Optimizing the CBCT Segmentation Pipeline with Intuition-Guided Processing

Qingyu Kuang[1,2][0009−0003−3830−0412]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

Abstract. In the past, general medical image models attracted considerable research interest. However, since medical imaging modalities vary widely and often fundamentally differ from RGB images, applying a general segmentation framework to specific tasks usually requires further optimization to achieve satisfactory performance. Cone-beam computed tomography (CBCT) is a commonly used medical imaging technique in dentistry. Optimizing the segmentation process for CBCT images can greatly enhance the effectiveness of computer-aided diagnostic systems in dental applications. In this work, we analyzed the ToothFairy3 dataset and proposed improvements to the nnU-Net framework. While preserving the auto-configuration capabilities of nnU-Net, we introduced targeted optimizations across the data preprocessing pipeline, network architecture, inference process, and postprocessing strategies to enhance performance for the CBCT multi-class segmentation task. Furthermore, the trained multi-class segmentation model can be integrated with user click prompts to train an interactive segmentation model. These modifications collectively reduced inference time, improved model effectiveness, and increased practical applicability. Code is available at https://github.com/kaoquanyu-for/formedseg.git.

Keywords: CBCT Segmentation · Deep Learning · Medical Images.

1 Introduction

Over the past decade, digital dentistry has advanced rapidly, with its key focus being the acquisition and segmentation of complete three-dimensional dental models and related structures. Currently, the mainstream technologies for obtaining 3D dental models mainly include intraoral scanning (IOS) or desktop scanning, and CBCT. Among these, intraoral or desktop scanning can conveniently capture the geometric morphology of the teeth crown surface but is limited to recording the external structure of teeth [6, 8]. In contrast, CBCT not only provides teeth surface information but also acquires internal 3D data such as jawbones, dental roots, and surrounding bone structures, offering greater advantages in clinical diagnosis and complex treatment planning [3]. As a result, it is widely used in oral and maxillofacial examinations and dental diagnostics [10]. However,

manual segmentation of CBCT images is time-consuming and demands specialized expertise and experience. Therefore, training deep learning-based models to automatically segment structures such as the maxillofacial bones and teeth in CBCT images can significantly streamline the process of diagnosis, evaluation, and surgical planning for dentists, while also providing critical references for applications such as dental crown design and the fabrication of surgical guides [11, 9].

However, current segmentation methods still face challenges due to variations in oral cavity opening states caused by different examination purposes, as well as considerable variations in morphological characteristics across different structures, which adversely affect segmentation accuracy and robustness. Examples of images with different oral cavity states from the public dataset ToothFairy3 [2, 1, 7] are illustrated in Fig. 1. Furthermore, the practical application of these models is limited by computational resources and time constraints, underscoring the gap that remains between theoretical research and clinical deployment.

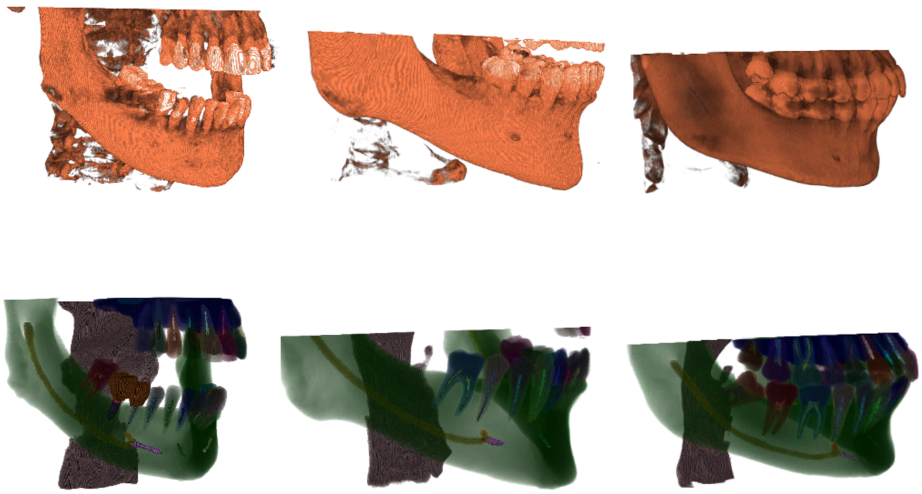

Fig. 1. The CBCT images and annotations in the Toothfairy3 dataset are shown in the figure. The top row displays the images under different oral cavity opening states, and the bottom row shows the corresponding annotated label images.

To address these issues, we re-examined the fundamental differences between CBCT images and RGB images. Inspired by human perceptual intuition for segmentation, we developed a novel processing pipeline for CBCT volume segmentation based on the nnU-Net [4] framework.

The main contributions of our work can be summarized as follows:

– Based on the nnU-Net framework, we introduced several modifications to enhance segmentation accuracy. These include adjusting the type and depth

of deep supervision loss computation, and adding category prediction heads to supervise the encoding process, thereby encouraging the extraction of more discriminative features.
– Furthermore, the trained multi-class segmentation model can be integrated with user click prompts to train a single-class segmentation model, enabling interactive prompt-based segmentation.
– The rules for constraining flip augmentation were adjusted to mitigate mis-segmentation caused by symmetrically similar structures in the images. And a novel augmentation method termed "tooth eraser" was introduced to increase data diversity.
– New designed post-processing workflow was optimized to align with the structural features of the segmentation output, balancing trade-offs between accuracy and inference time. And to enable large-patch inference, we optimized the inference process.

2 Methods

2.1 Overview

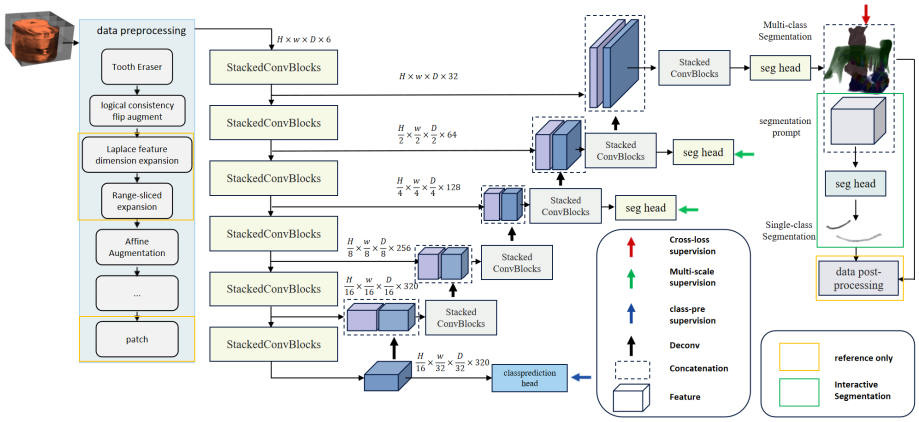

Fig. 2. Overview of the segmentation processing pipeline, illustrating the key stages and components of both the training and inference processes.

We have revisited the entire training and inference pipeline for medical image segmentation. Building upon the nnU-Net framework, we further improved a multi-class segmentation pipeline tailored for CBCT images, as illustrated in Fig. 2.

For the segmentation task on the ToothFairy3 dataset, the network utilized the architecture configuration derived from nnU-Net's automated parameter configuration, including the network depth and feature dimensionality across different layers. Building on this foundation, we introduced modifications to the data

preprocessing pipeline, network architecture, postprocessing strategies, and inference procedure to enhance its performance specifically for the ToothFairy3 segmentation task.

2.2 Dataset

The ToothFairy3 dataset comprises a large collection of 3D-annotated CBCT scans covering 77 anatomical structures that are highly relevant to orthodontics. In addition, the associated challenge not only focuses on segmentation accuracy but also incorporates inference efficiency as an evaluation metric and introduces an interactive segmentation task for the Inferior Alveolar Canal, addressing both automation and clinical needs. These features make ToothFairy3 particularly suitable for developing and evaluating segmentation models with strong clinical applicability.

The dataset contains 532 images, each with an isotropic resolution of 0.3 along all axes, but we manually selected 507 samples, discarding some extreme cases. The dataset contains three different sets, and their corresponding images are shown in Fig. 1.

2.3 Data Preprocessing

Tooth Eraser. Based on image characteristics, given that missing teeth are always present in the images, we randomly remove complete lower teeth without crowns in the images to enhance image diversity. Its effects are shown in Fig. 3.

Fig. 3. The visualization shows the processing effects of manually removing teeth according to image features.

Logical Consistency Flip Augment. Flipping augmentation of 3D images is commonly employed as a standard processing step to enhance data diversity. However, as noted in the article [5], when the segmentation targets include symmetrically similar structures, applying flipping augmentation without appropriate restrictions may cause confusion between these symmetrical parts in the images. This issue is particularly prominent in CBCT data, where multiple anatomical structures such as teeth and the inferior alveolar canal (IAC) display inherent symmetry. Moreover, when working with small patches, it becomes difficult to distinguish between upper and lower teeth based on structural features alone.

This issue is seldom encountered when processing images of other body parts, primarily for two reasons. First, most anatomical regions possess sufficient structural features with low morphological similarity between distinct structure. Second, in many cases there is no clinical need to differentiate between symmetric categories.

Intuitively, to mitigate the mis-segmentation caused by symmetrical structures, we diverged from the experimental setup described in article by retaining the flipping augmentation operation but imposing a key constraint: the number of flipping operations must always be even. When allowing flips along the x, y, and z axes, this means either performing no flips or flipping across an even number of axes.

The intuition behind this is straightforward: an odd number of flips results in a completely symmetrical version of the image, which can disrupt the perception of anatomical orientation. When the model is sufficiently complex, it may still learn to distinguish such flipped samples. However, both the model and human observers are likely to struggle when dealing with inherently symmetrical anatomical regions. The effects of applying different numbers of flipping operations are illustrated in Fig. 4.

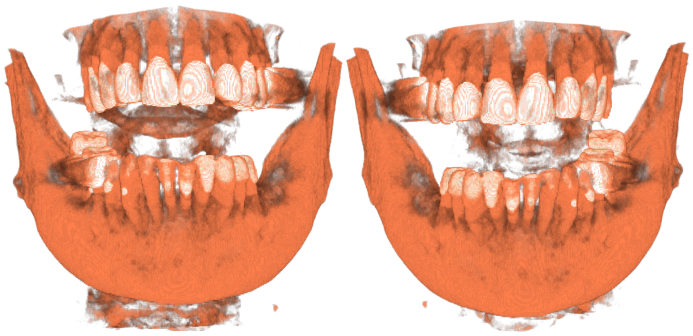

Fig. 4. The left figure shows the results of an even number of flipping operations, while the right figure displays the visualizations generated by an odd number of flipping operations.

Dimension Expansion. The imaging principle of CBCT differs significantly from that of RGB images, as its pixel values carry specific physical meanings. Since the ToothFairy3 dataset used in this study extends ToothFairy2 with additional annotation categories, we randomly sampled points across each category and recorded their Hounsfield Unit (HU) values in the ToothFairy2 dataset and visualized the statistics in Fig. 5. Based on this analysis, we divided the intensity values into multiple channels using the following intervals: [-600, 0], [0, 1000], [0, 2000], [1000, 3000], and [3000, maximum]. Different HU values may correspond to different tissue types, and in clinical practice, different intensity ranges are commonly used to capture images of specific tissues. Therefore, we analyzed the data ranges for each anatomical label and established the divisions described above. Given that teeth generally exhibit high HU values, the low-HU regions that are challenging to distinguish were split into multiple channels to provide the model with more detailed information. To enhance edge information, we computed a boundary channel by applying the Laplacian operator to the image restricted to the intensity range [-600, 3000], and incorporated it as an additional channel. At this stage, each channel is individually normalized to [0,1]. This multi-channel partitioning strategy constitutes one key component of our data preprocessing pipeline.

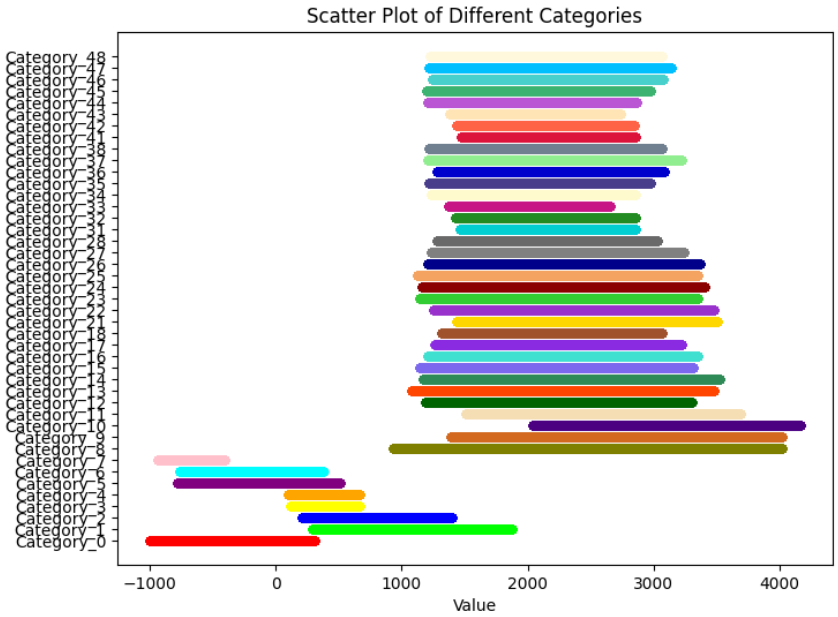

Fig. 5. The figure shows the Hounsfield Unit (HU) value ranges for different anatomical structures.

2.4 Network Architecture.

To enable accurate identification of different segmentation categories, we refined the deep supervision mechanism in the nnU-Net by reducing the number of supervised decoding layers from supervision at each stage to only the final three layers. Additionally, a category prediction head was added to the final encoder layer to perform 77-class prediction for each input image patch in the ToothFairy3 segmentation task. This design enhances discriminative feature learning through explicit category-wise supervision.

For this task, we employed a composite loss function consisting of cross-entropy loss, Dice loss, and focal cross-entropy loss. The i denotes the outputs at different levels of deep supervision. The formulation is as follows:

$$loss = \left(\sum_{i=0}^{2} \frac{1}{2^i} \left(l_{dice}^{i} + l_{fce}^{i} \right) \right) + l_{ice}^{class} \tag{1}$$

Dice loss and focal cross-entropy loss are computed for the outputs of the decoding stages at different levels, and cross-entropy loss is computed for the output of the category prediction head.

When training a single-class segmentation model with user click prompts, we froze the pretrained multi-class segmentation model and stacked the user click information with the multi-class inference outputs as an additional input channel to train a dedicated single-class segmentation head. When organizing the user click information, we only assign a value of 1 to the positions that the user clicked, and set all other positions to 0. At this stage, the model outputs predictions for only one class.

2.5 Model training

All experiments were conducted on 4 NVIDIA V100 GPUs (32 GB). During training, 20% of each set was used for validation. The models were trained for 30 epochs, with each epoch corresponding to a full traversal of all training data rather than random patch sampling. The training patch size was set to [160, 192, 192] and the batch size was set to 1. The model achieving the best performance on the validation set was retained.

For nnU-Net, we adopted the defaultUNet L configuration and further customized it. In addition to the preprocessing enhancements described above, RandAffine augmentation was applied. The model was optimized using AdamW with an initial learning rate of 1×10^{-4} and a ReduceLROnPlateau learning rate scheduler. We use a learning rate scheduler with a reduction factor of 0.8. The learning rate will not decrease below 1×10^{-6}, and the scheduler monitors the validation metric at every epoch.

2.6 Inference.

Due to the substantial computational requirements of training 3D data and the constraints on inference time and computing resources imposed by the Tooth-

Fairy3 challenge, we intuitively reasoned that increasing the patch size, during training and inference, could serve as an effective way to mitigate mis-segmentation in symmetrically similar structures. Although limited computational resources restricted the training patch size to [160, 192, 192], we re-examined the inference functions in both nnU-Net and MONAI and implemented strict memory management on the GPU during inference. This approach allowed the use of patch sizes consistent with those used in training, reduced the number of patches requiring inference under the same overlap ratio setting, shortened inference time, and ultimately improved the practical usability of the model.

Table 1. Inference time for different image sizes.

Image size	Inference Repeat (%)	Patch Num	Inference time(s) (without argmax and Post-processing)
262, 512, 512(F_001)	25	32	41
170, 352, 370(P_001)	25	18	19
188, 385, 462(S_0001)	25	18	17

Table 1 shows the inference time required on an RTX 4060 GPU(8GB). By keeping only one patch and the model in GPU memory at a time, memory consumption is relatively low. This enables the use of a larger patch size and significantly reduces the number of patches needed for inference. As tested, the patch size can be increased to [192, 192, 192].

Post-processing and the argmax operation are configured to run on the CPU. As their execution time is strongly affected by system RAM, these steps are excluded from the reported runtime.

2.7 Data Post-processing.

Determining the optimal post-processing strategy in nnU-Net requires repeated inference across the entire dataset to evaluate the retention of the largest connected component for each category, which is a highly time-consuming process. Based on the structural characteristics of the data, we employed a hybrid strategy combining projection-based connected region preservation with 3D connected-component labeling.

Compared to nnU-Net's automated strategy, our method requires configuration based on observational analysis of the dataset, but it reduces inference time and avoids repeated post-processing selection across different combination schemes. The specific procedure is shown in Fig. 6.

It can be observed that most segmentation errors in the model's output occur in the maxilla, mandible, and pharyngeal regions. During multi-class segmentation, we first retain the largest 3D connected components of the pharynx and mandible. The labels are then projected along the z-axis, and the two largest connected regions in the projection are preserved. These are back-projected into 3D

to remove erroneous segmentation areas outside the main anatomical structures, yielding the final segmentation result after post-processing.

When processing the output of a single-class segmentation, we assign labels to the results based on user prompts, as also illustrated in Fig. 6.

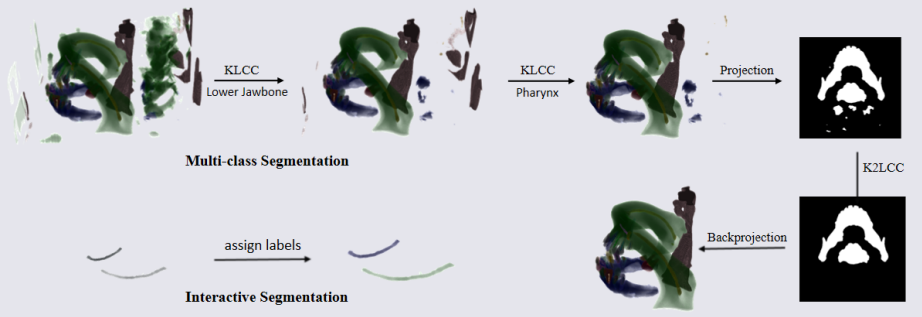

Fig. 6. The figure illustrates the post-processing pipeline for both multi-class and single-class segmentation. KLCC stands for "Keeping the Largest Connected Component".

3 Results

3.1 Validate the effectiveness of the flip constraint

We used a portion of the data as a validation set. For each patch, we performed inference to validate the Dice coefficient, then averaged the results. During training, all models were trained for the same duration, and the best performance results on the test set were recorded. The outcomes are presented in Table 2. Demonstrates the effectiveness of the constraint rules.

Table 2. Results of different constraint rules.

Model	Dice
default	0.9288
constraint	0.9409

3.2 Validation Results

The inference results for the three different sets are presented in Fig. 7.

On the current challenge leaderboard, the results from my final submission show that in the multi-instance segmentation task, the maximum Dice score reached 0.82, while the minimum was 0.14.

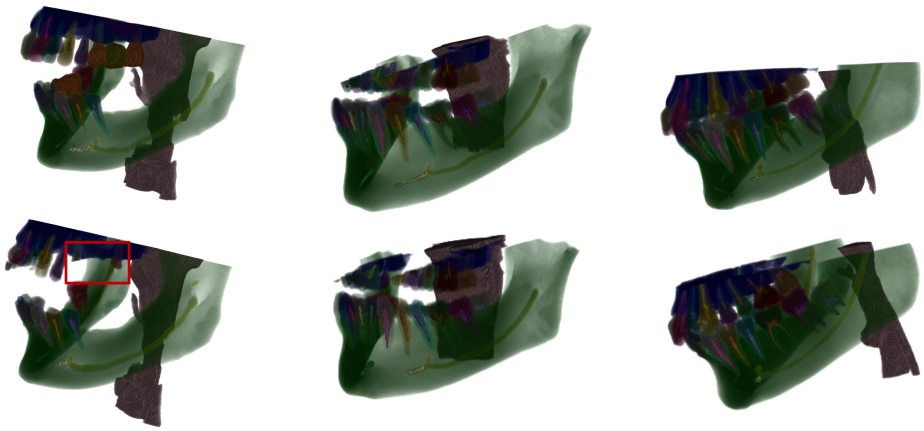

Fig. 7. The top row shows the labels, and the bottom row shows the model's inference results, corresponding from left to right to F_001, P_001, and S_0001 in ToothFairy3.

Similarly, for the interactive segmentation task, the highest Dice score achieved was 0.92, although several cases completely failed, producing no segmentation output whatsoever.

The best results of both tasks in the test phase are summarized in Table 3. The overlap ratio in the inference function was set to 25% for the multi-instance segmentation task and 50% for the interactive segmentation task. And our inference speed was ranked second in the interactive segmentation task.

Table 3. Evaluation results of both tasks in test phase. Dice similarity coefficient and HD95 are reported.

Metric	Statistic	Multi-class	Interactive
Dice Average	Min	0.149	0.0
	25%	0.516	0.66
	50%	0.652	0.83
	75%	0.718	0.88
	Max	0.820	0.92
	Mean	0.594	0.72
	Std	0.185	0.25
HD95 Average	Min	78.40	1.0
	25%	107.09	1.73
	50%	142.38	2.80
	75%	185.98	45.11
	Max	358.28	607.86
	Mean	163.31	76.82
	Std	69.14	159.13

Inspection of the inference results shows that certain data setup errors remain to be resolved, since the model identify the implants and the maxillary sinus. It remains a challenge for the model to accurately distinguish between dental crowns and bridges. In addition, in interactive segmentation tasks, the model may fail when processing certain images. This failure is possibly caused by variations in oral cavity opening states across images or by overfitting, which prevents the model from generalizing to different conditions. However, preliminary experiments have already achieved promising progress in both multi-class segmentation and interactive segmentation, providing valuable insights and ideas for further studies.

Acknowledgments. We would like to express our sincere gratitude to the organizers of the ToothFairy Challenge for their continuous efforts in creating and updating the publicly available CBCT datasets. Their dedicated work has made this study possible.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bolelli, F., Lumetti, L., Vinayahalingam, S., Di Bartolomeo, M., Pellacani, A., Marchesini, K., van Nistelrooij, N., van Lierop, P., Xi, T., Liu, Y., Xin, R., Yang, T., Wang, L., Wang, H., Xu, C., Cui, Z., Wodzinski, M., Müller, H., Kirchhoff, Y., Rokuss, M.R., Maier-Hein, K., Han, J., Kim, W., Ahn, H.G., Szczepański, T., Grzeszczyk, M.K., Korzeniowski, P., Caselles-Ballester, V., Paolo Burgos-Artizzu, X., Prados Carrasco, F., Berge', S., van Ginneken, B., Anesi, A., Grana, C.: Segmenting the inferior alveolar canal in cbcts volumes: The toothfairy challenge. IEEE Transactions on Medical Imaging **44**(4), 1890–1906 (2025). https://doi.org/10.1109/TMI.2024.3523096
2. Bolelli, F., Marchesini, K., van Nistelrooij, N., Lumetti, L., Pipoli, V., Ficarra, E., Vinayahalingam, S., Grana, C.: Segmenting maxillofacial structures in cbct volumes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5238–5248 (2025)
3. Bornstein, M.M., Scarfe, W.C., Vaughn, V.M., Jacobs, R.: Cone beam computed tomography in implant dentistry: a systematic review focusing on guidelines, indications, and radiation dose risks. International Journal of Oral & Maxillofacial Implants **29**(Suppl), 55–77 (2014)
4. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
5. Isensee, F., Kirchhoff, Y., Kraemer, L., Rokuss, M., Ulrich, C., Maier-Hein, K.H.: Scaling nnu-net for cbct segmentation. In: Supervised and Semi-supervised Multi-structure Segmentation and Landmark Detection in Dental Data. pp. 13–20. Springer, Cham (2025)
6. Joda, T., Ferrari, M., Gallucci, G., Wittneben, J., Brägger, U.: Digital technology in fixed implant prosthodontics. Periodontology 2000 **73**(1), 178–192 (2017)
7. Lumetti, L., Pipoli, V., Bolelli, F., Ficarra, E., Grana, C.: Enhancing patch-based learning for the segmentation of the mandibular canal. IEEE Access **12**, 79014–79024 (2024). https://doi.org/10.1109/ACCESS.2024.3408629

8. Mangano, F., Gandolfi, A., Luongo, G., Logozzo, S.: Intraoral scanners in dentistry: a re-view of the current literature. BMC Oral Health **17**(1), 149 (2017)
9. Revilla-León, M., Gómez-Polo, M., Vyas, S., Barmak, B.A., Gallucci, G.O., Att, W., Krishnamurthy, V.R.: Artificial intelligence applications in implant dentistry: a systematic review. Journal of Prosthetic Dentistry **129**(2), 293–300 (2023). https://doi.org/10.1016/j.prosdent.2021.05.008
10. Scarfe, W.C., Angelopoulos, C.: Maxillofacial Cone Beam Computed Tomography: Principles, Techniques and Clinical Applications. Springer, Cham (2018)
11. Torosdagli, N., Liberton, D.K., Verma, P., Sincan, M., Lee, J.S., Bagci, U.: Deep geodesic learning for segmentation and anatomical landmarking. IEEE Transactions on Medical Imaging **38**(4), 919–931 (2019)