IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 4, 2019, accepted February 9, 2020, date of publication February 24, 2020, date of current version March 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2975826

Automated CNN-Based Tooth Segmentation in Cone-Beam CT for Dental Implant Planning

S. LEE[1], S. WOO[1], J. YU[2], J. SEO[2], J. LEE[2], AND C. LEE[1], (Member, IEEE)
[1]School of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea
[2]Dio Implant, Busan 48058, South Korea

Corresponding author: C. Lee (chulhee@yonsei.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIT) under Grant 2018R1D1A1B07050345.

ABSTRACT Accurate tooth segmentation is an essential step for reconstructing the three-dimensional tooth models used in various clinical applications. In this paper, we propose a convolutional neural network (CNN) based method for fully-automatic tooth segmentation with multi-phase training and preprocessing. For multi-phase training, we defined and used sub-volumes of different sizes to produce stable and fast convergence. To deal with the cone-beam computed tomography (CBCT) images from various CBCT scanners, we used a histogram-based method as a preprocessing step to estimate the average gray density level of the bone and tooth regions. Also, we developed a posterior probability function. Regularizing the CNN models with spatial dropout layers and replacing the convolutional layers with dense convolution blocks further improved the segmentation performance. Experimental results showed that the proposed method compared favorably with existing methods.

INDEX TERMS Cone-beam computed tomography, convolutional neural network, network regularization, posterior probability, tooth segmentation.

I. INTRODUCTION

Cone-beam computed tomography (CBCT) has been widely used for dental diagnosis and treatment planning. CBCT provides three-dimensional tooth models that can be reconstructed for clinical applications, including orthodontic diagnosis [1]–[3] and implant dentistry [4]–[6]. In particular, many implant operations are performed in local dental clinics, where CBCT is often used due to the cost and size factors. CBCT can be used to design surgical guides to minimize angular deviations and displacements between the planned and placed implants. Figure 1 shows an example of a surgical guide. Accurate dental implant surgical guides can address both functional and aesthetic demands [7]. To reconstruct a three-dimensional tooth model, tooth segmentation is an essential step and is typically performed manually by a professional operator. Labeling tooth regions is a time-consuming task and accuracy depends on the operators. Thus, fully-automatic tooth segmentation methods have become an important issue.

FIGURE 1. **An example of surgical guide.**

However, CBCT tooth segmentation faces two main challenges. One challenge arises from dental anatomy. The root areas of the teeth are surrounded by alveolar bone and periodontal ligaments. The thickness of the periodontal ligaments is usually around 0.12 mm [8], which is too thin to be detected in many CBCT images due to the low spatial resolution of typical CBCT scanners [9]. Using a global thresholding method that considers the radiodensities of alveolar bone and

The associate editor coordinating the review of this manuscript and approving it for publication was Yuan Zhang.

teeth can be a solution since the radiodensities of the alveolar bone of maxillae and mandibles range from 800 to 1580 HU (Hounsfield unit) [10] whereas those of tooth components such as dentin and cementum show slightly higher values. However, in practice, the gray density values of teeth and alveolar bone are similar in the root areas. Moreover, even for objects with the same radiodensity, gray levels may appear differently according to their relative positions [11]. Therefore, teeth and their surrounding structures cannot easily be separated by a simple thresholding method [12]. To address this problem, some approaches have utilized adaptive thresholding methods [13], [14]. However, these approaches have some limitations since they require manual selection of the reference slice to select the initial contours. Also, the methods may not be used for CBCT images with metal artifacts. On the other hand, level-set based methods have shown more promising performance [15]–[17]. These methods are semi-automatic and require a manually selected initial contour or a seed point for each tooth.

Also, a starting slice where all the teeth are clearly separated must be manually selected, which may not be possible for CBCT data sets with severe metal artifacts. A graph cut based algorithm [18] also requires background and foreground initializations.

Another challenge is metallic objects that can produce severe streak artifacts. These artifacts can result in degraded CBCT image quality since transmitted X-rays are significantly attenuated or scattered by metallic objects [19]–[21]. Moreover, lower radiation doses in CBCT images cause more severe artifacts [20]. For tooth segmentation, the metal artifact problem is one of the most challenging problems. To solve this problem, several metal artifact reduction (MAR) methods have been developed [22]. Many MAR methods are based on sinograms that contain raw two-dimensional projections used in CBCT imaging. The back-projection of sinograms produces the reconstruction of the CBCT images and the reverse method is known as forward projection. The linear interpolation method [19] is a simple algorithm that replaces manually segmented metal objects with its neighboring objects in the sinogram domain. The normalized MAR method was introduced in [23], which utilizes an additional normalization method before linear interpolation for smooth approximation. As deep learning has shown good performance across many applications, several approaches based on deep convolutional neural networks (CNN) have been proposed. In [24], uncorrected and pre-corrected images were used as input and target images for fine-tuning pre-trained networks. In this method, the fine-tuned networks were used to predict a CNN prior image from an uncorrected image and the back-projection of the CNN prior image was used for the correction of the sinogram. Although there have been efforts to develop the MAR methods, there are still no solutions for irregularly shaped and highly attenuating metal objects [22]. Some researchers have studied CNN-based tooth segmentation methods for two-dimensional panoramic radiographs [25], three-dimensional mesh data [26] and detection algorithms [27]. Recently, Ma proposed a tooth segmentation method based on CNN and level-set methods [28]. However, this method was proposed mainly for dental root segmentation. Cui proposed ToothNet for tooth segmentation and classification in CBCT [29]. The method first extracts edge maps using a CNN model. Then, the method performs ROI extraction, segmentation, and classification using another CNN model that uses the edge maps.

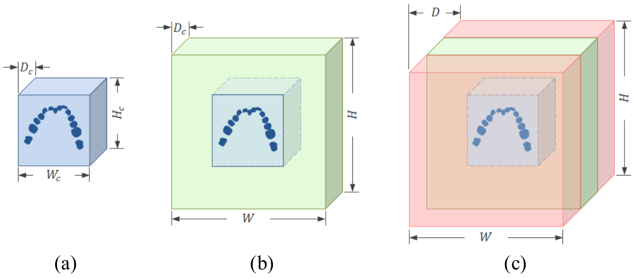


FIGURE 2. Three different sub-volumes of a CBCT volumetric data set: (a) teeth sub-volume, (b) teeth-containing slices, and (c) the entire CBCT volume.

In this paper, we proposed a fully automated CNN-based tooth segmentation method based on the U-Net structure [30]. For reliable training with a limited number of training samples, we proposed a multiphase learning method. In the preprocessing step, we proposed a histogram-based method to normalize the intensity differences to produce robust performance. Then, we computed the probability that a voxel belongs to the tooth region and defined a posterior probability map (PPM), which was inputted to the CNN to improve performance.

The rest of this paper is organized as follows. The multi-phase training method is proposed in Section II.A. The histogram-based preprocessing method and the posterior probability map (PPM) are presented in Section II.B. Section III shows the experimental results with over 100 CBCT data sets. Finally, some conclusions are presented in Section IV.

II. METHODOLOGY

A. TRAINING STRATEGY: MULTI-PHASE TRAINING

It has been reported that the number of tooth voxels is about 1% to 3% of the total voxels of CBCT data. Also, some CBCT slice images do not have any tooth voxels. This imbalanced class distribution can slow down the network convergence during training. A possible solution is to under-sample the major class (i.e. the non-tooth voxels). Since tooth voxels in the CBCT volume are locally concentrated, data sets can be easily under-sampled by cropping the volume. However, networks trained with only under-sampled data can produce many false positives in the non-tooth regions. To solve this problem, we propose multi-stage training. We first defined three different sub-volumes (s_a: teeth sub-volume, s_b: teeth-containing slices, s_c: entire CBCT volume) and sequentially used them to train the networks. We assumed that the CBCT data set had D slices of $W \times H$ pixel images (Figure 2(c)), which represents the entire CBCT volume. The teeth-containing slice (D_c) has at least one tooth voxel (Figure 2(b)). The teeth sub-volume is defined as a smaller sub-volume containing all the tooth voxels (in Figure 2(a)). These three different volumes (teeth sub-volume, teeth-containing slices, and the entire CBCT volume) were sequentially used to train networks, starting first with the teeth sub-volume images. The network was first trained using the slices from volume s_a. Then, the slices from volume s_b were used to further train the network. Then, volume s_c was used. In the three phases, we used the same patch size (192 × 192) and the patches (2D slices) were randomly cropped from the different volumes.

B. PREPROCESSING

As described in [16], the intensity prior is one of the most important factors for tooth segmentation in traditional approaches such as level-set based methods. In this section, we describe the preprocessing method we used to automatically integrate the intensity prior to the CNN models. This method consists of CBCT normalization and posterior probability estimation.

1) CBCT NORMALIZATION

We used 102 CBCT data sets, obtained by using 14 different CBCT scanner models from 7 different manufacturers. Most of the data sets were low-dose data sets since low-dose CBCT is typically used for dental implant planning. Due to the wide range of CBCT scanner models and low-dose CBCT, the image quality varied greatly in terms of resolution and quality. Some images were of poor quality.

Most CBCT devices use the gray density value scale, which is similar to the HU scale. The difference between the HU scale and the gray density value scale is that the HU values are absolute whereas the gray density values are not. Thus, CBCT device manufacturers have their ways of interpreting gray density values as HU-like values. Also, CBCT imaging quality differs depending on the types of X-ray emitters and sensors, scan time, types of field-of-view (FOV), voxel size, etc. Moreover, the parameters used for CBCT scanners can be adaptively chosen according to the patient's conditions. To minimize the gray level differences between the CBCT data sets, we used a normalization technique.

The goal of this normalization technique was to find the soft tissue HU value and the global contrast value of each CBCT data set.

Once these two values are estimated, the HU values of other objects can be easily calculated. To adaptively find the appropriate HU values for soft tissues, we used a histogram-based method. For each CBCT data set, a 256-bin histogram was computed (blue curves in Figure 3(a-d)). The histograms usually had two peaks and one spike, as shown in Figure 3(a). Two peaks near -1000 HU and 100 HU represent the air and soft tissue regions, respectively. Spikes occurred since the voxels outside of the field of view (FOV) had almost the same gray level. On the other hand, there were no peaks for bone, tooth, and metal artifacts. However, as can be seen in Figure 3(b-d), the spikes appeared anywhere on the histograms. After these spikes were suppressed by median filtering, the histograms were renormalized (SR-hist, red lines in Figure 3). Then the soft tissue HU value (x_s) was obtained by finding the center of the soft tissue peak. It is reported that the soft tissue radiodensity is around 20-40 HU and that the bone radiodensity is around 1000 HU [31]. Thus, the bone HU $x_{b,fixed}$ was calculated as follows:

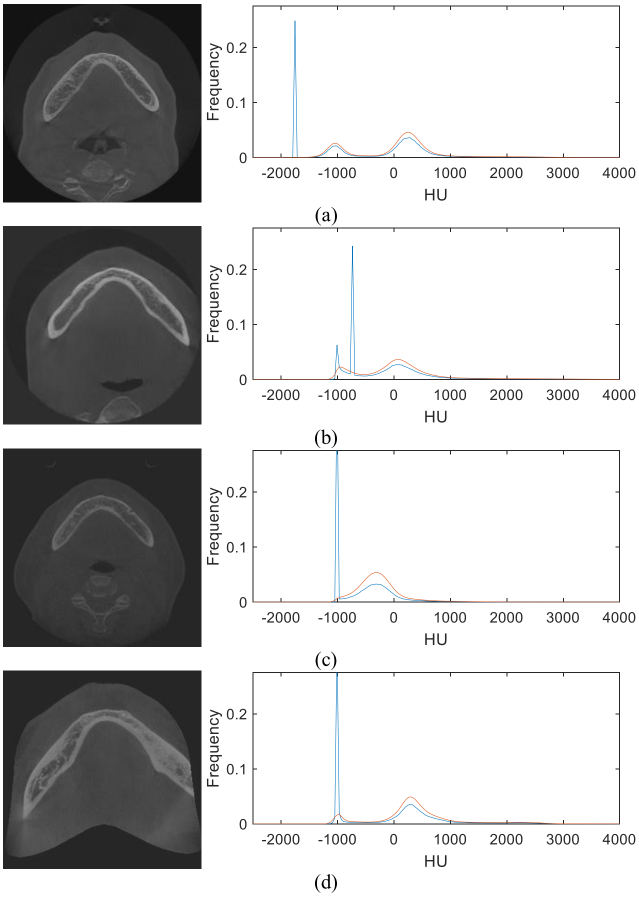


FIGURE 3. Four types of CBCT histograms with their corresponding CBCT slice images (Blue: histogram, Red: normalized histogram): (a) a spike outside of two peaks, (b) a spike between two peaks, (c) a spike at the tail of one peak, and (d) a spike far from a peak.

$$x_{b,fixed} = x_s + d \tag{1}$$

where d represents a constant that denotes the difference between the soft tissue and bone HU. Although the bone HU values of CBCT showed some variations, we observed that $d = 950$ was applicable to most of the CBCT data sets.

However, there were some data sets in low contrast so that the HU differences between bone and soft tissue regions were relatively small compared to other data sets. To address this problem, we estimated the global contrast of each CBCT data set. Since the HU values for air and water are defined as -1000 HU and 0 HU, the image contrast can be estimated by the HU difference between air and water. Alternatively, the air and soft tissue HU values of the SR-hist can be used

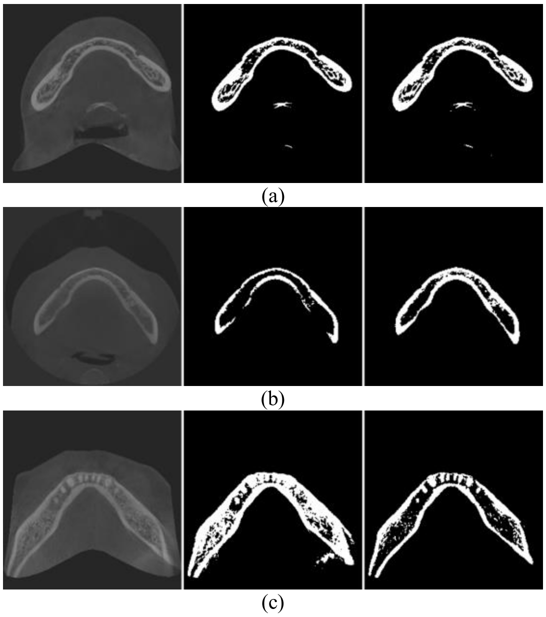


FIGURE 4. Some examples of image thresholding results with fixed and adaptive bone HUs. (a) Well segmented, (b) under-segmented, and (c) over-segmented examples with original CBCT slice images (left), segmentation results with fixed bone HUs (middle) and adaptive bone HUs (right).

for the estimation. However, as discussed above, there were some cases when the air peaks were blurred or invisible. Since "soft tissue peaks" existed for all cases, we conjectured that the HU range of the soft tissue voxels may be narrow if the global image contrast is low. Therefore, in the proposed method, the Gaussian function was used for the soft tissue peak estimation:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-b)^2}{2\sigma^2}} \tag{2}$$

where b represents the mean of the peak and σ denotes the standard deviation. We checked the following conditions:

(1) There were two peaks for the air and soft tissue

(2) The peak difference between the air and soft tissue was 1000-1100 HU.

If the CBCT set met these conditions, the bone HU value (x_b) was calculated as follows:

$$x_{b,adapt} = x_s + \rho_c^d \tag{3}$$

where $\rho_c = \frac{\sigma}{\mu_c}$ is the estimated relative contrast and μ_c denotes the mean of σ of the CBCT data sets in normal conditions. In this paper, $\mu_c = 131.7647$ was used. Figure 4 shows some examples of thresholding results when a fixed bone HU (the middle column) and adaptive bone HUs (the right column) were used. Applying the adaptive method to the CBCT data sets produced more stable performance.

2) POSTERIOR PROBABILITY ESTIMATION

The bone and tooth radiodensities (except for the pulp tissue in a tooth) were relatively higher than those of other objects (Figure 5(a)). The bone and tooth components are similar in radiodensity values as those shown in Figure 5(b). On the other hand, enamel and metal objects have higher values than any other objects. Thus, the probability that a voxel belonging to tooth regions was high if its HU value was high.

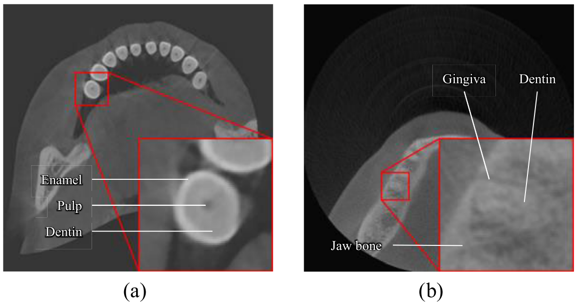


FIGURE 5. Examples of CBCT image slices. (a) Individual teeth appear clearly in tooth neck area. (b) Teeth with a jaw bone (mandible) in tooth root area.

For an arbitrary CBCT data set, consider a random variable s, which denotes the category of a voxel where $s = t$ for teeth and $s = t_c$ for non-teeth. We let $P(t)$ be the prior probability that a voxel belongs to the tooth region and $p(x)$ denote the probability density function of x. Then, $p\,(t|x)$ can be computed using the Bayes formula:

$$p\,(t|x) = \frac{p\,(x|t)\,P(t)}{p(x)} \tag{4}$$

The probability density functions $p(x)$ and $p\,(x|t)$ were estimated by SR-hist (Section II.B.1) and the teeth histogram. A CBCT data set was regarded as a set of observations (voxels) $\{x_i\}_{i=1,2,\cdots,N}$ of a random variable X where N represents the number of voxels. Then, the n-bin histogram with the bin width h is defined as:

$$\hat{f}_n(x) = \frac{\sum_{i=1}^{N} \mathbf{1}\{x_a \le x_i < x_b\}}{Nh} \tag{5}$$

where $x_a = x - h/2$, $x_b = x + h/2$, and $\mathbf{1}\{\cdot\}$ denotes the indicator function. Then, $p\,(x|t)$ was estimated as follows:

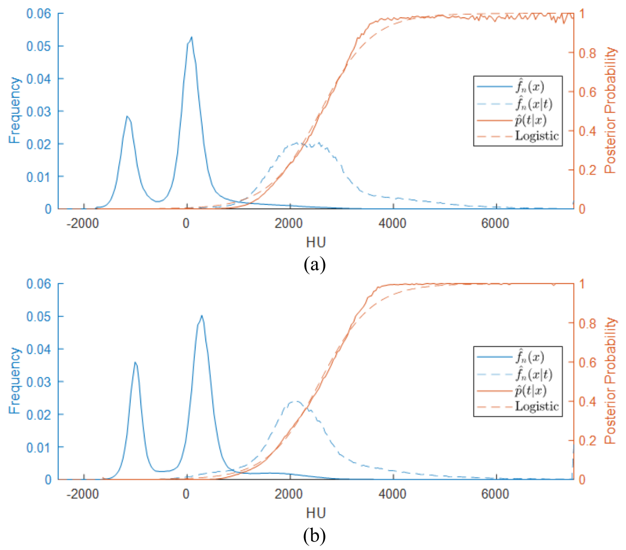
$$p\,(x|t) \approx \hat{f}_n(x|t) = \frac{\sum_{i=1}^{M} \mathbf{1}\{x_a \le x_i < x_b\}}{Mh}. \tag{6}$$

where M represents the number of voxels in the tooth region. Then, by substituting (5) and (6) into (4),

$$p\,(t|x) \approx \hat{p}\,(t|x) = \frac{\hat{f}_n(x|t)P(t)}{\hat{f}_n(x)} = \frac{\sum_{i=1}^{M} \mathbf{1}\{x_a \le x_i < x_b\}}{\sum_{i=1}^{N} \mathbf{1}\{x_a \le x_i < x_b\}} \tag{7}$$

Figure 6 shows some examples of $\hat{f}_n(x)$, $\hat{f}_n(x|t)$, and $\hat{p}\,(t|x)$.

For a given CBCT data set, only $\hat{f}_n(x)$ is given and the other functions ($\hat{f}_n(x|t)$, $P(t)$, and $\hat{p}\,(t|x)$) were unknown unless the teeth ground-truth is available. In other words, there is no prior information about voxels.



(a)

(b)

FIGURE 6. **Histograms of HU values for tooth voxels and the entire CBCT with their estimated posterior probabilities computed with (a) the training set and (b) the validation set.**

To resolve this problem, we assumed that $\hat{f}_n(x)$ was a mixture of the weighted Gaussian distributions of the various objects in a CBCT data set:

$$\hat{f}_n(x) = \sum_i w_i n\left(x; \mu_i, \sigma_i\right) \tag{8}$$

where w_i is a weight of category i and $n\left(x; \mu_i, \sigma_i\right)$ denotes the Gaussian distribution with the mean μ_i and standard deviation σ_i. Here, the category includes air, soft tissue, bone (b), teeth (t), etc. Then, $\hat{f}_n(x|t)$ can be written as:

$$\hat{f}_n(x|t) = n\left(x; \mu_t, \sigma_t\right) \tag{9}$$

By substituting (8) and (9) into (7),

$$\hat{p}\left(t|x\right) = \frac{n\left(x; \mu_t, \sigma_t\right)\left(w_t / \sum_i w_i\right)}{\sum_i w_i n\left(x; \mu_i, \sigma_i\right)} = \frac{w_t n\left(x; \mu_t, \sigma_t\right)}{\sum_i w_i n\left(x; \mu_i, \sigma_i\right)} \tag{10}$$

since $P(t) = w_t / \sum_i w_i = w_t$ by the definitions. Then (10) is written as:

$$\hat{p}\left(t|x\right) = \frac{1}{1 + \sum_{i \neq teeth} \frac{w_i n(x;\mu_i,\sigma_i)}{w_t n(x;\mu_t,\sigma_t)}}. \tag{11}$$

Since the radiodensities of bone and teeth (dentin and cementum) are similar, μ_t and μ_b may be similar. We assumed that the effect of the air and soft tissue for $\hat{p}\left(t|x\right)$ was ignorable and $\sigma_t \approx \sigma_b \approx \sigma$. Then, (11) was simplified as:

$$\hat{p}\left(t|x\right) \approx \frac{1}{1 + \frac{w_b \exp\left(-(x-\mu_b)^2/2\sigma^2\right)}{w_t \exp\left(-(x-\mu_t)^2/2\sigma^2\right)}} = \frac{1}{1 + \exp\left(-k\left(x - x_0\right)\right)} \tag{12}$$

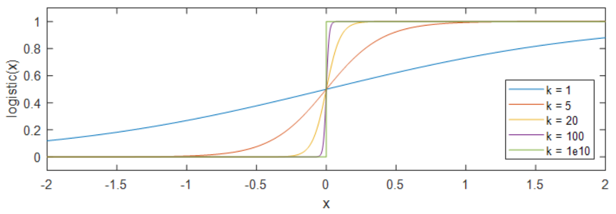


FIGURE 7. **Logistic functions with various values of k where $x_0 = 0$.**

where $k = \log \frac{w_b}{w_t} \cdot \frac{\mu_t - \mu_b}{\sigma^2}$, $x_0 = \frac{\mu_t + \mu_b}{2}$. From (12), we modeled the posterior probability $p\left(t|x\right)$ as a logistic function $L(x; x_o, k)$:

$$L(x; x_0, k) = 1/\left(1 + \exp\left(-k\left(x - x_0\right)\right)\right) \tag{13}$$

$$p\left(t|x\right) \approx \hat{p}\left(t|x\right) = L(x; \mu_t, k) \tag{14}$$

where x_0 and k denote the midpoint and steepness of the logistic curve. Figure 7 shows some examples of the logistic functions with various values of k with $x_0 = 0$. The logistic functions were used for the estimation of $\hat{p}\left(t|x\right)$ (red dashed curves in Figure 6(a-b)). We set $L(x_0; x_0, k) = 0.5$, which led to normalization effects. Using the estimated soft tissue HU and contrast values computed in Section II.B.1, μ_t and μ_b were estimated. However, the calculation of k required w_b, w_t, and σ, which are unknown. Thus, k was empirically selected.

3) POSTERIOR PROBABILITY MAP

To use the posterior probability in CNNs, a posterior probability map (PPM) was computed by applying the logistic function to all the voxels in the CBCT volume. Figure 8 shows an original CBCT slice image and the corresponding PPMs with various k values.

When high k values were used, image contrast was enhanced, which improved tooth segmentation. Mandibles, cervical vertebrae, and teeth appeared more discernible for $k \geq 10$. Air regions and soft tissue regions almost disappeared when $k = 20$ and $k = 50$.

To evaluate the potential benefits of multi-channel inputs, experiments were conducted using different settings. Along with the original U-Net used in Section III.A, we also tested the U-Net with two-channel inputs, which used the original images and the normalized images:

$$f_{norm}(x) = \rho_c\left(x - \mu_t\right) + d_t \tag{15}$$

where μ_t and ρ_c are defined in Section II.B.1, and d_t denotes the HU mean difference between the soft tissue regions and teeth under normal conditions.

Figure 9 shows a performance comparison. The red and green boxes show the U-Net performance with different numbers of input channels. Increasing the number of input channels was effective for the first phase with the volume s_a whereas there were negligible differences between the U-Nets with one and two-channel inputs in the second and third phases. It appears that using $f_{norm}(x)$ was helpful for the

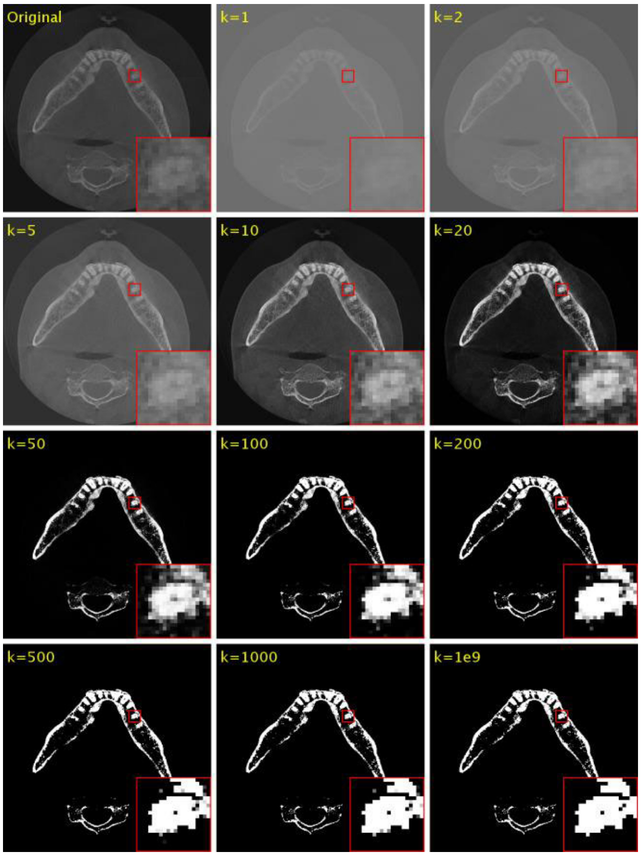


FIGURE 8. CBCT examples after applying logistic functions with various k values.

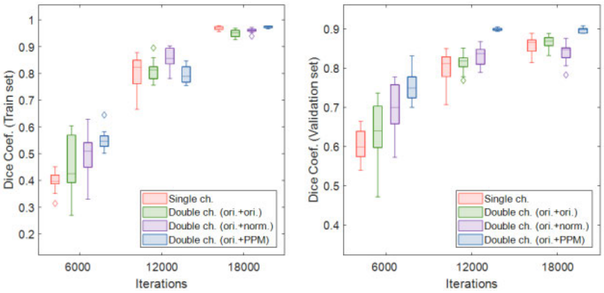


FIGURE 9. Performances of the U-Net with different inputs with the training set (left) and the validation set (right).

first and second phases. On the other hand, the U-Net with the PPM showed improved performance.

We also conducted experiments with various k values while keeping the other parameters unchanged. Figure 10 shows the dice coefficient values. It appears that using $k = 20$ may be the best solution. Therefore, we used $k = 20$ in this paper.

4) ADDITIVE UNIFORM NOISE FOR NETWORK STABILITY

In the real world, input images are prone to various types of noise, which may lead to inaccurate estimations of μ_t. Such inaccurate estimations may cause performance degradation. To understand the effects of inaccurate estimations, we added additive noise to μ_t as follows:

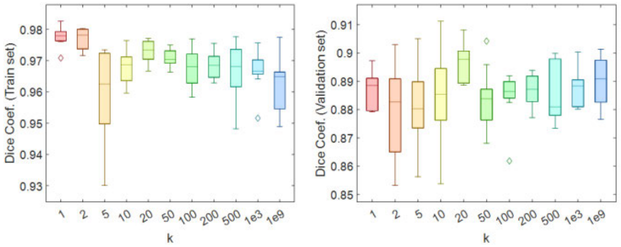
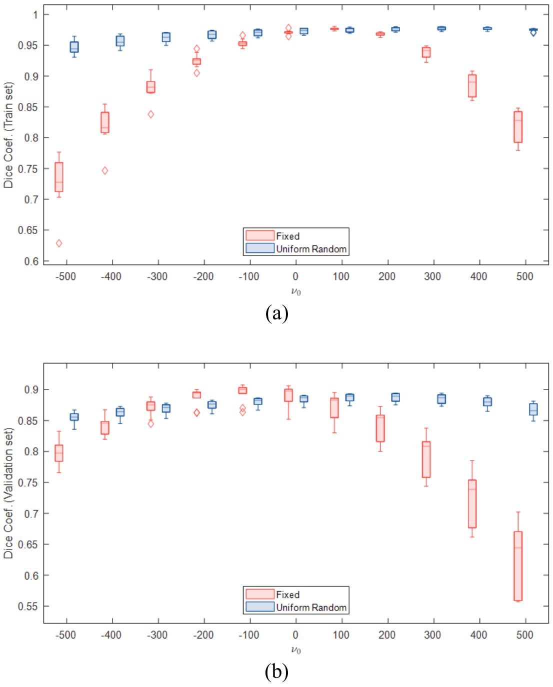


FIGURE 10. U-Net performances with different k values with the training set (left) and the validation set (right).



(a)

(b)

FIGURE 11. U-Net performances with different additive noise factors ν_0 (blue box: the network trained with the added noise, red box: the network trained without the added noise).

$$L(x; \mu_t, k, \nu_0) = 1/\left(1 + \exp\left(-k\left(x - (\mu_t + \nu_0)\right)\right)\right) \quad (16)$$

where ν_0 is the noise factor.

The red boxes in Figure 11 show how performance was affected by the noise. As expected, performance was significantly degraded as $|\nu_0|$ increased. To solve this problem, we added noise of various levels (uniform distribution: $U(-200, 200)$) during the training process. Figure 11 shows the performance of the network trained with the added noise. The network trained with the added noise showed stable performance even when the noise levels were high.

C. NETWORK ARCHITECTURE: UDS-NET

The proposed network is based on the U-Net architecture [30].

As shown in Figure 12, the dense blocks and spatial dropout layers are the main differences between the U-Net architecture [30] and the proposed network. In the proposed network, some convolutional layers were replaced with

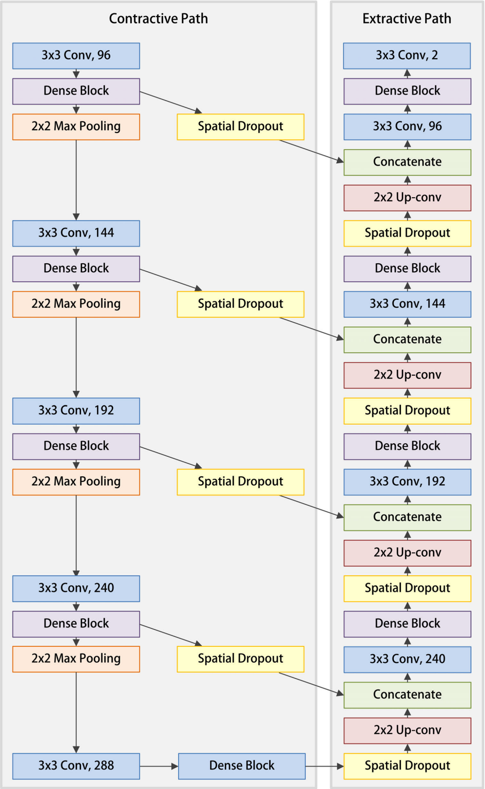


FIGURE 12. The proposed network.

dense blocks. Spatial dropout layers were inserted between the dense blocks and up-convolution layers in the extractive path. Also, spatial dropout layers were added to skip connections between the contractive and extractive paths. Compared to the original U-Net, the proposed UDS-Net (U-Net + dense block + spatial dropout) achieved better segmentation performance with a reduced number of parameters.

1) DENSE BLOCK

Dense connectivity, introduced in [32], has shown performance improvement in several image classification tasks using fewer parameters compared to other methods such as ResNet [33]. In a dense block, a convolutional layer receives features of its preceding layers. In the UDS-Net method, the dense blocks improved tooth segmentation and reduced the number of parameters. Figure 13 shows the dense block used in the proposed method where 1×1 convolutions with 192 output channels and 3×3 convolutions with 48 output channels were applied. Then, the output features were combined with the input features by concatenation.

2) SPATIAL DROPOUT

In U-Net [30], the output features from the contractive path were cropped and concatenated with the expansive path. Since overfitting sometimes occurred, the spatial dropout [34], a network regularization technique, was applied to the proposed UDS-Net method (Figure 12).

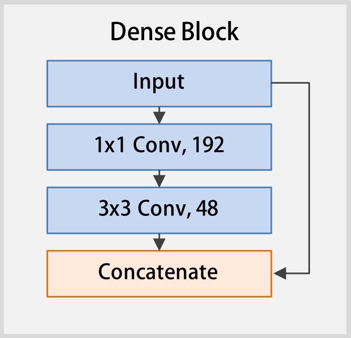


FIGURE 13. Dense block used in the proposed UDS-Net.

III. EXPERIMENTAL RESULTS

A. DATABASES

In the experiments, we used 102 CBCT data sets, obtained using 14 different CBCT scanner models produced by seven different manufacturers (including unknown scanner models and manufacturers). Each CBCT data set consists of multiple two-dimensional CBCT slices (16-bit data) in the DICOM file format. The width and height were identical, which ranged from 264 to 800 pixels. The number of slices ranged from 264 to 727 slices. Also, the physical sizes of a voxel in terms of width, height, and depth were always the same (from 0.15 mm to 0.30 mm).

We manually labeled all the tooth voxels for two sets (CBCT_#1, CBCT_#19). From the other 100 CBCT data sets, we selected and manually labeled five slices (5-slice sets):

1) A slice containing mandibles but not teeth.
2) A slice containing mandibular teeth surrounded by mandibles.
3) A slice containing maxillary teeth surrounded by the maxilla.
4) A slice containing mandibular teeth but not surrounded by mandibles.
5) A slice containing maxillary teeth but not surrounded by the maxilla.

Since some CBCT data sets failed to satisfy all the five conditions, they might have a fewer slices.

First, all the CBCT image slices were resized to 0.4 mm in the direction of the sagittal and coronal axes. For example, a 264×264 pixel image with a voxel size of 0.30×0.30 was resized with $264 \times (0.3/0.4) = 198$ pixels in each direction. Since both the U-Net and UDS-Net methods had four 2×2 max-pooling layers, we resized the images so that both the width and height were multiples of 16. For example, $\lceil 198/16 \rceil \times 16 = 208$ pixels. For training, missing pixels were extrapolated with the mirrored images. For evaluation, missing pixels were zero-filled.

B. EXPERIMENTS ON TRAINING METHODS

1) SELECTING THE SEQUENCE OF VOLUMES FOR MULTIPHASE LEARNING

Multi-stage training provided stable and faster convergence. The proposed training method was tested using the U-Net

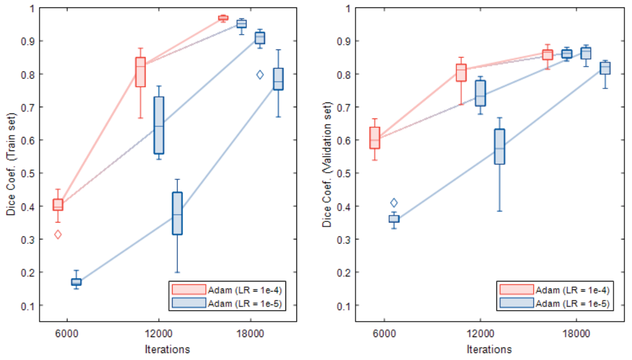


FIGURE 14. Performance of U-Net with multi-phase training $\langle s_a, s_b, s_c \rangle$ on training and validation set for the four learning rate schedules.

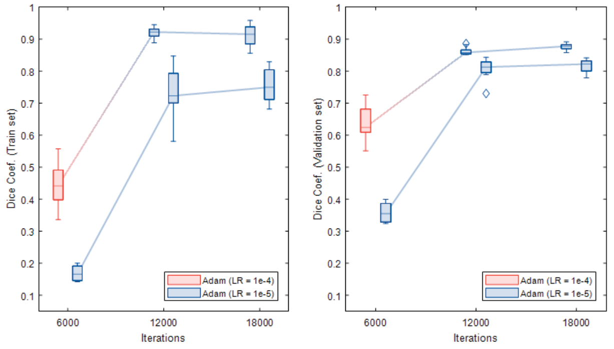


FIGURE 15. Performance of U-Net with multi-phase training $\langle s_a, s_c, s_c \rangle$ on training and validation set with different learning rate combinations: $[10^{-4}, 10^{-5}, 10^{-5}]$, $[10^{-5}, 10^{-5}, 10^{-5}]$.

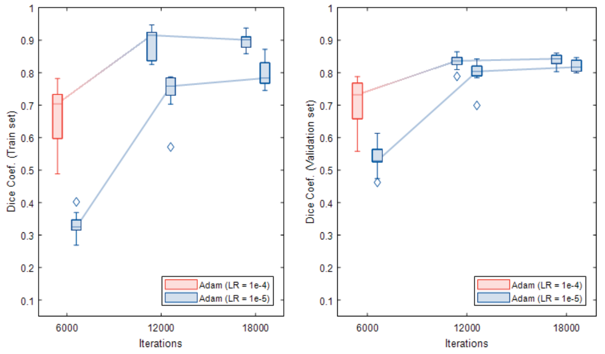


FIGURE 16. Performance of U-Net with multi-phase training $\langle s_b, s_c, s_c \rangle$ on training and validation set with different learning rate combinations: $[10^{-4}, 10^{-5}, 10^{-5}]$, $[10^{-5}, 10^{-5}, 10^{-5}]$.

architecture [30]. We used the Adam optimizer [35], which is commonly used in several biomedical segmentation methods based on U-Net [36]–[39]. To evaluate the multi-phase training method, we conducted preliminary tests. We compared four different sequences of multi-phase training: $\langle s_a, s_b, s_c \rangle$ (Figure 14), $\langle s_a, s_c, s_c \rangle$ (Figure 15), $\langle s_b, s_c, s_c \rangle$ (Figure 16) and $\langle s_c, s_c, s_c \rangle$ (Figure 17). For each phase, the number of iterations was set to 6000 and the learning rates were set to either 10^{-4} or 10^{-5}. Depending on the learning rate, the network may not have converged.

To find the working range of learning rates, we tested several combinations of different learning rates. In particular,

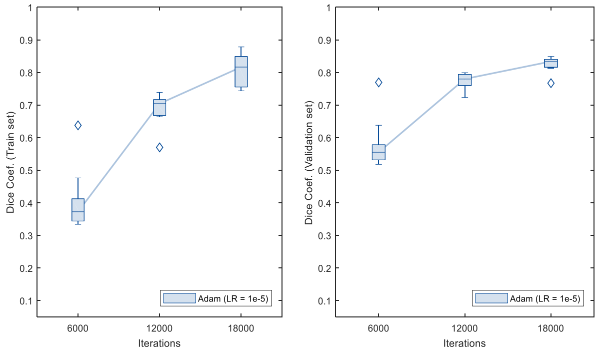


FIGURE 17. Performance of U-Net without multi-phase training $\langle s_c, s_c, s_c \rangle$ on training and validation set: $[10^{-5}, 10^{-5}, 10^{-5}]$.

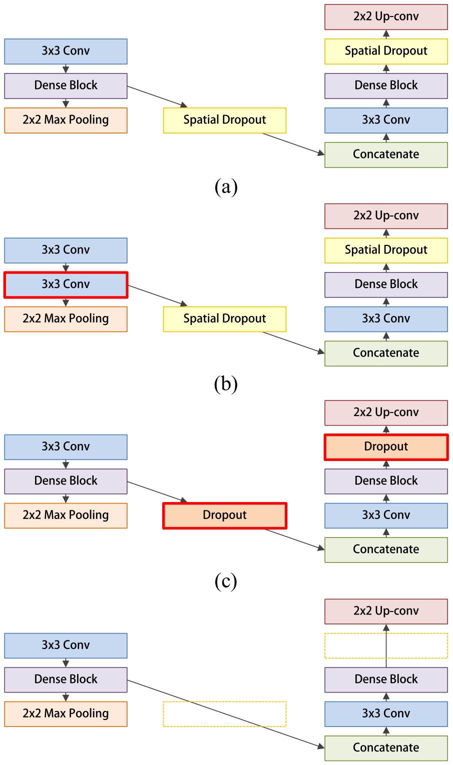


FIGURE 18. Some modifications of the UDS-Net.

we tested four learning rate schedules: $[10^{-4}, 10^{-4}, 10^{-4}]$, $[10^{-4}, 10^{-4}, 10^{-5}]$, $[10^{-4}, 10^{-5}, 10^{-5}]$, and $[10^{-5}, 10^{-5}, 10^{-5}]$. Each experiment was conducted 10 times. Figures 14-17 show performance comparison. The networks that failed to converge were not shown in the Figures. For example, $\langle s_c, s_c, s_c \rangle$ converged only for $[10^{-5}, 10^{-5}, 10^{-5}]$ (Figure 17). Based on these results, we used $\langle s_a, s_b, s_c \rangle$ and $[10^{-4}, 10^{-4}, 10^{-5}]$.

2) ABLATION STUDY ON DENSE BLOCKS AND SPATIAL DROPOUT

To evaluate the benefits of the dense blocks and spatial dropout layers in the proposed UDS-Net method, several experiments were conducted with different modifications to the proposed UDS-Net method. Figure 18(a) represents a part of the proposed UDS-Net method and Figure 18(b-d)

TABLE 1. Tooth loss statistics of the data.

#Tooth Loss	# Patients
1	3
2	2
3	5
4	9
5	17
6	18
7	16
8	16
9	9
10	2
11	2
12	1
13	1
17	1

TABLE 2. Statistics of metal artifacts (MA).

#Patients with MA		#Patients without MA
#Patients with Dental Implant	#Patients with Metal Structures	
26	73	3

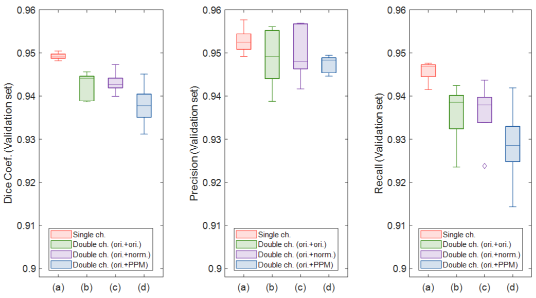
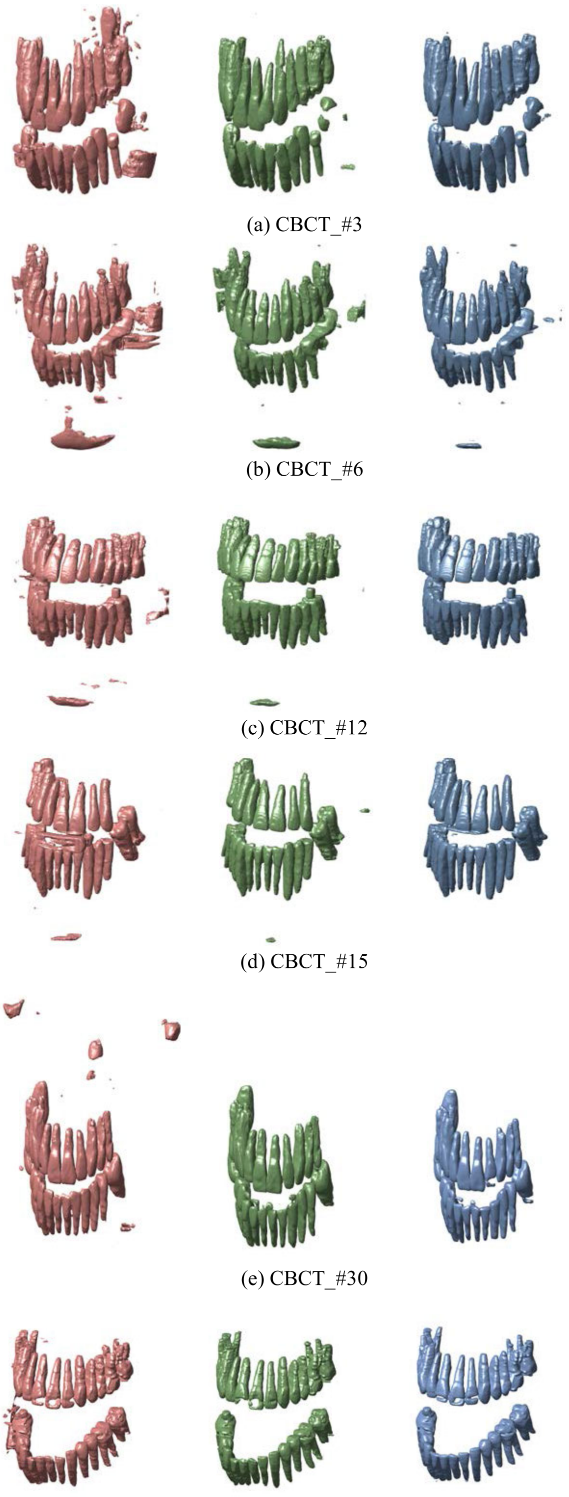


FIGURE 19. Performance comparison of the modifications (dice, precision, recall). (a) the proposed UDS-Net, (b) the dense blocks are replaced with convolutional layers, (c) the spatial dropout layers are replaced by dropout layers, (d) without dropout layers.

shows some modifications. First, the dense blocks in the proposed UDS-Net method were replaced with convolutional layers (Figure 18(b)), which is the same as the original U-Net architecture except for the spatial dropout layers. Also, the spatial dropout layers in Figure 18(a) were replaced by dropout layers (Figure 18(c)). The networks without dropout layers were also tested (Figure 18(d)). For all the cases, five experiments were conducted. As shown in Figure 19, the proposed UDS-Net method (Figure 18(a)) showed the best performance.

C. PERFORMANCE ANALYSIS OF TOOTH SEGMENTATION

The CBCT data sets were divided into three sets: training, validation, and test. For training, we used CBCT_#1 and the 5-slice sets from CBCT_#35 to CBCT_#102 (69 datasets, 1066 images). CBCT_#19 was used for validation (1 dataset, 400 images). We used the 5-slice sets from CBCT_#2 to CBCT_#18 and from CBCT_#20 to CBCT_#34 for testing (32 datasets, 151 images). We also applied the trained network to the entire slices of CBCT_#2∼CBCT_#18 and CBCT_#20∼CBCT_#34 data sets.


(a) CBCT_#3

(b) CBCT_#6

(c) CBCT_#12

(d) CBCT_#15

(e) CBCT_#30

(f) CBCT_#32

FIGURE 20. Some examples of tooth segmentation results with different CNN models (red: baseline, green: baseline with PPM and multi-phase training, blue: proposed method).

TABLE 3. Detailed information of network training with or without multi-phase training.

Phase	#Iteration	Training set	w/o multi-phase		w/ multi-phase	
			Sub-volume	Adam LR	Sub-volume	Adam LR
1st	6k	CBCT_#1	s_c	10^{-5}	s_a	10^{-4}
2nd	6k	CBCT_#1	s_c	10^{-5}	s_b	10^{-4}
3rd	6k	CBCT_#1	s_c	10^{-5}	s_c	10^{-5}
4th	60k	5slice sets	s_c	10^{-5}	s_c	10^{-5}

TABLE 4. Performance comparison with combinations of the proposed methods.

Model	# Parameters	Validation set			Test set		
		Dice	Recall	Precision	Dice	Recall	Precision
Level-set based method [15]	-	0.820	0.730	0.933	0.790	0.769	0.813
Baseline	30M	0.889	0.953	0.833	0.891	0.938	0.848
Proposed method	14M	0.938	0.952	0.924	0.918	0.932	0.904
Proposed method w/o spatial dropout	14M	0.930	0.930	0.930	0.916	0.918	0.913
Proposed method w/o multi-phase	14M	0.934	0.954	0.916	0.910	0.943*	0.878*
Proposed method w/o PPM	14M	0.935	0.956	0.915	0.912	0.945*	0.881*

* indicates that the results of the ablation study using the test set showed statistically significant differences compared to the proposed method (spatial dropout, multi-phase and PPM). We used the t-test with a confidence level of 95%.

Generally, patients who need implant operations are likely to be elderly and they may have experienced multiple tooth losses. Table 1 shows the tooth loss statistics of the data used in this paper. All the patients had at least one tooth loss. Also, many patients had tooth decay treatments and other dental treatments, which might have left some metal structures. Some patients already had implant operations. Consequently, metal artifacts were observed in most patients. Table 2 shows the statistics of metal artifacts. Ma's method was proposed for tooth root segmentation and may not be suitable for tooth segmentation for implant planning. Also, due to missing teeth and heavy metal artifacts, Cui's method might need to be significantly modified for this kind of application, which performs segmentation and classification using edge maps.

Thus, for performance comparison, we implemented two methods. The first one was a level-set based method based on [15], which required manual selection of initial conditions.

We also modified the original U-Net method [30] for tooth segmentation and used it as the baseline method. There were four training phases used in the training (Table 3).

After the three phases of multi-phase training with CBCT_#1, the 5-slices sets of the training set (CBCT_#35 to CBCT_#102) were used for the fourth phase training.

Different sub-volumes and different learning rates were used (Table 3).

We used data augmentation methods: translations (±5%), resize (±30%), vertical flip and horizontal flip. In other words, the translations and image resizing were limited to 5% and 30% of the original image size, respectively.

For all the networks, the Adam optimizer [35] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ was used and the mini-batch size was set to 2. Also, the binary cross-entropy function was used as a loss function. TensorFlow [40] and Keras [41] libraries were used for network implementation. We used an Intel®Xeon E5-2620v3 CPU and an NVidia Titan X (Pascal) with 12GB graphic RAM and 48GB RAM.

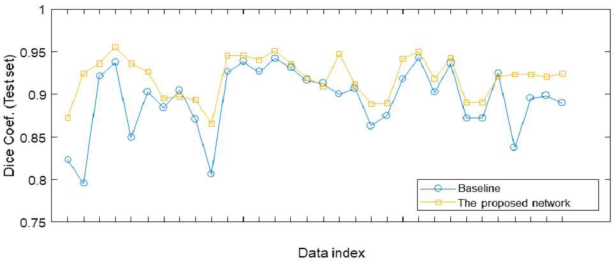
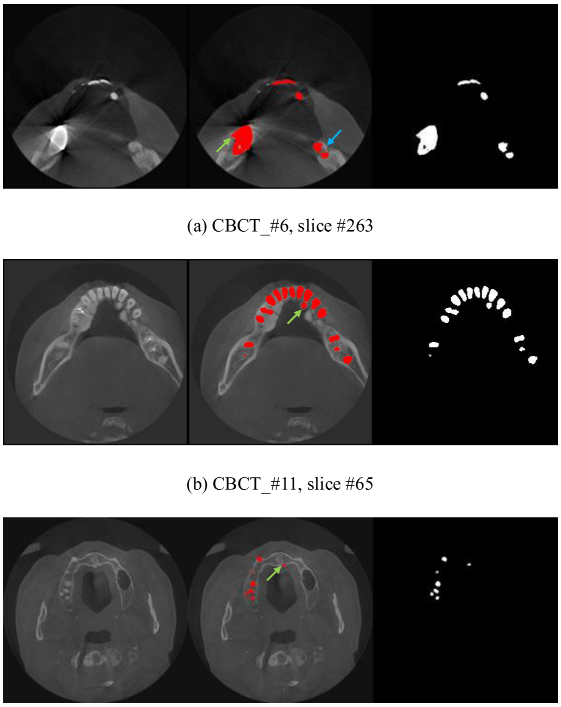


FIGURE 21. Performance comparison for each CBCT data set.

Table 4 shows the evaluation results on the validation and test sets. The level-set based method showed inferior performance compared to CNN based methods (the dice value was 0.820 for the validation set and 0.790 for the test set). Although the baseline method showed relatively good recall performance (0.953 for the validation and 0.938 for the test set), it produced many false positives (validation precision: 0.833; test precision: 0.848).

The proposed method noticeably improved the precision performance. In Table 4, we also tested the proposed method without one of the three key features (PPM, multi-phase training, spatial dropout). Also, the proposed UDS-Net method produced better performance while using fewer parameters than the baseline method.



(a) CBCT_#6, slice #263

(b) CBCT_#11, slice #65

(c) CBCT_#20, slice #336

FIGURE 22. Examples of false positives and false negatives. Red areas represent the tooth segmentation results. The green arrow indicates false positives whereas the blue arrow indicates false negatives.

Figure 20 shows some segmentation results (red: baseline, green: baseline with PPM and multi-phase training, blue: proposed method). The false positives were significantly reduced in the proposed method. Also, errors due to metal artifacts were noticeably reduced in the proposed method (Figure 20(d)), though we did not perform specific operations to remove metal artifacts. It appears that the posterior probability map might provide additional pixel-wise information to the network, which could be helpful for discriminating teeth from metal artifacts.

Figure 21 shows the performance comparison of all the test data. The proposed method produced noticeably improved segmentation performance. Some examples of false positives and false negatives are shown in Figure 22. Red areas represent the tooth segmentation results. The green arrow indicates false positives whereas the blue arrow indicates false negatives. Some portions of the wisdom teeth were usually undetected (Figure 22 (a)), mainly because they were not labeled in the ground truth images. On the other hand, false positives usually appeared on the mandible (Figure 22(b)), maxilla (Figure 22(c)), and streak artifacts (Figure 22(a)).

IV. CONCLUSION

In this paper, a fully automated CNN based tooth segmentation method was proposed for dental CBCT images. In the proposed method, a multi-phase training strategy was used to gradually expand the target volume of the CBCT images. As a preprocessing step, a histogram-based method was proposed to calculate the HU values for the bones and teeth in various CBCT data sets. Based on this information, the posterior probability of a voxel being in a tooth region was estimated when the voxel's HU value was given. Then we defined the posterior probability map (PPM) and used it along with the original CBCT images. We used the spatial dropout technique. With dense block and spatial dropout layers, the proposed method showed improved performance compared to the conventional U-Net architecture.

REFERENCES

[1] P. Mozzo, C. Procacci, A. Tacconi, P. T. Martini, and I. A. B. Andreis, "A new volumetric CT machine for dental imaging based on the cone-beam technique: Preliminary results," *Eur. Radiol.*, vol. 8, no. 9, pp. 1558–1564, Nov. 1998.

[2] S. J. Merrett, N. A. Drage, and P. Durning, "Cone beam computed tomography: A useful tool in orthodontic diagnosis and treatment planning," *J. Orthodontics*, vol. 36, no. 3, pp. 202–210, 2009.

[3] S. L. Hechler, "Cone-beam CT: Applications in orthodontics," *Dental Clinics North Amer.*, vol. 52, no. 4, pp. 809–823, Oct. 2008.

[4] W. C. Scarfe, A. G. Farman, and P. Sukovic, "Clinical applications of cone-beam computed tomography in dental practice," *J.-Can. Dental Assoc.*, vol. 72, no. 1, p. 75, 2006.

[5] C. Angelopoulos, S. Thomas, S. Hechler, N. Parissis, and M. Hlavacek, "Comparison between digital panoramic radiography and cone-beam computed tomography for the identification of the mandibular canal as part of presurgical dental implant assessment," *J. Oral Maxillofacial Surg.*, vol. 66, no. 10, pp. 2130–2135, Oct. 2008.

[6] E. Benavides, H. F. Rios, S. D. Ganz, C.-H. An, R. Resnik, G. T. Reardon, S. J. Feldman, J. K. Mah, D. Hatcher, M.-J. Kim, D.-S. Sohn, A. Palti, M. L. Perel, K. W. M. Judy, C. E. Misch, and H.-L. Wang, "Use of cone beam computed tomography in implant dentistry: The international congress of oral implantologists consensus report," *Implant Dentistry*, vol. 21, no. 2, pp. 78–86, Apr. 2012.

[7] W. Geng, C. Liu, Y. Su, J. Li, and Y. Zhou, "Accuracy of different types of computer-aided design/computer-aided manufacturing surgical guides for dental implant placement," *Int. J. Clin. Exp. Med.*, vol. 8, no. 6, p. 8442, 2015.

[8] W. J. Ralph and J. R. Jefferies, "The minimal width of the periodontal space," *J. Oral Rehabil.*, vol. 11, no. 5, pp. 415–418, Sep. 1984.

[9] D. Brüllmann and R. K. W. Schulze, "Spatial resolution in CBCT machines for dental/maxillofacial applications—What do we know today?" *Dentomaxillofacial Radiol.*, vol. 44, no. 1, Jan. 2015, Art. no. 20140204.

[10] H.-S. Park, Y.-J. Lee, S.-H. Jeong, and T.-G. Kwon, "Density of the alveolar and basal bones of the maxilla and the mandible," *Amer. J. Orthodontics Dentofacial Orthopedics*, vol. 133, no. 1, pp. 30–37, Jan. 2008.

[11] R. Pauwels, R. Jacobs, S. R. Singer, and M. Mupparapu, "CBCT-based bone quality assessment: Are hounsfield units applicable?" *Dentomaxillofacial Radiol.*, vol. 44, no. 1, Jan. 2015, Art. no. 20140238.

[12] O. Nackaerts, M. Depypere, G. Zhang, B. Vandenberghe, F. Maes, R. Jacobs, and S. Consortium, "Segmentation of trabecular jaw bone on cone beam CT datasets," *Clin. Implant Dentistry Rel. Res.*, vol. 17, no. 6, pp. 1082–1091, 2015.

[13] H. Heo and O.-S. Chae, "Segmentation of tooth in CT images for the 3D reconstruction of teeth," *Proc. SPIE*, vol. 5298, pp. 455–467, May 2004.

[14] H. C. Kang, C. Choi, J. Shin, J. Lee, and Y.-G. Shin, "Fast and accurate semiautomatic segmentation of individual teeth from dental CT images," *Comput. Math. Methods Med.*, vol. 2015, pp. 1–12, Aug. 2015.

[15] Z. Xia, Y. Gan, L. Chang, J. Xiong, and Q. Zhao, "Individual tooth segmentation from CT images scanned with contacts of maxillary and mandible teeth," *Comput. Methods Programs Biomed.*, vol. 138, pp. 1–12, Jan. 2017.

[16] H. Gao and O. Chae, "Individual tooth segmentation from CT images using level set method with shape and intensity prior," *Pattern Recognit.*, vol. 43, no. 7, pp. 2406–2417, Jul. 2010.

[17] D. X. Ji, S. H. Ong, and K. W. C. Foong, "A level-set based approach for anterior teeth segmentation in cone beam computed tomography images," *Comput. Biol. Med.*, vol. 50, pp. 116–128, Jul. 2014.

[18] L. Hiew, S. Ong, and K. W. Foong, "Tooth segmentation from cone-beam CT using graph cut," in *Proc. 2nd APSIPA Annu. Summit Conf.*, 2010, pp. 272–275.

[19] W. A. Kalender, R. Hebel, and J. Ebersberger, "Reduction of CT artifacts caused by metallic implants," *Radiology*, vol. 164, no. 2, pp. 576–577, Aug. 1987.

[20] R. Schulze, U. Heil, D. Groß, D. Bruellmann, E. Dranischnikow, U. Schwanecke, and E. Schoemer, "Artifacts in CBCT: A review," *Dentomaxillofacial Radiol.*, vol. 40, no. 5, pp. 265–273, 2011.

[21] F. E. Boas and D. Fleischmann, "CT artifacts: Causes and reduction techniques," *Imag. Med.*, vol. 4, no. 2, pp. 229–240, Apr. 2012.

[22] L. Gjesteby, B. De Man, Y. Jin, H. Paganetti, J. Verburg, D. Giantsoudi, and G. Wang, "Metal artifact reduction in CT: Where are we after four decades?" *IEEE Access*, vol. 4, pp. 5826–5849, 2016.

[23] E. Meyer, R. Raupach, M. Lell, B. Schmidt, and M. Kachelrieß, "Normalized metal artifact reduction (NMAR) in computed tomography," *Med. Phys.*, vol. 37, no. 10, pp. 5482–5493, Sep. 2010.

[24] Y. Zhang and H. Yu, "Convolutional neural network based metal artifact reduction in X-ray computed tomography," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1370–1381, Jun. 2018.

[25] H. Eun and C. Kim, "Oriented tooth localization for periapical dental X-ray images via convolutional neural network," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2016, pp. 1–7.

[26] X. Xu, C. Liu, and Y. Zheng, "3D tooth segmentation and labeling using deep convolutional neural networks," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 7, pp. 2336–2348, Jul. 2019.

[27] Y. Miki, C. Muramatsu, T. Hayashi, X. Zhou, T. Hara, A. Katsumata, and H. Fujita, "Tooth labeling in cone-beam CT using deep convolutional neural network for forensic identification," *Proc. SPIE*, vol. 10134, Mar. 2017, Art. no. 101343.

[28] J. Ma and X. Yang, "Automatic dental root CBCT image segmentation based on CNN and level set method," *Proc. SPIE*, vol. 10949, Mar. 2019, Art. no. 109492N.

[29] Z. Cui, C. Li, and W. Wang, "ToothNet: Automatic tooth instance segmentation and identification from cone beam CT images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6368–6377.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[31] G. N. Hounsfield, "Computed medical imaging," *Med. Phys.*, vol. 7, no. 4, pp. 283–290, 1980.

[32] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 648–656.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[36] A. de Gelder and H. Huisman, "Autoencoders for multi-label prostate MR segmentation," 2018, *arXiv:1806.08216*. [Online]. Available: http://arxiv.org/abs/1806.08216

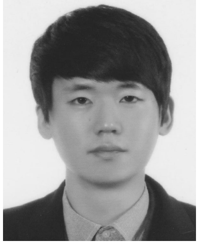
[37] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks," in *Proc. Annu. Conf. Med. Image Understand. Anal.*, 2017, pp. 506–517.

[38] V. Iglovikov and A. Shvets, "TernausNet: U-Net with VGG11 encoder pretrained on ImageNet for image segmentation," 2018, *arXiv:1801.05746*. [Online]. Available: http://arxiv.org/abs/1801.05746

[39] F. Pugliese, N. R. A. Mollet, G. Runza, C. van Mieghem, W. B. Meijboom, P. Malagutti, T. Baks, G. P. Krestin, P. J. deFeyter, and F. Cademartiri, "Diagnostic accuracy of non-invasive 64-slice CT coronary angiography in patients with stable angina pectoris," *Eur. Radiol.*, vol. 16, no. 3, pp. 575–582, Nov. 2005.

[40] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th Symp. Oper. Syst. Des. Implement.*, 2016, pp. 265–283.

[41] F. Chollet. (2015). *Keras*. Github Repository. [Online]. Available: https://github.com/fchollet/keras



S. LEE received the B.S. and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2013 and 2019, respectively. His research interests include image processing and pattern recognition.

S. WOO received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2014, where he is currently pursuing the Ph.D. degree in electrical and electronic engineering. His research interests include image/signal processing, pattern recognition, and neural networks.

J. YU received the B.S. and M.S. degrees in information and electronic engineering from Soongsil University, South Korea, in 2007 and 2009, respectively. He is currently working with the DIO IT R&D Center. His research interests include artificial intelligence and image processing.

J. SEO received the M.S. degree in computer science from Kyonggi University, Suwon, South Korea. He is currently working with the DIO IT R&D Center. His research interests include image processing, computer graphics, and computer vision.

J. LEE received the M.S. degree in electronic engineering from Hanyang University, Seoul, South Korea. He is currently working with the DIO IT R&D Center. His research interests include image processing, volume rendering, and machine learning.

C. LEE (Member, IEEE) received the B.S. and M.S. degrees in electronic engineering from Seoul National University, in 1984 and 1986, respectively, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1992. From 1986 to 1987, he was a Researcher with the Acoustic Laboratory, Technical University of Denmark (DTH). From 1993 to 1996, he worked with the National Institutes of Health, Bethesda, MD, USA. In 1996, he joined the faculty of the Department of Electrical and Computer Engineering, Yonsei University, Seoul, South Korea. His research interests include image/signal processing, pattern recognition, and neural networks.