Hierarchical Multi-Label Object Detection to Analyze Panoramic Dental X-rays

Bhavya Singh 200040036
Amruta Parulekar 20D070009
Annie D'Souza 20D070028
Bhavya Kohli 20D070021
Tejaswee Sulekh 20D070082
Sanjhi Priya 20D070070

I. Introduction

In the field of dental radiology, panoramic X-rays stand out as specialized images capturing a wide view of the entire mouth, including teeth and surrounding structures. These images provide comprehensive insights crucial for treatment planning and diagnosis. However, interpreting panoramic X-rays can be time-consuming and prone to misdiagnosis or miscommunication due to the exhaustive nature of the process.

Multi-label object detection is a critical technique employed in dental radiology, specifically for identifying and locating various objects within panoramic X-rays. In this context, each object typically represents a tooth, and the detection process involves assigning multiple labels or categories to each tooth. This method, structured hierarchically to represent different levels of detail, starts with identifying the mouth quadrant, then enumerating individual teeth within that quadrant, and finally, providing a diagnosis for each tooth.

Despite the benefits of AI in dental radiology, challenges persist, including anatomical variations and a lack of publicly available annotated data. However, the integration of AI technologies in dental practices offers promising opportunities to assist doctors, improve treatment outcomes, and enhance patient satisfaction.

II. Related works

One related work[1] in object detection involves the utilization of wavelets. A novel method proposes gray-scale object detection using the direct wavelet transform. The innovation lies in multilevel processing of both the object and the image, combined with analyzing wavelet transform coefficients. This approach results in a significant improvement in computation time.

Another related work[2] in object detection utilizes a diffusion-based approach. This method incorporates a Swin-Transformer as the image encoder, leveraging its capabilities learned from ImageNet-22k. The detection decoder, Modified DiffusionDet, employs three classification heads and leverages learning from previous stages. Additionally, it employs multi-label detection by freezing the classification heads.

III. Problem definition

Our objective is to analyse panoramic dental x-rays in a hierarchical fashion to detect all teeth and assign multiple labels. These dental x-rays are panoramic to give us a complete view of the jaw and clear views of every singular tooth. Upon these panoramic x-rays, we want to use object detection models to detect abnormal teeth and its positions based on the FDI system. The FDI system labels teeth according to the quadrant they are present in and assigns position labels 1-8. This problem is handled hierarchically. We are to begin by detecting the quadrants for each x-ray. Following that, the teeth are to be detected in all these quadrants and classified based on their type/position. Finally, these teeth must be identified as normal or having one of the diagnosis classes such as caries, deep caries, periapical lesions, and impacted teeth. Thus, each tooth has multiple labels, obtained in a step-by-step fashion.

IV. Objectives

1) To develop algorithms that can accurately detect abnormal teeth with dental enumeration and associated diagnosis , aiding in accurate treatment planning and helping practitioners carry out procedures with a low margin of error.
2) To improve computation time and efficiency by using techniques like wavelet transforms for preprocessing and feature extraction.
3) To address class imbalance by using loss functions like focal loss, intelligent data subset selection and geometric augmentations.
4) To experiment with different object detection techniques and get better performance than the baselines.

V. The Dataset

The dataset used for this project comprises X-rays from patients aged 12 and above, randomly selected from the hospital's database to ensure patient privacy and confidentiality. It includes diagnosis classes such as caries, deep caries, periapical lesions, and impacted teeth. Additionally, there are 1571 unlabeled X-rays for pre-training. The dataset is hierarchically organized into three categories:

1) 693 X-rays labeled only for quadrant detection and quadrant classes,

2) 634 X-rays labeled for only tooth detection with quadrant and tooth enumeration classes, and
3) 1005 X-rays fully labeled for only abnormal tooth detection with quadrant, tooth enumeration, and diagnosis classes.

The hierarchical labeling system allows for a structured organization of the dataset. X-rays labeled for quadrant detection and quadrant classes provide a foundational level of information. This is followed by X-rays labeled for tooth detection, which includes quadrant and tooth enumeration classes, allowing for a more detailed analysis. Finally, X-rays fully labeled for abnormal tooth detection contain quadrant, tooth enumeration, and diagnosis classes, offering comprehensive insights into dental abnormalities.

The dataset follows the FDI system, a globally-used dental enumeration system that assigns a number from 1 through 4 to each quadrant of the mouth, as well as numbers 1 through 8 to each of the eight teeth and each molar. This system significantly simplifies the dental enumeration task, providing a standardized approach for identifying and labeling teeth in the X-rays.

VI. Methods Overview

A. Pipeline

Three models are utilized in this project: M1, M2, and M3. M1, an object detection model, detects and annotates quadrants in raw panoramic images using only quadrant annotations. M2, another object detection model, detects teeth within quadrant patches and requires tooth annotations, as well as either quadrant annotations or a well-trained M1. Finally, M3 is a CNN classifier that classifies tooth patches into disease classes, relying on annotated teeth and their respective disease ground truths for training. This hierarchical approach allows for comprehensive analysis of dental images, starting from quadrant detection, moving to tooth detection, and concluding with disease classification.

B. Training and Inference

The training setup for the three models follows a hierarchical process. First, annotations for quadrants are used to train M1, which detects and annotates quadrants in raw panoramic images. Then, M1 is used to extract quadrant patches from images with tooth annotations. These quadrant patches, along with tooth annotations, are used to train M2, which detects teeth within the quadrant patches. Finally, M2 is used to extract tooth patches from images with annotations and disease ground truths, which are then used to train M3, a CNN classifier that classifies tooth patches into disease classes. This sequential approach ensures that each model builds upon the outputs of the previous one, leading to a comprehensive analysis of dental images. The inference process begins with a raw panorama, where M1 extracts quadrant patches. These patches are then inputted into M2, which retrieves tooth patches. Finally, M3 classifies individual teeth patches into disease classes. This sequential approach ensures a thorough analysis of dental images, starting from the broader view of quadrants down to the detailed classification of individual teeth based on their conditions.

C. Pros and cons

The proposed method offers several advantages. Firstly, it doesn't require all images to have all types of annotations; a subset of panoramas annotated for quadrants can be used to train M1, which can then be employed to annotate quadrants on images with only tooth annotations, and similarly for M2. This flexibility reduces annotation requirements and facilitates scalability. Secondly, the approach allows for the use of smaller, more specialized models instead of a single large model, enhancing interpretability and efficiency. Lastly, targeted human supervision can be utilized to fine-tune specific aspects of the workflow without the need to re-train the entire model each time.

However, there are also drawbacks. If the performance of M1 is suboptimal, it may result in poor annotations used for training M2, leading to reduced performance in subsequent stages. This dependency on the accuracy of M1 could potentially limit the overall effectiveness of the approach.

VII. Methods in Detail

A. Preprocessing with Wavelet Compression

We used Haar wavelet compression for lossless compression of the input data. This involved iteratively applying the Haar wavelet transform on the data, effectively decomposing it into multiple frequency components.

To capture the most significant information, we computed a weighted average across all the resulting components from the Haar transform. These weights were meticulously chosen using a grid search approach, ensuring optimal selection for our specific model.

Further, for training purposes, the bounding box annotations had to be scaled according to the level of compression applied i.e., for N-level compression, the images were scaled by $\frac{1}{2^N}$ and so were the bounding box coordinates.

B. Object detection models M1 and M2

We use the mmdetection library to implement the object detection models M1 and M2 because of its modular workflow and ease of setting up. The library offers an extensive list of usable models including the current state-of-the-art, the Co-DETR model [6]. This model is an improvement over the base DETection Transformer (DETR) and uses collaborative hybrid assignments for object detection along with the typical DETR heads as the readout layer (transformer decoder).

Multi-GPU training and inference are also carried out using the utility functions defined in the library and on the Co-DETR official github page here

C. Processing between models

As described in the section about our dataset, the third dataset only contains bounding boxes for diseased teeth and the images in the datasets do not overlap. So without this hierarchical pipeline, we can't obtain bounding boxes for normal

teeth. In order to train each model, some pre-processing was required based on the inferences of the previous model.

1) For training M2: The second dataset (containing only tooth bounding boxes) is passed through a trained M1 to obtain quadrant bounding boxes. Using these, the image is cropped to obtain 4 different patches, one for each quadrant. Next, we reflect the patches for quadrant 2, 3, and 4 to align with quadrant 1 and introduce some uniformity in the tooth positions and to minimise the effects of the differing quadrants, hoping to make the training for M2 easier.

For the bounding box labels in the annotations file for the second dataset, we begin by shifting the coordinates according to quadrant positions as the cropping changes the origin of the images. Next, we have to apply a transformation on the bounding box coordinates for teeth corresponding to quadrants 2, 3, and 4 due to the reflections performed on the images.

Finally, an updated annotations file is generated which refers to the cropped patches as the images under consideration and the corresponding bounding boxes.

2) For training M3: The third dataset (containing tooth bounding boxes only for abnormal teeth) is passed through a trained M1 to obtain quadrant bounding boxes. We obtain quadrant patches and shift bounding box labels in the updated annotations file as described in the previous section.

Next, we run these quadrant only images through a trained M2 to obtain bounding boxes for all teeth. For bounding boxes with valid confidence scores, we crop to obtain patches and resize them to prepare for M3 training.

Finally, we go through the annotations file and for existing teeth without any assigned labels, we assign normal.

D. Classifier with long tail solutions

In refining our tooth patch classifier, we focused on improving both the model itself and how we trained it. We chose the EfficientNetB0 model and paired it with a method called ADAM to make it work better. We also adjusted how fast the model learned and stopped training it early to avoid overdoing it. Our dataset had five types of tooth patches, but one type called "Normal" was way more common than the others, with 16,595 examples. To make sure the model learned from all types equally, we used a method called Focal Loss. We also carefully picked a smaller set of images from the Normal class using intelligent subset selection to keep things fair. Plus, we made more examples for the less common types of patches using simple tricks of geometric augmentation. These changes made our classifier much better at recognizing different types of tooth patches, which is important for dentistry.

E. Feature extraction with wavelets

We thought of utilizing wavelet feature information of the images and concatenating these features with the pre-final feature layer of the classifier i.e. efficient-net.
The method employed was as follows:

- Obtained wavelet features of the images by recursively passing them through wavelet transformation
- Extracted features from the pre-final layer of the efficient-net model
- Step 3 was performed in two different ways:
  1. Concatenate the features from steps 1 and 2 directly
  2. Pass the features obtained from step 1 through a linear layer and concatenate the output with the features from step 2.
- Train this revised model with validation and perform inference

The different compression ratios used were: 3,4
The linear layer divisions used were: For R=3:9 (3196 to 348) and for R=4:3 (784 to 261).
The best results were obtained with the compression ratio of 3.

TABLE I
RESULTS FOR M1 AND M2

Model	Compression Degree	AP $IoU_{0.50:0.95}$	AP $IoU_{0.75}$	AR $IoU_{0.50:0.95}$
M1	0	0.713	0.907	0.797
	3	0.713	0.890	0.795
	4	0.699	0.867	0.791
	5	0.698	0.900	0.776
	SOTA (2020)	0.651	0.524	0.727
M2	0	0.543	0.583	0.718
	3	0.542	0.587	0.727
	4	0.343	0.275	0.650
	SOTA (2020)	0.494	0.394	0.668

VIII. RESULTS

A. M1 and M2

The results are shown in Table VIII-A. As can be seen, for M1, results are comparative to the SOTA model with slight decrease for every additonal each degree of compression. For M2, we find compression degree 3 to be optimal, and degree 4 onwards, the model suffers considerably.

B. M3

The results are tabulated in Table VIII-B. We can observe a significant improvement in class-wise accuracy for the diseases with less representation by using focal loss and geometric augmentation, even with compressed data.

TABLE II
RESULTS FOR M3

Experiments	Class-wise Accuracy(%)					Overall
Baseline	83.22	75.44	8.54	31.7	90.66	72.7%
Geometric augmentations	80.95	83.50	28.57	63.74	82.26	75.6%
Focal Loss	85.39	70.85	41.67	42.65	86.04	74.88%
Wavelets compression degree 3 w GA	91.35	80.24	24.78	34.59	87.32	78.9%
Wavelets compression degree 4 w GA	89.87	78.18	31.82	30.34	86.22	76.0%
(Number of Images)	604	2189	158	578	2000	5529

IX. Conclusion

In conclusion, this work demonstrates the successful implementation of hierarchical multi-label object detection for analyzing panoramic dental X-rays. By incorporating Haar wavelet compression as a preprocessing step, we achieved performance comparable to models that did not utilize compression. Notably, the compressed data significantly reduced training time across subsequent models (M1, M2, and M3). For instance, training M1 without compression took 1 hour, while using compressed data reduced this to 28 minutes. Similar reductions were observed in M2 (4 hours to 2 hours). This reduction highlights the efficiency gained by working with lower-resolution images, a key factor for deploying the architecture on edge devices with limited processing power and memory constraints. The reduced training time and potential for deployment on edge devices suggest this approach holds promise for real-world applications in dental diagnostics. Our code can be found on Github.

X. References

[1] Bohush, Rykhard & S., Maltsev & A, Aniskovich. (2005). Object Detection Using Wavelet Transform.

[2] Ibrahim Ethem Hamamci et al., (2023). Diffusion-Based Hierarchical Multi-Label Object Detection to Analyze Panoramic Dental X-ray.

[3] Dentex Challenge, 2023, https://dentex.grand-challenge.org/

[4] Dataset: https://zenodo.org/records/7812323#.ZDQE1uxBwUG

[5] OpenMMLab: https://github.com/open-mmlab/mmdetection

[6] Zong, Zhuofan, Guanglu Song, and Yu Liu. "Detrs with collaborative hybrid assignments training." Proceedings of the IEEE/CVF international conference on computer vision. 2023.