CHAPTER 3

TEETH IMAGE SEGMENTATION METHODS USING MULTIMODAL CNN ARCHITECTURE AND ATTENTION GUIDED DEEP NETWORK

In this chapter, two deep learning methods have been proposed for automatic teeth area segmentation from dental panoramic X-ray images. The first method entitled "Exploiting Multimodal CNN Architecture for Automated Teeth Segmentation on Dental Panoramic X-ray Images" is designed to exploit the multi-CNN features from different CNN streams. These multi features are combined to gather rich dental information to improve the quality of segmentation. The second method "Cascaded Deep Neural Network with Attention Guidance for Teeth Segmentation on Dental Panoramic Radiographs "consists of cascading two deep models. This model extracts rich teeth contextual information from variants of CNN. The attention mechanism is designed to pay attention on the teeth boundaries and preserve shape information. The second model is designed to learn three important features: luminance, contrast and structural properties to improve the quality of segmented teeth map. The model presented is able to segment proper teeth region, preserve edges, shape information as well removes the blur pixels. The proposed models are evaluated on the publicly available benchmark dataset comprising 1500 dental panoramic X-ray images. In addition, an extensive ablation study is performed to show the effectiveness of the proposed models.

3.1 Background

Dental panoramic X-ray images are the major source used in dentistry to analyse the patient's dental structure and to identify various dental irregularities like carries, bone abnormalities, infections in gums and teeth, tumours, cysts etc[87],[88]. The accurate teeth segmentation from these images provides useful information to dentists, which helps draw better and faster treatment strategies for patients. In addition, segmented dental radiographs are also used in the forensic identification of a person while detecting their age and gender based on their dental features[89],[90].

Dentists manually analyse these panoramic X-ray images or use semi-automatic methods for segmenting teeth which requires expert knowledge and thus can sometimes lead to an error or misjudgement[2]. This is because panoramic radiographs are taken from outside of the mouth and cover the entire area surrounding teeth, including nasal bones, jawbones, spinal bone, and other details of the face area and the images obtained are of low quality[91][92]. Thus, requires high demand for automatic and accurate teeth segmentation in the dentistry community.

3.2 Related Work

For teeth segmentation, till now many traditional methods such as morphological operators[1], level-set[31], contour[93] and active contours[94] have been presented. These methods were used to improve segmentation quality by removing noises and enhancing the contrast but require image pre-processing steps thus affecting the efficiency. A binary support vector machine[95] with handcrafted features were also used but its performance relies on the quality of the feature extracted. Also, these methods suffer to give clear boundary details. Additionally, if applied on large datasets these methods perform poorly.

The emergence of deep learning-based methodologies has resulted in substantial advancements over traditional methodologies. Deep learning due to its great success in medical image analysis tasks, can be applied to teeth segmentation. Some researchers used Mask RCNN for tooth segmentation and detection[96],[97]. For segmentation, this approach requires specialized and precise instance labels that are usually unavailable and is a complex task. Some researchers modified U-Net[32],[98] and Dense-ASSP[99] but lacked in providing clear teeth boundaries are complex and has high number of parameters. Most models used conventional CNN thus leaving scope to explore different types of CNN.

The deep learning approaches have shortcomings as some approaches focuses on teeth region while ignoring the teeth boundaries, some approaches are prone to misclassify jaw bones as teeth, suffers from over or under segmentation and produce noisy results. Furthermore, some of these approaches utilize limited and annotated image datasets which affect contextual feature extraction and some are designed with complex network architecture having large number of trainable parameters.

Thus, to address these issues, this chapter presents two novel deep learning methods. The detailed explanation of the two proposed methods is illustrated in the subsequent sections.

3.3 Proposed Methods

3.3.1 Exploiting Multimodal CNN architecture for automated teeth segmentation on dental panoramic X-ray images

3.3.1.1 Background

Panoramic X-ray images are the major source used in field of dental image segmentation. However, such images suffer from the disturbances like low contrast,

presence of jaw bones, nose bones, spinal bone, and artifacts. Thus, to observe these images manually is a tedious task, requires expertise of dentist and is time consuming. Hence, there is need to develop an automated tool for teeth segmentation. Recently, few deep models have been developed for dental image segmentation. But such models possess large number of training parameters, thus making the segmentation a very complex task. Also, these models are based only on conventional CNN and lacks in exploiting multimodal CNN features for dental image segmentation. Thus, to address these issues, a novel encoder-decoder model based on multimodal-feature extraction for automatic segmentation of teeth area is proposed in this section.

3.3.1.2 Proposed Method and Model

3.3.1.2.1 Overview

The proposed model is inspired form the multi-CNN encoder and decoder architecture to exploit the feature of different CNN based models. The encoder is designed to exploit features from different types of CNN models like Conventional CNN, Atrous-CNN and Separable CNN. The main motivation behind such design is to take advantages of individual networks and combined to generate fine grained contextual features for teeth segmentation. The following subsections explain the pipeline of the proposed model.

3.3.1.2.2 Preprocessing

The original size of each of the dental panoramic X-ray image is [1991 × 1127] in the chosen dataset. The original image is resized to [512,512,3] where first two column represents height and width while third represents the depth. The original size of ground truth was [1991 × 1127] which is also resized as [512 × 512]. The images are resized for better memory utilization.

3.3.1.2.3 Architecture of the Proposed Model

The proposed model has two stages encoder and decoder. The encoder of the proposed model consists of three modules. Each module in encoder will get same image as an input. The first module is based on the convolutional layers to extract the contextual features. This module includes three convolutional layers along with batch normalization and leaky ReLU (Rectified Linear Units) as activation function. The Max pooling layer of size [2×2] is used for down-sampling. The second module contains the three dilated convolutional layers along with batch normalization. Activation function used here is leaky ReLU. Max pooling layers sized [2×2] are used for down-sampling of images. The advantage of dilated convolution is, a larger receptive field thus no loss of coverage. Also dilated convolution layer is computationally efficient as larger coverage is provided at same cost. The use of dilated convolution layer ensures that there is no loss of resolution of output image. The third module includes the three depth-wise convolution layers along with batch normalization and leaky ReLU as activation. Max pooling is used for down-sampling. The advantage of depth-wise separable convolution when compared to standard CNNs is they have lesser parameters to adjust, which minimizes overfitting. Finally, the output of all the three modules is fused and passed to decoder. The decoder part consists of a single path in which four deconvolution layers are used to reconstruct the image. The first three deconvolution layer are used along with batch normalization and leaky ReLU as an activation function. The last deconvolution layer uses sigmoid as an activation function for classification of the pixels. The up-sampling layers are used to reconstruct the image to original size. The proposed can be trained in end-to-end fashion. Figure. 3.1 shows the detail architecture design of proposed model and Table 3.1 shows the configuration of different layers used in the proposed model.

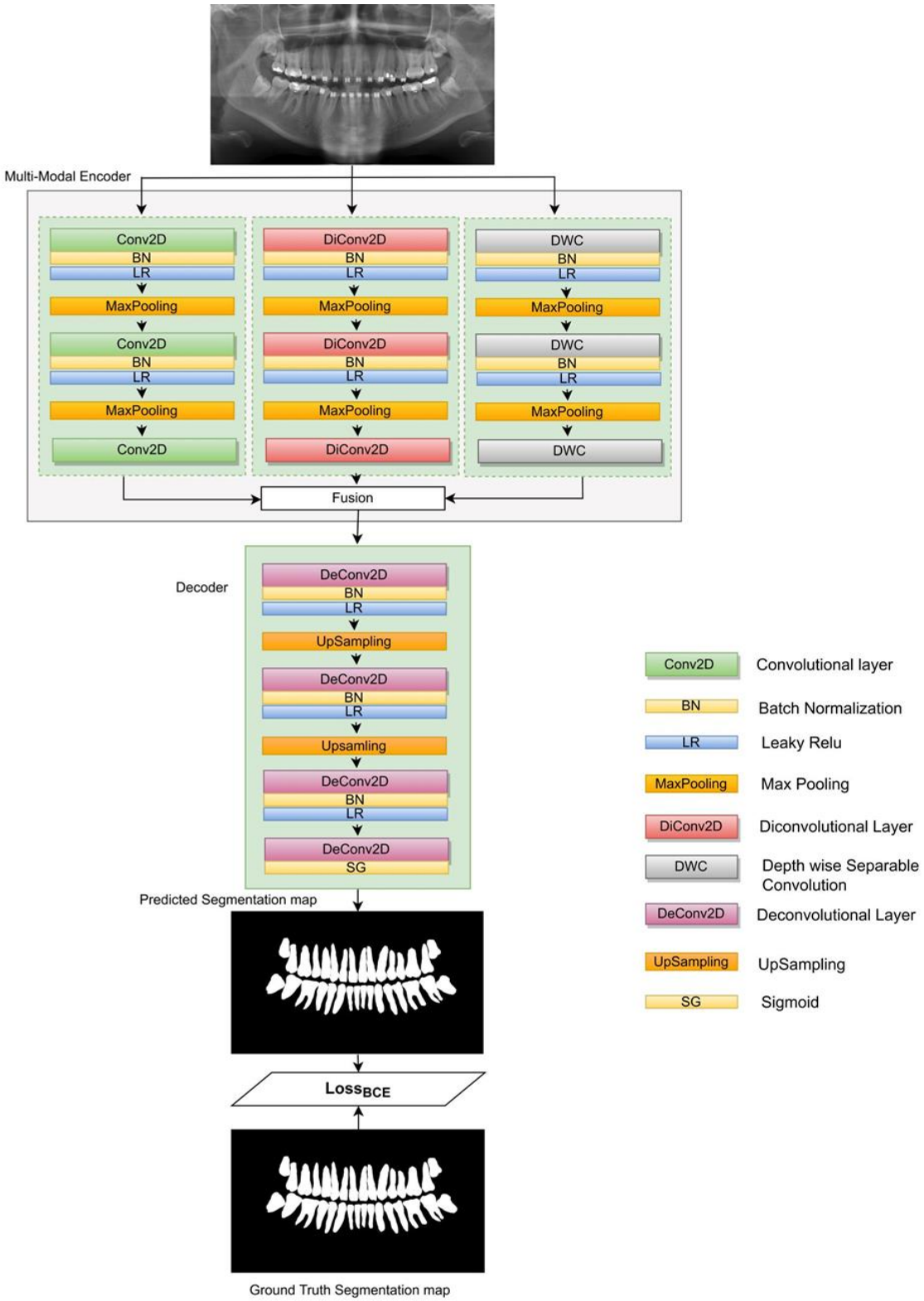


Figure 3.1 Architecture of the proposed model

Table 3.1 The details of the layers used in proposed model.

Layer Name	Number of Kernel	Kernel Size	Dilation Rate	Depth Multiplier
Convolution2D	64	(5,5)	-	-
Convolution2D	48	(3,3)	-	-
Convolution2D	32	(3,3)	-	-
Diconvolution2D	64	(5,5)	4	-
Diconvoluoion2D	48	(4,5)	4	-
Diconvolution2D	32	(5,6)	4	-
Depthwise Separable Convolution	-	(5,5)	-	2
Depthwise Separable Convolution	-	(5,5)	-	2
Depthwise Separable Convolution	-	(5,5)	-	2
DeConvolution2D	32	(5,6)	-	-
DeConvolution2D	48	(5,4)	-	-
DeConvolution2D	64	(5,4)	-	-
DeConvolution2D	1	(1,1)	-	-

3.3.1.2.4 Optimizing the Proposed Model

Let θ be all the trainable parameter of proposed model. Let the set $P = \{P_1, P_2, P_3, \times P_n\}$ be the predicted segmented map output and the set $G = \{G_1, G_2, G_3, \times G_n\}$ be the ground truth. The loss between P and G are obtained by using the binary cross-entropy loss. The following equation describes the loss between P and G:

$$Loss_{BCE} = -(G_i(log\ P_i) + (1 - G_i)\log(1 - P_i)) \quad (3.1)$$

where $P_i \in \{0, 1\}$ is predicted segmented map of input image and $G_i \in \{0, 1\}$ is ground truth of input image. Further, the loss obtained by equation (3.1) is minimized by using Adam optimizer[100].

3.3.1.3 Experiments and Results

3.3.1.3.1 Dataset

The UFBA_UESC[2] dental images dataset is used to evaluate the proposed model. This dataset includes 1500 dental panoramic X-ray images. The images of this dataset are categorized into 10 different classes of which 1200 images were randomly selected for training set while 150 images each were selected as validation set and test set respectively.

3.3.1.3.2 Experimental Setup

The proposed model has been implemented on 4 NVIDIA GeForce GTX 1080i GPUs in Python by using Keras API and TensorFlow as background in end-to-end fashion. The batch size, learning rate, momentum of batch normalization is set to 8, 0.0001 and 0.95 respectively. The number of epochs of the proposed model was set to 500. For avoiding overfitting this paper adopted early stopping method to halt the training of the proposed model. The patience parameter of the early stopping is set to 4. The early stopping method observe the value of training loss at every epoch and it stops training when the loss is not changed after 4 consecutive epochs. It has been observed that at epoch number 39, the model stops its training. Figure 3.2 plots training and validation losses for every epoch.

Figure 3.2 Training versus Validation Loss. The plot shows the training and validation loss of the model for each epoch.

3.3.1.3.3 Results and Discussion

The proposed model achieves accuracy, precision, recall and IoU of 96.06%, 95.01%, 94.06% and 76.62% respectively. The total parameters of the proposed model are 0.338M. For comparative analysis, the performance of the proposed model is compared with current state-of-the- art deep learning and conventional approaches.

3.3.1.3.3.1 Comparison with state-of-the art deep approaches models

The proposed model is compared with U-Net[39], BiseNet[101], Dense-ASPP[102], SegNet[103] and TSASNet[54] on different metrics and number of parameters which is presented in Table 3.2. It is observed that for accuracy of the proposed method stood third in the table while proposed outperforms other approaches in terms of precision and recall. The specificity of the model is also high which demonstrates the model has focused more on background pixels thus could be the reason for hampering the performance in terms of accuracy and IoU. The recall for proposed model is better among the other models which depicts that model has also paid attention towards the foreground pixels. The best part of the proposed model is, it totally out performs the state-of-the-art methods in terms of number of parameters. The number of parameters used by the proposed model is only 0.338 which is far less when compared with [39],[101],[102],[103],[54]. The visual results of the proposed method can be seen in Figure 3.3. It can be observed that the proposed can correctly identify the segmentation of tooth area while removing the other parts like jaw bones, nasal bones and neck bones for ease of use in clinical practices.

Table 3.2 Performance comparison with state-of-the-art deep learning approaches

Model Names	Accuracy	Specificity	Precision	Recall	Parameters (M)
U-Net [39]	96.04	97.68	89.89	90.18	31.04
BiseNet [101]	95.05	95.98	85.83	92.48	12.2
Dense-ASPP[102]	95.50	97.76	90.09	86.88	46.16
SegNet[103]	96.38	98.32	92.56	89.05	29.44
TSASNet[54]	**96.94**	97.81	94.97	93.77	78.27
Proposed Method	96.06	**99.92**	**95.01**	**94.06**	**0.338**

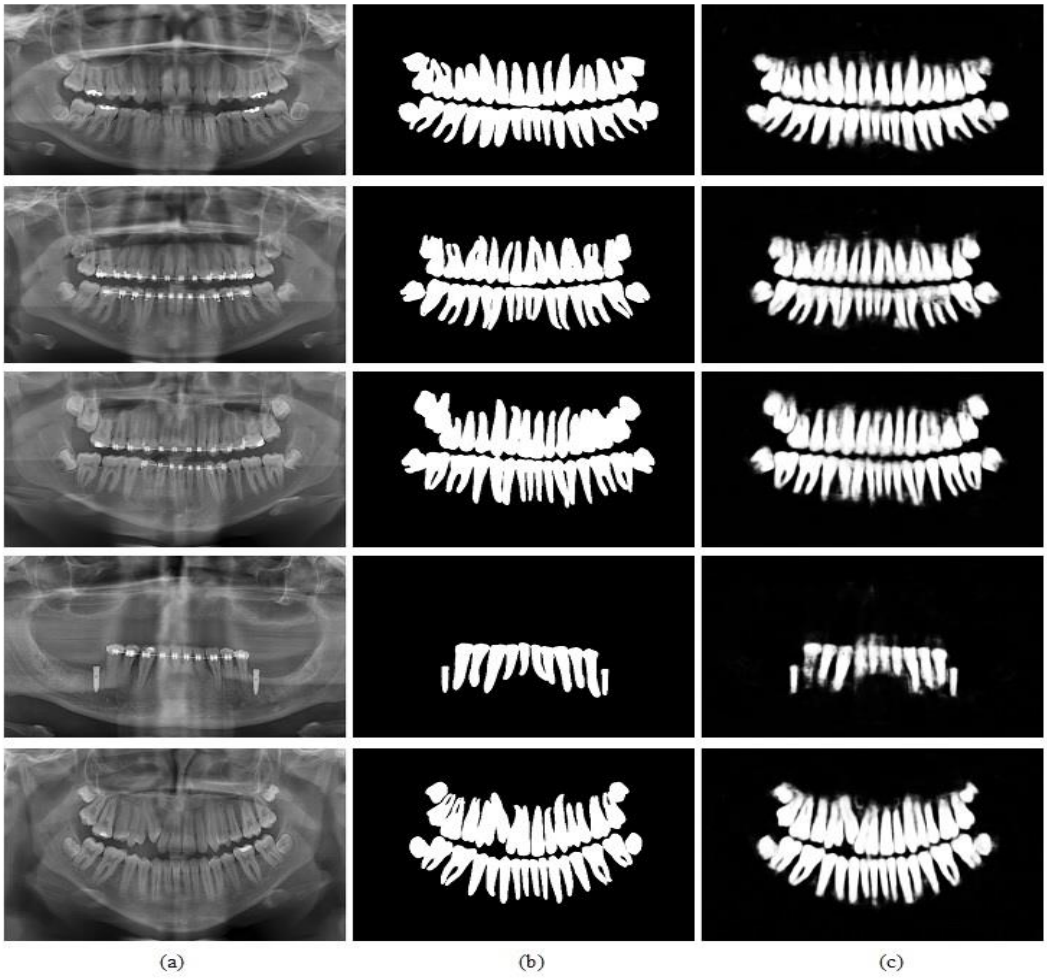


Figure 3.3 Qualitative results of the predicted segmentation maps. (a) Original Images (b) Ground truth segmentation maps of corresponding maps (c) Predicted segmentation maps

3.3.1.3.3.2 Comparison with state-of-the art conventional methods

The proposed model is also compared with the conventional approaches and the results are shown in Table 3.3. The proposed method out-performs traditional methods which requires the prior knowledge for segmentation while the proposed method can perform automatic segmentation and can be used in practical applications.

Table 3.3 Performance comparison with conventional approaches

Model Names	Accuracy	Specificity	Precision	Recall
FCM [11]	82.15	90.78	61.42	45.03
Level set [31]	75.50	78.45	48.37	68.49
Splitting & Merging [109]	81.33	99.15	81.24	8.14
Region growing [2]	68.10	69.48	35.53	63.41
Global Thresholding [2]	79.29	81.91	52.02	69.31
Proposed Model	**96.06**	**99.92**	**95.01**	**94.06**

3.3.1.3.3.3 Ablation Study

The ablation study of the proposed model is conducted to validate the proposed model. The same dataset has been used and evaluation has performed on accuracy, precision, recall, and intersection over union. Six different modules are creates based on the several combinations of three columns of the encoder while the decoder is same of all the six modules. These are,

M1-C1: The Module-1 Column-1 contains first column of the encoder.

M2-C2: The Module-2 Column-2 contains second column of the encoder.

M3-C3: The Module-3 Column-3 contains third column of the encoder.

M4-C13: The Module-1 Column-1 contains column one and three of the encoder.

M5-C12: The Module-1 Column-1 contains column one and two of the encoder.

M6-C23: The Module-1 Column-1 contains column two and three of the encoder.

Table 3.4 shows the quantitative result of each module where module M5-C12 has high accuracy, precision and IoU while module M1-C1 has high recall. The individual modules of the proposed architecture predict background pixels more (specificity) as compared to foreground pixels which result in high specificity when compared to recall for individual modules. Thus, the representation power of individual modules is less but when all the modules are combined, the feature encoding get enhanced, which results in higher recall and IoU and provides better results when compared to all other modules. The proposed performs better than individual modules in terms of all performance metrics. The qualitative results are shown in Figure 3.4. It can be seen that module five has better visibility as compared to other modules and it closer to proposed method.

Table 3. 4 Comparison of several modules in ablation study

Modules	**Accuracy**	**Precision**	**Recall**	**IoU**
M1-C1	93.29	88.26	80.36	62.03
M2-C2	93.93	92.83	78.44	62.97
M3-C3	93.75	90.19	80.14	63.21
M4-C13	93.49	94.07	74.83	62.64
M5-C12	94.53	94.96	79.37	68.44
M6-C23	93.63	92.17	77.57	63.36
Proposed Model	**96.06**	**95.05**	**94.06**	**76.62**

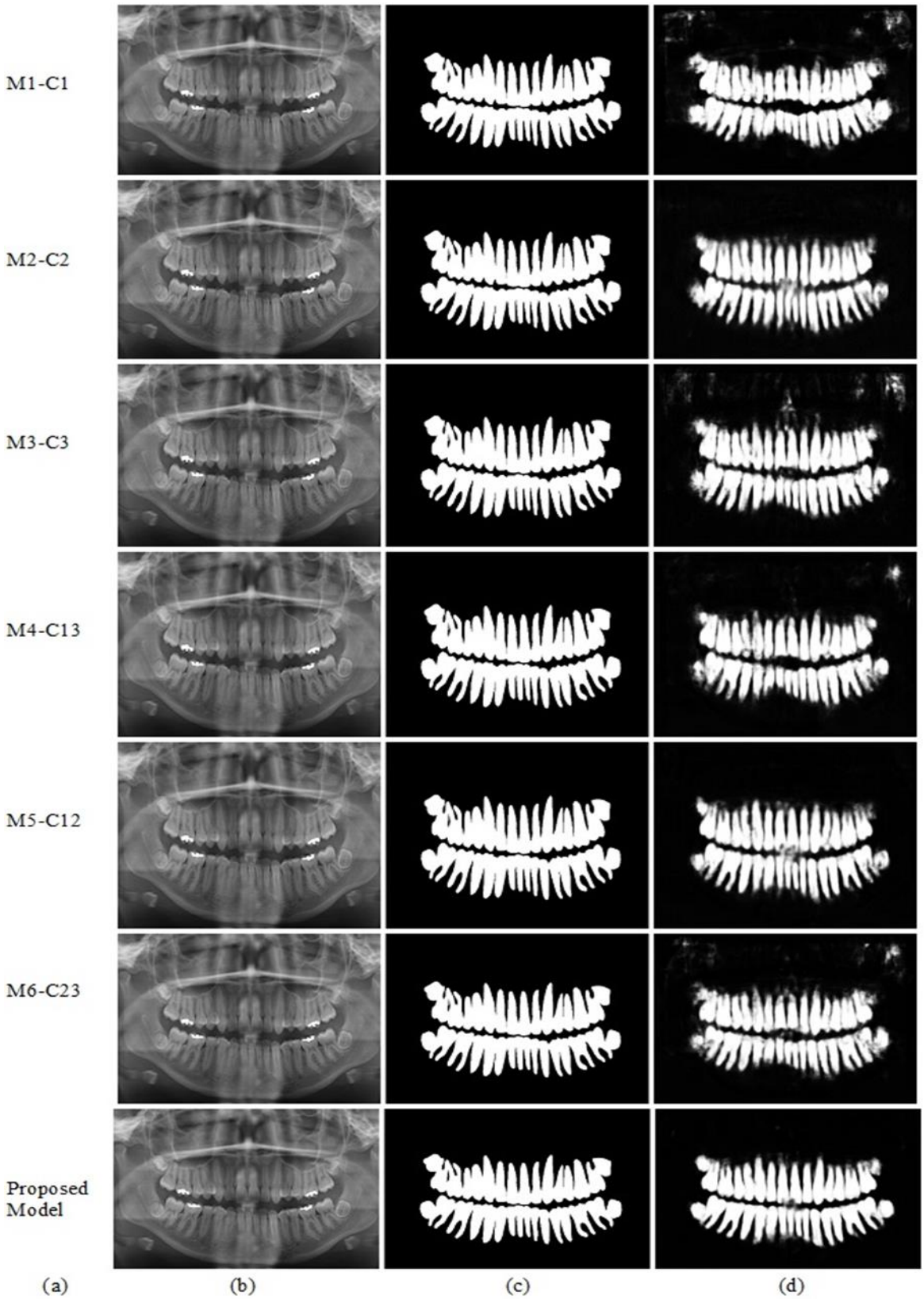


Figure 3.4 Qualitative results of the predicted segmentation maps of the modules. (a) Modules (b) Original Images (c) Ground truth (d) Predicted segmentation maps

3.3.1.3.4 Summary

This section proposed an end-to-end multimodal CNN architecture based on encoder and decoder structure for automatic segmentation of teeth area. The encoder part consists of three different CNN based architectures for extracting multiple features and contextual information from dental panoramic X-ray images. The decoder is simple single stream of deconvolutional layers. The experiment results shows that proposed model requires very less number of parameters and it out performs state-of-art segmentation method. The limitation of this work is, the model focuses on the background pixels due to which result is affected. In future, performance will be further improved by focusing on foreground pixels by using attention mechanisms. Also, weak boundaries to be determined by giving attention to the tooth boundaries.

3.3.2 Cascaded Deep Neural Network with Attention Guidance for Teeth Segmentation on Dental Panoramic Radiographs

3.3.2.1 Background

Teeth segmentation plays a vital role in dental imaging to diagnose and develop better treatment strategies. Few deep learning methodologies are proposed in literature based on CNN to segment the teeth region. However, these methods usually have large numbers of parameters, lack in exploiting fine-grained features and suffer in obtaining sharp teeth boundaries.This section presents a novel attention-guided deep cascaded network for automatic teeth segmentation to address these issues.

3.3.2.2 Proposed Method and Model

3.3.2.2.1 Overview

Teeth segmentation with better visibility of teeth boundaries and teeth area is always in greater need for clinical diagnosis. The proposed model addresses issues by

designing cascading two deep architecture for teeth segmentation. The block diagram of the presented methodology is shown in Figure 3.5. The first deep architecture is an encoder-decoder model which performs the following important tasks,

- Enhance the quality of the features by extracting and fusing features from multimodal CNN architecture in the encoder module. The main motive is to utilize the advantages of several CNN architectures. In this study, three different types of CNN networks are utilized: normal CNN, Spatial Separable CNN and Atrous (Dilation) CNN.
- To give attention to teeth boundaries, two Tooth Boundary Attention Blocks (TBAB) i.e., TBAB1 in the encoder and TBAB2 in the decoder are proposed. These two attention blocks map the feature to ground-truth teeth boundaries to give more emphasis to the teeth boundaries.
- To preserve the low-level features from teeth images, a skip connection is developed to surpass the multi-modal low-level features from the encoder to the decoder, concatenated with up-sampled map for input to TBAB2.
- The segmentation block (SegBlock) receives the fused features to generate the predicted teeth segmented map.

The second model of the cascaded architecture is solely designed to enhance the quality of the teeth segmentation map by proposing stacks of CNN layers. The model can also able to remove isolated or blurred pixels of the first model's output.

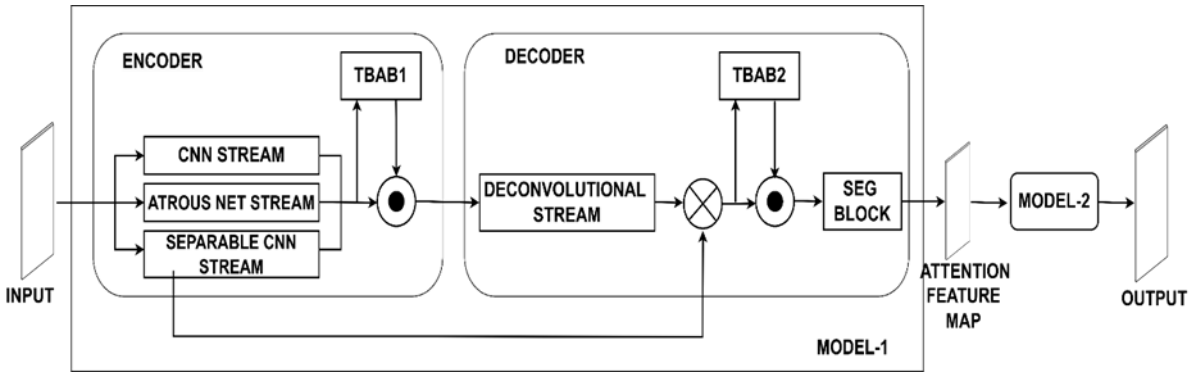


Figure 3.5 Block Diagram of the presented methodology

3.3.2.2.2 Details of the Proposed Architecture

The detailed architecture of the presented model is represented in Figure 3.6. The proposed model comprises of two cascading models. The first model is on multiple feature extraction from different CNN architectures. The encoder part is designed in such a way that it has three streams of different CNN architectures. All three streams receive the same dental radiograph as input.

The first stream comprises of four convolutional layers having different size of kernels to retrieve the contextual features from the images. An activation layer and batch normalization layer come after each convolution layer. The leaky ReLU is used as an activation function. For down-sampling Max pooling layer of dimension [2 ×2] is utilized. The second stream of encoder includes four dilated convolutional layers having different dilation rates with activation as leaky ReLU and batch normalization layer. Here also Max pooling of dimension [2 ×2] is applied to down-sample the feature maps. Dilated convolution has the advantage of having a more enormous receptive field and consequently no loss of coverage. Also, they are computationally efficient since it provides more coverage for the similar cost. The inclusion of an Atrous Net guarantees that the output image retains its resolution. The third stream consists of four separable convolutional layers. The activation function used here is leaky ReLU with batch

normalization. Again, Max pooling is applied in this stream for down sampling. Compared to normal CNNs, the advantage of separable convolution is that it spatially divides the convolution kernel and perform convolution based on spatially separable kernels followed by point convolution to merge the obtained features. This will result in fewer number of matrix multiplication hence making it computationally efficient and reduce overfitting. The output of each stream is fused and elementwise multiplication is performed with the output of TBAB1 and then passed it to the encoder part of the module. The decoder part comprises a single stream having four deconvolution layers with leaky ReLU as activation and each followed by a batch normalization. The upsampling layers are utilized here to generate a reconstructed image to its original dimensions.

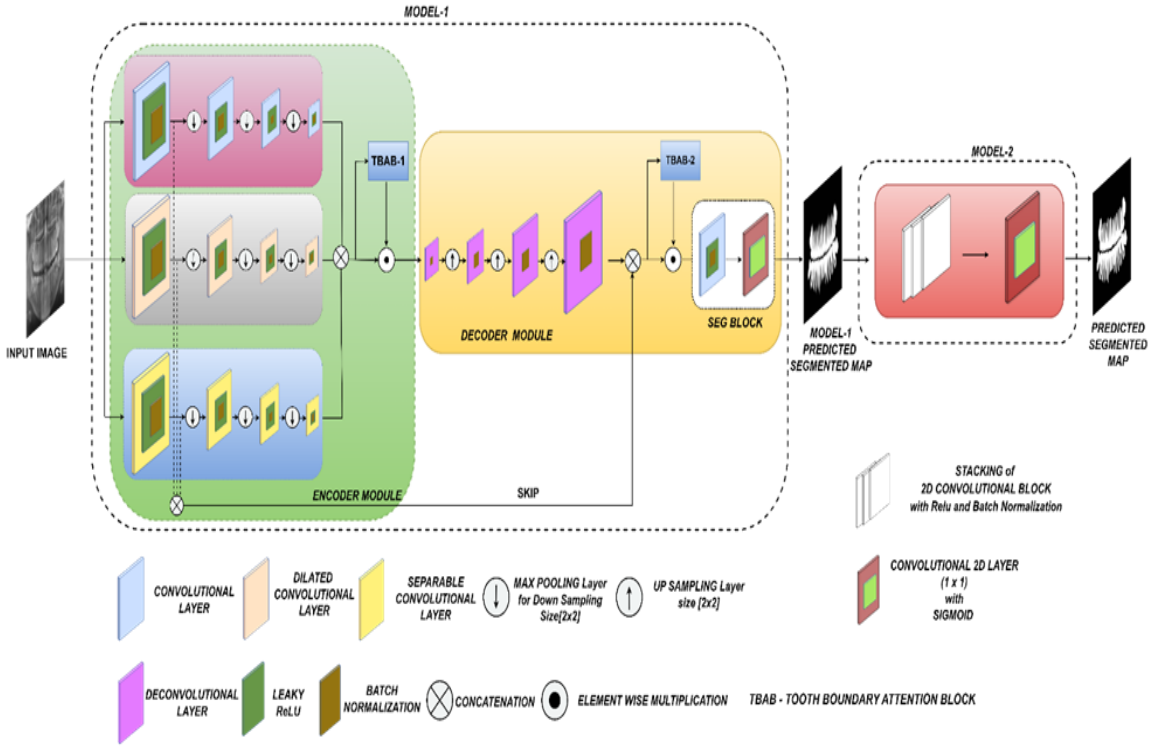


Figure 3.6 The detailed architecture of the proposed model.

The encoded features are given attention to learn teeth boundaries using these Teeth Boundary Attention Block (TBAB). The illustration of TBAB is given in Figure 3.7. The input to TBAB-1 is the fusion of output from all three streams of CNN. TBAB1 constitutes two layers of convolution block. The first one having kernel of dimension [3×3] with activation as leaky ReLU and batch normalization layer while the second convolution layer's kernel size is [1×1]. The loss is obtained by using binary cross-entropy between the generated edge map and its corresponding ground truth. Similarly, during decoding, the low-level features are given attention using TBAB-2 which has same structure as TBAB1. The long-skip connection is used to forward the low-level encoding information to the decoding module to enhance the feature learning process. The output of the first layer of each stream is concatenated and forwarded further using a long skip connection. This output is then fused with decoder output and is passed to the TBAB-2. The elementwise multiplication is again performed with fusion (of decoder and skip) and output of TBAB2 and the resultant feature maps are passed to SegBlock where the segmentation is performed.

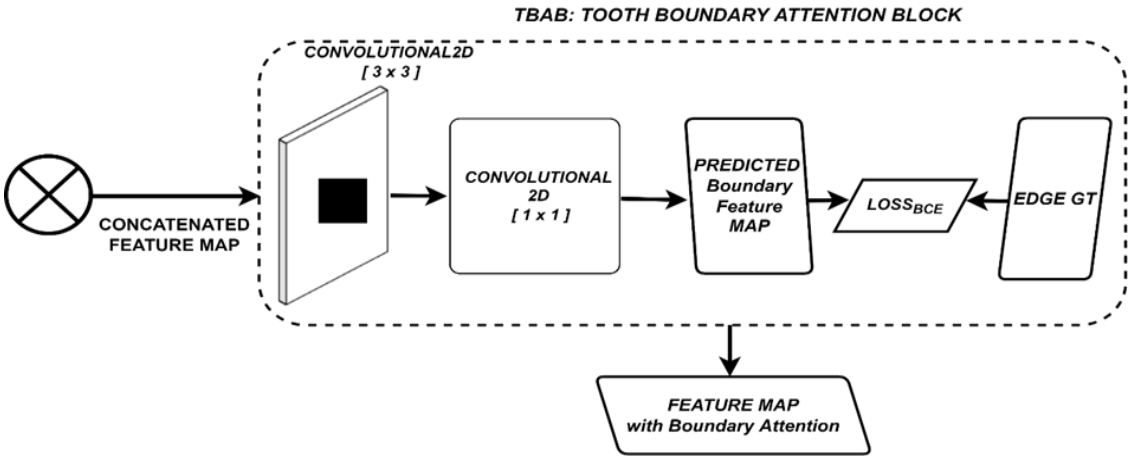
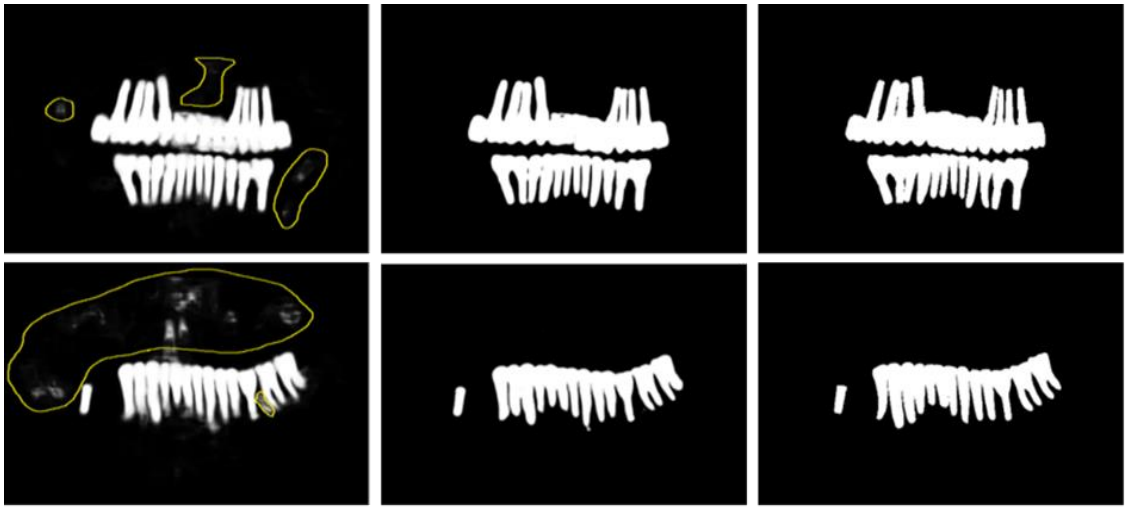


Figure 3.7 TBAB: Teeth boundary attention block

The SegBlock consists of [1×1] convolutional layer to predict the segmented teeth maps. The segmented map obtained from the first model lacks in quality and has corrupted pixels which are enclosed by yellow line in Figure 3.7. Therefore, a second cascaded model is designed to remove the blurriness and noisy pixels. The predicted segmented teeth maps obtained from the first model are further given as an input to the second model to enhance the predicted teeth region and teeth boundaries. The second model comprises of stacking of three convolutional layer that uses ReLU as activation function with batch normalization. The last layer is a [1×1] convolutional layer with sigmoid activation function utilized for classification of pixels. The second model employs the Structural Similarity Index Measure (SSIM) which examines the visual effect of three characteristics of an image: contrast, structure and luminance. The model is trained in an end-to-end way in step-by-step manner. Figure 3.8 shows the different outputs of Model-1 and Model-2 along with ground truth.



(a) Predicted Segmented Map of Model-1 (b) Predicted Segmented Map of Model-2 (c) Ground Truth

Figure 3.8 The example of noisy and blurry segmented map predicted by Model-1 marked by yellow lines whereas predicted segmented map by Model-2 removes corrupted pixels. (a) Predicted segmented predicted by Model-1. (b) Segmented Map of Teeth Area predicted by Model-2. (c) The ground truth of the correspondent images.

3.3.2.2.3 Optimizing the Proposed model

Two scales [64 X 64] and [512 X 512] of ground truth teeth boundaries are obtained from the ground truth segmentation maps by applying Canny edge detector followed by morphological operator using dilation to connect the missed edges. Let these are represented by two set: S = {s_1, s_2, s_3,……,s_k} and T = {t_1, t_2, t_3,……,t_k} where k being the total number of images. The segmentation maps of ground truth are represented by a set G = {g_1, g_2, g_3,……, g_n}.

The model is trained in minibatch manner. Let sets θ_{TBAB1} ,θ_{TBAB2} and θ_{SegBlock} represents all the trainable parameters for module TBAB1, TBAB2 and SegBlock respectively. The losses for three modules of first model are derived by the binary cross entropy (BCE) loss. The losses for each of three modules for each sample i is represented as:

$$Loss_{TBAB1_i} = Loss_i(s_i, P_{TBAB1_i}) = -\sum_{k=1}^{size} s_{i_k} log P_{TBAB1_{i_k}} \tag{3.2}$$

where $P_{TBAB1_{i_k}}$ is predicted edge segmented map for TBAB1. 'size' is the size of predicted feature map $P_{TBAB1_{i_k}}$.

The loss in decoder Tooth Boundary Attention Module is obtained by using binary cross-entropy loss as follows.

$$Loss_{TBAB2_i} = Loss_i(t_i, P_{TBAB2_i}) = -\sum_{k=1}^{size} s_{i_k} log P_{TBAB2_{i_k}} \tag{3.3}$$

where P_{TBAB2_i} is predicted edge segmented map for TBAB2. 'size' is the size of predicted feature map P_{TBAB2_i}.

The loss in the segmentation block is also calculated by using BCE loss defined below:

$$Loss_{SegMap_i} = Loss_i(g_i, P_{SegBlock_i}) = -\sum_{k=1}^{size} g_{i_k} log P_{SegBlock_{i_k}} \tag{3.4}$$

where and $G_i \in \{0,1\}$ is considered as input image's ground truth and $P_{SegBlock_i} \in \{0,1\}$ represents predicted segmentation map of input image.

The final loss function of the proposed model is obtained by adding all the above mentioned three loss functions and further minimized which is represented as follows:

$$\underset{\theta model1}{argmin} Loss_{final_i} \tag{3.5}$$

where

$$Loss_{final_i} = Loss_{TBAB1_i} + Loss_{TBAB2_i} + Loss_{SegBlock_i} \tag{3.6}$$

$$\theta_{model1} = [\theta_{TBAB1}, \theta_{TBAB2}, \theta_{SegBlock}] \tag{3.7}$$

The proposed model is optimized using minibatch based gradient descent approach due to its better properties like more stable converge towards the global minimum, faster learning and comparable or even better computational efficiency in comparison to other optimization approaches like batch-gradient and stochastic gradient approaches.

If the batch size is b and the number of samples of each batch is l , then the equation (6) can be represented as:

$$\underset{\theta_{model1}}{argmin} [Loss_{final}]^b \tag{3.8}$$

where,

$$[Loss_{final}]^b = \frac{1}{l} \times \sum_{i=1}^{l} Loss_{TBAB1_i} + Loss_{TBAB2_i} + Loss_{SegBlock_i} \tag{3.9}$$

Algorithm 3.1 shows the training and optimizing for Model-1

Algorithm 3.1 Training and Optimizing the Model-1

Input: Dental panoramic images of size 512 X 512

Ground-Truth Segmentation Maps: The set $G = \{g_1, g_2, \ldots\ldots, n\}$ represents ground truth for segmentation map for n images

Ground-Truth Teeth Boundary Maps: The set S= $\{s_1, s_2, \ldots\ldots, s_n\}$ represents ground truth of teeth boundary maps for n images of size 64 X 64. Set T= $\{t_1, t_2, \ldots\ldots, t_n\}$ represents ground truth of teeth boundary maps for n images of size 512 X 512.

Parameters: $learning\ rate\ (\eta)$, $Learnable\ parameters\ (\theta_{model1})$, $momentum$.

Initialisation: $Max\_Iter = 500$, $Batch_{Size} = 8, kernel$ regularize value of $L_2 - norm = 0.01$ $counter = 1, \eta = 0.001,$ and $patience\ value\ of\ early\_stopping = 5$.

Output: Optimize the Model 1

While *early_stopping* or $counter = Max\_Iter$ is fulfilled, do

  For each batch b *= 1 to* $\left\lceil \frac{N}{Batch\_Size} \right\rceil$ do

    For each sample i in batch bt do

      1. Find the loss of TBAB1 i.e., , $Loss_{TBAB1_i}$ loss of TBAB2 i.e., $Loss_{TBAB2_i}$ and

      Loss of SegBlock i.e.,$Loss_{SegBlock_i}$ using Equation- 1, 2 and 3 respectively.

    end for

    2. Obtain the mean of cumulative of loss for the given batch b using Equation-8.

    3. Find the gradients of the loss $\left[Loss_{final}\right]^b$ using [45].

    4. Obtain gradients, and optimize Equation-7 by updating Model-1's learnable parameters θ_{model1} using Adam[46] optimizer.

  End For

  5. $counter+= 1$

End While

The loss for the second model is calculated by SSIM to check the structural similarity between the predicted segmented map obtained from model 1 and its ground-truth. Let θ_{model2} denotes the trainable parameters of model2. Let the set $Y = \{y_1, y_2, y_3, \ldots\ldots, y_k\}$ be the predicted teeth segmented map from first model and the set $G =$

$\{g_1, g_2, g_3, \ldots\ldots, g_k\}$be the ground truth and represents the number of images. The loss is defined as:

$$Loss_{model2} = Loss\ (y_i, g_i) = 1 - \frac{(2\mu_{y_i}\mu_{g_i} + c_1)(2\sigma_{y_i g_i} + c_2)}{(\mu_{y_i}^2 + \mu_{g_i}^2 + c_1)(\sigma_{y_i}^2 + \sigma_{g_i}^2 + c_2)} \tag{3.10}$$

where μ_{y_i}, μ_{g_i} denotes mean and σ_{y_i}, σ_{g_i} represents standard deviations of y and g respectively while $\sigma_{y_i g_i}$ is their co-variance. c_1 and c_2 are constants to ensure stability if denominator becomes zero. The default value of c_1 = 0.01 and c_2 = 0.03.

The Model-2 is also optimized by using minibatch based gradient descent. Let the batch size be *b* and l be the number of samples. Hence the equation 9 can be represented as:

$$\underset{\theta_{model2}}{argmin}[Loss_{model2}]^b \tag{3.11}$$

where

$$[Loss_{model2}]^b = \frac{1}{l} \times \sum_{i=1}^{l} (1 - \frac{(2\mu_{y_i}\mu_{g_i} + c_1)(2\sigma_{y_i g_i} + c_2)}{(\mu_{y_i}^2 + \mu_{g_i}^2 + c_1)(\sigma_{y_i}^2 + \sigma_{g_i}^2 + c_2)}) \tag{3.12}$$

Algorithm 3.2 shows the training and optimizing for Model-2

Algorithm 3.2 Training and Optimizing the Model-2

Input: Predicted Segmented Map from Model-1 of size 512 X 512

Ground-Truth Segmentation Maps: The set $G = \{g_1, g_2, \ldots\ldots, n\}$ represents ground truth for segmentation map for n images

Parameters: $learning\ rate\ (\eta)$, $Learnable\ parameters\ (\theta_{model2})$, $momentum$.

Initialisation: $Max\_Iter = 500$, $Batch_{Size} = 8, kernel$ regularize value of $L_2 - norm = 0.01$ $counter = 1, \eta = 0.001$, and $patience\ value\ of\ early\_stopping = 5$.

Output: Optimize the Model-2

While *early_stopping* or $counter = Max\_Iter$ is fulfilled, do

For each batch b = *1 to* $\left\lceil \frac{N}{Batch_Size} \right\rceil$do

For each sample i in batch bt do

1. Find the loss of Model-2 i.e., $Loss_{model2}$

end for

2. Obtain the mean of cumulative of loss for the given batch b using Equation-11.

3. Find the gradients of the loss $[Loss_{model2}]^b$ using [104].

4. Obtain gradients, and optimize Equation-10 by updating Model-2's learnable parameters θ_{model2} using Adam[100] optimizer.

End For

5. $counter += 1$

End While

3.3.2.3 Experiments and Results

3.3.2.3.1 Dataset

The presented model is evaluated using the UFBA_UESC[2] dental imaging dataset. There are 1500 dental panoramic X-ray images in this dataset which are divided

into 10 categories of which 1200 images are chosen at random for the training set and 150 images chosen for the validation and test sets, respectively.

3.3.2.3.2 Experimental Setup

The presented model's code is written in Python with Keras API and TensorFlow as backend which is executed in different computing nodes of the Super Computer: The Param Shivay. The, learning rate, batch size and batch normalisation momentum are all set to 0.001, 8 and 0.95, respectively. The suggested model's number of epochs was set at 500. To avoid overfitting, this work used an early stopping approach to discontinue training of suggested model. The early terminating patience parameter is set to 5. The training loss value is monitored by the early stopping approach and if there is no change in loss after five consecutive epochs the training will be terminated.

3.3.2.3.3 Result and Discussion

The dice coefficient, IoU, accuracy, recall and precision of the proposed model is 92.2%, 85.7%, 96.4%, 91.3%, and 92.3%. Also, the number of total parameters is 0.560 million(M). The presented model is compared with the recent state-of-the-art deep learning-based methods as well as with conventional segmentation methods. The comparative analysis is as follows:

3.3.2.3.3.1 Comparative Analysis with recent deep learning methods

The presented model is compared with several state-of-the art deep learning based methods SegNet[103], U-Net[39], BiseNet[101], CEnet[105], Attention U-Net[106], U-Net ++[107] and DeepLabV3+[108] for teeth segmentation on different evaluation metrics and number of parameters shown in Table 3.5 .

Table 3.5 Quantitative result comparison with state-ot-the-art deep learning methods

Models	Accuracy	Precision	Recall	IoU	Dice Coefficient	Parameters (M)
SegNet [103]	96.1	92.1	88.6	82.4	90.3	29.44
U-Net [39]	96.2	94.2	86.8	82.4	90.3	31.04
BiseNet [101]	92.7	92.5	69.6	78.7	79.4	23.06
CEnet [105]	96.3	93.3	90.2	84.7	91.7	38.69
U-Net++ [107]	95.1	92.5	82.6	77.4	87.3	9.20
Attention U-Net [106]	94.0	**96.1**	73.8	71.6	83.4	34.87
DeepLabV3+ [108]	96.2	90.2	87.2	84.2	91.4	60.99
Proposed Model	**96.4**	92.3	**91.3**	**85.7**	**92.2**	**.560**

It can be inferred that the proposed model performs better in comparison to the state-of-the-art methods in terms of accuracy, IoU and dice coefficient which are considered as crucial segmentation performance measurements. The trade-off between precision and recall is well balanced thus resulting in high IoU and dice score. Another major advantage of the proposed model is that it completely outperforms state-of-the-art methodologies in terms of parameter count. The suggested model employs just 0.560 million parameters, which is significantly fewer than methods as given in Table 3.5. The qualitative results of several existing models with proposed model are presented in Figure 3.9. It is observed that BiseNet performs poorly on these images for segmentation task. The visual results CENet and DeeplLabV3+ are good but teeth boundaries are not smooth whereas the proposed method can properly detect teeth segmentation regions while eliminating other elements such as the jaw, nasal and neck bones for simplicity of usage in clinical settings thus indicating that the obtained results for the presented model are closer to ground truth. Also, the tooth boundary predictions are smoother and sharper, with less blurring.

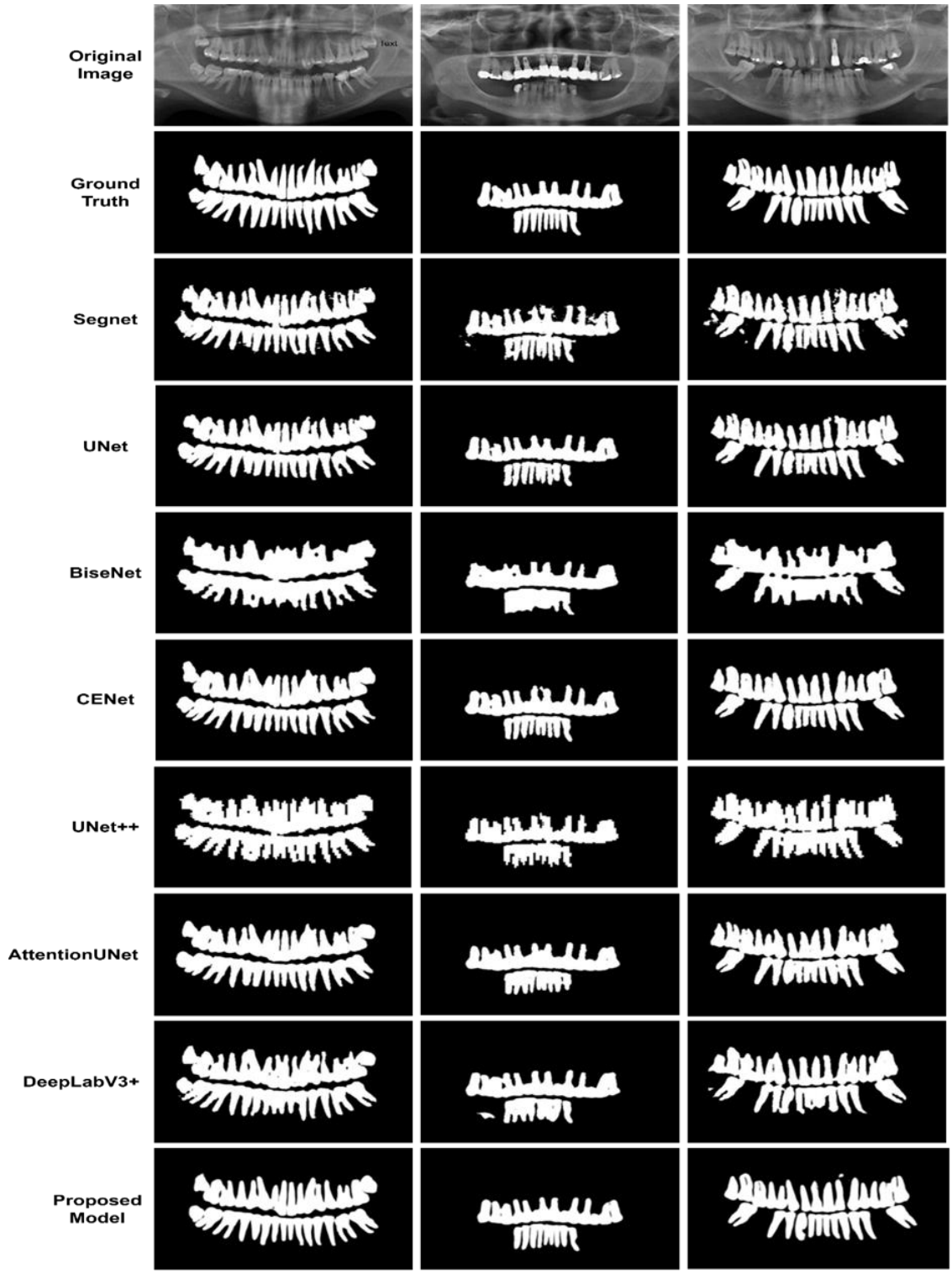


Figure 3 9 Visual results obtained from the proposed model (a) Original Images from the dataset (b) Ground Truth map of the respective image (c) Predicted Segmented Teeth Area by the proposed model.

3.3.2.3.3.2 Comparative Analysis with conventional approaches

The presented model is also compared against traditional methodologies, with the result presented in Table 3.6. Among the conventional approaches, fuzzy c-means approach has the highest accuracy while splitting and merging performs best in the case of precision. Global thresholding methods have high recall whereas the level set method has a high dice coefficient. Overall, the proposed method outperforms all the traditional methods which require pre-knowledge for segmentation in practicality while the presented method can automatically segment the teeth region and if of great use for dentists.

Table 3.6 Quantitative result comparison with conventional methods

Methods	Accuracy	Precision	Recall	Dice Coefficient
FCM [11]	82.1	61.4	45.0	49.4
Level set [31]	75.5	48.3	68.4	59.1
Splitting & Merging [109]	81.3	81.2	8.1	14.8
Region growing [2]	68.1	35.5	63.4	44.1
Global Thresholding [2]	79.2	52.0	69.3	56.2
Proposed Model	**96.4**	**92.3**	**91.3**	**92.2**

3.3.2.3.4 Ablation Studies

The proposed model's ablation analysis is performed to verify the efficacy of the components of model. The similar dataset of 1500 dental images have been utilised, and the metrics evaluated are dice coefficient, IoU, accuracy, recall and precision. Several modules are created based on variations of Model-1, attention block and skip connection but the decoder and Model-2 is the same for all modules except for first module. The 0A represents the module without attention block, 1A denotes that attention block is present, 0S means there is no skip connection, and 1S means with skip connection. The modules are categorized as follows:

- Combinations of Modules with all streams

- M1-M123 – This module has all three streams of different CNN architectures with attention block and skip connection but without second deep cascaded model.
- M2-M123-0A1S – This module also has all three different CNN architectures with skip but without attention block.
- M3-M123-1A0S - This module also includes all three different CNN architectures but with attention block and without skip.
- M4-M123-0A0S - This module has all three CNN architectures but there is no attention block and skip in this module.

- Combinations of Modules based on Conventional CNN and Atrous Net (Di convolutional layers)
  - M5-M12-0A0S – This contains Conventional CNN and Atrous Net in encoder but without attention block and skip connections.
  - M6-M12-1A1S – This module also has both streams along with attention block and skip connection.
- Combinations of Modules based on Conventional CNN and Separable CNN
  - M7-M13-0A0S – This module contains two streams in encoder but without attention block and skip connections.
  - M8-M13-1A1S – This module comprises of Conventional CNN and Separable CNN along with attention block and skip connection.
- Combinations of Modules based on Atrous Net and Separable CNN
  - M9-M23-0A0S – This module has two streams Atrous Net and separable CNN but without attention block and skip connections.
  - M10-M23-1A1S – This module is a combination of Atrous Net and separable CNN with attention block and skip connection.

- Combinations of Modules based only on Conventional CNN
    - M11-M1-0A0S – This module has only one stream i.e. Conventional CNN in encoder but without attention block and skip connections.
    - M12-M1-1A1S – This module also has only one stream Conventional CNN with attention block and skip connection.
- Combinations of Modules based only on Conventional CNN
    - M13-M2-0A0S – This module has only on Atrous Net. without attention block and skip connections.
    - M14-M2-1A1S – This module also has only one stream of Atrous Net but with attention block and skip connection.
- Combinations of Modules based only on Conventional CNN
    - M15-M3-0A0S – This module has only separable CNN without attention block and skip connections.
    - M16-M3-1A1S – This module also has only Separable CNN but with attention block and skip connection.

Table 3.7 Quantitative results of ablation study of different modules.

Modules	Accuracy	Precision	Recall	IoU	Dice Coefficient
M1-M123	96.3	91.2	92.1	78.7	88.0
M2-M123-0A1S	94.6	93.8	80.7	68.0	80.9
M3-M123-1A0S	94.6	92.4	82.1	64.9	78.7
M4-M123-0A0S	95.9	94.4	87.4	75.7	86.2
M5-M12-0A0S	94.5	94.9	79.3	52.3	68.7
M6-M12-1A1S	94.0	85.4	86.9	59.0	74.2
M7-M13-0A0S	92.7	90.2	74.5	55.9	71.5
M8-M13-1A1S	94.2	91.0	81.3	59.8	74.8
M9-M23-0A0S	93.6	92.1	77.5	63.3	76.8
M10-M23-1A1S	95.0	89.9	86.9	70.5	82.6
M11-M1-0A0S	90.2	91.4	59.3	43.1	60.2
M12-M1-1A1S	91.7	91.5	67.7	54.2	70.2
M13-M2-0A0S	90.9	94.9	60.4	52.4	64.3
M14-M2-1A1S	91.3	72.8	93.8	60.6	75.4

M15-M3-0A0S	92.2	89.8	72.4	53.1	66.5
M16-M3-1A1S	93.5	89.5	79.2	62.2	76.7
Proposed Model	**96.4**	**92.3**	**91.3**	**85.7**	**92.2**

The quantitative outcomes of each module are demonstrated in Table 3.7. Some modules have high precision but low recall and vice versa demonstrating that these modules have focused more either on foreground pixels or background. The modules with both attention block and skip have performed better in comparison with those without attention and skip which proves the efficacy of attention block. Module- M1- M123 has the highest accuracy and has the balanced precision and recall also its dice score and IoU is better among all the modules but in the qualitative results it can be seen that this module has corrupted pixels and unclear boundaries. The proposed model performs better than each individual modules and produce sharp and clear boundaries. The visual results of the ablation performed are demonstrated in Fig. 3.10.

3.3.2.4 Summary

In this section, a novel deep cascaded network with teeth boundary attention block was proposed to segment the teeth region automatically from the dental panoramic images. The proposed architecture comprised two models: an attention-guided encoder & decoder network and a deep convolutional architecture. Firstly, an encoder-decoder model was proposed to extract multiple CNN features and fine-grained contextual information from dental radiographs and the attention mechanism was designed to obtain teeth boundary information. Thereafter a stack of simple convolutional layers was used to enhance the predicted segmented map by removing the blurry and noisy pixels. A novel proposed methodology was evaluated in terms of qualitative and quantitative analysis on 1500 panoramic radiographs. The presented methodology was compared with other

existing methodologies in terms accuracy, precision, recall, IoU and dice score. From the obtained results, it showed that the proposed methodology is segmenting the proper region of teeth, preserving edges, shape information and removing blur pixels in a single frame work.

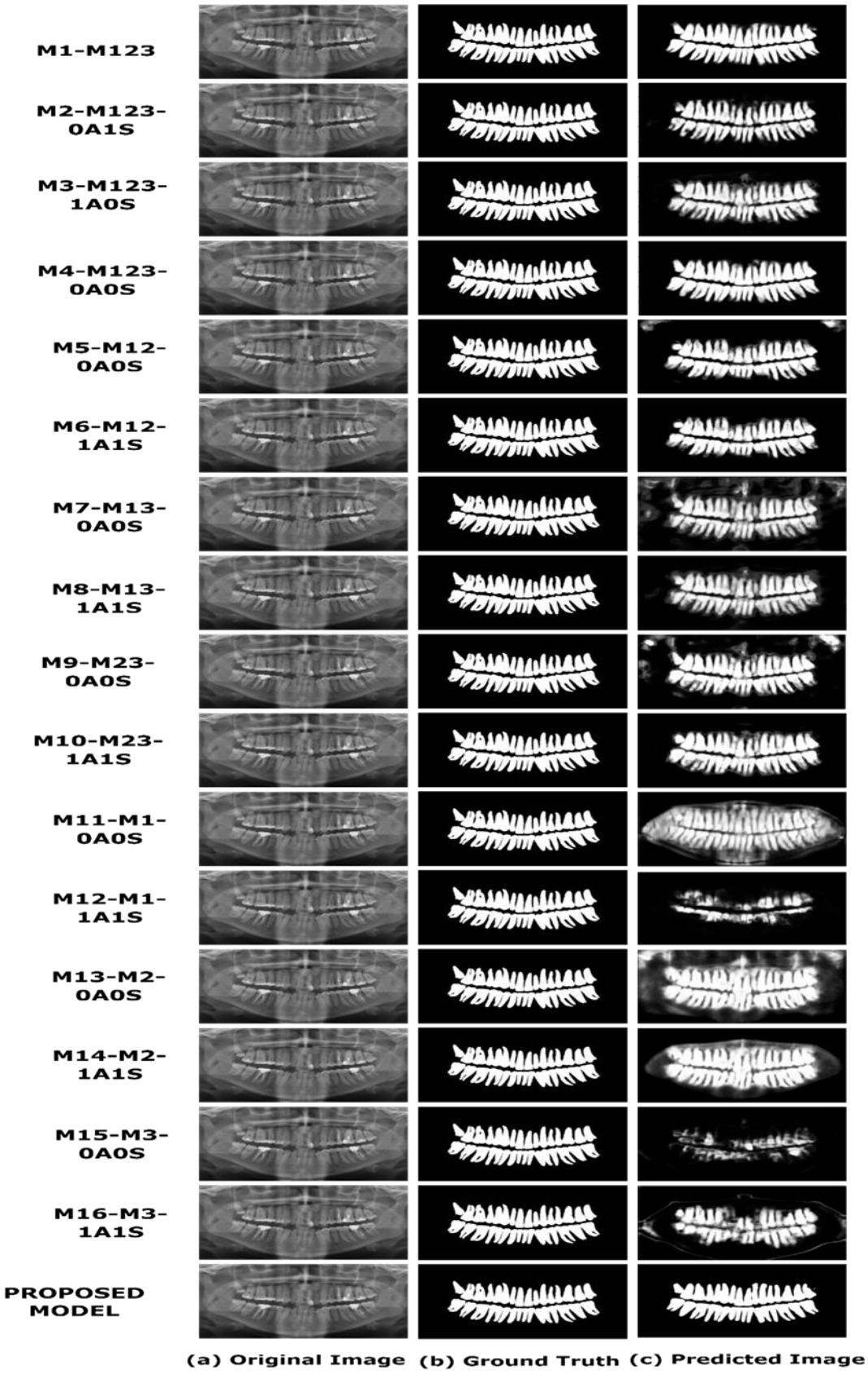


Figure 3.10 Visual results of the ablation study of different modules (a) Original Image of Dataset (b) Ground truth of respective image (c) Predicted segmented image by each module.

3.4 Conclusion

This chapter proposed two deep learning methodologies for segmenting teeth from dental panoramic X-ray images. The first method improves the segmentation performance by extracting multiple features and combining them to gather rich dental information. The model focused on the background pixels thus affecting the overall result. In the second deep methodology the teeth contextual information is gathered. Also, the attention mechanism is added to preserve boundary information. The model segmented proper teeth region while maintain edge and shape information and eliminated the blurry pixels. The limitation of this approach is that edge groundtruth are obtained separately thus limits the model to a particular task. The future work is to generate the edge groundtruth within a framework. Both models were tested on publicly available UFBA_UESC dental dataset. The first model got similar results to the state-of-the-art methods with less number of parameters whereas second model outperforms the state-of-the-art methods. An extensive ablation study was conducted to show the efficacy of different modules.