2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

HTLNet: A Segmentation-Free Multi-View Approach for Robust 3D Tooth Landmark Localization

Chentao Wang[1], Jing Du[2], Ran Fan[1] and Fuchang Liu[1(✉)]

[1] School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China
[2] Department of Media and Communication, Kangwon National University, Chuncheon 24341, Republic of Korea
liufuchang9437@163.com

Abstract. Tooth landmark localization plays a pivotal role in digital orthodontics, providing the computational foundation for generating alignment coordinates and guiding precise treatment planning. However, the limited availability of high-quality 3D tooth landmark datasets and the prevalent reliance on segmentation-based methods hinder the accuracy and scalability of current approaches. In this work, we construct a high-quality 3D tooth landmark dataset through manual annotation, specifically designed for training and evaluating tooth landmark localization models in real-world clinical scenarios. To overcome the limitations of existing methods, we propose HTLNet (Heatmap-based Tooth Landmark Localization Network), a novel segmentation-free localization framework based on multi-view 2D heatmap regression. HTLNet eliminates the dependency on prior segmentation and reduces error propagation in the processing pipeline. Experimental results demonstrate that HTLNet outperforms state-of-the-art 3D models, such as PointNet-Reg, in terms of accuracy and robustness, especially under challenging conditions such as missing teeth or misaligned dentition. Our method provides a generalizable, scalable, and efficient solution, making it well-suited for integration into intelligent dental digital systems and advancing the application of computer vision technologies in digital healthcare.

Keywords: Neural networks, Heatmap regression, Multi-view learning, 3D landmark localization, Segmentation-free, Orthodontic applications, Tooth landmark datasets.

1 Introduction

With the rapid development of digital oral technologies, computer-assisted diagnosis and treatment have been widely adopted in modern dentistry[1][2][3][4]. In orthodontics, 3D intraoral scanners are commonly used to obtain detailed tooth morphology, laying the foundation for digital tooth arrangement. To perform precise alignment, it is essential to establish a dental coordinate system based on anatomical landmarks. Accurate localization of these landmarks directly affects treatment quality and efficiency.

Tooth landmark localization is essentially a form of 3D landmark localization, capable of directly processing three-dimensional reconstructed data and providing a more intuitive representation. In recent years, advances in deep learning models have significantly improved the accuracy and robustness of this technology. Wang et al.[5] proposed using a graph convolutional network[6] to construct 3D heatmaps based on Gaussian distances and learn their geometric information for accurate landmark prediction. They also introduced a nearest surface matching technique to optimize the inference process. To address the limitations of traditional graph convolutional networks, such as limited receptive fields and shared transformation matrices, Zhao et al.[7] introduced the Semantic Graph Convolutional Network, enabling end-to-end training of local and global relationships. Zou et al.[8] further enhanced graph convolutional performance through weight and similarity modulation.

However, 3D dental models are structurally more complex than general 3D models. To ensure accuracy, a single 3D tooth model often contains up to 84 landmarks and approximately 70,000 triangular facets. Directly applying the above methods would lead to excessive computational demands and reduced accuracy. Therefore, most methods rely on prior 3D segmentation, such as PointNet-Reg[9]. Yet even after downsampling, models like PointNet[10] still face challenges in handling the data without sacrificing precision. Jiang et al.[11] proposed a segmentation method combining a concavity-aware harmonic field with heuristic feature line extraction, which showed promising results. However, current methods still cannot guarantee ideal segmentation for all teeth, and the segmentation quality significantly affects landmark localization accuracy. Moreover, structural differences among incisors, canines, and molars[12][13] lead to inconsistent results in mixed-type training. Missing teeth further complicate boundary determination, resulting in poor generalization. To improve robustness, Wei et al.[14] introduced a multi-scale latent feature extraction module based on point cloud networks, but its performance remains dependent on segmentation quality.

Dataset limitations further hinder progress. Most publicly available datasets focus on 2D images such as clinical photographs or X-rays[15][16], while 3D datasets like Teeth3DS[17] lack annotated landmarks. Although a recent work by Wang and Lei et al.[18] provides 3D tooth landmarks, it only includes 200 samples, which limits model training and generalization. Thus, there is an urgent need for high-quality 3D tooth landmark datasets.

In response to the limitations of segmentation-dependent methods and the scarcity of annotated 3D dental landmark data, we propose a novel segmentation-free multi-view **H**eatmap-based **T**ooth **L**andmark Localization Network (HTLNet). Our method leverages multi-view rendering to project 3D tooth meshes into multiple 2D views, from which HTLNet predicts landmarks. The optimal landmark positions are then determined through a view matching mechanism guided by consistency and confidence estimation, followed by an inverse mapping step to recover the corresponding 3D coordinates. This pipeline eliminates the reliance on tooth-level segmentation and significantly enhances robustness in challenging clinical scenarios such as tooth crowding, deformities, and missing teeth. We further evaluate HTLNet against commonly used 2D landmark detection networks such as Pose-ResNet[19], Hourglass[20], CPN[21], and ViTPose[22], as well as ConvNeXt[23], a modern convolutional architecture that excels at capturing long-range dependencies and achieves performance comparable to transformer-based models like ViT[24]. To support research and development in this field, we also introduce a new public dataset, LandMark-Teeth3DS, which consists of 752 3D tooth models annotated with detailed landmark coordinates across various tooth types. The key contributions of this work are summarized as follows:

1) We propose HTLNet, a novel segmentation-free 3D tooth landmark localization framework that transforms the 3D detection task into a multi-view 2D heatmap regression problem, thereby reducing computational complexity and improving robustness. To further ensure the accuracy of 3D landmark reconstruction, we introduce an optimal view selection strategy coupled with an error detection mechanism.

2) A new dataset, LandMark-Teeth3DS, containing 752 annotated 3D tooth models, which we plan to make publicly available.

3) Extensive comparative and ablation experiments demonstrating our method's advantages in accuracy, robustness, and generalizability.

2 Method

2.1 Overall Framework

To address the limitations of existing 3D methods and eliminate the reliance on tooth segmentation for landmark localization, this paper proposes a tooth landmark localization method based on multi-view projection, as illustrated in Fig. 1. First, we synthesize multiple 2D images from a 3D tooth model using specified virtual cameras. The generated 2D images are then input into our 2D landmark localization network, HTLNet, for training, enabling it to output the 2D coordinates of the tooth landmarks and its network architecture is illustrated in Fig. 2. Since a single viewpoint may result in occluded landmarks and incomplete localization, we select five different viewpoints to ensure full coverage of all landmarks. The final 2D landmarks are obtained by selecting the predictions from the viewpoint with the lowest matching cost, based on the matching costs of each landmark across the five viewpoints. Finally, the 2D coordinates are mapped back to the 3D model to obtain the 3D coordinates of the tooth landmarks.

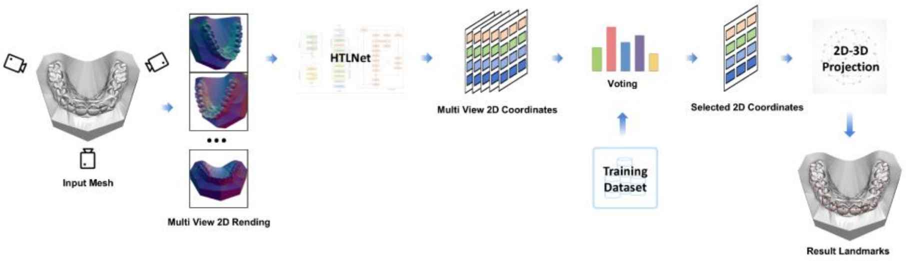


Fig. 1. The pipeline of segmentation-free multi-view heatmap-based tooth landmark localization.

2.2 HTLNet Architecture

Relative to the 3D mesh model, 2D images have fixed sizes and require relatively fewer computations. Therefore, we use Pytorch3D to render the 3D mesh model into 2D images. Considering that a single viewpoint may lead to some landmarks being occluded, we selected five viewpoints that can better cover all the landmarks of the teeth for rendering. Each rendered 2D image from these viewpoints is encoded in RGB, containing the appearance information of the teeth.

The dental image consists of three parts: the teeth, the dental arch, and the background. However, for landmarks localization, only the teeth contain useful information. Therefore, we introduce the Attention Block to focus on the dental region in the image. The Attention Block first extracts features using the DRes, and then combines multi-scale features and attention mechanisms through the EMSA Module[25] to focus on features at different scales, improving the model's performance. It automatically learns the weighted importance of features at different scales, enabling the network to focus more effectively on key regions.

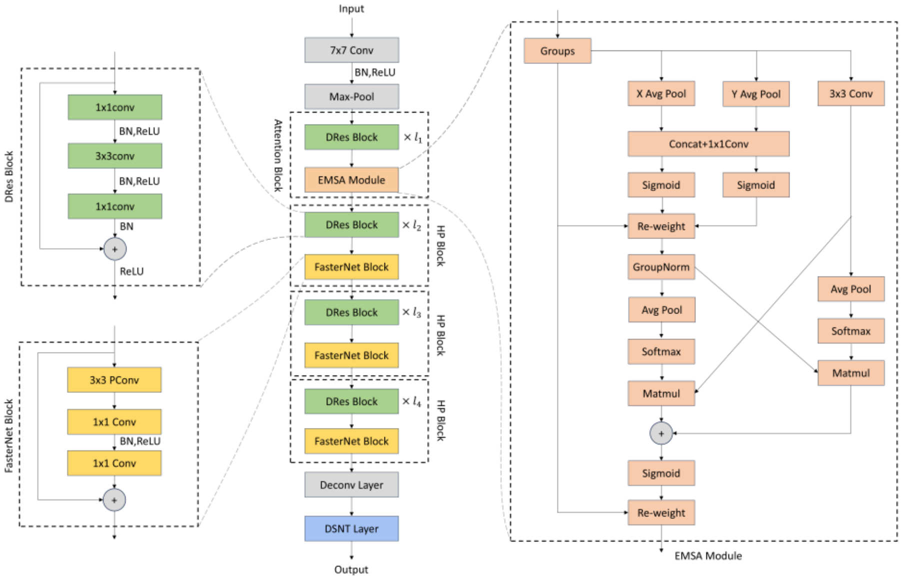


Fig. 2. The structure of HTLNet. The proposed network mainly consists of an Attention block and an HP block. The Attention block is composed of a DRes block and an EMSA module, while the HP block is composed of a DRes block and a FastNet block.

The HP Block is a key component of this network and is composed of two main parts. The first part is the DRes[26] block, which introduces a low-dimensional bottleneck in a specific layer of the network. This compresses the input features via a convolutional layer with a lower output dimension, and then recovers these features through one or more higher-dimensional convolutional layers. This approach reduces the network's computational load while preserving its expressive power. The second part is the FasterNet Block[27], which consists of a 3x3 partial convolution (PConv) and two

1x1 convolutions. PConv takes advantage of redundancy in the feature map by applying convolution operations on only a subset of the channels while leaving some input channels unchanged. During memory access, only the first or last consecutive channels are used as representatives for the entire feature map in the computation. Furthermore, PConv preserves part of the original input features within the feature map, providing richer semantic information for high-level feature fusion in the subsequent layers. Importantly, the input and output feature maps maintain the same number of channels, ensuring that PConv requires fewer floating-point operations and less memory access than standard convolution. In 2D dental images, where occlusion may occur, PConv computes only within the valid regions, skipping over missing areas, which helps avoid interference from missing values in the convolution results.

We also use deconvolution to transform the features into a heatmap of the same size as the input image, facilitating more accurate localization of tooth landmark coordinates.

At the end of the model, we apply the DSNT[28] layer to convert the output 2D heatmap into accurate keypoint targets. Through this weighted summation process, the model not only leverages the heatmap's characteristics but also performs direct coordinate regression for the keypoints, thus enhancing both the efficiency and accuracy of the model.

2.3 Inverse Mapping from 2D to 3D Coordinate Points

In our method, a set of 2D images of a tooth is captured from five different viewpoints. Each 2D image from a viewpoint is input into HTLNet, which then outputs a set of 2D tooth landmark coordinates. However, each viewpoint may suffer from occlusion issues, and sometimes the displacement of 2D coordinates can cause a misalignment in the corresponding 3D coordinates. Therefore, the results obtained from each viewpoint cannot be directly used as they are. The key step is to reasonably select the best viewpoint for each tooth landmark by using prior matching cost from training dataset.

The inverse mapping of the 2D landmarks from the five viewpoints to the final 3D landmarks can essentially be reduced to a maximum matching problem in a bipartite graph, which is solved using a greedy algorithm. The key lies in defining the matching cost between each 2D landmarks and the 3D landmarks. We estimate the matching cost using the following method: First, we use the HTLNet network to predict the 2D landmark coordinates from each of the five 2D images of the teeth in the training set. We then map these 2D coordinates back to the 3D tooth surface to form five sets of 3D coordinates (with known camera parameters). We use the error between these coordinates and the ground truth coordinates as the matching cost, and sort them in ascending order. After statistics are collected for all data in the training set, We can then obtain the matching cost of each landmark with the five viewpoints. For each 2D landmark, we choose the 2D coordinate from the viewpoint with the lowest matching cost as the resulting coordinate. By utilizing this matching cost, we can select the most suitable 2D coordinates in other datasets.If the z-axis of a landmark significantly differs from its neighbors, we treat it as an outlier and replace its coordinates with those from the second most likely viewpoint.

3 Experimental setup and environment

3.1 Dataset

The definition of anatomical landmarks on tooth is very important in dental anatomy and orthodontics. We adopted the same annotation standards as described in reference[17], where each tooth has 6 landmarks defined as shown in Table 1, and the overall tooth landmarks are shown in Fig. 3. Following this study[17], our method does not differentiate between upper and lower jaws, allowing for direct prediction of 84 landmarks for the 14 teeth.

Since there is currently no publicly available 3D tooth landmark localization dataset, we created a dataset called LandMark-Teeth3DS for experimentation and will make it publicly available. This dataset consists of two parts: 3D dental models and tooth landmark 3D coordinates. It covers a variety of dental types, including normal tooth models, missing tooth models, crowded tooth models, and models with missing and crowded tooth, among others. Some examples are shown in Fig. 4, covering these four types of dental models. We display them from five different viewpoints, with the camera perspectives defined in Table 2. The 3D tooth models come from two sources: the first part is from a subset of the dataset in the MICCAI 2022 competition data, namely Teeth3DS[17], and the second part comes from real cases from dental clinics, including 752 upper or lower tooth models in total, of which 693 are normal tooth models and 59 are abnormal tooth models. Among the normal tooth models, 668 are from Teeth3DS and 25 are collected by us. Among the abnormal tooth models, 30 are from Teeth3DS and 29 are collected by us. The annotation of the tooth landmark coordinates is carried out by three participants, and the process, guided by dental professionals, took a total of 20 days to complete. We adopt the same approach as Wang et al.[29], dividing the 693 normal tooth models in the dataset into training, testing, and validation sets with a 7:2:1 ratio. This results in 489 samples in the training set, 143 samples in the testing set and 61 samples in the validation set, and the test set includes 25 tooth model that collected by us. The 59 abnormal tooth models are used to test the stability of our method in dealing with abnormal dental conditions.

Since our experiment involves predicting 2D landmarks based on 2D images, we need to project the 3D tooth models in LandMark-Teeth3DS onto 2D images. This generates a 2D image tooth landmark dataset, we call it LandMark-Teeth2D, which will be used for training and testing our network. For angle selection, considering the visibility of different landmarks and minimizing computational cost, we adjusted the camera's elevation, azimuth, and distance to select five different views, and render the 2D image using Pytorch3D, the specific definitions and explanations provided in Table 2. Here, dist refers to the distance from the camera to the centroid of the mesh model, elev refers to the angle between the vector from the mesh model's centroid to the camera and the X_z plane, and azim refers to the angle between the projection of the vector from the mesh model's centroid to the camera onto the X_z plane and the (0,0,1) vector.

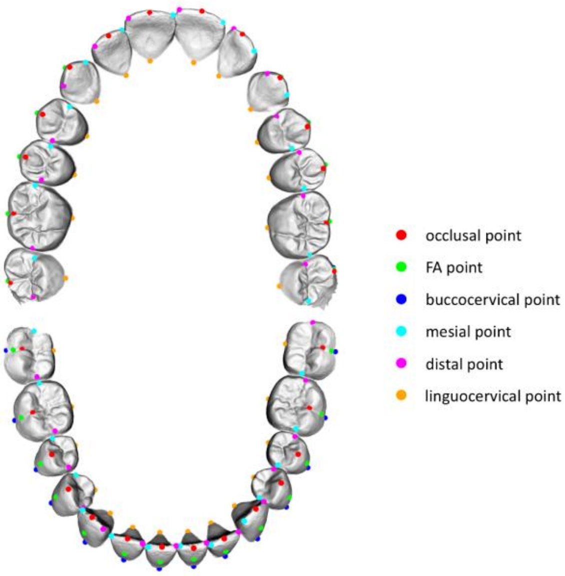


Fig. 3. Schematic diagram of tooth landmarks.

Table 1. Tooth landmark description.

Index	*Name*	*Definition*
1	Occlusal point	The midpoint of the edge of the incisors.
2	FA point	The midpoint of the facial axis of the clinical crown.
3	Buccocervical point	The lowest and most concave point in the buccogingival line.
4	Mesial point	The most mesial point.
5	Distal point	The most distal point.
6	Linguocervical point	The lowest and most concave point in the linguogingival line.

Table 2. Tooth landmark description.

Index	*Name*	*Arguments*
1	Top view	dist=7, elev=0, azim=0
2	Front view	dist=7, elev=-45, azim=0
3	Rear view	dist=7, elev=45, azim=0
4	Right view	dist=6, elev=0, azim=40
5	Left view	dist=6, elev=0, azim=-40

To enhance the LandMark-Teeth2D, we employed data augmentation techniques. Since our training data are derived from 3D mesh models mapped to 2D images, we applied augmentation from both 3D and 2D perspectives. For 3D augmentation, we referenced methods from literature[10][30] for point cloud augmentation, adjusting camera-to-model distances to generate different-sized 2D images. For 2D augmentation, we followed techniques summarized in literature[31], performing vertical and horizontal mirroring of original images. After projecting from 5 viewpoints and applying both forms of augmentation, we obtained a total of 10,395 2D dental images, with 7335 images used for training and 3060 images for testing.

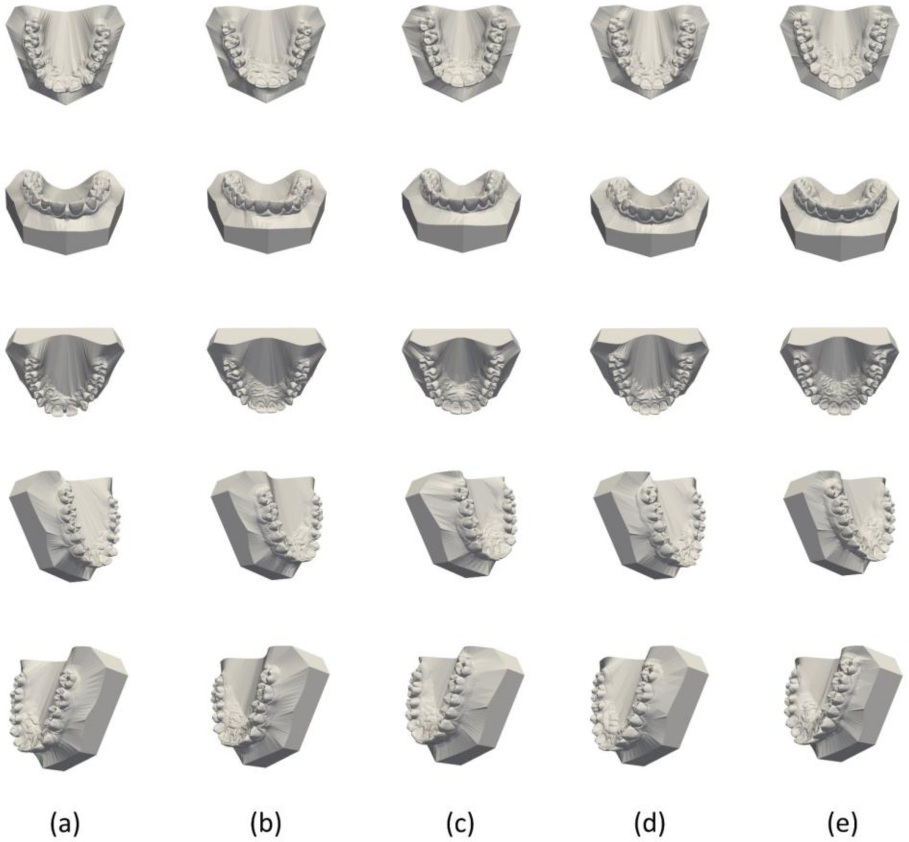


Fig. 4. Part of the LandMark-Teeth3DS dataset examples. The dataset displays five columns, where columns (a) and (b) show instances of normal upper and lower jaw models, column (c) shows instances of models with missing teeth, column (d) shows instances of models with crowded teeth, and column (e) shows instances of models with both missing and crowded teeth.

3.2 Experimental Setup and Evaluation Metrics

The training environment for the proposed dental key point detection method is as follows: 11th Gen Intel® Core™ i7-11700@2.50GHz CPU, 135.1 GB RAM, 4 GeForce RTX 3090 GPUs with 24GB VRAM each. The operating system is Ubuntu 18.04.5 LTS, programming language is Python 3.8, and Pytorch framework version is 2.0. The training details are show in Table 3.

This paper employs two metrics to evaluate experimental results. The first metric is the distance loss, which measures the mean Euclidean distance between the coordinates of predicted and actual tooth landmarks, quantifying the degree of deviation between predicted and actual landmarks. The second metric is the standard deviation, which represents the sample standard deviation of all tested landmark loss values.

Table 3. Training details and GPU memory usage.

Model	*Input Image Resolution*	*Weights*	*Batch Size*	GPU Memory Usage	*Epochs*	*Training* time
PointNet-Reg[9]	NA	5.59MB	65	48G	150	4h
Wei et al. [14]	NA	5.4MB	64	48G	150	8h4m
ConvNeXt[23]	1024^2	827.5MB	8	48G	80	3d1h
ViT[24]	1024^2	167.2MB	128	48G	80	8d2h
ViTHeatmap	1024^2	624.5MB	2	72G	42	12d
TeethHourglass-512	1024^2	109MB	24	48G	109	5d2h
TeethHourglass-1024	1024^2	110MB	3	48G	90	15d
TeethDSNT	1024^2	115.4MB	3	48G	24	1d1h
TeethAttention	1024^2	115.5MB	2	48G	24	1d2h
TeethPConv	1024^2	216MB	2	48G	24	1d4h
HTLNet	1024^2	216MB	2	48G	24	1d5h

4 Experimental results and analysis

4.1 Results of HTLNet

To further validate the superiority of the HTLNet network proposed in this paper, a series of comparative experiments were conducted. All experiments were performed under the same hardware and software, using the same dataset. PointNet-Reg[9] is a method based on 3D heatmap regression and is currently a widely used approach. Wei et al.[14] introduced an improved method by adding preprocessing on top of point cloud networks. In our method, an important process is predicting tooth landmarks on the 2D image, specifically in the structure of HTLNet in Fig. 1. We compared our network with five existing networks. TeethHourglass-512 and TeethHourglass-1024 apply the Hourglass[20] network for heatmap regression, with output resolutions of 512x512 and 1024x1024, respectively, to generate 2D heatmaps. The argmax method is applied to the heatmap to find the 2D tooth landmark coordinates with the highest heat value. Vision Transformer (ViT)[24] is a popular deep learning model for image processing in recent years. We adopted both point coordinate regression and heatmap regression methods to predict landmarks on the 2D image of teeth, corresponding to experiments named ViT and ViTHeatmap. Finally, we also tested the latest ConvNeXt[23] network for point regression to predict keypoints on the 2D image of teeth. The training details are shown in Table 3.

To demonstrate the general applicability of the method proposed in this paper, we prepared two sets of validation data. The first set consists of normal tooth models, sourced from the normal tooth validation set of LandMark-Teeth3DS, with a total of 61 samples. The second set consists of abnormal tooth models, sourced from the abnormal tooth validation set of LandMark-Teeth3DS, with a total of 59 samples. The experimental results are shown in Table 4 and Table 5. From Table 4, it can be seen that our HTLNet has a significant advantage in both average loss and standard deviation, showing the best performance. From Table 5, it is clear that when dealing with abnormal teeth data, our HTLNet achieves significantly lower distance loss and standard deviation compared to other methods, demonstrating a considerable advantage in handling abnormal data as well.

Table 4. Evaluation of normal tooth landmarks prediction.

Model	*Method type*	*Average loss*	Standard deviation
PointNet-Reg[9]	3D heatmap regression	3.38	23.38
Wei et al.[14]	3D heatmap regression	3.23	29.23
ConvNeXt[23]	2D point regression	26.56	215.82
ViT[24]	2D point regression	30.82	148.74
ViTHeatmap	2D heatmap regression	2.29	7.74
TeethHourglass-512	2D heatmap regression	1.51	3.70
TeethHourglass-1024	2D heatmap regression	1.18	2.63
HTLNet	2D Numerical Coordinate Regression	**0.45**	**1.01**

Table 5. Evaluation of abnormal tooth landmarks prediction.

Model	*Method type*	*Average loss*	Standard deviation
PointNet-Reg[9]	3D heatmap regression	22.81	94.15
Wei et al.[14]	3D heatmap regression	22.67	91.56
ConvNeXt[23]	2D point regression	23.96	175.43
ViT[24]	2D point regression	23.41	164.12
ViTHeatmap	2D heatmap regression	3.56	9.71
TeethHourglass-512	2D heatmap regression	2.49	9.11
TeethHourglass-1024	2D heatmap regression	2.77	11.77
HTLNet	2D Numerical Coordinate Regression	**0.84**	**5.98**

In addition, we have also visualized the experimental results for comparison, as shown in Fig. 5 and Fig. 6. Fig. 5 presents the results on the normal tooth model, where it can be seen that HTLNet outperforms other methods and is closest to the real tooth landmarks. Fig. 6 shows the results on the abnormal tooth model, where HTLNet still achieves the best performance. Even in cases of missing teeth, fewer teeth, or dental misalignment, it still yields results close to the real tooth landmarks. Furthermore, the poor tooth segmentation does not affect the performance of our method, demonstrating better robustness. On the other hand, the 3D methods such as PointNet-Reg[9] and the approach by Guangshu Wei et al.[18] show strong dependency on the tooth segmentation results. Poor segmentation significantly affects their detection performance, even leading to detection failure in some cases.

To demonstrate the validity of the data in this paper, we randomly selected a portion of the data from LandMark-Teeth3DS, consisting of 300 samples, including 280 normal tooth models and 20 abnormal tooth models. The above mentioned method was used to predict the tooth landmarks, and the results are shown in Table 6. The table displays the prediction errors for each tooth landmark. From the results, it can be observed that our method maintains relatively stable errors across different tooth types, which, on the other hand, indicates that our data does not show a significant bias towards any specific type of tooth.

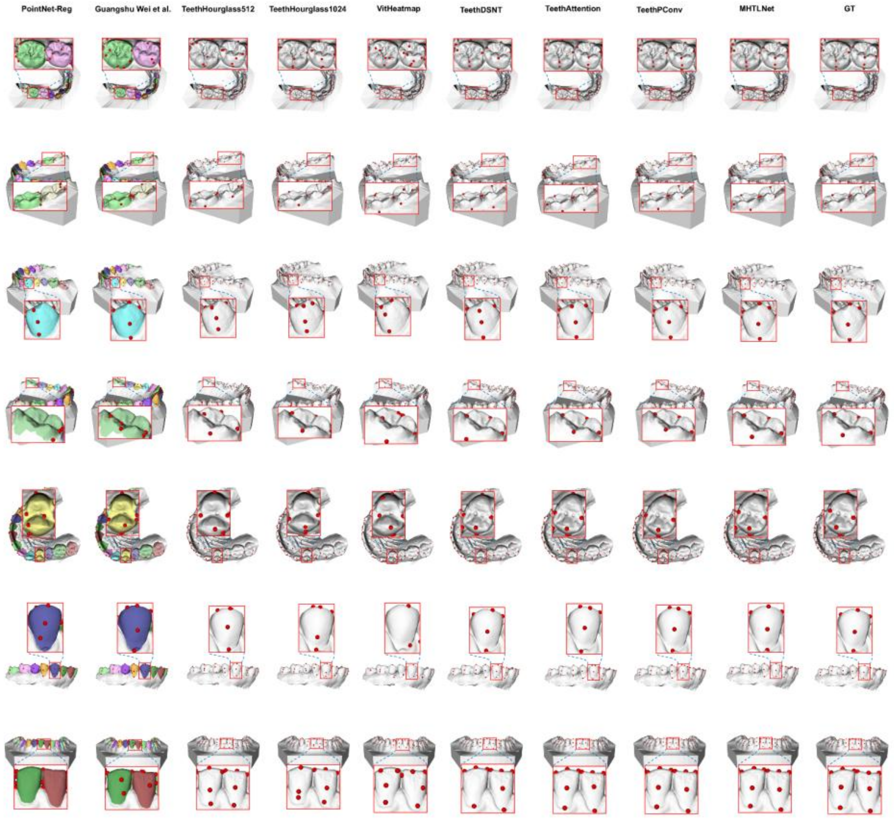


Fig. 5. Comparison of experimental results of different methods under normal dental conditions.

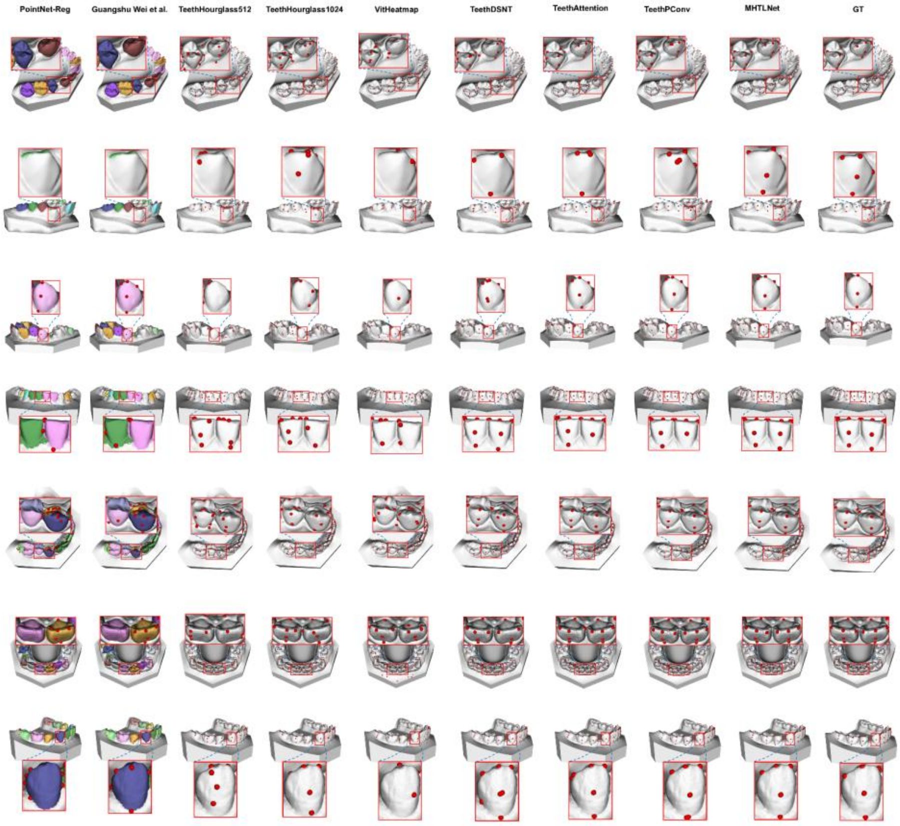


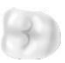
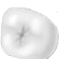
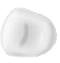
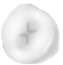
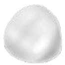
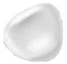
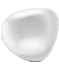
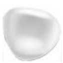
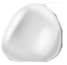
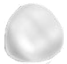
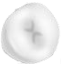
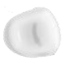
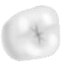
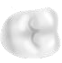
Fig. 6. Comparison of experimental results of different methods under abnormal dental conditions.

4.2 Ablation Study

To verify the performance of each module, detailed ablation studies were conducted using validation set of the LandMark-Teeth3DS dataset. The application of DSNT[28] for tooth landmark localization is used as the baseline, as shown in Experiment 1. And experiments were carried out under the same environment and parameter settings. In each experiment, different modules are added to evaluate the impact of each module on detection performance. We also visualized the results of the ablation study, the visualization results on the normal tooth models are shown in the Fig. 5, and the visualization results on the abnormal tooth models are shown in the Fig. 6.

We conduct four experiments, and the results are presented in Table 7 and Table 8. Experiment 1 demonstrates the performance of the DSNT network in tooth landmark localization task, providing a reference for evaluating the effects of the improvements introduced by each module.

Table 6. Test results of different types of teeth in the LandMark-Teeth3DS dataset.

Experiment															*Mean*
PointNet-Reg[9]	3.31	4.97	4.16	5.78	3.14	4.24	3.13	4.96	4.33	6.31	3.28	4.61	3.09	4.42	4.27
Wei et al.[14]	3.23	3.11	3.72	2.97	3.3	3.19	3.25	3.13	3.39	3.12	3.51	3.03	3.17	3.25	3.24
ConvNeXt[23]	29.5	28.3	31.3	28.1	26.5	25.7	26.3	26.4	29.9	21.8	23.1	24.9	25.8	22.8	26.5
ViT[24]	31.5	28.9	32.4	31.1	32.9	30.5	30.4	28.1	31.4	29.4	32.7	30.3	31.2	30.6	30.8
ViTHeatmap	2.42	1.83	2.84	2.02	2.4	2.24	2.78	1.83	2.37	1.85	2.21	2.21	2.7	2.32	2.21
TeethHourglass-512	1.97	1.32	1.48	1.58	1.56	2.56	1.49	1.29	1.62	1.32	1.93	1.49	1.6	1.75	1.64
TeethHourglass-1024	1.41	1.27	1.37	1.39	1.33	1.37	1.32	1.23	1.33	1.17	1.3	1.23	1.32	1.2	1.3
HTLNet (Ours)	**0.39**	**0.47**	**0.43**	**0.53**	**0.44**	**0.45**	**0.37**	**0.49**	**0.65**	**0.44**	**0.39**	**0.42**	**0.43**	**0.49**	**0.48**

Experiments 2 and 3 demonstrate the individual performance of each module within the DSNT network. Experiment 2 introduces the HP Block with EMSA module to focus attention on the key parts of the input data. Experiment 3 reconstructs the feature extraction module in the backbone network with an HP Block that contains the FasterNet block, which can extract features more effectively. Compared to Experiment 1, HP Block with EMSA module significantly reduce the loss and deviation, especially on abnormal teeth. This suggests that the block focuses attention on the key parts of the input data, which helps subsequent modules perform better feature extraction. Experiment 3 show a significant reduction in loss and deviation, especially on abnormal teeth, indicating that the module composed of HP Blocks has better feature extraction capabilities.

Experiment 4 represents the HTLNet network proposed in this paper. It can be observed that this network integrates the advantages of each module,and further reduce the loss and deviation,especially on abnormal teeth. Compared to Experiment 1, the loss decreased by 0.09 and the deviation decreased by 0.5 for normal teeth, and it is more pronounced in abnormal teeth, the loss decreased by 0.86 and the deviation decreased by 7.23.

Table 7. Result of ablation studies on normal teeth.

Experiment	*Model*	*Average loss*	Standard deviation
1	DSNT	0.54	1.48
2	DSNT+EMSA	0.50	1.39
3	DSNT+FasterNet	0.50	1.01
4	Ours	**0.45**	**0.98**

Table 8. Result of ablation studies on abnormal teeth.

Experiment	*Model*	*Average loss*	Standard deviation
1	DSNT	1.80	13.21
2	DSNT+EMSA	1.39	8.32
3	DSNT+FasterNet	1.38	7.46
4	Ours	**0.94**	**5.98**

4.3 Practical Application – Dental Arch Estimation

In the process of dental orthodontics, the calculation of dental arches is a crucial step used to determine the correct alignment and positioning of teeth, providing a reliable basis for developing effective orthodontic treatment plans. The dental arch refers to the longitudinal axis of the teeth, typically an imaginary line from the top of the crown to the tip of the root. Accurately calculating and adjusting the dental arch is essential for achieving ideal occlusion and aesthetic outcomes.

We used the predictions from HTLNet, which includes 6 predicted landmarks for each tooth, to establish local coordinate systems for teeth. For visualization purposes, we projected the estimated dental arch onto the tooth surface, as shown in Fig. 7. It can be observed that our method achieved good results across different dental samples, accurately estimating the direction of the dental arch based on the predicted 3D landmarks.

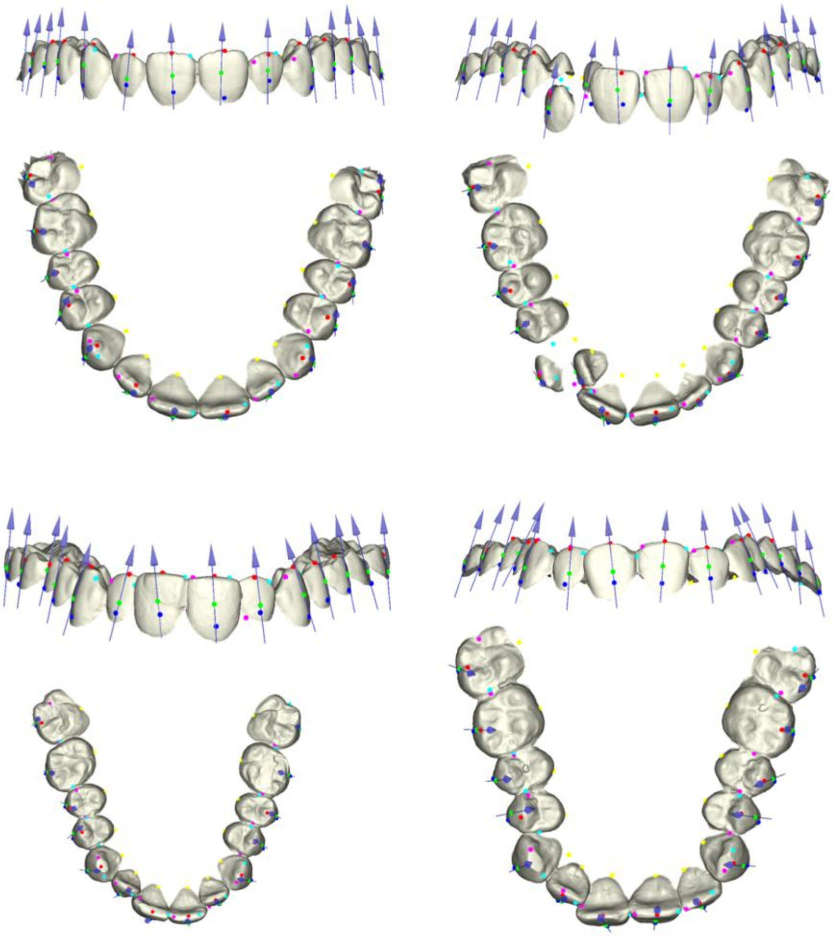


Fig. 7. The visualization of tooth axis estimation results. There are four distinct instances, each with estimated landmarks and axes.

5 Conclusions and Prospective

For the task of tooth landmark localization, the majority of previous studies have used three-dimensional methods for prediction. While these methods have achieved certain effectiveness, their results are often influenced by the precision of the 3D tooth model (e.g., due to downsampling). Moreover, they typically require tooth segmentation as a preprocessing step, where the quality of segmentation directly impacts the final landmark localization accuracy. Tooth segmentation processes are prone to issues such as dental caries, missing teeth, and misalignment, significantly affecting the accuracy of tooth segmentation and thereby hindering ideal tooth landmark localization.

We present a novel segmentation-free multi-view framework for 3D tooth landmark localization. We manually constructed a high-quality 3D dental landmark dataset and developed the HTLNet network architecture, which predicts 2D landmarks from five rendered views of a 3D tooth model. A maximum matching strategy, combined with an error detection mechanism, is employed to select the optimal viewpoint for each landmark. The predicted 2D coordinates are then accurately mapped back to 3D space. Through extensive experiments, including comparisons with state-of-the-art 2D and 3D methods, as well as ablation studies, we demonstrate that our approach significantly improves localization accuracy and exhibits strong robustness against dental abnormalities such as missing and misaligned teeth.

However, the localization accuracy of our method is also affected by image resolution; lower resolutions can introduce certain deviations. Higher image resolutions require larger storage space, and correspondingly increase GPU requirements and training time. The datasets used in this paper were annotated by ourselves, and due to limitations in human resources and time, the dataset size is currently relatively small. Future work could consider expanding the dataset further. Our network can only perform end-to-end processing for a single view, which has certain limitations. In future work, we will continue to research and make further improvements to the network to enable end-to-end processing with multiple view selections.

References

[1] Im, J., Lee, S., Kim, H.: Comparison of Virtual and Manual Tooth Setups with Digital and Plaster Models in Extraction Cases. Amer. J. Orthod. Dentofacial Orthop. 145(4), 434–442 (2014)

[2] Lim, S.-W., Kim, J., Park, H.: Can We Estimate Root Axis Using a 3-Dimensional Tooth Model via Lingual-Surface Intraoral Scanning? Amer. J. Orthod. Dentofacial Orthop. 158(5), e99–e109 (2020)

[3] Li, J., Zhang, M., Liu, Y.: A Fine-Grained Orthodontics Segmentation Model for 3D Intraoral Scan Data. Comput. Biol. Med. 168, 107821 (2024)

[4] Fontana, M., Rossi, F., Bianchi, G.: Correlation Between Mesio-Distal Angulation and Bucco-Lingual Inclination of First and Second Maxillary Premolars Evaluated with Panoramic Radiography and Cone-Beam Computed Tomography. Appl. Sci. 11(5), 2374 (2021)

[5] Wang, Y., Zhang, L., Xu, T.: Learning to Detect 3D Facial Landmarks via Heatmap Regression with Graph Convolutional Network. In: Proc. AAAI Conf. Artif. Intell. 36, 2595–2603 (2022)

[6] Kipf, T. N., Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks. arXiv preprint arXiv:1609.02907 (2016)

[7] Zhao, L., Peng, X., Tian, Y.: Semantic Graph Convolutional Networks for 3D Human Pose Regression. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 3425–3435 (2019)

[8] Zou, Z., Tang, W.: Modulated Graph Convolutional Network for 3D Human Pose Estimation. In: Proc. IEEE/CVF Int. Conf. Comput. Vis., 11477–11487 (2021)

[9] Wu, T.-H., Lin, Y.-C., Huang, T.-K.: Two-Stage Mesh Deep Learning for Automated Tooth Segmentation and Landmark Localization on 3D Intraoral Scans. IEEE Trans. Med. Imag. 41(11), 3158–3166 (2022)

[10] Qi, C. R., Su, H., Mo, K., Guibas, L. J.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 652–660 (2017)

[11] Jiang, X., Wang, Q., Liu, B.: C2F-3DToothSeg: Coarse-to-Fine 3D Tooth Segmentation via Intuitive Single Clicks. Comput. Graphics 102, 601–609 (2022)

[12] Bhargavi, A., Reddy, K., Rao, V.: Comparative Tooth Anatomy—A Review. Int. J. Dental Sci. Res. 1(1), 34–37 (2013)

[13] Black, G. V.: Descriptive Anatomy of the Human Teeth (1897)

[14] Wei, G., Zhang, R., Chen, Y.: Dense Representative Tooth Landmark/Axis Detection Network on 3D Model. Comput. Aided Geom. Des. 94, 102077 (2022)

[15] Chaudhary, S. D., Singh, R., Kumar, N.: Teeth or Dental Image Dataset. Mendeley Data 1 (2024). https://doi.org/10.17632/6zsnhrds9t.1

[16] Loop, H. I. T.: Teeth Segmentation on Dental X-Ray Images (2023). [Online]. Available: https://www.kaggle.com/dsv/5884500

[17] Ben-Hamadou, A., Ahmed, M., Ali, T.: 3DTeethSeg’22: 3D Teeth Scan Segmentation and Labeling Challenge. arXiv preprint arXiv:2305.18277 (2023)

[18] Wang, S., Li, F., Zhou, Q.: A 3D Dental Model Dataset with Pre/Post-Orthodontic Treatment for Automatic Tooth Alignment. Sci. Data 11(1), 1277 (2024)

[19] Xiao, B., Wu, H., Wei, Y.: Simple Baselines for Human Pose Estimation and Tracking. In: Proc. Eur. Conf. Comput. Vis. (ECCV), 466–481 (2018)

[20] Xu, T., Takano, W.: Graph Stacked Hourglass Networks for 3D Human Pose Estimation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 16105–16114 (2021)

[21] Wu, L., Feng, Y., Lu, J.: CPN: Complementary Proposal Network for Unconstrained Text Detection. In: Proc. AAAI Conf. Artif. Intell. 38, 6057–6065 (2024)

[22] Xu, Y., Zhang, Y., Liu, X.: ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. Adv. Neural Inf. Process. Syst. 35, 38571–38584 (2022)

[23] Liu, Z., Lin, Y., Hu, H.: A ConvNet for the 2020s. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 11976–11986 (2022)

[24] Dosovitskiy, A., Beyer, L., Kolesnikov, A.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929 (2020)

[25] Ouyang, D., Jin, X., Wang, L.: Efficient Multi-Scale Attention Module with Cross-Spatial Learning. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 1–5 (2023)

[26] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 770–778 (2016)

[27] Chen, J., Zhu, Y., Zhang, T.: Run, Don’t Walk: Chasing Higher FLOPs for Faster Neural Networks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 12021–12031 (2023)

[28] Nibali, A., He, Z., Wollkopf, M.: Numerical Coordinate Regression with Convolutional Neural Networks. arXiv preprint arXiv:1801.07372 (2018)

[29] Wang, X., Peng, Y., Lu, L.: ChestX-ray8: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2097–2106 (2017)

[30] Wang, Y., Sun, Y., Liu, Z.: Dynamic Graph CNN for Learning on Point Clouds. ACM Trans. Graph. 38(5), 1–12 (2019)

[31] Shorten, C., Khoshgoftaar, T. M.: A Survey on Image Data Augmentation for Deep Learning. J. Big Data 6(1), 1–48 (2019)