ToothInpaintor: Tooth Inpainting from Partial 3D Dental Model and 2D Panoramic Image

Yuezhi Yang[1] , Zhiming Cui[1,2], Changjian Li[3], Wenping Wang[1]

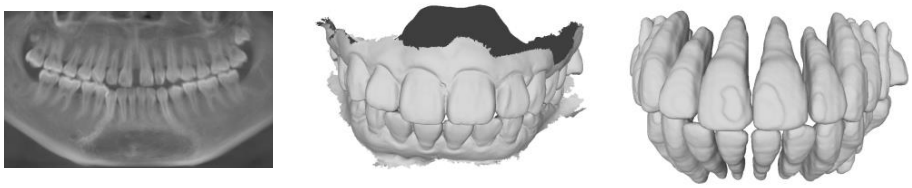
[1] Department of Computer Science, The University of Hong Kong, Hong Kong
[2] Shanghai United Imaging Intelligence Co., Ltd., China
[3] Department of Computer Science, University College London, London, UK

Abstract. In orthodontic treatment, a full tooth model consisting of both the crown and root is indispensable in making the treatment plan. However, acquiring tooth root information to obtain the full tooth model from CBCT images is sometimes restricted due to the massive radiation of CBCT scanning. Thus, reconstructing the full tooth shape from the ready-to-use input, e.g., the partial intra-oral scan and the 2D panoramic image, is an applicable and valuable solution. In this paper, we propose a neural network, called *ToothInpaintor*, that takes as input a partial 3D dental model and a 2D panoramic image and reconstructs the full tooth model with high-quality root(s). Technically, we utilize the implicit representation for both the 3D and 2D inputs, and learn a latent space of the full tooth shapes. At test time, given an input, we successfully project it to the learned latent space via neural optimization to obtain the full tooth model conditioned on the input. To help find the robust projection, a novel adversarial learning module is exploited in our pipeline. We extensively evaluate our method on a dataset collected from real-world clinics. The evaluation, comparison, and comprehensive ablation studies demonstrate that our approach produces accurate complete tooth models robustly and outperforms the state-of-the-art methods.

1 Introduction

In computer-aided orthodontic dentistry, a full tooth shape with roots is indispensable in evaluating the past treatment, measuring the teeth movement, and making the new treatment plan[7, 8]. Cone-bean computed tomography (CBCT) images are the only data modality containing 3D tooth root information. However, considering the massive radiation of CBCT scanning, utilizing a CBCT to acquire the full tooth shape of the patient in orthodontic treatment is not allowed in many countries in the world. Thus, reconstructing complete 3D tooth shape from the more accessible data modalities, i.e., the 2D panoramic image and the partial intra-oral dental model, is an applicable and promising direction.

It is, however, a challenging task due to the following facts. First, the panoramic image only contains the projected tooth information (Fig. 1(a)), which has the inherent ambiguity from 2D to 3D. And second, the intra-oral scan contains the crown shape solely (Fig. 1(b)). Previous works [2, 11, 1] have been explored to



(a) (b) (c)

Fig. 1. Given (a) a panoramic image and (b) several partial dental models as inputs, (c) the faithful full tooth models are expected to be predicted.

utilize template-fitting based methods to deform templates to match the tooth contour and the crown shape. However, these methods require a set of pre-defined tooth templates, which cannot represent the varying tooth shape appearance flexibly. Recently, many learning-based methods have been proposed for general 3D shape generation and completion [10, 5]. But these methods only consider the partial 3D information, which is not applicable for our specific task where the panoramic image provides the trusted tooth contour as an additional reference.

In this paper, we present a novel learning-based method for tooth reconstruction from a partial 3D dental model and a 2D panoramic image. Briefly, inspired by the successful attempts of the implicit representation [10] in deep learning, we represent both the 2D and 3D tooth information via the corresponding signed distance function (SDF) fields. With the SDF representation, we then build a latent space of full tooth models using an auto-decoder neural network, where various tooth types and shapes are encoded and each code corresponds well to a high-quality full tooth shape. At test time, given the 3D SDF of the partial dental model and the 2D SDF of the panoramic image as inputs, the goal is to project them into the latent space to find a faithful tooth shape code to derive the full model. This is achieved by neural optimization via the auto-decoder network, where the network parameters are fixed and only the resulting code is optimized. Furthermore, to help maintain the global tooth shape during the code optimization, adversarial learning is introduced to assist the robust projection. The extensive experiments show that our method can reconstruct high-quality complete tooth models with the panoramic image and a partial dental model, which has significantly outperformed the state-of-the-art performance and offer the potential usability of our framework in real-world clinics.

2 Method

Fig. 2 shows an overview of our *ToothInpaintor*, which consists of the data processing, multi-dimensional SDF, and adversarial learning modules. We elaborate on the pipeline in this section.

2.1 Data Processing

As shown in Fig. 2, given the inputs of CBCT images, the panoramic image, and the intra-oral scan, we process them to obtain the required data as follows.

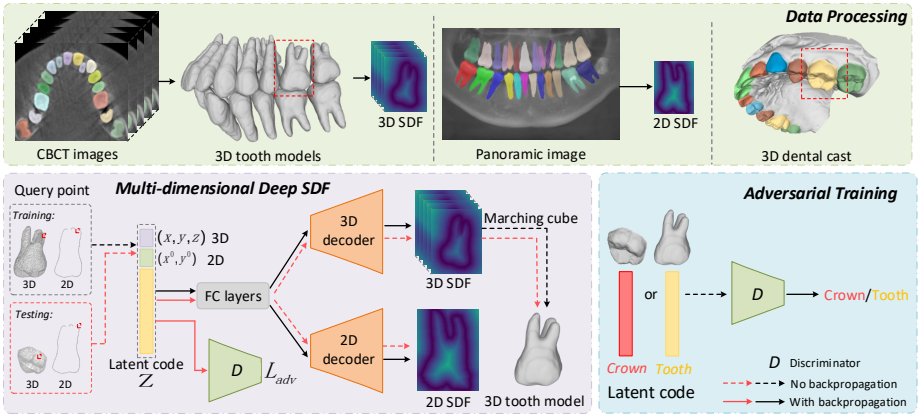


Fig. 2. The pipeline of *ToothInpaintor*, which consists of the data processing (Sec. 2.1), mutli-dimensional deep SDF and the adversarial learning (Sec. 2.2) modules. Note here, a dotted line means no back-propagation in the corresponding data flow, as illustrated in the legend.

Firstly, we adopt ToothNet [4] to accurately segment tooth individuals from CBCT images and reconstruct them as the ground truth full tooth shapes. We then utilize TSegNet [3] to faithfully segment tooth crown from the dental model. Finally, Mask R-CNN [6] is employed to segment the individual tooth from the panoramic image. To map the panoramic image back to the tooth models, we first calculate the biquadratic curve of the tooth center points to fit a dental arch curve. Then, all 2D tooth centers from the panoramic image are obtained and fitted to the dental arch curve so that each tooth contour can align well to its corresponding tooth model. To better calculate the SDF field, we normalize the extracted full tooth individuals within a unit ball, and roughly scale and place the partial crown model to fit a full tooth shape.

2.2 Multi-dimensional Deep SDF

In our approach, we utilize the signed distance function (SDF) to represent both the 2D and 3D tooth shapes. Take the 3D tooth shape as an example, given a spatial point, its SDF value indicates the distance to the closest surface and the sign refers to whether the point is inside (negative) or outside (positive) of the watertight surface. With the SDF representation, the tooth model can be directly instantiated by Marching Cubes [9].

Since we have both the 3D tooth and 2D contour shapes, we exploit multi-dimensional SDFs as in the following.

3D SDF. Inspired by DeepSDF [10], given a tooth shape, we prepare a set of pairs X consists of the 3D point samples and their SDF values:

$$X = \{((x, y, z), s) : s = SDF((x, y, z))\}. \tag{1}$$

Intuitively, having these data pairs, we can train a neural network to produce the SDF value given any spatial query point for a specific tooth model. However,

we want a neural network that can represent various tooth shapes, and embed them together in a low-dimensional latent space. Thus, as shown in Fig. 2, we utilize a shape code z (256-dim vector) as the additional input to the network and map it to the desired shape represented by the continuous SDF field using the network $f_\theta((x, y, z), z_i)$, defined as:

$$f_\theta((x, y, z), z_i) \approx SDF^i((x, y, z)) \tag{2}$$

where x is an arbitrary spatial point, z_i refers to the shape code of the ith shape that needs to be learned by the network, and $SDF^i((x, y, z))$ is the ground truth SDF value of the ith shape. Note here, f_θ is an auto-decoder network and composed of several Fully Connected (FC) layers to produce the SDF values.

2D SDF. Recall that, the 2D tooth contour shape serves as an important shape reference in our task, and we further use the 2D SDF derived from the tooth contour. Similarly, we calculate a 2D SDF field from each tooth contour, where the value of each pixel indicates the signed distance to the iso zero-contour. To incorporate the 2D SDF input, as can be seen from Fig. 2, we simply concatenate the 2D *pixel-level* query point (x^0, y^0), the latent code z, and the 3D spatial point (x, y, z) together. And our auto-decoder network has an extra output branch to predict the 2D SDF. In this way, the network can automatically build the intrinsic relationship between the 3D tooth shape and the projected 2D contour.

Loss Function. Under the configuration of deep SDF learning, the network parameters and latent code z should be optimized jointly, and we supervise the prediction of the 2D and 3D SDFs. Additionally, we introduce a regularization term on the latent code to improve the generalization ability. All together, the loss function $\mathcal{L}$ is formulated as:

$$\min_{\theta,z} \mathcal{L} = \mathcal{L}_{SDF}^{3D} + \mathcal{L}_{SDF}^{2D} + \lambda \left\| z \right\|_2 , \tag{3}$$

where $\mathcal{L}_{SDF}^{3D}$ and $\mathcal{L}_{SDF}^{2D}$ refer to the L1 loss of the 3D and 2D SDFs prediction, respectively. And λ is a balancing weight that is set as 0.0001 in all experiments.

2.3 Shape Code Optimization with Adversarial Learning

At the testing stage, the goal is to project the inputs (i.e., the crown model and the panoramic image) to the learned latent space to find a correct code that encodes the full tooth shape respecting the inputs. We achieve it using neural optimization. Specifically, we fix the network parameters since it has captured the tooth shape priors when building the latent space, and only update the code given the supervision of the crown shape and the contour shape. See Fig. 2 for an illustration, where the red lines present the data flow at the testing time and dotted lines mean no gradient back-propagation. Yet, when solely supervising the prediction of the SDFs to optimize the code, the crown shape tends to dominate the optimization so that the resulting code cannot maintain the global tooth shape well, e.g., with irregular tooth roots. To resolve this issue, we further propose adversarial learning in this optimization step.

Adversarial Learning. To use the discriminator, we should first train it at the training time. Specifically, we have the code z_{full} representing the full tooth shape, and also optimize the partial shape code $z_{partial}$ by feeding the network with the crown model. To learn $z_{partial}$, we disable the gradient back-propagation of the auto-decoder network parameters when supervising the reconstruction of the partial shape to update the code solely. As illustrated in Fig. 2, we take z_{full} as the positive example, and $z_{partial}$ as the negative example to train the discriminator. Note here, the discriminator loss is only used to optimize itself instead of the learning of the code and auto-decoder network parameters. After training, the discriminator has the ability to tell whether a shape code represents a regular full shape or not.

Testing Stage Optimization. With the learned network f_θ and the discriminator, we optimize the resulting shape code $\widehat{z}$ by supervising the reconstruction and the discriminator losses, defined as:

$$\widehat{z} = \underset{z}{argmin}(\mathcal{L}_{SDF}^{3D} + \mathcal{L}_{SDF}^{2D} + \lambda \left\| z \right\|_2 + \mathcal{L}_{adv}), \tag{4}$$

where the $\mathcal{L}_{SDF}^{3D}$ and the $\mathcal{L}_{SDF}^{2D}$ are the same L1 loss as in the training stage, but the only difference is that the 3D SDF comes from the partial crown model. And $\mathcal{L}_{adv}$ is the common adversarial loss.

Stopping Criteria. The well-learned discriminator is the key module to help maintain the global tooth shape in neural optimization. Hence, instead of stopping the optimization by checking the SDF reconstruction error, e.g., lower than a proper threshold, we refer to the discriminator loss. That is, given the supervision of positive labels, in case the discriminator converges to a low error, we then stop the optimization. In this way, the SDF loss reaches an acceptable low value indicating the satisfactory reconstruction of the inputs, while the resulting code successfully fools the discriminator demonstrating that the code represents a regular full tooth shape.

Once we have the optimized latent code $\widehat{z}$, we directly feed forward the network to calculate the 3D SDF value of each query point in a spatial grid with 512^3 resolution. Then, the predicted complete tooth shape can be reconstructed by Marching Cubes [9].

Post-nonrigid Deformation. The network cannot always guarantee to reproduce the input crown shape due to the optimization nature, we thus exploit a post-nonrigid deformation step to further refine the tooth crown shape to have the geometric details from the crown input. Specifically, laplacian surface editing [12] is employed with a set of deformation handles built by nearest neighbor searching between the predicted tooth model and the crown model. Finally, we take the tooth models after non-rigid deformation as our final results.

2.4 Implementation Details and Network Training

Our framework was implemented using PyTorch, and we used Adam optimizer to train the framework, where the learning rates are set as 10^{-3} and 10^{-4},

Table 1. Quantitative results of alternative ablation networks. The smaller the value, the better the reconstruction accuracy.

Method	CD↓	HD↓	ASD↓
bNet	0.96	2.56	0.47
bNet-P	0.79	2.39	0.37
bNet-P-D	0.78	2.38	0.36
FullNet	**0.75**	**2.36**	**0.33**

respectively. In total, we trained the whole network 1000 epochs in the training stage for about 12 hours using a Nvidia 1080Ti GPU. The auto-decoder network has 7 FC layers as backbone and 7 and 5 FC layers for the following 3D and 2D branches, respectively. The discriminator has 5 FC layers.

3 Experiments

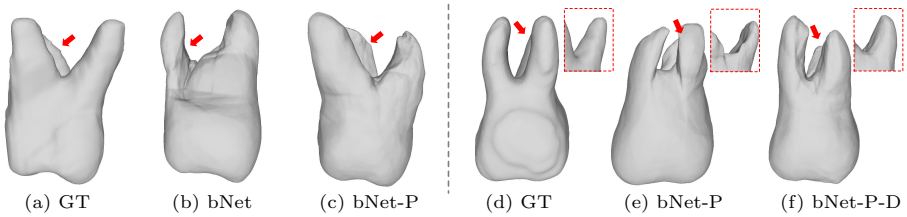
Dataset. To train and test our approach, we have collected a dataset from real-world clinics, where different data modalities can be obtained by scanning. In total, there are 135 paired CBCT images, panoramic images, and intra-oral scans in our dataset, where the resolution of the CBCT images and panoramic images are $0.35mm$ and $0.1mm$, respectively. We randomly split it into 100, 10, and 25 for training, validating, and testing, respectively.

Evaluation Metrics. To quantitatively evaluate the performance of our method, we make use of several metrics to measure the reconstruction accuracy. Concretely, Chamfer distance (CD), Hausdorff distance (HD), as well as average surface distance (ASD) are chosen and are reported in Table 1.

3.1 Ablation Study

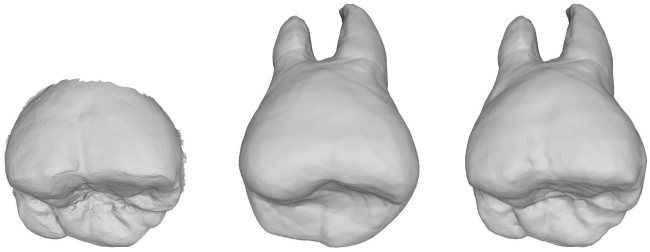
To demonstrate the efficacy of the key network components or loss terms, we have conducted additional experiments with alternative configurations. Specifically, we first build the baseline network denoted as bNet, with only the 3D full tooth shapes to build the manifold and the partial scan solely to optimize a code to produce the full tooth shape. All alternative configurations are derived by augmenting bNet with one network component or loss term and trained using the same training dataset from scratch.

Benefits of 2D SDF from Panoramic Image. Reconstructing the full 3D tooth shape from the partial crown model is a highly ill-posed problem, but the 2D contour from the panoramic image can alleviate the ambiguity to some extent. To validate its effectiveness, we augment bNet with the extra 2D SDF input, denoted as bNet-P. The quantitative results are presented in Table 1, where compared to bNet, the CD error drops from 0.96 to 0.79 (i.e., 0.17 increasing). The improvement is consistent with the qualitative comparison in the left half of Fig. 3. It can be seen that guided by the tooth contour from the 2D SDF, bNet-P produces a tooth with correct root numbers and a reasonable shape.



(a) GT (b) bNet (c) bNet-P (d) GT (e) bNet-P (f) bNet-P-D

Fig. 3. The visual results of ablation experiments. The left half shows the results w/wo 2D SDF input, and the right half shows the results w/wo adversarial learning.



(a) dental crown (b) bNet-P-D (c) FullNet

Fig. 4. The visual comparison of tooth inpainting with (c) or without (b) the post-deformation, compared to (a) the input dental crown.

Benefits of Adversarial Learning. Adversarial learning plays an important role in our framework, to validate, we further augment bNet-P with adversarial learning (denoted as bNet-P-D) to validate its efficacy. Statistically, although bNet-P-D gains only a little improvement (e.g., 0.01) in terms of all three metrics as can be seen from Tab. 1, the visual results in the right half of Fig. 3 can reveal the benefit more clearly. Since the panoramic image only provides the projected contour shape, for molar teeth with three roots, only a tiny part of the third root can be seen due to occlusion. Thus, without adversarial learning to maintain the global tooth shape, bNet-P generates the unsatisfactory result (Fig. 3(e)) with irregular roots, especially the sunken third root. However, the root with a relatively small area contributes a tiny portion of the error. Instead, bNet-P-D produces almost perfect result (Fig. 3(f)).

Benefits of Post-deformation. To validate the effectiveness of the post-deformation step, we augment bNet-P-D with this step as our FullNet. As can be seen from Tab. 1, all the metrics are consistently improved, which is consistent with the visual results in Fig. 4, where the geometric details of the crown part are kept. The statistical and visual results demonstrate the efficacy of our FullNet, which offers the potential usability of our framework in real-world clinical scenarios. More results can be seen in the result gallery of Fig. 5.

3.2 Comparison with State-of-the-art Methods

Most of the state-of-the-art methods [2, 11, 1] adopt a template-fitting framework to reconstruct the complete tooth shape from the crown and panoramic image. To conduct a fair comparison, we implement [11] (denoted as TSEst) to present

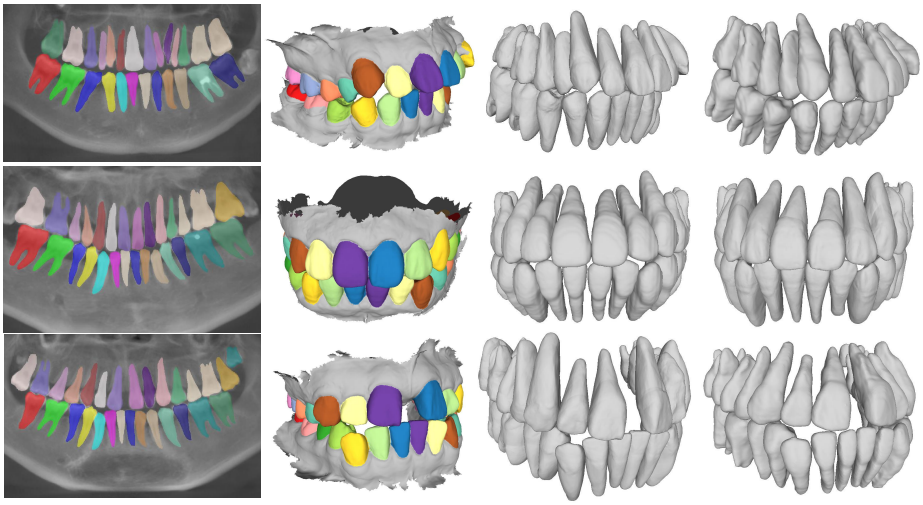


Fig. 5. Three typical examples of the tooth inpainting from the 2D panoramic image and 3D partial dental model.

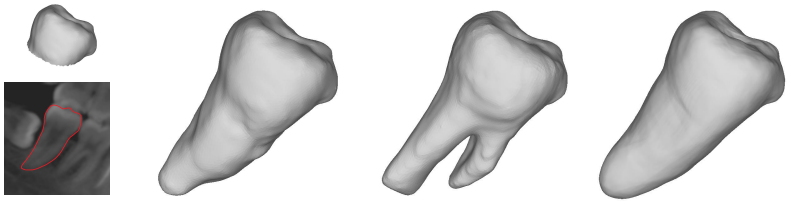


Fig. 6. Visual comparison with TSEst [11].

the visual results. Notably, although the overall tooth shape looks plausible, the inherent drawback of template fitting is that it cannot automatically deviate the template to match the input faithfully. As shown in Fig. 6, it happens to have a molar tooth with only one root as represented in the panoramic image. Not surprisingly, TSEst produces a tooth with two roots due to the pre-selected tooth template even if it disagrees with the contour shape. Instead, we automatically obtain a superior result similar to the ground truth.

4 Conclusion

In this paper, we proposed a neural solution to reconstruct a full tooth shape from a partial dental model and a panoramic image. Our method is fully automatic without any tooth template and user interaction. It produces promising results by first building a faithful complete tooth shape latent space and then projecting the inputs to find a full tooth shape code that respects the inputs. We have evaluated our approach both qualitatively and quantitatively, and compared it against state-of-the-art methods, where our approach produces superior results

and outperforms others. The outstanding results offer the potential usability of our framework in real-world clinical scenarios.

References

1. Barone, S., Paoli, A., Razionale, A.V., Savignano, R.: 3d reconstruction of individual tooth shapes by integrating dental cad templates and patient-specific anatomy. In: ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers Digital Collection (2014)
2. Barone, S., Paoli, A., Razionale, A.V.: Geometrical modeling of complete dental shapes by using panoramic x-ray, digital mouth data and anatomical templates. Computerized Medical Imaging and Graphics **43**, 112–121 (2015)
3. Cui, Z., Li, C., Chen, N., Wei, G., Chen, R., Zhou, Y., Wang, W.: Tsegnet: an efficient and accurate tooth segmentation network on 3d dental model. Medical Image Analysis p. 101949 (2020)
4. Cui, Z., Li, C., Wang, W.: Toothnet: Automatic tooth instance segmentation and identification from cone beam ct images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6368–6377 (2019)
5. Dai, A., Ruizhongtai Qi, C., Nießner, M.: Shape completion using 3d-encoder-predictor cnns and shape synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5868–5877 (2017)
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
7. Hu, K.S., Kang, M.K., Kim, T.W., Kim, K.H., Kim, H.J.: Relationships between dental roots and surrounding tissues for orthodontic miniscrew installation. The Angle Orthodontist **79**(1), 37–45 (2009)
8. Liang, Y., Song, W., Yang, J., Qiu, L., Wang, K., He, L.: X2teeth: 3d teeth reconstruction from a single panoramic radiograph. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 400–409. Springer (2020)
9. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. ACM siggraph computer graphics **21**(4), 163–169 (1987)
10. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 165–174 (2019)
11. Pei, Y., Shi, F., Chen, H., Wei, J., Zha, H., Jiang, R., Xu, T.: Personalized tooth shape estimation from radiograph and cast. IEEE transactions on biomedical engineering **59**(9), 2400–2411 (2011)
12. Sorkine, O., Cohen-Or, D., Lipman, Y., Alexa, M., Rössl, C., Seidel, H.P.: Laplacian surface editing. In: Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing. pp. 175–184 (2004)