Learning-based CBCT–IOS Registration with PointNet++ and SVD

Changkai Ji[1 [0009−0007−7090−7360]], Yusheng Liu[1 [0009−0004−2624−9223]], Yuxian Jiang[1 [0009−0002−7689−5333] [0009−0009−3223−0082]], and Lisheng Wang[1 [0000−0003−3234−7511]]

School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China
{changkaiji, lswang}@sjtu.edu.cn

Abstract. Accurate registration of intraoral scans (IOS) and cone-beam computed tomography (CBCT) is a critical prerequisite for precise diagnosis and treatment planning in dentistry. However, large modality discrepancies and dense point clouds make this task challenging in practice. In this work, we propose a learning-based framework for CBCT–IOS registration, developed in the context of the MICCAI STSR Task 2 2025 Challenge. Our method leverages dual PointNet++ encoders to extract modality-specific features, followed by a differentiable SVD head that execute rigid-body constraints in the predicted transformation. To enhance robustness, we design geometric data augmentation strategies, while point cloud sampling and simplification are employed to accelerate inference. Ablation studies demonstrate that augmentation substantially reduces registration errors, while relaxing CBCT filtering thresholds further improves alignment by preserving richer anatomical cues. Overall, our approach achieves competitive performance, ranking second on the validation leaderboard, and provides a practical balance between accuracy and efficiency.

Keywords: CBCT-IOS registration · PointNet++ · Rigid transformation · Data augmentation · Inference acceleration

1 Introduction

Three-dimensional registration of dental data plays a crucial role in computer-aided diagnosis, treatment planning, and surgical guidance [3,9]. In clinical practice, intraoral scans (IOS) provide high-resolution crown geometry, while cone-beam computed tomography (CBCT) offers comprehensive information on both crowns and roots [14]. Accurate alignment of these heterogeneous modalities is essential for integrating complementary anatomical details, thereby enhancing the precision and reliability of dental treatment [12]. To promote the development of robust registration algorithms, the MICCAI STSR 2025 Challenge Task 2 was organized to benchmark algorithms that can effectively handle multi-modal data discrepancies and to encourage practical solutions that may translate into real-world clinical applications.

Despite its importance, CBCT–IOS registration remains a challenging task. The two modalities differ substantially in terms of resolution, field of view, and information content [8]. IOS captures only the visible crowns with fine detail but lacks root structures, whereas CBCT provides full jaw coverage but contains significant noise and redundant information. These discrepancies introduce difficulties in establishing reliable correspondences and estimating robust transformations. Moreover, limited availability of paired ground-truth annotations further complicates the training of data-driven approaches.

Recent advances in deep learning have achieved remarkable success across imaging tasks [11,2,7,10]. Researchers have increasingly applied deep learning methods to multi-modal 3D registration problems [8]. Such methods alleviate the need for handcrafted descriptors and have achieved promising results in various medical imaging domains. However, deep learning-based approaches often require large annotated datasets [15,4], and their inference pipelines may still suffer from inefficiency due to the high dimensionality of volumetric data and dense point clouds. Therefore, it remains an open question how to design a framework that is both accurate and computationally efficient [6,5].

In this work, we propose a learning-based registration framework specifically designed for CBCT–IOS alignment in the MICCAI STSR 2025 Challenge. Our method employs PointNet++ encoders to extract modality-specific features from IOS and CBCT point clouds [13], followed by a transformation head based on singular value decomposition (SVD) that enforces rigid-body constraints in the predicted matrix [18]. To enhance robustness, we incorporate extensive data augmentation during training, enabling the model to generalize well across diverse clinical cases. Additionally, we use point cloud sampling and simplification to accelerate inference, reducing computational overhead and enabling fast inference without compromising accuracy. As a result, we achieve competitive performance on the validation leaderboard. Our contributions can be summarized as follows:

- We design data augmentation strategies to improve model robustness and registration accuracy under diverse clinical conditions.
- We adopt point cloud sampling and simplification techniques to accelerate inference while maintaining accuracy.
- Our method achieves second place on the validation leaderboard of the STSR 2025 Task 2, demonstrating both effectiveness and efficiency.

2 Method

2.1 Framework Overview

Figure 1 illustrates the overall architecture of our proposed framework for CBCT-IOS registration. The framework follows a learning-based paradigm that takes as input two point clouds: one sampled from the IOS mesh and the other from the CBCT volume. Both point clouds are independently encoded by two PointNet++ encoders, which are responsible for extracting hierarchical geometric features. The extracted features are subsequently aligned through a feature matching module, followed by a SVD head that estimates the rigid transformation

matrix between the two modalities. This transformation is then used to map the IOS points into the CBCT coordinate system. During training, multiple loss terms are employed to jointly supervise the transformation prediction, including point-based losses, Chamfer distance, and penalties on rotation and translation. This design ensures that the model captures both global and local geometric correspondences in a computationally efficient manner.

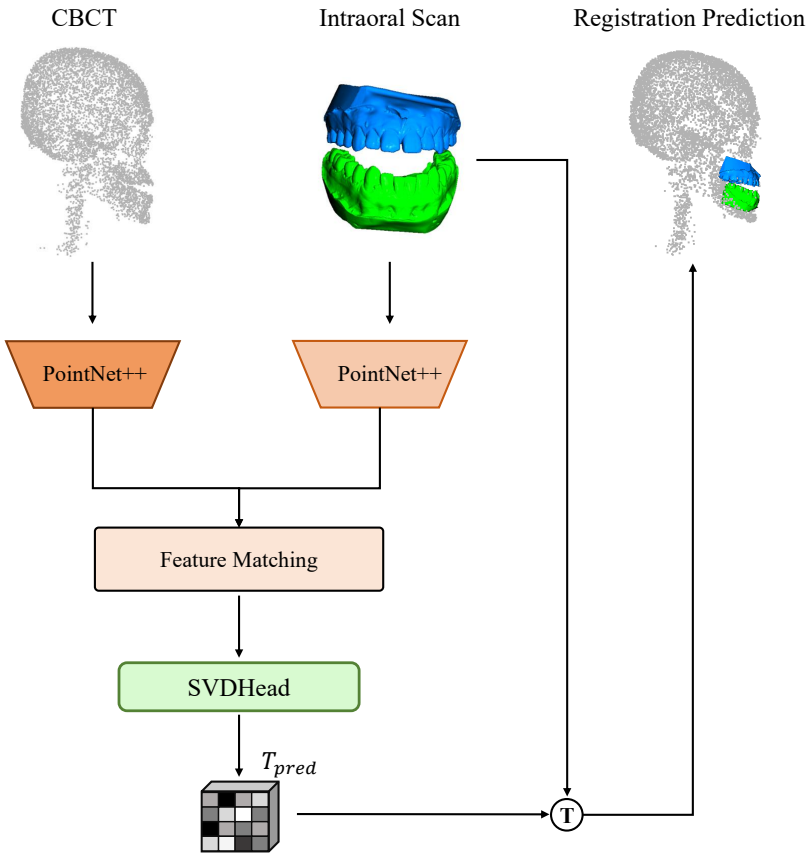

Fig. 1. Overview of our registration model. CBCT scans and intraoral scans are separately processed by PointNet++, followed by feature matching and rigid transformation estimation using SVD. The predicted transformation is applied to align the intraoral scans with CBCT data.

2.2 Data Augmentation

To enhance the robustness and generalization of the proposed model, we employed a series of data augmentation strategies tailored to 3D point clouds. Specifically, random rigid transformations, including rotations and translations, were applied independently to both the IOS-derived point sets. These augmentation techniques enrich the diversity of the training dataset and mitigate the risk of overfitting, particularly in scenarios where annotated data is limited.

To better illustrate the effectiveness of our augmentation strategies, Figure 2 provides a visual example. The first column presents the original CBCT and IOS pairs prior to augmentation, while the subsequent three columns demonstrate augmented versions of the same case. These examples highlight how the applied transformations produce diverse yet clinically plausible variations, enabling the model to learn invariances that are essential for accurate and robust registration.

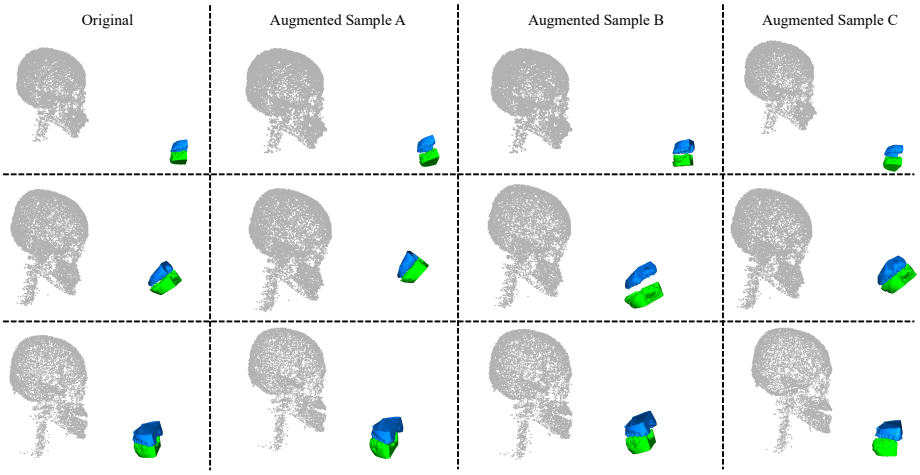

Fig. 2. The visualization of data augmentation strategies. The first column shows the original CBCT and IOS pairs, while the subsequent columns display augmented versions of the same case, demonstrating the diverse and clinically plausible variations produced by the applied transformations.

2.3 Model Training

The training of our framework follows a fully supervised paradigm, where the objective is to learn accurate rigid transformations between CBCT and IOS point clouds. A central component of the architecture is the feature extraction stage, implemented via PointNet++ encoders. PointNet++ extends the original PointNet architecture by introducing hierarchical feature learning, where local

neighborhood information is progressively aggregated at multiple scales. This design allows the network to capture both fine-grained geometric details and global structural context, which is essential for modeling complex dental anatomy. In our framework, two independent PointNet++ encoders are employed, one for the CBCT point cloud and the other for the IOS point cloud. These dual encoders extract modality-specific features while preserving their geometric consistency.

The extracted feature representations are then passed to the transformation estimation module, referred to as the SVDHead. This module aligns the latent embeddings of the two modalities by constructing a correspondence matrix and applying a differentiable SVD. The SVDHead directly estimates the optimal rigid transformation matrix, decomposed into a rotation matrix and a translation vector, which maps the IOS point cloud onto the CBCT reference. Compared with regression-based alternatives, the SVD-based formulation offers improved stability and guarantees the orthogonality of the predicted rotation matrix.

Optimization is performed using the Adam optimizer. During training, the network parameters are updated to minimize a composite loss function that jointly enforces geometric alignment and transformation accuracy, which will be detailed in the following subsection. This training strategy tries the network to converge reliably and generalize well to unseen test data.

2.4 Loss Function

To achieve robust and accurate registration, we adopt a composite loss function that integrates multiple complementary objectives. Each component of the loss is designed to address a specific aspect of the alignment problem, ensuring both local geometric consistency and global rigid transformation accuracy.

Point Loss. This term enforces point-wise consistency between the transformed source point cloud $\hat{\mathbf{P}}_{src}$ and the ground truth aligned point cloud $\mathbf{P}_{gt}$. It is formulated as a mean squared error (MSE), directly penalizing local misalignments:

$$\mathcal{L}_{point} = \frac{1}{N} \sum_{i=1}^{N} \left\| \hat{\mathbf{p}}_{src}^{(i)} - \mathbf{p}_{gt}^{(i)} \right\|^2 .$$

Chamfer Distance. To capture global shape similarity, we compute the bidirectional Chamfer distance between the predicted source $\hat{\mathbf{P}}_{src}$ and the target CBCT $\mathbf{P}_{tgt}$:

$$\mathcal{L}_{chamfer} = \sum_{p \in \hat{\mathbf{P}}_{src}} \min_{q \in \mathbf{P}_{tgt}} \|p - q\|^2 + \sum_{q \in \mathbf{P}_{tgt}} \min_{p \in \hat{\mathbf{P}}_{src}} \|q - p\|^2.$$

This term encourages the transformed point sets to occupy the same geometric space.

Rotation Loss. We explicitly constrain the predicted rotation $\hat{R}$ to be consistent with the ground truth R_{gt}. This is measured by the geodesic distance on $SO(3)$:

$$\mathcal{L}_{rot} = \arccos\left(\frac{\mathrm{Tr}(\hat{R} R_{gt}^{\top}) - 1}{2} \right).$$

Translation Loss. As translation misalignment often dominates registration error in clinical practice, we emphasize translation accuracy by computing the Euclidean distance between the predicted $\hat{t}$ and ground truth t_{gt} translation vectors:

$$\mathcal{L}_{trans} = \left\|\hat{t} - t_{gt}\right\|_2^2.$$

Matrix Regularization. To ensure the predicted transformation matrix remains a valid rigid body transformation, we introduce a regularization term that penalizes deviations from orthogonality and unit determinant:

$$\mathcal{L}_{mat} = \left\|\hat{R}^\top \hat{R} - I\right\|_F^2.$$

Overall Loss. The total loss integrates all components in a weighted sum:

$$\mathcal{L} = \lambda_p \cdot \mathcal{L}_{point} + \lambda_c \cdot \mathcal{L}_{chamfer} + \lambda_r \cdot \mathcal{L}_{rot} + \lambda_t \cdot \mathcal{L}_{trans} + \lambda_m \cdot \mathcal{L}_{mat},$$

Where, the weighting coefficients $\lambda_p = 0.5$, $\lambda_c = 1.0$, $\lambda_r = 1.0$, $\lambda_t = 3.0$, and $\lambda_m = 0.3$ are employed. The relatively higher weight assigned to the translation loss reflects its critical importance for achieving clinically meaningful registration accuracy.

2.5 Inference Acceleration

To ensure computational efficiency and enable practical deployment, we implemented a point cloud sampling and simplification strategy. During inference, the original CBCT scans often produce dense point sets, which substantially increase computational cost without proportionally improving accuracy. To address this, we uniformly subsampled the CBCT point clouds to a fixed number of points, while IOS meshes were converted to point clouds with a comparable resolution. This design ensures balanced complexity between modalities, reduces GPU memory consumption, and accelerates inference speed.

Importantly, this balance between efficiency and precision makes the framework more applicable in real-world clinical scenarios, where both accuracy and time efficiency are crucial.

3 Experiments and Results

3.1 Dataset and Assessment Metrics

The dataset provided by the STSR 2025 challenge comprises paired CBCT volumes and IOS meshes [17,16]. In the training phase, two subsets are available: a labeled set, where each CBCT-IOS pair is annotated with an affine transformation matrix aligning the upper and lower dentition, and an unlabeled set containing paired CBCT volumes and IOS meshes. In addition, a validation set is released without annotations, serving as the benchmark for leaderboard evaluation.

For quantitative evaluation, two complementary metrics are used: the mean translation error, which measures the Euclidean distance between predicted and ground-truth translation vectors, and the mean rotation error, computed as the geodesic distance between the predicted and reference rotation matrices. These metrics directly reflect the fidelity of the registration outcome, with lower values indicating higher accuracy. Although computational efficiency, such as inference time and GPU memory usage, is not explicitly scored in the validation phase due to the limitations of the challenge platform, it remains a practical consideration when deploying the methods in real clinical workflows.

3.2 Implementation details

Environments and Requirements. All experiments were conducted on a workstation, and the details of the hardware and software configuration are summarized in Table 1. The model was trained using the PyTorch framework for a total of 200 epochs.

Table 1. System Configuration

Ubuntu version	Ubuntu 24.04 LTS
CPU	Intel(R) Xeon(R) Platinum 8352S CPU @ 2.20GHz
RAM	503 GB
GPU	1 NVIDIA GeForce RTX 4090 (24G)
CUDA version	12.4
Programming language	Python 3.9.19
Deep learning framework	PyTorch (torch 1.12.1, torchvision 0.19.1)
Codes available at	https://github.com/duola-wa/MICCAI-2025-STSR-Task-2

3.3 Results and Analysis

To evaluate the effectiveness of our method, we present a series of ablation studies focusing on different design choices. As shown in Table 2, applying data augmentation substantially improves registration accuracy. Both translation and rotation errors are reduced, highlighting the importance of introducing geometric variability during training. By exposing the model to diverse transformations, augmentation enhances robustness to unseen cases and prevents overfitting, leading to a more generalizable registration framework.

Table 2. Effect of data augmentation on registration accuracy.

Setting	**Mean Translation Error (mm)**	**Mean Rotation Error (°)**
w/o Augmentation	230.80	37.54
w/ Augmentation	165.57	24.00

As shown in Table 3, incorporating Iterative Closest Point (ICP) refinement reduces the mean translation error relative to the baseline prediction [1]. However, given the limited overall gain and additional computational cost, ICP was not included in our final pipeline.

Table 3. Effect of ICP refinement on registration accuracy.

Method	Mean Translation Error (mm)	Mean Rotation Error (°)
w/o ICP	165.57	24.00
w/ ICP	157.68	43.56

Table 4 further compares the performance under different CBCT filtering thresholds. The threshold refers to the intensity cutoff applied to CBCT voxels when extracting point clouds. A higher threshold retains densest regions such as enamel and cortical bone, while a lower threshold preserves a larger portion of anatomical structures, including lower-density bone. Relaxing the criterion from 800 to 600 therefore increases the number of target points available for alignment, which leads to a modest improvement in both translation and rotation accuracy. This suggests that incorporating a richer set of structural cues benefits the registration process.

Table 4. Effect of CBCT filtering threshold on registration accuracy (w/o ICP).

Filtering Condition	Mean Translation Error (mm)	Mean Rotation Error (°)
CBCT > 800	165.57	24.00
CBCT > 600	164.46	23.71

4 Conclusion

In this paper, we present a learning-based framework for CBCT–IOS registration, tailored to the MICCAI STSR Task 2 2025 Challenge. The framework integrates dual PointNet++ encoders with a differentiable SVD head to estimate rigid transformations under orthogonality constraints. By leveraging tailored data augmentation and efficient point cloud sampling, our approach seeks to balance accuracy and inference speed. Experimental results demonstrate the effectiveness of the proposed augmentation strategies. Ultimately, our method achieved second place on the validation leaderboard. These results highlight the potential of our framework for clinical applications that demand rapid and reliable responses. In future work, we plan to further explore semi-supervised strategies to better leverage unlabeled data and to investigate lightweight architectures that further reduce computational overhead for deployment in clinical settings.

References

1. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures. vol. 1611, pp. 586–606. Spie (1992)
2. Bolelli, F., Lumetti, L., Vinayahalingam, S., Di Bartolomeo, M., Pellacani, A., Marchesini, K., Van Nistelrooij, N., Van Lierop, P., Xi, T., Liu, Y., et al.: Segmenting the inferior alveolar canal in cbcts volumes: the toothfairy challenge. IEEE Transactions on Medical Imaging (2024)
3. Flügge, T., Derksen, W., Te Poel, J., Hassan, B., Nelson, K., Wismeijer, D.: Registration of cone beam computed tomography data and intraoral surface scans–a prerequisite for guided implant surgery with cad/cam drilling guides. Clinical Oral Implants Research **28**(9), 1113–1118 (2017)
4. Ji, C., Du, C., Zhang, Q., Wang, S., Ma, C., Xie, J., Zhou, Y., He, H., Shen, D.: Mammo-net: Integrating gaze supervision and interactive information in multi-view mammogram classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 68–78. Springer (2023)
5. Ji, C., Liu, Y., He, L., Jiang, Y., Huang, C., Wang, L.: Two-stage semi-supervised nnu-net framework for tooth segmentation in cbct images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 100–109. Springer (2024)
6. Ji, C., Liu, Y., He, L., Jiang, Y., Huang, C., Wang, L.: A two-stage semi-supervised nnu-net model for automated tooth segmentation in panoramic x-ray images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 91–99. Springer (2024)
7. Jiang, Y., Liu, Y., Ji, C., Wang, L.: Enhanced multi-structure segmentation in cbct images with adaptive structure optimization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 30–40. Springer (2024)
8. Kim, S., Choi, Y., Na, J., Song, I.S., Lee, Y.S., Hwang, B.Y., Lim, H.K., Baek, S.J.: Best of both modalities: Fusing cbct and intraoral scan data into a single tooth image. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 553–563. Springer (2024)
9. Lim, S.W., Hwang, H.S., Cho, I.S., Baek, S.H., Cho, J.H.: Registration accuracy between intraoral-scanned and cone-beam computed tomography–scanned crowns in various registration methods. American Journal of Orthodontics and Dentofacial Orthopedics **157**(3), 348–356 (2020)
10. Liu, Y., Xin, R., Yang, T., Wang, L.: Inferior alveolar nerve segmentation in cbct images using connectivity-based selective re-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–12. Springer (2024)
11. Liu, Y., Zhang, S., Wu, X., Yang, T., Pei, Y., Guo, H., Jiang, Y., Feng, Z., Xiao, W., Wang, Y.P., et al.: Individual graph representation learning for pediatric tooth segmentation from dental cbct. IEEE Transactions on Medical Imaging (2024)
12. Olczyk, A., Malicka, B., Skośkiewicz-Malinowska, K.: Retrospective study of the morphology of third maxillary molars among the population of lower silesia based on analysis of cone beam computed tomography. Plos one **19**(2), e0299123 (2024)
13. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)

14. Su, S., Liu, Y.m., Zhan, L.p., Gao, S.y., He, C., Zhang, Q., Huang, X.f.: Evaluation of the accuracy of cone-beam computed tomography image segmentation of isolated tooth roots based on the dynamic threshold method. BMC oral health **23**(1), 752 (2023)
15. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: using gaze to supervise computer-aided diagnosis. IEEE Transactions on Medical Imaging **41**(7), 1688–1698 (2022)
16. Wang, Y., Chen, X., Qian, D., Ye, F., Wang, S., Zhang, H.: Semi-supervised Tooth Segmentation: First MICCAI Challenge, SemiToothSeg 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8, 2023, Proceedings, vol. 14623. Springer Nature (2024)
17. Wang, Y., Zhang, Y., Chen, X., Wang, S., Qian, D., Ye, F., Xu, F., Zhang, H., Zhang, Q., Wu, C., et al.: Sts miccai 2023 challenge: grand challenge on 2d and 3d semi-supervised tooth segmentation. arXiv preprint arXiv:2407.13246 (2024)
18. Wang, Y., Solomon, J.M.: Deep closest point: Learning representations for point cloud registration. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3523–3532 (2019)