치의과학박사 학위논문

Automated Dental Image Analysis Using Deep Neural Networks

심층신경망을 이용한 자동화된 치과 의료영상 분석

2021년 8월

서울대학교 치의학대학원

치의과학과 치과보철학 전공

차 준 영

Automated Dental Image Analysis Using Deep Neural Networks

지도교수 한 중 석

이 논문을 치의과학박사 학위논문으로 제출함
2021년 6월

서울대학교 치의학대학원
치의과학과 치과보철학 전공
차 준 영

차준영의 박사 학위논문을 인준함
2021년 7월

위 원 장 _____

부위원장 _____

위 원 _____

위 원 _____

위 원 _____

Abstract

Purpose: In dentistry, deep neural network models have been applied in areas such as implant classification or lesion detection in radiographs. However, few studies have applied the recently developed keypoint detection model or panoptic segmentation model to medical or dental images. The purpose of this study is to train two neural network models to be used as aids in clinical practice and evaluate them: a model to determine the extent of implant bone loss using keypoint detection in periapical radiographs and a model that segments various structures on panoramic radiographs using panoptic segmentation.

Methods: Mask-RCNN, a widely studied convolutional neural network for object detection and instance segmentation, was constructed in a form that is capable of keypoint detection, and trained to detect six points of an implant in a periapical radiograph: left and right of the top, apex, and bone level. Next, a test dataset was used to evaluate the inference results. Object keypoint similarity (OKS), a metric to evaluate the keypoint detection task, and average precision (AP), based on the OKS values, were calculated. Furthermore, the results of the model and those arrived at by a dentist were compared using the mean OKS. Based on the detected keypoint, the peri-implant bone loss ratio was obtained from the radiograph.

For panoptic segmentation, Panoptic DeepLab, a neural network model ranked high in the previous benchmark, was trained to segment key structures in panoramic radiographs: maxillary sinus, maxilla, mandibular canal, mandible, natural tooth, treated tooth, and dental implant. Then, each evaluation metric of panoptic, semantic, and

instance segmentation was applied to the inference results of the test dataset. Finally, the confusion matrix for the ground truth class of pixels and the class inferred by the model was obtained.

Results: The AP of keypoint detection for the average of all OKS thresholds was 0.761 for the upper implants and 0.786 for the lower implants. The mean OKS was 0.8885 for the model and 0.9012 for the dentist; thus, the difference was not statistically significant ($p = 0.41$). The mean OKS of the model was in the top 66.92% of the normal distribution of human keypoint annotations.

In panoramic radiograph segmentation, the average panoptic quality (PQ) of all classes was 80.47. The treated teeth showed the lowest PQ of 57.13, and the mandibular canal showed the second lowest PQ of 65.97. The Intersection over Union (IoU) was 0.795 on average for all classes, where the mandibular canal showed the lowest IoU of 0.639, and the treated tooth showed the second lowest IoU of 0.656. In the confusion matrix, the proportion of correctly inferred pixels among the ground truth pixels was the lowest in the mandibular canal at 0.802. The AP, averaged for all IoU thresholds, was 0.316 for the treated tooth, 0.414 for the dental implant, and 0.520 for the normal tooth.

Conclusion: Using the keypoint detection neural network model, it was possible to detect major landmarks around dental implants in periapical radiographs to a degree similar to that of human experts. In addition, it was possible to automate the calculation of the peri-

implant bone loss ratio on periapical radiographs based on the detected keypoints, and this value could be used to classify the degree of peri-implantitis. In panoramic radiographs, the major structures including the maxillary sinus and the mandibular canal could be segmented using a neural network model capable of panoptic segmentation. Thus, if deep neural networks suitable for each task are trained using suitable datasets, the proposed approach can be used to assist dental clinicians.

Keyword : keypoint detection; panoptic segmentation; machine learning; deep learning; medical image analysis
Student Number : 2017-33206

Contents

Abstract i

Chapter 1. Introduction 1

Chapter 2. Materials and methods 5

Chapter 3. Results 23

Chapter 4. Discussion 32

Chapter 5. Conclusions 45

Published papers related to this study 46

References 47

Abbreviations 52

Abstract in Korean 53

Acknowledgements 56

List of Tables

[Table 1] 23
[Table 2] 25
[Table 3] 28
[Table 4] 30
[Table 5] 30
[Table 6] 31

List of Figures

[Figure 1] 6
[Figure 2] 9
[Figure 3] 14
[Figure 4] 16
[Figure 5] 18
[Figure 6] 24
[Figure 7] 26
[Figure 8] 27
[Figure 9] 29

1. Introduction

As computing power has increased with the development of hardware such as GPUs, it has become possible to train deep neural networks composed of a large number of parameters. Accordingly, in recent years, the field of machine learning has seen rapid advances. In particular, after AlexNet[1] won the ImageNet Large Scale Visual Recognition Challenge[2] in 2012, deep learning achieved considerable success in the field of computer vision. As deep neural networks, especially CNN, can successfully classify general images, numerous CNNs have been developed and applied to medical images. Several studies have focused on using CNNs for the binary classification of radiographic images, such as pulmonary tuberculosis[3], osteoporosis[4], or periodontal bone loss[5] to aid clinicians in diagnosis.

In recent years, neural networks have been developed not only for classifying images but also for recognizing various objects or regions within an image. Object detection involves localizing each object, while instance segmentation involves detection and segmentation of the object. Some CNNs were developed for object detection: a series of neural networks based on the region-based CNN (R-CNN)[6-8], You Only Look Once (YOLO)[9], and Single Shot MultiBox Detector (SSD)[10] can detect objects by predicting the bounding boxes around each object and predict the class of the object simultaneously. Furthermore, Mask R-CNN, which uses a modified architecture of R-CNN[6], can predict segmentation masks or keypoints within the bounding boxes.[11]

Studies on neural network models for semantic segmentation, which involves classifying each pixel of an image into various classes, have also been conducted. Fully convolutional networks[12] or u-shaped networks[13] have been developed.

Several studies have sought to apply these methods on medical images: the application of semantic segmentation has been attempted in some studies[14-16] where each pixel in the image is classified, while object detection[17, 18] or instance segmentation[19] have been tried in others. However, little research has been conducted on detecting specific points in medical and dental images. To the best of our knowledge, the detection of individual dental implants and location of important landmarks, such as the implant marginal bone level, using fully end-to-end deep learning methods has not yet been attempted. In general, the identification of the peri-implant marginal bone level in conventional radiographs is difficult because it involves comprehending the three-dimensional bone shape based on a two-dimensional image[20, 21]. Often, the boundaries of the bones around the implant are obscure or the heights of the buccal and lingual bone levels are different.

Therefore, the first aim of the present study is to address this lacuna in research. We employed a deep learning model, namely, the Mask R-CNN, for localizing the implants and finding keypoints within the detected implant site on periapical radiographs. Based on the results, the marginal bone loss ratio was calculated and the corresponding classification was performed. Such a classification may assist dentists in the analyses of periapical radiographs.

Several studies on applying CNNs to radiographs have dealt with panoramic radiographs.[22] Semantic segmentation[23], object detection[24-26], and instance segmentation[27] have been applied to dental panoramic radiographs.

Panoramic radiographs are commonly used during conventional dental treatment. It would be useful to automatically detect and segment various structures belonging to multiple categories on dental panoramic radiographs for the diagnosis and treatment planning of various diseases. To the best of our knowledge, the use of a machine

learning method, including deep neural networks, for this kind of task has not been studied in the dental field.

Detecting and segmenting various kinds of structures in panoramic radiographs is a complex task because structures of various sizes and shapes overlap each other. In some cases, a ghost or double image interferes with the identification of a specific structure.

Furthermore, the boundaries of some important structures, such as the mandibular canal or maxillary sinus, are difficult to distinguish in a panoramic radiograph. For a successful dental implant surgery, it is important to identify the boundaries of the maxillary sinus and mandibular canal to determine the surgical method and the type of implant that are most suitable. Nonetheless, no previous studies have attempted to automatically segment these structures on a panoramic radiograph.

Therefore, the second aim of this study is to automatically segment various types of structures, including the maxillary sinus and mandibular canal, in an orthopantomogram. The following classes were detected and segmented: maxillary sinus, maxilla, mandibular canal, mandible, normal tooth, treated tooth, and dental implants.

To do this, a new concept called "panoptic segmentation," which was recently proposed for integrating multiple tasks in computer vision, was applied.[28] Panoptic segmentation combines semantic segmentation and instance segmentation and involves predicting not only the semantic label but also the instance id for each pixel. Regions that are not countable, such as grass or road, must be segmented using semantic label classification on each pixel as in semantic segmentation. By contrast, countable objects, such as a person or car, must be segmented using both semantic label and instance id classification on each pixel, which produces results similar to those of instance segmentation.

Panoptic segmentation is one of the most challenging tasks in the field of computer vision; hence, a recently developed and verified deep convolutional neural network model designed for panoptic segmentation was adopted[29]. The proposed deep learning–based automatic method might assist dental practitioners in diagnosing various oral and maxillofacial diseases and developing appropriate treatment plans.

2. Materials and methods

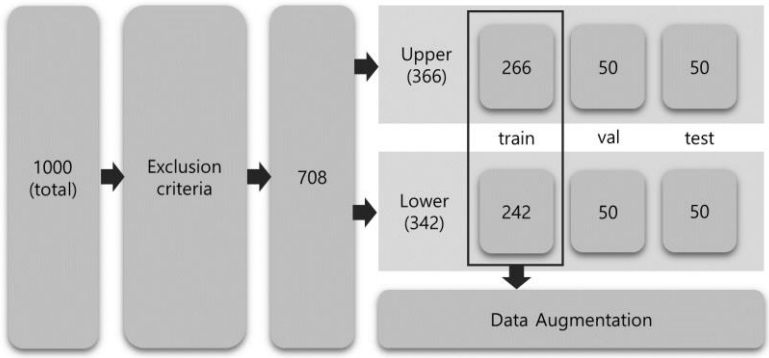
2.1 Ethics Statement

This study was conducted in accordance with the deliberation results by the Institutional Review Board of Seoul National University Dental Hospital, Dental Life Science Research Institute (IRB No. ERI20024, Notification of deliberation exemption in August 2020). Ethical review, approval, and patient consent were waived for this study owing to the following reasons: this study is not a human subjects' research project as specified in the Bioethics and Safety Act; it is practically impossible to obtain the consent of the research subjects in this study; and the risk to the subjects is extremely low because of the retrospective nature of the study.

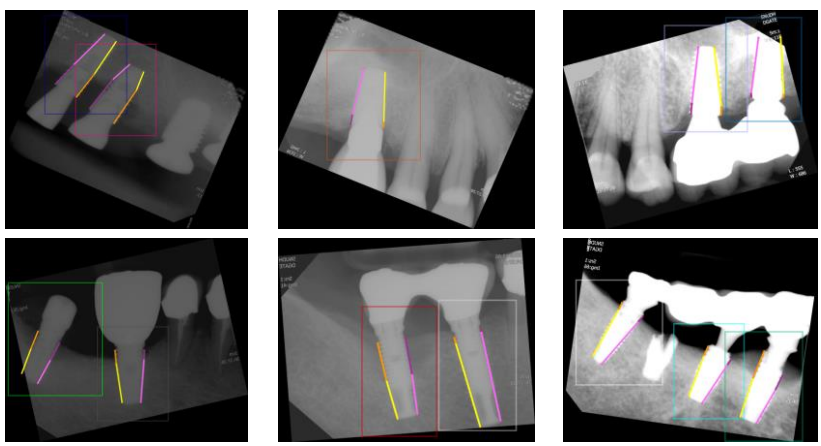
2.2 Keypoint detection on periapical radiographs

2.2.1 Datasets

In order to create datasets, 1000 periapical radiographic images were obtained between December 2018 and June 2020 from the Picture Archiving and Communication System (PACS) in Seoul National University Dental Hospital. Among these, radiographs that were not properly captured in parallel or those that were out of focus were excluded. Furthermore, if the graft material hindered the correct observation of the alveolar bone in a radiograph, it was excluded. Ultimately, 292 periapical radiographs were excluded. The remaining 708 images were separated into upper and lower periapical radiographs. For each of the datasets, the images were further divided into training, validation, and test datasets. The overview of the datasets is shown in Figure 1.



(a)



(b)

Figure 1. (a) Overview of the datasets. After the exclusion criteria were used to remove certain data, the remaining data were separated into training, validation, and test datasets. The training dataset was used for training the model, while the validation dataset was used for assessing overfitting. The test dataset was used for evaluation. The numbers given are the number of periapical radiographs. The numbers of upper and lower periapical radiographs are specified separately. (b) Visualized examples of the data augmentation process including random horizontal flip, rotation, contrast and brightness shift.

Training data were enriched by randomly generated data augmentation, leveraged with a horizontal flip, rotation, contrast, and brightness shift. The size of the augmented dataset is not fixed, as the data augmentation process used in this study could potentially generate an infinite number of modified input images. However, based on an analysis of the evaluation result of the validation dataset, the training was stopped after 128,000 iterations to avoid overfitting, with the learning rates ranging between 0.0005-0.00005. Implant bounding boxes and keypoint annotations for ground truth were performed by a dental practitioner (the author). An oral and maxillofacial radiologist reviewed and confirmed the annotation results.

2.2.2 Neural Network Model Architecture

For better results in the prediction of the landmarks around the implants, we prepared three separate neural networks and connected them. The first is for identifying whether the image shows the upper or lower jaw. After the radiographic image was classified, it was fed into one of the other two networks, which are responsible for either the upper or lower implants. These two parallel networks apply the core logic, detect implants using bounding box regression, and predict landmarks. They each have the same architecture but were separately trained for upper and lower implants.

2.2.2.1 Upper and Lower Implant Classifier

Based on a 152-layer ResNet[30], a classification model was created for sorting upper and lower periapical radiographs. Weights pretrained on ImageNet[31] were used, but the last fully connected layer was switched so that there were two final output nodes, i.e. one for each of the classes (upper and lower implants).

2.2.2.2 Mask R-CNN with Keypoint Head

After the 152-layer model classified whether the given periapical radiograph contained an image of the upper or lower jaw, the radiograph image was fed into another model that was trained specifically for upper or lower implants.

At this stage, individual implants are detected using bounding boxes. Based on the detected region, the six keypoints, including mesial and distal marginal bone level, were predicted.

For this procedure, a modified R-CNN architecture, Mask R-CNN, was used. Mask R-CNN, the latest descendant of the R-CNN model, comprised a "backbone" and "heads".[11] The backbone network is a CNN that outputs feature maps from the original input image. Among various options, the FPN[32] based on ResNet,[30] known for robust results when used for Mask R-CNN, was adopted in this study.

Using the feature maps from the backbone network, the box head performed object classification and bounding box regression and the mask head performed object segmentation. By attaching a keypoint detection head and properly training the network, the model can predict specific keypoints on the objects that were detected by the box head. As demonstrated in a previous study, this method with a keypoint head can be used for human pose estimation, wherein the model picks some keypoints of the human body, such as eyes, elbows, and knees.[11] In the present study, we adopted this architecture, i.e., the Mask R-CNN based on ResNet-FPN backbone with a keypoint detection head. The scheme of the model excluding the upper and lower jaw classification is shown in Figure 2.

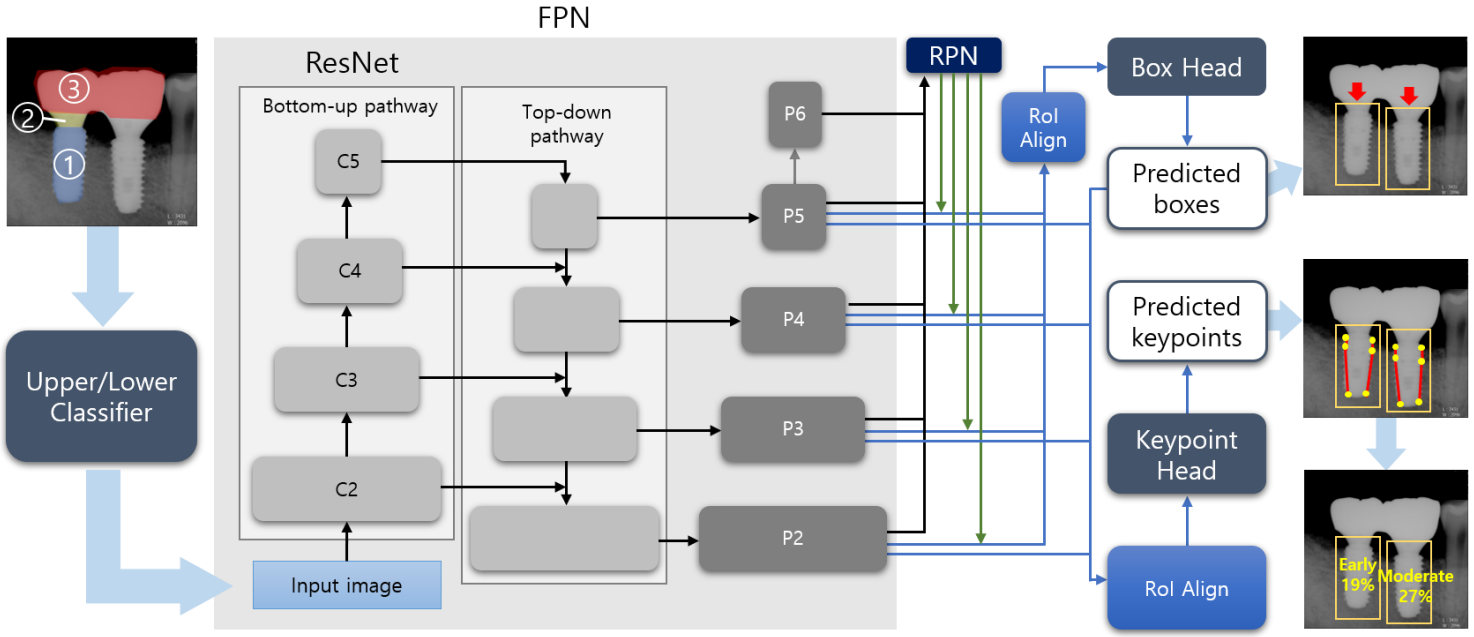


Figure 2. Architecture of the model used in this study. The FPN was constructed with the ResNet as a backbone CNN. The RPN takes feature maps from the FPN and proposes RoI. The box head further refines the proposals and predicts final bounding boxes (red arrows). In addition, the keypoint head localizes the keypoints (yellow dots in the middle of the radiographs on the right) based on the predicted bounding boxes. C2–5 and P2–6 denotes the output feature maps of the ResNet and FPN, respectively. Implant (①), Abutment (②) and Superstructure (③) are shown in the left radiograph.

The model was trained and tested for detecting implants and locating the six positions of the detected implant on dental periapical radiographs. The six keypoints were the right and left sides of the peri-implant bone level, the implant apex, and the implant top. To cover various types of implants, the most coronal thread was annotated as the top of the implant. We refer to these six positions as "keypoints" because it is a widely used term in point-detecting tasks, such as human pose estimation[33, 34] or facial keypoint detection[35].

2.2.2.3 Bone Loss Ratio and Classification

Some studies on classification systems for peri-implantitis use radiographic bone loss along with clinical indicators, such as bleeding/suppuration on probing or probing depth.[36-38] They also use the ratio of the radiographic bone loss over the total implant length to classify the peri-implantitis. Based on their suggested criteria, we calculate and classify the bone loss ratio so that dental practitioners can easily refer to it.

Using the coordinates of the six keypoints obtained from the prediction, the total length of the implant and the implant length that is not surrounded by sound bone can be calculated. The total length is measured from the center of the apex to the center of the implant top, and the length corresponding to the radiographic bone loss is measured from the center of the implant top to the center of the two marginal bone level keypoints. From these values, the percentage of the implant length in the bone defect site over the total length is calculated.

Based on this percentage, the severity of the bone loss around the implant is classified into one of four groups: normal, if the percentage is $\leq$10%, early, if the percentage is >10% and $\leq$25%, moderate, if the percentage is >25% and $\leq$50%, and severe, if the

percentage is >50%.

2.2.3 Evaluation Methods

As the prediction process comprises two phases, i.e., implant bounding box regression and implant keypoint localization, two different metrics are used for evaluating each task.

2.2.3.1 Intersection over Union (IoU)

For evaluating the model's performance of detecting implants, a metric that can measure how close the model's bounding box is to the ground truth bounding box is needed. IoU, also known as the Jaccard index, is used as the requisite metric. IoU is calculated by dividing the overlapping area of the ground truth box (A) and the model-predicted box (B) by the total area of the coverage of the two boxes.

$$\mathrm{IoU} = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

At various values of the IoU thresholds, the model's AP and AR can be obtained.

2.2.3.2 Object Keypoint Similarity (OKS)

To evaluate the model's keypoint detection performance, OKS[39] is used as an analogous option to IoU. OKS is calculated for each object, and it ranges from 0 to 1. The value tends closer to 1 as the model's prediction approaches the ground truth. This metric can be used in a manner similar to IoU, which is generally used for evaluating object detection tasks. The OKS for each implant is defined as follows.

$$\mathrm{OKS}_j = \frac{\sum_i \left[exp\left(\frac{-d_{ji}^2}{2s_j^2 k_i^2}\right) \delta(v_{ji} > 0) \right]}{\sum_i \delta(v_{ji} > 0)} \quad (2)$$

In equation (2), j represents each individual implant and i corresponds to each keypoint type. In the present study, the i represents the Lt. bone level, Rt. bone level, Lt. apex, Rt. apex, Lt. implant top, and Rt. implant top. Furthermore, v denotes the visibility flags (v = 0: not labeled, v = 1: labeled but not visible, and v = 2: labeled and visible) of the ground truth. For each implant, the ground truth and predicted keypoints have the form [x1, y1, v1, ⋯, x6, y6, v6], where x and y are the keypoint locations and v is a visibility flag.

Consider a vector $\vec{d}_{ji}$ that starts from a ground truth keypoint and ends at the detected keypoint. Variable d_{ji} represents the distance between the ground truth keypoint and the detected keypoint. Furthermore, s_j is the scale of the implant j and is defined as the square root of the ground truth segmented area of the implant.

k_i can be regarded the per−keypoint standard deviation but multiplied by some constant, which is 2 here $(k_i = 2\sigma_i)$. To obtain the per−keypoint standard deviation σ_i, standardized to implant scale s, redundantly annotated images in the validation dataset were used to calculate $\sigma_i^2 = \mathrm{E}_{(j)}\left[d_{ji}^2/s_j^2\right]$. Here, $\mathrm{E}_{(j)}$ represents an average over j. As the mean of $\frac{\vec{d}_{ji}}{\mathrm{s_j}}$ over j becomes a zero vector, the per−keypoint standard deviation σ_i can be obtained by calculating the mean of d_{ji}^2/s_j^2 over j.

OKS can be used as a threshold when determining precision and recall based on keypoint detection. Among the keypoint−detected implants, only those whose OKS values are higher than the OKS threshold are considered true positives. Using different settings of

the OKS thresholds, the corresponding precision–recall curves as well as AP and AR can be obtained.

2.2.3.3 Mean OKS

To compare the prediction result of the model with results of humans, all the OKS values of detected implants were averaged to calculate the mean OKS. While the precision–recall graphs reflect the model's prediction confidence scores, the mean OKS does not include the information of various confidence scores. When comparing with a human, confidence scores cannot be used unless the human who performs the detection task provides specific confidence scores in a similar way to the model. Instead, only one threshold value of 0.7 was used for implant detection, and only the detections that had a confidence score above it were regarded valid when calculating the mean OKS. To verify the validity of the model, this metric was used to compare its performance to that of a dentist.

2.2.4 Keypoint Heatmap Visualization

The keypoint detection output of the deep learning model used in this study comprises six points. These points are determined by selecting the highest logit of the neural network's output. By converting the logit values to colored keypoint heat maps, each pixel's likelihood of being a keypoint can be visualized. Hence, the pixels that were given high scores by the model become easier to find. The keypoint logits were converted to values between 0.0 and 1.0, which were consequently assigned to a specific color. Examples of the keypoint heat maps are shown in Figure 3.

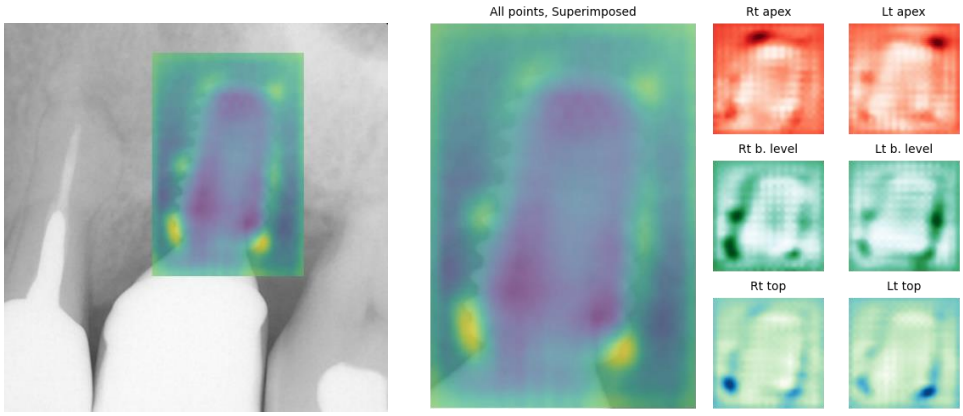


Figure 3. Examples of the keypoint heat maps. Each pixel's likelihood of being a keypoint is mapped into the heat map. Heat maps that combine all keypoints are superimposed on the original radiograph. Individual heat maps for each corresponding keypoint are shown.

2.2.5 Statistical Analysis

Varying the threshold for the model's confidence scores on bounding box regression with different IoU thresholds, the AP and AR on implant detection were obtained (IoU threshold 0.50-0.95, increased by 0.05). In addition, using different thresholds for the model's confidence scores on keypoint detection with different OKS thresholds, the corresponding precision-recall curves and the AP and AR were obtained (OKS threshold of 0.50-0.95, and increased by 0.05). An independent t-test was used to compare the mean OKS values between a dentist and our model with the total, upper jaw, and lower jaw dataset. p-value<0.05 was considered to be statistically significant. The calculations of the AP, AR, precision-recall curves, and the independent t-test results were performed using Python (Python 3.6.9, Python Software Foundation). The software code used for the evaluation was modified from an open-source project created during previous research.[40]

2.3 Panoptic segmentation on panoramic radiographs

2.3.1 Datasets

Ninety dental panoramic radiographs were obtained from the PACS in the Seoul National University Dental Hospital. In some instances, all the regions could not be annotated accurately; for example, the boundary of the medial wall of the maxillary sinus might be unclear, the border of the mandibular canal might not be visible, or the patient might not be properly positioned inside the focal trough. Such panoramic radiographs were excluded from the study. Similarly, radiographs of patients who had undergone unusual treatments, such as mandibular reconstruction, or those containing radiopaque materials that hindered the discrimination of the jaw structures, were excluded.

Finally, 51 panoramic radiographs were separated into three groups: training (n = 30), validation (n = 11), and test (n = 10) datasets. The ground truths were annotated by a dental practitioner (the author), while an oral and maxillofacial radiologist reviewed, corrected, and confirmed the annotations. Visualized examples of the annotations are shown in Figure 4.

For panoptic segmentation, the classes that are subjected to semantic segmentation are referred to as "stuff," whereas those subjected to instance segmentation are referred to as "thing".[28] The eight classes (five stuff and three things) included in the current study were as follows: maxilla, maxillary sinus, mandible, mandibular canal, normal tooth, treated tooth, dental implant, and unlabeled. Among these, normal tooth, treated tooth, and dental implant were assigned as things, so all objects in these classes were segmented individually.

Some classes were categorized to help understand the results easily: the maxilla and mandible were categorized as "bone,"

whereas the normal tooth, treated tooth, and dental implant were categorized as "tooth." The maxillary sinus and mandibular canal were not categorized because their morphology is not similar to the other classes.

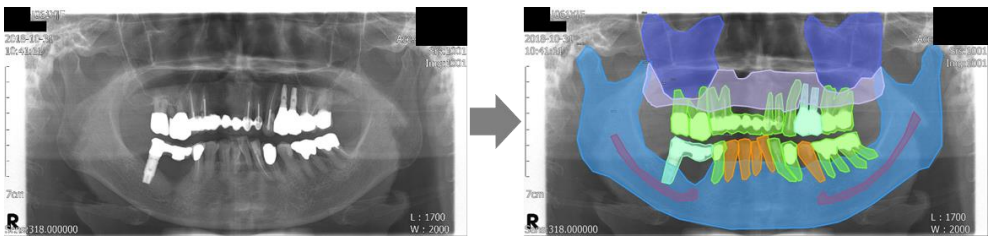


Figure 4. Visualized examples of the annotation results. Eight classes, including a background class, were used. Semantic segmentation was applied to four classes: maxillary sinus, maxilla, mandibular canal, and mandible. Instance segmentation was applied to three classes: normal tooth, treated tooth, and dental implant.

2.3.2 Neural Network Model Architecture

Panoptic segmentation is challenging and has not been applied to panoramic radiographs; therefore, it is important to use a high-performance artificial neural network model. The results of the Cityscapes dataset benchmark[41] were investigated to select the appropriate state-of-the-art deep neural network. Based on the evaluation results of various models, a high-performance model, Panoptic DeepLab[29], was selected. It comprises a semantic segmentation branch and an instance segmentation branch, both of which share the same encoder as the backbone. The instance segmentation branch predicts the center of the mass for each instance and the offset vector, which starts from each foreground pixel and points to the corresponding center of mass.

Based on the center and the offset vector, the instance id of each

foreground pixel can be determined. Each pixel is relocated by its offset vector and the distance between the predicted instance center and the relocated pixel is calculated. Next, the index of the closest instance center is allocated to the pixel as its instance id, which yields the result of the instance segmentation branch. Merging the prediction result of the semantic segmentation branch with that of the instance segmentation branch gives the final panoptic segmentation result. An overview of the model is shown in Figure 5.

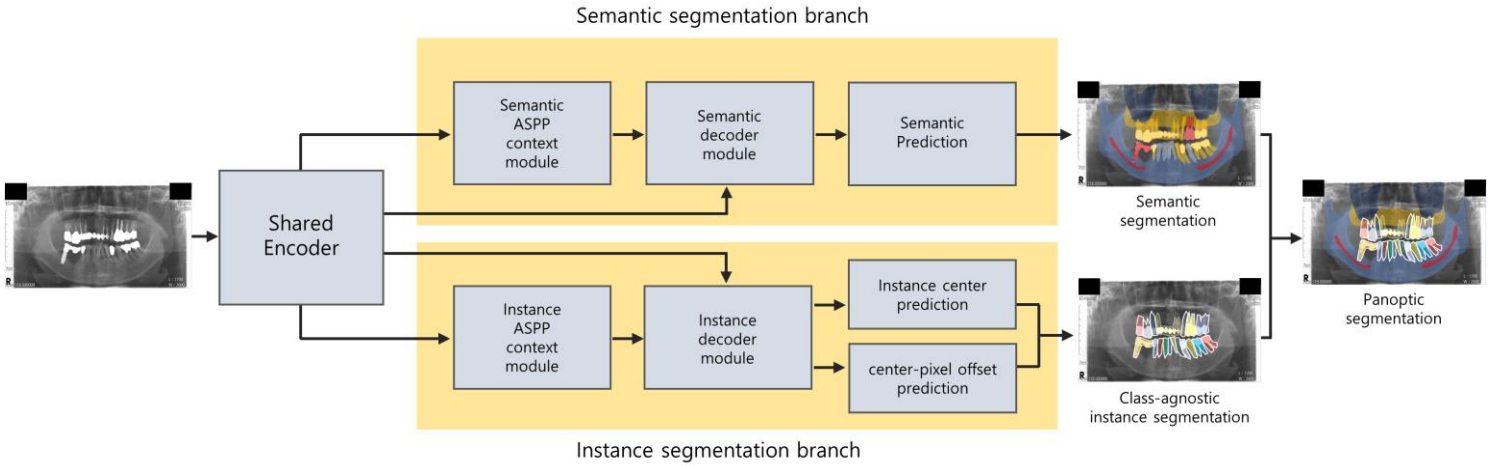


Figure 5. Overview of the model used in this study. The semantic and instance segmentation branches use the same feature map and share the encoder. Each branch has its multi-scale context module and decoder, both of which use ASPP. The instance segmentation branch predicts the center of mass for each instance and the offset vector between the center and each pixel. With the predicted center and the offset, the instance id of each pixel can be determined, thereby resulting in class-agnostic instance segmentation. The final panoptic segmentation is obtained by fusing the results of the semantic segmentation and class-agnostic instance segmentation.

2.3.3 Evaluation Methods

Panoptic segmentation encompasses semantic and instance segmentation; thus, the inference results of the model in this study can be evaluated from diverse perspectives. The metrics designed for panoptic segmentation as well as those for semantic and instance segmentation were selected by considering all these perspectives.

First, the PQ, SQ, and RQ were obtained, as proposed in a previous study[28]. These metrics consider the semantic and instance perspectives in a comprehensive manner and are widely used for the evaluation of the results from panoptic segmentation[41, 42].

PQ is defined as

$$PQ = \frac{\sum_{(p,g)\in TP} IoU(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \quad (3)$$

where p, g, and IoU represent the predicted segment, ground truth segment, and IoU, respectively. TP, FP, and FN represent the true positives, false positives, and false negatives, respectively, at the instance level. Specifically, TP, FP, and FN are matched segment pairs, unmatched predicted segments, and unmatched ground truth segments, respectively, found by segment matching after calculating the IoU. PQ is calculated for each class and can be averaged over classes, as shown in the result section.

In equation (3), IoU, which is also known as the Jaccard index, is defined as

$$IoU(p,g) = \frac{|p \cap g|}{|p \cup g|} \quad (4)$$

where p and g represent the predicted segment and ground truth segment, respectively. Only segment pairs with IoUs higher than 0.5 were considered to be matched pairs.

PQ can be expressed as a product of SQ and RQ:

$$PQ = SQ \times RQ \tag{5}$$

where SQ is defined as

$$SQ = \frac{\sum_{(p,g)\in TP} IoU(p,g)}{|TP|} \tag{6}$$

and RQ as

$$RQ = \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{7}$$

In equation (6), SQ is the averaged IoU of all the matched segment pairs. Furthermore, in equation (7), RQ is the same as the F1 score, the harmonic mean of precision and recall. Thus, decomposing PQ into these two terms helps in interpreting the PQ.

The IoU and iIoU[43] were calculated to evaluate the model's inference results from the perspective of the pixel-level semantic segmentation. Unlike the IoU in equation (3), where each IoU was calculated for each segment pair, the IoU, in this case, was calculated for each class, globally across all the panoramic radiographs in the test dataset:

$$IoU = \frac{TP}{TP + FP + FN} \tag{8}$$

The essential concept of IoU is the same in equations (4) and (8), but we used another notation to emphasize the difference described above. Unlike in equations (3), (6), and (7), TP, FP, and FN in equation (8) represent the number of pixels of true positive, false positive, and false negative, calculated for one class summed over all the panoramic radiographs.

However, the IoU in equation (8) has some bias toward large objects. To address this, iIoU uses values that are adjusted as per the scale of each object:

$$iIoU = \frac{iTP}{iTP + FP + iFN} \tag{9}$$

where iTP and iFN are respectively TP and FN weighted by the ratio of the average instance area of the class to the area of each ground truth instance:

$$iTP = \sum_i \left[TP_i \times \frac{(average\ instance\ area\ of\ the\ class)}{(area\ of\ the\ ground\ truth\ instance\ i)} \right] \tag{10}$$

$$iFN = \sum_i \left[FN_i \times \frac{(average\ instance\ area\ of\ the\ class)}{(area\ of\ the\ ground\ truth\ instance\ i)} \right] \tag{11}$$

In equations (10) and (11), TP_i and FN_i represent the number of pixels of true positive and false negative, respectively, that belong only to the corresponding ground truth instance i.

As this metric assumes that the model's output does not include any information about distinguishing among the individual instances, the pixels that correspond to FP do not belong to a specific instance. Thus, FP is not weighted. Furthermore, the IoU and iIoU were calculated for each class and each category.

Finally, to evaluate the model's inference results from the perspective of the instance segmentation, the AP for each "thing" was calculated and averaged across 10 different IoU threshold values ranging from 0.5 to 0.95 in steps of 0.05, as it is a widely used method for avoiding bias toward a specific threshold[34].

2.3.4 Neural Network Training Specifications

The training data were enriched using randomly generated data augmentation, which included horizontal flipping and randomized cropping. The hyperparameters of the model, such as base learning rate and the number of total iterations, were chosen based on the

evaluation results of the validation dataset. Consequently, the base learning rate was set to 0.001, warmed up for 1000 iterations, and gradually decreased as the iteration progressed. The total number of iterations was 65,000. The Adam method[44] was used for neural network optimization. The neural network model was trained on a cloud machine (Colaboratory, Google Research) with a 16 GB GPU accelerator (Tesla V100, Nvidia).

The software codes for preprocessing the data, training and running the model for inference results, and computing the evaluation metrics were mostly written using Python (Python 3.7.10, Python Software Foundation). To facilitate the model training and inferencing, an open-source project[45] was modified and used as a library, based on a machine learning library with GPU support (PyTorch 1.8.1, Facebook AI Research). An annotation tool[46] was used to prepare the datasets, whose results were preprocessed to proper formats in order to be fed into the model. The model used in this study was adopted from a previous study[29]. Evaluation metrics from previous studies[28, 43] and the related open-source codes[47, 48] with some modifications were used.

3. Results

3.1 Keypoint detection on periapical radiographs

3.1.1 Implant Detection Evaluation

For evaluation of implant detection (bounding boxes around implants), AP and AR at various IoU thresholds were calculated. The AP and AR for upper implant detection averaged for all IoU thresholds, increased in steps of 0.05 from 0.5 to 0.95, are 0.627 and 0.684, respectively. The AP and AR for lower implant detection averaged for all IoU thresholds are 0.657 and 0.728, respectively. The results are presented in Table 1.

Table 1. AP and AR on various IoU and OKS thresholds.

		AP (all)	AP (50)	AP (75)	AR (all)
bounding box	upper	0.627	1.000	0.746	0.684
	lower	0.657	1.000	0.714	0.728
keypoints	upper	0.761	1.000	0.832	0.810
	lower	0.786	1.000	0.907	0.845

AP (all): AP averaged over all IoU/OKS (0.50-0.95). AP(50): AP at IoU/OKS 0.50. AP(75): AP at IoU/OKS 0.75. AR (all). AR averaged over all IoU/OKS (0.50-0.95).

3.1.2 Keypoint Detection Evaluation

To obtain the OKS value in Equation (2), the standard deviations σ_i for each keypoint type i around the implant were calculated by annotating the test dataset twice. The calculated σ_i for this study were as follows: σ_i = [0.0895, 0.0816, 0.0193, 0.0196, 0.0209, and 0.0273] for Lt. bone level, Rt. bone level, Lt. apex, Rt. apex, Lt. implant top, and Rt. implant top, respectively. Based on these standard deviations, the OKS values for each ground truth and prediction pair were calculated. Similar to IoU, the OKS value can be

interpreted as how close the model’ s prediction is to the ground truth keypoints.

Using these results, AP and AR at various OKS thresholds were calculated for the evaluation of keypoint detection around the detected implant. The results are summarized in Table 1. In addition, precision-recall curves were crafted by varying the OKS threshold from 0.50 to 0.95 in increments of 0.05. The graphs are shown in Figure 6.

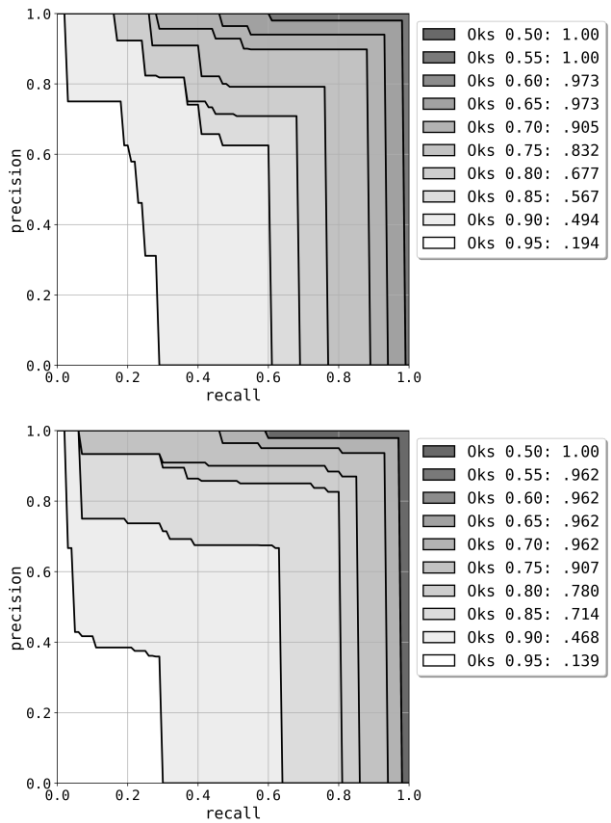


Figure 6. Precision-recall graph for various OKS thresholds. Each colored graph represents each OKS threshold, and each point in the graph corresponds to a specific confidence score threshold of the model. Top: result of the upper images. Bottom: result of the lower images.

3.1.3 Mean OKS

For the test dataset, the mean OKS values of the model used in this study were 0.8748, 0.9029, and 0.8885 for upper, lower, and total test datasets, respectively, while the mean OKS of a dentist for the total test dataset was 0.9012. From Equation (2), the individual keypoint similarity is $exp(-d_i^2/2s^2{k_i}^2)$ and $k_i = 2\sigma_i$. Thus, given the mean OKS, where the prediction belongs within the normal distribution of human annotated keypoints can be estimated. The mean OKS of 0.8885 for the total test dataset corresponds to $d_i/s \approx 0.9725\sigma_i$, on average, and this indicates that approximately 66.92% of human keypoint annotations are better than those of the model used herein. To compare the mean OKS values between a dentist and our model, an independent t-test was performed. All pairs showed no statistically significant difference. The results are summarized in Table 2.

Table 2. Mean OKS values of a dentist and the model.

	Mean OKS		p-value
Dentist	0.9012	Dentist-Model (total)	0.4095
Model (total)	0.8885	Dentist-Model (upper)	0.1441
Model (upper)	0.8748	Dentist-Model (lower)	0.9125
Model (lower)	0.9029	Model (upper)-Model (lower)	0.1543

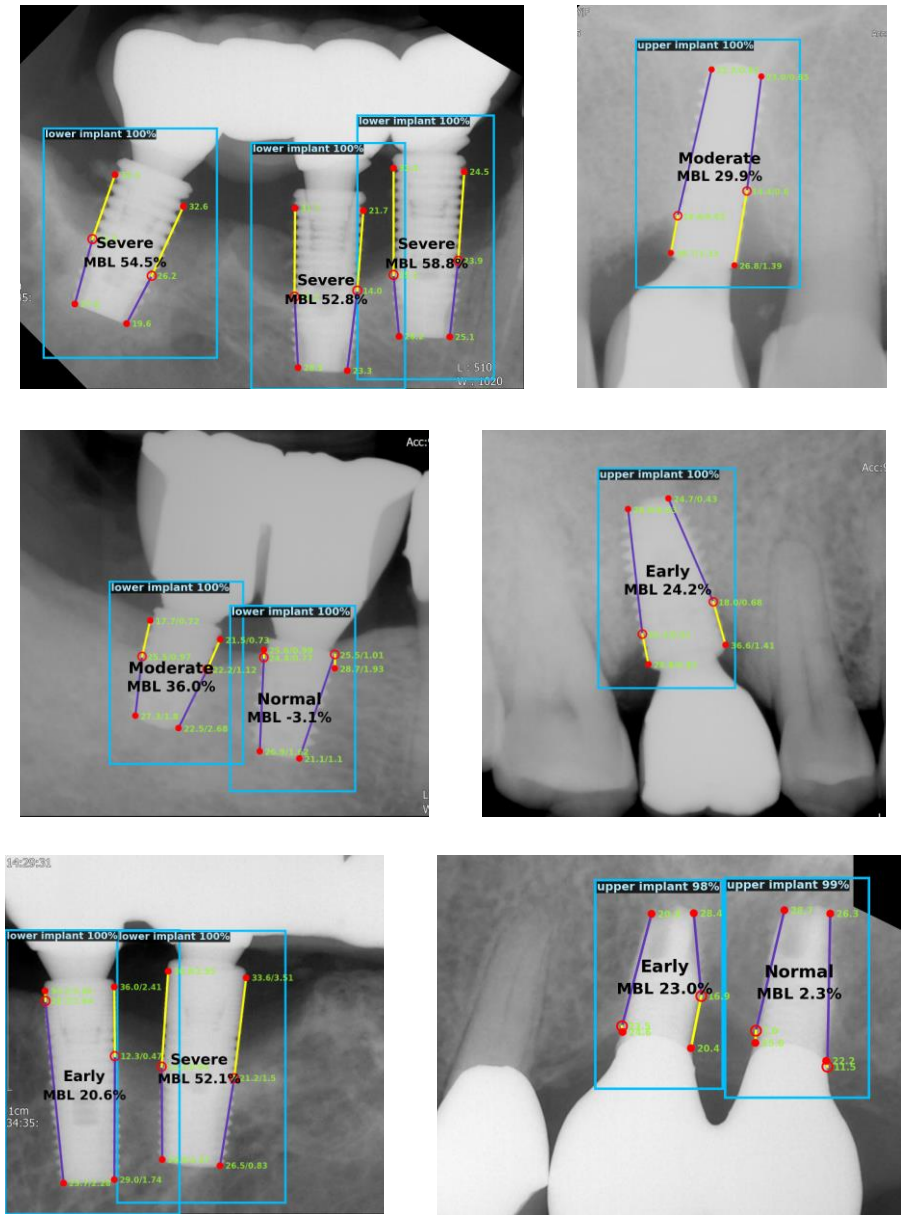


Figure 7. Examples of the predicted results. Each implant is detected using a bounding box and predicted keypoints are shown within the box. Radiographic bone loss ratio is calculated based on the keypoint locations. Confidence scores of the implant and keypoint detection are also shown.

3.2 Panoptic segmentation on panoramic radiographs

3.2.1 Visualization of the Inference Results

In order to visually examine the inference results, the output values of the model were visualized and superimposed on the original inputs of the panoramic radiographs (Figure 8).

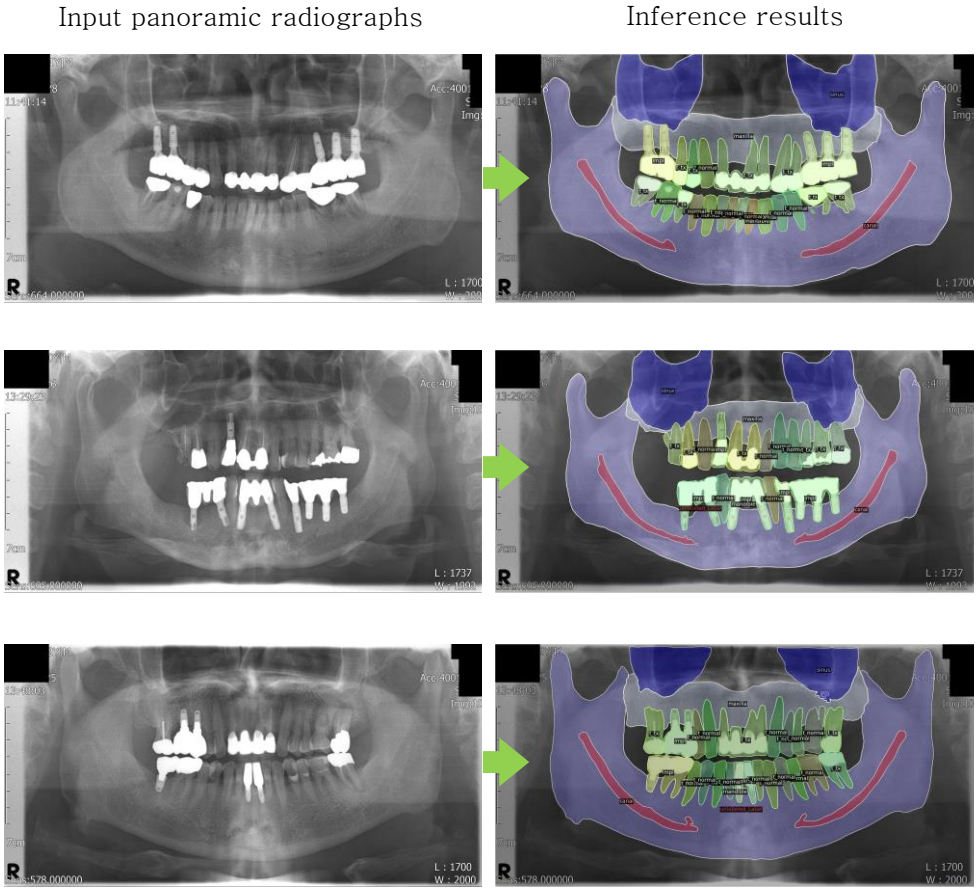


Figure 8. Visualization of the inference results. Some structures, including the mandibular canal and the medial wall of the maxillary sinus, are difficult to distinguish on the original panoramic radiographs but are fairly well segmented. Note that the normal tooth, treated tooth, and dental implant classes are individually segmented for each object; this is not possible with conventional semantic segmentation. Left: original input panoramic

radiographs. Right: visualized results of the model's inference.

3.2.2 Evaluation with Panoptic Segmentation Metrics

The metrics for panoptic segmentation, PQ, SQ, and RQ, were computed for each class and averaged over three categories: all classes, things, and stuff. The PQ, SQ, and RQ values averaged over all classes were 74.9, 83.2, and 90.0, respectively. The evaluation results are presented in Table 3.

Table 3. PQ, SQ, and RQ.

	PQ	SQ	RQ	N
Unlabeled	95.77	95.77	100.00	–
Mandible	89.73	89.73	100.00	–
Maxilla	82.77	82.77	100.00	–
Sinus	90.74	90.74	100.00	–
Canal	65.97	65.97	100.00	–
Stuff	85.00	85.00	100.00	5
Normal tooth	84.28	87.29	96.55	–
Treated tooth	57.13	85.69	66.67	–
Dental implant	77.34	87.89	88.00	–
Things	72.92	86.96	83.74	3
All	80.47	85.73	93.90	8

N: the number of classes.

3.2.3 Evaluation with Semantic Segmentation Metrics

The metrics for pixel-level semantic segmentation, IoU, and iIoU, were computed for each class and category. The iIoU was computed only for "things" and the corresponding category. A confusion matrix is presented to simplify the evaluation results (Figure 9). In the confusion matrix, rows represent the ground truth classes, whereas columns represent the classes predicted by the model. Each value in a cell represents a ratio of the number of pixels predicted by the model (as a class of the column among the pixels of the ground truth class) to the number of pixels of the ground truth class. A prior

was computed for each row in the matrix, which represented a ratio of the number of pixels of the corresponding ground truth class to the total number of pixels. The confusion matrix for each class is shown in Figure 9. The IoU and iIoU for each class and each category are shown in Table 4 and Table 5, respectively.

Ground truth \ Prediction	UL	Mn	Mx	Sinus	Canal	T_n	T_{tx}	Impl	Prior
UL	0.981	0.006	0.002	0.006	0.000	0.002	0.002	0.001	0.537
Mn	0.037	0.941	0.002	0.000	0.016	0.002	0.001	0.001	0.222
Mx	0.018	0.012	0.902	0.015	0.000	0.018	0.030	0.005	0.046
Sinus	0.023	0.000	0.017	0.959	0.000	0.000	0.000	0.001	0.071
Canal	0.001	0.194	0.000	0.000	0.802	0.000	0.000	0.002	0.017
T_n	0.018	0.025	0.017	0.000	0.000	0.897	0.043	0.000	0.039
T_{tx}	0.019	0.041	0.011	0.000	0.000	0.054	0.867	0.008	0.038
Impl	0.029	0.004	0.003	0.000	0.000	0.001	0.038	0.924	0.030

Figure 9. Confusion matrix. Each row and column represent the ground truth class and the predicted class, respectively. Each cell represents the ratio of the number of predicted pixels (column class) among the ground truth pixels (row class) to the number of ground truth pixels. Priors were computed for each row to represent the ratio of the number of pixels in the corresponding ground truth class to the total number of pixels. UL, unlabeled class; Man, mandible; Max, maxilla; T_n, normal tooth; T_{tx}, treated tooth; Impl, dental implant.

Table 4. IoU and iIoU for each class.

	IoU	iIoU
Unlabeled	0.954	–
Mandible	0.898	–
Maxilla	0.812	–
Sinus	0.898	–
Canal	0.639	–
Normal tooth	0.727	0.744
Treated tooth	0.656	0.611
Dental implant	0.775	0.827
Average	0.795	0.727

Table 5. IoU and iIoU for each category.

	IoU	iIoU
Unlabeled	0.954	–
Bone	0.886	–
Sinus	0.898	–
Canal	0.639	–
Tooth	0.895	0.890
Average	0.854	0.890

Bone: a category including maxilla and mandible. Tooth: a category including normal tooth, treated tooth, and dental implant.

3.2.4 Evaluation with Instance Segmentation Metrics

The metrics for instance segmentation, AP, were computed for each thing. The AP values averaged over all the IoU thresholds for the normal tooth, treated tooth, and dental implant were 0.520, 0.316, and 0.414, respectively (Table 6). In addition, the AP value at the IoU threshold of 0.5, which is widely used for evaluation in the object detection task, has been presented for ease of reference.

Table 6. AP for all the "thing" classes.

	AP (50)	AP (all)
Normal tooth	0.772	0.520
Treated tooth	0.490	0.316
Dental implant	0.714	0.414
Average	0.658	0.417

AP (50): AP at IoU threshold 0.5. AP (all): AP averaged across all IoU thresholds.

4. Discussion

4.1 Keypoint detection on periapical radiographs

A dental clinician needs to locate appropriate landmarks to analyze radiographs and diagnose peri-implantitis. Furthermore, as suggested by previous studies, the severity of peri-implantitis can be categorized based on the percentage of the radiographic bone loss.[36-38] For these tasks, the proposed automated system can be used for assisting dental researchers or practitioners.

AP and AR are popular evaluation metrics in the field of object detection and instance segmentation. However, these metrics are intended for model evaluation and not human evaluation because AP and AR are calculated using scores, indicating the model's confidence. Thus, the application of these metrics for evaluating humans is difficult. For a long time, the prediction results of machines did not match those of humans with regard to object detection or instance segmentation.[2] Therefore, the metrics comparing AI and human experts in these fields have not been extensively studied.

In segmenting biological images, such as in cell segmentation, the average IoU[49-51], or the mean Dice coefficient[52, 53] are frequently used. These metrics can also be applied to humans because they do not require a confidence score. As OKS was designed analogous to IoU, OKS can be averaged and interpreted in a manner similar to the average IoU. Here, to compare the prediction results of the model with those of a human expert, the mean OKS was calculated over entire implants that were detected in the test dataset. The result of the independent t-test showed no statistically significant difference between the results given by a dentist and those given by the model. Given the p-values, this deep learning model was considered to be helpful in detecting peri-implantitis under clinical situations.

To measure the amount of bone loss, a reference position that can be considered as a threshold for a sound bone level should be present. According to the 7th and 8th European Workshop on Periodontology[54, 55], the use of a baseline radiograph is recommended after physiologic remodeling, which is usually captured at the time of prosthesis installation unless immediate loading is performed, to assess the changes in the level of the crestal bone. However, in many clinical situations, a baseline radiograph is unavailable. Such cases were discussed in previous studies. The 8th European Workshop on Periodontology reached a consensus that a vertical distance of 2 mm from the expected marginal bone level is recommended as a threshold when no prior radiograph is present[54]. In the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions, the threshold was set to 3-mm apical of the most coronal portion of the intra-osseous part of the implant in the absence of a prior radiograph[56]. Other studies suggested a threshold from a fixed reference point, such as 2 mm apically from the implant platform for bone-level implants or 2 mm apically from the apical termination of the polished collar for tissue-level implants[36, 57].

As we did not use baseline radiographs, as is the usual case in many clinical situations, setting a reference point that can be used as a bone loss threshold is important. Owing to the fact that finding the expected marginal bone level can be subjective and the 2-mm distance in the radiographs may vary owing to distortion or magnification, a clear landmark is required that can be identified on radiographs.

Our dataset included a wide variety of implants, which have diverse shapes and implant-abutment junctions. If the implant is a bone-level implant and adopted platform switching, the most coronal position of the implant can be clearly observed. However, some types of implants, such as tissue-level implants, have shapes for which

identifying the most coronal point of the rough surface on periapical radiographs is difficult.

To cover all the implants in the dataset and those used widely in clinical practice, the most coronal thread of the implant was adopted as a threshold position. In other words, the most coronal thread was considered as a reference point that is supposed to be a sound bone level if bone resorption was not more than 0.2 mm annually after the physiologic remodeling, which occurs mostly during the first year after implant placement. This can help avoid radiographic distortion or magnification that will make it difficult to determine a 2-mm distance in the radiographs. Furthermore, this approach can be applied to various types of implants. However, this method has a limitation in that some types of implants have the most coronal thread at a more apical position than others. For instance, the reference point of the implant top for tissue-level implants should be around 2 mm below the apical termination of the polished collar but, often, the most coronal thread is located below that. Nevertheless, identifying the point that is the end of the rough surface area or the apical termination of the polished collar on periapical radiographs is difficult even for a human expert. Further research is necessary to overcome this challenge.

Numerous previous studies have stated certain diagnosis criteria for peri-implantitis[36, 54-57]. Most of them use radiographic marginal bone loss and bleeding on probing and/or suppuration as the criteria. As the diagnosis of peri-implantitis requires radiographic as well as clinical information, diagnosing the disease only with periapical radiographs is not practical. This is a limitation of the suggested system, and further research needs to be conducted using more general information, such as clinical information, to assist peri-implantitis diagnosis. In addition, some information such as the length of the implant is needed to obtain the absolute depth of the bone

defect site. Using the ratio calculated by the automated system suggested in this study, the absolute bone loss length can be obtained by multiplying it with the real length of the implant.

It is also important to note that two-dimensional images are not the only tool to evaluate the marginal bone loss of the dental implant. Cone-beam computed tomography (CBCT) is helpful when assessing the peri-implant bone loss because it can provide a three-dimensional relationship between a dental implant and the surrounding alveolar bone. Some previous studies have sought to identify bone conditions around implants using periapical radiographs and CBCT images together[58, 59], and other studies have reported that CBCT is highly accurate for detecting peri-implant bone defects[60, 61]. Thus, it will be more meaningful if a machine learning system can utilize the information from the CBCT as well as two-dimensional radiographs when evaluating the peri-implant bone conditions. This should be further studied in the future. Still, measuring the amount of the bone defect on conventional radiographic methods is important because two-dimensional radiographic images are widely used in the field and dental clinicians often encounter situations in which CBCT cannot be used owing to patients' financial constraints or the circumstances of dental clinics.

Although the mean OKS between the model and a human expert are comparable, the proposed approach has certain limitations. First, a threshold for the model still needs to be set, over which the model will output the bounding box results. Hence, object detection results and the mean OKS value can vary with the threshold. In addition, as the information of the model's confidence score above the threshold is not used, the comparison is limited between two models whose confidence scores are different.

Second, the standard deviations σ_i for each keypoint type i around the implant were calculated based on human (i.e., dentist)

annotations. However, even for a dentist, bone level detection is a challenge, as seen from the following values: σ = [0.0895, 0.0816, 0.0193, 0.0196, 0.0209, 0.0273]. The first two figures show the standard deviations of the left and right bone level annotations, which are the highest among the six values.

Thus, if the model precisely locates the bone level and locates other keypoints far from the ground truth, the OKS value decreases even though the model was successful at more challenging tasks. As the standard deviations σ_i of human annotations are larger for bone level keypoints, the distances of the detected bone level keypoints from the ground truth contribute to a relatively lower degree. As localizing the bone level is considered the most important task, this evaluation metric should be modified in further research.

Unlike the common objects in a context evaluation method[39] where the object detection score of the model is used for applying a threshold when plotting the precision-recall curves, the keypoint localizing score was used here. For tasks where the difficulty of detecting objects is proportional to that of localizing keypoints, using the object detection score for the precision-recall curve makes sense. However, in our task, detecting the implant is easier than detecting the keypoint position and the confidence score of implant detection was quite high for many cases. This implies that the confidence score of implant detection does not indicate the confidence rate of keypoint detection. The plotted results of the precision-recall curves using the two different confidence scores were different in our task. When using the implant detection score, even when the recall value decreased, the precision did not approach 1.0 when the OKS threshold was high (>0.70). After changing the confidence score to represent the keypoint score instead of the implant detection score, the precision reached 1.0 at all OKS thresholds as the confidence score threshold increased. However, the results may vary because

of the randomness of the data augmentation process during training.

In this study, the keypoint annotation that was compared with the inference result of the neural network model was performed by the same dentist who performed the ground truth annotation of the test dataset for the comparison between the neural network and the dentist. Given that the same person performed the annotation, the works were done at a certain time interval, as in the previous study comparing artificial intelligence and humans. Nevertheless, the OKS values of the dentist could be high because of this reason. The comparison will be more reliable if multiple dentists participate. However, as in this study, if the same person performs the ground truth annotation of the test dataset and the annotation representing the dentist, the OKS of the dentist will be higher. Thus, in this study, the comparison was more unfavorable to the neural network model. If the results of several dentists are used in the future, the inference results of the neural network model may become higher than in the present study in the OKS distribution.

The extent of bone loss around the implant varies depending on the type of implant–abutment junction or platform switching. If such clinical information is also added and used for the training of neural networks, more reliable inference results can be obtained. This should be addressed in future research.

Although studies have shown that periapical radiographs are more reliable than panoramic radiographs for determining bone level around implants, periapical radiographs have more variation in actual clinical practice than panoramic radiographs in which the patient's posture and angle are relatively more easily standardized during imaging. In this study, 708 periapical radiographs that met the inclusion criteria out of 1,000 were used, but there were several cases that were not taken in parallel, and there were many issues such as not showing the apex of the implant even if they were parallel.

Thus, if more strict exclusion criteria are applied, the proportion of radiographs passing the criteria will be extremely low. In future studies, with more stringent exclusion criteria and radiographs taken parallel to clearly visible implant threads, the inference accuracy of the deep learning model will be further improved.

4.2 Panoptic segmentation on panoramic radiographs

Dental panoramic radiographs are used to detect and diagnose diseases of the oral and maxillofacial area and to create various treatment plans in the dental clinic. Therefore, accurately reading a panoramic radiograph is important in the field of dentistry.

Several studies on artificial intelligence models targeting dental panoramic radiographs have been published in the past. Some of them aimed to binary-classify the presence of specific diseases[62] or classify the types of diseases using panoramic images[63]; however, the deep neural network models used in these studies could classify images but not locate the disease within the panoramic radiographs. Furthermore, most of these models must be trained with cropped images, which needed be done manually so that the RoI was located in the center of the cropped image.

Certain studies attempted to perform object detection, which predicts the location of the disease and classifies it, using panoramic radiographs[24-26]. Furthermore, instance segmentation, which not only locates the object or lesion but also segments its outline, was applied on panoramic radiographs[27]. However, most instance segmentation models were not designed to segment extremely wide or long objects, such as the jaw or mandibular canal shown in panoramic radiographs. For example, one of the most widely studied and used CNNs for instance segmentation, the Mask R-CNN[11], uses anchors of various sizes and aspect ratios to predict the RoI before further regression.

Although it is possible to adjust the size, aspect ratio, and angle of the anchor to fit the jaws or mandibular canal in the panoramic radiograph, the model was not designed to detect such unusual objects. Some researchers applied semantic segmentation[23], which classifies each pixel but does not distinguish between individual teeth in the panoramic radiograph.

Each of these approaches has its advantages and disadvantages. Panoptic segmentation was recently proposed to combine the different types of tasks[28], and deep learning models that can achieve this are currently being evaluated[29, 64]. However, to the best of our knowledge, panoptic segmentation has not been applied in the fields of medicine and dentistry.

In the current study, a state-of-the-art deep learning model capable of panoptic segmentation was applied to dental panoramic radiographs. Good results were obtained, as observed from the evaluation and visualization results. It is difficult to distinguish the various structures and the double as well as ghost images visible on the panoramic radiograph. The outlines of the maxillary sinus and mandibular canal are often difficult to find, even for an experienced dentist. However, it is important to identify the boundaries of these structures, especially during treatment planning. Therefore, unlike many previous studies that mainly focused only on teeth segmentation[23, 27], the present study examined whether the maxillary sinus and mandibular canal can be identified on dental panoramic radiographs using a deep neural network model.

The segmentation of the mandibular canal showed the lowest PQ and SQ among the "stuff" classes, which was consistent with the results of the IoU. However, as shown in Figure 3, most of the original input panoramic radiographs tested were faint, making it difficult to read and identify the mandibular canal. Furthermore, as the confusion matrix indicates, the prior value of the mandibular canal

was the lowest among all classes, which showed that the area covered by the mandibular canals was less than 2% of all the pixels. Nevertheless, our model accurately detected almost 80% of the ground truth pixels of the mandibular canal. A previous study, which proposed the concept of panoptic segmentation, tried to compare the artificial neural networks to human annotators and showed that human annotators outperformed the machines[28]. Given that artificial intelligence has not surpassed humans in this task so far, the recognition of the mandibular canal in the current study might be considered a noteworthy achievement.

Among the stuff classes, the maxilla showed the second lowest evaluation score after the mandibular canal, as can be seen from the PQ and IoU results. The reason for this is presumed to be the unclear boundary between the maxillary sinus and maxilla; in addition, some structures such as the hard palate and zygomatic arch often interfere with the readings. There are cases where the central part of the maxilla is unclear because of the overlap of a ghost image caused by the cervical spine. In addition, the number of pixels in the area covered by the maxilla was the second smallest (<5%) after the number of pixels covered by the mandibular canal, among the stuff classes. Nonetheless, the model detected almost 90% of the ground truth pixels of the maxilla correctly, as can be seen in the confusion matrix. Surprisingly, the results for the maxillary sinus were even better (96% of the total ground truth pixels were accurately detected) considering the fact that the medial wall and the floor of the maxillary sinus are difficult to identify and often interfere with other structures, such as the innominate line and nasal floor. It is very important to determine the location and shape of the maxillary sinus and mandibular canal during dental implant surgery; therefore, these results suggest that artificial intelligence offers considerable potential to be of assistance in dental clinics in the future.

In the case of the thing classes, the treated tooth class showed the second lowest IoU after the mandibular canal. This is presumed to be because of the disadvantage in training because the number of treated teeth was relatively smaller than the number of normal teeth. A similar trend was observed with the iIoU and AP values. Furthermore, the treated tooth showed the lowest PQ among all classes because of the low RQ. A key reason for this is that many of the treated teeth and dental implants are bridges; thus, the model may find it difficult to distinguish between an individual tooth and an implant. In the case of a bridge without a pontic, or a single crown adjacent to a bridge or another single crown, it is sometimes difficult to ascertain whether they are simply adjacent or connected to each other.

In addition, in the case of a bridge that connects a natural tooth to a dental implant, a lower evaluation score is inevitable because the boundary between the treated tooth class and the implant class cannot be distinguished. The presence of a pontic in a bridge that connects the same type of abutment teeth, i.e., only natural teeth or only dental implants, ensures that the abutments are one connected instance. However, when the types of abutments are diverse, i.e., when a natural tooth is connected to an implant, it is not possible to identify the boundary between the treated tooth class and the dental implant class, in spite of the presence of a pontic.

Note that the incorrect detection of even a very small segment can negatively affect the RQ value to the same extent as a very large segment can. This is because the segment area itself does not affect the RQ. Unlike normal teeth and dental implants, the treated tooth class covers various treatments, such as fixed dental prostheses, diverse types of restorations, and root canal treatment. Owing to this characteristic of the treated tooth class, there are several cases where the segments predicted by the model are split and fragmented,

thereby reducing the RQ.

These issues seem to affect the AP in a similar manner, as can be seen from the low AP values in the treated tooth class. However, these issues do not decrease the IoU, which is calculated independently of the individual instances being distinguished from each other.

The IoU in equation (6) is a widely used metric in semantic segmentation; however, it does not reflect whether each object is detected, and is biased toward large objects. To alleviate this shortcoming, the iIoU, which normalizes the IoU using the area of each instance, was used in the current study. Nevertheless, the ranking of the thing classes according to the evaluation score was the same for IoU as well as iIoU, in this study. This might be attributed to the small differences in the areas between the objects belonging to the thing classes, on the panoramic radiographs. If there was a large difference in area between individual instances, the iIoU value could have been very different from the IoU value.

The thing classes showed lower PQ values than the stuff classes did, mainly owing to the difference in RQ; in addition, the treated tooth class played a major role, as described earlier. Some degree of loss of RQ was observed even in the normal tooth class, where there was no need to distinguish between the bridge and the adjacent single crowns. The nature of the dental panoramic radiograph is probably a key reason for this, i.e., multiple teeth of similar shape and size being located close to each other. In many computer vision tasks, if a large number of objects belonging to a particular class are clustered in one location, they are separated into another class because it is difficult to distinguish between each object. Taking this into account, it is quite natural to have a low RQ value for the thing classes, because crowded teeth are frequently observed in dental panoramic radiographs.

Unlike the instance segmentation task, where each segment can overlap other segments, each pixel has only one value in the panoptic segmentation class. This is not a major concern in typical images because if a specific object covers the object behind it, the covered part of the object is not visible. However, in radiographs, even if the object in the front covers the object behind, the object behind may be visible, so a situation arises in which a specific pixel must have multiple values. This situation occurs frequently in normal tooth classes because the crowns are very often overlapped and the hidden parts are visible. Thus, annotating the ground truth for the normal tooth class requires some compromise. When two crowns overlap each other, the midpoint of the overlapped portion is assumed as the boundary between the two crowns because a pixel can have only one instance id. Alternatively, all the teeth can be treated as a single instance if the crowns overlap. However, this method was not used because, in some cases, it was not clear whether the crowns were overlapped. The evaluation index results differ considerably depending on whether the teeth are viewed as one instance or separate instances. This issue could adversely affect not only the RQ but also the SQ of the normal tooth class because SQ can be interpreted as the averaged IoU over all matched segment pairs.

It is worth noting that satisfactory results were obtained in the current study, despite a significantly smaller number of datasets being used than those used for general machine learning and deep learning training projects. This might be attributed to the transfer learning, data augmentation, and standardized imaging methods used for dental panoramic radiographs. Unlike general computer vision tasks, in the case of panoramic radiographs, the radiograph is taken in a consistent manner with the patient positioned at a certain location and angle. Furthermore, the structures in the image are arranged in a specific pattern. Therefore, considering the number of radiographs

used in this study, we believe that better results can be obtained if a larger amount of data is used.

Another factor that must be mentioned is that in this study, the panoramic radiographs were captured using a specific machine in a single hospital. The brand and company of the machine used to take orthopantomograms affect the quality and characteristics of the radiographs considerably. A previous study demonstrated that the performance of the model can be improved when panoramic radiographs from multiple hospitals are mixed (cross-center training)[65]. Therefore, it is possible to develop a more generalized deep neural network by using radiographs captured by various types of panoramic radiograph machines. Further research is needed to improve generalization and avoid overfitting of neural networks.

Several pieces of radiographic equipment that take better-quality panoramic radiographs have been developed in recent years, and many of them are being introduced in dental clinics. Given that panoramic radiographs used in this study have inferior quality compared to those taken using recently developed machines, the inference results will be further improved if the model is trained and tested using higher-quality images. Thus, as radiographic equipment continue to improve, artificial intelligence might be of assistance for the reading of panoramic radiographs in the future.

Although the machine learning method described in this study can segment many important structures in panoramic radiographs, there remain many other structures that have not been considered. Once the above-mentioned improvements are applied, the proposed method can be the foundation for future studies in detecting a diverse range of structures in dental panoramic radiographs.

5. Conclusions

A keypoint detection model, after being fine-tuned with a machine learning method based on transfer learning, demonstrated the ability to determine the extent of bone loss on radiographs for diagnosing peri-implantitis almost as accurately as human experts.

In addition, a deep neural network model designed for panoptic segmentation could detect and segment various structures in dental panoramic radiographs. It could even segment the maxillary sinus and mandibular canal, which are often difficult to distinguish on a radiograph.

Thus, after fine-tuning with a suitable dataset, these machine learning methods can potentially assist dental practitioners while diagnosing and categorizing peri-implantitis, as well as setting up treatment plans and diagnosing oral and maxillofacial diseases.

Published papers related to this study

1. Cha J-Y, Yoon H-I, Yeo I-S, Huh K-H, Han J-S. Peri-Implant Bone Loss Measurement Using a Region-Based Convolutional Neural Network on Dental Periapical Radiographs. *Journal of Clinical Medicine*. 2021; 10(5):1009. https://doi.org/10.3390/jcm10051009

2. Cha J-Y, Yoon H-I, Yeo I-S, Huh K-H, Han J-S. Panoptic Segmentation on Panoramic Radiographs: Deep Learning-Based Segmentation of Various Structures Including Maxillary Sinus and Mandibular Canal. *Journal of Clinical Medicine*. 2021; 10(12):2577. https://doi.org/10.3390/jcm10122577

References

1. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. presented at: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1; 2012; Lake Tahoe, Nevada.
2. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*. 2015;115(3):211-252.
3. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology*. Aug 2017;284(2):574-582. doi:10.1148/radiol.2017162326
4. Lee KS, Jung SK, Ryu JJ, Shin SW, Choi J. Evaluation of Transfer Learning with Deep Convolutional Neural Networks for Screening Osteoporosis in Dental Panoramic Radiographs. *J Clin Med*. Feb 1 2020;9(2)doi:10.3390/jcm9020392
5. Krois J, Ekert T, Meinhold L, et al. Deep Learning for the Radiographic Detection of Periodontal Bone Loss. *Sci Rep*. Jun 11 2019;9(1):8495. doi:10.1038/s41598-019-44839-3
6. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014:580-587.
7. Girshick R. Fast r-cnn. 2015:1440-1448.
8. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:150601497*. 2015;
9. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. 2016:779-788.
10. Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector. 2015:arXiv:1512.02325. Accessed December 01, 2015. https://ui.adsabs.harvard.edu/abs/2015arXiv151202325L
11. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. 2017:arXiv:1703.06870. Accessed March 01, 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv170306870H
12. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. 2015:3431-3440.
13. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Springer; 2015:234-241.
14. Tran PV. A fully convolutional neural network for cardiac segmentation in short-axis MRI. *arXiv preprint arXiv:160400494*. 2016;
15. Memiş A, Varlı S, Bilgili F. Semantic segmentation of the multiform

proximal femur and femoral head bones with the deep convolutional neural networks in low quality MRI sections acquired in different MRI protocols. *Computerized Medical Imaging and Graphics*. 2020;81:101715.
16. Pekala M, Joshi N, Liu TA, Bressler NM, DeBuc DC, Burlina P. Deep learning based retinal OCT segmentation. *Computers in biology and medicine*. 2019;114:103445.
17. Sa R, Owens W, Wiegand R, et al. Intervertebral disc detection in X-ray images using faster R-CNN. IEEE; 2017:564-567.
18. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. *Radiology: Artificial Intelligence*. 2019;1(1):e180001.
19. Wang J, Li Z, Jiang R, Xie Z. Instance segmentation of anatomical structures in chest radiographs. IEEE; 2019:441-446.
20. De Bruyn H, Vandeweghe S, Ruyffelaert C, Cosyn J, Sennerby L. Radiographic evaluation of modern oral implants with emphasis on crestal bone level and relevance to peri-implant health. *Periodontol 2000*. Jun 2013;62(1):256-70. doi:10.1111/prd.12004
21. Lofthag-Hansen S, Huumonen S, Grondahl K, Grondahl HG. Limited cone-beam CT and intraoral radiography for the diagnosis of periapical pathology. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod*. Jan 2007;103(1):114-9. doi:10.1016/j.tripleo.2006.01.001
22. Schwendicke F, Golla T, Dreher M, Krois J. Convolutional neural networks for dental image diagnostics: A scoping review. *J Dent*. Dec 2019;91:103226. doi:10.1016/j.jdent.2019.103226
23. Koch TL, Perslev M, Igel C, Brandt SS. Accurate segmentation of dental panoramic radiographs with U-NETS. IEEE; 2019:15-19.
24. Ariji Y, Yanashita Y, Kutsuna S, et al. Automatic detection and classification of radiolucent lesions in the mandible on panoramic radiographs using a deep learning object detection technique. *Oral Surg Oral Med Oral Pathol Oral Radiol*. Oct 2019;128(4):424-430. doi:10.1016/j.oooo.2019.05.014
25. Yang H, Jo E, Kim HJ, et al. Deep learning for automated detection of cyst and tumors of the jaw in panoramic radiographs. *Journal of clinical medicine*. 2020;9(6):1839.
26. Kwon O, Yong T-H, Kang S-R, et al. Automatic diagnosis for cysts and tumors of both jaws on panoramic radiographs using a deep convolution neural network. *Dentomaxillofacial Radiology*. 2020;49(8):20200185.
27. Lee JH, Han SS, Kim YH, Lee C, Kim I. Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs. *Oral Surg Oral Med Oral Pathol Oral Radiol*. Jun 2020;129(6):635-642. doi:10.1016/j.oooo.2019.11.007
28. Kirillov A, He K, Girshick R, Rother C, Dollár P. Panoptic

segmentation. 2019:9404-9413.
29. Cheng B, Collins MD, Zhu Y, et al. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. 2020:12475-12485.
30. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016:770-778.
31. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. Ieee; 2009:248-255.
32. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. 2017:2117-2125.
33. Sun K, Xiao B, Liu D, Wang J. Deep high-resolution representation learning for human pose estimation. 2019:5693-5703.
34. Lin T, Maire M, Belongie SJ, et al. Microsoft COCO: Common objects in context. arXiv 2014. *arXiv preprint arXiv:14050312*. 2014;
35. Korshunova I, Shi W, Dambre J, Theis L. Fast face-swap using convolutional neural networks. 2017:3677-3685.
36. Lin G-H, Kapila Y, Wang H-L. Parameters to define peri-implantitis: A review and a proposed multi-domain scale. *Journal of Oral Implantology*. 2017;43(6):491-496.
37. Decker AM, Sheridan R, Lin GH, Sutthiboonyapan P, Carroll W, Wang HL. A Prognosis System for Periimplant Diseases. *Implant Dent*. Aug 2015;24(4):416-21. doi:10.1097/ID.0000000000000276
38. Froum SJ, Rosen PS. A proposed classification for peri-implantitis. *International Journal of Periodontics and Restorative Dentistry*. 2012;32(5):533.
39. COCO - Common Objects in Context: Keypoint Evaluation. Jun 30, 2020. Accessed Jun 30, 2020, 2020. https://cocodataset.org/#keypoints-eval
40. Ruggero Ronchi M, Perona P. Benchmarking and error diagnosis in multi-instance pose estimation. 2017:369-378.
41. Cityscapes Dataset Benchmark Suite. Accessed 28 Feb 2021, 2021. https://www.cityscapes-dataset.com/benchmarks/
42. COCO Dataset - Panoptic Leaderboard. Accessed 28 Feb 2021, https://cocodataset.org/#panoptic-leaderboard
43. Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding. 2016:3213-3223.
44. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:14126980*. 2014;
45. Girshick YWaAKaFMaW-YLaR. Detectron2. Accessed 17 Mar 2021, https://github.com/facebookresearch/detectron2
46. Brooks J. COCO Annotator. Accessed 7 Mar 2021, https://github.com/jsbroks/coco-annotator/
47. The Cityscapes Dataset. Accessed 20 Mar 2021, https://github.com/mcordts/cityscapesScripts

48. COCO 2018 Panoptic Segmentation Task API. Accessed 20 Mar 2021, https://github.com/cocodataset/panopticapi
49. Yi J, Tang H, Wu P, et al. Object-guided instance segmentation for biological images. 2020:12677-12684.
50. Qu H, Wu P, Huang Q, et al. Weakly Supervised Deep Nuclei Segmentation Using Partial Points Annotation in Histopathology Images. *IEEE Transactions on Medical Imaging*. 2020;39(11):3655-3666.
51. Guerrero-Peña FA, Fernandez PDM, Ren TI, Cunha A. A weakly supervised method for instance segmentation of biological cells. *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Springer; 2019:216-224.
52. Lee H, Jeong W-K. Scribble2Label: Scribble-Supervised Cell Segmentation via Self-generating Pseudo-Labels with Consistency. Springer; 2020:14-23.
53. Nishimura K, Bise R. Weakly supervised cell instance segmentation by propagating from detection response. Springer; 2019:649-657.
54. Sanz M, Chapple IL, Periodontology* WGotVEWo. Clinical research on peri-implant diseases: consensus report of Working Group 4. *Journal of clinical periodontology*. 2012;39:202-206.
55. Lang NP, Berglundh T, Periodontology WGotSEWo. Periimplant diseases: where are we now?--Consensus of the Seventh European Workshop on Periodontology. *Journal of clinical periodontology*. 2011;38:178-181.
56. Berglundh T, Armitage G, Araujo MG, et al. Peri-implant diseases and conditions: Consensus report of workgroup 4 of the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions. *J Periodontol*. Jun 2018;89 Suppl 1:S313-S318. doi:10.1002/JPER.17-0739
57. Konstantinidis IK, Kotsakis GA, Gerdes S, Walter MH. Cross-sectional study on the prevalence and risk indicators of peri-implant diseases. *Eur J Oral Implantol*. 2015;8(1):75-88.
58. Hadzik J, Krawiec M, Slawecki K, Kunert-Keil C, Dominiak M, Gedrange T. The Influence of the Crown-Implant Ratio on the Crestal Bone Level and Implant Secondary Stability: 36-Month Clinical Study. *Biomed Res Int*. 2018;2018:4246874. doi:10.1155/2018/4246874
59. Hadzik J, Krawiec M, Kubasiewicz-Ross P, Prylinska-Czyzewska A, Gedrange T, Dominiak M. Short Implants and Conventional Implants in The Residual Maxillary Alveolar Ridge: A 36-Month Follow-Up Observation. *Med Sci Monit*. Aug 14 2018;24:5645-5652. doi:10.12659/MSM.910404
60. Song D, Shujaat S, de Faria Vasconcelos K, et al. Diagnostic accuracy of CBCT versus intraoral imaging for assessment of peri-implant bone defects. *BMC Med Imaging*. Feb 10 2021;21(1):23. doi:10.1186/s12880-021-00557-9

61. Schwindling FS, Hilgenfeld T, Weber D, Kosinski MA, Rammelsberg P, Tasaka A. In vitro diagnostic accuracy of low-dose CBCT for evaluation of peri-implant bone lesions. *Clin Oral Implants Res*. Dec 2019;30(12):1200-1208. doi:10.1111/clr.13533
62. Murata M, Ariji Y, Ohashi Y, et al. Deep-learning classification using convolutional neural network for evaluation of maxillary sinusitis on panoramic radiography. *Oral radiology*. 2019;35(3):301-307.
63. Lee JH, Kim DH, Jeong SN. Diagnosis of cystic lesions using panoramic and cone beam computed tomographic images based on deep learning neural network. *Oral Dis*. Jan 2020;26(1):152-158. doi:10.1111/odi.13223
64. Kirillov A, Girshick R, He K, Dollár P. Panoptic feature pyramid networks. 2019:6399-6408.
65. Krois J, Cantu AG, Chaurasia A, et al. Generalizability of deep learning models for dental image analysis. *Scientific reports*. 2021;11(1):1-7.

Abbreviations

1. AP: average precision
2. AR: average recall
3. ASPP: atrous spatial pyramid pooling
4. CNN: convolutional neural network
5. FPN: feature pyramid network
6. GPU: graphics processing unit
7. iIoU: instance-level IoU
8. IoU: Intersection over Union
9. OKS: object keypoint similarity
10. PQ: panoptic quality
11. ResNet: residual neural network
12. RoI: region of interest
13. RPN: region proposal network
14. RQ: recognition quality
15. SQ: segmentation quality

-국문초록-

심층신경망을 이용한 자동화된 치과 의료영상 분석

서울대학교 치의학대학원 치의과학과 치과보철학 전공

(지도교수 한 중 석)

차 준 영

목 적: 치과 영역에서도 심층신경망(Deep Neural Network) 모델을 이용한 방사선사진에서의 임플란트 분류, 병소 위치 탐지 등의 연구들이 진행되었으나, 최근 개발된 키포인트 탐지(keypoint detection) 모델 또는 전체적 구획화(panoptic segmentation) 모델을 의료분야에 적용한 연구는 아직 미비하다. 본 연구의 목적은 치근단 방사선사진에서 키포인트 탐지를 이용해 임플란트 골 소실 정도를 파악하는 모델과 panoptic segmentation을 파노라마영상에 적용하여 다양한 구조물들을 구획화하는 모델을 학습시켜 진료에 보조적으로 활용되도록 만들어보고, 이 모델들의 추론결과를 평가해보는 것이다.

방 법: 객체 탐지 및 구획화에 있어 널리 연구된 합성곱 신경망 모델인 Mask-RCNN을 키포인트 탐지가 가능한 형태로 준비하여 치근단 방사선사진에서 임플란트의 top, apex, 그리고 bone level 지점을 좌우로 총 6지점 탐지하게끔 학습시킨 뒤, 학습에 사용되지 않은 시험 데이터셋을 대상으로 탐지시킨다. 키포인트 탐지 평가용 지표인 object keypoint similarity (OKS) 및 이를 이용한 average precision (AP) 값을 계산하고, 평균 OKS값을 통해 모델 및 치과의사의 결과를 비교한다. 또한, 탐

지된 키포인트를 바탕으로 방사선사진상에서의 골 소실 정도를 수치화한다.

Panoptic segmentation을 위해서는 기존의 벤치마크에서 우수한 성적을 거둔 신경망 모델인 Panoptic DeepLab을 파노라마영상에서 주요 구조물(상악동, 상악골, 하악관, 하악골, 자연치, 치료된 치아, 임플란트)을 구획화하도록 학습시킨 뒤, 시험 데이터셋에서의 구획화 결과에 panoptic / semantic / instance segmentation 각각의 평가지표들을 적용하고, 픽셀들의 정답(ground truth) 클래스와 모델이 추론한 클래스에 대한 confusion matrix를 계산한다.

결 과: OKS값을 기반으로 계산한 키포인트 탐지 AP는, 모든 OKS threshold에 대한 평균의 경우, 상악 임플란트에서는 0.761, 하악 임플란트에서는 0.786이었다. 평균 OKS는 모델이 0.8885, 치과의사가 0.9012로, 통계적으로 유의미한 차이가 없었다 (p = 0.41). 모델의 평균 OKS 값은 사람의 키포인트 어노테이션 정규분포상에서 상위 66.92% 수준이었다.

파노라마영상 구조물 구획화에서는, panoptic segmentation 평가지표인 panoptic quality 값의 경우 모든 클래스의 평균은 80.47이었으며, 치료된 치아가 57.13으로 가장 낮았고 하악관이 65.97로 두번째로 낮은 값을 보였다. Semantic segmentation 평가지표인 global한 Intersection over Union (IoU) 값은 모든 클래스 평균 0.795였으며, 하악관이 0.639로 가장 낮았고 치료된 치아가 0.656으로 두번째로 낮은 값을 보였다. Confusion matrix 계산 결과, ground truth 픽셀들 중 올바르게 추론된 픽셀들의 비율은 하악관이 0.802로 가장 낮았다. 개별 객체에 대한 IoU를 기반으로 계산한 Instance segmentation 평가지표인 AP값은, 모든 IoU threshold에 대한 평균의 경우, 치료된 치아가 0.316,

임플란트가 0.414, 자연치가 0.520이었다.

결 론: 키포인트 탐지 신경망 모델을 이용하여, 치근단 방사선사진에서 임플란트의 주요 지점을 사람과 다소 유사한 수준으로 탐지할 수 있었다. 또한, 탐지된 지점들을 기반으로 방사선사진상에서의 임플란트 주위 골소실 비율 계산을 자동화할 수 있고, 이 값은 임플란트 주위염의 심도 분류에 사용될 수 있다. 파노라마 영상에서는 panoptic segmentation이 가능한 신경망 모델을 이용하여 상악동과 하악관을 포함한 주요 구조물들을 구획화할 수 있었다. 따라서, 이와 같이 각 작업에 맞는 심층신경망을 적절한 데이터로 학습시킨다면 진료 보조 수단으로 활용될 수 있다.

주요어 : 의료 인공지능; keypoint detection; panoptic segmentation; 머신러닝; 딥러닝
학 번 : 2017-33206

감사의 글

먼저, 인공지능에 관한 논문을 써 보겠다고 처음 말씀드렸을 때 아직 임상적 의미가 크지 않고 선행연구들도 얼마 없었음에도 불구하고 적극적으로 격려해 주시고 도움 주신 한중석 교수님께 감사드립니다. 한중석 교수님의 긍정적인 조언이 없었다면 제가 이렇게 오랜 기간 연구를 지속하기 어려웠을 것입니다.

또한, 연구에 필수적인 데이터들을 수집할 수 있게끔 도와주시고 올바르게 작업할 수 있도록 지도해주신 허경회 교수님과, 전반적인 논문 작성 및 투고에 대해 가르쳐 주시고 신경 써 주시며 보완할 점들을 일깨워 주신 여인성 교수님, 그리고 참고할 연구들이 많지 않았던 상황에서 어떤 식으로 연구를 진행하면 좋을지 함께 고민하고 조언해주신 윤형인 교수님께 다시 한번 감사의 말씀을 전합니다.

그리고 항상 열과 성을 다해 보철학교실을 지도해주시는 허성주 교수님, 곽재영 교수님, 김성균 교수님, 임영준 교수님, 김명주 교수님, 권호범 교수님, 이재현 교수님께 이 자리를 빌려 감사드립니다. 덧붙여, 함께 고생하고 도와주신 보철학교실 선생님들과, 데이터 준비에 도움 주신 영상치의학교실 선생님들께도 감사드립니다.

또한, 집안을 이끌어 나가느라 고생하시면서도 제가 어떤 선택을 하든 믿어주신 아버지, 힘든 시간 속에서도 식구들을 챙기시고 한결같이 저를 지지해주신 어머니, 컴퓨터에 대해 무지했던 제가 프로그래밍과 인공지능을 공부하면서 묻는 질문들에 답해준 동생 준범에게 감사한 마음을 전합니다. 관련 공부에 큰 도움 된 인터넷상의 수많은 분들과 책 저자분들, 그리고 여러 오픈소스 기여자분들께도 감사드리고 싶습니다.

마지막으로, 처음부터 지금까지, 힘들었던 나날에도 끝까지 곁에서 함께해주었고 지금도 언제나 에너지와 설렘이 되어주는 나의 사랑 선형에게 이 연구를 바칩니다. 우리, 꼭 짜장면 먹으러 갑시다.

2021년 6월

차 준 영