Dual-stage Course-to-fine Tooth Alignment Framework with Decoupled Position, Orientation and Shape Features

Xiaoxian Jin[1], Yichi Wu[1], and Hongkai Wang[1,2,3]

[1] School of Biomedical Engineering, Faculty of Medicine, Dalian University of Technology, Dalian 116024, China
[2] Central Hospital of Dalian University of Technology (Dalian Municipal Central Hospital), Dalian, China
[3] Dalian Key Laboratory of Intelligent Health and Digital Twin, Dalian, China
wang.hongkai@dlut.edu.cn

Abstract. Automated tooth alignment from oral scan data is critical for improving efficiency and objectiveness of digital orthodontics. Existing methods struggle with clinically complex cases like severe misalignments caused by the mixing of position, orientation and shape irregularities. We propose a new strategy which decouples the complex irregularities into simpler factors and addresses them sequentially via a dual-stage coarse-to-fine framework. The first stage addresses the positional irregularities using a cost-effective lightweight bidirectional RNN. Afterwards, the second stage employs a diffusion model whose regression core is a Vision Transformer (ViT) equipped with long-range skip connections, to progressively refine tooth orientation and centroid positions. To precisely capture the shape and orientation features which are essential for fine-scale adjustment, we construct the tooth statistical shape models and register them to the oral scan data. Experiments demonstrate improved accuracy and robustness across diverse scenarios while maintaining computational efficiency. Thanks to the decoupled adjustment strategy, our method successfully addresses the challenging clinical scenarios with severe misalignment and large tooth extraction gaps, which commonly fail the compared approaches. The fine-scale capturing of orientation and shape also yield a reduction of reconstruction error, achieving a Target Alignment Error (TRE) of 0.66 mm—approximately 20% lower than state-of-the-art methods—highlighting our framework's potential for precise orthodontic planning.

Keywords: Tooth alignment, Dual-stage framework, Decoupled features.

1 Introduction

The growing demand for efficient and precise orthodontic treatment has spurred the integration of artificial intelligence into digital dentistry. Among these advances, automated tooth alignment has emerged as a critical component. It not only supports clinicians in treatment planning and reduces manual effort but also helps patients visualize

potential outcomes, highlighting the importance of studying automated alignment techniques.

Early research on automatic tooth alignment primarily relied on 2D images, such as using StyleGAN for alignment prediction from facial portraits [1]. Recent advancements have shifted to 3D point clouds [2,3,4] and multi-modal data integration, combining 2D images, 3D meshes, point clouds, and anatomical landmarks [5,6,7]. While these methods improve accuracy, their strict data requirements limit practical applicability.

The automated alignment task can be defined as predicting the rigid transformation of each tooth, which involves translation and rotation. Existing methods generally follow three-step pipeline: (1) Feature extraction using PointNet/PointNet++ [8,9] for 3D point cloud data; (2) Feature propagation between teeth employing graph neural networks (GNNs) [4,10,11], capturing spatial relationships at multiple scales; (3) Transformation parameter regression based on MLP [2,4,6,10,11], where recent methods [7,12] leverage diffusion probabilistic models (DPMs) [13] to mirror the progressive nature of orthodontic tooth movement. To represent rigid-body transformations in tooth alignment, different parameterization methods are used, including 4×4 transformation matrices [3] and six degrees of freedom (6-DoF) transformation parameters [2,4,6,7].

Although existing methods have advanced tooth alignment automation but still struggle with severe malocclusions caused by positional and orientation deviations. The presence of missing teeth, especially when repositioning adjacent teeth after an extraction, poses a significant challenge in tooth alignment. Some studies [4] can only process non-missing teeth, limiting clinical applicability.

Inspired by these limitations and related research, we propose an automatic tooth alignment network for 3D point clouds. By capturing more comprehensive tooth features, including shape and orientation, we decouple translation and rotation and address them progressively in a coarse-to-fine network.

To sum up, we make the following contributions:

(1) We propose a two-stage automatic tooth alignment framework. The first stage uses an ultra-lightweight bidirectional recurrent neural network (Bi-RNN) [14] for initial centroid alignment, reducing the second stage's complexity, along with an improved centroid sequencing strategy.

(2) Our method integrates tooth shape and orientation using a statistical shape model to enhance orientation representation.

(3) The second stage employs a diffusion probabilistic model to simulate orthodontic alignment, predicting quaternion-based rotations and translations for improved accuracy and reliability.

2 Method

2.1 Overview

The overall framework architecture is illustrated in Fig. 1, which addresses centroid positions and orientation sequentially via a dual-stage coarse-to-fine framework.

In the first stage, a Bi-RNN predicts tooth centroid movements by modeling contextual relationships. The second stage refines alignment through three steps: (1) a feature extraction module combines axial information from statistical shape models (SSM) fitting, surface morphology from PointNet++, and local feature aggregation via transformers; (2) a diffusion-based regression module predicts 7D transformation parameters using global/local features and centroids as inputs; (3) the transformation module applies these parameters to the original point cloud, producing the final aligned output.

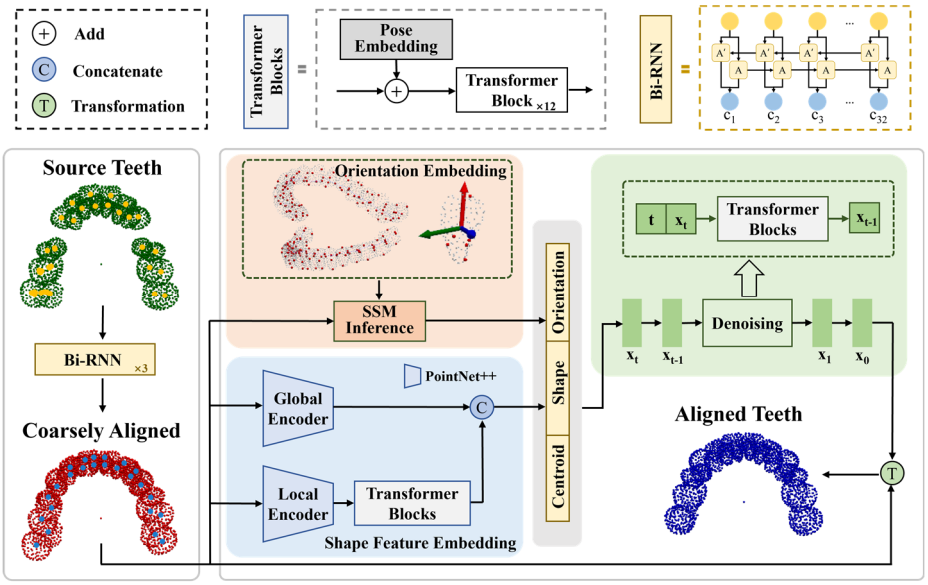

Fig. 1. The overall architecture of the proposed framework.

2.2 Network

Coarse Centroid Alignment. By incorporating a Bidirectional Recurrent Neural Network (Bi-RNN), we effectively capture contextual information in the tooth centroids sequence for movement prediction. The input tooth point cloud is structured as an ordered set of points $P = \{P_v \subseteq \mathbb{R}^{N_v \times 3}\}$, where v denotes the index of the tooth according to the Universal Number System, and N_v is the number of points sampled per tooth. We compute the centroids $c = \{c_v \subseteq \mathbb{R}^3\}$ for each tooth and apply decentralization to reduce the impact of sparse spatial distribution. Missing teeth are handled with zero-filling, ensuring the consistent sequence length while excluding zero-filled points during decentralization and normalization.

Typically, the centroid sequence starts and ends with the third molar, but missing teeth at these endpoints can affect accuracy of centroid estimation for adjacent teeth. To overcome this, we create multiple centroid sequences starting from the anterior regions, considering interactions between sequences for more reliable predictions. This

approach handles missing teeth data and ensures a reasonable coarse alignment, setting the stage for fine adjustments.

Feature Extraction Module. The key characteristics of a tooth mainly include its surface morphology and orientation information in 3D space. To represent tooth orientation, we construct SSM [15] for each tooth, the specific construction details of which are described in our previous work [16]. We select three pairs of feature points from the SSM average model, corresponding to the independent principal 3D directions as shown in Fig. 1. The axial directions of the sample data are then determined by fitting the SSM. This approach ensures higher accuracy in representing directional axes, particularly in cases of irregular tooth shapes or severe misalignments, by utilizing learned deformation patterns from the SSM rather than relying on fixed, non-deformable templates.

For tooth morphology feature extraction, we integrate PointNet++ and Transformer architectures to comprehensively capture both geometric details and spatial relationships. The design integrates global and local features, with Transformer blocks enhancing feature interactions through self-attention. Additionally, spatial position encoding is incorporated to refine feature representation.

Parameter Regression Module. Instead of predicting the intertwined 4×4 transformation matrix directly, we decomposed the rigid transformation into separate rotation around centroid (represented by quaternions) and translation parameters, both predicted independently through regression as Φ. To simulate gradual orthodontic adjustments, we use a diffusion model D with an U-ViT [17] network as the regression core. It progressively refines tooth postures by denoising, leveraging global/local features and centroid positions as condition, and outputs the precise transformation parameters. We employ Denoising Diffusion Implicit Models (DDIM) [18] for efficient sampling, requiring only minimal iterations due to the pre-aligned teeth from the initial stage.

$$\Phi = D(F_{orientation} \oplus F_{shape} \oplus c^{coarse}) \tag{1}$$

where $F_{orientation}$ and F_{shape} extracted orientation and morphology features, and c^{coarse} is the predicted centroid coordinates in the first stage.

Transformation Module. The transformation module applies the network-predicted tooth transformations to the point cloud data, resulting in a finely tuned aligned point cloud P^{fine}. To simplify computation, we convert the predicted rotation quaternions into a 3×3 rotation matrix T_{rot}, which is then combined with the predicted translation parameters T_{trans} to transform the coarse aligned point cloud P^{coarse}. It is worth noting that the center of rotation of the T_{rot} is the coarse aligned point cloud centroid c^{coarse}.

$$\begin{pmatrix} (P^{fine})^T \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & T_{trans} \\ & 1 \end{pmatrix} \begin{pmatrix} 1 & c^{coarse} \\ & 1 \end{pmatrix} \begin{pmatrix} T_{rot} & \\ & 1 \end{pmatrix} \begin{pmatrix} 1 & -c^{coarse} \\ & 1 \end{pmatrix} \begin{pmatrix} (P^{coarse})^T \\ 1 \end{pmatrix} \tag{2}$$

2.3 Loss Function

The ground truth aligned point cloud P_{gt}^{fine} and centroids c_{gt} can be derived from the provided transformation matrix T_{gt}.

During the centroid coarse alignment phase, the mean squared error between the predicted and ground truth centroid coordinates serves as the supervisor:

$$\mathcal{L}_{coarse} = \frac{1}{n}\sum\left\|c^{coarse} - c_{gt}\right\|_2^2 \tag{3}$$

where n is the number of teeth.

The loss function for the second stage $\mathcal{L}_{fine}$ can be expressed as:

$$\mathcal{L}_{fine} = \mathcal{L}_{pts} + \mathcal{L}_c + \mathcal{L}_{rot} \tag{4}$$

where $\mathcal{L}_{pts}$ and $\mathcal{L}_c$ respectively denote the coordinates supervision of the point cloud and centroid, both of which use the same computational method as $\mathcal{L}_{coarse}$. $\mathcal{L}_{rot}$ denotes the supervision of the rotation parameters.

As described in Sec. 2.2, we use quaternion-based rotation. To ensure consistency with it, we convert the 3×3 rotation matrix ground truth to quaternion form and compute Cosine Similarity Accuracy (CSA). A regularization term is added to penalize non-unit-length quaternions.

$$\mathcal{L}_{rot} = (1 - \left|\left\langle \Phi^{quater}, \Phi_{gt}^{quater} \right\rangle\right|) + \lambda(\left\|\Phi^{quater}\right\| - 1)^2 \tag{5}$$

where $|\langle\cdot\rangle|$ denotes the absolute value of the inner product, $\|\cdot\|$ denotes the Euclidean norm, and the weight factor for regularization λ is set to 1 in in our experiment.

2.4 Data Augmentation

To enhance training data diversity and improve model robustness, we applied several data augmentation techniques: (a) Integral Dentition Rotation simulates scanner or operator variations by applying random rotations of [-15, +15] degrees along the x, y, and z axes. (b) Individual Tooth Rotation and Translation were performed by randomly selecting up to three teeth per case, applying random rotations of [-10, +10] degrees and translation vectors sampled from a Gaussian distribution $\mathcal{N}(1.5,0.01)$ mm. Anterior teeth received more frequent transformations, aligning with clinical observations. (c) To ensure consistency, the ground truth was updated: global dentition rotation was applied to both pre- and post-orthodontic point clouds, while individual tooth transformations kept the post-orthodontic ground truth unchanged. Each training case underwent one global dentition rotation and two individual tooth transformations, simulating subtle changes and reducing underfitting.

3 Experiments and Results

3.1 Dataset and Evaluation Metrics

We used the ISICDM-ATRC Challenge dataset [19], which includes 94 cases of pre-orthodontic oral scans with rigid-body transformation matrices for each tooth. The dataset features mild to moderate tooth irregularities such as overbite, crossbite, crowding, and spacing (see Fig. 2(a-d)), with 20 to 31 teeth per case, and wisdom teeth in about 10% of cases. Approximately 21% of first premolars were extracted, aligning with typical clinical practice. Missing teeth were treated with zero fillings, and the point cloud size for a single tooth is 128, with 4096 for the entire dentition. The dataset was split into 74 training and 20 testing cases, with random grouping and data augmentation applied during training (see Fig. 2(d-f)).

To evaluate prediction accuracy, we use two metrics: Cosine Similarity Accuracy (CSA), which measures directional consistency between the predicted and ground truth transformation parameters, and Target Alignment Error (TRE) [3], defined as the point-by-point Euclidean distance between the transformed point cloud and ground truth.

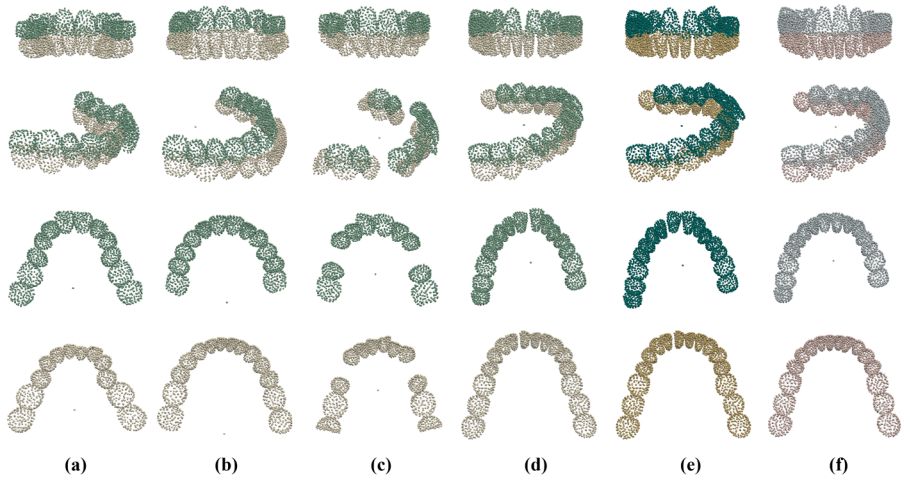

Fig. 2. Training data preview: The four rows illustrate anterior, lateral, maxillary, and mandibular views. (a-d) represent cases with deep overbite, crossbite, tooth extraction, and spacing. (e) shows data augmentation from (d), and (f) is the common ground truth for (d) and (e).

3.2 Implementation Details

In the coarse alignment stage, three Bi-RNNs process tooth centroid sequences, each with three hidden layers of size 32, without weight sharing. In the refined stage, PointNet++ extracts global and local features (1024 dimensions), with tooth position encoding set to 32. The diffusion model uses 60 noise addition steps and 20 sampling steps.

Both training phases use the Adam optimizer with a learning rate of 1e-3, decaying by 0.7 every 20 epochs, and a batch size of 4. The first phase trains for 200 epochs and

the second for 300. All networks are implemented in PyTorch and trained on a NVIDIA RTX 3090 GPU.

3.3 Comparisons

To validate the first-stage centroids coarse alignment method, we tested MLP [20], LSTM [21], GRU [22], Transformer [23], Bi-RNN-s (bidirectional RNN with a single sequence), and RNN (unidirectional RNN with multiple sequences), comparing them with our Bi-RNN model, with all methods using three hidden layers for fairness. As shown in Fig. 3, our method outperforms others in predicting tooth centroids, particularly in handling large gaps from tooth extractions, which simplifies fine-tuning in the second stage. We computed the TRE between predicted and ground truth centroids, showing that our Bi-RNN model achieves the most accurate and stable predictions (0.58±0.16 mm) with a parameter size of just 0.17MB.

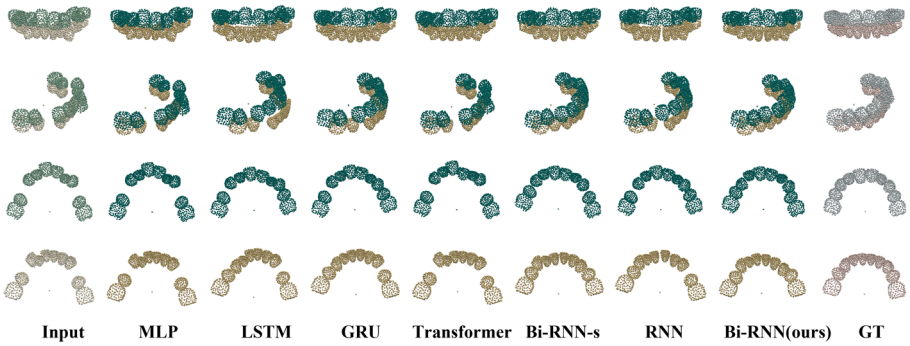

Fig. 3. Visualization of the results comparing the first stage network with other methods. The four rows illustrate anterior, lateral, maxillary, and mandibular views.

To validate the effectiveness of our method, we compared it with PSTN [3], TAligNet [1], and TANet [2], which also rely solely on point cloud data. Fig. 4 visualizes the alignment results of these methods. Quantitatively, as shown in Fig. 5(a), our approach achieves the best performance in both TRE and CSA of rotational parameters. Additionally, Fig. 5(b) presents the point-by-point TRE cumulative distribution with errors below each threshold. Our method achieves a higher cumulative distribution at lower TRE thresholds, indicating higher accuracy and robustness.

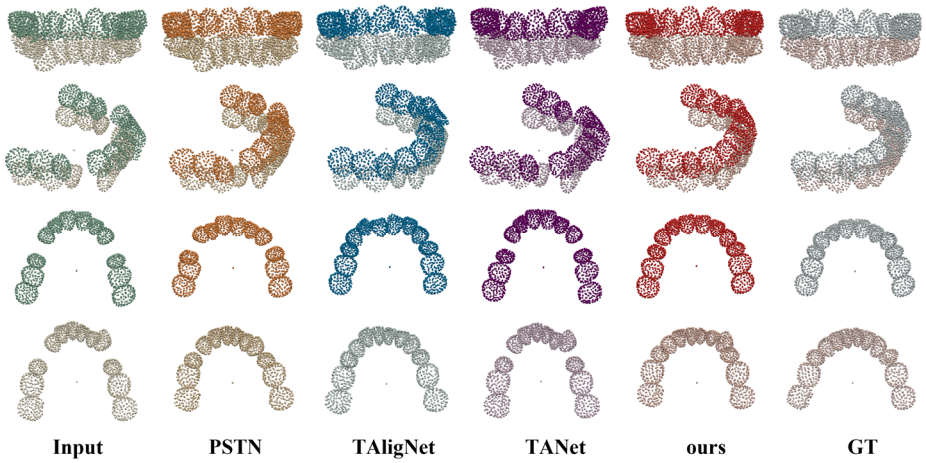

Fig. 4. Visualization of the second stage results comparing with other methods. The four rows illustrate anterior, lateral, maxillary, and mandibular views.

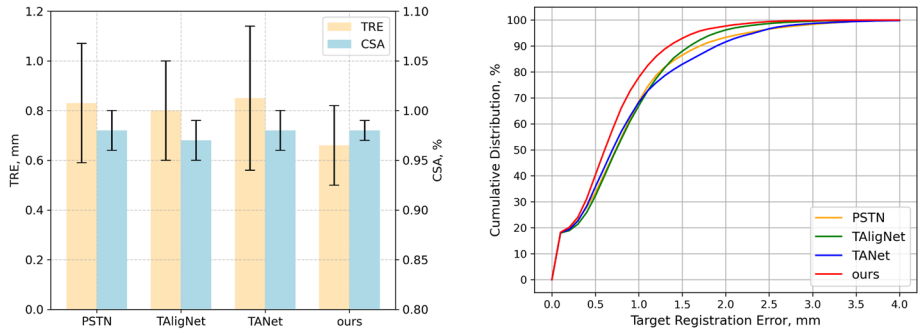

Fig. 5. Quantitative comparison with other methods. (a) Comparison of TRE and CSA. (b) Point-by-point TRE cumulative distribution curve, where the horizontal coordinate is the threshold of TRE, and the vertical coordinate indicates the cumulative distribution of point clouds with errors below threshold.

4 Conclusion

In this study, we propose an innovative two-stage framework for automated tooth alignment that significantly enhances precision and clinical applicability. Our key innovation lies in decoupling the intertwined irregularities of tooth position, orientation, and shape into sequential sub-tasks, thereby addressing complex clinical scenarios more effectively. Our framework has dual strengths: the Bi-RNN's cost-effective positional correction and the diffusion model's clinical-inspired iterative refinement. Additionally, we integrate tooth SSMs to model the orientation features essential for fine-scale

adjustments. Experiments demonstrate that our approach offers enhanced reliability for cases traditionally deemed challenging, such as case with missing teeth.

For future work, we aim to further optimize the diffusion model's convergence efficiency and consider integrating more tooth features (e.g., surface texture, occlusal surface morphology, etc.) to enhance the model's generalizability across complex pathological presentations through validation on more datasets, aligning it more closely to actual medical needs.

Acknowledgments. This work was supported in part by the Joint Funds of the National Natural Science Foundation of China (No. U24A20753), the general program of National Natural Science Fund of China (No. 81971693), the funding of Dalian Key Laboratory of Intellectual Health and Human Digital Twin and Dalian Engineering Research Center for Artificial Intelligence in Medical Imaging and Liaoning technology innovation center of hyperpolarized MRI.

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Chen, B., Fu, H., Zhou, K., Zheng, Y.: Orthoaligner: image-based teeth alignment prediction via latent style manipulation. IEEE Transactions on Visualization and Computer Graphics **29**(8), 3617–3629 (2022)
2. Wei, G., Cui, Z., Liu, Y., Chen, N., Chen, R., Li, G., Wang, W.: Tanet: towards fully automatic tooth arrangement. In: European conference on computer vision. pp. 481–497. Springer (2020)
3. Li, X., Bi, L., Kim, J., Li, T., Li, P., Tian, Y., Sheng, B., Feng, D.: Malocclusion treatment planning via pointnet based spatial transformation network. In: International conference on medical image computing and computer-assisted intervention. pp. 105–114. Springer (2020)
4. Deng, Q., Yang, X., Huang, M., Jiang, L., Zhang, D.: Taposenet: Teeth alignment based on pose estimation via multi-scale graph convolutional network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 314–323. Springer (2024)
5. Yang, L., Shi, Z., Wang, Y., et al.: iortho-predictor: model-guided deep prediction of teeth alignment. ACM Transactions on Graphics **39**(6), 216 (2020)
6. Wang, C., Wei, G., Wei, G., Wang, W., Zhou, Y.: Tooth alignment network based on landmark constraints and hierarchical graph structure. IEEE Transactions on Visualization and Computer Graphics **30**(2), 1457–1469 (2022)
7. Lei, C., Xia, M., Wang, S., Liang, Y., Yi, R., Wen, Y.H., Liu, Y.J.: Automatic tooth arrangement with joint features of point and mesh representations via diffusion probabilistic models. Computer Aided Geometric Design **111**, 102293 (2024)
8. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
9. 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)

10. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE transactions on neural networks **20**(1), 61–80 (2008)
11. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
12. Fan, Y., Wei, G., Wang, C., Zhuang, S., Wang, W., Zhou, Y.: Collaborative tooth motion diffusion model in digital orthodontics. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1679–1687 (2024)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Schuster, M., Palwal, K.K.: Bidirectional recurrent neural networks. IEEE transactions on Signal Processing **45**(11), 2673–2681 (1997)
15. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer vision and image understanding **61**(1), 38–59 (1995)
16. Jin, X., Wang, S., Hu, J., et al. Automated tooth crown design with optimized shape and biomechanics properties. Frontiers in Bioengineering and Biotechnology 11, 1216651 (2023).
17. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A vit backbone for diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22669–22679 (2023)
18. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
19. Isicdm 2024 challenge. https://www.imagecomputing.org/isicdm2024/index.html#/Challenge/Two, accessed: 2024-04-09
20. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. nature **323**(6088), 533–536 (1986)
21. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
22. Cho, K., Van Merriénboer, B., Gulçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1724–1734 (2014)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)