Transformer-based Tooth Alignment Prediction with Occlusion and Collision Constraints

Anonymous ICCV submission

Paper ID 7185

Abstract

The planning of digital orthodontic treatment requires providing tooth alignment, which relays clinical experiences heavily and consumes a lot of time and labor to determine manually. In this work, we proposed an automatic tooth alignment neural network based on Swin-transformer. We first re-organized 3D point clouds based on dental arch lines and converted them into order-sorted multi-channel textures, improving both accuracy and efficiency. We then designed two new orthodontic loss functions that quantitatively evaluate the occlusal relationship between the upper and lower jaws. They are important clinical constraints, first introduced and lead to cutting-edge prediction accuracy. To train our network, we collected a large digital orthodontic dataset in more than 2 years, including various complex clinical cases. We will release this dataset after the paper's publishment and believe it will benefit the community. Furthermore, we proposed two new orthodontic dataset augmentation methods considering tooth spatial distribution and occlusion. We compared our method with most SOTA methods using this dataset, and extensive ablation studies and experiments demonstrated the high accuracy and efficiency of our method.

1. Introduction

Tooth correction, medically known as orthodontics[32], primarily involves the use of metal braces[9] or clear aligners[18] to alleviate or rectify the conditions of dental misalignment and malformation. With the widespread adoption of digital acquisition technologies, computer-aided alignment design has been paid extensive attention, such as those based on intraoral scanners[35] and cone-beam computed tomography (CBCT)[2]. 3D tooth models are initially segmented individually[16, 24, 45, 46], and then repositioned by the clinician considering various alignment factors such as the extent of dental protrusion, dental skeletal relationship, and periodontal conditions of the patient, etc. It heavily relies on the clinical expertise of orthodontists and is time-consuming, thereby significantly increasing the duration and cost of orthodontic treatment planning.

With the advancement of artificial intelligence, learning-based methods for tooth alignment are emerging rapidly[20], aiming at achieving fully automated tooth alignment. Among these methods, PointNet-based[33] ones are particularly representative[21, 23, 25, 44]. TANet[44] employs PointNet to construct a feature extraction module, encoding both jaw global information and teeth local information, and utilizes MLP to design regressors for predicting the position of each tooth. PSTN[23] utilizes both PointNet[33] and PointNet++[34] for feature encoding, refining features based on a combination of local and global latent vectors to regress tooth transformation parameters. TAligNet[25], also based on PointNet encoders and MLP decoders, employs Squeeze-and-Excitation Blocks[13] and shared FC sequences for feature propagation to predict alignment parameters.

PointNet-based tooth alignment prediction methods showed great potential, but limitations have been revealed in representing local features of point clouds[40] recently. This paper introduces a more advanced shift window transformer (referred to as Swin-T). It incorporates sliding window operations and hierarchical merging design on the foundation of traditional vision transformers, addressing issues such as lower precision due to the variability of objects, excessive pixel count leading to high computational complexity, and low computational efficiency encountered in transformer-based methods. As teeth share similar sizes and structures, they are sampled into uniformly sized 3D point clouds and transformed into regular multi-channel data, forming ordered data. Building upon this data organization, this paper proposes a multi-level channel compression structure based on Swin-T (SWTBS) and an SWTP module to respectively extract global information of tooth centers and local information of tooth point clouds. Benefiting from the performance optimization of shift windows and communication between windows, features of individual teeth can mutually inform one another, gradually expanding the receptive field and exhibiting excellent global control over entire dental arches[14], thereby achieving higher prediction accuracy.

We have collected a dataset of 855 orthodontic alignment plans in two years, to be released with the paper and source code. We also introduce constraint data augmentation during preprocessing, evaluated in our experiments. Our dataset, consisting of manually aligned initial tooth models, offers two advantages over previous methods using intraoral scan data: (1) the aligned labels correspond directly with the original scan data, avoiding complex point cloud matching, and (2) dentists' manual alignment is considered a more reasonable ground truth than post-orthodontic scan data.

To summarize, our contributions are as follows:

- A lightweight tooth alignment network based on Swin-T is designed to replace traditional 3D point cloud feature extraction encoders. It organizes scattered point clouds into regularly sized and orderly sorted multi-channel texture forms, ensuring high efficiency and seamless compatibility with complex scenarios such as missing teeth and wisdom teeth, surpassing the accuracy of the SOTA method in tooth alignment.
- Two occlusal loss functions, the occlusal projecting overlap loss and occlusal distance uniformity loss, are designed based on medical domain knowledge. These functions enable more accurate and efficient quantitative measurement for the occlusal relationship between the upper and lower jaws.
- An extensively annotated orthodontic alignment dataset, tailored to better suit the requirements of orthodontists, has been labeled. It will be released after the paper's publishment and benefits the community. Additionally, two new orthodontic data augmentation methods considering tooth spatial distribution and occlusion are proposed to further increase the scale of training data.

2. Related Works

2.1. Learning-based tooth alignment

Existing AI tooth alignment methods primarily use 3D point cloud data as input, rather than mesh or voxel data[27, 49]. Early AI tooth alignment methods were mainly based on the PointNet[33] structure and its derivatives. TANet[44] utilizes PointNet[33] to encode the point cloud features of intraoral scan segmentation models, including global and local features. It then employs graph neural networks to connect and communicate tooth local features, regressing the 6DOF information[41] of teeth. PSTN[23] uses PointNet[33] to encode global and local features and PointNet++[34] to encode local features. After fusion, it uses a decoder designed based on PointNet[33] to regress orthodontic transformations of teeth. TAligNet[25] achieves feature extraction of 3D tooth models and tooth arrangement. It utilizes PointNet[33] as the feature encoder and employs fully connected layer sequences and SE blocks[13] for feature propagation, finally using fully connected layers to regress rotation and translation.

Besides PointNet, recent AI tooth alignment methods also adopted emerging network structures such as DGCNN[43] and diffusion models[12]. Wang et al. proposed an improvement to TANet using tooth landmarks, where DGCNN was utilized to extract point cloud information[40]. They proposed a hierarchical regression using a three-layer graph neural network structure[36] to better predict the displacement transformation of each tooth, the landmark serves as a component of the tooth frame. TAPoseNet employed DGCNN to predict the local coordinate axes of teeth and utilized an autoencoder to extract geometric information[5]. Additionally, they proposed a multi-scale GCN[50] to characterize the spatial relationships between teeth at different levels, enabling more accurate prediction of the target positions of teeth. LETA[37] extracts features through latent encoding of dual branches (original data & ground truth), and predicts 6DoF of teeth by utilizing encoding differences. During training, GT point clouds are required, while only original point clouds are needed to complete prediction.

Lei et al. employed probabilistic diffusion models[38] to iteratively denoise random variables, learning the distribution of transformation matrices for dental transitions from malocclusion to normal occlusion, thus achieving more realistic orthodontic predictions[21]. Furthermore, the network structure of TAligNet mentioned above was actually proposed in image-based tooth alignment methods, iOrthoPredictor[25], which utilize three-dimensional geometric information encoded in the unsupervised generative model StyleGAN[15]. Through meaningful paths in latent space normals, alignment processes in image space are generated. Due to the complexity of the occlusal action between the upper and lower teeth, the calculation methods based on angles or center points have large errors and limited effectiveness, thus having deficiencies.

2.2. Shift window transformer

Transformers[39] have achieved major advancements in NLP[29] and computer vision domains[19]. Vision Transformer (ViT)[6] directly applies self-attention to image patches, achieving strong results in classification without CNNs[17]. Swin-T[26] builds on ViT with movable windows, limiting sub-attention to non-overlapping local regions for greater efficiency. Swin3D[51] adapts Swin-T for 3D point clouds, converting sparse points into voxel grids and using farthest point sampling (FPS)[8, 28] and KNN pooling[22], though this is computationally intensive and misses advantages from serializing relative positions. Our

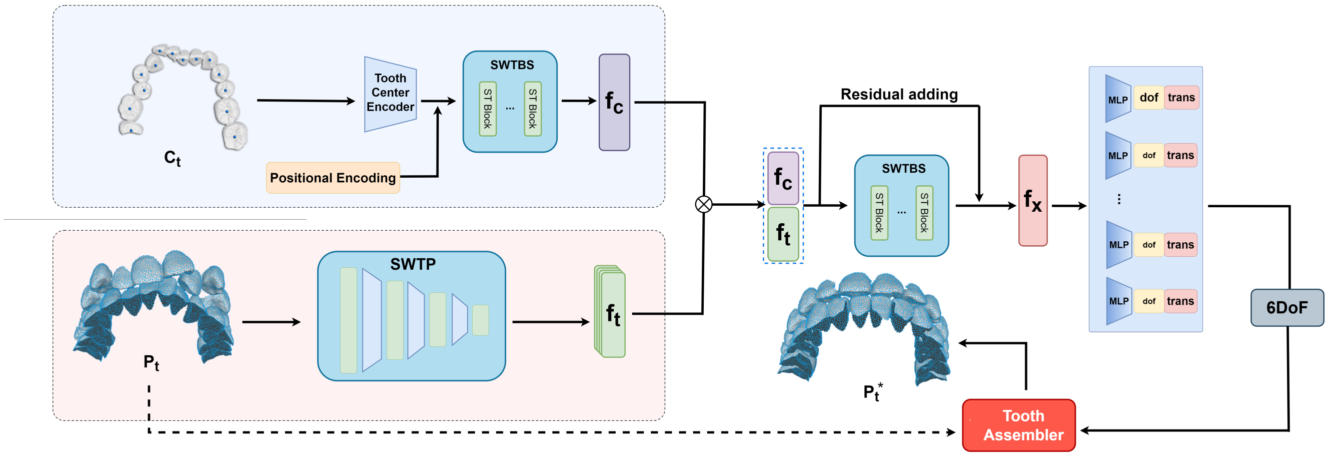


Figure 1. Network architecture overview. The encoding module has two branches: one for global features from the tooth center and one for local features from the tooth point cloud. Global features are extracted using SWTBS with shared Swin-T blocks, while local features are processed via SWTP with multi-stage hierarchical fusion. The features are merged, passed through SWTBS propagation, and then regressed by an MLP to predict the 6DOF transformation parameters for orthodontics

approach enhances the multi-level fusion network in Swin-T[26] by using sliding windows to reduce data size layer by layer, optimizing efficiency and expanding global receptive fields. We use FPS[28] to uniformly sample and structure each tooth's data, avoiding direct KNN downsampling as in[53] and[47], to improve information propagation.

3. Methodology

3.1. Network overview

We segment the patient's intraoral scan model to obtain the gingival point cloud G and crown point clouds T for 32 teeth. These include up to 16 upper teeth (numbered 1 to 16) and 16 lower teeth (numbered 17 to 32), following the universal tooth naming standard. Missing teeth are discussed in Section 3.2. To ensure training efficiency, farthest point sampling[30] is applied with $N = 512$ to balance sampled point quantity and network performance.

$$T = \{t_i | 1 \leq i \leq 32\} \tag{1}$$

$$P = \{p_j^t | t \in T, 1 \leq j \leq 512\} \tag{2}$$

Each tooth's centroid is defined as C_t, where c_t is the geometric centroid of the sampled point cloud P_t.

$$C_t = \{c_t | t \in T, c_t = \underset{p \in P_t}{Ave}(p)\} \tag{3}$$

The goal of this study is to predict the 6Dof[7] pose transformation parameters for each tooth model, using both pre/post-orthodontic data. To ensure accurate loss calculation during training, the dataset includes corresponding ground truth data P_t^*, with consistent sampling positions and orders for P_t and P_t^*.

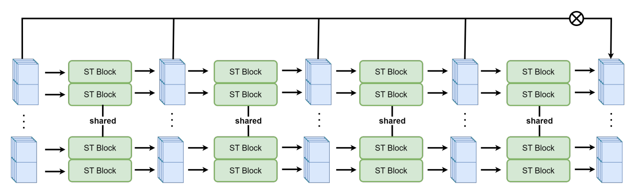


Figure 2. SWTBS module: Four groups of shared Swin-T blocks, each with 16 channels, with residuals added to the final output.

This paper proposes a dual-module architecture for feature extraction, as illustrated in Figure 1. The global module encodes tooth center points using MLP layers and positional encoding, followed by Swin-T blocks (SWTBS) to produce f_c. The local module processes the 3D tooth point cloud through the Swin transformer pipeline (SWTP), producing f_t. After pooling and merging f_t with f_c, the resulting high-dimensional vector $F = \{f_c, f_t\}$ is processed by SWTBS to yield f_x, enhancing feature optimization. Finally, a downsampling regression module obtains the 6DoF transformation parameters for orthodontics.

The tooth point cloud is processed through hierarchical downsampling, similar to the Swin Transformer[26], involving patch partitioning and feature merging across four stages, as shown in Figure 3. Unlike the original network, we keep the channel count constant to prevent excessive feature dimension and loss during MLP conversion. Data columns, rather than rows, are merged during feature processing, as the first dimension represents the number of teeth, and each tooth's rotations and translations are unique.

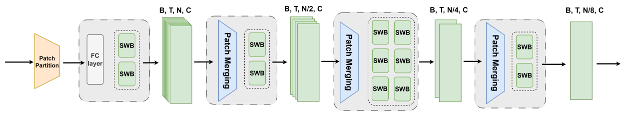


Figure 3. SWTP module, featuring a multi-stage feature fusion mechanism.

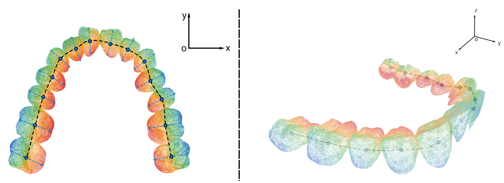


Figure 4. Points cloud are serialized according to their distance from simulated dental arch line, values on lingual side set as positive while labial side set as negative.

3.2. Data re-organization and augmentation

We designed a serialization method based on a simulated dental arch line, created by fitting central points from a tooth segmentation model and connecting them with Hermite curves. Points are sorted by their distance from the arch, with labial points positive and lingual points negative. Results in Figure 4 maintain relative positions among the 512 sampled points, improving network performance.

Regular data augmentation applies random rotation and translation on teeth based on a Gaussian distribution. It generates pre-orthodontic data while preserving the ground truth as post-orthodontic data. However, regular augmentation may produce clinical-illogical cases, including too far away from the arch lines and teeth collision. For this sake, we propose a constrained data augmentation that involves two relevant clinical constraints:

3.2.1. Jaw regularization constraint

If the distance between two teeth exceeds 2.35 mm, the farther tooth is moved towards the central incisor along the dental arch line until the gap is within the threshold, as shown by the red teeth in Figure 5. Teeth exceeding 2.2 mm from the arch are pulled inward, as shown by the blue teeth, based on dataset statistics. Our strategy is to move the distal tooth towards the mesial tooth, specifically in the direction of the central incisor until the inter-tooth gap is within the threshold. This approach minimizes large gaps in the augmented dataset, ensuring more reasonable data. If the movement increases inter-tooth distance in the opposite direction or causes tooth collisions, the method in Section 3.2.2 is used for detection and avoidance, all along the simulated dental arch line.

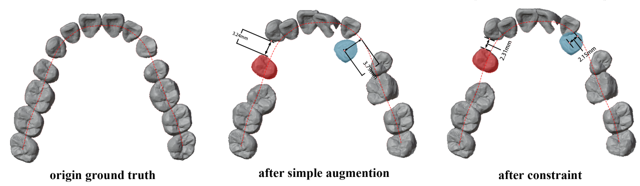


Figure 5. The maxillofacial regularization corrects excessive gaps or deviations based on dataset statistics.

3.2.2. Collision detection constraints

We use a BVH collision detection algorithm[10, 31] to detect collisions and identify the colliding parts. Upon collision, the simulated dental arch line is used to avoid interlocking while preserving the arch shape.

For the efficiency purpose, we parallelize the BVH construction with a tooth-wise multi-threaded acceleration. It is worth noting that the BVH construction and collision detection are only employed in the pre-processing stage, which has more tolerance on the performance.

It is worth mentioning that missing teeth are specially handled, thus have very limited impact on both serialization and data augmentation. Though Missing teeth will affect the feature extraction in a single window, the impact is negligible for the sliding window that moves in the data.

3.3. Loss functions

The global loss function of the network consists of four components, the latter two are specifically designed in this paper to address dental occlusion. Each part of the loss function will be elaborated in the following part of this subsection.

$$L = \delta_0 * L_{recon} + \delta_1 * L_{fit} + \delta_2 * L_{uni} + \delta_3 * L_{val} \quad (4)$$

The hyperparameters δ_0, δ_1, δ_2 and δ_3 are used to weight each component accordingly.

3.3.1. Reconstruction loss

We utilized the model reconstruction loss mentioned in[40]. Different from[40], the post-orthodontic data in our dataset was manually adjusted by orthodontists using the pre-orthodontic data. Therefore, the vertex positions and orders of the models before and after orthodontic treatment are correspond.

$$L_{recon}^{point} = \sum_{t \in T} \left(\sum_{i=0, p \in P_t} ||\overline{p}_i - p_i^*||_2^2 + ||\overline{c}_t - c_t^*||_2^2 \right) \quad (5)$$

3.3.2. Transformation parameter loss

The transformation parameter loss comprises two components: rotation loss and translation loss. We computes

ICCV #7185

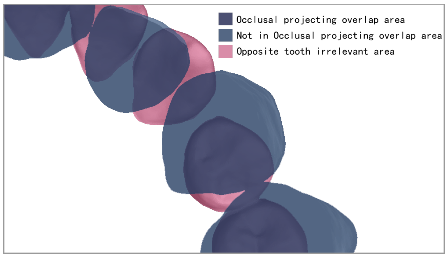


Figure 6. Visualization of the occlusion projection range.

weights for each tooth during training, based on the magnitude of misalignment. These weights are cumulatively added to the original loss, as shown in Equation 7 & 8. Drawing from[52], we emphasize that more severe misalignments should receive greater attention, corresponding to larger loss values.

$$L_{val} = \omega * L_{rotate} + L_{trans} \tag{6}$$

$$L_{rotate} = \sum_{t \in T} \left[\sum_{i=0}^{3} L_1 \left(\overline{rotate}_t(i), rotate_t^*(i) \right) * \left(1.0 + \zeta_t^{rotate} \right) \right] \tag{7}$$

$$L_{trans} = \sum_{t \in T} \left[\sum_{i=0}^{2} L_1 \left(\overline{trans}_t(i), trans_t^*(i) \right) * \left(1.0 + \zeta_t^{trans} \right) \right] \tag{8}$$

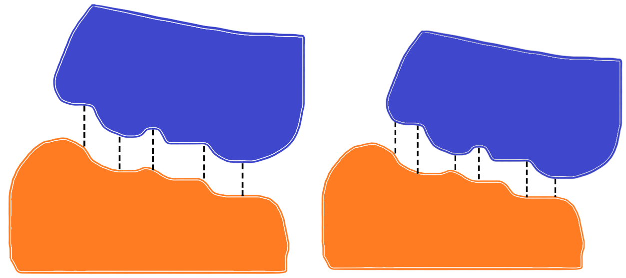
Due to the relatively small numerical values of rotation loss, an additional parameter w is introduced to amplify the impact of quaternion rotation loss during actual training.

3.3.3. Occlusal projecting overlap

Occlusion projection range consistency loss represents whether the interocclusal region between the predicted results of upper and lower jaws matches the ground truth. The definition of occlusion projection range is as follows: Let tooth t in one jaw have a corresponding area β_t in the opposite jaw. Project all points of t and β_t onto the occlusal plane.

$$m_i = \underset{p_j \in P_{\beta_t}^f}{Argmin} \left\| p_i - p_j \right\|_2, p_i \in P_t^f \tag{9}$$

If the closest distance between points from t's point cloud and β's point cloud (on the occlusal plane) is less than a threshold τ, then points from t's point cloud within this distance are considered part of t's occlusion projection range. We introduce the concept of occlusion projection range to bring predicted results closer to the ground truth at the occlusion projection range level. In the Figure 6, P_t^f represents the point cloud of tooth t projected onto the occlusal plane, $P_{\beta_t}^f$ represents the point cloud of region β_t projected onto the occlusal plane, and m_i denotes the minimum two-dimensional plane distance between a point p_i from the point cloud of t and the nearest point p_j from the point cloud of β_t.



(a) Aligned (b) Misaligned

Figure 7. An illustration of occlusal distance uniformity. The variation of occlusion distances (blck dotted lines) in aligned scenarios (a) is much smaller than the one in misaligned scenarios (b).

τ is the threshold used to divide the occlusion projection range. X is a binary sequence where $X_t(i)$ records whether point p_i from the point cloud of tooth t belongs to the occlusion projection range based on the relationship between m_i and τ. If m_i is less than τ, then $X_t(i)$ takes the value of 1; otherwise, it takes the value of 0.

$$L_{fit} = \underset{t \in T}{Ave} \left(\sum_{i=0}^{n-1} \left| \overline{X}_t(i) - X_t^*(i) \right| \right) \tag{10}$$

3.3.4. Occlusal distance uniformity

Due to the completely different morphologies and occlusion patterns between anterior and posterior teeth, we have proposed distinct loss function designs for anterior and posterior teeth based on discussions in [1, 4, 11]. Therefore, the occlusal distance uniformity loss L_{uni} across upper and lower jaws is composed of two parts: L_{uni}^{ant} for anterior teeth and L_{uni}^{pior} for posterior teeth. As shown in the following formula, we introduce a weighting parameter w^{pior} to balance.

$$L_{uni} = L_{uni}^{ant} + w^{pior} \cdot L_{uni}^{pior} \tag{11}$$

This paper evaluates the consistency and similarity of vectors connecting corresponding points in the occlusal regions of posterior teeth, based on the projection ambiguity constraint discussed in[4]. The occlusal distance uniformity is calculated within the occlusal projection range defined in Section 3.3.3. For a point p_i in the occlusal projection range of tooth t, the distance to the nearest point p_j in the corresponding range β_t on the opposing jaw is denoted as d, and the collection of all such distances forms set D. The uniformity of D represents the degree to which the occlusal ranges match concavely or convexly, defining the occlusal distance uniformity loss function for posterior teeth.

$$L_{uni}^{pior} = \sum_{t \in T_{pior}} \underset{X_t(i)=1}{Var} \left(\min_{X_{\beta_t}(j)=1} \|p_i - p_j\|_2 \right) \tag{12}$$

Due to the more prominent crowns of the upper incisors, their shape differs significantly from the molars, so the distance uniformity calculation method for molars cannot be applied. Based on the upper and lower anterior tooth correspondence described in[1, 11], we propose the vertex coordinate difference and angular difference between the tooth axis vector and the ground truth, denoted as L_{uni}^{ant1} and L_{uni}^{ant2}, respectively.

$$L_{uni}^{ant} = L_{uni}^{ant1} + \omega^{ant} * L_{uni}^{ant2} \tag{13}$$

$\overline{Peak}_t$ and $Peak_t^*$ correspond to the highest incisal points in the predicted and ground truth data, respectively.

$$L_{uni}^{ant1} = \sum_{t \in T_{ant}} \left\|\bar{c}_t - c_t^*\right\|_2 + \sum_{t \in T_{ant}} \left\|\overline{Peak}_t - Peak_t^*\right\|_2 \tag{14}$$

$$L_{uni}^{ant2} = \sum_{t \in T_{ant}} arccos \left(\frac{\left(\overline{Peak}_t - \bar{c}_t\right) \cdot \left(Peak_t^* - c_t^*\right)}{\left\|\overline{Peak}_t - \bar{c}_t\right\|_2 * \left\|Peak_t^* - c_t^*\right\|_2} \right) \tag{15}$$

4. Experiments

4.1. Dataset pre-processing and evaluation metrics

Our dataset contains 855 sets of dental data, which are derived from the 3D models of the upper and lower jaw teeth constructed through oral scans. Each set of dental data also includes data from multiple stages during the orthodontic treatment process, usually divided into about 30 stages. We can take the final orthodontic result of each patient as the ground tooth, and multiple treatment stages can respectively form multiple pairs of pre- and post-orthodontic data with the ground tooth. These data are first preliminarily segmented using the semantic segmentation network TSegNet[3], and then manually optimized for the mesh and edges. The optimized crown models are then aligned and arranged by experienced orthodontists to obtain the corresponding ground truth data. We randomly selected 700 samples for training, 35 samples for validation, and the remaining 120 samples for testing. Our labeled data do not come from the intraoral scan after orthodontic, because this would introduce differences in topology aspects on points before and after treatment.

We used an NVIDIA GeForce RTX 3090 (24GB VRAM) for 500-epoch training (batch size 8). Set N=512, initial learning rate 1.5e-4, w for the rotation loss in the transformation parameter loss was set to 10.0, The threshold τ in Section 3.3.3 was empirically set to 0.07mm, shift window size 8×8. Our evaluation metrics used ADD/AUC from TANet[44] and landmark[40].

Table 1. Comparison of evaluation metrics between the proposed method and the SOTA method. Note that * represents the effect of our method on the dataset[42].

Model	Test result			
	ADD ↓	ADD/AUC ↑	ME_{rotate} ↓	$ME_{translate}$ ↓
TAligNet	1.5307	0.72	7.5461	2.0392
TANet	1.0075	0.81	6.9274	1.6815
PSTN	1.5889	0.71	8.6938	2.2155
Ptv3	1.2136	0.78	7.0663	1.7581
Landmark	0.8139	0.84	7.8277	1.3764
TADPM	1.1815	0.76	7.7426	1.7351
Ours*	**0.8115**	**0.84**	**2.9338**	**1.5904**
Ours	**0.6584**	**0.89**	**2.7678**	**1.1584**

4.2. Comparisons with SOTA methods

We tested the performance of some advanced methods on our dataset, including TANet[44], PSTN[23], TAligNet[25], Landmark[40], and TADPM[21]. Among them, the results of Landmark were obtained by the Wang et al. when they ran our dataset. We debugged the open-source code released by Lei et al. and reformatted our dataset according to the dataset format they published[42] to obtain the test results of TADPM. Although due to the limitations of the equipment, we used a more simplified dental mesh model and a smaller batch size, which may lead to some differences compared with the results published by Lei et al.[21], conducting the training and testing on the same hardware allows for a more rigorous comparison of the performance of all the methods. In addition, we processed their dataset into our format and conducted training and testing using our method, as shown in Table 1.

Training and testing data and specifications were consistent with Section 4.1. We compared ADD/AUC, average rotation error, and average translation error. Table 1 shows that our method performs best in all aspects, whether for AUC or rotational and translational deviations. Additionally, compared with the TADPM method, during the training and testing process, our method takes much less time to process a single case. Moreover, under the same time and equipment conditions, the quality of our method is better.

We compared the curves of average point distances, as shown in Figure 8. It is evident from the figure that our method achieves the highest accuracy under different definitions of average point distance. It is noteworthy that beyond an average point distance of 2.5, all curves converge to nearly 1.0. Therefore, the chart only displays curves for $k \leq 2.5$.

Figure 9 shows aligned tooth models achieved by our method versus others, with views from the front, side, and top. Our occlusal projection range alignment loss and occlusal distance uniformity loss help correct upper-lower jaw gaps and occlusal misalignment better than Chamfer vector loss. Landmarks improve focus on joint points, resolving misaligned gaps. Additionally, data serialization ensures ac-

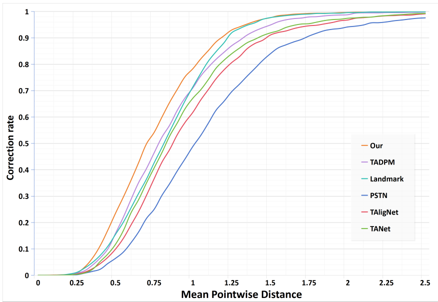


Figure 8. Comparison of accuracy curves between the proposed method and the SOTA method.

Table 2. Ablation experiment results of loss functions, testing the impact of different loss function combinations on final test results.

Loss fuc	Test result		
	$ADD/AUC \uparrow$	$ME_{rotate} \downarrow$	$ME_{translate} \downarrow$
L_{recon}	0.64	9.6	2.7
L_{val}	0.62	10.5	3.1
$L_{recon} + L_{val}$	0.79	8.3	2.2
$L_{recon} + L_{val} + L_{uni}$	0.81	5.9	1.7
$L_{recon} + L_{val} + L_{fit}$	0.83	5.3	1.4
$L_{recon} + L_{val} + L_{uni} + L_{fit}$	**0.89**	**2.7**	**1.1**

curate recognition of inter-tooth positions, even with incomplete models, effectively handling issues with wisdom teeth and intra-jaw misalignment.

4.3. Visual results with complex cases

Figure 10 shows a comparison of our tooth alignment results. It can be seen that our method produces very neat alignments even in complex situations such as large gaps, crossbite, and triangular tooth arrangements. The introduction of serialization enhances the transformer's ability to perceive the positions of teeth within the jaw, and the shift window efficiently extracts local features of the teeth, supported by a comprehensive combination of loss functions. And our network specifying a maximum of 16 teeth per jaw, can handle cases with wisdom teeth or missing teeth.

4.4. Ablation study

4.4.1. Loss functions

We discussed several loss functions in Section 3. Results of ablation experiments validating their effectiveness are as follows. Model reconstruction and transformation parameter losses are effective. Our proposed occlusal losses improve prediction accuracy based on medical principles, which ensures that the upper and lower jaws are closely aligned according to natural occlusion laws and gradually move to their correct positions, despite longer training time.

4.4.2. Network architecture

Table 3 shows that using the tooth center feature module, especially SWTBS, yields better accuracy, as Swin-T's sliding

Table 3. Ablation experiment results of network architecture.

Methods	w/o SWTP	$ADD/AUC \uparrow$	$ME_{rotate} \downarrow$	$ME_{translate} \downarrow$
VTBS	✓	0.79	7.10	1.83
PTv3	✓	/	/	/
SWTBS(Ours)	✓	**0.90**	**2.70**	**1.10**
VTBS	×	0.75	8.50	2.20
PTv3	×	0.73	9.80	2.40
SWTBS(Ours)	×	0.81	7.20	2.00

Table 4. Ablation experiment results of serialization.

Serialization Function	Test result		
	$ADD/AUC \uparrow$	$ME_{rotate} \downarrow$	$ME_{translate} \downarrow$
Random Order	0.77	6.1	1.9
Based on dental local z-axis	0.80	5.4	1.7
Based on dental arch center	0.82	5.6	1.3
Based on virtual arch line	**0.89**	**2.7**	**1.1**

window and merging mechanisms enhance point cloud fusion. Without SWTP, accuracy decreases across modules, showing that SWTP's multi-stage architecture better captures global features. Using PTv3 alone further reduces performance due to lacking mechanisms like sliding windows, critical for effective inter-tooth feature extraction.

It can be observed that compared to network structures using only Swin blocks or Vision blocks, the multi-stage Swin block structure yields higher accuracy. This is because the multi-stage approach reduces the size of the latent vector progressively through dimensional merging, which is more effective in retaining task-specific dental features than directly passing down through averaging.

4.4.3. Point cloud serialization

We discussed sorting points of individual teeth for input into the Swin-T multi-layer feature fusion module in Section 3. Serialization ensures points selected by the window corresponds to the same local region of the teeth, can better extract relative position features between teeth[48].

As shown in Table 4, the random sorting method was worst as it couldn't use sequential information to boost the transformer's performance. Sorting by the local Z-axis of the teeth had sequence benefits and better performance, but it was still not enough because of multi-peaked tooth crowns (common in posterior teeth), making sequentially arranged points in different local regions. The center-point-based sorting method solved this problem but had angular deviations for posterior teeth since the teeth are U-shaped. Our method based on the simulated dental arch curve for U-shaped jaws achieved best prediction results.

4.4.4. Data augmentation

Our constrained data augmentation increases training the data scale, but excessive augmentation may lead to network distortion. To avoid it, we restrict the mount of augmented cases. Figure 11 shows the perdiction accuracy with varying augmentation ratio of augmented cases to original ones, it reaches the peak when the ratio is 54%. The accuray

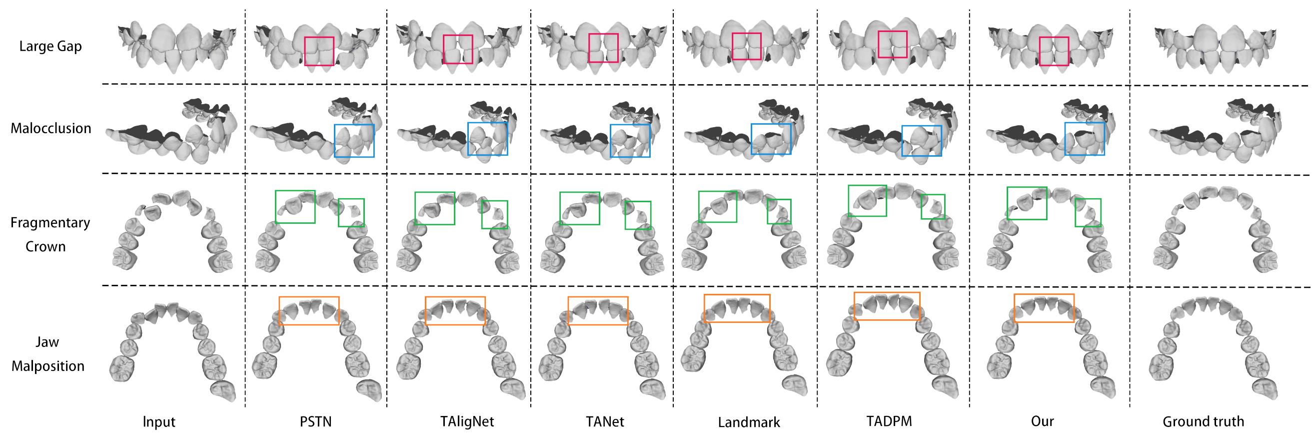


Figure 9. Comparison of prediction results between the proposed method and other methods. Here, four typical and challenging orthodontic problems are selected, listed from top to bottom: large gap, malocclusion, fragmentary crown, and jaw malposition.

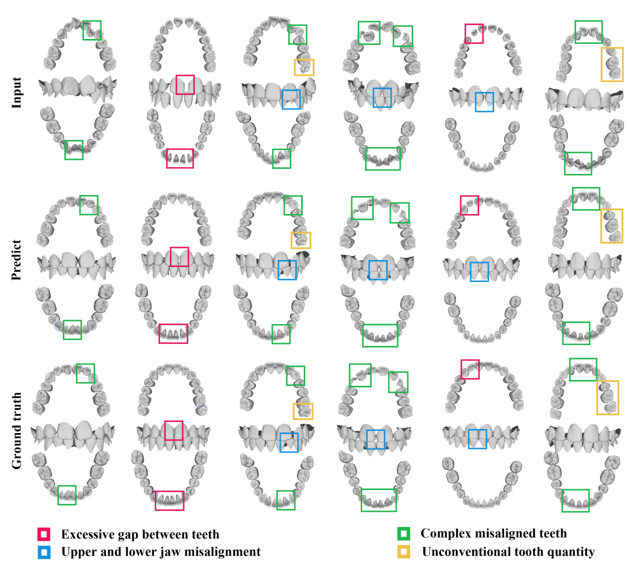
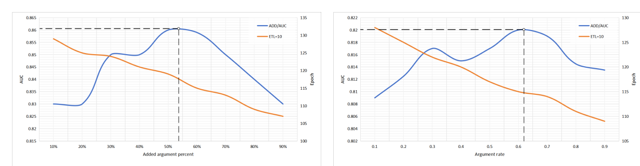


Figure 10. The figure shows the alignment prediction results of our method on 6 data cases in 6 columns.



(a) Constrained augmentation (b) Regular augmentation

Figure 11. The effect of data augmentation intensity on final prediction accuracy (blue curves) and training convergence speed (orange curves).

Table 5. Ablation experiment results of data augmentation.

Augmentation		$ADD/AUC \uparrow$	$ETL = 10 \downarrow$	$SigAugTime \downarrow$
None		0.83	145	/
Source Data	Regular	0.84	125	0.84
	Constrained	0.86	141	3.53
Target Data	Regular	0.85	109	0.84
	Constrained	**0.90**	**134**	**3.51**

of regular augmentation reaches the peak when the ratio is 62%. We compared our constrained augmentation with regular augmentation, on both source data and target data. Table 5 shows accuracy statistics with these scenarios, it gives two insights: (1) constraint augmentation contributes more significantly than regular augmentation, as regular augmentation ignores clinical requirements and introduces more biases, and (2) augmentation on target data provides better training quality than source data.

5. Conclusions

This paper proposes a novel, high-precision and efficient neural network approach for tooth alignment prediction. It uses the multi-level feature fusion structure of Swin-T as its core, supplemented by a tooth center feature extraction module that emphasizes global features. Two occlusion evaluation loss functions are designed to effectively describe the occlusal relationships between upper and lower jaws. Furthermore, this paper constructed a open dataset in the field of tooth arrangement. This dataset includes over 855 fully annotated data pairs, consisting of point clouds sampled from tooth crowns, addressing the issue of a lack of public datasets in this field. A new constrained augmentation method is proposed to further augment the datasets. For future work, we plan to consider other stomatologic constraints for the tooth alignment task, and predict the full path of the tooth orthodontic treatment instead of the target position.

References

[1] Lawrence F Andrews. The six keys to normal occlusion. *Am J orthod*, 62(3):296–309, 1972. 5, 6

[2] Cheng Cheng, Xiaosheng Cheng, Ning Dai, Yi Liu, Qilei Fan, Yulin Hou, and Xiaotong Jiang. Personalized orthodontic accurate tooth arrangement system with complete teeth model. *Journal of medical systems*, 39:1–12, 2015. 1

[3] Zhiming Cui, Changjian Li, Nenglun Chen, Guodong Wei, Runnan Chen, Yuanfeng Zhou, Dinggang Shen, and Wenping Wang. Tsegnet: An efficient and accurate tooth segmentation network on 3d dental model. *Medical Image Analysis*, 69: 101949, 2021. 6

[4] S Davies and RMJ Gray. What is occlusion? *British dental journal*, 191(5):235–245, 2001. 5

[5] Qingxin Deng, Xunyu Yang, Minghan Huang, Landu Jiang, and Dian Zhang. Taposenet: Teeth alignment based on pose estimation via multi-scale graph convolutional network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 314–323. Springer, 2024. 2

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[7] Guoguang Du, Kai Wang, Shiguo Lian, and Kaiyong Zhao. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. *Artificial Intelligence Review*, 54(3):1677–1734, 2021. 3

[8] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y Zeevi. The farthest point strategy for progressive image sampling. *IEEE transactions on image processing*, 6 (9):1305–1315, 1997. 2

[9] Theodore Eliades, Spiros Zinelis, Christoph Bourauel, and George Eliades. Manufacturing of orthodontic brackets: a review of metallurgical perspectives and applications. *Recent Patents on Materials Science*, 1(2):135–139, 2008. 1

[10] Yan Gu, Yong He, Kayvon Fatahalian, and Guy Blelloch. Efficient bvh construction via approximate agglomerative clustering. In *Proceedings of the 5th High-Performance Graphics Conference*, pages 81–88, 2013. 4

[11] LT Hiew, SH Ong, and Kelvin WC Foong. Optimal occlusion of teeth. In *2006 9th International Conference on Control, Automation, Robotics and Vision*, pages 1–5. IEEE, 2006. 5, 6

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1, 2

[14] Silvana Allegrini Kairalla, Giuseppe Scuzzo, Tarcila Triviño, Leandro Velasco, Luca Lombardo, and Luiz Renato Paranhos. Determining shapes and dimensions of dental arches for the use of straight-wire arches in lingual technique. *Dental press journal of orthodontics*, 19:116–122, 2014. 2

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[16] Suzi Kim and Sunghee Choi. Automatic tooth segmentation of dental mesh using a transverse plane. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4122–4125. IEEE, 2018. 1

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[18] Kislaya Kumar, Shivani Bhardwaj, and Vishal Garg. Invisalign: A transparent braces. *Journal of Advanced Medical and Dental Sciences Research*, 6(7):148–150, 2018. 1

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[20] Min Kyeong Lee, Veerasathpurush Allareddy, Sankeerth Rampa, Mohammed H Elnagar, Maysaa Oubaidin, Sumit Yadav, and Shankar Rengasamy. Applications and challenges of implementing artificial intelligence in orthodontics: A primer for orthodontists. In *Seminars in Orthodontics*. Elsevier, 2024. 1

[21] Changsong Lei, Mengfei Xia, Shaofeng Wang, Yaqian Liang, Ran Yi, Yuhui Wen, and Yongjin Liu. Automatic tooth arrangement with joint features of point and mesh representations via diffusion probabilistic models. *arXiv preprint arXiv:2312.15139*, 2023. 1, 2, 6

[22] Jiaxin Li, Ben M Chen, and Gim Hee Lee. So-net: Self-organizing network for point cloud analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9397–9406, 2018. 2

[23] Xiaoshuang Li, Lei Bi, Jinman Kim, Tingyao Li, Peng Li, Ye Tian, Bin Sheng, and Dagan Feng. Malocclusion treatment planning via pointnet based spatial transformation network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 105–114. Springer, 2020. 1, 2, 6

[24] Sheng-hui Liao, Shi-jian Liu, Bei-ji Zou, Xi Ding, Ye Liang, and Jun-hui Huang. Automatic tooth segmentation of dental mesh based on harmonic fields. *BioMed research international*, 2015(1):187173, 2015. 1

[25] YANG Lingchen, SHI Zefeng, Wu Yiqian, LI Xiang, ZHOU Kun, FU Hongbo, and Youyi Zheng. iorthopredictor: model-guided deep prediction of teeth alignment. *ACM Transactions on Graphics*, 39(6):216, 2020. 1, 2, 6

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 3

[27] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition.

In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 922–928. IEEE, 2015. 2

[28] Facundo Mémoli and Guillermo Sapiro. A theoretical and computational framework for isometry invariant recognition of point cloud data. *Foundations of Computational Mathematics*, 5:313–347, 2005. 2, 3

[29] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 2

[30] Carsten Moenning and Neil A Dodgson. Fast marching farthest point sampling. Technical report, University of Cambridge, Computer Laboratory, 2003. 3

[31] Jia Pan, Sachin Chitta, and Dinesh Manocha. Fcl: A general purpose library for collision and proximity queries. In *2012 IEEE International Conference on Robotics and Automation*, pages 3859–3866. IEEE, 2012. 4

[32] William R Proffit, Henry Fields, Brent Larson, and David M Sarver. *Contemporary Orthodontics, 6e: South Asia Edition-E-Book*. Elsevier Health Sciences, 2019. 1

[33] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1, 2

[34] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 1, 2

[35] Marta Revilla-León, Dean E Kois, Jonathan M Zeitler, Wael Att, and John C Kois. An overview of the digital occlusion technologies: Intraoral scanners, jaw tracking systems, and computerized occlusal analysis devices. *Journal of Esthetic and Restorative Dentistry*, 35(5):735–744, 2023. 1

[36] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80, 2008. 2

[37] Zefeng Shi, Zijie Meng, Ruizhe Chen, Yang Feng, Zeyu Zhao, Jin Hao, Bing Fang, Zuozhu Liu, and Youyi Zheng. Leta: Tooth alignment prediction based on dual-branch latent encoding. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015. 2

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[40] Chen Wang, Guangshun Wei, Guodong Wei, Wenping Wang, and Yuanfeng Zhou. Tooth alignment network based on landmark constraints and hierarchical graph structure. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 1, 2, 4, 6

[41] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2642–2651, 2019. 2

[42] Shaofeng Wang, Changsong Lei, Yaqian Liang, Jun Sun, Xianju Xie, Yajie Wang, Feifei Zuo, Yuxin Bai, Song Li, and Yong-Jin Liu. A 3d dental model dataset with pre/post-orthodontic treatment for automatic tooth alignment. *Scientific Data*, 11(1):1277, 2024. 6

[43] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019. 2

[44] Guodong Wei, Zhiming Cui, Yumeng Liu, Nenglun Chen, Runnan Chen, Guiqing Li, and Wenping Wang. Tanet: towards fully automatic tooth arrangement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 481–497. Springer, 2020. 1, 2, 6

[45] Kan Wu, Li Chen, Jing Li, and Yanheng Zhou. Tooth segmentation on dental meshes using morphologic skeleton. *Computers & Graphics*, 38:199–211, 2014. 1

[46] Tai-Hsien Wu, Chunfeng Lian, Sanghee Lee, Matthew Pastewait, Christian Piers, Jie Liu, Fan Wang, Li Wang, Chiung-Ying Chiu, Wenchi Wang, et al. Two-stage mesh deep learning for automated tooth segmentation and landmark localization on 3d intraoral scans. *IEEE transactions on medical imaging*, 41(11):3158–3166, 2022. 1

[47] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022. 3

[48] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. *arXiv preprint arXiv:2312.10035*, 2023. 7

[49] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 2

[50] Zhilong Xiong and Jia Cai. Multi-scale graph convolutional networks with self-attention. *arXiv preprint arXiv:2112.03262*, 2021. 2

[51] Yu-Qi Yang, Yu-Xiao Guo, Jian-Yu Xiong, Yang Liu, Hao Pan, Peng-Shuai Wang, Xin Tong, and Baining Guo. Swin3d: A pretrained transformer backbone for 3d indoor scene understanding. *arXiv preprint arXiv:2304.06906*, 2023. 2

[52] yeyuxmf. Auto tooth arrangement. `https://github.com/yeyuxmf/auto_tooth_arrangement`, 2022. 5

[53] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 3