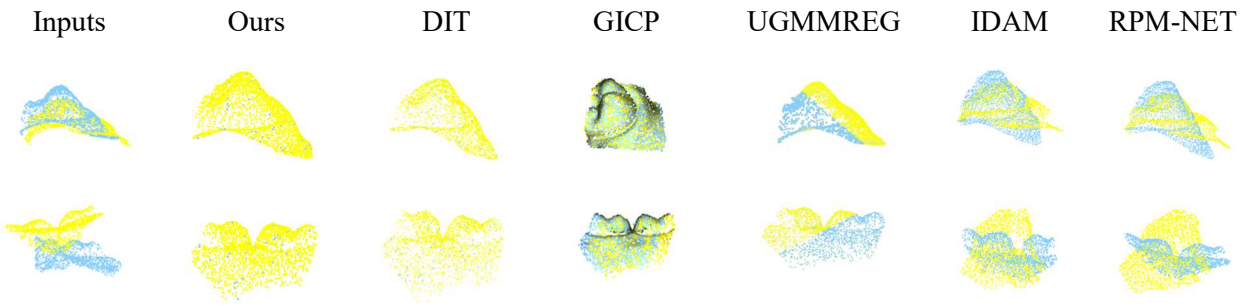

Figure 6. Comparison of Model Performance

Table 1. Comparison of Registration Errors Across Models

models	Scenario 1 (Yellow Plaster Model)				Scenario 2 (Real Oral Cavity)			
	R_{RMSE}	R_{MAE}	T_{RMSE}	T_{MAE}	R_{RMSE}	R_{MAE}	T_{RMSE}	T_{MAE}
GICP	0.1384	0.0096	0.0036	0.0025	0.0776	0.0674	0.0511	0.0476
UGMMREG	0.2964	0.2601	0.0003	0.0002	0.3508	0.3069	0.0003	0.0003
IDAM	5.285	2.318	0.0653	0.0254	27.47	11.44	0.1932	0.1157
RPM	5.721	2.5668	0.0766	0.0298	26.87	10.92	0.1865	0.1096
DIT	0.0081	0.0071	1.3e-4	1.1e-4	0.0086	0.0075	1.4e-4	1.2e-4
Ours	0.0006	0.0005	4.1e-5	3.5e-5	0.0007	0.0006	4.3e-5	3.5e-5

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0002

Table 2. Ablation Study Results

models	Scenario 1 (Yellow Plaster Model)				Scenario 2 (Real Oral Cavity)			
	R_{RMSE}	R_{MAE}	T_{RMSE}	T_{MAE}	R_{RMSE}	R_{MAE}	T_{RMSE}	T_{MAE}
Baseline (DIT)	0.0081	0.0071	1.3e-4	1.1e-4	0.0086	0.0075	1.4e-4	1.2e-4
New Loss Only	0.0007	0.0006	4.8e-5	4e-5	0.001	0.0007	5e-5	4.1e-5
Module Replacement	0.0009	0.0007	4.6e-5	4e-5	0.0012	0.0008	5.2e-5	4.1e-5
Full Model (HIT)	0.0006	0.0005	4.1e-5	3.5e-5	0.0007	0.0006	4.3e-5	3.5e-5

3.4 Ablation Study

To evaluate the individual contributions of the oral-specific loss design and the hybrid interaction attention mechanism, we conducted a set of ablation experiments across two representative scenarios[20]: the registration of yellow plaster dental models and real intraoral single-frame point clouds. The experimental settings are organized around four configurations: (1) the baseline model, i.e., the original DIT with standard loss; (2) a variant using the new loss function only; (3) a variant using the proposed attention module only; and (4) the full HIT model, which integrates both improvements. Quantitative results are summarized in Table 2.

Compared to the baseline, the introduction of the tailored loss function alone leads to substantial improvements across all metrics. On the plaster model dataset, the rotation root mean square error (RRMSE) drops from 0.0081 to 0.0007, and the translation root mean square error (TRMSE) is reduced by more than 60%, from 1.3e-4 to 4.8e-5. A similar trend is observed on real intraoral data, where RRMSE falls from 0.0086 to 0.001, and TMAE from 1.2e-4 to 4.1e-5. These results demonstrate that the loss function-designed to capture domain-specific structural priors such as dental arch symmetry and contour consistency-plays a key role in guiding the optimization process toward more anatomically plausible transformations, even when the feature extraction module remains unchanged.

The effect of replacing the baseline attention module with the hybrid interaction attention mechanism is also notable. Although the improvements are slightly less pronounced than those achieved by the loss design, they remain consistent across metrics and datasets. On the plaster model data, the RRMSE decreases from 0.0081 to 0.0009, and on the real intraoral data, from 0.0086 to 0.0012. The hybrid attention mechanism thus improves the model's ability to extract robust and discriminative spatial features, particularly in low-texture or partially occluded regions where traditional point-wise attention may struggle.

The full HIT model, which incorporates both the oral-specific loss and the hybrid attention module, achieves the best performance across all metrics and both datasets. For plaster models, RRMSE and TRMSE are reduced to 0.0006 and 4.1e-5, respectively, while in the real intraoral scenario, these values are 0.0007 and 4.3e-5. The consistent superiority of the full model suggests a strong complementarity between the two components: the loss function provides precise supervision aligned with oral anatomy, while the hybrid attention mechanism enhances the quality of geometric feature extraction. Together, they enable more accurate estimation of rigid transformations, particularly under challenging conditions such as sparse sampling, soft-tissue deformation, or scanning noise.

Taken together, the ablation study confirms that both the attention structure and the loss design are necessary to unlock the full performance of the proposed HIT framework. Their respective contributions are evident in both synthetic and real-world settings, supporting the model’s robustness and generalization ability. These findings also reinforce the value of designing domain-specific architectures in medical point cloud registration tasks, where generic solutions often fall short of the accuracy required for clinical applicati.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0002

4. Conclusion

In this study, we introduce a novel point cloud registration framework-Hybrid Interaction Transformer (HIT)-specifically tailored for the challenges inherent in oral cavity data. The proposed model leverages a hybrid interaction attention mechanism that combines point convolution with relative attention to achieve more effective geometric feature modeling and global spatial interaction. Furthermore, we design a multi-component loss function incorporating contour consistency, cycle consistency, feature discrimination, and dental arch symmetry, each targeting the anatomical and structural characteristics unique to oral point clouds.

Through extensive experiments conducted on both synthetic dental models and real intraoral scan data, the HIT model demonstrates superior performance in terms of registration accuracy and robustness. Compared to both traditional methods (e.g., GICP) and contemporary learning-based approaches (e.g., DIT, RPM-Net), our model achieves significantly lower rotation and translation errors, highlighting its capability to capture fine-grained spatial correspondences and maintain geometric fidelity during alignment. The ablation studies further validate the contribution of each architectural component, confirming the synergistic effect of the hybrid attention design and loss formulation.

Importantly, the proposed approach addresses several key limitations of prior methods: it enhances robustness to sparsity, improves generalization across diverse dental structures, and reduces the dependency on high-quality initialization. These characteristics make it particularly suitable for clinical contexts where point clouds may be noisy, incomplete, or vary in density due to scanning conditions. Moreover, the model's ability to generalize across different materials (e.g., plaster, resin, soft tissue) and intraoral structures suggests strong adaptability for heterogeneous datasets in real-world dental applications.

Beyond its core registration function, HIT also provides a foundational framework upon which other downstream tasks-such as 3D reconstruction, dynamic occlusion modeling, and temporal tracking-can be built. Given its architecture's modularity, HIT could be naturally extended to multi-frame alignment, offering potential applications in continuous intraoral scanning and time-resolved dental monitoring. Additionally, its attention-based formulation opens up promising opportunities for future cross-modal fusion tasks, such as integrating RGB, depth, or radiographic data with point cloud geometry to enrich feature representations.

Nevertheless, the current model exhibits certain limitations when aligning point clouds with substantial density imbalance-such as dense targets versus sparse sources-a scenario frequently encountered in practical intraoral scanning. Under such conditions, the establishment of accurate and complete correspondences becomes more challenging, and the model's performance may degrade without sufficient data or extended training. Future improvements will require architectural refinements that promote better information propagation across variable-density regions and mechanisms that explicitly handle confidence estimation under data uncertainty.

Looking ahead, future research will focus on several directions. First, we aim to extend the HIT architecture to better accommodate unbalanced point densities through adaptive sampling strategies, multi-resolution encoding, and contrastive feature alignment mechanisms. Second, we plan to evaluate the clinical applicability of the proposed framework in real-world dental workflows, such as preoperative planning, prosthetic modeling, digital occlusion analysis, and orthodontic treatment monitoring, by integrating it into existing CAD/CAM systems and evaluating its usability in collaboration with dental professionals. Third, there is potential to scale the model to large-scale oral databases and benchmark it on standardized clinical datasets to further verify its generalizability and real-time performance.

Taken together, this work demonstrates that combining domain-specific geometric constraints with hybrid attention mechanisms can significantly advance the accuracy, robustness, and clinical relevance of point cloud registration in digital oral healthcare. As oral medicine increasingly shifts

ISSN: 2414-1895 DOI: 10.6919/ICJE.202604_12(4).0002

toward precision and automation, HIT represents a practical and theoretically grounded step forward in bridging algorithmic innovation with real-world dental applications.

References

[1] Dhanda M, Kukreja A, Pande S S. Region-based efficient computer numerical control machining using point cloud data[J]. Journal of Computing and Information Science in Engineering, 2021, 21(4): 041005.

[2] P. J. Besl and N. D. McKay, “Method for registration of 3-D shapes,” Proc. SPIE, vol. 1611, pp. 586–606, Apr. 1992.

[3] J. Yang, H. Li, D. Campbell, and Y. Jia, “Go-ICP: A globally optimal solution to 3D ICP point-set registration,”IEEE Trans. Pattern Anal.Mach.Intell.,vol.38, no.11, pp. 2241–2254, Nov. 2016.

[4] Y. Aoki, H. Goforth, R. A. Srivatsan, and S. Lucey, “PointNetLK: Robust & efficient point cloud registration using PointNet,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 7163–7172.

[5] X. Li, J. K. Pontes, and S. Lucey, “PointNetLK revisited,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 12763–12772.

[6] Z. J. Yew and G. H. Lee, “RPM-Net: Robust point matching using learned features,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog nit. (CVPR), Jun. 2020, pp. 11824–11833.

[7] Z. J. Yew and G. H. Lee, “REGTR: End-to-end point cloud corre spondences with transformers,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 6677–6686.

[8] K. Fu, S. Liu, X. Luo, and M. Wang, “Robust point cloud registration framework based on deep graph matching,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 8893–8902.

[9] G. Chen, M. Wang, Q. Zhang, L. Yuan and Y. Yue , “Full Transformer Framework for Robust Point Cloud Registration With Deep Information Interaction ” IEEE Trans Neural Netw Learn Syst 2024 Vol. 35 Issue 10 Pages 13368-13382, Oct. 2024

[10] Ligas M, Prochniewicz D. Procrustes based closed-form solution to the point-wise weighted rigid-body transformation in asymmetric and symmetric cases[J]. Journal of Spatial Science, 2021, 66(3): 445-457.

[11] A. Vaswani et al,“Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.

[12] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, “Point transformer,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 16259–16268.

[13] Jin X, Xie Y, Wei X S, et al. Delving deep into spatial pooling for squeeze-and-excitation networks[J]. Pattern Recognition, 2022, 121: 108159.

[14] Deerwester S, Dumais S T, Fraser Marglin D, et al.Indexing by Latent Semantic Analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

[15] **Feng, Y., Xu, F., & Xu, R**, "3D Shape Recognition via Deep Learning". *IEEE Transactions on Image Processing*, 2009, 29, 5870-5880.

[16] **Peng, H., & Zhang, X,** "An Efficient Algorithm for 3D Surface Reconstruction with Smoothing Constraints". *Computer Graphics Forum*, 2009,35(7), 191-201.

[17] **Aigerman, N., & Efrat, A,**"Generalized-ICP: A Novel Framework for Point Set Registration". "Generalized-ICP". Robotics: Science and Systems (RSS), 2009.

[18] **Dong, Q., & Wang, Z,**"An Unsupervised Learning-based Gaussian Mixture Model for Point Cloud Registration". *Computer Vision and Image Understanding*,2022, 192, 102897.

[19] J. Li, C. Zhang, Z. Xu, H. Zhou, and C. Zhang, “Iterative distance-aware similarity matrix convolution with mutual-supervised point elimination for efficient point cloud registration,” in Proc. 16th Eur. Conf. Comput. Vis. Glasgow, U.K.: Springer, Aug. 2020, pp. 378–394.

[20] **Zhou, Y., Wang, Y., Hu, Q., & Huang, J,**"Ablation Studies in Machine Learning: The Importance of Feature Selection". *Journal of Machine Learning Research*, 2006,21(154), 1-26.