CHAPTER 4

ENHANCING TEETH SEGMENTATION USING MULTIFUSION DEEP NEURAL NET IN PANORAMIC X-RAYS

This chapter presents a novel multifusion deep neural network for enhancing teeth segmentation in panoramic X-ray images to provide accurate teeth region. The model extracts texture rich semantic information of teeth region by multistage fusion. The dual type skip is designed to minimize the semantic information gaps to get clear tooth boundaries. Also, the model's performance is enhanced by introducing the hybrid loss. The model is tested on two benchmark dental datasets UFBA_UESC dataset and Tufts dataset. In addition, an extensive ablation study is performed to show the effectiveness of proposed model.

4.1 Background

In the oral field, analysing dental images is crucial as it provides information about the patient's oral conditions and helps in drawing better treatment plans. Teeth segmentation is one of the prominent tasks for analysing tooth status in dental imaging[92]. The segmented dental images convey essential information for monitoring teeth development, detecting dental structures and location of teeth for teeth numbering and implant planning[43]. Another application of teeth segmentation is in the field of forensics to identify personal details like age and gender based on dental features[110],[111].

Among many dental imaging modalities, panoramic X-ray images are mostly utilized for diagnosing dental diseases. Dentists examine these X-ray images by visualizing them to determine the presence of various dental problems like cavities, fractured teeth, gum diseases, bone abnormalities and oral cancer[36],[5]. Analysing panoramic X-ray images manually is difficult and time-consuming, requiring expertise and experience. Even for experts examining X-ray images is more challenging because such images suffer from low contrast at boundaries, the noise produced by machines, overlapped images of teeth etc. thus may lead to incorrect judgements. Also, the presence of jaw and nasal bones makes this task difficult for dentists. Thus, to summarize, there are several issues with panoramic images such as poor contrast, blurring borders of teeth and the presence of jaw bones and other mouth elements. Thus, examining such images manually is challenging and requires experience and time Hence, a precise and automated segmentation technique is required.

4.2 Related Work

With the introduction of deep learning in medical image analysis field, the limitations of traditional approaches for dental image segmentation can be greatly enhanced. The deep learning approaches have an advantage over traditional methods as these approaches has ability to learn features straight from unprocessed data instead of using hand-designed features. Several image segmentation models have been proposed are based on convolutional neural networks of which U-Net[39] and SegNet[103] have demonstrated impressive performance in the field of medical image segmentation, including dentistry.

Yang et al.[112] utilized conventional CNN as the backbone and proposed a pipeline for automatic analysis to diagnose tooth diseases using a small dataset having 196 periapical images. A couple-shaped model with the deep neural net was presented by

Wirtz et al.[113] for automatic and robust segmentation of each tooth on low-quality dental panoramic radiographs. Due to insufficient data and shape inconsistencies, their approach results in low segmentation accuracy. To show the efficacy of transfer learning in segmenting dental radiographs Caylak et al.[49] employed the pre-trained Inception-ResNet-v2 and pre-trained U-Net models on small datasets. Only 131 dental panoramic radiographs were used. Nishitani et al.[32] presented a teeth segmentation model using U-Net on 162 panoramic images. This method focuses on the hybrid loss function to train U-Net architecture for tooth edges. However, the drawback of this method is suboptimal hyperparameters tuning and the need for optimization of edge width.

Silva et al[2]. proposed the standard dental dataset of panoramic X-ray images which consists of 1500 images. The authors implemented some traditional methods to segment dental images. Additionally, the Mask-RCNN deep neural network was deployed on this dataset in different categories of this dataset. The results obtained were not up to the mark thus leaving scope for other neural networks to be evaluated on this dataset. Jader et al[96]. used mask regional convolutional networks with ResNet-101 as a backbone for instance segmentation on 1500 panoramic radiographs. The main objective is to detect individual teeth or missing teeth. The limitation of this approach is that it lacks in capability for segmenting mouth and teeth components due to the presence of tooth appliances. Oktay et al.[114] and Pinheiro et al.[115] also used the Mask RCNN for segmenting individual teeth for tooth numbering in panoramic X-ray images. These approaches suffer because of the presence of dental implants and the overlapping of teeth. These models are utilized for instance segmentation in which the focus is on segmenting individual teeth which is a complex and laborious task when compared to semantic segmentation as it requires intensive and careful labelling. In this study, the focus is on semantic segmentation.

Modified U-Net was presented by Koch et al.[98] for accurate tooth segmentation. The authors made patches of original images and combines several techniques to improve segmentation performance. The annotated and patches of images were used which hampers model's performance. Zhao et al.[54] presented a two-stage attention model based on CNN for accurate tooth segmentation. To locate the tooth area both global and local features were extracted using attention modules in first stage. A fully convolutional network was employed to segment dental region in second stage. The model has approximately 78 million trainable parameters and was evaluated on 1500 panoramic radiographs. The limitation of this method is in handling high number of trainable parameters.

Cui et al.[57] investigated the generative adversarial network with certain conditions on dental dataset comprises 1500 panoramic images for tooth segmentation. This method is highly dependent on annotation which compromises genuine feature extraction. Lin et al.[37] presented a lightweight neural network strategy using knowledge distillation for segmenting dental radiographs to deploy on edge devices. 1321 resized dental X-ray images were utilized to measure the performance. Model designed was lightweight but needs significant improvement in terms of dice and IoU score.

Panetta et al.[59] developed a new benchmarking multimodal dental database consisting of 1000 panoramic X-ray images. Authors implemented recent state-of-the-art segmentation techniques with image enhancement methodologies to evaluate the dataset. The deep models consisting of Atrous Net are yet to be explored on this dataset.

Vision-based automatic dental segmentation techniques employing both traditional and deep learning models have been presented in the literature. Deep learning techniques performed better as compared to traditional methods. These techniques are mainly based on conventional CNN, Mask RCNN and modified U-Net architecture for

dental image segmentation. Thus, leaving scope to explore different types of CNN specially Diconvolutional (Atrous Net) layers for accurate and automatic dental panoramic X-ray image segmentation. Models based on Mask RCNN are generally used for instance segmentation while in this study, the focus is on the semantic segmentation of dental panoramic X-ray images which helps provide more clarity of tooth structure for better analysis. The deep models especially based on the attention mechanism possess a huge number of trainable parameters and takes longer training and testing time for model since there is a trade-off between accuracy and trainable model parameters. A lightweight deep model was proposed but it suffers with performance on dice score and IoU score which are important evaluation criteria for segmentation.

Thus, to address above mentioned issues, in this chapter a multifusion deep neural network with less trainable parameters is presented for precise and automatic segmentation of teeth region from dental panoramic X-rays. The details of the proposed model are discussed in the subsequent section.

4.3 Proposed Methods

4.3.1 Overview

The proposed multistage fusion deep model primarily concentrates on the automatically segmenting the entire teeth region from dental panoramic radiographs. The proposed model is based on encoder-decoder architecture. The encoder module is responsible for encoding texture semantic information of dental region by fusing information gathered from two different CNN-based streams i.e., conventional CNN and Atrous (Diconvolutional) Net at each step. Two types of skip connections Long Skip and Short skips are designed to preserve the low-level features of teeth which help in the

reconstruction of the segmented teeth region map at the decoder module. The training of the proposed multistage fusion deep net is done in an end-to-end manner.

4.3.2 Detailed Architecture of Proposed Network

The major components of the proposed network are convolutional-dilated convolutional based encoder, deconvolutional decoder and dual type of skip connections. The proposed model is trained in an end-to-end way using hybrid loss function. The detailed architecture of the proposed multifusion deep model is shown Figure 4.1.

Encoder: The encoder part of the presented model is composed of two different CNN-based streams. Each stream of the encoder receives the same input in the form of original panoramic X-ray image. The first stream uses simple convolutional layers to capture essential dental information from panoramic radiographs. This stream comprises of six convolution layers having different kernel sizes followed by batch normalization and uses an activation function ReLU. In this stream three Max pooling layers of dimensions [2 x 2] are used to minimise the spatial dimensionality of image. Second stream of the encoder part contains six dilated-convolutional (Diconvolution) layers having different dilation rates. These layers also use activation function ReLU with batch normalization. Here also, three Max pooling layers of dimensions [2 x 2] are used to downsample feature maps. Dilated convolution has the advantage of capturing contextual semantic information of teeth region in bigger receptive field. Additionally, it suppresses non-relevant background region without increase in the number of network parameters. The skip connections are also designed in the encoder module. Short-type skip connections are utilized in both streams of encoder that is in simple CNN stream as well as dilated convolution (Atrous) net stream to enhance the training of proposed deep network. Long-type skip connection are present from an encoder part to the decoder part

of the network. To exploit the features of multi-CNN, feature maps obtained from each convolution layer of the CNN stream and dilated convolutional layer of the Atrous Net stream are fused. These fused feature maps are provided as input to another stream called fused stream which includes three simple convolutional layers having filter size [3x3] with ReLU as an activation function followed by batch normalization layer. The feature maps obtained from all streams are fused at each stage while maintaining symmetry in the encoder. This enables a network to learn more context from feature maps of dental panoramic images. Finally, all the features of each stream are fused and are passed as input to the decoder.

Decoder: The features extracted by the encoder are in compressed form and are the representation of original dental panoramic X-ray images. These compressed features are then fed to the decoder to reconstruct the final segmented teeth image. The decoder accumulates the feature information from various layers and concatenates encoder-extracted features with upsampled feature maps. The symmetry is maintained throughout the encoder and the decoder module. The decoder includes eight deconvolution layers with different kernel sizes. The first seven layers utilizes activation function ReLU and batch normalization. The upsampling layers of dimensions [2x2] are added to increase the spatial dimensions which help in reconstructing the original size of the image. The final layer in the decoder is a deconvolution layer of channel 1 and has a [1x1] kernel size and uses the sigmoid activation function for pixel-wise classification. The training of proposed model is completed in an end-to-end way.

Skip Connections: There are two types of skip connections presented in the proposed deep network. The first are the Short-type skip connections and the second are Long-type skip connections. The objective of these skip connections is to enhance information flow between feature maps. The short-type skip connections are proposed in

the encoder module while the Long-type skip connections are between the encoder module and decoder module. The short-type skip connections are utilized in conventional CNN and Atrous Net streams to reuse the dental semantic features obtained from earlier layers in the other layers. Additionally, they help in faster convergence of the learning process by

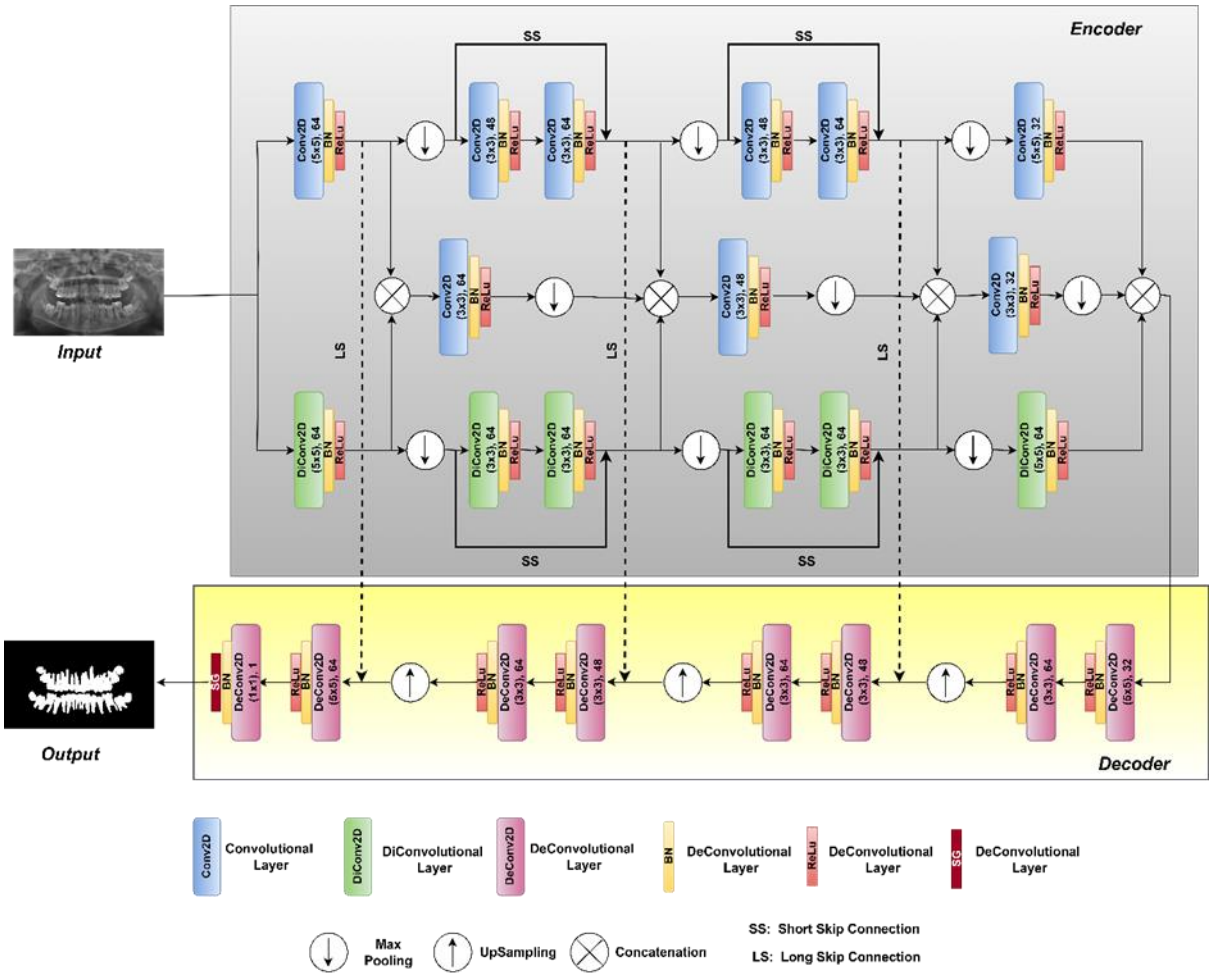


Figure 4.1 The detailed architecture of the proposed model.

stabilizing gradient updates and well distribution of parameter updates. Downsampling in the encoder part results in a loss of spatial structural information due to different types and shapes of teeth. Thus, to preserve low-level spatial features Long-type skip

connections are introduced which originate from an encoder and meet the decoder to enhance the restructuring of the segmented teeth map.

Hybrid Loss Function: Two different loss functions are employed, in order to optimize the proposed model. Binary cross-entropy (BCE) loss is the first loss function, which is based on distribution loss and is commonly utilized in binary classification and segmentation. BCE is expressed as:

$$Loss_{BCE} = -(Gt_i(\log Pr_i + (1 - Gt_i)\log(1 - Pr_i)) \quad (4.1)$$

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.

The second loss function is a dice loss inspired by the dice coefficient score is based on region loss and is utilized to refine the overlapping between the two images. This loss is mostly utilized in optimizing medical images. It is described as follows:

$$Loss_{DICE} = 1 - \frac{2Pr_i\, Gt_i}{Pr_i + Gt_i} \quad (4.2)$$

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.

In order to obtain clear teeth segmented maps a hybrid loss function is presented and is defined as:

$$Loss_{MODEL} = \lambda\, .Loss_{BCE} + (1 - \lambda).Loss_{DICE} \quad (4.3)$$

where $Loss_{BCE}$ is binary cross entropy loss, mostly employed by deep learning-based techniques to classify pixels or data into two categories. $Loss_{DICE}$ is dice loss which measures the quality of an image. λ is a positive constant range between 0 and 1. The value of λ is empirically set to 0.8.

The BCE loss is considered as a pixel-by-pixel loss that is utilized to minimise the gap between each pixel. The dice loss is considered as a loss of foreground and becomes zero

as the prediction of foreground increases assisting the proposed network to emphasize on the foreground. Both loss functions have different benefits thus utilizing them jointly optimizes the efficacy of the proposed network.

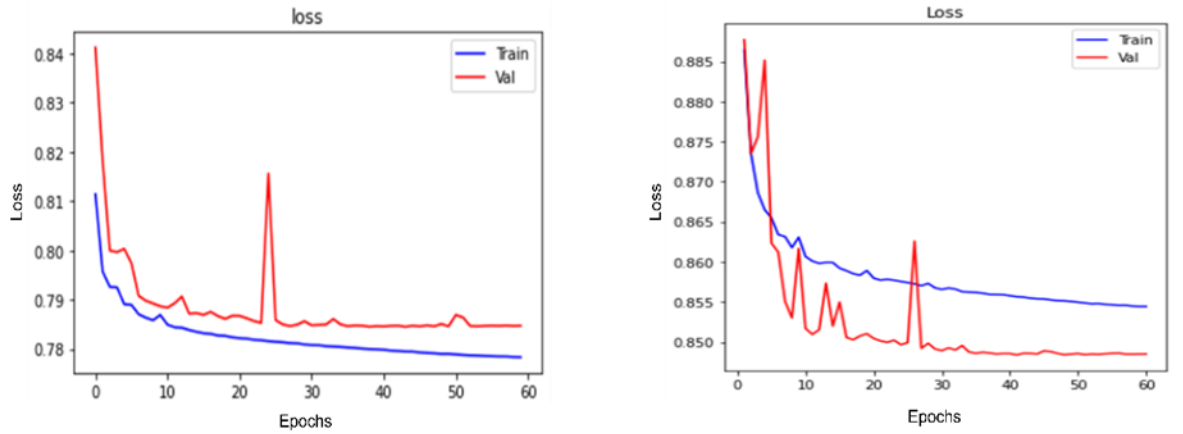
4.4 Experiments and Results

4.4.1 Datasets

The two-benchmark dataset are used to assess the proposed deep model. The first dataset is UFBA_UESC[2] dental dataset. For training, validation and testing, dataset 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 set. The second benchmark dataset[59] used to assess the suggested model is called Tufts dental database. This dataset is also segregated in 3 parts with similar split ratio having 800 randomly selected images in training set while 100 image each in test and validation set.

4.4.2 Experimental Setup

The presented deep model is implemented in an end-to-end way on four NVIDIA GeForce GTX 1080i GPUs using Pytorch framework in python. The learning rate, batch size and the number of epochs is set to 0.0001, 16 and 200 respectively. To avoid overfitting, the early stopping method is adopted with the patience parameter set to 4 that is training will be halted if there will be no change in training loss for 4 consecutive epochs. It can be inferred that the proposed network converges at approximately 60 epochs. The training and validation loss for every epoch is plotted and is shown in Figure 4.2. Further, the hybrid loss function is optimized by using RMSprop optimizer.



(a) Training Vs Validation Loss of Proposed Model for Dataset 1

(b) Training Vs Validation Loss of Proposed Model for Dataset 2

Figure 4. 2 Losses of proposed model for both datasets. (a) The plot demonstrates the training loss and validation loss for dataset 1 it seems the model converges approximately at 55th epoch. (b) This plot demonstrates training loss and validation loss for dataset 2 which seems to be converged at epoch 60.

4.4.3 Result and Discussion

The proposed model's performance is compared with recent existing deep learning models all of which can be utilized in segmenting dental panoramic X-ray images. These deep models are SegNet[103], U-Net[39], BiseNet[101], CENet[105], U-Net++[107] and Nanonet[4] which were applied to the same benchmark dental image datasets. To have a fair comparison, the same parameter setting has been used and datasets are split in the same ratio for training, testing and validation set. The assessment of the models is done on five different metrics, number of parameters and two different panoramic radiographs datasets. For dataset 1, accuracy, precision, recall, IoU and dice score of the proposed model are 97%, 96.4%, 90.6, 91.1% and 92.4% respectively with only 0.380 million (M) parameters. On dataset 2 the proposed model attains 97.7% accuracy, 95.4 % precision, 92.8% recall, 90.2% IoU and 90.7% dice coefficient while the number of parameters is same.

Table 4.1 Performance summary of the proposed model compared with deep segment state-of-the-art methods for Dataset 1

	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]	96.7	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
NanoNet [4]	96.6	95.0	82.8	89.9	91.3	**.235**
Proposed Model	**97.0**	**96.4**	**90.6**	**91.1**	**92.4**	.380

Table 4.2: Performance summary of the proposed model compared with deep segment state-of-the-art methods for Dataset 2.

	Accuracy	Precision	Recall	IoU	Dice Coefficient	Parameters (M)
SegNet[103]	96.4	88.5	92.1	88.2	89.3	29.44
U-Net [39]	96.5	89.7	91.8	88.3	88.5	31.04
BiseNet [101]	96.7	83.4	90.9	86.6	82.0	23.06
CENet [105]	96.7	93.3	90.2	84.7	91.7	38.69
U-Net++ [107]	97.1	86.2	90.3	78.9	88.2	9.20
NanoNet [4]	97.4	89.7	91.2	90.0	90.4	**.235**
Proposed Model	**97.7**	**95.4**	**92.8**	**90.2**	**90.7**	.380

The quantitative comparison results with other deep models for dataset 1 is represented in Table 4.1 and for dataset 2 is represented in Table 4.2. It is observed that the presented deep model completely outperforms the state-of-the-art methods on all evaluation metrics except the number of parameters for both datasets. The presented model is better than all the deep segment state-of-the-art methods in terms of parameters except Nanonet which has only 0.235 trainable parameters. Though Nanonet has shown good performance while having less number of parameters its dice coefficient and IoU score are quite less compared to the proposed model which are considered important segmentation metrics.

The visual results of the proposed model are demonstrated in Figure 4.3 for dataset 1 and Figure 4.4 for dataset 2. It can be observed that U-Net, SegNet, BiseNet, U-Net++ and CEnet are prone to misclassify the correct teeth pixels with jaws bone and other facial bone pixels whereas the proposed method correctly identifies segmented teeth region. It can be seen that BiseNet is comparatively weak in dealing with low-contrast images. SegNet, CEnet, and U-Net++ suffer in predicting the tooth boundaries while NanoNet seems to perform better in detecting tooth boundaries but shows the teeth structure slightly thicker while the proposed method has predicted the clear and smooth teeth region segmented map by suppressing the other background details of mouth.

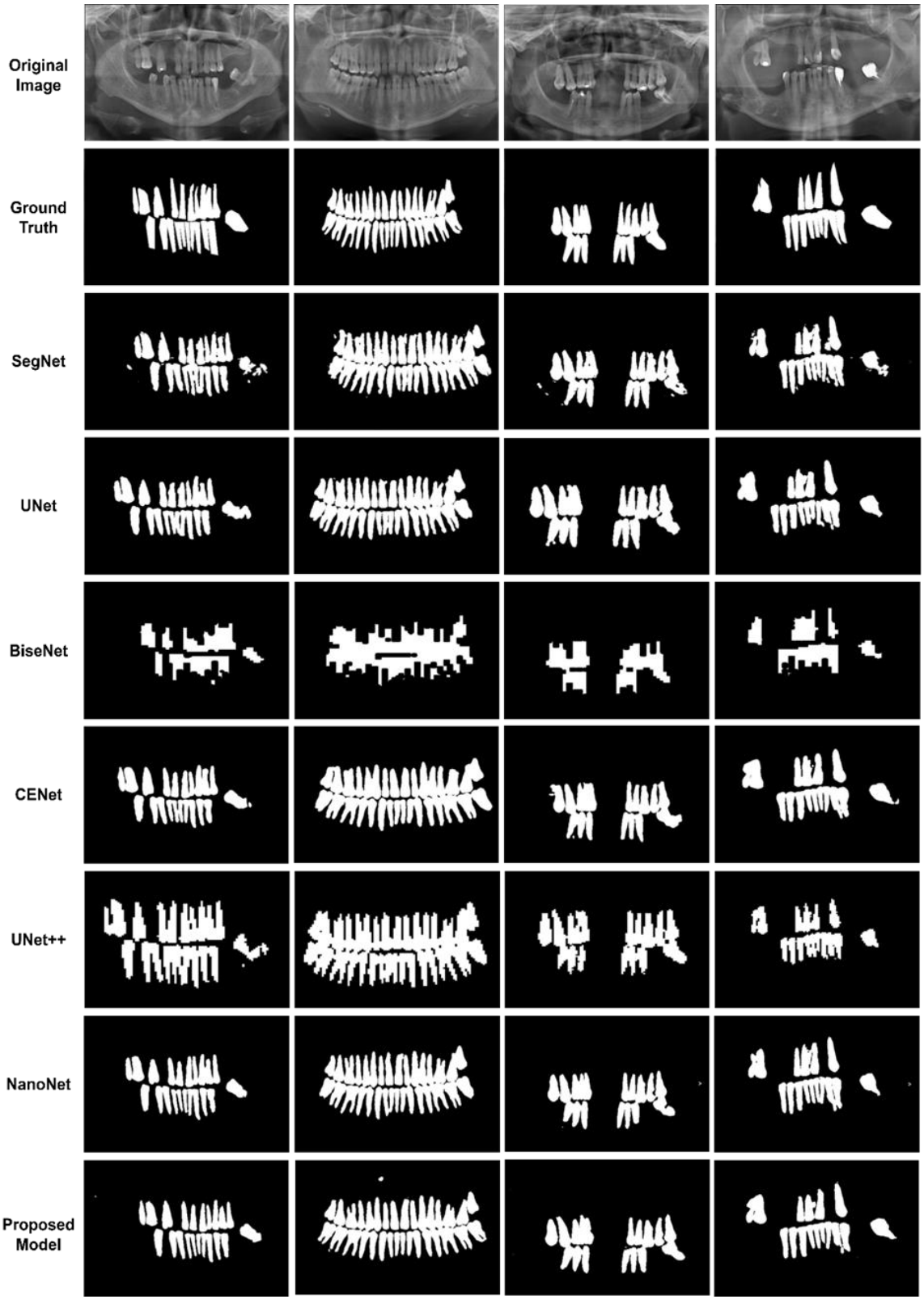


Figure 4.3 Comparison of visual results of proposed model with different segmentation based deep models on Dataset 1. The first two rows represent original image and its corresponding ground truth while the remaining rows represents the predicted segmented teeth map by different methods including the proposed model.

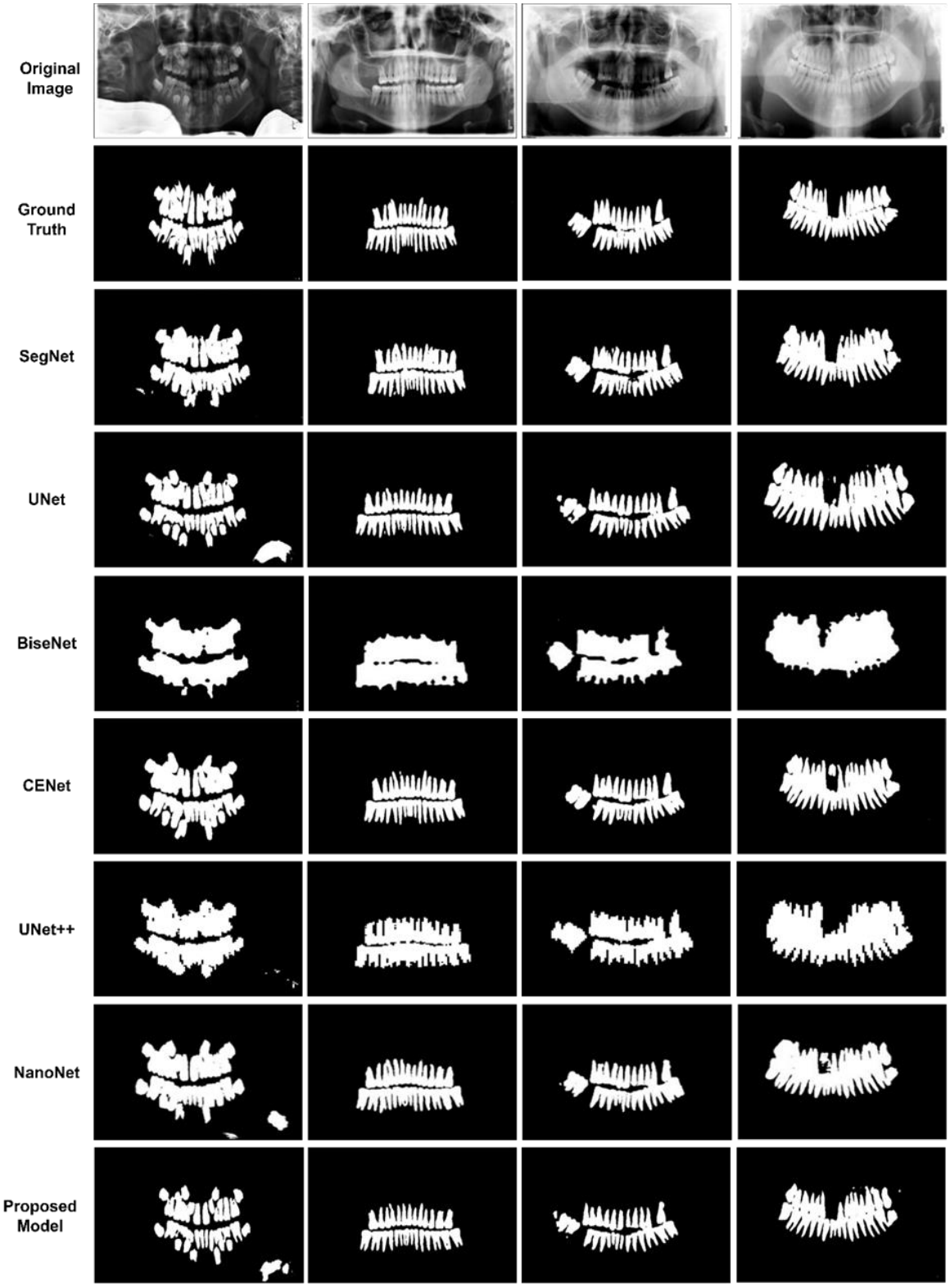


Figure 4.4 Comparison of visual results of proposed model with different segmentation based deep models on Dataset 2. The first two rows represent original image and its corresponding ground truth respectively while rest shows segmented teeth region mask predicted by various deep models.

4.4.4 Ablation Study

The ablation study is carried out to verify the efficacy of the components of the suggested deep model. The experiments are conducted with similar parameter settings but with different components on the same datasets. The taring and validation loss of each module is shown in Figure 4.5. and Figure 4.6 for dataset1 and dataset 2 respectively from which one can infer that all the modules are converged at epoch 60. The assessment metrics used are precision, accuracy, recall, IoU and dice coefficient. Three different modules are developed from the proposed model based on skip connections and are as follows:

- M1-WS: Module1-Without Skip connections. The first module is composed of the encoder and decoder part of the presented deep network but there are no skip connections present, neither in an encoder nor from an encoder to decoder.
- M2-LS: Module2-Long Skip. This is the second module which consists of only long skip connections from an encoder to the decoder.
- M3-SS: Module3-Short Skip. The third module contains only short skip connections which are present in both conventional CNN stream and Atrous Net stream of the encoder module.

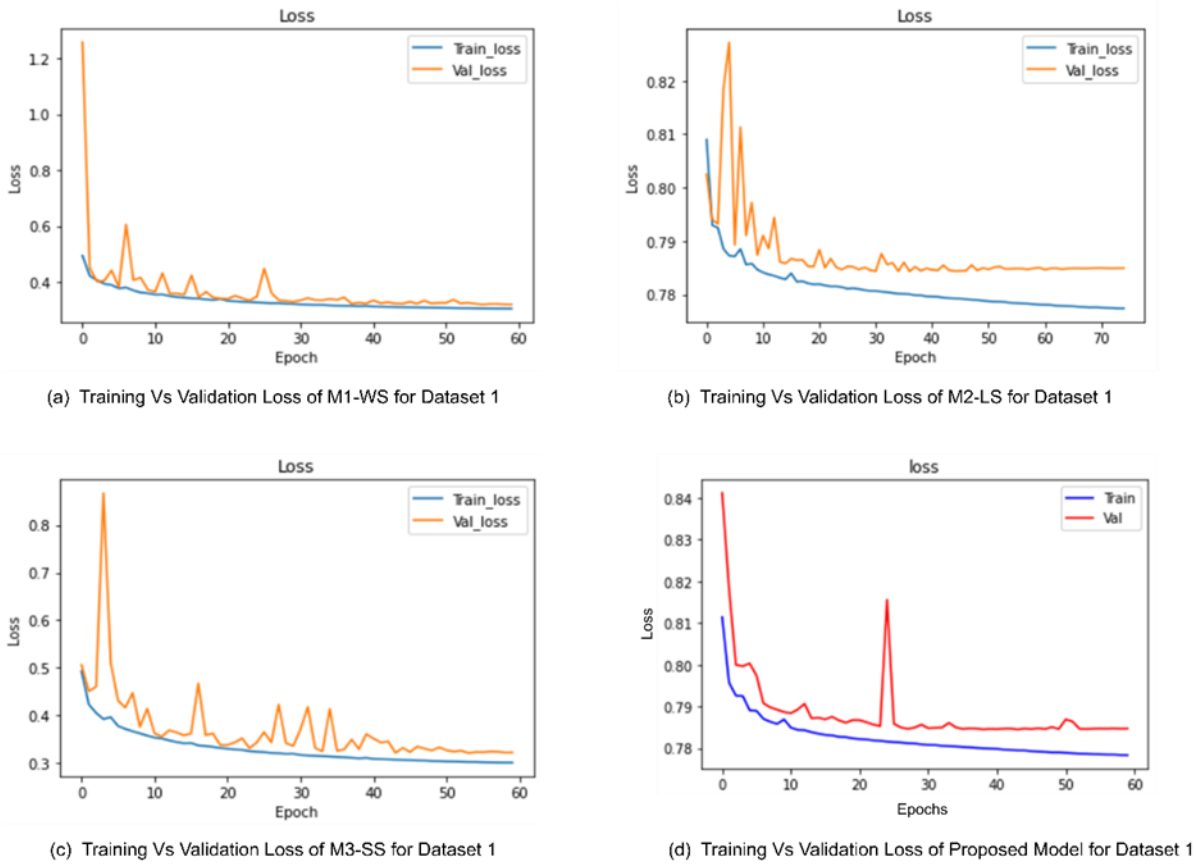


Figure 4.5 Training vs Validation loss for Dataset 1. (a) This plot represents the training and validation loss for module without skip connections. (b) The loss for module 2 having long skip connections. (c) This plot shows loss for the third module having only short skip connections. (d) This represents the training and validation loss for proposed model. All modules including proposed model is converged at epoch 60.

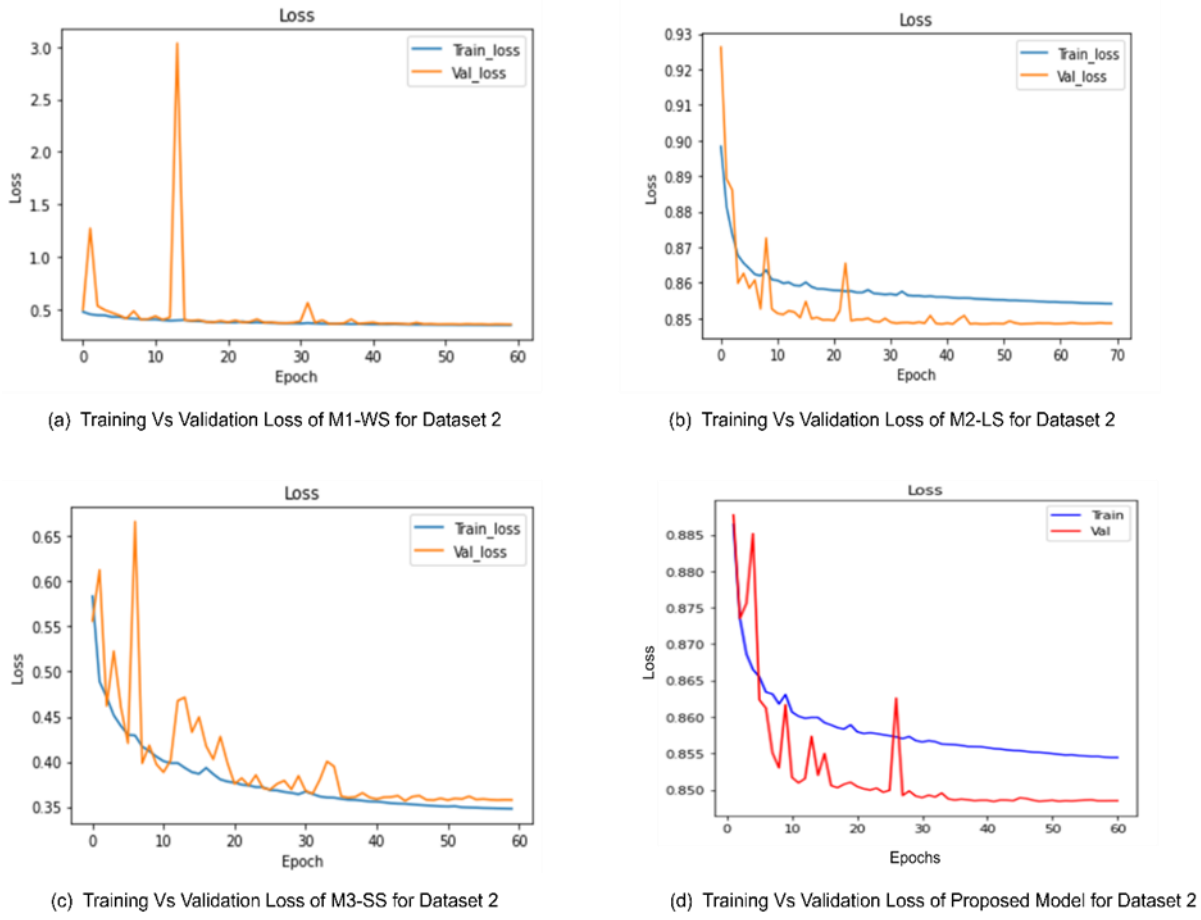


Figure 4.6 Training vs Validation loss for Dataset 2. (a) This plot represents the loss for first module with no skip connections. (b) The training and validation loss for second module having long skip connections is shown. (c) This plot shows loss for the third module having only short skip connections. (d) This represents the training and validation loss for proposed model. All modules including proposed model is converged at epoch 60.

Table 4 3 Ablation experiment comparison of different modules on Dataset 1.

Modules	**Accuracy**	**Precision**	**Recall**	**IoU**	**Dice Coefficient**
M1-WS	96.4	94.4	85.5	89.2	90.1
M2-LS	96.7	88.6	96.0	85.5	92.2
M3-SS	96.4	96.3	85.4	89.1	90.5
Proposed Model	**97.0**	**96.4**	**90.6**	**91.1**	**92.4**

Table 4.4 Ablation experiment comparison of different modules on Dataset 2.

	Accuracy	**Precision**	**Recall**	**IoU**	**Dice Coefficient**
M1-WS	97.3	88.8	81.5	87.7	87.9
M2-LS	97.7	88.9	92.4	82.8	90.6
M3-SS	97.4	94.9	83.1	88.6	88.3
Proposed Model	**97.7**	**95.4**	**92.8**	**90.2**	**90.7**

The effectiveness of each model is represented by the quantitative results are shown in Table 4.3 and Table 4.4 for dataset 1 and dataset 2 as well as by the qualitative results shown in Figure 4.7 for dataset 1 and Figure 4.8 for dataset 2. It can be inferred that without any type of skip connections, the dice score and IoU are less as compared to a network having only long skip connections. Modules without skip and with short skip connections have similar results except for precision which is high for networks with only short-type skip connection. The long-type skip connection module has the highest recall. The proposed model which includes both short-type skip connection and long-type skip connection has performed superior than each individual module in all evaluation metrics. The visual results demonstrate the proposed model's segmented teeth map has better visibility than other modules.

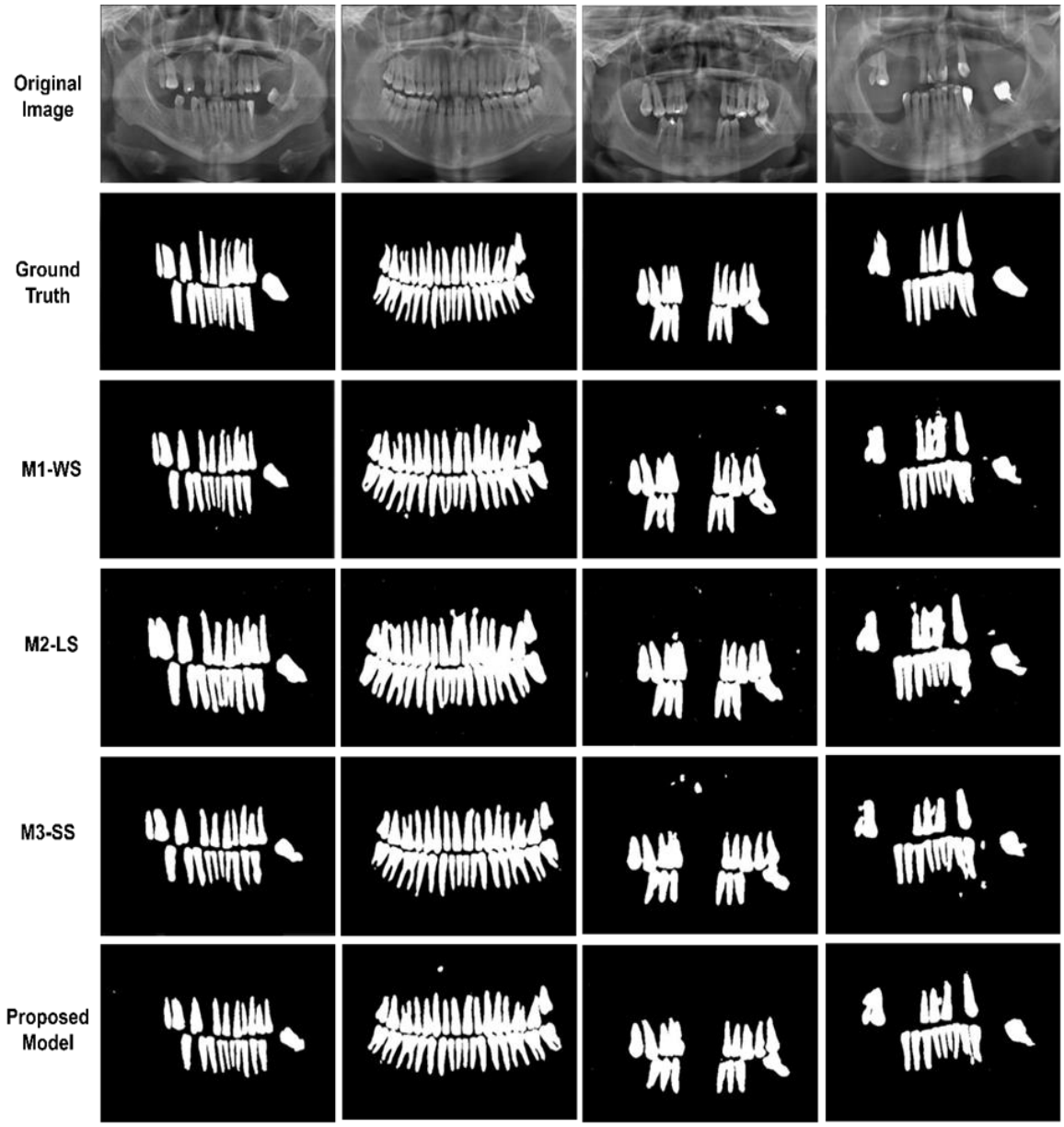


Figure 4.7 Comparison of visual results from ablation experiment on Dataset 1. The first two rows represent original image and its ground truth respectively. The next three rows demonstrate the predicted segmented map by different modules while the last row represents the predicted segmented map by proposed model.

4.5 Conclusion

This chapter proposed a novel multifusion deep model based on encoder-decoder structure in order to obtain clear and accurate teeth region segmented from dental panoramic radiographs. The encoder of proposed model comprised of two different CNN

based architectures with fusion of features at each step for extracting diverse features and contextual rich information. The two types of skip connections were proposed in encoder and from encoder to decoder to regain dental features using different connections. The decoder is made-up of deconvolutional layers to restructure the image. The experiments were executed to demonstrate efficacy of the presented model on two different datasets. The results obtained shows that the presented model out performs the state-of-the-art deep segmentation methods with requiring relatively a smaller number of trainable parameters. The proposed model also dealt with limitations of model proposed in chapter 3.