A Symmetric Fully Convolutional Residual Network With DCRF for Accurate Tooth Segmentation

YUNBO RAO,1 (Member, IEEE), YILIN WANG,1 FANMAN MENG,2 (Member, IEEE), JIANSU PU1, JIHONG SUN3, AND QIFEI WANG4, (Member, IEEE)
1School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
2School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
3School of Medicine, Zhejiang University, Zhejiang 310029, China
4Department of Electrical Engineering and Computer Sciences (EECS), University of California at Berkeley, Berkeley, CA 94720, USA

Corresponding author: Yunbo Rao (raoyb@uestc.edu.cn)

This work was supported in part by the Science and Technology Service Industry Demonstration Project of Sichuan under Grant 2019GFW126, in part by the Science and Technology Project of Sichuan under Grant 2019YFG0504, in part by the National Natural Science Foundation of China under Grant 61872606 and Grant U19A2078, and in part by the Sichuan Science and Technology Innovation Project under Grant 2019KJ06.

ABSTRACT Accurate tooth segmentation from CBCT images is a crucial step for specialist to perform quantitative analysis, clinical diagnosis and operation. In this paper, we present a symmetric full convolutional network with residual block and Dense Conditional Random Field (DCRF), which can achieve accurate segmentation automatically for tooth images. The proposed method can not only strengthen feature propagation, but also boost feature reuse, which can be credited to the contracting path and the expanding path that extract and recover pixel cues sufficiently. To this end, we apply special deep bottleneck architectures (DBAs) and summation-based skip connection into our network to ensure accurate segmentation for much deeper neural network. Compared with previous methods which are based on conditional random field for original image intensity, our approach applies DCRF to the posterior probability generated by the proposed network. To avoid the interferences of noises around the tooth, we combine the pixel-level prediction capability of DCRF, which further enhance the segmentation performance. In the experiments, we verify the capabilities of our methods based on four evaluation indicators, which demonstrates the superiority of our method.

INDEX TERMS Tooth segmentation, CBCT images, fully convolutional network, bottleneck architecture, conditional random field.

I. INTRODUCTION

With a bunch of virtues like low radiant quantity, Cone Beam Computer Tomography (CBCT) [1] can produce 3D images of tooth with high resolution by scanning only once, which makes it be one of the most important method for diagnosis of dental diseases. By mapping tooth data produced by CBCT into 3D model, details of tissue and structure inside the tooth can be revealed. However, segmentation of tooth is also one of the challenging task for dental diagnosis.

Different from traditional CT technology, CBCT uses the projection data of 2D surface to replace the projection data of 1D linear produced by traditional CT, i.e., the cone beam scanning of 3D takes the place of traditional fan beam scanning of 2D. The 3D images can be formed directly after CBCT image reconstruction. Besides, not only can CBCT technology generate clear images with high resolution, but also it can produce images within reasonable range, requiring a short period of time for data assumption and employing standard forms of data. This makes significant contributions to segmentation and reconstruction based on medical images.

The visualization technology of human tooth has become an essential tool for computer aided diagnosis. Xu et al. [39] proposed a 3D tooth segmentation model completed in two-level hierarchical: teeth-gingiva labeling and inter-teeth labeling, and it has been directly applicable to orthodontic CAD systems. To solve some oral problems in orthodontics and prosthetics, doctors are required to grasp the overall
anatomical information of the root and crown manually with segmentation map generated by CBCT images of the target tooth. With the help of accurate segmentation, they will be able to perform clinical diagnosis, making reasonable plans for surgical and repair more efficiently. Hosntabal et al. [11] proposed a level set model with 3D region as auxiliary to extract the surface of tooth. Considering the huge advantages brought by segmentation process, tooth segmentation based on CBCT images has now become a fundamental task.

The existing segmentation methods for medical images can be classified into two categories. (1) The traditional methods, including the template-based fitting method [3] and the level-set method [5], [6]. This type of method requires more participation of professional experts and suffers from low self-learning ability and high sensitivity to noise and vague cases. (2) The methods based on deep learning [2], [4], [7], [9]. Recently, due to the strong robustness and versatility of convolution neural networks (CNNs), research on medical image processing using deep learning algorithms has become a contemporary hot spot, which outperforms the traditional methods significantly [8], [10]. Miki et al. [40] used AlexNet [44] network architecture and image data enhancement strategy to realize the classification of teeth, which can be used for automatic filing of dental records for forensic identification. Tian et al. [35] presented a two-level hierarchical feature learning 3D tooth segmentation and classification model by combining sparse voxel octree and 3D CNNs, which solved the problem of misclassification of highly similar tooth categories. Cui et al. [41] obtained the edge map of CBCT images firstly, and then proposed a new learning similarity matrix based on Mask R-CNN [45] to realize the automatic segmentation and recognition of tooth from CBCT images. Zhou et al. [42] proposed an asymmetric architecture called LEDNet. The encoder of the network adopts the idea of the down sampling of FCN, and the decoder adopts the idea of attention of pyramid network to realize precise semantic segmentation. Gou et al. [37] annotated the tooth image as a dataset by improving the level-set method, and then trained the dataset by employing the U-Net network [31]. This method can use the unlabeled data in the experiment to achieve semi-supervised learning.

However, the deep learning based segmentation is limited by the high diversity and cost of medical images. For example, compared with other kinds of images, the tooth images have three challenges. Firstly, there are significant variation between conventional and unconventional cases, limiting the networks’ ability of generalization, as illustrated in Figure 1 (bottom left). Secondly, an image always contains multiple objects [19], requiring more manual annotation, as shown in Figure 1. Thirdly, similar gray values between tooth and similar structures around them make the tooth vague margin [11], [13], as shown in the bottom row of Figure 1. Therefore, an accurate and efficient automatic segmentation method without labeling is urgently demanded in clinical practice to reduce the workload of experts.

Based on these observations, this paper rebuilds and optimizes the deep bottleneck architectures such as ResNet via the network structure based on U-Net. Moreover, we introduce our novel deep neural network called symmetric fully convolutional residual network, with dense conditional random fields to refine the posterior probability map. The overall architecture is showed as Figure 2. To summarize, our main contributions are listed as follows.

(1) To be more specific, we design a novel DBA to replace the general convolutional layer, which can not only increase the number of layers in the network, improving feature extraction, but also reduce the amount of network parameters.

(2) We skillfully employ our modified DBAs into every convolution layer of U-Net and introduce skip connection structure to enhance the propagation and reuse of the features.

(3) In addition, instead of directly using DCRF to operate segmentation on original intensity information, we apply DCRF to the segmentation probability map produced by the proposed network.

(4) Based on the probability map produced by our network, we make full use of DCRF for overall structured prediction to get rid of the noises in the images, which can locate the tooth contour precisely. In addition, the refinement of the tooth boundary further boosts the performance of segmentation.

II. RELATED WORK

Because the level set method has advantages in dealing with topological changes and contour propagation, researchers have applied the level set method to biomedical images. Broadly speaking, in the existing methods for image segmentation, the level set method [13], [15] is in a pivotal status. Gan et al. [5] proposed a hybrid level set method based on global energy, local energy, tooth shape constraint energy and edge detection energy of tooth to segment tooth from CT images. An active contour model based on edge is proposed by Li and Xu [16]. Their main contribution is to use distance regularized level set. Gao and Oksam [6] proposed a level set model based on the shape and strength prior of
tooth to segment the crown and root of individual tooth. Xia et al. [14] proposed a method to segment the contacts of maxillary and mandible teeth. Although the result of traditional methods is acceptable, these methods need tedious expert annotations and repeated iterations to obtain an available single image segmentation. In addition, it is hard to produce satisfactory result when the maxillary and mandible teeth are occluded [13].

Thanks to the development of computer hardware, deep learning has achieved outstanding performance in the field of image processing. By training an end-to-end network, deep neural network performs superiorly in several image processing fields such as feature extraction, image classification [18], semantic segmentation [8] and so on. Among multiple excellent models, ResNet [20], [24] proposes a unique jump structure called residual unit. Employing bottleneck structure with residual unit enhances its performance by a meaningful margin, it achieves a significant breakthrough in the field of deep neural network. With its strong property over image classification, image location and object detection, ResNet won the championship of these three events of Large Scale Visual Recognition Challenge (ILSVRC) in 2015. Besides, ResNet also demonstrates dominated performance in the field of semantic segmentation [25]–[27].

For segmentation task, fully convolutional networks (FCNs) [28] are proposed and then applied into medical image segmentation [29], [30]. In order to achieve better performance, some researchers introduce U-Net [31], which uses a contracting path and expansive path to extract feature and produce pixel-wise feature map. The unique structure of U-Net contributes to its outstanding performance over FCNs, making U-Net popular in the field of medical image segmentation [32]–[34]. However, with just a few layers, U-Net suffers from lack of the ability to extract feature from deep layers at contracting path while at expansive path it cannot supplement enough information to the feature maps, limiting the accuracy of the segmentation. Adding the number of layers in U-Net to improve its ability of extracting feature is also a promising direction [36]. However, the structure and means of optimization of those models still need to be improved. Besides, always with large amount of noise, the fact that gray scale information provided by medical images are in a low-quality feature space makes different kinds of tissue representing the same grey level, bring about low precision in segmentation.

Context cues represent the spatial relationships between category labels and play an important role in the tasks of structured prediction. Context cue or higher-level information is the key in semantic image segmentation. In recent years, conditional random field (CRF) acts as an effective way to optimize results from medical image segmentation [21]–[23]. Combining DNN’s ability of feature extraction and CRF’s ability of structured modeling can contribute to better performance in image segmentation task. Zhou et al. [43] designed a multi-scale deep context convolutional network named MDCCNet for semantic segmentation. It utilized convolution to combine feature maps from different levels of networks in a holistic manner. On this basis, the author used dense connected conditional random fields to optimize the output of the MDCCNet and achieved efficient experimental results on the PASCAL VOC 2012 and SIFTFlow dataset. Based on the traditional multi-scale input and sliding pyramid network, Lin et al. [38] combined with CRF to obtain background context information, constructed a deep learning model. However, lots of methods for training medical images still rely on the adjustment of limited parameters and us original intensity information as the main feature space. The original intensity always provides low quality feature space for CRF. Also, with large amount of noise, different tissues may display the same grey level, which makes the precise segmentation of medical images based on grey level be a challenge task.

III. METHODS

In this section, we introduce in detail our proposed method for accurate tooth segmentation from CBCT images. We start by extracting and organizing the information of the CBCT images. Furthermore, a novel symmetrical network architecture designed with DBAs is introduced to perform the training task. Finally, a post processing procedure named DCRF is applied to further optimize the contours of tooth segmentation results.

A. SYMMETRIC FULLY CONVOLUTIONAL RESIDUAL NETWORK

U-Net [31], composed of expansive path and contracting path, is used as our basic network structure. The contracting

FIGURE 2. The overview of our framework. Given CBCT tooth images, we first put the preprocessed images into the symmetric fully convolutional residual network to get the segmentation probability map. Then DCRF is directly applied to the probability map, and the segmentation result of tooth is obtained through iteration and optimization. Finally, the segmentation task is completed.
path can help the extraction of the feature of the context while the expansive path achieving the pixel-wise location and prediction precisely. It is true that increasing the number of hidden layers can be beneficial to the performance of neural networks. Increasing the number of layers, however, will not only accelerate the number of arguments, but also require more computing resource, which lead to overfitting. To get rid of the defect, we apply modified deep bottleneck architectures to deepen the network while decreasing the number of arguments.

In detail, we designed three DBAs to replace the convolutional layers that are employed in the original U-Net structure, as shown in Figure 3, where the left side is the original structure, and the right side is our proposed DBA. The original structure has only one path. In our proposed structure instead, each DBA can be divided into two branches. The left one only contains a 1 × 1 convolutional layer with the stride of 1 while the right one is composed of three connected convolutional layers, which are 1 × 1, 3 × 3 and 1 × 1 with the stride of 1 respectively. In particular, we added an identical mapping path to DBA1 to ensure that the features are effectively propagated to solve the problem of gradient disappearance or explosion.

After each operation of convolution, ReLU activation function is directly applied so as to pass the non-linear feature. Before the employment of ReLU function, the Batch Normalization (BN) is used. The detailed step is that we firstly assume \( k(n) = k(1), k(2), \ldots, k(n) \) to represent the outputs of \( n \) nodes within a layer, then we calculate the average of these outputs in terms of \( \text{avg}_k \), defined as

\[
\text{avg}_k = \frac{\sum_{m=1}^{n} k_m}{n} \quad (1)
\]

The variance is also required in terms of \( \text{var}_k^2 \), i.e.,

\[
\text{var}_k^2 = \frac{\sum_{m=1}^{n} (k_m - \text{avg}_k)^2}{n} \quad (2)
\]

We employ the learnable parameters \( \vartheta \) and \( \gamma \). and then based on the standard variation calculated by the outputs of \( n \) nodes, we perform unbiased estimation to get the output \( Y_{BN} \) of BN

\[
Y_{BN} = \frac{\vartheta}{\sqrt{\text{var}_k^2 + \epsilon}} \cdot \text{avg}_k + \gamma \quad (3)
\]

Here, \( \epsilon \) is a constant. The advantage by employing BN operating is that it can boost the speed of training, accelerate the process of convergence and alleviate the vanishing of gradient.

We define the input of DBAs as \( x(w, h, n) \), where \( w, h, n \) represent the width, height and number of channels respectively. The weight of the filter in the left branch is denoted as \( \omega_{L1} \). And \( \omega_{R1}, \omega_{R2}, \omega_{R3} \) denote the weight of the filters in the right branch. Then each layer of the network can be defined as \( t^{(1)}(x, \omega_{R1}), t^{(2)}(x, \omega_{R2}), t^{(3)}(x, \omega_{R3}) \). The functions of BN and ReLU are defined uniformly as \( G(x) \). As a result, the output of DBA can be represented as

\[
DBA_n = x + f(x, \omega)_L + f(x, \omega)_R \quad (4)
\]

where \( x \) denotes the input data \( x(w, h, n) \), which means the identity mapping in DBA1, \( n \) denotes the number of channels of the DBA output. Each kind of DBA produce different parameter \( n \). \( f(x, \omega)_L \) is the output of the left branch, and can be represented as

\[
f(x, \omega)_L = G(t(x, \omega_{L1})) \quad (5)
\]

where \( f(x, \omega)_R \) is the output of the right branch, i.e.,

\[
f(x, \omega)_R = G\{t^{(3)}[G\{t^{(2)}[G\{t^{(1)}(x, \omega_{R1})], \omega_{R2}\}], \omega_{R3}\]\}

Among the three types of DBAs, DBA1 can keep the number of channels while performing a series of convolution, which replaces the normal convolution in U-Net.

\[
DBA1 = DBA_n \quad (7)
\]

DBA2 can double the number of channels after a series of convolution, which replaces the convolution that doubles the number of channels in U-Net. Here is no identity mapping \( x(w, h, n) \) in DBA2, so parameter \( x \) should be subtracted.

\[
DBA2 = DBA_{2n} - x \quad (8)
\]
Y. Rao et al.: Symmetric Fully Convolutional Residual Network With DCRF for Accurate Tooth Segmentation

**FIGURE 4.** The overview of the symmetric fully convolutional residual network. Three different DBAs replace different convolutional layers in traditional U-net. DBA1 and DBA2 are used in the downsampling; DBA1 and DBA3 are used in the upsampling. Finally, a deeper network structure is formed by combining several common convolution layers.

**TABLE 1.** Comparison of the number of parameters between the original structure and three different DBAs.

| DBAs   | Number of parameters of DBA | Number of parameters of original | Ratio (DBA/Original) |
|--------|----------------------------|----------------------------------|----------------------|
| DBA1   | $4.5C^2$                   | $9C^2$                           | 0.50                 |
| DBA2   | $3.5C^2$                   | $18C^2$                          | 0.50                 |
| DBA3   | $4.5C^2$                   | $4.5C^2$                         | 0.78                 |

DBA3 can halve the number of channels after convolution, replacing the convolution that halves the number of channels after performing multi-scale contextual features fusion in U-Net. Here is no identity mapping $x(w, h, n)$ in DBA3, so parameter $x$ should be subtracted.

$$\text{DBA3} = \text{DBA}_{(n/2)} - x$$

Replacing the ordinary convolutional layers in U-Net with three types of DBAs, we build a network with 54 layers, which effectively deepens the network, optimizes the process of feature extraction and improves the accuracy of location. The detailed information of the network structure is showed in Figure 4. Table 1 compares the number of parameters between the original structure and three different DBAs. It is worth noting that DBA1 and DBA2 have only half the amount of parameters of the original structure. The results show that although the number of layers in our network increases, the amount of our total parameters decreases, which indicates the efficiency of our strategy.

**B. DCRF FOR POST PROCESSING PROCEDURE**

The conditional random fields model [12], [17] is graph model made up of unary potential and adjacent pixels, which ignores the information of the whole space. In this paper, we apply DCRF to the posterior probability map directly instead of the original intensity of the tooth. The segmentation probability map obtained from the symmetric fully convolutional residual network is used as the input of DCRF model. DCRF can not only make use of the relation between adjacent pixels, but also master and use the pixel information of the whole space to judge and predict the local pixels. At the same time, the model can be built according to the relationship between the length and distance of pixels in the space to fully grasp the context of the whole space, and the Gibbs energy function is defined as

$$E(A) = \sum_i P_i(a_i) + \sum_{i<j} P_{ij}(a_i, a_j), \quad i, j \in \{1, 2, \ldots, N\}$$

The DCRF energy function is composed of unary potential function $P_i(a_i)$ and binary potential function $P_{ij}(a_i, a_j)$, where $N$ denotes the number of pixels in the whole space. The unary potential function is a status characteristic function defined at the position $i$ of the observed sequence, which can be further defined as:

$$P_i(a_i) = -\log_2 \varphi(a_i)$$

where $\varphi(a_i)$ is to calculate the probability that the $i$th pixel in the input image belongs to a certain category $a_i$. In this paper, $a_i$ is the segmentation probability map got by the symmetric fully convolutional residual network.

The binary potential function is a transfer eigenfunction defined at different observation positions, which is used to
characterize the correlation between variables and the effect of observation sequences on them, and is represented as

$$P_{ij}(a_i, a_j) = \theta(a_i, a_j) \sum_{m=1}^{k} \omega_m k^m(f_i, f_j) \quad (12)$$

where $\theta(a_i, a_j) = \begin{cases} 1 & \text{if } (a_i \neq a_j) \\ 0 & \text{otherwise} \end{cases}$. Since all pixels in the DCRF model are fully connected, there is a correspondence between each pair of pixels $i$ and $j$ in the image, no matter where they are relative to each other. Here, $f_i$ and $f_j$ are the eigenvectors of $i$ and $j$, respectively. $k^m$ is Gaussian kernel, which depends on the eigenvectors of the pixels $i$ and $j$, and $\omega_m$ is its corresponding weight.

The bilateral relationship is a popular pairwise relationship in image processing, which roughly indicates that pixels with similar colors or positions may belong to the same class. The binary potential function is defined by the reference to bilateral relationship and gray intensity, i.e.,

$$P_{ij}(a_i, a_j) = \omega_1 \exp \left( -\frac{||\delta_i - \delta_j||^2}{2\sigma_a^2} - \frac{||I_i - I_j||^2}{2\sigma_p^2} \right) + \omega_2 \exp \left( -\frac{||\delta_i - \delta_j||^2}{2\sigma_g^2} \right) \quad (13)$$

where the first kernel depends on both the pixel position in term of $\delta$ and the pixel gray intensity in term of $I$, while the second kernel only depends on the pixel position, and $\alpha$, $\beta$ and $\gamma$ the super parameters to control the value of the Gaussian kernel.

The binary potential function is to describe the relationship between pixels, and encourage similar pixels to assign the same label, while pixels with larger differences are assigned by different labels. The definition of “relationship” between pixels is related to the gray intensity and the actual relative distance, so that the segmentation at the boundary can be achieved, and the effect of refining the boundary can be realized, and the accurate segmentation map can be obtained finally.

C. DATASET AND PRE-PROCESSING

The original dataset of CBCT tooth images is provided by West China Hospital, Sichuan University. The images of the dataset are all in DCM format, as shown in Figure 5(a). In order to transform these images into the data that can be passed into the neural network, we extract and reorganize the information of these DCM files. The overall procedure of processing is showed in Figure 5.

Firstly, the useful parts of CBCT tooth images are extracted and transformed into trainable images. In details, we first calculate the windowing of the CBCT file, and then use the windowing to map the gray value of the image to 0-255 to get the imgMap. Next we use the minimum value of the imgMap to window the image data and convert the data to uint8 format. These generated images are store in PNG format also with resolution of $401 \times 401$. We use these images as training dataset shown in Figure 5(b). Algorithm 1 summarizes the available training images extracted from CBCT tooth images.

Secondly, we use annotation tool called LabelMe to annotate these images as shown in Figure 5(c). Thirdly, considering that the grey-scale of the annotated images produced by LabelMe ranges from 0 to 1 while the network requires images range from 0 to 255, we expand the gray-scale. Finally, we use these images as the labels as shown in Figure 5(d).

After preprocessing of the CBCT images of tooth, the training dataset is composed of 86 images (conventional / unconventional = 51/35) with ground truth annotations. Testing data contains 24 images. In order to achieve better performance, data augmentation strategies are applied within our
Y. Rao et al.: Symmetric Fully Convolutional Residual Network With DCRF for Accurate Tooth Segmentation

Algorithm 1 Training Images Obtained by Extracting and Transforming Relevant Information in CBCT Images

Input: The matrix of grey value of the CBCT image: \( \text{img} \)
Output: PNG image of size 401 * 401: \( \text{pngImg} \)

1: \( \text{windowing} = \max (\text{img}) - \min (\text{img}) \)
2: \( \text{imgMap} = \frac{256}{\text{windowing}} \ast \text{img} \)
3: \( \text{windowImg} = \text{imgMap} + \min (\text{imgMap}) \)
4: \( \text{uint8Img} \leftarrow \text{windowImg} //\text{Convert to uint8 format} //\text{image} \)
5: \( \text{pngImg} \leftarrow \text{uint8Img} //\text{Save as PNG image of size} //401 * 401 \)

training process. We perform rotation, horizontal translation, vertical translation, scaling and normalization to the images to enrich the training dataset.

IV. EXPERIMENTS AND RESULTS

A. NETWORK TRAINING

We build a novel symmetric fully convolutional residual network which is composed of nine convolution groups. Each group contains two DBAs structure, excepting that we only apply one DBA1 structure in the first block. At the stage of downsampling, DBA2 and DBA1 are used for the second to fifth blocks, followed by a single max-pooling layer. In order to alleviate the concerns about overfitting, a dropout layer is employed before the max-pooling layer within the fourth block. With the purpose of saving the result of downsampling, instead of adding a max-pooling layer, we directly employ dropout layer after the fifth block. To achieve multi-scale feature fusion, we employ DBA3 and DBA1 for the sixth to ninth blocks, after which we perform deconvolution and concat channels. A convolutional layer with \( 1 \times 1 \) kernel (stride 1), is added at the end of the ninth block. We adopt sigmoid activation function to the \( 1 \times 1 \) layer and get the final segmentation map of probability. Adam is used as the optimizer for parameter tuning with a fixed learning rate 0.01. Binary cross entropy is employed as loss function of the network. The whole network model is stored as HDF5 format file after training. Thus, the well-trained symmetric fully convolutional residual network model is obtained.
FIGURE 8. The segmentation results of tooth of unconventional cases. From top to bottom: original images, the ground truth of images, the segmentation probability map without DCRF, and segmentation results with DCRF.

B. IMPLEMENTATION DETAILS
After the training procedure, we input the dataset to our trained network model. In symmetric fully convolutional residual network, fully connected layers are removed, which eliminates the restriction to the size of input images. The probability map of tooth segmentation is generated through the whole network, and then this probability map is used as the input of DCRF model. After DCRF model optimization, the final accurate result of tooth segmentation is generated. All the works are implemented in Keras and ran on a server with Nvidia GeForce 1070 GPU. During the experiment, the batch size for training is set to be 100 and the network is trained for 10 epochs. Specifically, Figure 6 shows the training accuracy and loss of our tooth data set for every 100 iterations. It can be seen that with the increase of the number of iterations, the segmentation accuracy is gradually improving and converging.

C. QUALITATIVE EVALUATION
To better evaluate the effectiveness of our proposed method qualitatively, the segmentation results of partial test data are shown in Figure 7 (conventional case) and Figure 8 (unconventional case). In order to demonstrate the fact that DCRF has to refine the outline information, we show the segmentation probability map generated by the symmetric fully convolutional residual network. The qualitative results are shown at the middle row in Figure 7 and Figure 8. As can be seen from the segmentation results, our method can accurately segment tooth objects by using a special and efficient neural network architecture to fuse multi-level contextual features and refine tooth boundaries in combination with DCRF. What should be mentioned is that our method can accurately segment the tooth objects in both conventional and unconventional cases. However, some teeth are not completely separated due to the excessive contact between tooth and a large number of noises of the same structure around the tooth. In comparison, our method is capable of separating these touching tooth objects clearly. Under the framework of unified multitasking, the symmetric full convolution residual network and DCRF can qualitatively discover and predict the filling of dental information, which enhances its superiority.
### TABLE 2. Accuracy comparison of without DCRF and with DCRF.

| Method       | VD (mm$^3$) | DSC (%) | ASSD (mm) | MSSD (mm) |
|--------------|-------------|---------|-----------|-----------|
| Without DCRF | 19.98       | 91.37   | 0.28      | 1.22      |
| With DCRF    | 18.86       | 91.66   | 0.25      | 1.18      |

**D. QUANTITATIVE EVALUATION AND COMPARISON**

In order to evaluate tooth segmentation accuracy, we employ four widely used performance metrics, i.e. Volume Difference (VD), Dice Similarity Coefficient (DSC), Average Symmetric Surface Distance (ASSD) and Maximum Symmetric Surface Distance (MSSD). For convenience, the real and predicted region of a tooth are denoted as ‘A’ and ‘B’ respectively.

VD is the volume difference between segmentation results and groundtruth. If the value of DSC equals to 0, it means perfect segmentation, and the formulation is:

$$VD = |A-B|$$  \hspace{1cm} (14)

DSC is the most widely used evaluation criteria for segmentation, which ranges from 0 to 1. If the value of DSC equals to 1, it means perfect segmentation, and the formulation is defined as:

$$DSC = \frac{2 \times |A \cap B|}{|A \cup B|}$$  \hspace{1cm} (15)

ASSD is based on surface voxels. If the value of ASSD equals to 0, it means perfect segmentation, and the formulation is defined as:

$$ASSD(A, B) = \frac{1}{|S(A) + S(B)|} \times \left( \sum_{S_A \in S(A)} dist(S_A, S(B)) + \sum_{S_B \in S(A)} dist(S_B, S(A)) \right)$$  \hspace{1cm} (16)

where $S(A)$ denotes the set of surface voxels. The $dist(v, S(A))$ is the shortest distance from a voxel $v$ to the whole set $S(A)$, which is defined as $dist(v, S(A)) = \min_{S_A \in S(A)} \|v - S_A\|$, where $\|•\|$ is the Euclidean norm.

MSSD is also known as the hausdorff distance. If the value of MSSD equals to 0, it means perfect segmentation, and the formulation is defined as

$$MSSD(A, B) = \max \left\{ \max_{S_A \in S(A)} dist(S_A, S(B)), \right\} \left\{ \max_{S_B \in S(B)} dist(S_B, S(A)) \right\}$$  \hspace{1cm} (17)

where $S(A)$ denotes the set of surface voxels. The $dist(v, S(A))$ is the shortest distance from a voxel $v$ to the whole set $S(A)$, which is defined as $dist(v, S(A)) = \min_{S_A \in S(A)} \|v - S_A\|$, where $\|•\|$ is the Euclidean norm.

To validate the effect of the DCRF model, we augment the base network with DCRF model detection study. We use the segmentation probability map output by the basic network as the result without DCRF (see the third row of Figure 8), and then compare it with the segmentation result with DCRF (as the fourth row of Figure 8), as shown in Table 2.

The quantitative analysis of our proposed tooth segmentation method is compared with the traditional methods of Gan et al. [5], Xia et al. [14], Gao et al. [6], and Hosntalab et al. [11]. And the quantitative tooth segmentation accuracy of our method and four traditional methods on the test dataset is shown in Figure 9. It is worth noting that the traditional method is not inferior to the deep learning-based method. Xia et al. [14] used the level set method to obtain desired effect, but slightly worse than the method of our proposed.

### TABLE 3. The CBCT tooth segmentation accuracy comparison by different algorithms.

| Method       | VD (mm$^3$) | DSC (%) | ASSD (mm) | MSSD (mm) |
|--------------|-------------|---------|-----------|-----------|
| Miki’s [40]  | 46.15       | 88.80   | 0.48      | 2.06      |
| Original U-Net | 41.11       | 89.14   | 0.42      | 1.79      |
| Gou’s [37]   | 32.74       | 89.37   | 0.31      | 1.68      |
| Tian’s [35]  | 25.56       | 89.81   | 0.29      | 1.39      |
| Our method   | 18.86       | 91.66   | 0.25      | 1.18      |
In order to further verify the superiority of our method, we modified the original U-Net method [31] to achieve CBCT tooth segmentation. Not only that, we also compared with several latest learning-based segmentation methods [35], [37], [40] for tooth segmentation. The accuracy comparison between our proposed method and other excellent algorithms is shown in Table 3. We can see that compared with other latest learning-based advanced methods, the accuracy of our method is higher in the four performance metrics. Especially, the value of ASSD is close to the size of a pixel, which means that our method achieves high performance in pixel level segmentation. And in the most authoritative indicator DSC, our method also has the highest accuracy. By testing some unconventional cases of tooth, we discovered that when there are large-scale similar confusing structures around the tooth, the confusing structures are excessively discriminated as a segmentation object, as shown in Figure 8 (fourth column).

In summary, our method for accurate tooth segmentation from CBCT achieved the best segmentation result on all testing data, which proved the effectiveness of our proposed method consistently.

V. CONCLUSION

In this paper, we have presented a symmetric fully convolutional residual network with DCRF to accurately segment tooth from CBCT images. We apply symmetric fully convolutional residual network to generate the segmentation probability map from input images of tooth and use DCRF as post processing measure, which solves the problem of boundary smoothing in the field of deep learning based segmentation. In symmetric fully convolutional residual network, special DBAs and BN layers are applied to enhance the propagation and reuse of feature, improving the network ability to extract feature and perform pixel-wise prediction. Rather than the original information, we use DCRF on the segmentation probability map produced by our network. We make full use of DCRF to locate the outline of tooth and refine its boundary, which improves the accuracy of segmentation of tooth images. According to the experimental result, our method has a good performance compared with existing methods.

REFERENCES

[1] L. Lechuga and G. A. Weidlich, “Cone beam CT vs. Fan beam CT: A comparison of image quality and dose delivered between two differing CT imaging modalities,” *Cereus*, vol. 8, no. 9, p. 778, Sep. 2016.

[2] Z. Zhang, L. Yang, and Y. Zheng, “Translating and segmenting multimodal medical volumes with Cycle- and shape-consistency generative adversarial network,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9242–9251.

[3] S. Barone, A. Paoli, and A. V. Razionale, “CT segmentation of dental shapes by anatomy-driven reformation imaging and B-spline modeling,” *Int. J. Numer. Methods Biomed. Eng.*, vol. 32, no. 6, Jun. 2016, Art. no. e02747.

[4] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, “Fully convolutional instance-aware semantic segmentation,” 2016, arXiv:1611.07709. [Online]. Available: http://arxiv.org/abs/1611.07709

[5] Y. Gan, Z. Xia, J. Xiong, G. Li, and Q. Zhao, “Tooth and alveolar bone segmentation from dental computed tomography images,” *IEEE J. Biomed. Health Inform.*, vol. 22, no. 1, pp. 198–204, Jan. 2018.

[6] 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.

[7] J. Dai, K. He, and J. Sun, “Instance-aware semantic segmentation via multi-task network cascades,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3150–3158.

[8] L. Mou, Y. Hua, and X. X. Zhu, “A relation-Augmented fully convolutional network for semantic segmentation in aerial scenes,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12416–12425.

[9] J. Dai, K. He, Y. Li, S. Ren, and J. Sun, “Instance-sensitive fully convolutional networks,” in *Proc. ICCV*. Cham, Switzerland: Springer, 2016, pp. 534–543.

[10] C. H. Zhou, S. C. Chen, C. X. Ding, and D. C. Tao, “Learning contextual and attentive information for brain tumor segmentation,” in *Proc. Int. MICCAI Brainlesion Workshop*, 2018, pp. 497–507.

[11] M. Hosntalab, R. A. Zoroofi, A. A. Tehrani-Fard, and G. Shirani, “Segmentation of teeth in CT volumetric dataset by panoramic projection and variational level set,” *Int. J. Comput. Assist. Radiol. Surg.*, vol. 3, nos. 3–4, pp. 257–265, Sep. 2008.

[12] Y. Gan, Z. Xia, J. Xiong, Q. Zhao, and F. Aua, “Are spatial and global constraints really necessary for segmentation?” in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 9–16.

[13] 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.

[14] Z. X. Xi, 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.

[15] Y. Gan, Z. Xia, J. Xiong, Q. Zhao, Y. Hu, and J. Zhang, “Toward accurate tooth segmentation from computed tomography images using a hybrid level set model,” *Med. Phys.*, vol. 42, no. 1, pp. 14–27, Dec. 2014.

[16] C. Li, C. Xu, C. Gui, and M. D. Fox, “Distance regularized level set evolution and its application to image segmentation,” *IEEE Trans. Image Process.*, vol. 19, no. 8, pp. 2343–2354, Dec. 2010.

[17] M. Cogswell, X. Lin, S. Purushwalkam, and D. Batra, “Combining the best of graphical models and ConvNets for semantic segmentation,” 2014, arXiv:1412.4313. [Online]. Available: http://arxiv.org/abs/1412.4313

[18] J. Gu, “Recent advances in convolutional neural networks,” *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[19] X. He and S. Gould, “An exemplar-based CRF for multi-instance object segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 296–303.

[20] 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.

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[22] W. Luo and M. Yang, “Fast skin lesion segmentation via fully convolutional network with residual architecture and CRF,” in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1438–1443.

[23] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, “Group consistent similarity learning via deep CRF for person re-identification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8649–8658.

[24] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, “Identity mappings in deep residual networks,” in *Proc. ECCV*. Cham, Switzerland: Springer, 2016, pp. 630–645.

[25] A. Dadashzadeh and A. Tavakoli Targhi, “Multi-level contextual network for biomedical image segmentation,” 2018, arXiv:1810.00327. [Online]. Available: http://arxiv.org/abs/1810.00327

[26] L. Yang, F. Zhang, C. Chen, and D. Z. Chen, “Suggestive annotation: A deep active learning framework for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 399–407.

[27] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

[28] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 600–607.
[29] H. Chen, X. Qi, L. Yu, and P.-A. Heng, “DCAN: Deep contour-aware networks for accurate gland segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2487–2496.

[30] X. Xu, Q. Lu, L. Yang, S. Hu, D. Chen, Y. Hu, and Y. Shi, “Quantization of fully convolutional networks for accurate biomedical image segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8300–8308.

[31] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. MICCAI, Munich, Germany, Oct. 2015, pp. 234–241.

[32] H. Li, A. Zhigalloy, and H. Bjorn Menze, “Automatic brain structures segmentation using deep residual dilated u-net,” in Proc. MICCAI, 2018, pp. 385–393.

[33] H. Tang, X. Pei, S. Huang, X. Li, and C. Liu, “Automatic lumbar spinal CT image segmentation with a dual densely connected U-Net,” 2019, arXiv:1910.09198. [Online]. Available: http://arxiv.org/abs/1910.09198

[34] W. Wang, K. Yu, J. Hugonot, P. Fua, and M. Salzmann, “Recurrent U-net for resource-constrained segmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 2142–2151.

[35] S. Tian, N. Dai, B. Zhang, F. Yuan, Q. Yu, and X. Cheng, “Automatic classification and segmentation of teeth on 3D dental model using hierarchical deep learning networks,” IEEE Access, vol. 7, pp. 84817–84828, 2019.

[36] T. M. Quan, G. C. D. Hildebrand, and W.-K. Jeong, “FusionNet: A well-known international journals and conferences. His current research interests include image segmentation and object detection. He is a member of the IEEE Circuits and Systems Society. He was a recipient of the Best Student Paper Honorable Mention Award at the 12th Asian Conference on Computer Vision, Singapore, in 2014, and the Top 10 Paper Award at the IEEE International Conference on Image Processing, Paris, France, in 2014.

[37] Q. Zhou, Y. Wang, J. Liu, X. Jin, and L. J. Latecki, “An open-source project for real-time image semantic segmentation,” Sci. China Inf. Sci., vol. 62, no. 12, Dec. 2019, Art. no. 227101.

[38] Q. Zhou, W. Yang, G. Gao, W. Ou, H. Lu, J. Chen, and L. J. Latecki, “Multi-scale deep context convolutional networks for semantic segmentation,” World Wide Web, vol. 22, no. 2, pp. 555–570, Mar. 2019.

[39] A. Krizhevsky, I. Sutskever, and E. Geoffrey Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. NIPS, 2012, pp. 1106–1114.

[40] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2980–2988.

YUNBO RAO (Member, IEEE) received the B.S. degree from Sichuan Normal University, in 2003, and the M.E. and Ph.D. degrees from the School of Computer Science and Engineering, University of Electronic Science and Technology of China, in 2006 and 2012, respectively. From 2009 to 2011, he was a Visiting Scholar of electrical engineering with the University of Washington, Seattle, USA. Since 2012, he has been with the School of Information and Software Engineering, University of Electronic Science and Technology of China, where he is currently an Associate Professor. His research interests include image segmentation, 3-D reconstruction, video enhancement, and medical image processing.

JANSU PU received the M.D. degree in clinical medicine from the School of Medicine, Zhejiang University, Hangzhou, China, in 2009. He is currently an Attending Radiologist and a Chief Physician with the Department of Radiology, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University. He has authored or coauthored numerous medical and nanomaterial articles in well-known international journals, such as Radiology, Nature Communications, and Advanced Materials. His current research interests include molecular imaging, artificial intelligence, and radionics.

Dr. Sun is a member of the Radiological Society of North America. He was a recipient of the Travel Award for Young Investigators in Molecular Imaging at the 96th (2010) and 98th (2012) Scientific Assembly and the Annual Meeting of Radiological Society of North America (RSNA), USA.

QIFEI WANG (Member, IEEE) received the B.S. degree in information and computing science from the Beijing University of Posts and Telecommunications, China, in 2007, and the Ph.D. degree in control science and engineering from Tsinghua University, China, in 2013. He joined the Department of Electrical Engineering and Computer Sciences (EECS), University of California at Berkeley, Berkeley, CA, USA, in 2014. His current research interests include computer vision, machine learning, video processing, and communications.

YILIN WANG received the B.S. degree in the academic of information science and technology from Shijiazhuang Tiedao University, Shijiazhuang, China, in 2018. He is currently pursuing the master’s degree majoring in software engineering with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include computer vision, machine learning, and image segmentation.

FANMAN MENG (Member, IEEE) received the Ph.D. degree in signal and information processing from the University of Electronic Science and Technology of China, Chengdu, China, in 2014. From 2013 to 2014, he was a Research Assistant with the Division of Visual and Interactive Computing, Nanyang Technological University, Singapore. He is currently an Associate Professor with the School of Electronic Engineering, University of Electronic Science and Technology of China. He has authored or coauthored numerous technical articles in well-known international journals and conferences. His current research interests include image segmentation and object detection. He is a member of the IEEE Circuits and Systems Society. He was a recipient of the Best Student Paper Honorable Mention Award at the 12th Asian Conference on Computer Vision, Singapore, in 2014, and the Top 10 Paper Award at the IEEE International Conference on Image Processing, Paris, France, in 2014.

JIHONG SUN received the M.D. degree in clinical medicine from the School of Medicine, Zhejiang University, Hangzhou, China, in 2009. He is currently an Attending Radiologist and a Chief Physician with the Department of Radiology, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University. He has authored or coauthored numerous medical and nanomaterial articles in well-known international journals, such as Radiology, Nature Communications, and Advanced Materials. His current research interests include molecular imaging, artificial intelligence, and radionics.

YUNBO RAO (Member, IEEE) received the B.S. degree from Sichuan Normal University, in 2003, and the M.E. and Ph.D. degrees from the School of Computer Science and Engineering, University of Electronic Science and Technology of China, in 2006 and 2012, respectively. From 2009 to 2011, he was a Visiting Scholar of electrical engineering with the University of Washington, Seattle, USA. Since 2012, he has been with the School of Information and Software Engineering, University of Electronic Science and Technology of China, where he is currently an Associate Professor. His research interests include image segmentation, 3-D reconstruction, video enhancement, and medical image processing.