This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3247459

ImTooth: Neural Implicit Tooth for Dental Augmented Reality

Hai Li[*], Hongjia Zhai[*], Xingrui Yang, Zhirong Wu, Yihao Zheng, Haofan Wang, Jianchao Wu[†], Hujun Bao, Guofeng Zhang[†]

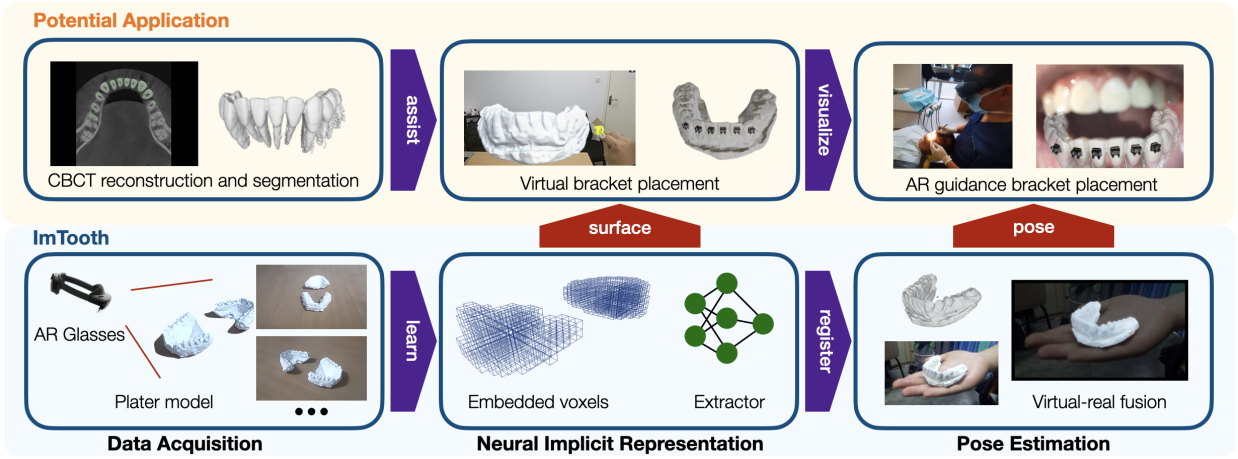

Fig. 1. Overview of ImTooth system and its potential application in dental augmented reality. Our system provides three major functions, i.e. data acquisition, neural implicit reconstruction and pose estimation. We also show the potential application of our system for tooth bracket placement navigation in orthodontics.

Abstract— The combination of augmented reality (AR) and medicine is an important trend in current research. The powerful display and interaction capabilities of the AR system can assist doctors to perform more complex operations. Since the tooth itself is an exposed rigid body structure, dental AR is a relatively hot research direction with application potential. However, none of the existing dental AR solutions are designed for wearable AR devices such as AR glasses. At the same time, these methods rely on high-precision scanning equipment or auxiliary positioning markers, which greatly increases the operational complexity and cost of clinical AR. In this work, we propose a simple and accurate neural-implicit model-driven dental AR system, named ImTooth, and adapted for AR glasses. Based on the modeling capabilities and differentiable optimization properties of state-of-the-art neural implicit representations, our system fuses reconstruction and registration in a single network, greatly simplifying the existing dental AR solutions and enabling reconstruction, registration, and interaction. Specifically, our method learns a scale-preserving voxel-based neural implicit model from multi-view images captured from a textureless plaster model of the tooth. Apart from color and surface, we also learn the consistent edge feature inside our representation. By leveraging the depth and edge information, our system can register the model to real images without additional training. In practice, our system uses a single Microsoft HoloLens 2 as the only sensor and display device. Experiments show that our method can reconstruct high-precision models and accomplish accurate registration. It is also robust to weak, repeating and inconsistent textures. We also show that our system can be easily integrated into dental diagnostic and therapeutic procedures, such as bracket placement guidance.

Index Terms—Artificial intelligence, Neural implicit representation, Dental Mixed / Augmented reality

1 Introduction

In recent years, augmented reality (AR) technology has made great progress and is gradually integrated into multiple application fields. Wearable AR devices, especially AR glasses, have greatly improved existing visualization methods and laid the foundation for a new generation of interactive displays.

For the medical field, AR technology is particularly important. Whether it is preoperative planning or intraoperative navigation, AR's powerful virtual-real integration and interaction ability can effectively assist doctors in decision-making and completing accurate surgery [34, 42, 59]. The current medical AR is still in the early stage. Due to the complex structure of human organs and strict requirements for accuracy, further research and development are required for both algorithms and hardware. Among all medical disciplines, dental AR is one of the most promising research directions [21, 30, 60]. The obvious structural features and rigid body properties of the teeth make them more suitable for reconstruction and localization than other organs, which is also the basis for building practical AR applications. However, the existing dental AR solutions [48, 51, 56, 57] still need to rely on high-precision sensors or external positioning markers, which undoubt-

- *Hai Li, Hongjia Zhai, Hujun Bao, and Guofeng Zhang are with the State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China. E-mails: {garyli, zhj1999, baohujun, zhangguofeng}@zju.edu.cn*
- *Xingrui Yang is with the High-speed Aerodynamics Institute, China Aerodynamics Research and Development Center, China. E-mails: xingruiy@gmail.com*
- *Zhirong Wu, Yihao Zheng and Haofan Wang are with the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. E-mails: {zhirongwu, zhengyihao, wanghaofan}@zju.edu.cn*
- *Jianchao Wu is with the Department of Stomatology, Hangzhou Mingzhou Hospital, Hangzhou, China, and Taizhou Stomatology Hospital, Taizhou, China. E-mail: wujianchao555@163.com*

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

[*]Equal contribution

[†]Guofeng Zhang and Jianchao Wu are corresponding authors.

edly increases the cost and operation difficulty of AR. Moreover, these solutions cannot be applied to AR glasses, which weakens their application scope. In this work, we mainly focus on designing a lightweight dental AR system that is suitable for AR glasses and support basic functions required by dental AR.

In common dental practices, such as orthodontics, a complete 3D reconstruction of the patient's teeth is a prerequisite before diagnosis and treatment. This step is usually performed with conventional reconstruction methods from high-accuracy intraoral structured-light sensors, which makes the process considerably expensive. In comparison, plaster casts of teeth are more versatile and inexpensive, but traditional reconstruction methods are often impractical due to their indistinct texture. The recent emergence of neural implicit representation [35, 36, 58, 62] provides a solution to handle such objects. By considering per-pixel color information, the Multi-Layer Perceptron (MLP) based model can easily capture the consistency across all views and learn the geometry and appearance information implicitly. Not only that, but recent works have also shown the pose optimization ability [29] benefits from the fully differentiable architecture of neural implicit representation. This inspired us to think about AR solutions in dentistry from a new perspective.

We design an integrated dental AR system that supports reconstruction, registration, and interaction based on AR glasses (Microsoft HoloLens 2) without additional sensors. We call this system ImTooth, for it leverages the most state-of-the-art neural implicit representation as the core component. As shown in Figure 1, our system uses a single AR Glasses as the only sensor and display device, which enhanced portability and versatility. Different from other methods, we use the plaster teeth models as our reconstruction target and learn a voxels-based neural implicit representation [27] which is editable and flexible. The learned representation can be directly used for pose estimation without extra training.

To make our system easier to operate, we propose a scale-preserving reconstruction algorithm, which takes raw data captured from HoloLens 2 and reconstructs it in real size without the complicated pre-processing and post-processing procedures used in most of the existing methods. To compensate for the domain gap from the plaster model to the actual oral cavity, we also incorporate edge information into the neural implicit representation during reconstruction. We demonstrate the potential application of our AR system with a navigation application for tooth bracket placement.

The major contributions of our proposed approach are summarized as follows:

- We propose a lightweight dental AR system based on neural implicit representation that supports AR glasses-based reconstruction, registration, and interaction. It provides a low-cost and high-precision solution for AR applications in dentistry.
- We propose a scale-preserving implicit reconstruction method alone with consistent edge representation, which supports the accurate reconstruction and registration.
- We propose a method that implicitly learns consistent edges and enables direct pose registration via edge alignment without additional networks.
- We build a potential AR application for bracket placement guidance based on our AR system.

2 Related Works

In this section, we review the works most relevant to the proposed method, including neural implicit surface reconstruction, visual localization, and dental augmented reality.

2.1 Neural Implicit Surface Reconstruction

Neural implicit representations of 3D surfaces have attracted a lot of attention in the field of 3D reconstruction. Compared to traditional method [47], the geometry information of the scene is represented by a neural network that outputs the signed distance field or occupancy field, which can be used to generate consistent novel views. Neural implicit approaches can reconstruct the scene with a set of posed images, which don't need 3D supervision information. NeRF [35] is the representative work that introduces the use of volumetric rendering and neural networks to represent the spatial density and appearance of the 3D scene and achieves impressive results for novel view synthesis. However, NeRF-based methods [11, 32, 35, 64] can't obtain the high-quality 3D geometrical structure of the scene. To solve this issue, some works are proposed to regress the signed distance value and use it in the volume rendering equation. NeuS [58] proposes to replace the density with neural SDF representation which can handle the self-occlusion situation and build a bridge between the SDF and volume rendering. Based on [58], some works are developed to utilize the additional geometry information to improve the reconstructed neural implicit surfaces, such as depth [5], and normal [65]. These works leverage the capacity of full-connected layers to encode entire scenes with a learned mapping function.

However, a single MLP has limited representation ability and cannot scale well to large scenes, and the coordinate input can be unstable on large scales. Therefore, most of the works pre-scale the potential scene to a smaller area, usually a unit sphere or cube. To reconstruct high-quality results of the large-scale scene, Vox-Surf [27] adopts a sparse voxel structure to divide the spatial regions and store the geometry features in the nodes of the voxel. This way can save computational resources by only reconstructing occupied voxels, which can also be subdivided to recover finer details in large-scale scenes. In this work, we move one step further toward practicality. We consider incorporating joint camera pose and scene optimization into a voxel-based neural implicit representation, thereby alleviating complex pre-processing.

2.2 Visual Localization

Visual localization is the fundamental step for many augmented reality applications, which aims to estimate the precise position and orientation of the query image with respect to a pre-built 3D map. Generally, the visual localization approaches can be classified into pose regression, coordinate regression, and feature-based methods. Pose regression methods [9, 24, 25] directly regress the pose from the extracted feature of a single image, however, which is not competitive in terms of localization accuracy. The coordinate regression methods [7, 8, 49] aim to estimate the 3d coordinates of the pixel in the camera view and apply Perspective-n-Point (PnP) to solve the pose of the query image. Different from pose regression methods, the coordinate-based methods use PnP to replace the pose regression process, which is more interpretable and achieves better performance. Finally, the feature-based visual localization methods [18, 44, 52] usually consist of two steps, global image retrieval [2, 3, 13, 16, 20, 22, 38, 39], and local feature matching [12, 23, 41, 50, 53]. This kind of method applies a hierarchical localization way to support large-scale scene localization and has the best generalization ability compared to pose and coordinate regression methods. Among the above three kinds of approaches, the feature-based approach is the most practical and has the best generalization performance.

For the global image retrieval step, feature-based visual localization methods usually use retrieval-based methods to provide an approximate pose of a given query image. NetVLAD [2] proposes to aggregate patch-level features through a trainable VLAD layer. Other retrieval-based methods propose new approaches to handle weakly supervised information [17] and patch-level features [20]. For the local feature matching step, keypoint detection and description are first performed on the query image to obtain discriminative and repeatable keypoints. The hand-crafted feature [31] and deep learning-based feature [14, 15] can be used in this process. After obtaining the descriptor of the keypoints, feature matching is performed between the query image and the database to obtain 2D-3D matching pairs. Finally, the pose of the query image can be obtained via the RANSAC and PnP process. HLoc [44] is currently the most mainstream approach for feature-based localization methods. It utilizes the best feature point descriptor, SuperPoint [14], and a feature matching method, SuperGlue [45], to achieve good generalization results in different scenarios.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3247459

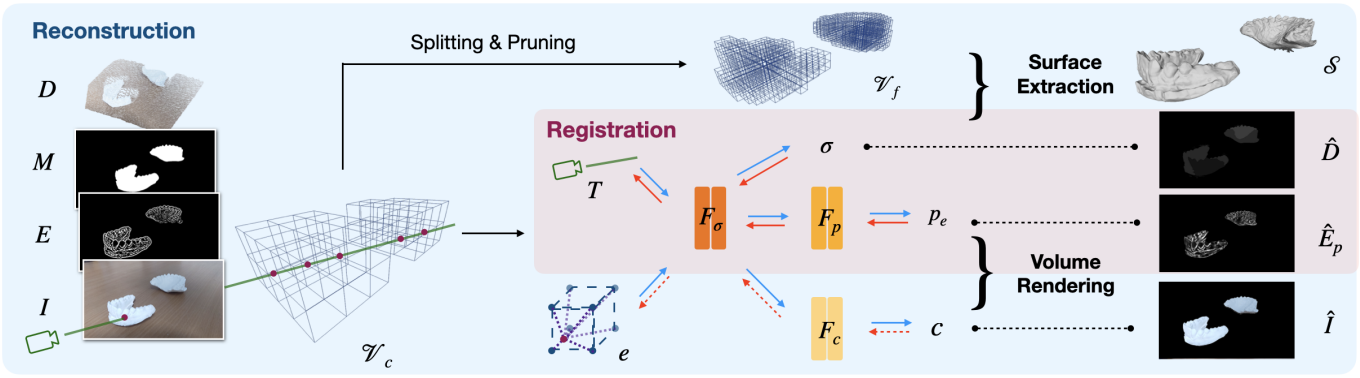

Fig. 2. Demonstration of the shared core module of ImTooth. The reconstruction module learns the embedded voxels and the extractor of the implicit scene and the registration module leverages the learned scene for direct pose optimization via backpropagation to the camera pose. The blue arrow is the forward propagation, the red is the backpropagation and the dotted line indicates the reconstruction phase only.

2.3 Dental Augmented Reality

For dentistry, augmented reality (AR) technology can superimpose the virtual tooth model reconstructed from the oral image on the corresponding oral position of the patient, so as to enhance the doctor's visual system and improve the surgical ability through some information interaction and guidance. Besides, some dental education applications [40, 55] are proposed for the training process of the doctors, which guide the doctor to correct the posture with the dental extraction simulator. In the orthodontic filed [1], AR technology was also used for guiding the bracket placement in orthodontics correction according to the edge information of the teeth. Among those AR applications, registration technology plays an extremely important role, which decides the accuracy of the orientation and position of the virtual model placed in the real world.

Some approaches use manual marker points or artificially placed objects to achieve good registration performance. The camera can identify the markers and object points in the image and use this information to track and register the virtual models. Some common image markers are, ARToolKit, ARTag, Visual Code, and AR Studio. However, those marker-based methods need to artificially place some objects in the real world, which may be impossible or harmful during medical surgery. Another kind of approach is natural feature-based tracking registration technology. Natural feature-based methods are proposed to extract the reference points from the real world and consist of feature point extraction and matching algorithms. The registration speed and accuracy correlate to the key points detection and description algorithms. The existing algorithms that are commonly used for keypoint detection are SIFT [31], SURF [6], and FAST [54] and description algorithms are ORB [43], BRIEF [10], and BRISK [26]. In recent years, methods based on deep learning have gradually emerged, such as SuperPoint [14] and D2Net [15] etc. These methods have strong generalization ability and have been widely used in academia and industry. However, it is sensitive to scene and texture changes, and such methods based on indirect features are usually unable to further optimize the matching position, which will affect the accuracy of the pose. Therefore, we propose to directly optimize the pose via pixel-level feature differences for higher accuracy.

3 Methods

The proposed ImTooth system consists of two modules, an implicit reconstruction module, and a pose registration module. They share a common core module as shown in Figure 2. This module learns multiple information about the scene implicitly during reconstruction and can be directly used to optimize the pose during registration.

To reconstruct the detail of the teeth model, we adopt the idea of voxel-based neural implicit representation proposed in Vox-Surf [27], which is flexible and lightweight compared to other methods [36, 58, 62, 63]. We first briefly review several basic concepts.

Voxel-based Neural Implicit Representation. In voxel-based neural implicit representation, the scene is divided into a set of axis-aligned voxels $\{\mathscr{V}_i\}$. For each voxel, $\mathscr{V}_i$, its eight corners store independent embedding vectors, $\mathbf{e}$, which encodes the geometrical and appearance information of the scene and can be optimized during the training process. For point $p \in \mathbb{R}^3$ within one voxel $\mathscr{V}_i$, its embedded feature vector can be obtained via the retrieval function Γ, which tri-linear interpolate the embedding vector from eight corners according to point coordinate. Vox-Surf adopts two MLP-based networks F_σ and F_c to extract view-dependent radiance, c, and signed distance value, σ.

$$F_\sigma, F_c : (\mathbf{e}, d) \rightarrow (c, \sigma), \tag{1}$$

where d and $\mathbf{e} = \Gamma(p)$ are the 2D view direction and feature vector at 3D position p, respectively. The SDF value σ represents the distance from point p to the closest point on the underlying surface $\mathscr{S}$. Thus, the surface can be easily extracted by the following equation:

$$\mathscr{S} = \{\, p \in \mathbb{R}^3 \mid F_\sigma(\Gamma(p)) = 0 \,\}. \tag{2}$$

Volume Rendering. Rendering with neural implicit representation is usually based on volume rendering [33, 35], which is the accumulation of the color c of the N_p sample points along the ray r with respect to the density α (Equation 3). Each ray is emitted from the center of the camera o in direction d and passes through a pixel on the image. The points on the ray are denoted as $r(t) = o + t \cdot d$, where t is the depth for each sampled point.

$$C(r) = \sum_{i=1}^{N_p} T(t_i)\alpha_i(t_i)c_i, \tag{3}$$

$$T(t_i) = \prod_{j=1}^{i-1}(1 - \alpha(t_j)). \tag{4}$$

To convert the SDF value σ to density α, we leveraged the density conversion function Equation 5 proposed in [58].

$$\alpha(t_i) = ReLU\left(\frac{\Phi_s(\sigma_i) - \Phi_s(\sigma_{i+1}))}{\Phi_s(\sigma_i)}\right), \tag{5}$$

where $ReLU(\cdot)$ is the Rectified Linear Unit, and $\Phi_s(\cdot)$ is the Sigmoid function.

Limitation in Existing Reconstruction. Most existing neural implicit reconstruction methods [36, 58, 62, 63] requires a complicated preprocessing procedure to stabilize the training process. A common procedure is to pre-reconstruct the scene with structure-from-motion algorithms, such as COLMAP [46] to get the camera poses. These poses are further transformed to make sure the latent surface is inside a

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3247459

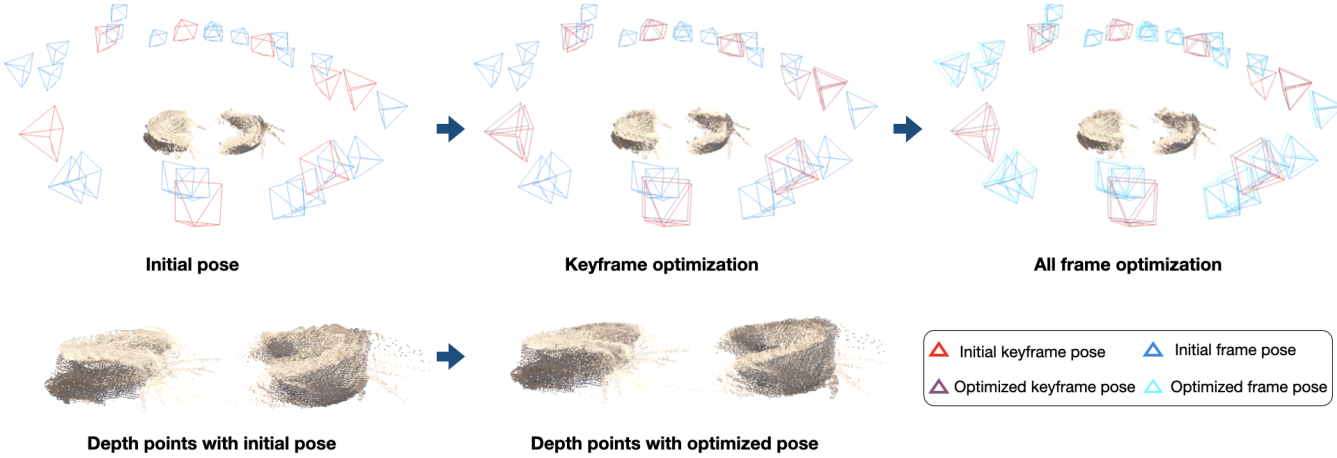

Fig. 3. Frame-wise pose optimization. We propose a coarse-to-fine reconstruction manner to optimize the raw pose from HoloLens 2 in two steps. The top line shows the camera pose optimization result and the bottom line shows the depth point visualization of multiple frames during optimization.

unit sphere. Although Vox-Surf [27] relieves the latter step, an accurate 6 DoF pose is still necessary. However, this pre-processing method has two problems. Firstly, existing SfM algorithms mainly rely on feature matching, which is vulnerable in weak or repeated texture scenes, such as in our case. Secondly, up-to-scale reconstruction cannot match the real scale.

3.1 Implicit Reconstruction

Toward practical usage, our system is designed to replace the complicated pre-process procedure with a noise sensor input, and reconstruct the tooth model in an end-to-end manner. Apart from multi-view images, our system also takes rough poses and sparse depth as input to assist the scale estimation. The rough poses can be obtained through inertial sensors (IMU) or odometry of mobile devices. However, these poses usually have a drift problem, resulting in inconsistent scales, thus we also leverage the low-resolution depth map for initialization. As shown in the bottom left of Figure 3, the depth points of multiple input poses and depth are noisy and sparse, which is far from the accuracy required for dental model reconstruction. To obtain the scale-consistent and detail-preserved reconstruction results, we design a coarse-to-fine reconstruction manner to optimize the camera pose T and neural implicit representation, simultaneously.

Coarse-Level Reconstruction. In this part, we aim to quickly construct the potential surface with real scale and correct the noisy poses provided by the odometry. The coarse reconstruction part consists of the following steps.

1) Real scale alignment. We first extract the predicted potential tooth mask M for each input image, this step is done by the off-the-shelf tool, rembg*, which is developed based on [37]. Based on the depth inside the predicted mask, we can construct the initial coarse voxels $\mathcal{V}_c$ with large voxel size. The scale of the voxels is aligned with the real world.

2) Key-frames selection. To maintain a consistent scale during the reconstruction process, we select several key-frames according to certain overlap thresholds. Specifically, we take the first frame to make up the world coordinate and fix its pose during reconstruction. To establish enough co-visibility with a small number of images, we borrow the concept of key-frames from SLAM methods [61], and select the key-frames based on the mutual visibility of voxels.

3) Frame-wise pose optimization. For l-th iteration, if $l < l_{key}$, we only optimize the rays sampled from key-frames and recover the coarse scene with consistent scale as shown in top-middle of Figure 3. After l_{key} iterations of key-frame optimization, we optimize the rays from all frames but fix the pose of key-frames for another l_{all} iterations to rectify the pose for all frames (Figure 3 top-right). As shown in the bottom-right of Figure 3, after the coarse-level reconstruction, the fused depth map is more aggregated, and some structural details have emerged.

4) Rendering losses and optimization. To optimize the frame-wise pose and neural network, we use volume rendering to obtain the rendering results, then minimize the difference between the rendering results and ground truth observations.

Here, we show the training losses used in our approach during the reconstruction and registration modules. As shown in the Figure 2, we take the interpolated embedding $\mathbf{e}$ and posed view direction d as input, the per point SDF value σ and color c on each ray are computed by F_σ and F_c. The accumulated color $C(r)$ for each ray r is obtained through Equation 3. So, the photometric loss between the rendered color, $C(r)$, and input color, $\hat{C}(r)$, from input image I is computed by Equation 6:

$$\mathcal{L}_{color} = \sum_r \|\hat{C}(r) - C(r)\|_1. \tag{6}$$

However, photometric information is insufficient to represent the characteristics of teeth. Therefore, we consider finding a more salient feature that is suitable for our teeth model matching. As shown in Figure 4, the edge pattern can be easily recognized regardless of the domain, therefore we proposed to integrate the edge feature inside our ImTooth representation during reconstruction.

Fig. 4. Example images of teeth in different domains and their corresponding edge maps.

We extract the binary edge map E from image I and add a new edge prediction branch F_p which predicts the probability of point p appearing on the edge. The edge probability $P_e(r)$ for each ray r is also accumulated with Equation 3 and trained by a binary cross entropy edge loss with ground truth $\hat{P}_e(r) \in E$ according to Equation 7:

$$\mathcal{L}_{edge} = \sum_r BCE(\hat{P}_e(r), P_e(r)). \tag{7}$$

*https://github.com/danielgatis/rembg

Apart from the appearance information, we also consider the geometrical information to supervise the reconstruction process. To constrain the scale with noise input depth, we leverage the full-depth supervision, which splits the points $p(t)$ along the ray r into three intervals according to depth. For the points in front of the observed depth $t < \hat{t} - \delta t$, we consider these points to be outside the surface. The depth loss for outside points is formulated as:

$$\mathscr{L}_{outside} = \sum_{t} \|1 - \Phi(-s \cdot \sigma_{p(t)})\|_2, \tag{8}$$

where s is the scale factor to control the truncated margin of SDF value, δt is a noise-tolerated depth range, and $\hat{t}$ is the observed depth value from the depth sensor of HoloLens 2.

For the points behind the observed depth $t > \hat{t} + \delta t$, we consider them to be inside the surface.

$$\mathscr{L}_{inside} = \sum_{t} \|\Phi(-s \cdot \sigma_{p(t)})\|_2. \tag{9}$$

For the points between the near range $\hat{t} - \delta t \leq t \leq \hat{t} + \delta t$, we directly constrain their SDF values to be 0.

$$\mathscr{L}_{near} = \sum_{t} \|\sigma_{p(t)}\|_2. \tag{10}$$

By combining the above three losses, the total depth loss is defined as:

$$\mathscr{L}_{full\_depth} = \mathscr{L}_{outside} + \mathscr{L}_{near} + \mathscr{L}_{inside}. \tag{11}$$

Additionally, to constrain the regulated field, we also leverage the eikonal term [19].

$$\mathscr{L}_{eikonal} = \sum_{t} (\|\nabla \sigma_{p(t)}\|_2 - 1)^2. \tag{12}$$

Finally, the total loss used for optimization at coarse-level reconstruction is defined as follows:

$$\mathscr{L}_{coarse} = \mathscr{L}_{color} + \mathscr{L}_{full\_depth} + \mathscr{L}_{edge} + \mathscr{L}_{eikonal}. \tag{13}$$

Fine-Level Reconstruction. After the coarse-level reconstruction, we already get the coarse structure of the latent tooth model and exact scaled camera poses. So, in this part, we aim to model the detailed structure of the teeth model. The fine-level reconstruction part consists of the following steps.

1) Coarse voxel splitting. Representing teeth models with a set of coarse voxels prolongs the reconstruction process, and will lead to overly-smoothed surface reconstruction due to the sparsely sampled points. To reconstruct finer details, we periodically divide each existing voxel into eight sub-voxels and assign new embedded features via interpolation. We use the embedding retrieval function $\Gamma(\cdot)$ to compute the initial embeddings for newly generated voxel vertices. These embeddings will be optimized together with the neural network.

2) Redundant voxel pruning. To preserve the sparse structure of our method, when splitting a coarse-level voxel into eight fine-level voxels, we need to remove some redundant voxels which do not contain any surface. Since each voxel contains a continuous signed distance field, we can prune the redundant voxels according to the SDF values. We first uniformly sample 3D points inside each voxel, and the SDF values of these points are obtained from the geometry extractor F_σ. We use a predefined distance threshold, τ, to decide whether to retain or prune the voxel. If the SDF values of all sampled 3D points in the voxel satisfying $|F_\sigma(\Gamma(p))| < \tau$, we will prune the corresponding voxel from $\mathscr{V}$.

3) Fine tuning. During the fine-level reconstruction, we fine-tune our model with the above-mentioned losses except $\mathscr{L}_{full\_depth}$. Since the noise depth supervision could only learn a coarse geometrical structure and is harmful in detail structure reconstruction. So, the total loss used for fine-level reconstruction is shown in the following:

$$\mathscr{L}_{fine} = \mathscr{L}_{color} + \mathscr{L}_{edge} + \mathscr{L}_{eikonal}. \tag{14}$$

By periodically pruning the empty voxels and splitting the remaining voxels, we get the fine voxels $\mathscr{V}_f$ which can be used to extract a much more delicate surface. For NeuS and many other coordinate-based methods, they have to reconstruct them separately, but in our method, we can reconstruct them together, and separate them later by moving their supporting voxels around, as shown in Figure 5.

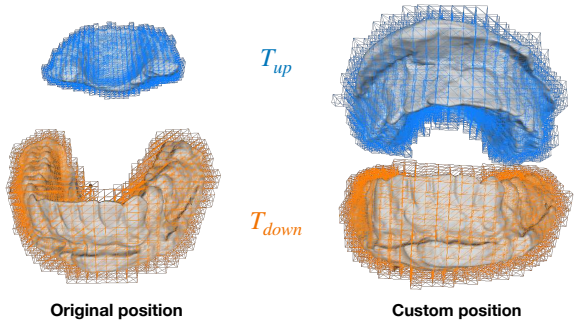

Fig. 5. We can separate the supporting voxels of both arches, and transform their coordinates with two different transformations T_{up} and T_{down}.

3.2 Pose Registration

Another critical component of successful dental AR applications is pose registration, which aligns the reconstructed model to the real teeth image for virtual-real fusion. Existing solutions either use the intermediate features for matching or use a deep model for direct pose regression. However, in our case, the teeth model does not contain enough color information for feature extraction and matching, which will decrease the performance of pose registration.

To handle this issue, we use the trained model in Sec 3.1 to render edge and geometrical information for estimated pose optimization. Specifically, given a query image I_q, and its initial pose, $T_q \in \mathrm{SE}(3)$, we first sample a set of rays according to the estimated pose and obtain the rendering results. Here, we fix the trained MLP model and voxel embedding vectors in Sec 3.1 and send the view direction, d, and interpolated embedding $\mathbf{e}$ of sampled points along the rays into the trained model and render the probability edge map, P_e, and depth $D(r)$.

Since our model is totally differentiable, the error between predicted edge map P_e and extracted edge map E can also back-propagate to the camera pose, which allows us to iteratively optimize the camera pose during registration without additional training. However, in practice, optimization purely depending on the edge is prone to a local minimum due to the noise imposed by lighting conditions, camera blurring, or other environmental problems. Thus, we also leverage the depth loss Equation 15 for pose optimization. Unlike the full depth loss, the depth used here is also the accumulated depth through Equation 3.

$$\mathscr{L}_{depth} = \sum_{r} \|\hat{D}(r) - D(r)\|_1. \tag{15}$$

So, the total loss used in pose registration is Equation 16, where β controls the weight of two losses according to the quality of input depth and edge.

$$\mathscr{L}_{reg} = \mathscr{L}_{depth} + \beta \mathscr{L}_{edge}. \tag{16}$$

3.3 Implementation Details

The detailed network architecture is shown in Figure 7. We set the voxel embedding length to 16, and use a 3-layer MLP with 128 hidden units as the geometry extractor F_σ. it takes ray direction and geometry feature vector as input and outputs 3-dimensional RGB color. The edge extractor F_p is a 2-layer MLP that takes the geometry feature vector as input and outputs 1-dimensional edge probability. Here $PE(\cdot)$ denotes the positional encoding [35]. We use 6 frequencies on ray

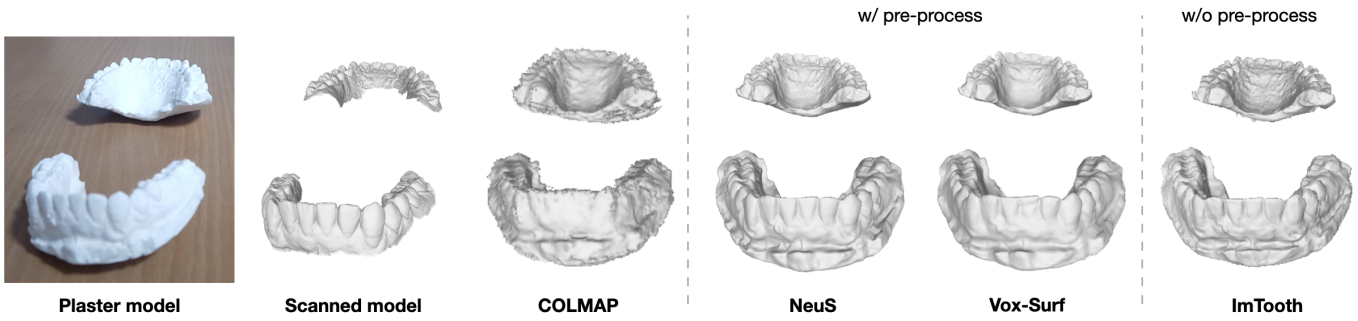

Fig. 6. Surface reconstruction results. From left to right, we show the RGB image of the plaster model, scanned surface (ground truth), and the surface reconstructed by the COLMAP [46, 47], NeuS [58], Vox-Surf [27], our proposed ImTooth, respectively.

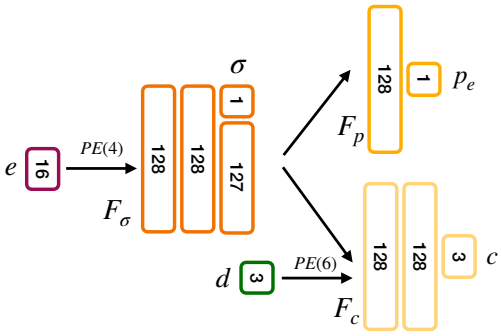

Fig. 7. Detailed architectures of our neural network.

Table 1. Evaluation result of the reconstruction module. We report the chamfer distance and parameters used for different methods.

	Chamfer Dist.	Net Param.	Extra Param.
COLMAP	0.5160mm	-	-
NeuS (pre-process)	0.3251mm	1.41M	-
Vox-Surf (pre-process)	0.3509mm	0.36M	0.50M (8113 emb)
Ours (w/o pre-process)	0.3926mm	0.42M	0.46M (7545 emb)

direction d and 4 frequencies on voxel embeddings $\mathbf{e}$. However, we found positional encoding could be harmful in the early stage. Inspired by recent work [28], we adopt a weighted factor $w_k(l)$ for frequency component according to iteration number l as shown in Equation 17 and Equation 18:

$$\gamma_k(e;l) = w_k(l) \cdot [\cos(2^k \pi e), \sin(2^k \pi e)], \tag{17}$$

$$w_k(l) = \begin{cases} 0 & ,if\ l < k \\ \dfrac{1-\cos((l-k)\pi)}{2} & ,if\ k \le l < k+1 \\ 1 & ,if\ l \ge k+1 \end{cases} \tag{18}$$

We generate the initial octree using voxels of size 0.1 and set the maximum number of voxel hits to 20. The initial ray marching step is 0.001 and decreases with the voxel size. We apply the voxel pruning and splitting at 10,000, 20,000, 30,000, 50,000, and 100,000 iterations, respectively, the pruning threshold $\tau = 0.01$. We adopt the surface-aware voxel resampling strategy [27] after 30,000 iterations. We set l_{key} to 5,000 for scene initialization. For each iteration, we random sample 4,096 rays from select frames. The learning rate for the network and embeddings is 0.001 and 0.00001 for the pose. For pose registration, we set the β in Equation 16 to 0.01 to prevent optimization to a local minimum in the early stage. All experiments are run on a single NVIDIA V100 graphics card.

4 Experiments

4.1 Dataset

The data used in all experiments are captured by a Microsoft HoloLens 2. The raw RGB image sequence is captured by PhotoVideo (PV) camera in size 1920×1080. To maintain the real scale, we also record the depth map in size 512×512 in Articulated HAnd Tracking (AHAT) mode and directly project to the PV frame. We also save the raw poses estimated by HoloLens 2 for initialization.

We first collect a pair of plaster tooth casts from a volunteer and also the corresponding scanned model generated by iTero scanner [†] as ground truth. We then record three datasets for evaluation of our system: OnTable, OnHand, and Real. The OnTable dataset is used for reconstruction, and we place the upper and lower teeth model on the table. The OnHand dataset is used for the evaluation of registration, and we place the lower teeth model on hand and move it freely. The Real dataset is a set of images captured from the real mouth of the volunteer for real-world testing.

4.2 Evaluation of Surface Reconstruction

In this part, we evaluate the performance of the proposed surface reconstruction module. The compared baseline methods, evaluation metrics, and results of this part are shown in the following.

Baseline and Metrics. We compare our method with the traditional reconstruction method, COLMAP [46, 47], and the recent implicit surface reconstruction method, NeuS [58], Vox-Surf [27]. We show the reconstruction error of the point clouds between the reconstructed result and ground truth model and also the number of used parameters of the neural network and additional embedding vectors. For NeuS and Vox-Surf, we use preprocessed poses following the procedure in NeuS, which is complex and requires human assistance. Instead, our method directly uses the raw poses collected by HoloLens 2 as input and continuously optimizes the raw poses during the reconstruction.

The metric used for the evaluation of the reconstruction quality is shown in the following equation:

$$\begin{aligned} \mathscr{D}_{recon} &= \frac{1}{|P|} \sum_{(p,q)\in\Lambda_{P,Q}} \|p-q\|^2, \\ \Lambda_{P,Q} &= \{(p, argmin_q \|p-q\|)\}, \end{aligned} \tag{19}$$

where $p \in P$ and $q \in Q$ are two sets of points sampled from the reconstructed and ground truth meshes.

[†] https://itero.com

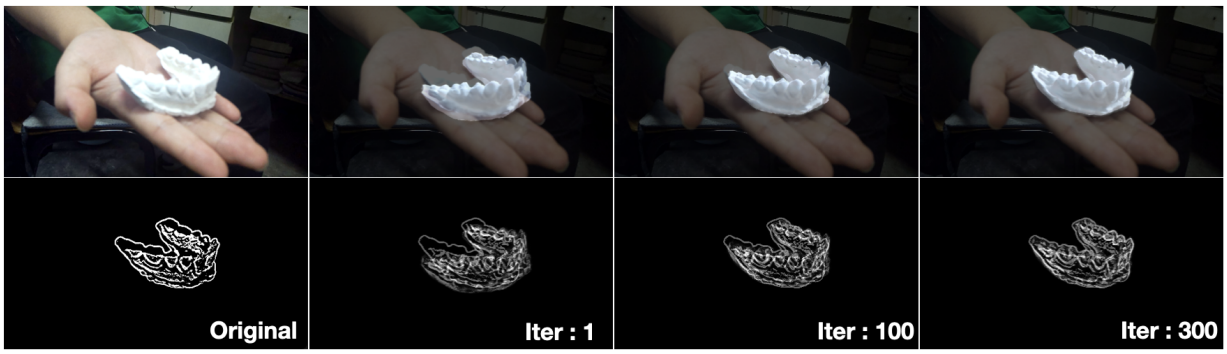

Fig. 8. Projection results for iterative pose optimization. As shown in the figure, as the number of iterations increases, the reconstructed model is gradually aligned with the model of the image, and the edges coincided.

Table 2. Runtime performance of different steps in our proposed method.

per iter	ray intersection	ray sample	emb	sdf	norm	edge	render	backward	full
time (ms)	2.149	1.121	0.741	1.528	7.032	0.558	1.016	24.815	41.767

Qualitative and Quantitative Results. The overall quantitative results are shown in Table 1. Even though we do not use the ground truth pose data for reconstruction, we also can achieve comparable performance with two recent implicit reconstruction approaches. The small gap is mainly due to the surface defects in some under-observed areas caused by the early depth noise. For the parameters (both neural network and extra features) comparison, our requirement is much lower than NeuS (0.88M vs. 1.41M), which indicates that we can apply the model on the mobile device with a faster inference speed. We also show the qualitative reconstruction results in Figure 6. As shown in the figure, our proposed ImTooth is able to reconstruct surfaces comparable to other methods without accurate pose and outperforms the traditional method like COLMAP.

4.3 Evaluation of Pose Registration

In this part, we evaluate the performance of our proposed pose registration module. The compared baseline methods, evaluation metrics, and results of this part are shown in the following. We use the OnHand dataset for evaluating the registration. The GT poses are obtained by reconstructing the OnHand dataset. During reconstruction, we mask out the background to eliminate the inconsistency caused by hand moving.

Baseline and Metrics. We compare our method with the visual localization method HLoc [44], which estimates the camera pose by global image retrieval and local feature matching. HLoc combines a state-of-the-art learning-based keypoint descriptor SuperPoint [14], and feature matcher SuperGlue [45]. Besides, we also compare the iterative closest point (ICP) [4] approach, which uses depth from HoloLens 2 and reconstructed models for pose optimization. To verify the effectiveness of our optimization loss in Sec 3.2, we show both the performance of only using depth and using both depth and edge information. For the evaluation metrics, we follow the setting in [44], which estimates the errors of the position and orientation values between the predicted and ground truth data.

Table 3. Evaluation results of the registration module. We report the recall [%] at different distance and orientation thresholds.

dist. [cm] ori. [deg]	2.5 / 5.0 / 7.5 5 / 10 / 15
HLoc (SP+SG) [14, 44, 45]	5.88 / 12.94 / 17.65
ICP [4]	17.64 / 38.82 / 51.76
Ours (Depth)	76.47 / 84.70 / 92.94
Ours (Edge + Depth)	80.00 / 90.58 / 96.47

Qualitative and Quantitative Results. To validate the effectiveness of our proposed method, we perform per-frame pose estimation with the query images captured by a HoloLens 2. The query images are captured at different locations and viewpoints. We manually initialize the pose by finding an approximate transformation and add a random noise within [-5, 5] (cm) translation and [-5, 5] (deg) rotation in three directions.

In order to measure the localization accuracy of the proposed method, we use the commonly used relative translation error (Equation 20) and relative rotation error (Equation 21):

$$M_{RTE} = |t - \hat{t}|, \tag{20}$$

$$M_{RRE} = \arccos((trace(R^T \hat{R}) - 1)/2), \tag{21}$$

where $\hat{t}$ and $\hat{R}$ are the ground-truth translation and rotation, respectively, and t and R are the estimated ones. For the ground-truth pose, we first estimate the pose through the structure-from-motion algorithm, and then scale the pose according to the actual scanning model.

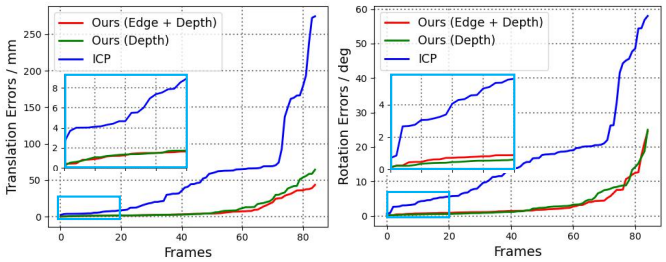

Fig. 9. Translation (mm.) and rotation (deg.) errors of different methods. We show the error of ICP and the proposed method with different settings. For better visualization, we zoom in on some curve results inside the blue box.

The pose recalls at different translation and orientation thresholds are shown in Table 3. As shown in the table, our approach achieves the best results for different thresholds. The feature-based visual localization methods can not achieve good performance on our data due to the textureless region of the teeth model. As shown in the last two rows in the table, combining two kinds of geometrical information (depth and edge) can lead to the best registration performance. Additionally, we show translation and rotation errors for different test frames taken from different viewpoints in Figure 9. To get a better view of the curve, we

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3247459

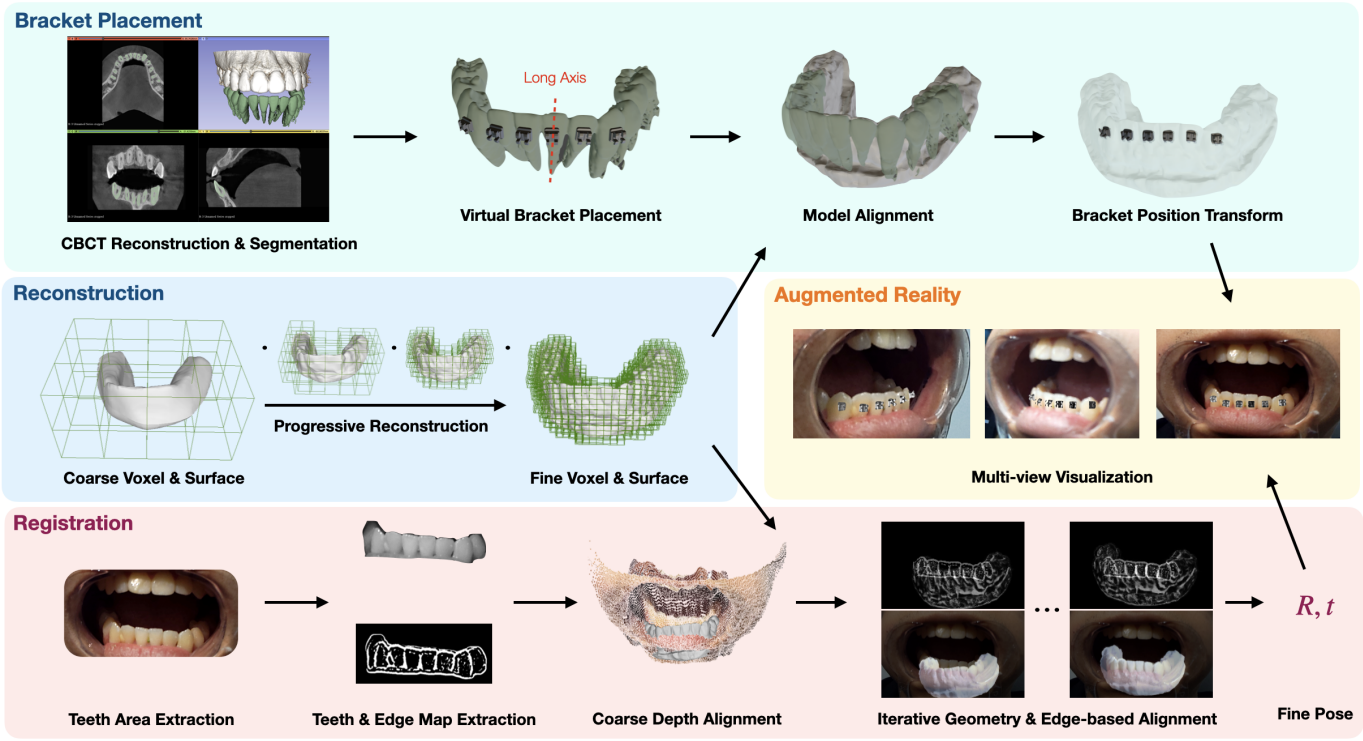

Fig. 10. The full process of dental brackets guidance system based on ImTooth. The whole application is divided into four modules. The reconstruction module is used to reconstruct the tooth model based on neural implicit representation, the bracket placement module is used to determine the position of the bracket on the reconstructed model, the registration module is used to align the actual tooth image with the reconstructed model, the augmented reality module performs virtual-real fusion display according to the estimated position.

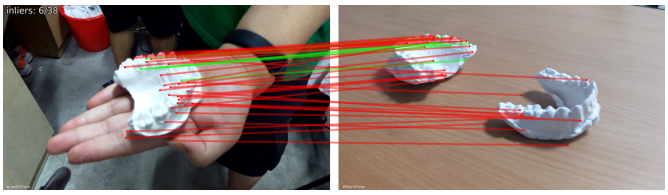

Fig. 11. Failure case of HLoc feature matching. The red (green) lines represent the outliers (inliers) matching between feature points between different images. As we can see, feature-based methods can be fragile in such textureless objects, even the "inliers" are still mismatched.

zoom in on the graph of the first 20 frames sorted by error. It can be seen from the figure that our Imtooth can achieve very good registration results (translation error is less than 2mm, rotation error is less than 5 degrees), which is sufficient to meet medical needs. However, due to the fixed-focus camera and low-resolution depth camera in consumer-grade AR glasses, there will be blurring in the short distance, and the depth in the long distance is not reliable, resulting in poor registration accuracy for other viewing angles.

Also, we give the visualization results of the iterative pose optimization process in Figure 8. We overlay the reconstructed teeth model and edge map on the original RGB and edge images with the estimated pose from different optimization iterations. It can be seen from the figure that with gradual optimization, the estimated pose gradually reaches the result of ground truth.

Additionally, we show some failed feature matching visualization results for HLoc (SuperPoint + SuperGlue) in Figure 11. According to the feature point matching results, we can know that due to the textureless structure, feature-based methods may produce very unreliable results, leading to the failure of registration. Although our method achieves good performance using geometry-informed optimization, it fails when the overlap between the initial and actual position is too small, as it may get stuck in a locally optimal solution. Therefore, our method is more suitable for local pose adjustment during tracking.

We also report the runtime cost for each part during registration as shown in Table 2. Each optimization iteration takes 40 ms, and we optimize each scene with 100 iterations. In practice, we find that with a good initialization position, the iteration number can be decreased.

5 Potential Applications

We develop a teeth bracket guidance system based on our ImTooth system. The full process is shown in Figure 10. We divide the full system into four modules. We will elaborate on the specific steps of each module below:

Reconstruction Module The goal of this module is to obtain a reconstructed neural implicit model. First, we need to obtain a plaster model of the patient's teeth, which is usually a routine procedure in dental clinics. We then use AR glasses to capture multi-view images of plaster casts and run the automatic scale-preserving reconstruction algorithm proposed in subsection 3.1. After this step, a set of embedded voxels and a corresponding extractor are generated, which is the core part of our system.

Bracket Placement Module This module is used to predetermine the placement position on the reconstructed model of the bracket, which is usually based on the reconstructed model of CBCT. We first extract approximate tooth regions from CBCT. Since the CT model has a clear root and crown structure, the position of the long axis of the tooth can be estimated more accurately, and the bracket position can be determined accordingly. Since the scale of our model is consistent with the CT model, the bracket position can be easily transformed into the ImTooth model coordinate system by global alignment.

Registration Module This module is used to estimate the relative pose of the reconstructed model and the real teeth in the current view. For actual patients, we first find candidate tooth regions, which can be done by template matching or some off-the-shelf facial key point detec-

tion methods. Then, we generate edge maps and segmentation masks [37] through edge detection and background removal. As described in subsection 4.3, the initial pose is important to our method. Therefore, we design an interactive coarse position initialization method. Through the gesture interaction of HoloLens 2, Doctors can manually place a virtual teeth model near the patient's teeth to bootstrap the registration algorithm. After initialization, our algorithm will use the geometry and pre-learned edge information to iteratively optimize the relative pose so that the reconstructed tooth can fit the actual tooth area. It is worth noticing that the initialization can also be done automatically using the depth alignment method, However, this can be less stable due to insufficient viewing area.

Augmented Reality Module This module mainly performs a virtual-real fusion display on the AR glasses according to the position of the bracket preset in the previous steps and the final optimized pose and assists the doctor in the treatment. Since AR glasses usually have their own SLAM module, the number of iterative optimizations can be reduced by merging with the real-time pose estimated by the glasses.

5.1 Implementation Details

The system uses HoloLens 2 as data acquisition, display, and interaction equipment. We first run the reconstruction procedure in offline mode and store the reconstructed model in AR glasses for interactive virtual bracket placement. The registration steps are completed on a computing server with a single NVIDIA RTX 2080Ti graphics card and transmitted the optimized pose to the glasses through the wireless network. Although the optimization cannot run in real-time, with the real-time pose estimated by HoloLens 2, we adopt a delayed optimization strategy that performs the optimization for every n frame to compensate for the drift from HoloLens 2's pose. This strategy allows us to minimize optimization steps in practical applications.

6 Limitation and Future Works

Currently, our method still has some limitations in practical use. Firstly, for real-world applications, the texture is prone to be affected by the environment, the shadow or overexposure could affect the accuracy of edge extraction, which directly affects the pose estimation result. For this problem, a more robust edge detection method, especially for teeth, would be helpful for registration. Secondly, the initial position is important for getting good pose optimization. However, identifying the teeth from occluded images is challenging. We believe that a descriptor for teeth is a good choice for precise localization without human interaction. Thirdly, our system still has yet to achieve true real-time performance ($\sim$ 30 Hz). We will further improve and optimize our method for better performance.

7 Conclusion

In this work, we propose a novel dental AR system named ImTooth. Our system leverages neural implicit representation techniques for modeling and tracking. Our system greatly simplifies the current AR process and does not require additional high-precision scanners, significantly reducing the overall cost. Our system can easily be integrated into the dental diagnosis and treatment process, and provide better assistance to dentists.

8 Acknowledgments

This work was partially supported by NSF of China (No. 61932003).

References

[1] A. Aichert, W. Wein, A. Ladikos, T. Reichl, and N. Navab. Image-based tracking of the teeth for orthodontic augmented reality. In *Medical Image Computing and Computer-Assisted Intervention*, vol. 7511, pp. 601–608, 2012.

[2] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5297–5307, 2016.

[3] R. Arandjelovic and A. Zisserman. All about VLAD. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1578–1585, 2013.

[4] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9:698–700, 1987.

[5] D. Azinovic, R. Martin-Brualla, D. B. Goldman, M. Nießner, and J. Thies. Neural RGB-D surface reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6280–6291, 2022.

[6] H. Bay, T. Tuytelaars, and L. V. Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision*, vol. 3951, pp. 404–417, 2006.

[7] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother. DSAC - differentiable RANSAC for camera localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2492–2500, 2017.

[8] E. Brachmann and C. Rother. Learning less is more - 6D camera localization via 3d surface regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4654–4662, 2018.

[9] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz. Geometry-aware learning of maps for camera localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2616–2625, 2018.

[10] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF: Binary robust independent elementary features. In *European Conference on Computer Vision*, vol. 6314, pp. 778–792, 2010.

[11] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su. MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo. In *IEEE International Conference on Computer Vision*, pp. 14104–14113, 2021.

[12] S. Cho, S. Hong, S. Jeon, Y. Lee, K. Sohn, and S. Kim. Cats: Cost aggregation transformers for visual correspondence. *Advances in Neural Information Processing Systems*, 34:9011–9023, 2021.

[13] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *European Conference on Computer Vision Workshops*, pp. 1–2, 2004.

[14] D. DeTone, T. Malisiewicz, and A. Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018.

[15] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler. D2-Net: A trainable CNN for joint description and detection of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8092–8101, 2019.

[16] S. Garg, N. Suenderhauf, and M. Milford. Semantic-geometric visual place recognition: a new perspective for reconciling opposing views. *The International Journal of Robotics Research*, 41:573–598, 2022.

[17] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li. Self-supervising fine-grained region similarities for large-scale image localization. In *European Conference on Computer Vision*, vol. 12349, pp. 369–386, 2020.

[18] M. Geppert, P. Liu, Z. Cui, M. Pollefeys, and T. Sattler. Efficient 2D-3D matching for multi-camera visual localization. In *International Conference on Robotics and Automation*, pp. 5972–5978, 2019.

[19] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman. Implicit geometric regularization for learning shapes. In *International Conference on Machine Learning*, vol. 119, pp. 3789–3799, 2020.

[20] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer. Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14141–14152, 2021.

[21] T.-K. Huang, C.-H. Yang, Y.-H. Hsieh, J.-C. Wang, and C.-C. Hung. Augmented reality (ar) and virtual reality (vr) applied in dentistry. *The Kaohsiung journal of medical sciences*, 34:243–248, 2018.

[22] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3304–3311, 2010.

[23] W. Jiang, E. Trulls, J. Hosang, A. Tagliasacchi, and K. M. Yi. COTR: Correspondence transformer for matching across images. In *IEEE International Conference on Computer Vision*, pp. 6187–6197, 2021.

[24] A. Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6555–6564, 2017.

[25] A. Kendall, M. Grimes, and R. Cipolla. PoseNet: A convolutional network for real-time 6-dof camera relocalization. In *IEEE International Conference on Computer Vision*, pp. 2938–2946, 2015.

[26] S. Leutenegger, M. Chli, and R. Siegwart. BRISK: Binary robust invariant scalable keypoints. In *IEEE International Conference on Computer Vision*,

pp. 2548–2555, 2011.

[27] H. Li, X. Yang, H. Zhai, Y. Liu, H. Bao, and G. Zhang. Vox-Surf: Voxel-based implicit surface representation. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–12, 2022.

[28] C. Lin, W. Ma, A. Torralba, and S. Lucey. BARF: Bundle-adjusting neural radiance fields. In *IEEE International Conference on Computer Vision*, pp. 5721–5731, 2021.

[29] Y. Lin, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T. Lin. iNeRF: Inverting neural radiance fields for pose estimation. In *IEEE International Conference on Intelligent Robots and Systems*, pp. 1323–1330, 2021.

[30] C. Llena, S. Folguera, L. Forner, and F. Rodríguez-Lozano. Implementation of augmented reality in operative dentistry learning. *European Journal of Dental Education*, 22:e122–e130, 2018.

[31] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[32] R. Martin-Brualla, N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth. NeRF in the Wild: Neural radiance fields for unconstrained photo collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7210–7219, 2021.

[33] N. L. Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1:99–108, 1995.

[34] M. Mehran and B. T. Leila. A novel augmented reality system of image projection for image-guided neurosurgery. *Acta Neurochirurgica*, 155:943–947, 2013.

[35] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, vol. 12346, pp. 405–421, 2020.

[36] M. Oechsle, S. Peng, and A. Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *IEEE International Conference on Computer Vision*, pp. 5569–5579, 2021.

[37] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaïane, and M. Jägersand. U^2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020.

[38] F. Radenović, G. Tolias, and O. Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1655–1668, 2018.

[39] J. Revaud, J. Almazán, R. S. Rezende, and C. R. d. Souza. Learning with average precision: Training image retrieval with a listwise loss. In *IEEE International Conference on Computer Vision*, pp. 5107–5116, 2019.

[40] P. Rhienmora, K. Gajananan, P. Haddawy, M. N. Dailey, and S. Suebnukarn. Augmented reality haptics system for dental surgical skills training. In *ACM Symposium on Virtual Reality Software and Technology*, pp. 97–98, 2010.

[41] B. Roessle and M. Nießner. End2End multi-view feature matching using differentiable pose optimization. *arXiv preprint arXiv:2205.01694*, 2022.

[42] J. P. Rolland and H. Fuchs. Optical versus video see-through head-mounted displays in medical visualization. *Presence: Teleoperators and Virtual Environments*, 9:287–309, 2000.

[43] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski. ORB: An efficient alternative to SIFT or SURF. In *IEEE International Conference on Computer Vision*, pp. 2564–2571, 2011.

[44] P. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk. From Coarse to Fine: Robust hierarchical localization at large scale. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12716–12725, 2019.

[45] P. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4937–4946, 2020.

[46] J. L. Schönberger and J. Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016.

[47] J. L. Schönberger, E. Zheng, J. Frahm, and M. Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, vol. 9907, pp. 501–518, 2016.

[48] J. Shi, S. Liu, Z. Zhu, Z. Deng, G. Bian, and B. He. Augmented reality for oral and maxillofacial surgery: The feasibility of a marker-free registration method. *The International Journal of Medical Robotics and Computer Assisted Surgery*, p. e2401, 2022.

[49] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. W. Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2930–2937, 2013.

[50] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou. LoFTR: Detector-free local feature matching with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8922–8931, 2021.

[51] L. Sun, H.-S. Hwang, and K.-M. Lee. Registration area and accuracy when integrating laser-scanned and maxillofacial cone-beam computed tomography images. *American Journal of Orthodontics and Dentofacial Orthopedics*, 153:355–361, 2018.

[52] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii. InLoc: Indoor visual localization with dense matching and view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:1293–1307, 2021.

[53] D. Tan, J.-J. Liu, X. Chen, C. Chen, R. Zhang, Y. Shen, S. Ding, and R. Ji. ECO-TR: Efficient correspondences finding via coarse-to-fine refinement. *arXiv preprint arXiv:2209.12213*, 2022.

[54] M. Trajkovic and M. Hedley. Fast corner detection. *Image and Vision Computing*, 16:75–87, 1998.

[55] D. Wang, H. Tong, Y. Shi, and Y. Zhang. Interactive haptic simulation of tooth extraction by a constraint-based haptic rendering approach. In *IEEE International Conference on Robotics and Automation*, pp. 278–284, 2015.

[56] J. Wang, Y. Shen, and S. Yang. A practical marker-less image registration method for augmented reality oral and maxillofacial surgery. *International Journal of Computer Assisted Radiology and Surgery*, 14:763–773, 2019.

[57] J. Wang, H. Suenaga, K. Hoshi, L. Yang, E. Kobayashi, I. Sakuma, and H. Liao. Augmented reality navigation with automatic marker-free image registration using 3D image overlay for dental surgery. *IEEE Transactions on Biomedical Engineering*, 61:1295–1304, 2014.

[58] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems*, pp. 27171–27183, 2021.

[59] K. Wong, H. M. Yee, B. A. Xavier, and G. A. Grillone. Applications of augmented reality in otolaryngology: A systematic review. *Otolaryngology–Head and Neck Surgery*, 159:956–967, 2018.

[60] P. Xia, A. M. Lopes, and M. T. Restivo. Virtual reality and haptics for dental surgery: a personal review. *The Visual Computer*, 29:433–447, 2013.

[61] X. Yang, H. Li, H. Zhai, Y. Ming, Y. Liu, and G. Zhang. Vox-Fusion: Dense tracking and mapping with voxel-based neural implicit representation. In *IEEE International Symposium on Mixed and Augmented Reality*, pp. 80–89, 2021.

[62] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems*, pp. 4805–4815, 2021.

[63] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, R. Basri, and Y. Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems*, 2020.

[64] A. Yu, V. Ye, M. Tancik, and A. Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4578–4587, 2021.

[65] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger. MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv:2022.00665*, 2022.