Masked Self-Supervised Pretraining for Semantic Segmentation of Dental Radiographs

Tejeswar Pokuri, Laalenthika Konthalapalli*, **Sarvesh Kumar***
Department of Computer Science and Engineering
`tejeswar.mitmpl2022, konthalapalli.mitmpl2023, sarvesh.mitmpl2023`
`@learner.manipal.edu`

Karthik Bhimavarapu
Department of Data Science and Engineering
`bhimavarapu.mitmpl2023@learner.manipal.edu`

Manipal Academy of Higer Education
Manipal, Karnataka, India

Abstract

In dental diagnosis and treatment planning, teeth segmentation is essential. Semantic segmentation of 2D dental radiographs helps to analyze dental structures precisely, detect anomalies in teeth, and evaluate oral health issues. However, creating segmentation masks is a time-consuming process and is prone to inaccuracies due to complex tooth structures. In this research, we propose a self-supervised learning approach for teeth segmentation using Modified ResUNet and Random Block Masking as the pretext task, where random blocks in dental radiographs are masked, and the model is trained to reconstruct the entire radiograph. We utilize only 20% of the samples for training the down-stream task of semantic segmentation. Our proposed approach outperforms state-of-the-art models such as U-Net, PSPNet and performs comparably to LinkNet, trained on 80% of the samples. Modified ResUNet trained on our approach is able to produce an accurate segmentation mask even when the ground truth mask contains errors.

1 Introduction

Dental radiographs are helpful for dentists to identify different infections, cysts, fractures, tumors, and dental caries [1], However, most dental radiographs have low image contrast, reduced brightness, and noise, making it difficult to accurately evaluate them [2]. As a result, teeth segmentation serves as a critical prerequisite for clinical dental analysis and surgical procedures, enabling dentists to comprehensively assess oral conditions and subsequently diagnose conditions [3]. Solutions developed by researchers have employed various supervised learning architectures, including Fully Convolutional Networks (FCNs) like U-Net [3] and PSPNet [16], as well as Generative Adversarial Networks (GANs) [19], Transformers [20], and Mask R-CNNs [18]. Although these approaches achieve a higher Dice score and Jaccard index, they often necessitate a larger training dataset that requires extensive manual annotation of segmentation masks, which is time-consuming, error prone and costly [5].

To address these limitations, the need for a teeth segmentation method that can effectively process dental radiographs using minimal annotated data is evident. Self-supervised learning techniques

---

*Equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

offer a promising solution to this challenge [6,7]. By leveraging the inherent structure and patterns within dental X-ray images, self-supervised methods can learn meaningful representations without relying on large amounts of labeled data. In this research, we propose Random Block Masking as a self-supervised approach inspired by [6,7] for enhanced teeth segmentation with minimal annotated data.

2 Related Works

With the recent developments in attention and transformer architecture, self-supervised learning has gained traction in computer vision domains with scarce amounts of manually annotated data. Conventional approaches for teeth segmentation employ encoder-decoder network architectures like U-Net [9], U-Net++ [14], LinkNet [15] or pyramid-based architectures like PSPNet [16] or FPN [17]. These models specialise in enhancing and blending high-level and low-level features which is crucial in dental segmentation where fine details and broader spatial information are key.

A more recent approach [18] to teeth segmentation uses Mask-RCNNs fine-tuned with a deep ResNet101+FPN backbone. Toothpix [19] utilises cGANs for automatic pixel-level teeth segmentation in dental radiographs. The generator synthesises accurate annotations, while the discriminator distinguishes between real annotations and generated annotations. [20] performs self-supervised learning using SimMIM and UM-MAE with a Swin Transformer encoder, which then serves as the backbone for Cascade Mask R-CNN with FPN in detection and instance segmentation tasks. [20] splits their dataset into 5 folds, with 20% of the dataset in each fold, using 4 folds to train the model. It is worth to mention that transformers are computationally expensive and memory intensive due to their ability of attending to long sequences of tokens [25]. [21] incorporates an attention module into U-Net to enhance the accuracy of tooth segmentation in CBCT radiographs by concentrating on the most relevant features.

3 Methodology

3.1 Modified ResUNet

In this section we propose Modified ResUNet which is used for both pretext and down-stream tasks, inspired by [8,9,10]. The architecture follows an encoder-decoder design, with four encoder and decoder blocks, incorporating residual connections to address the vanishing gradient issue in deep convolutional networks [10]. We also utilize skip connections as described in [9], which help maintain spatial features between the encoder and decoder sections. Instead of concatenating the final output of a residual block with its input, we enhance feature extraction by concatenating the output of the first BatchNorm2D layer with the output of the residual block.

3.2 Random Block Masking

We propose a self-supervised pretext task called Random Block Masking, inspired from inpainting and Masked Auto Encoder [6,7]. This task involves masking random blocks in dental radiographs and training the model to reconstruct the entire dental X-ray. This approach helps in extracting structural features of the teeth, which are essential for the downstream task of teeth segmentation. Masking ratio can dynamically vary from 15% to 75% as the number of blocks can vary from 2 to 5, each covering 15% of the image, while conventional masking techniques like MAE [6] use a constant masking ratio of 25% across all images.

Additionally, since our dataset contains radiographs from different age groups, the presence of a full set of 32 teeth is not consistent across all images. Our solution involves passing the Sobel [12] filtered output of the image, concatenated with the masked X-ray. This guides the model in generating the entire dental radiograph.

3.3 Teeth Segmentation

For the down-stream task of teeth segmentation, Modified ResUNet trained on the pretext task of Random Block Masking is used without modifications to its architecture. The model is trained using a combination of Dice and Binary Cross-Entropy losses. Dice loss helps in achieving accuracy in an

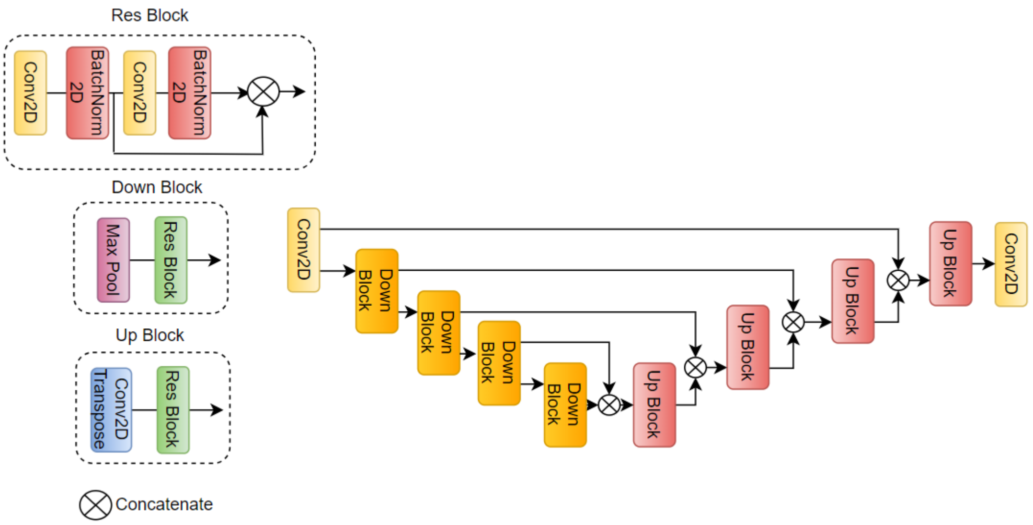
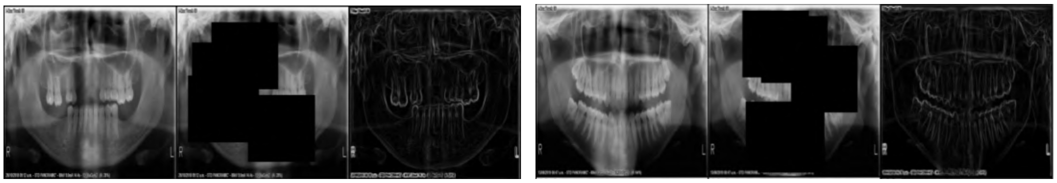


Figure 1: Model overview.



Dental X-ray **Masked X-ray** **Sobel Filter** **Dental X-ray** **Masked X-ray** **Sobel Filter**

Figure 2: The above image shows sample masked X-ray image, X-ray image and Sobel filter.

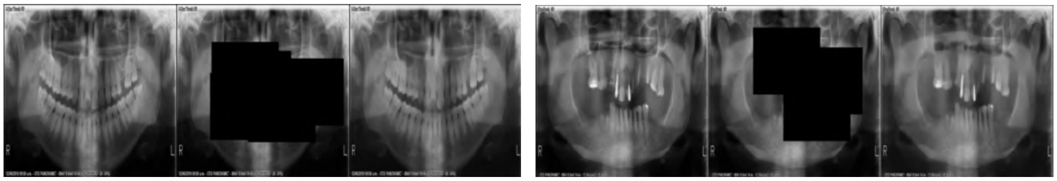
imbalanced dataset and Binary Cross-Entropy treats it as a pixel-wise classification task. Additional information is provided in the appendix.

4 Results and Discussion

We have utilized the publicly available dataset [11], consisting of 598 dental radiographs. This dataset has been annotated by 12 individuals and includes a total of 15,318 segmented teeth. The model is trained on the NVIDIA Tesla P100 GPU, with 16GB memory and 3584 CUDA cores.

4.1 Results related to Random Block Masking

We have achieved an impressive Structural Similarity Index [22] of 0.912, Learned Perceptual Image Patch Similarity [23] of 0.088, and Features Similarity Index Matrix score [24] of 0.956 in the pretext task of Random Block Masking. We trained our Modified ResUNet model for 80 epochs using the Adam optimizer [26] with an initial learning rate of 0.0001, utilizing L1 loss [13] for training. A qualitative analysis is included in the appendix.



Dental X-ray **Masked X-ray** **Generated X-ray** **Dental X-ray** **Masked X-ray** **Generated X-ray**

Figure 3: Sample output images of the pretext task.

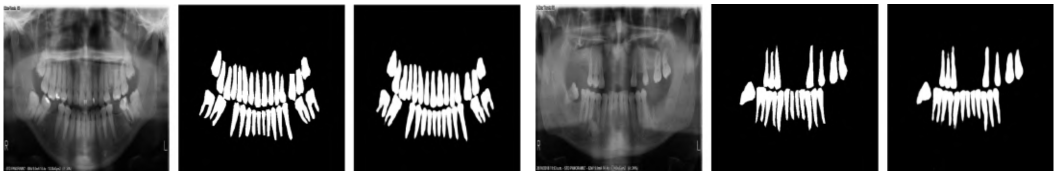
4.2 Results related to Teeth Segmentation

Our self-supervised approach trained on 20% of the samples in the dataset, has achieved a higher accuracy in terms of Dice score and Jaccard score compared to models such as U-Net, LinkNet and PSPNet trained on 80% of the samples. A qualitative and quantitative analysis with the aforementioned models is included in the appendix.

Table 1: Performance Comparison of SOTA models and our approach

Model	Train. Data	Dice	Jaccard
U-Net	0.2	0.673	0.564
	0.3	0.715	0.643
	0.8	0.844	0.753
PSPNet	0.2	0.824	0.700
	0.3	0.779	0.638
	0.8	0.847	0.734
LinkNet	0.2	0.805	0.675
	0.3	0.865	0.763
	0.8	**0.899**	**0.818**

Model	Train. Data	Dice	Jaccard
Modified ResUNet	0.2	0.785	0.698
	0.3	0.814	0.736
	0.8	0.867	0.794
Our approach	0.2	**0.874**	**0.778**
	0.3	**0.881**	**0.786**
	0.8	0.892	0.804

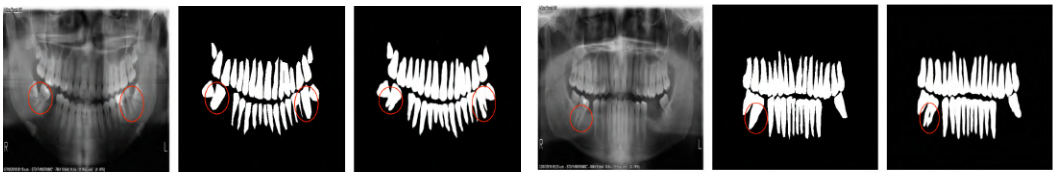


Dental X-ray **Ground Truth** **Generated mask** **Dental X-ray** **Ground Truth** **Generated mask**

Figure 4: Sample output of the downstream task trained on 20% of samples in the dataset with pretext task of Random Block Masking.

4.3 Model Performance with Inaccurate Ground Truth

Due to human involvement in annotation, errors exist in the ground truth segmentation mask. Our approach has predicted the correct segmentation mask despite incorrect ground truth, due to our novel self supervised technique of Random Block Masking.



Dental X-ray **Incorrect mask** **Generated mask** **Dental X-ray** **Incorrect mask** **Generated mask**

Figure 5: Instances demonstrating our model's ability to produce precise segmentation masks despite flaws in the ground truth mask.

5 Conclusion

In this paper we introduce Modified ResUNet for our self-supervised learning technique of Random Block Masking, followed by the down-stream task of teeth segmentation. We achieve accuracy scores rivaling traditional segmentation architectures while simultaneously learning features marked incorrectly in the ground truth mask. We believe that the proposed method can encourage more applications of self-supervised learning in the field of healthcare, reducing the amount of manually annotated data required. As a continuation to our work, we propose multi-class segmentation to distinguish between different classes of teeth - incisors, canines or wisdom teeth.

References

[1] Solanki, A. J., & Mahant, P. M. (07 2017). A Review on Dental Radiographic Images. International Journal of Engineering Research and Applications, 07, 49–53. doi:10.9790/9622-0707074953

[2] Rahmi-Fajrin, H., Puspita, S., Riyadi, S., & Sofiani, E. (2018). Dental radiography image enhancement for treatment evaluation through digital image processing. Journal of clinical and experimental dentistry, 10(7), e629–e634. https://doi.org/10.4317/jced.54607

[3] Chen, X., Ma, N., Xu, T., & Xu, C. (2024). Deep learning-based tooth segmentation methods in medical imaging: A review. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, 238(2), 115–131. doi:10.1177/09544119231217603

[4] Grujev, M., Ilic, M., Milosavljevic, A., & Ilic, A. S. (2024). Review of Teeth Image Segmentation. 2024 11th International Conference on Electrical, Electronic and Computing Engineering (IcETRAN), 1–6. doi:10.1109/IcETRAN62308.2024.10645091

[5] Kunzo, T., Kocur, V., Gajdošech, L., & Madaras, M. (2023, September). Processing and Segmentation of Human Teeth from 2D Images using Weakly Supervised Learning. 2023 World Symposium on Digital Intelligence for Systems and Machines (DISA), 24, 133–139. doi:10.1109/disa59116.2023.10308924

[6] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2021). Masked Autoencoders Are Scalable Vision Learners. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2111.06377

[7] Jenni, S., Jin, H., & Favaro, P. (2020). Steering Self-Supervised Feature Learning Beyond Local Pixel Statistics. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6407–6416. doi:10.1109/CVPR42600.2020.00644

[8] Zhang, Z., Liu, Q., & Wang, Y. (2018). Road Extraction by Deep Residual U-Net. IEEE Geoscience and Remote Sensing Letters, 15(5), 749–753. doi:10.1109/LGRS.2018.2802944

[9] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (pp. 234–241). Cham: Springer International Publishing.

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[11] Humans In The Loop. (2023). Teeth Segmentation on dental X-ray images [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/5884500

[12] Sobel, I., & Feldman, G. (01 1973). A 3×3 isotropic gradient operator for image processing. Pattern Classification and Scene Analysis, 271–272.

[13] Shalev-Shwartz, S., & Tewari, A. (2009). Stochastic methods for l1 regularized loss minimization. Proceedings of the 26th Annual International Conference on Machine Learning, 929–936. Presented at the Montreal, Quebec, Canada. doi:10.1145/1553374.1553493

[14] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4 (pp. 3-11). Springer International Publishing.

[15] Chaurasia, A., & Culurciello, E. (2017, December). Linknet: Exploiting encoder representations for efficient semantic segmentation. In 2017 IEEE visual communications and image processing (VCIP) (pp. 1-4). IEEE.

[16] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).

[17] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).

[18] Zhu, G., Zewen, P., & Kim, S. (02 2020). Tooth Detection and Segmentation with Mask R-CNN. 070–072. doi:10.1109/ICAIIC48513.2020.9065216

[19] Cui, W., Zeng, L., Chong, B., & Zhang, Q. (2021). Toothpix: Pixel-Level Tooth Segmentation in Panoramic X-Ray Images based on Generative Adversarial Networks. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 1346–1350. doi:10.1109/ISBI48211.2021.9433807

[20] Almalki, A., & Latecki, L. J. (2023). Self-supervised learning with masked image modeling for teeth numbering, detection of dental restorations, and instance segmentation in dental panoramic radiographs. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 5594-5603).

[21] Tao, S., & Wang, Z. (2022). Tooth CT Image Segmentation Method Based on the U-Net Network and Attention Module. Computational and mathematical methods in medicine, 2022, 3289663. https://doi.org/10.1155/2022/3289663

[22] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600–612. doi:10.1109/TIP.2003.819861

[23] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 586–595. doi:10.1109/CVPR.2018.00068

[24] Zhang, L., Zhang, L., Mou, X., & Zhang, D. (2011). FSIM: A Feature Similarity Index for Image Quality Assessment. IEEE Transactions on Image Processing, 20(8), 2378–2386. doi:10.1109/TIP.2011.2109730

[25] Fuad, K. A. A., & Chen, L. (2023). A Survey on Sparsity Exploration in Transformer-Based Accelerators. Electronics, 12(10). doi:10.3390/electronics12102299

[26] Kingma, D.P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. CoRR, abs/1412.6980.

[27] Picard, D. (2021). Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. ArXiv, abs/2109.08203.

Appendix

The following sections will provide additional details on the methodology, experimentation, and qualitative analysis.

A Random Block Masking

The dataset [11] includes patients across age demographics and many suffer from common dental abnormalities like missing or misaligned teeth or implants. The Sobel filtered input helps the model generate the entire dental radiograph by providing information on edges, number of teeth, structure, and shape of the teeth behind the random masked block. The filter also helps avoid generating anatomically incorrect shapes or the blurring of sharp boundaries in the reconstructed output. Consequently, providing the Sobel filtered input enhances structural definition and clarity of output images, providing the model with additional cues for interpreting masked regions.

We have trained the model using L1 loss [13], that aids in generating X-rays that are pixel-wise accurate with smoother outputs. L1 loss focuses on minimizing the average deviation across pixels and it is less sensitive to large outliers such as noise and aberrant regions like high contrast or metallic implants, making it suitable for generating X-ray images from masked X-ray images.

L1 loss is defined as follows,

$$\mathcal{L}_1 = \frac{1}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} |I_{ij} - \hat{I}_{ij}| \tag{1}$$

where N and M are the dimensions of the image, I_{ij} is the pixel value of the X-ray at position (i, j), and $\hat{I}_{ij}$ is the pixel value of the generated image at the same position.

B Teeth Segmentation

With the weights from the pretext task of Random Block Masking, we train Modified ResUNet for the down-stream task of segmentation using a combination of Dice and Binary Cross-Entropy (BCE) loss. The combination of Dice and BCE loss optimises for both region overlap and pixel-wise accuracy.

Dice loss is defined as follows,

$$\mathcal{L}_{Dice} = 1 - \frac{2 \sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij} g_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij} + \sum_{i=1}^{N} \sum_{j=1}^{M} g_{ij}} \tag{2}$$

where p_{ij} and g_{ij} are the predicted value and the ground truth respectively, for pixel (i, j).

Binary Cross-Entropy (BCE) loss is defined as follows,

$$\mathcal{L}_{BCE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} [g_{ij} \log(p_{ij}) + (1 - g_{ij}) \log(1 - p_{ij})] \tag{3}$$

where N is the number of pixels, p_{ij} is the predicted probability for pixel (i, j), and g_{ij} is the ground truth. The combination of Dice loss and BCE loss where α and β are hyperparameters are set to 0.7 and 0.3 respectively,

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{Dice} + \beta \cdot \mathcal{L}_{BCE} \tag{4}$$

$$\alpha = 0.7, \beta = 0.3 \tag{5}$$

C Extended Experimentation and Discussion

This section presents a comprehensive analysis, using quantitative and qualitative methodologies to provide a more in-depth examination on our approach. Since we have utilized only 20% of the samples in the dataset for training the down stream task, to ensure better robustness we tested with multiple seed values and we achieved similar accuracy [27]. We evaluate our pretext task's results using the Structural Similarity Index (SSIM). SSIM is typically used when evaluation of structures should not be influenced by luminance, as is the case in this task.

SSIM is defined as follows,

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{6}$$

where x is the reference image, y is the generated image, μ_x is the mean of x, σ_x^2 is the variance of x, σ_{xy} is the covariance of x and y, c_1 and c_2 are constants.

Learned Perceptual Image Patch Similarity (LPIPS) is used to calculate perceptual similarity. It uses the activations of the network for evaluation as the representational space corresponds to human-like perception. The formula for LPIPS is mentioned below,

$$\text{LPIPS}(x, y) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \| w_l \odot (\hat{y}_{hw}^l - \hat{x}_{hw}^l) \|_2^2 \tag{7}$$

Where x, y are input images; l is a network layer; $\hat{y}_{hw}^l$, $\hat{x}_{hw}^l$ are activations at position (h, w).

Feature SIMilarity (FSIM) index was proposed for Image Quality Assessment (IQA) and is composed of 2 components - Phase Congruency (PC) and Gradient Magnitude (GM). PC is important as it is invariant to light variation while GM is computed to store contrast information.

$$\text{FSIM} = \frac{\sum_{x \in \Omega} S_L(x) \cdot \text{PC}_m(x)}{\sum_{x \in \Omega} \text{PC}_m(x)} \tag{8}$$

$$\text{where } S_L(x) = [S_{PC}(x)]^\alpha \cdot [S_G(x)]^\beta S_{PC}(x) = \frac{2\text{PC}_1(x) \cdot \text{PC}_2(x) + T_1}{\text{PC}_1^2(x) + \text{PC}_2^2(x) + T_1} \tag{9}$$

$$S_G(x) = \frac{2G_1(x) \cdot G_2(x) + T_2}{G_1^2(x) + G_2^2(x) + T_2} \text{PC}_m(x) = \max(\text{PC}_1(x), \text{PC}_2(x)) \tag{10}$$

For teeth segmentation the metrics used are Dice score, Jaccard index, Precision and Recall. Dice score is defined as follows. Dice Loss is defined in equation 2.

$$\text{Dice Score} = 1 - \text{Dice Loss} \tag{11}$$

The Jaccard Index, also known as Intersection over Union (IoU), measures the similarity between two sets by dividing the size of their intersection by the size of their union, indicating the overlap between the sets. Jaccard index is defined below,

$$\text{Jaccard Index} = \frac{|A \cap B|}{|A \cup B|} \tag{12}$$

Precision refers to the number of true positives divided by the total number of positive predictions (i.e., the number of true positives plus the number of false positives). Precision is defined as follows,

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{13}$$

Recall is calculated by dividing the number of true positives by the total number of positive instances, which includes true positives and false negatives. Recall is defined as follows,

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{14}$$

D Limitations

Our approach has not been tested on multiple dental radiograph datasets, and we have not performed a quantitative analysis to verify the model's accuracy when the ground truth mask contains errors. Furthermore, the effectiveness of Random Block Masking as a pretext task remains unexplored in other healthcare tasks and across different domains.

E Quantitative and Qualitative Analysis

The tables below provide a more detailed quantitative analysis. In general, the variation across all metrics is minimal [27] at just 0.4%, when training our Random Block Masking approach for segmentation with the modified ResUNet using different seed values and 20% of the samples in the dataset.

Table 2: Performance Comparison of SOTA models and our approach.

Model	Training Dataset	Dice	Jaccard	Precision	Recall
U-Net [9]	0.1	0.628	0.547	0.613	0.712
	0.2	0.673	0.564	0.665	0.748
	0.3	0.715	0.643	0.709	0.798
	0.8	0.844	0.753	0.845	0.898
	0.9	0.876	0.789	0.867	0.876
PSPNet [16]	0.1	0.787	0.648	0.744	0.835
	0.2	0.824	0.700	0.835	0.812
	0.3	0.779	0.638	0.757	0.802
	0.8	0.847	0.734	0.832	0.862
	0.9	0.787	0.649	0.819	0.757
LinkNet [15]	0.1	0.600	0.431	0.832	0.441
	0.2	0.805	0.675	0.840	0.706
	0.3	0.865	0.763	0.861	0.871
	0.8	**0.899**	**0.818**	**0.909**	0.892
	0.9	0.893	0.807	0.909	0.880
Modified ResUNet	0.1	0.765	0.687	0.745	0.834
	0.2	0.785	0.698	0.779	0.865
	0.3	0.814	0.736	0.807	0.827
	0.8	0.867	0.794	0.857	0.857
	0.9	0.907	0.823	0.892	0.867
Random Block Masking + Modified ResUNet	0.1	**0.851**	**0.740**	**0.851**	**0.870**
	0.2	**0.874**	**0.778**	**0.869**	**0.882**
	0.3	**0.881**	**0.786**	**0.874**	**0.889**
	0.8	0.892	0.806	0.878	**0.907**
	0.9	**0.911**	**0.82**7	**0.910**	**0.912**

Table 3: Performance metrics of dataset with different seed values.

Seed	Dataset	Dice	Jaccard	Precision	Recall
42	0.1	0.846	0.727	0.823	0.863
	0.2	0.872	0.774	0.870	0.876
	0.3	0.881	0.787	0.877	0.886
	0.8	0.896	0.812	0.888	0.905
	0.9	0.912	0.834	0.934	0.910
3407	0.1	0.855	0.748	0.842	0.872
	0.2	0.875	0.779	0.870	0.882
	0.3	0.882	0.786	0.869	0.894
	0.8	0.891	0.804	0.873	0.911
	0.9	0.910	0.825	0.912	0.915
0	0.1	0.852	0.745	0.832	0.876
	0.2	0.878	0.782	0.867	0.889
	0.3	0.881	0.787	0.878	0.886
	0.8	0.890	0.802	0.874	0.907
	0.9	0.911	0.824	0.884	0.912

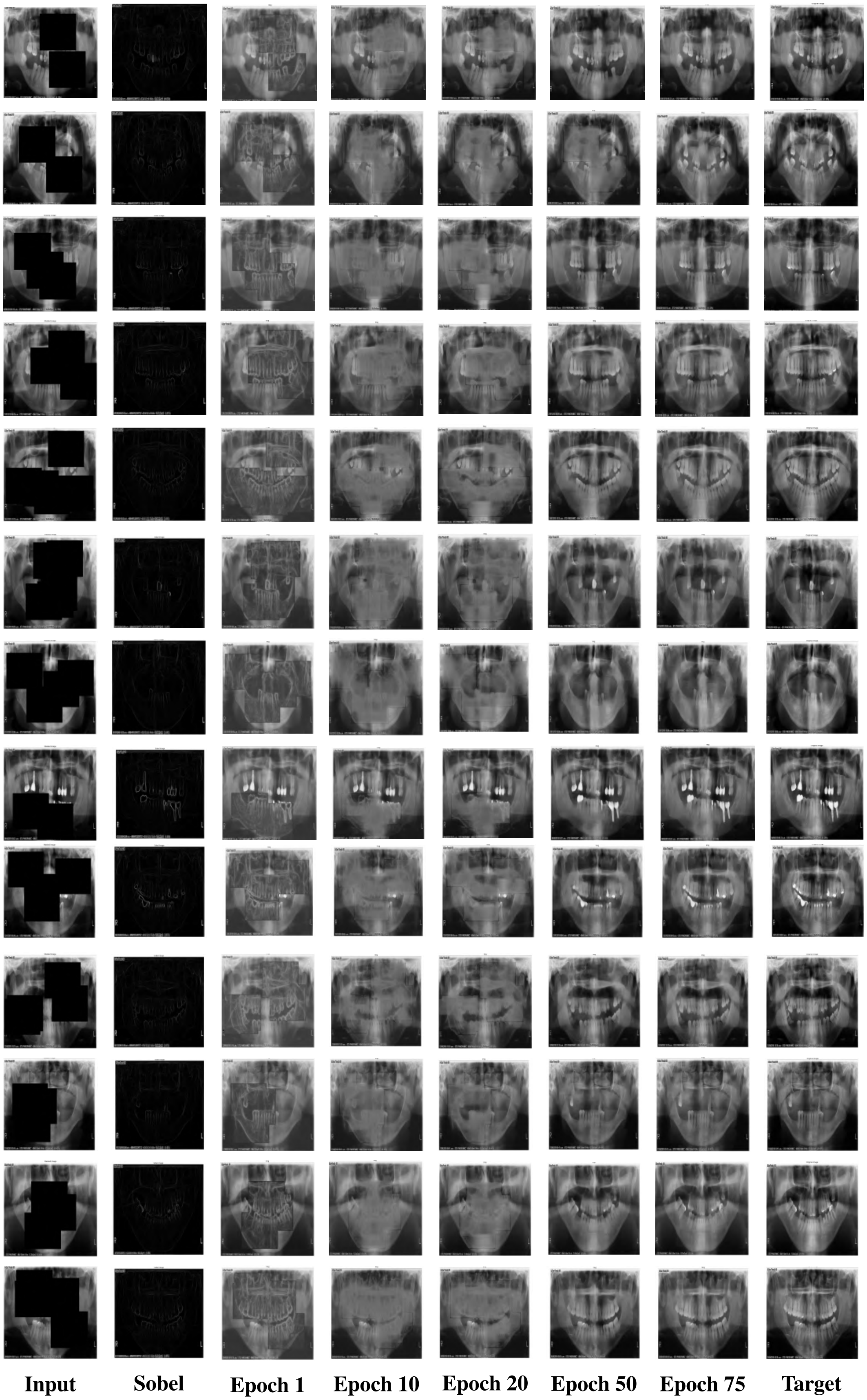
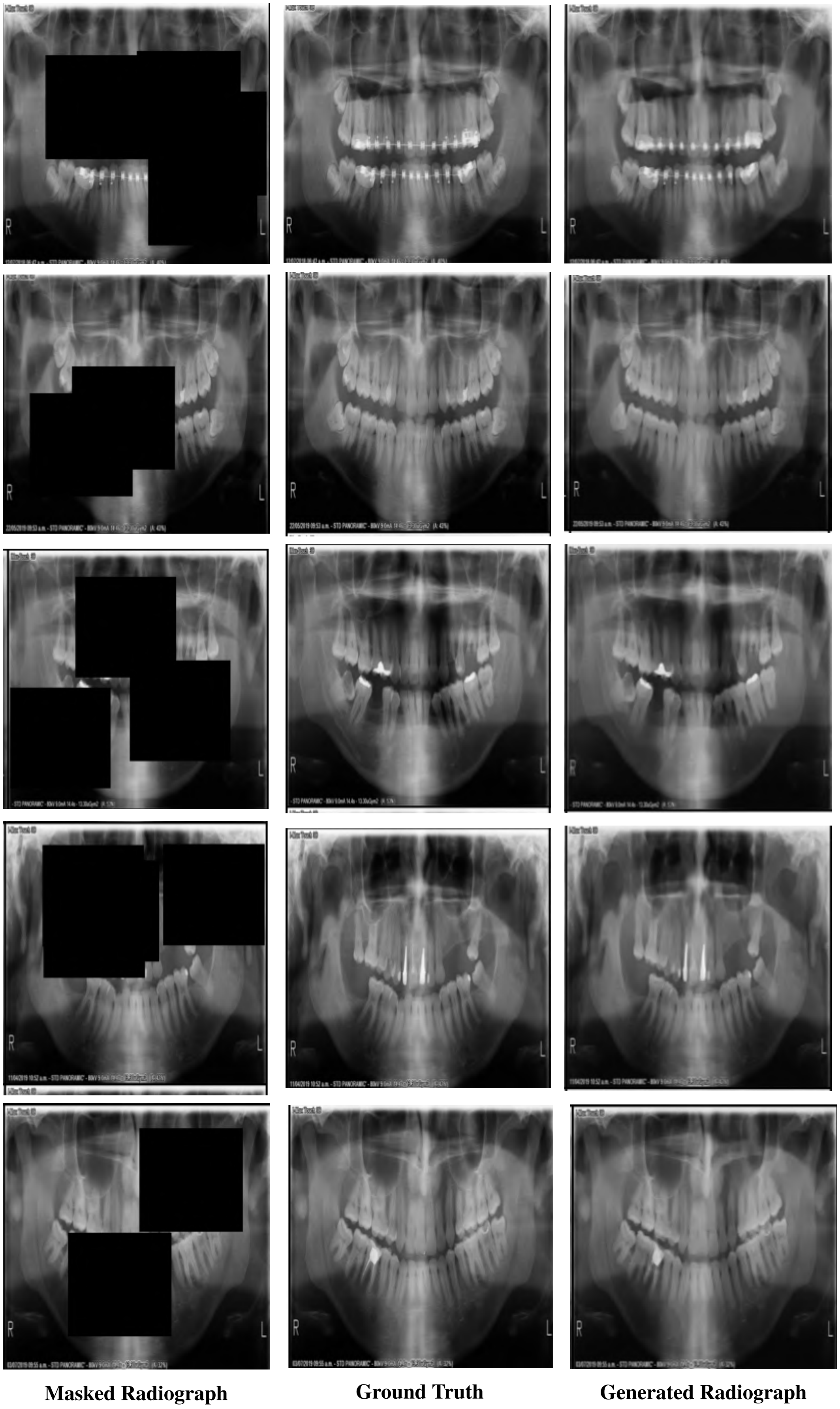


Figure 6: Sequential learning of our model on the Random Block Masking pretext task.



Masked Radiograph **Ground Truth** **Generated Radiograph**

Figure 7: Qualitative Analysis of Random Block Masking.

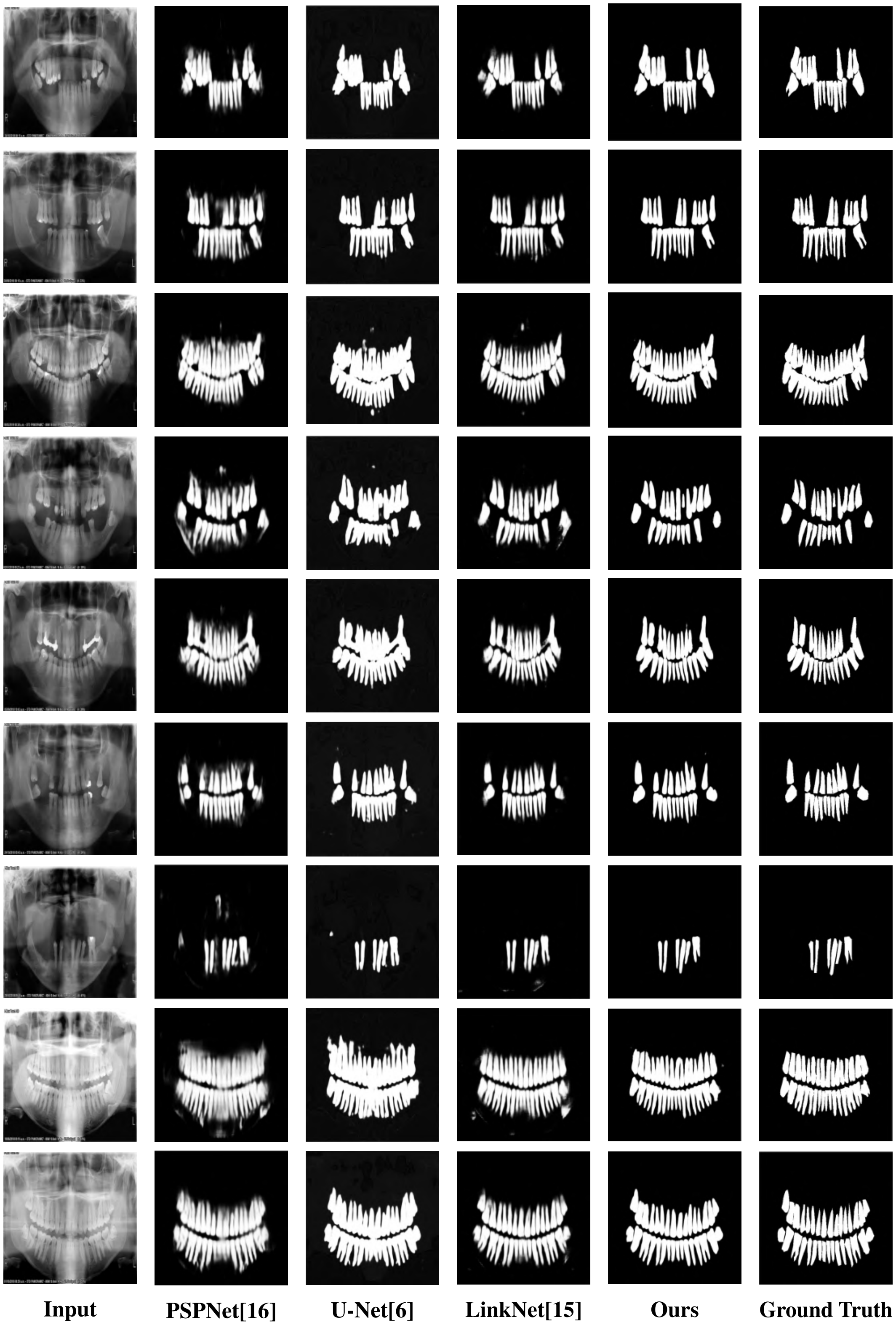
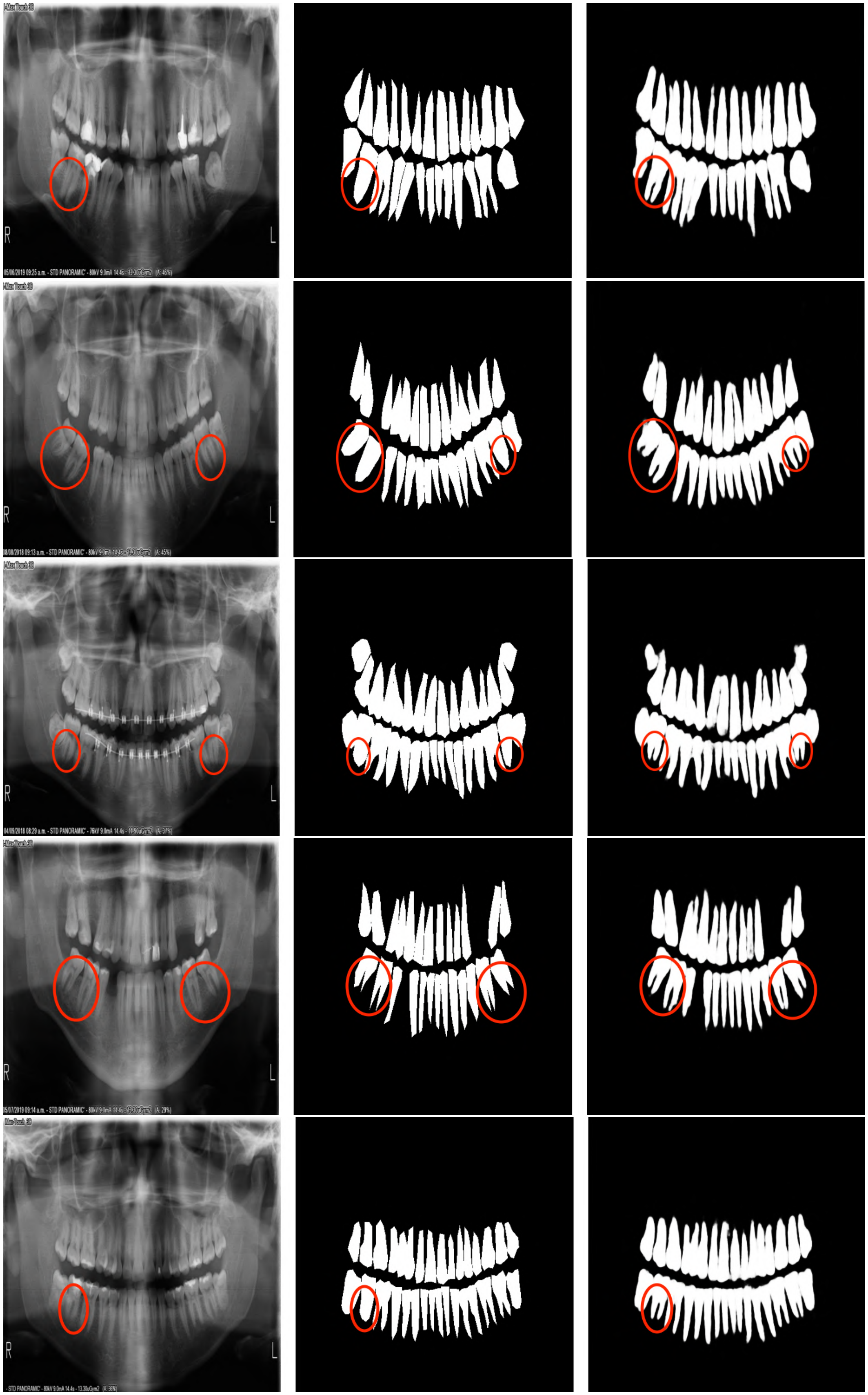


Figure 8: Qualitative Comparison with SOTA models.



Input Radiograph **Ground Truth** **Generated Mask**

Figure 9: Instances demonstrating imperfect ground truth, however our model performs better due to self supervised learning.