CHAPTER 6

INVESTIGATION OF TRANSFORMER-BASED NEURAL NETWORK FOR DENTAL IMAGE SEGMENTATION

This chapter presents a novel deep neural network based on transformers for precise teeth segmentation. The encoder-decoder architecture is fused with transformer structure to manage spatial information in dental panoramic X-rays. The self-attention mechanism of transformer is utilized to capture long-range information. The model utilizes the skip connections to maintain the characteristics of original image, guaranteeing sufficient image reconstruction. Variants of encoder-decoder architecture using transformer were also investigated. The model is validated on two benchmark dental datasets UFBA_UESC dataset and Tufts dataset.

6.1 Background

The transformer was initially developed in natural language processing by Vaswani et al.[117], comprising Multi-Head Attention (MHA) and Position-wise Feed-Forward Networks. The transformers solely rely on a self-attention mechanism. This inspired the researcher to utilize the transformer for computer vision tasks. Several modifications have been made to transformers to make them suitable for computer vision jobs. Dosovitskiy et al.[118] developed a vision transformer (ViT) for image recognition tasks, which shows remarkable performance in image classification. In ViT, the images were divided into equal-sized patches as in transformers, one-dimensional sequences are only passed. After that, these patches were flattened to a two-dimensional patch sequence.

The patch embeddings are obtained after the linear projection, before passing them to encoder position embeddings are added to it. The Multi-layer Perceptron (MLP) is added into the transformer encoder while keeping the MSA transformer intact. At last, the MLP Head module produces the image classification. After the success in classification task, the ViT is utilized by the researchers for the segmentation task. The transformers were introduced to segmentation tasks to gather long-range dependency and global context information to avoid the loss of basic features and local information.

6.2 Related Work

After showing its potential in image processing, the transformer has motivated medical researchers to explore its capability in segmenting medical images. Chen et al.[119] applied transformer blocks to U-Net architecture for the first time for medical image segmentation and proposed TransUnet. The transformer and convolutional layers were combined as encoder to minimize the loss of low-level details. This methodology was applied to segment eight different organs, but the teeth region segmentation was not carried out. Chen et al.[120] proposed Transformer-based attention guided network called TransAttUnet, in this a multilevel guided attention and multiscale skip connection were designed jointly to enhance performance of semantic segmentation. Multiple datasets were used to evaluate the performance of model. The work was not carried on dental dataset.

Karacan et al. [60] performed the segmentation task on dental panoramic images for segmenting teeth and maxillomandibular region. The authors used three different deep models based on attention mechanism. The two models are based on transformer network namely Vision Transfromer and Segmneter while the other model is convNext. The model was trained and tested on Tufts dataset having 1000 dental panoramic radiographs.

The authors have shown that these models perform better than U-net architecture but suffer from precise segmentation.

Very few methods has been used for dental panoramic image segmentation using transformer. The existing methods segmentation result are not precise thus leaving scope of improvement. To overcome the above issues transformer-based neural network for teeth segmentation is proposed.

6.3 Proposed Method

6.3.1 Proposed Network Architecture

The proposed model is based on an encoder-decoder architecture. The encoder encodes the high-level dental information from the panoramic images. The encoder consists of five convolutional layers with different kernel sizes, each layer is followed by batch normalization and a ReLU activation function. Max pooling operations are applied after each convolutional layer to reduce the spatial dimensions of the feature maps. The transformer is integrated into the proposed model after the last encoder layer. The transformer gathers the information from patches of images with which the pose information is also embedded. The transformer captures long-range dependencies and contextual information for precise dental image segmentation. Following the transformer, the decoder part comprises of the network performs upsampling operations to reconstruct the segmented image. Each upsampling step is performed using max unpooling, which uses the indices stored during the corresponding pooling operation to upsample the feature maps. Similar to the encoder, the decoder consists of deconvolutional layers, batch normalization, and ReLU activations. The final layer of the decoder uses a convolutional layer followed by a sigmoid activation to produce the teeth segmentation map. The

architecture of the proposed model is shown in Figure 6.1. The network is trained with the dice loss.

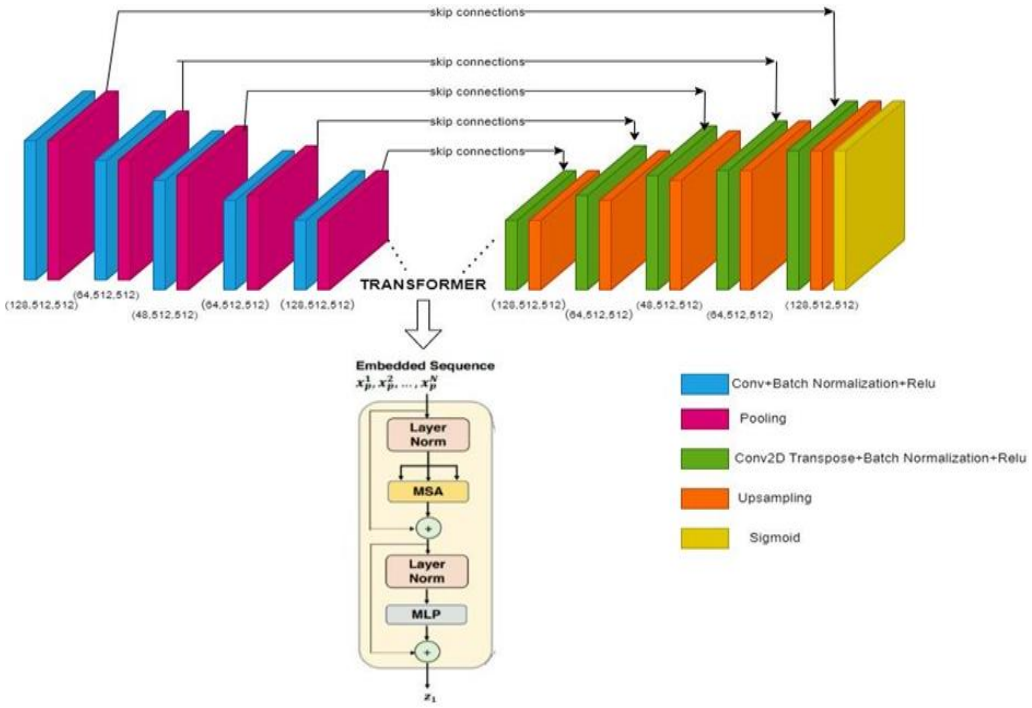

Figure 6.1 Architecture of the proposed model with transformer.

6.4 Experiment and Results

6.4.1 Datasets

The two-benchmark datasets UFBA_UESC[2] dental dataset and Tufts dental database are used to test the proposed deep model. There are 1200 randomly selected images in the training set and remaining 300 images are equally divided between validation set and test set for dataset 1. The second benchmark dataset have 800 randomly selected images in training set while 100 image each in test and validation set.

6.4.2 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 150 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.

6.4.3 Result and Discussion

The proposed model's performance is compared with recent deep learning utilizing the transformers. The deep models are TransUnet[119] and TransAttUnet[120]. Apart the variations of the proposed model have also been implemented. These models are based on the components like using dilated convolutions and with or without skip connections. Also, the modification of TransUnet is done by introducing dilated convolution layers in place of conventional convolution layers. 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 evaluation of the models is done on five different metrics dice score, IoU, accuracy, precision and recall and two different panoramic radiograph datasets.

Table 6.1Performance comparison of deep methods using a transformer structure for dataset 1

Models	Dice Score	IoU	Accuracy	Precision	Recall
Trans-Seg Basic	91.87%	88.83%	96.91%	95.05%	92.13
Trans-Seg Dilated	91.17%	88.08%	96.56%	93.03%	88.13%
Tran-Seg Skip	91.94%	89.83%	96.90%	94.67%	94.61%
Dilated Skip	91.10%	87.50%	96.68%	96.25%	91.72%
TransUnet.[119]	92.06%	90.46%	97.07%	94.67%	94.15%
TransUnet Dilated	92.64%	89.96%	97.17%	93.59%	93.36%
TransATTUnet[120]	91.68%	88.62%	96.92%	94.33%	92.20%

Table 6.2 Performance comparison of deep methods using a transformer structure for dataset 2

Models	Dice Score	IoU	Accuracy	Precision	Recall
Trans-Seg Basic	85.76%	83.58%	97.54%	90.49%	91.25%
Trans-Seg Dilated	85.22%	85.29%	97.43%	91.84%	91.26%
Tran-Seg Skip	85.52%	85.97%	97.64%	91.84%	91.55%
Dilated Skip	82.28%	84.33%	97.26%	90.68%	90.41%
TransUnet.[119]	87.62%	88.38%	98.06%	93.05%	93.62%
TransUnet Dilated	86.77%	87.53%	97.87%	90.83%	90.86%
TransATTUnet[120]	83.47%	84.45%	97.78%	90.30%	94.83%

The results for the dataset 1 is demonstrated in Table 6.1 while for dataset 2 it is demonstrated in Table 6.2. For dataset 1 proposed TransUnet with dialted convolution performs better than the other methods. Specially it beat the performance of original TransUnet. For dataset 2 TransUnet performs better than all the models while the modified TransUnet with dilation is the next best. The simple proposed method and its variant produced the results at par. TransAttUnet performs poorly for dental image segmentation. The visual results for all the models for both datasets are shown from Figure 6.2 to Figure 6.5.

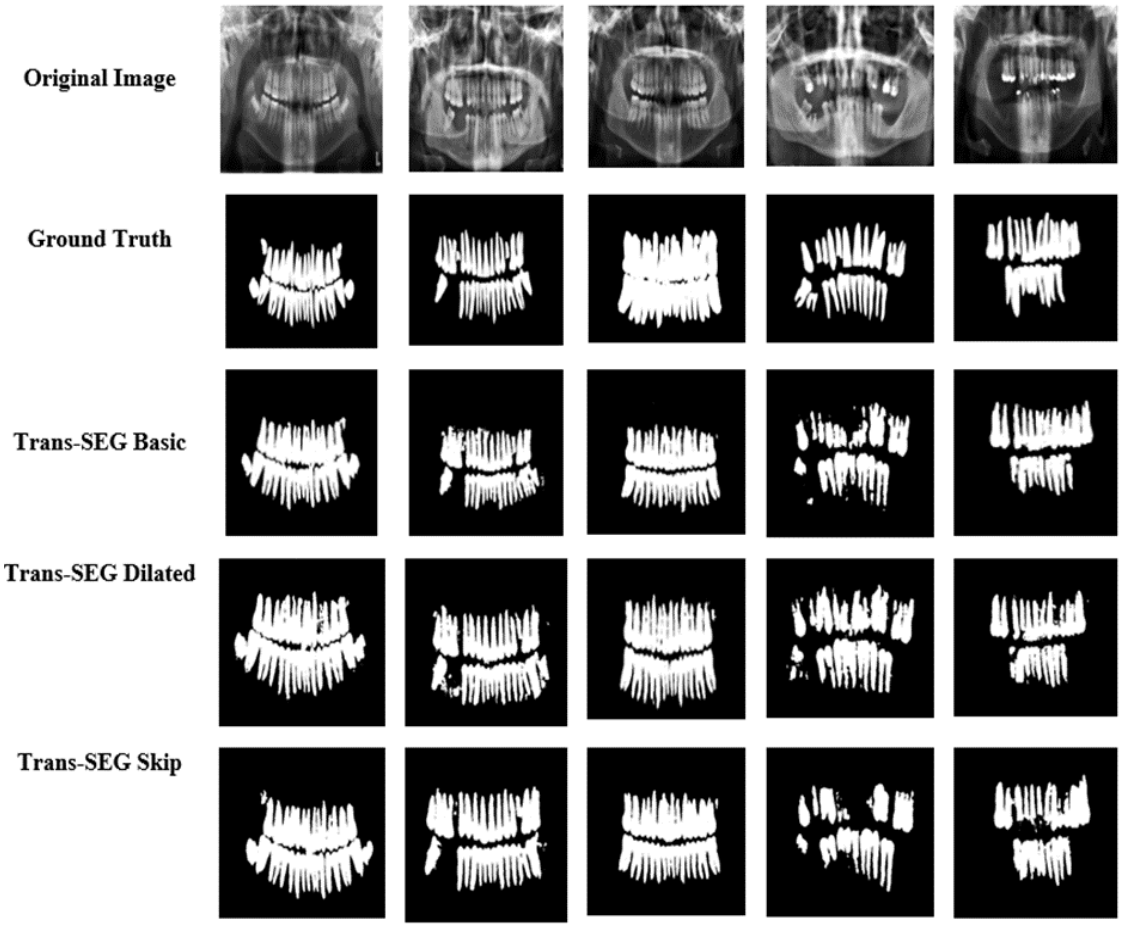

Figure 6.2 Visual results of the transformer based deep models for dataset 1.

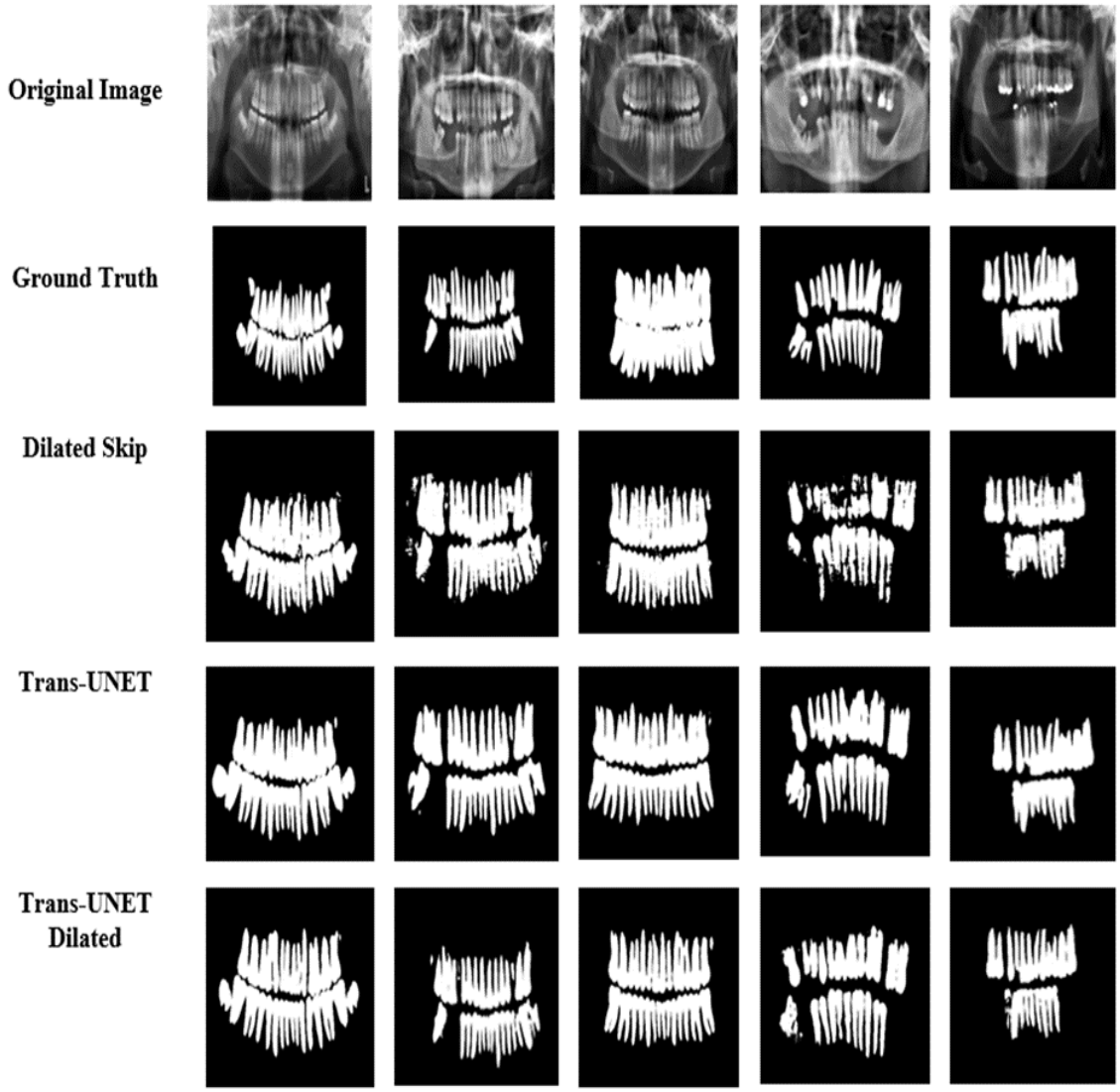

Figure 6. 3 Visual results of the transformer based deep models for dataset 1.

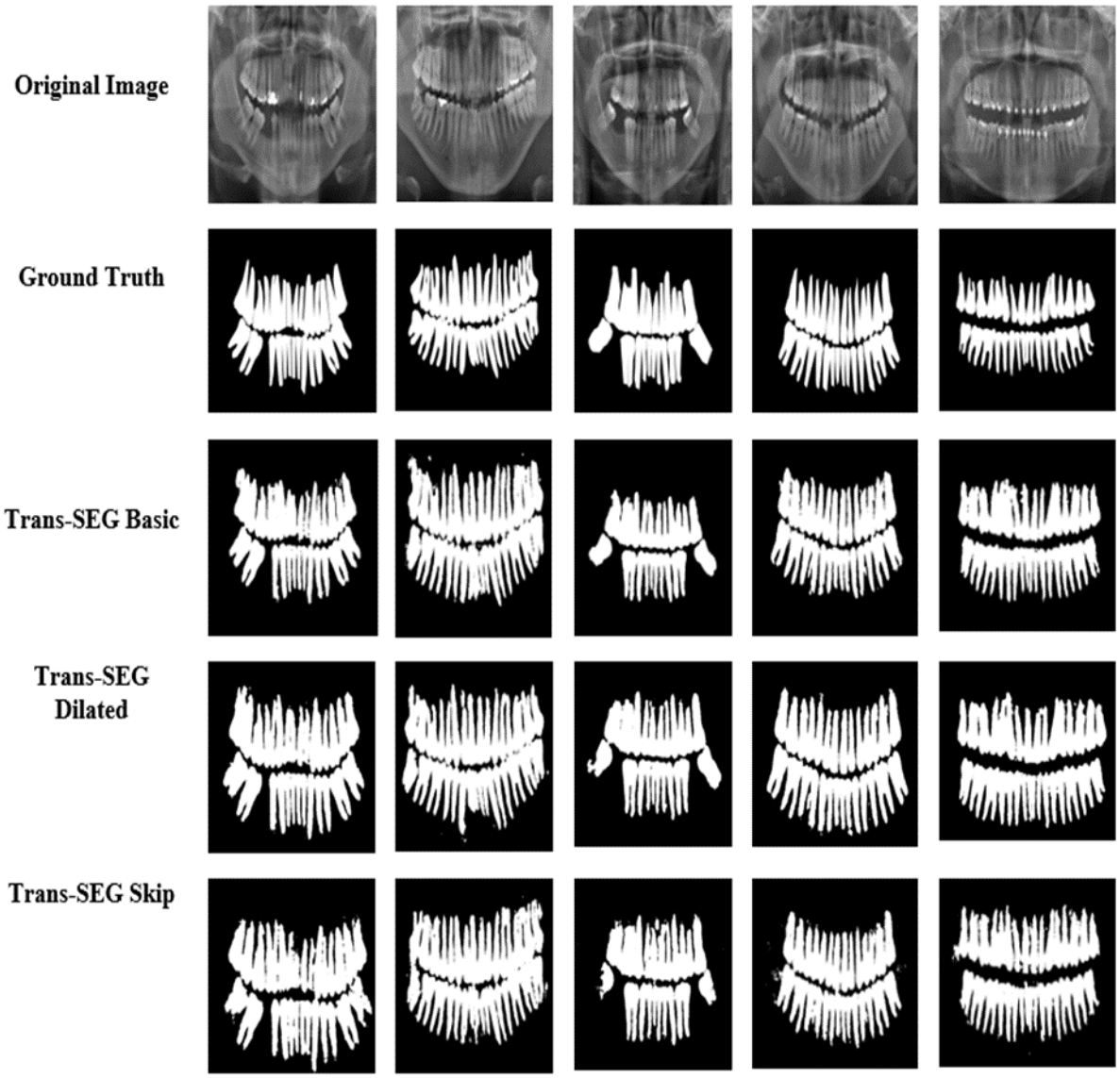

Figure 6.4 Visual results of the transformer based deep models for dataset 2.

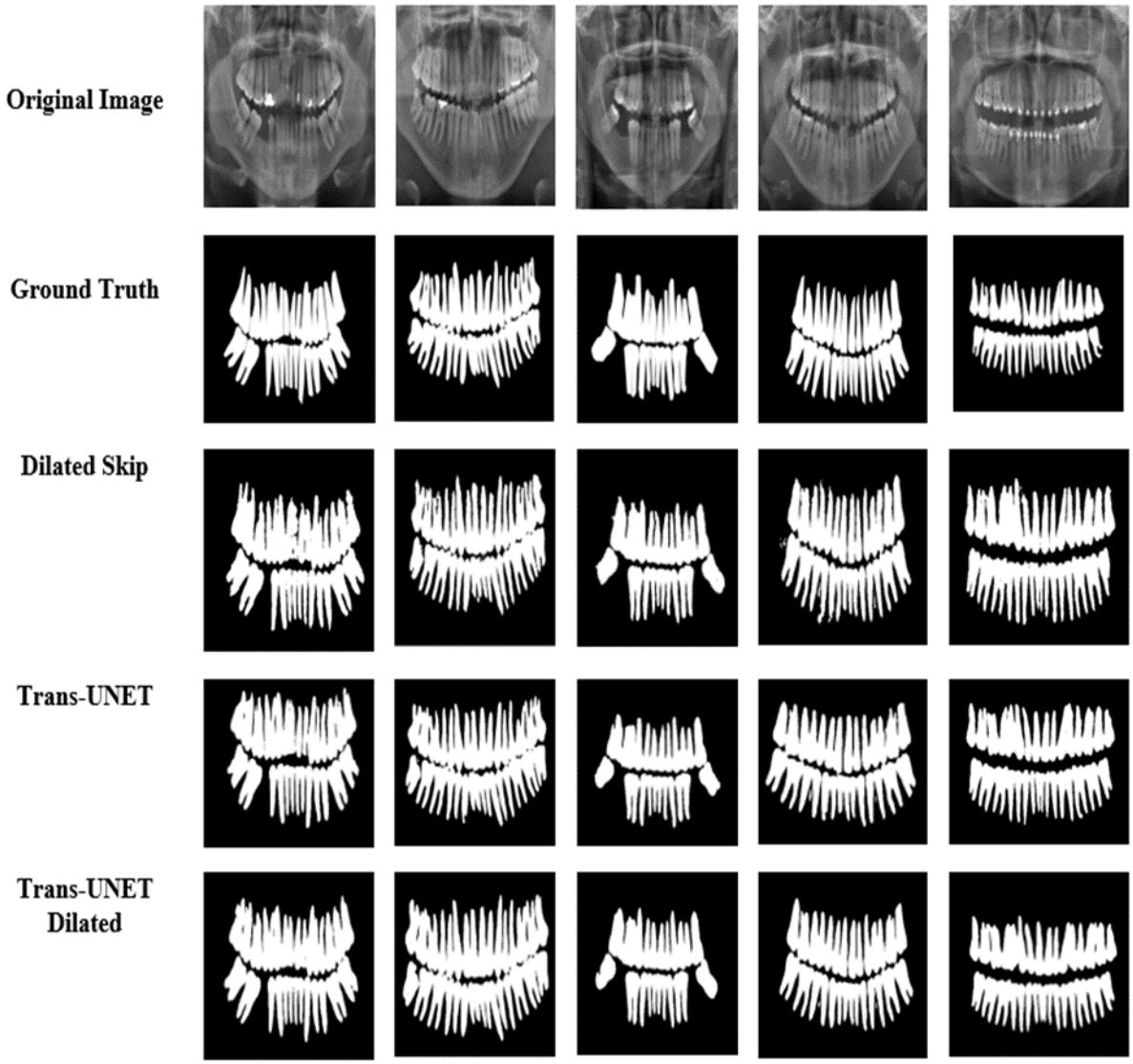

Figure 6.5 Visual results of the transformer based deep models for dataset 2.

6.5 Conclusion

This chapter presented a novel deep neural network using transformers for precise teeth segmentation from dental panoramic radiographs. The transformer structure was integrated with the encoder-decoder architecture to manage spatial information in dental panoramic X-rays. The self-attention mechanism of transformer was utilized to capture long-range information. The skip connections were designed to maintain the

characteristics of original image, which helped in the reconstruction of images. Variants of encoder-decoder architecture using transformer were also investigated. The models were validated on two benchmark dental datasets UFBA_UESC dataset and Tufts dataset. The results showed that transformer-based U-Net architectures performed better than other transformer-based encoder-decoder architectures for dental image segmentation.