Multi-Objective Processing of Dental Panoramic Radiographs

Langston Nashold, Poojan Pandya, Ting Lin
Department of Computer Science, Stanford University
{lnashold, poojanp, linting}@stanford.edu

Abstract

The field of dentistry is highly dependent on the analysis of dental radiographs produced from X-Rays to guide diagnosis of disease. However, the task of observing a radiograph and producing a disease diagnosis must be completed by individuals in the dental field with a high degree of specialization. In this work, we attempt to detect if a radiograph contains a disease (or another "abnormality") in a binary classification problem, in addition to performing pixel-wise segmentation of the teeth contained in the radiograph. While this task has been attempted previously, no approach has yet produced results reliable enough for use in practice.

Here, we present two novel architectures ("Multi-Headed" and "Fusion") for the analysis of dental panoramic radiographs as an attempt to further research in the field. This task makes for an interesting computer vision research problem due its challenging nature yet high potential for impact. Using the newly released multimodal Tufts Dental Database, we aim to perform teeth segmentation as well as abnormality detection without manually injecting domain knowledge or pre-selecting regions of interest. Our experiments show that the multi-headed architecture significantly improves upon the standard ResNet used as a baseline and provides qualitatively sound and interpretable results.

1. Introduction

The field of dentistry is highly dependent on the use of visual data to facilitate and guide diagnosis of disease. Panoramic dental radiography is one of the most common examinations performed in dental clinics. Unlike traditional X-rays, panoramic dental X-rays are extraoral; they are able to quickly capture a single image that shows the patient's teeth, jawbones and surrounding facial structures, providing a comprehensive view of the patient's entire mouth. After obtaining a panoramic dental X-ray from new patients, dentists use the image to label tooth numbers, and identify pre-existing conditions by hand from panoramic radiographs. These tasks require hours of manual labor and require significant dental expertise.

The introduction of Artificial Intelligence (AI) to medical radiographic imaging can help automate many of these processes. Applied to dental radiographs, developments in deep learning and computer vision techniques for semantic segmentation, objection detection and classification can enable dentists to see more patients and augment their accuracy. Although there have been some promising results for diverse tasks such as teeth numbering [1], cavity detection [12], fine-grained abnormality classification [7], an impediment to further progress in the field is the lack of public datasets. Compared to standard computer vision tasks, the collection of medical image data requires dense expert annotation, while ethics surrounding patient privacy may discourage researchers from making data public. Most prior literature complete their own data collection and annotation processes, which takes from three months to over a year. Released less than two months ago, the Tufts Dental Database makes strides in filling in this gap [8]. It is the largest public dataset to date (though still limited in size), with 1,000 thoroughly annotated panoramic dental images. Using this database, we focus specifically on the task of teeth segmentation, abnormality detection, and abnormality localization. This involves accurately recognizing regions of the x-ray that contain teeth, classifying the image as normal or abnormal, then identifying regions of abnormality within the x-ray. Intuitively, the tasks of teeth segmentation, abnormality detection and localization should share objectives, since accurately learning features of teeth is essential for differentiating between healthy and abnormal teeth structures.

Based on this intuition, we propose two hybrid approaches: 1) A multi-head CNN model with one head for teeth segmentation and another for abnormality detection, 2) A fusion approach, using teeth segmentation masks or gaze plots as input in addition to the radiographs for abnormality detection. For teeth segmentation tasks, our output is a predicted binary teeth mask. For abnormality detection, our output is a binary classification for whether or not the given radiograph contains an abnormality. These ap-

proaches also take advantage of the fact that the dataset is small but detailed, with several annotations associated with a single image. We compare their performance on the tasks with each other as well as our baseline model. We find that our approach improves upon the baseline method of training a standard ResNet-50 for teeth segmentation and training a standard ResNet-18 for abnormality detection.

2. Related Work

Core tasks that dentists perform using panoramic dental radiography include teeth detection, teeth numbering, and abnormality detection and labeling. Although we focus specifically on the tasks of teeth and abnormality detection, we expand our discussion to literature that uses deep learning techniques to automate any number and combination of these tasks.

2.1. R-CNN based approaches

Abnormality detection, teeth segmentation and teeth numbering are often treated as object detection problems. Regional Based Convolutional Neural Networks (R-CNN) tackle merging region proposals with a convolutional neural network backbone, and modern variants give state-of-art-results while being relatively efficient. As a result, much prior literature on teeth segmentation and abnormality detection adopts R-CNN based approaches.

Chen, et al. (2019) [1] used Faster R-CNN for automatic teeth detection and numbering with three post-processing techniques to inject prior domain knowledge, including a filtering algorithm, a model to detect missing teeth, and a rule-base module. However, their approach requires significant hand-crafted rule-based processing.

Prados-Privado M, et al. (2021) [11] adopted an entirely neural model ensemble. They used Matterport Mask RCNN for object detection, and ResNet101 for the classification layer, which showed improvement over Faster RCNN. Their choice to use transfer learning from teeth segmentation for teeth numbering is also a valuable insight.

Similar to Prados-Privado M, et al. in using a by-step approach, Lee et. al. (2022) [7] built a four-part model that uses a DICOM converter, Faster RCNN for object detection, a position filtering module, and a polygon shaper, trained on about 23,000 anonymized panoramic dental images to detect 17 fine-grained dental anomalies. Their model could filter out normal images with high sensitivity, but their precision is relatively low.

Aiming to tackle the problem of small datasets in the dental field, Yang et al. (2018) [14] developed a different region-of-interest algorithm, which identifies proposed regions of interest. Then, these inputs are fed into a smaller CNN structure inspired by GooLeNet to classify radiographs in one of three categories, depending on the stage of treatment. In our research, we seek to build upon this by performing an adjacent task and decreasing the reliance on domain knowledge to develop regions of interest.

2.2. Attention and Transformer-based approaches

Following the state-of-the-art results achieved by transformers in the Natural Language Processing (NLP) field, the computer vision community has also increasingly turned to transformers for vision tasks. Vision transformers (ViT) embeds images patches linearly and then feeds the vectors to a standard Transformer encoder. In their survey paper on the usage of ViT in medical computer vision, Parvaiz et. al. [9] found that classification on X-rays was an especially popular task for researchers adopting transformer-based approaches.

Jiang et al. (2021) [6] propose a lightweight attention-based transformer model, with the goal of detecting caries and operating on mobile devices. This paper achieves 62.3% precision and 57.9% recall on caries detection.

Similarly, Sun & Chen (2022) [12] proposed an attention-based transformer model for dental caries (teeth decay) detection. Their model uses a sparse R-CNN with FPN backbone with a novel attention module, which improved upon the baseline Faster RCNN result. Their results show promise for small attention-based modules on smaller datasets; however, because they limit their abnormality detection to caries only, their approach would need to be modified to better handle broad-scope abnormality detection.

2.3. Self-Supervised Learning

Large annotated medical datasets often require significant expertise, time and effort to create. Self-supervised learning, which obtains supervisory signals from the data itself by leveraging the underlying structure in the data, is a method that can take advantage of large amounts of unlabeled medical images and greatly reduce the difficulty of data collection.

Applying self-supervised learning to dental caries classification, Taleb et. al. (2022) [13] trained three self-supervised algorithms on a large corpus (38k) of unlabeled bitewing radiographs (BWRs). They then applied the learned neural network representations on tooth-level dental caries classification, fine tuning the model using labels extracted from electronic health records. They found that using as few as 18 annotations can produce sensitivity comparable to human-level diagnostic performance.

2.4. Fusion Techniques

Surveying prior literature, we found that objection detection using Faster-RCNNs and deep model backbones show promising results when there is a sufficiently large dataset (¿=20k), while more novel, modular models work better with small datasets and narrow tasks. Because our dataset is

different than the datasets used by most prior literature, we aim to take advantage of the fact that the dataset is small but densely annotated using multi-objective training. Gadzicki et. al. [3] discovered that early multi-modal fusion showed significant performance improvement over uni-modal approaches for human activity recognition.

Applying multimodal fusion with deep neural networks to medical imaging, Huang et. al. [5] developed and compared different multimodal fusion model architectures for leveraging both Computed Tomography scans and electronic patient data. They found that a late fusion model (late Elastic Average) significantly outperformed both early fusion and uni-modal approaches. Because their dataset and objective is more similar to ours than Gadzicki et. al., we keep the favorable results obtained by late fusion in mind while developing our own approach.

3. Data

3.1. Dataset

We are using the Tufts Dental Database [8] for our project. This database, released in April 2022, provides 1000 panoramic x-rays labeled with bounding boxes for the maxillomandibular region (region of interest), bounding boxes for teeth, teeth numbers, and specific abnormalities in each x-ray. It also contains corresponding masks for teeth and for abnormalities. Moreover, the dataset contains gaze plots from eye-tracking data, which record the areas of the x-ray a dentist spent the most time analyzing, as well as short textual descriptions of the abnormalities present in each radiograph. Each image is of pixel size 840 $\times$ 1615, which we downsample and resize to pixel size 260 $\times$ 400.

Figure 1. Radiograph

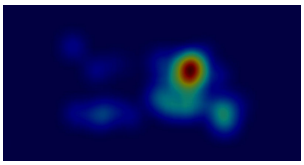

Figure 2. Gazemap

Figure 3. Teeth Mask

Figure 4. Abnormality Mask

We adopt a standard 80/10/10 train/validation/test split. Since the dataset is relatively small, we performed data augmentation with random crops, horizontal flips, and contrast adjustments, which allows us to effectively generate additional radiographs with varying size, brightness and contrast, as x-rays naturally have varying levels of exposure. We experimented with different combinations of data augmentation strategies during our experiments, and the above combination yielded the best results overall. We did not perform data normalization because the pixel values of our radiographs are already normalized zero to one.

Because the dataset doesn't provide labels for whether or not a given radiograph contains abnormality, we preprocess the data to create binary labels based on the provided abnormality masks. We found 340 images containing abnormalities, and 660 images without.

3.2. Single-Objective Baselines

For our baseline, we approach the two tasks separately: 1) teeth segmentation, and 2) abnormality detection. For both tasks, our input is a single dental panoramic radiograph resized to pixel size 260 $\times$ 400.

3.2.1 Teeth Segmentation

As a baseline method for teeth segmentation, we used a PyTorch built-in DeepLabV3 with a ResNet backbone (deeplabv3_resnet50), which is a SOTA model for segmantic segmentation. For a more detailed explanation of our baseline models, see Section 4.

We were able to achieve an IOU of 91.2%, a per-pixel accuracy of 94.9% and a F1 of 0.705.

3.2.2 Abnormality Detection

For abnormality detection, we used a ResNet18 (see Section 4) with pretrained weights. Because our dataset is small, we used a smaller ResNet, stronger regularization and early stopping to prevent overfitting. We achieved a baseline accuracy of 84% with early stopping.

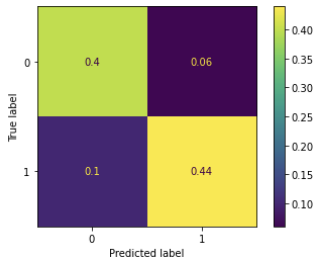

Figure 5. Confusion Matrix for Abnormality Detection

4. Methods

4.1. Problem Statement

Motivated by the stronger baseline results for teeth segmentation compared to abnormality detection, we explore methods that take advantage of the previous task for the later with a multi-objective model (or model ensemble). We

introduce two such models: 1) a multi-headed CNN model; 2) a fusion CNN model.

4.2. ResNet

The ResNet architecture has shown strong performance on a wide range of applications, including object detection, semantic segmetation, image classification [4]. The ResNet architecture is composed of multiple ResNet blocks. Within each ResNet block, a 3x3 convolution is applied, followed by BatchNorm, Relu, and another convolution (we can call this chain of operations F(x)). Finally, the key characteristic of ResNet blocks are the "skip connections", in which we add an identity mapping to the output of the block (so our final output is F(x) + x). This allows us the ResNet architecture to very easily learn the identity mapping, which in turn, has allowed ResNets to become much deeper than previous models. We use ResNet as a base model in both of our architectures.

4.3. DeepLabv3

DeepLabv3 was chosen as a base architecture because it has shown state-of-the-art results on semantic segmentation tasks and is well-supported by modern frameworks [2, 10]. It is built on a backbone architecture (in our case, ResNet50), with a Atrous Spatial Pyramid Pooling (ASPP) layer built on top of it. The ASPP relies on atrous convolutions – which are similar to normal convolutional layers, however, they add a "dilation" parameter, which widens the field of view while preserving spatial information.

Deeplabv3 builds a segmentation model by using four atrous convolutions in parallel with progressively larger dialation parameters. These preserve the dimensionality of the input, and allow the model to segment the object on different scales. These atrous layers allow for excellent performance on segmentation tasks.

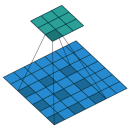

Figure 6. Atrous Layer

We also considered using a transformer-based architecture for the backbone, but found that in prior literature, CNN backbones yield better results in low-data environments.

4.4. Multi-Headed CNN with Hard Parameter Sharing

The multiheaded CNN architecture was implemented in PyTorch with a single ResNet-50 backbone [10]. On top of this, we have two heads – a two layer fully-connected network for classification and an Atrous Spatial Pooling Pyramid for segmentation (as in [2]). Each head produces a separate output, on which a loss for the respective task is computed.

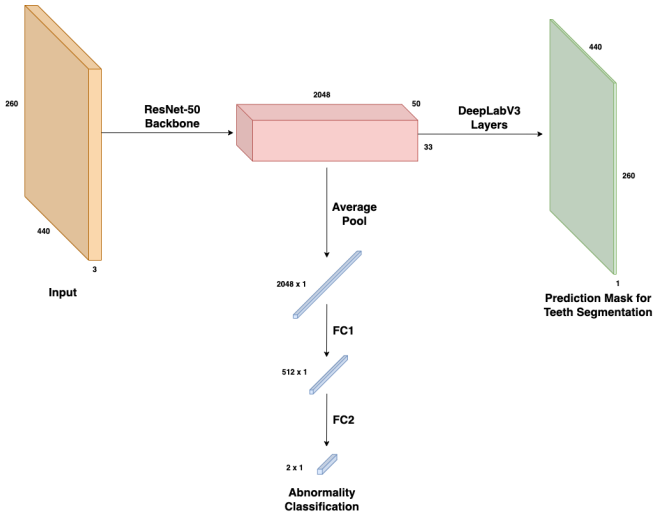

Figure 7. Multiheaded Model

The motivation for a multiheaded approach was several fold. First, parameters learned in the process of training one task can be shared by another task. This means that parameters learned by the segmentation task can be used by the classification task, and vice versa. Furthermore, a multiheaded approach also allows for implicit data augmentation – each task has a higher dataset size than it would if we were training on it individually. This is especially important considering the limited size of our dataset.

4.5. Fusion CNN

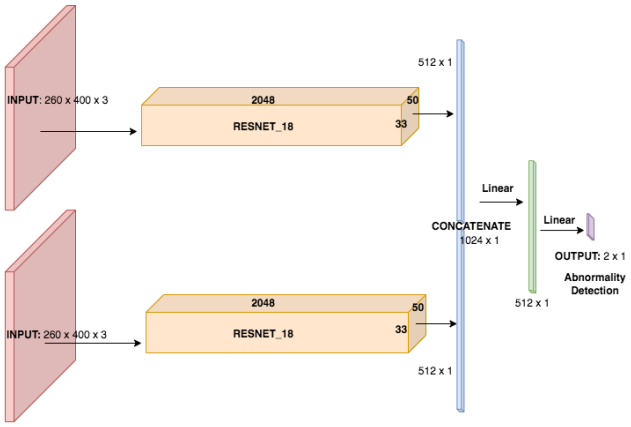

Figure 8. Fusion CNN Model

The fusion model sought to determine if it would be easier to predict abnormalities if the model had access to more data. Therefore, we created a model to train on a slightly different task – instead of producing both a predicted abnormality label and a predicted teeth mask, it took both the radiograph and *either* the ground-truth teeth mask *or* the

gaze map as input, and predicted an abnormality label. It was expected that this model would produce as good or better results than the baseline, since it had access to a strict superset of the supervised data.

The model was built on two ResNets, each pretrained on ImageNet. The outputs of these two ResNets were passed through average pooling layers, then concatenated together, before being passed through a FC (see Figure 7).

We chose a late fusion approach rather than an early fusion approach. This was supported by the work done by Huang et al., which found much better results in similar medical imagery. Furthermore, late fusion is also generally understood to produce better results for classification tasks, because errors between each data source are handled independently. Finally, we were working with a relatively small dataset, and late fusion allowed a better transfer from pretrained ImageNet weights.

4.6. Loss Functions

We experimented with three different loss functions for segmentation (detection always used weighted BCE Loss). Jaccard and Dice are both expressed in set notation as follows.

$$\text{Jaccard}(U, V) = \frac{|U \cap V|}{|U \cup V|}$$

$$\text{Dice}(U, V) = \frac{2|U \cap V|}{|U| + |V|}$$

$$\text{BCE}(y, \hat{y}) = -w * (y \log(p) + (1 - y) \log(1 - p))$$

Our selection of loss functions was largely guided by an attempt to mitigate the effects of class imbalance. Furthermore, Dice and IOU loss both correlate more strongly with the intended goal of our model to localize teeth and abnormalities as accurately as possible.

5. Experiments

5.1. Metrics

To evaluate our results for teeth and abnormality segmentation, we used 1) pixel accuracy (PA) and 2) intersection over union (IoU), and 3) F1 score. We report PA and IOU as is standard for semantic segmentation tasks; however, because our segmentation task for abnormality segmentation is severely imbalanced, we report F1 to quantify if we are able to correctly identify abnormal pixels. We also visualize our predictions for teeth and abnormality segmentation in comparison to the ground truth masks to qualitatively assess our model. For abnormality detection, which is a standard binary classification task, we report accuracy and F1 scores.

5.2. Hyperparameter Tuning

We chose to tune the following hyperparameters: learning rate, regularization strength, BCE loss weights (when applicable), and the number of layers of the ResNet frozen with pre-trained weights from ImageNet. We chose these parameters because they were commonly tuned in our literature review. Furthermore, they empirically had the greatest affect on our final result. We created a training and validation split as previously described, and each hyperparameter was randomly initialized either uniformly or on a logarithmic scale (learning rate and regularization strength). We performed two passes – a coarse grained approach using a wide range of variables, and a fine-grained approach on a narrower range of well-performing values.

We found the optimal values for our hyperparameters were as follows:

L.R.	Reg.	Frozen Layers	BCE Weight	Epochs
2e-4	1e-5	3	2.1	5

Table 1. Optimal Hyperparameters

5.3. Results

	Segmentation			Detection	
	F1	IOU	Accuracy	F1	Accuracy
Baseline	70.5	91.2	94.9		0.84
Multi-Headed	**76.0**	**92.4**	**95.6**	**0.92**	**0.92**
Fusion	N/A	N/A	N/A	0.69	0.80

Table 2. Comparative Results for Teeth Segmentation and Abnormality Detection

We evaluate our two models, "Multihead" and "Fusion", in comparison to the baseline and report results in Table 2. Overall, our multihead CNN model shows moderate improvement over the baseline model for teeth segmentation, and strong improvement over the baseline for abnormality detection. Our late-fusion CNN model was underperforming compared to our expectations. We tried both ground truth teeth masks and gaze maps as a second input, but ultimately decided on gaze maps. Although we hoped that the multi-input fusion of radiographs and gaze maps would allow the model to learn richer feature representations of healthy teeth while receiving guidance to locate potential regions of interest, the accuracy was in fact 4% lower than the baseline. We suspect that this is because our dataset was too small to successfully tune the final layers of the two ResNet18 models. In addition, a qualitative analysis of the gaze maps found that they sometimes, but did not always

correlate to regions of abnormalities, which may have introduced noise to our training process. In the future, we might attempt to use our fusion model using a different stream of input instead of the gaze maps, such as written descriptions of abnormalities. We might additionally select our input data more carefully to filter out gaze maps that did not correlate to regions of abnormality at all.

A qualitative analysis of our teeth segmentation predictions (Fig. 9 & 10) show that they are highly consistent with the ground truth teeth masks, although less fine-grained on the borders between teeth and background. Given the strong baseline for the task, the results are not surprising, though still noteworthy.

Figure 9. Prediction

Figure 10. Ground Truth

By looking at the confusion matrices for abnormality detection in Fig.5 (baseline) and Fig.11, we see that our fusion model is less sensitive to abnormalities compared to both our baseline and multihead model. This may be because the training for our fusion model was less effective due to backpropagation through the two ResNet18 backbones.

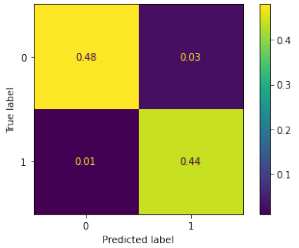

Figure 11. Confusion Matrix for Multi-Head Model

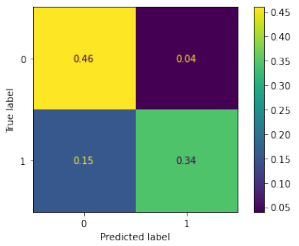

Figure 12. Confusion Matrix for Fusion Model

5.4. Ablation Study: Loss Functions

As previously mentioned, we experimented with weighted binary Cross Entropy loss, IOU loss, and Dice loss for segmentation tasks.

We can see that for all metrics, IOU produces the best result (although it is not significantly higher than Dice loss). IOU and Dice loss are very similar metrics, so this makes sense. We can also see Dice and IOU have significantly higher F1 scores than weighted BCE. One potential cause of this is because both these functions are better suited for segmentation tasks.

	W-BCE	IOU	Dice
IOU	0.914	**0.918**	0.918
Accuracy	0.952	**0.952**	0.951
F1	0.696	**0.767**	0.757

Table 3. Comparative Results for Loss Function

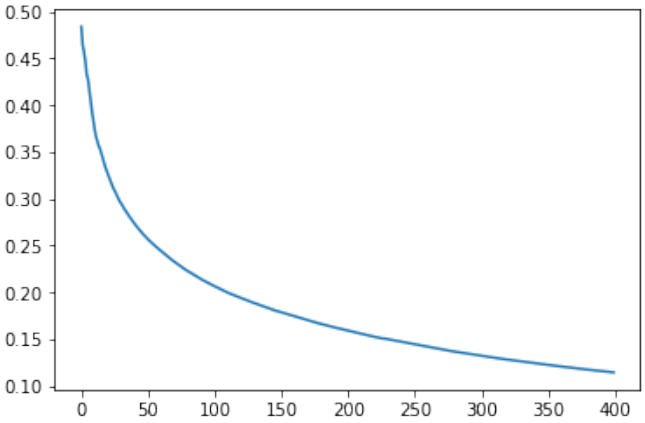

Figure 13. Training Loss Curve with Dice Loss

5.5. Abnormality Segmentation

We also preliminarily explore the task of abnormality segmentation. Compared to teeth segmentation and abnormality detection, this task is significantly more difficult as the dental radiographs in our small dataset contain a wide range of different non-severe, minor abnormalities.

Using our multi-modal model pretrained on teeth segmentation and abnormality detection, we then continued to train the segmentation head using our abnormality mask data. Because there were only around 34% images with abnormalities present, and each region of abnormality was very small compared to the full radiograph, the class imbalance was very severe—in fact, predicting a complete mask (with all black pixels) would achieve a 99.8% per pixel accuracy across the validation set. To remedy the class imbalance, we used weighted binary cross-entropy loss, where penalizing mis-classifying an abnormal pixel 100 times more heavily than mis-classifying a normal one.

Although we achieve high validation accuracy (94.5%) and IOU (94.3%), this is largely due to the model successfully masking normal segments. A qualitative analysis of our predictions shows that there is still large room for improvement. Future work should consider using fixed crops of images for abnormality detection, and then perform segmentation on the crops before piecing the cropped masks together.

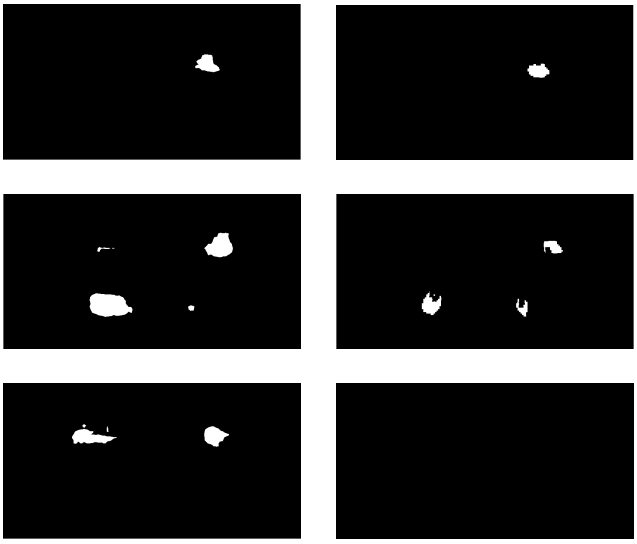

Figure 14. Abnormality Segmentation. Left: Predictions; Right: Ground Truths

Figure 15. Radiograph

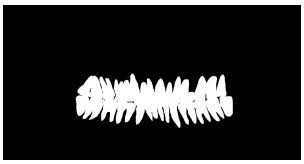

Figure 16. Ground Truth

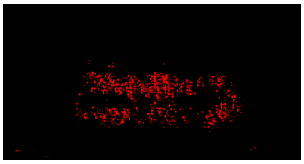

Figure 17. Saliency Map

Figure 18. Prediction

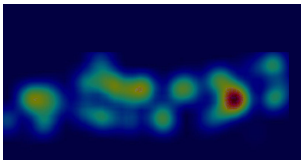

Figure 19. Gaze Plot

5.6. Saliency

To reduce the "black-box" nature of CNN models, we computed saliency maps for teeth segmentation outputs, which visualize the relative contribution of each pixel in the input image as a way to interpret how our model is performing segmentation. To compute the saliency maps, we back-propagate the gradient through the input image and compute the absolute value of the gradient with respect to each pixel. Then, we take the maximum value over the three color channels to create the saliency plot. It is important to note that, especially when working with medical data, model interpretability is a key aspect of effective computer vision research. To aid in this analysis, our dataset contains "gaze plots" for each of our examples, which show which areas of the radiograph dental experts focused on when evaluating the radiograph. Eye-tracking software was used to generate these plots while dentists evaluated and labeled the provided data. Figures 14–18 show the ground truth segmentation, our predicted segmentation, gaze plot, and saliency map for one radiograph. We found that our saliency maps on validation data generally aligned well with the maxillomandibular region of interest and the gaze plots provided, which demonstrates that the model is not overfitting to the training data and is generally evalutating the right features.

6. Conclusion

We have demonstrated a body of work comparing various methods tackling the analysis of dental panoramic radiographs to perform teeth segmentation and abnormality detection. Of the methods we tried, we found that the Multi-Headed CNN produced the best results for teeth segmentation and abnormality detection. This is likely because the learned features for each of these tasks are helpful for the other, so our multi-headed approach enables us to take advantage of these adjacent tasks with a paradigm inspired by and similar to transfer learning. Furthermore, we experimented with various hyperparameters, architectures, and design choices to find the optimal configurations for our particular task.

6.1. Future Work

While we aimed to take advantage of the rich annotations given in our dataset, we were unable to explore all combinations of multi-modal inputs due to time and computing constraints. One direction future work could explore is utilizing expert text descriptions of pre-existing conditions present in the radiograph, which are provided in the same dataset. It could involve processing the descriptions using a contextual attention module and concatenating or averaging the resulting vectors with the sequence vectors obtained through passing the radiographs through a ViT. Group ViT would be a good baseline model to start with. Furthermore, particularly when performing abnormality segmentation, we felt limited by the amount of labeled data we had access to. To mitigate this, it could be productive to try a self-supervised approach with unlabeled data to learn more features specific to dental radiographs.

7. Acknowledgements

The three authors of this paper contributed equally to this project. All three authors contributed to the two model im-

plementations and the final write-up. Langston Nashold implemented the Dataset and Dataloader implementations, as well as the training loop. He also implemented the hyperparameter tuning framework. Ting Lin ran baseline experiments, helped develop the approach, assisted with hyperparameter tuning, and kept research logs of all experiments performed. Poojan Pandya obtained the dataset, helped develop the approach, ran data analysis methods, assisted with hyperparameter tuning, and developed saliency maps. For implementing Dice and IOU loss, we used the code provided in the Kaggle notebook "Loss Function Library - Keras & PyTorch", which was released under the Apache 2.0 open source license. All other code was written using the PyTorch framework, including PyTorch implementations of resnet-18, resnet-50, and deeplabv3. We thank our project mentor Zhuoyi Huang, and CS231N TA William Shen for their help with this project.

References

[1] Hu Chen, Kailai Zhang, Peijun Lyu, Hong Li, Ludan Zhang, Ji Wu, and Chin-Hui Lee. A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films. *Scientific Reports*, 9(1), Mar. 2019. 1, 2

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 4

[3] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *23rd International Conference on Information Fusion (FUSION)*, pages 1–6. IEEE, 2020. 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 4

[5] Shih-Cheng Huang, Anuj Pareek, Roham Zamanian, Imon Banerjee, and Matthew P. Lungren. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific Reports*, 10(1), Dec. 2020. 3

[6] Hao Jiang, Peiliang Zhang, Chao Che, and Bo Jin. RDFNet: A fast caries detection method incorporating transformer mechanism. *Computational and Mathematical Methods in Medicine*, 2021:1–9, Nov. 2021. 2

[7] Sangyeon Lee, Donghyun Kim, and Ho-Gul Jeong. Detecting 17 fine-grained dental anomalies from panoramic dental radiography using artificial intelligence. *Scientific Reports*, 12(1), Mar. 2022. 1, 2

[8] Karen Panetta, Rahul Rajendran, Aruna Ramesh, Shishir Rao, and Sos Agaian. Tufts dental database: A multimodal panoramic x-ray dataset for benchmarking diagnostic systems. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1650–1659, Apr. 2022. 1, 3

[9] Arshi Parvaiz, Muhammad Khalid, Rukhsana Zafar, Huma Ameer, Muhammad Ali, and Muhammad Fraz. Vision transformers in medical computer vision – a contemplative retrospection. 03 2022. 2

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 4

[11] María Prados-Privado, Javier García Villalón, Antonio Blázquez Torres, Carlos Hugo Martínez-Martínez, and Carlos Ivorra. A convolutional neural network for automatic tooth numbering in panoramic images. *BioMed Research International*, 2021:1–7, Dec. 2021. 2

[12] Che Sun and Hu Chen. An attention-based transformer model for dental caries detection. In Guoqiang Zhong, editor, *International Conference on Electronic Information Engineering, Big Data, and Computer Technology (EIBDCT 2022)*, volume 12256, pages 673 – 679. International Society for Optics and Photonics, SPIE, 2022. 1, 2

[13] Aiham Taleb, Csaba Rohrer, Benjamin Bergner, Guilherme Leon, Jonas Rodrigues, Falk Schwendicke, Christoph Lippert, and Joachim Krois. Self-supervised learning methods for label-efficient dental caries classification. *Diagnostics*, 12:1237, 05 2022. 2

[14] Jie Yang, Yuchen Xie, Lin Liu, Bin Xia, Zhanqiang Cao, and Chuanbin Guo. Automated dental image analysis by deep learning on small dataset. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 01, pages 492–497, 2018. 2