CHAPTER 5

TEETHCAPS: A DENTAL IMAGE SEGMENTATION MODEL BASED ON CAPSULE NETWORK

In this chapter, a novel deep segmentation model based on a capsule network is designed for the segmentation of dental panoramic X-ray images. The proposed model combines the ability of convolution filters and attention mechanism with the capsule architecture for improving the segmentation performance on dental images. The model is validated on two open benchmark dental datasets UFBA_UESC dataset and Tufts dataset. The results obtain shows the potential of capsule-based network for segmentation tasks.

5.1 Background

The goal of computer vision area is to provide machines with vision capabilities equivalent to human eyesight. CNNs with the aid of automated feature learning techniques based on gradient descent, transformed the field of computer vision by outperforming human performance on several publicly available datasets. Though CNNs have demonstrated remarkable performance in computer vision-related applications, they possess certain limits. This marks the beginning of the researcher's society's gradual shift from CNNs to capsule networks. CNNs, unlike the human brain, have quite a few levels of hierarchy (arrangement of neurons layer-by-layer), which may adversely affect the extraction of high-level information from data. CNNs suffer from a significant amount of information loss as they automatically encode an object’s pattern without considering its

pose and positioning information in relation to other objects due to which they are unable to generate the actual information contained in the data. CNNs often utilize the pooling layers which are combined with convolution procedures, resulting in the loss of specific location and spatial information. Also, when compared to human vision CNNs require a large amount of data for training to develop effective and generalisable models. Thus, leaving scope for developing models that can closely resemble the human brain's functionality.

In recent years, the capsule network, due to its potential has gained a lot of attention in the research community. The motive behind employing capsules can be considered as a step forward to replicate the functioning of biological neuron. The capsules are a group of neurons that can perform sophisticated calculations to encode high-level features in vector form. It can estimate the object's presence and instantiation parameters at a particular position. Hinton et al. [] introduced the term 'capsules' in the computing area, where the authors proposed employing powerful neurons to encapsulate rich information rather than basic neurons, but no methodology was available at that time for training of such architecture. Sabour et al. [61] in the year 2017 proposed a neural network using capsules and introduced the dynamic routing algorithm for communication between the capsules. The performance of the CapsNet was remarkable on MNIST dataset. The advantage of CapsNet over CNN is that it represents features in vector form thus providing more relevant information with small amount of data. It can also preserve precise loss and spatial information lost by pooling layers in CNN. Thus, increased the interest of researchers to work with capsules for computer vision tasks.

5.2 Related Work

Deep learning techniques based on CNNs are being used extensively for medical image segmentation tasks. One of the most prominent architectures used is U-Net architecture. However, this architecture due to down-sampling lost relevant information like location information. Also, the base U-Net architecture faces challenges to learn long range dependencies and global information because of localization of convolution operations. These shortcomings are overcome by the different variants of U-Net-based architecture like attention UNet[106], U-Net++[107] etc. Apart from the U-Net network, some other state-of-the-art networks are SegNet[103], BiseNet[101], CENet[105] and NanoNet[4] which performed better for medical image segmentation. However, these networks utilize the CNN and thus suffer from capturing long-range dependencies.

To overcome the shortcomings of CNN Sabour et al.[61] proposed the CapsNet architecture in which the concept of dynamic routing was introduced for communication between the capsules. The CapsNet architecture has a complex nature, which limits it to the task of classification and detection in image analysis. Later, Lalonde et al.[63] extended the capability of capsule network for image segmentation task by introducing the concept of convolutional capsules and deconvolutional capsules in a novel architecture called SegCaps. The authors also introduced the concept of locally constraint routing algorithm which is a modification of the dynamic routing algorithm for the purpose of segmentation. Initially, the network was used for object segmentation and later for binary segmentation of medical images. Bonheur et al.[64] proposed MaTwo-CapsNet, for multiclass segmentation of medical images. The authors combined the two matrices that is pose information and appearance feature into a special capsule. The routing between capsules is performed by a novel dual-routing algorithm.

Komm et al.[66] proposed a deep-learning method which combines capsule network with the inception architecture for the task retinal vessel segmentation. Koresh et al.[67] proposed a modified capsule network to segment the three major boundaries of the corneal layer from the OCT corneal images. Bonheur et al[68]. proposed the capsule network called OnlyCaps-Net which utilizes the separable depth-wise convolution along with two squash functions softsquash and unit squash for multilabel semantic segmentation of medical images. Bragsten et al[69]. proposed an optimised MaTwoCapsNet by hyper-tuning the parameters for segmenting intravascular ultrasound images.

Nguyen et al.[71] proposed an efficient deep neural network based on 3D capsule network called 3D-UCaps to segment volumetric medical images. Moghaddasi et al. optimised the 3D-Ucaps architecture along with hybrid loss function to take advantage of capsule networks as well as CNN for automatically segmenting 3D mandibles. Tran et al. presented an efficient 3D encoder-decoder based on capsule architecture called 3D ConvCaps for medical image segmentation. The authors further extend their work by adding self-supervised learning to the 3D capsule network and proposed SS-3DCapsnet for medical image segmentation.

Huynh et al.[78] proposed a novel model, that combine a capsule architecture and ResNext architecture for 2D chest X-ray image segmentation and 3D kidney tumor segmentation. Zade et al. modified the SegCaps and introduced the curriculum learning approach for glioma segmentation. Pawan et al.[82] enhanced SegCaps by introducing dilation layers, residual connections, inception blocks and capsule pooling to segment sub-retinal fluid from OCT images. Wan et al. proposed an attention guided capsule network for medical image segmentation from four different medical image datasets.

From the literature it can be observed that deep learning methodologies based on capsule networks have shown its potential for medical image segmentation but it has not been used for teeth segmentation from panoramic X-rays. Thus, leaving scope for capsule network to be explored for dental image segmentation.

5.3 Proposed Method

5.3.1 Overview

The proposed capsule network termed as TeethCaps is inspired by SegCaps[63]. The primary objective of the proposed capsule network is to segment the teeth region and discard the irrelevant information of background parts such as spinal bones, nasal bones, jaw bones and other non-important parts from the dental panoramic X-ray images. The proposed TeethCaps has three major components feature extractor block, attention block and capsule block. Four convolutional layers with different dilation rates are introduced in the feature extractor block, which converts the original image to a feature map while maintaining the relevant information. After that, an attention mechanism is applied to these feature maps by using a convolutional block attention module (CBAM)[116], which refines the feature map in the attention block. At last, the refined features are passed to the capsule block comprising of the convolutional and deconvolutional capsules. The convolutional capsules extract and encode the features in vector form while deconvolution capsules up-samples the feature map. This block is responsible for maintaining precise and spatial information to perform finer segmentation.

5.3.2 Details of the Proposed Architecture TeethCaps

The major components of the proposed TeethCaps are a feature extractor block, attention block and capsule block. The proposed model is trained in end-to-end manner using the dice loss function. The detailed architecture of the proposed model is shown in Figure 5.1.

Feature Extractor Block: The input is passed initially to the feature extractor block which has four convolutional layers. The first two are conventional convolutional layers, while the last two are dilated convolutional layers. The advantage of this is that it provides more relevant features with high dimensions to the attention block. The layers are organised sequentially, with each layer followed by ReLU function. An input image of size [512 x 512] is passed to these layers to extract the prominent features for teeth region. The size of the kernel used is [5 x 5] while the number of channels is increased from 16 to 64. The dilation rate for the dilated convolution layers is set to 3. The output is a feature of size [512 x 512 x 16] which is passed to the attention block.

Attention Block: The attention block comprises of CBAM, which may be incorporated into any convolutional architecture without adding any additional complexity to model design. The main purpose of introducing this block is to provide more attention to teeth region portion instead of overall image and to minimize the spatial information loss caused by the convolution and pooling operations. The CBAM consists of two modules channel attention module (CAM) and spatial attention module (SAM) where CAM generates the channel attention map C_f having meaningful information whereas SAM produces spatial attention map S_f having information about location of the meaningful information. In attention block a feature map F_m obtained from feature extractor block is passed as an input and a refined feature map F_f' of the same dimension is generated as output. The refined features generated has local and vital information of

the teeth region which is then passed to the capsule block for segmentation task. The working of attention block is summarized as:

$$F_f' = S_f\big(\, C_f(\, F_m\,) \textcircled{X} F_m\big) \textcircled{X}\, F_m \qquad (5.1)$$

Where $\textcircled{X}$ represents element-wise multiplication. C_f denotes channel attention map, S_f denotes spatial attention map. F_m is feature map output from feature extractor block. F_f' is the refined feature map output.

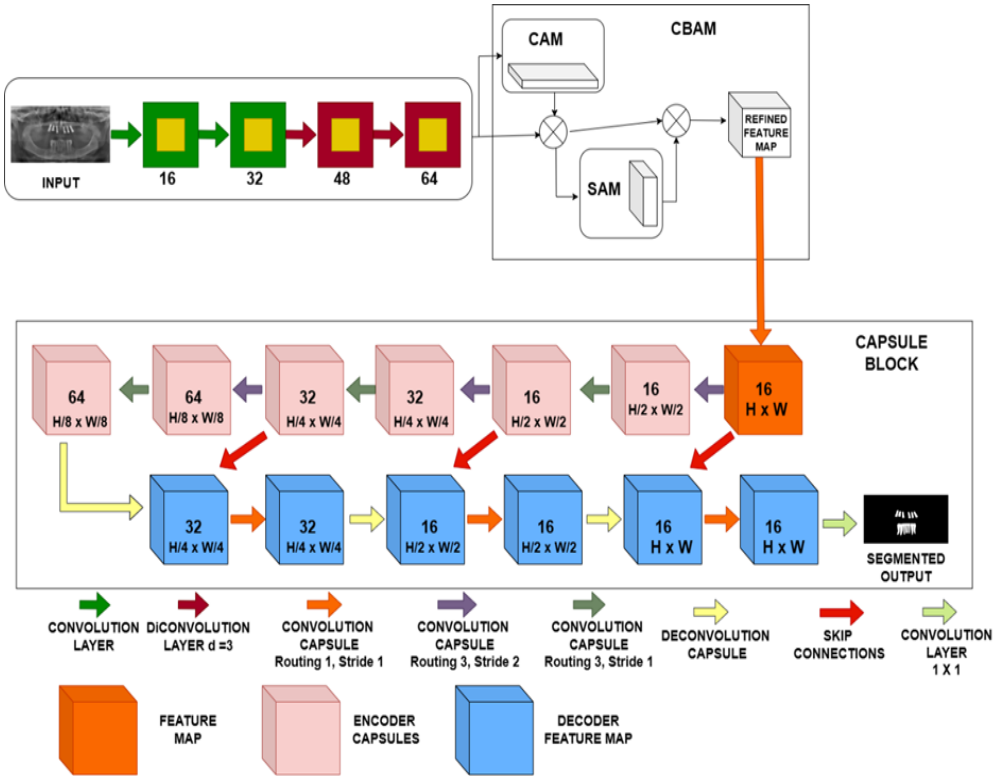

Figure 5.1 Architecture details of the proposed TeethCaps

Capsule Block: The refined feature map generated from the attention block is fed as an input to the capsule block, producing the predicted binary segmented map corresponding to the teeth region. The capsule block consists of the encoder and decoder parts. The encoder part comprises the convolutional capsules that extract and encode features in vector format. The decoder part consists of the deconvolutional capsules whose work is

to upscale the feature map generated from encoder part to reconstruct the final segmented teeth map. The skip connections are utilized to concatenate the features extracted from convolutional capsules in the encoder to corresponding deconvolutional capsules in the decoder. The network takes care of both local and global information for precise segmentation. In the capsule block, the locally constrained routing is utilized for the communication between the capsules. The algorithm 5.1 shows the locally constrained routing. In this, the child capsules are routed to the parent within a user-defined kernel rather than every child to every parent. The set of capsules is considered as capsule-type ct_i^m in layer m. Child capsules will attempt to resemble the parent capsule's output during first routing iteration. The prediction vector $\hat{u}_{a,b|ct_i^m}$ of the convolutional capsules are calculated by the matrix multiplication of transformation matrix $W_{ct_i^m|a,b}$ and the activation vector of the child capsule $u_{a,b|ct_i^m}$, which is represented by the following equation:

$$\hat{u}_{a,b|ct_i^m} = W_{ct_i^m|a,b} * u_{a,b|ct_i^m} \quad (5.2)$$

Parent capsule input is calculated by following equation:

$$p_{a,b} = \sum_n d_{ct_i^m|a,b} * \hat{u}_{a,b|ct_i^m} \quad (5.3)$$

Where $p_{a,b}$ is the parent capsule and $d_{ct_i^m|a,b}$ is the coupling coefficient between child capsule and parent capsules. The value of coupling coefficient is obtained by the routing softmax given in following equation:

$$d_{ct_i^m|a,b} = \frac{\exp(k_{ct_i^m|a,b})}{\sum_n \exp(k_{ct_i^m|n})} \quad (5.4)$$

Where $k_{ct_i^m|a,b}$ is considered as log prior probabilities that prediction vector should be routed to the parent capsules.

The output capsule is calculated by the squash function:

$$v_{a,b} = \frac{||p_{a,b}||^2}{1+||p_{a,b}||^2} \frac{p_{a,b}}{||p_{a,b}||} \quad (5.5)$$

From the second routing iteration onwards agreement between the child capsule and parent capsule is computed by equation 5.6 and child capsules which closely approximate parent capsule's output is maximized while for rest it is minimized.

$$o_{ij} = p_{a,b} \,.\, \hat{u}_{j|i} \quad (5.6)$$

Loss Function: The model is trained on dice loss defined as:

$$Loss_{DICE} = 1 - \frac{2Pr_i\, Gt_i}{Pr_i + Gt_i} \quad (5.7)$$

where $Gt_i \in \{0, 1\}$ is considered as input image's ground truth and $Pr_i \in \{0, 1\}$ represents the predicted segmentation map pixel from input image.

Algorithm 5.1 shows the procedure for locally constrained routing between the capsules

Algorithm 5.1 Locally Constrained Routing

Input: Feature Map of size 512 X 512 X 16

Output: Vector Output of Capsule at spatial location

1: **Routing (**$\hat{u}_{a,b|ct_i^m}, r, m, a, b$)

2: **for** all capsule-types ct_i^m at position (a, b) in layer m and capsule-type ct_j at layer $(m + 1)$ **do**

3: $k_{ct_i^m} \leftarrow 0$

4: **end** for

5: **for** iteration = 1, 2,..., r **do**

6: for all capsule-types ct_i^m in layer m : $d_{t_i^m} \leftarrow$ softmax ($k_{ct_i^m}$)

7: for all capsule-types ct_j in layer $(m + 1)$: $p_{a,b} \leftarrow \sum_n d_{ct_i^m|a,b}\, \hat{u}_{a,b|ct_i^m}$

8: for all capsule-types ct_j in layer $(m + 1)$: $v_{a,b} \leftarrow$ squash ($p_{a,b}$)

9: for all capsule-types t_i^l in layer l and capsule-type t_j at layer$(m+1)$:

$k_{ct_i^m|a,b} \leftarrow k_{ct_i^m|a,b} + \hat{u}_{a,by|ct_i^m} . v_{a,b}$

10: **end** for

11: **return** $v_{a,b}$

5.4 Experiments and Results

5.4.1 Datasets

The two-benchmark dataset are used to validate the proposed TeethCaps. The first dataset is UFBA_UESC[2] dental dataset. The second benchmark dataset is Tufts dental database. For training, validation and testing datasets is segregated in 3 parts with a split ratio of 8:1:1 that is there are 1200 randomly selected images in the training set and remaining 300 images are equally divided between validation set and test for first dataset whereas for second dataset 800 randomly selected images are in training set while 100 image each in test and validation set.

5.4.2 Experimental Setup

The proposed TeethCaps is implemented in an end-to-end way on 4 NVIDIA GeForce GTX 1080i GPUs in python using Pytorch libraries. The learning rate, batch size and the number of epochs is set to 0.001, 16 and 100 respectively and an Adam optimizer[100] was used for optimization. The routing iterations were set to 1 and 3.

5.4.3 Result and Discussion

The proposed TeethCaps achieved a segmentation performance with an accuracy of 97.4%, a dice score of 92.8%, IoU of 91.6%, a precision of 95.6% and a recall of 89.8%

for dataset 1. For dataset 2 segmentation performance achieved is accuracy 98%, dice score 91.8%, IoU 91.1%, precision 96.1% and recall 92.6%. The performance is compared with state-of-the-art deep learning models all of which are utilizes the CNNs and capable for segmentation of dental panoramic X-rays. These deep models are SegNet[103], UNet[39], BiseNet[101], CENet[105], Unet++[107], Nanonet[4]. The comparison is also made with capsule network based architecture called SegCaps[63]. The parameter setting for all the models compared is kept same for fair comparison. Both the datasets are split in the same ratio for training testing and validation sets.

Tables 5.1 and Table 5.2 demonstrate the quantitative result comparison of the proposed model with other deep models for dataset 1 and dataset 2. For dataset 1 the proposed model performs better in terms of accuracy, IoU and dice score compared to state-of-the-art whereas for the dataset 2 it out the state-of-the-art methods. The proposed model also beats the SegCaps for both datasets.

Table 5.1 Comparative analysis of different methods for Dataset 1with proposed capsule bed model

Models	**Accuracy**	**Precision**	**Recall**	**IoU**	**Dice Coefficient**
SegNet [103]	96.1	92.1	88.6	82.4	90.3
UNet[39]	96.2	94.2	86.8	82.4	90.3
BiseNet[101]	92.7	92.5	69.6	78.7	79.4
CENet[105]	96.7	93.3	**90.2**	84.7	91.7
Unet++[107]	95.1	92.5	82.6	77.4	87.3
NanoNet[4]	96.6	95.0	82.8	89.9	91.3
SegCaps[63]	96.5	**96.0**	86.5	89.6	91.0
Proposed Model	**97.4**	95.6	89.8	**91.6**	**92.8**

For visual inspection figures 5.2 and 5.3 shows the predicted segmented teeth map of the proposed model and the deep models for dataset 1 and dataset 2. It can be observed that the proposed model accurately segmented the teeth region while suppressing the other parts like jaw bone, nasal bones and spinal bones. The BiseNet performs poorly compare to others exhibiting its weakness to deal with low contrast images. It can be observed that SegCaps misclassifies correct teeth region pixels with the background pixels. Overall the proposed model produced the better visual results than the state-of-the-art methods.

Table 5. 2 Comparative analysis of different methods for Dataset 2 with proposed capsule based model

Models	**Accuracy**	**Precision**	**Recall**	**IoU**	**Dice Coefficient**
SegNet[103]	96.1	92.1	88.6	82.4	90.3
UNet[39]	96.2	94.2	86.8	82.4	90.3
BiseNet [101]	92.7	92.5	69.6	78.7	79.4
CENet [105]	96.7	93.3	90.2	84.7	91.7
Unet++ [107]	95.1	92.5	82.6	77.4	87.3
NanoNet [4]	96.6	95.0	82.8	89.9	91.3
SegCaps[63]	96.9	94.9	83.0	88.3	88.5
Proposed Model	**98.0**	**96.1**	**92.6**	**91.1**	**91.8**

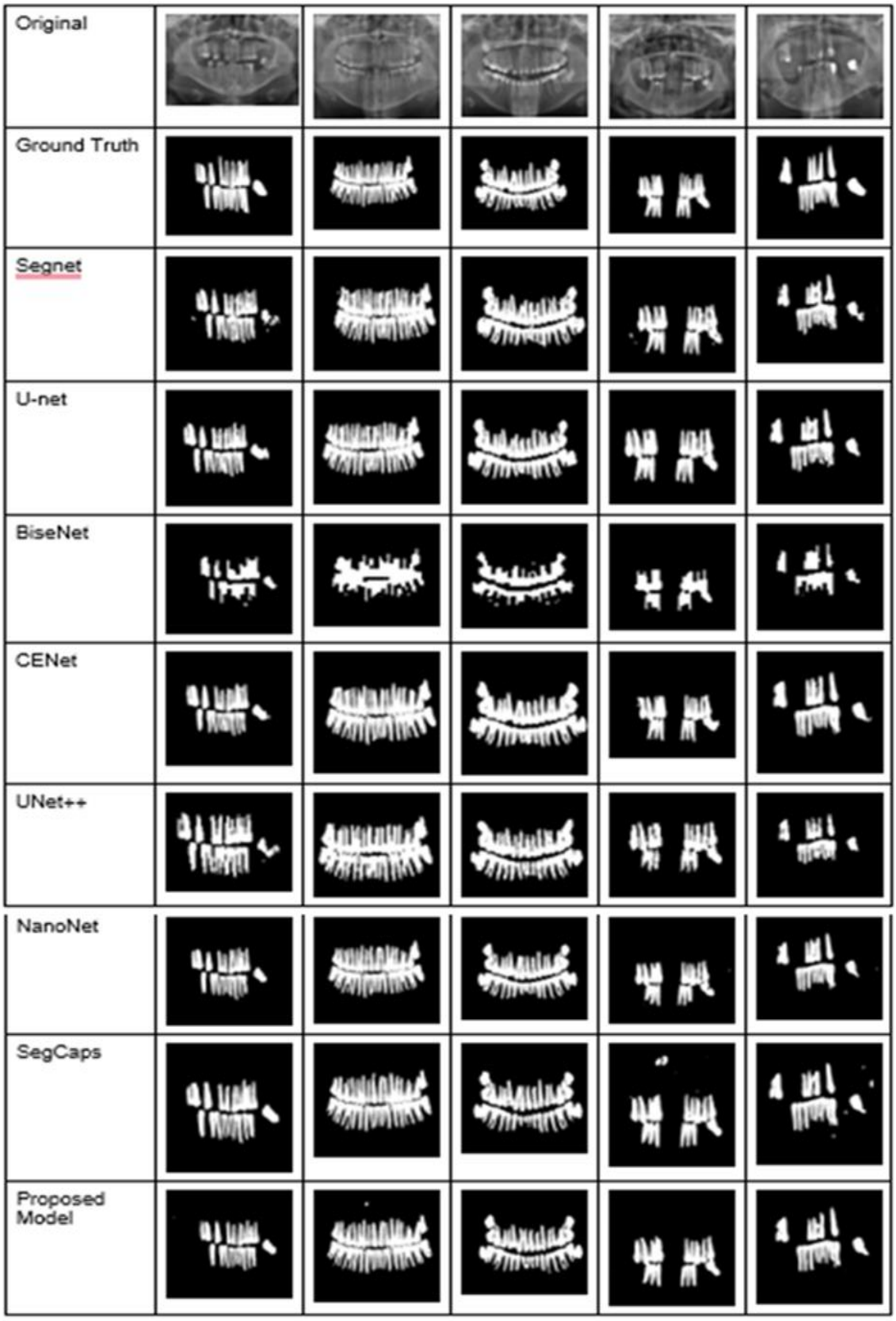

Figure 5.2 Visual results comparative analysis of different methods for Dataset 1 with proposed capsule-based model

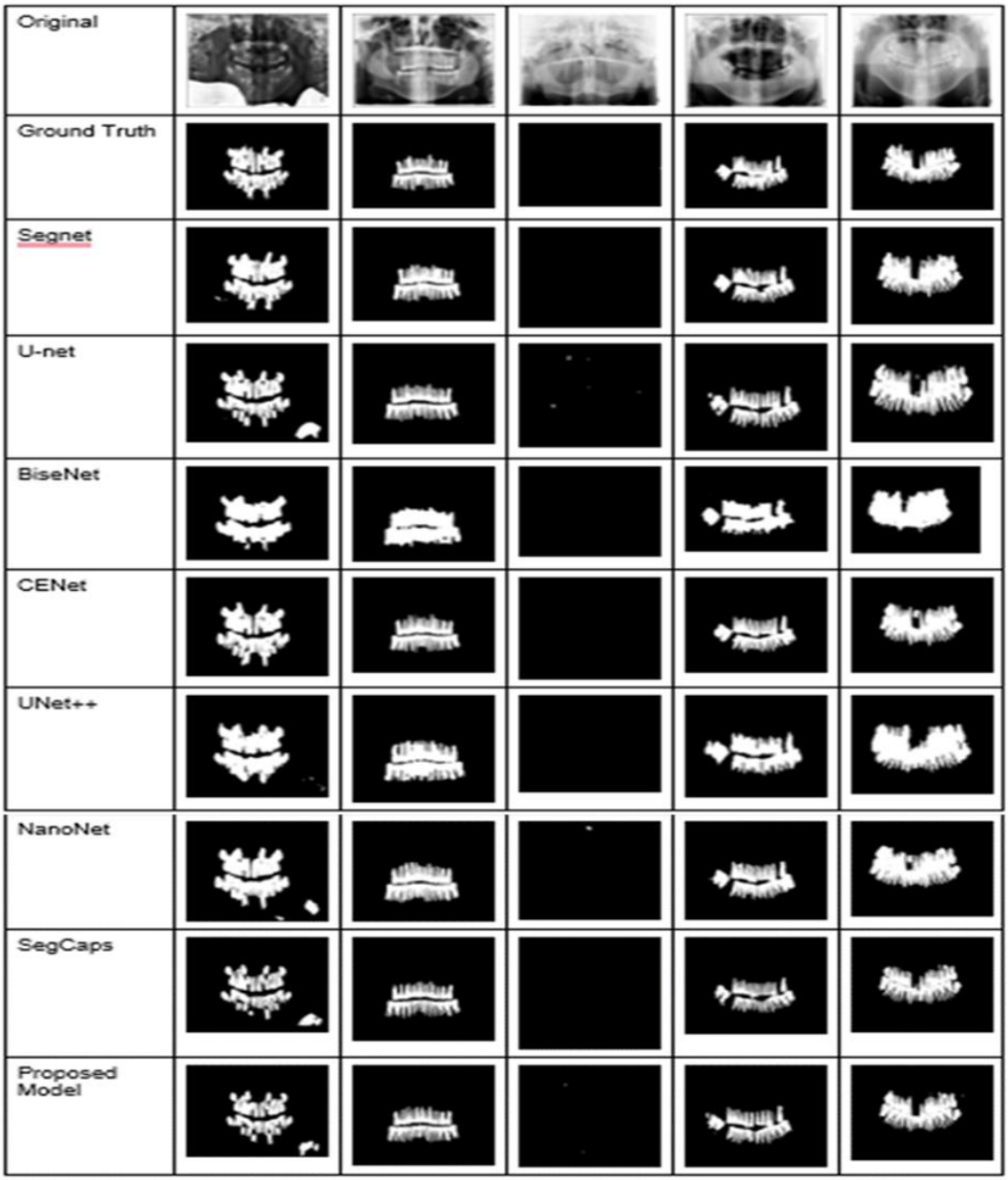

Figure 5.3 Visual results comparative analysis of different methods for Dataset 2 with proposed capsuled-based model

5.5 Conclusion

This chapter explored the potential of the capsule architecture for the automated and accurate teeth segmentation from dental panoramic X-ray images. A novel capsule-based architecture called TeethCaps was proposed for segmentation task. A series of convolutional layer with different dilation rate was introduced to maintain shallow

features and obtain rich feature map. Attention block using CBAM was added for improving segmentation accuracy. The segmentation task was performed by the capsule-based architecture using the locally constrained routing algorithm. The results shows that the proposed model outperforms stat-of-the-art methods. Overall, the proposed model performed well for segmentation of dental panoramic X-rays thus paving a way for capsule architecture for segmentation tasks in dental imaging.