Multi-Granularity Tooth Analysis via Faster Region-Convolutional Neural Networks for Effective Tooth Detection and Classification

Samah AbuSalim[1], Nordin Zakaria[2], Salama A Mostafa[3], Yew Kwang Hooi[4], Norehan Mokhtar[5], Said Jadid Abdulkadir[6]

Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Seri Iskandar 32160, Malaysia[1, 2, 4, 6]
Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400, Johor, Malaysia[3]
Dental Simulation and Virtual Learning Research Excellence Consortium-Department of Dental Science-Advanced Medical and Dental Institute, Universiti Sains Malaysia, Bertam, 13200, Kepala Batas, Penang, Malaysia[5]

***Abstract*—In image classification, multi-granularity refers to the ability to classify images with different levels of detail or resolution. This is a challenging task because the distinction between subcategories is often minimal, needing a high level of visual detail and precise representation of the features specific to each class. In dental informatics, and more specifically tooth classification poses many challenges due to overlapping teeth, varying sizes, shapes, and illumination levels. To address these issues, this paper considers various data granularity levels since a deeper level of details can be acquired with increased granularity. Three tooth granularity levels are considered in this study named Two Classes Granularity Level (2CGL), Four Classes Granularity Level (4CGL), and Seven Classes Granularity Level (7CGL) to analyze the performance of teeth detection and classification at multi-granularity levels in Granular Intra-Oral Image (GIOI) dataset. Subsequently, a Faster Region-Convolutional Neural Network (FR-CNN) based on three ResNet models is proposed for teeth detection and classification at multi-granularity levels from the GIOI dataset. The FR-CNN-ResNet models exploit the effect of the tooth classification granularity technique to empower the models with accurate features that lead to improved model performance. The results indicate a remarkable detection effect in investigating the granularity effect on the FR-CNN-ResNet model's performance. The FR-CNN-ResNet-50 model achieved 0.94 mAP for 2CGL, 0.74 mAP for 4CGL, and 0.69 mAP for 7CGL, respectively. The findings demonstrated that multi-granularity enables flexible and nuanced analysis of visual data, which can be useful in a wide range of applications.**

Keywords*—Dental informatics; intra-oral image; deep learning; faster region-convolutional neural network; classification; granularity level; tooth detection

I. Introduction

As living standards improve and dental health awareness increases, a growing number of individuals are pursuing dental treatments (such as orthodontics, dental implants, and restoration) as a means to maintain a healthy lifestyle [1]. According to the WHO Global Oral Health Status Report (2022), almost 3.5 billion people worldwide are affected by oral illnesses [2]. In underdeveloped nations, the lack of oral hygiene knowledge, limited access to dental care facilities, and high cost of treatment contribute to untreated dental issues, resulting in severe consequences for individuals in these regions [3].

Extensive research has been conducted to explore Deep learning-based object detection methods for dental disease diagnosis in various dental models including radiographic images, CBCT images, and intra-oral images [4] [5].

Radiographs and periodontal images are widely used as objective diagnostic tools for tooth disease diagnosing. This includes bitewing, periapical, and panoramic images. Despite their widespread use these images are known to have limitations. For example, they are likely to contain tooth ghost images, low resolution and contrast, overlaps, angulation, magnification, and other artifactual information which are sources of unwanted features and noise [6]. Alternatively, CBCT is employed for their high-quality three-dimensional volumetric information which addresses the issue of distortion and superimposition of bony and dental structures [7]. However, automatic segmentation using CBCT poses certain difficulties, such as noisy images, unclear edges, presence of a human skull [8].

Recently, intra-oral dental images are used for tooth disease diagnosis. They provide valuable insights into a patient's oral health status and help in formulating treatment plans [9]. This approach (i) does not necessitate specialized equipment for data acquisition, (ii) offer rich features despite small image size, and (iii) consequently requires low computational cost for image processing and object detection tasks. However, the identification and detection of individual teeth in these images present some challenges such as partial occlusion, overlapping, and varying Illumination [10] [11]. Another issue is unavalaibility of comprehensive intra-oral image datasets.

Deep learning (DL) has emerged as a powerful approach for overcoming the challenges in the dentistry domain, capable of autonomously extracting high-level and discriminative characteristics from a given dataset [12]. Convolutional neural networks (CNNs) have achieved significant appeal among DL approaches due to their well-established multilayer structure. CNN-based techniques for dental image processing have demonstrated outstanding performance in a variety of clinical tasks, most notably tooth detection and classification/

numbering across many dental imaging modalities, including cone-beam computed tomography (CBCT) [13] and radiography images [14]. However, the classification of tooth types in intra-oral dental images is a challenging task due to the complex and diverse structures found in these images [15]. These images have rich geometrical structure which makes it difficult to learn the discriminative features among the tooth classes. Despite some common morphological characteristics for distinguishing tooth type between individuals, there exist great variances in surface appearance with the same type of tooth [11]. Additionally, teeth classification in intra-oral images is demanding due to the inhomogeneous texture or color distribution of teeth. For example, even if the images represented the same incisor type, there are often strong differences in the directionality, granularity, or color tone of teeth. These variations make it challenging to classify the teeth accurately. Hence analysis of dental images using deep learning models has caught the attention of many researchers [16].

To overcome the aforementioned challenges, we hypothesize that the discriminative local detailed information of intra-oral images is naturally hidden in various granularity patches of the images. Multi-granularity in image classification is useful for applications such as object recognition, where objects may be present at different scales or levels of complexity. By classifying objects at multiple levels of granularity, it is possible to accurately identify objects of different sizes or shapes, which can be useful for tasks such as autonomous navigation or robotic manipulation. Thus, the underlying research work examines the effect of granularity in tooth detection and classification using intra-oral images. The multiple levels of granularity are used to specify the structural levels of the tooth. The granularity level changes based on the three tooth groups considered in this study. The Granular Intra-Oral Image (GIOI) dataset consists of three granular levels named Two Classes Granularity Level (2CGL) of the upper and lower jaw, Four Classes of Granularity Level (4CGL) of incisor, canine, premolar, and molar, and Seven Classes Granularity Level (7CGL) is used for tooth classification.

The following are the main contributions of this study:

- Modeling of faster region-convolutional neural network (FRCNN) based on three types of ResNet models for multi-granularity levels teeth classification from intra-oral images.
- Analysis of teeth detection and classification at multi-granularity levels via FRCNN.

The rest of the paper is divided into the following sections. Section II summarizes the previous research on tooth detection and classification, and granularity level classification. Section III offers the proposed methodology. The experimental findings are presented and discussed in Section IV. Section V analyses the effect of granularity levels on the tooth classification task. Section VI provides the conclusion of the research work and suggests opportunities for further study.

II. Related Work

The core of this research work is to analyze teeth detection and classification at multi-granularity levels via FRCNN from Granular Intra-Oral Image (GIOI) dataset. Therefore, to get a better understanding of the existing research work, this section presents a review of related work on the topic of (i) tooth classification using deep learning models including Faster R-CNN, AlexNet, and VGG; and (ii) the effect of multi-granularity on classification accuracy.

A. Tooth Classification using Deep Learning Models

In the context of deep learning, this study used Convolutional Neural Network (CNN). A CNN is a type of Artificial Neural Network (ANN) that is commonly used in Deep learning for image, text, object recognition, and classification. CNNs have been widely used in computer vision tasks such as object detection, face recognition, and image segmentation [17]. They have also been applied in other domains, such as natural language processing and dentistry.

In an automated diagnostic procedure, classifying teeth is a crucial task. Researchers have examined the classification task using a small sample of tooth periapical pictures; one such study was carried out by Zhang et al. [18] employed a cascade network structure for the automated identification of 32 teeth positions. Their approach utilized multiple CNNs as the fundamental modules and achieved an F1-Score, precision, and recall of 80.4, 80.3, and 80.6, respectively. Oktay [19] introduced a CNN-based method for tooth detection in dental panoramic X-ray images. The approach accurately determines the potential positions of three tooth types (incisors, premolars, and molars), achieving a remarkable accuracy level of over 0.92. Similarly to this, Miki et al. [13] used 52 CBCT images to categorize teeth into seven types based on their location. AlexNet was employed as the CNN structure in this study, and it achieved a classification accuracy of 88.8%. In research on automated detection and labeling of 2D teeth, Zhang et al. [18] and Chen et al. [20] used CNN to identify teeth in periapical radiographs, and experimental findings indicated that their precision rates were 95% and 90% respectively. These findings ensured the importance of deep learning models such as AlexNet [13] [19] and VGG [18] in achieving accurate and efficient detection for automated dental charting and proper surgical and treatment planning.

Another model based on GoogleNet, a fully convolutional network (FCN) was proposed to detect teeth by Muramatsu et al. [21]. The classification of teeth by type (i.e., incisors, canines, premolars, and molar) and tooth condition was performed using a ResNet-50-based pre-trained network. Görürgöz et al. [22] applied transfer learning with a pre-trained GoogLeNet Inception v3 CNN and developed an algorithm consisting of jaw classification, region detection, and final classification models. The proposed algorithm achieved an F1 score, precision, and sensitivity of 0.8720, 0.7812, and 0.9867, respectively. These findings demonstrate the potential of CNN algorithms for efficient and precise tooth detection and numbering in dental imaging, which could lead to more reliable diagnoses and treatments.

These studies show that CNN models are trained on a large dataset of dental images, where each image is labeled with the coordinates or bounding boxes representing the location of each tooth. CNN learns to recognize patterns and features that differentiate teeth from the background and other structures in the image [23]. It's important to note that different studies proposed specific modifications or variations of CNN architectures to optimize tooth detection and classification performance such as AlexNet, Faster R-CNN, GoogLeNet, RCNN, and ResNet.

Tooth detection and classification using Faster R-CNN (Faster Region-based Convolutional Neural Network) [24] is an area of research that focuses on automating the process of identifying and categorizing teeth in dental images. The effectiveness of Faster R-CNN in tooth numbering and identifying dental cavities on oral radiographs was studied by Tuzoff et al. [25]. They used the Faster R-CNN architecture for teeth detection using 1,352 adult panoramic radiographs. A two-stage system was proposed, in which faster R-CNN is used to detect the teeth followed by a VGG-16 network to identify and number. Nonetheless, the study encountered misclassification errors resulting from similarities between adjacent teeth. Chen et al. [20] suggested employing Faster R-CNN for tooth detection and recognition in dental periapical films. The test dataset demonstrated precision and recall values of over 90%. However, the study faced challenges due to complications such as missing teeth and root canal treatments in the images from regular clinical work. Mahdi et al. [26] presented an automatic teeth recognition model that leverages the Faster R-CNN technique based on the residual network. This model represents a significant step forward in dental image analysis, achieving impressive results with high mean Average Precision (mAP) scores of 0.974 and 0.981 for ResNet-50 and ResNet-101, respectively. In a similar vein, Bilgir et al. [27] developed a Faster R-CNN model that automated tooth numbering over a dataset of 2,482 panoramic radiographs with a precision of 0.96. Estai et al. [28] proposed a three-step method for automatically detecting and counting teeth in digital orthopantomography (OPG) pictures. They used U-Net, Faster R-CNN, and VGG-16 CNN models. The results showed that it had a high recall and precision score of 0.99 for tooth detection and 0.98 for tooth numbering, indicating its potential importance in general dentistry and forensic medicine applications.

It is concluded that Faster R-CNN is sensitive to objects with missing features i.e., broken tooth [20], overlapping [29], occlusion [25], blur, and noise [30]. These issues distort the fine details of the tooth [26] [31]. This leads to low classification accuracy [32] and limits the model's generalization ability on other imaging modalities or dental issues [13]. Despite these issues, there are various advantages to using Faster R-CNN for tooth identification and classification tasks. It enables exact tooth localization in dental pictures while effectively handling size and aspect ratio variations [33]. Research has shown that Faster R-CNN accurately identifies the position of teeth with a high IOU value [20]. The model exhibits significant potential in dental image processing tasks, assisting in dental diagnosis, treatment planning, and various other applications within the dental field [34].

B. Effect of Multi-Granularity on Classification Tasks

The ability to analyze or portray data at numerous levels of detail or abstraction is referred to as multi-granularity. Multi-granularity is important for a range of applications since it provides for more nuanced and flexible data processing. It has been the focus of recent studies in fields such as scene classification [35], land change detection [36], and brain image analysis [37]. This section reviews relevant research on the role of granularity in various classification problems.

Several researchers have recently investigated the use of multi-granularity in medical image classification, employing a range of approaches and techniques. Within these investigations, Wu et al. [38] focused their research on lung nodule classification. Their study evaluated a novel approach using a publicly available lung nodule dataset, and the results demonstrated that employing the multi-granularity approach resulted in enhanced classification accuracy. In addition, Wang et al. [39] provided a unique method for producing generalized visual representations for medical images using multi-granularity cross-modal alignment. To assess the effectiveness of their approach, they used a variety of medical imaging datasets, including chest X-rays and mammograms. The results showed that the proposed model outperformed existing methods in a variety of classification and retrieval tasks, highlighting the effectiveness of multi-granularity cross-modal alignment in acquiring comprehensive visual representations for medical images. Wang et al. [36] suggested a multi-granularity framework for extracting latent ontologies from remote sensing datasets, which they tested in six different scenarios. The results showed that combining three granularity levels produced the best results, with the second level of granularity providing the highest accuracy. On the other hand, the third level of granularity exhibited comparatively lower accuracy. Furthermore, the study highlighted that fine-scale cropping increased classification accuracy whereas excessive cropping degraded performance. Additionally, Zuo et al. [40] presented an innovative method for fine-grained crop disease classification that combines multi-granularity feature aggregation with self-attention and spatial reasoning to improve accuracy. The evaluation outcomes showcase the effectiveness of incorporating multi-granularity feature aggregation, self-attention, and spatial reasoning in the field of fine-grained crop disease classification.

In the past few years, significant progress has been made in deep learning-based image classification and object re-identification. For instance, Ouyang et al. [41] introduced a hybrid methodology that merges a CNN with a modified capsule network for remote sensing image classification. Their model incorporated spatial-spectral attention and multi-granularity features, allowing it to effectively capture precise spatial and spectral information. Likewise, Tu et al. [42] introduced the Multi-granularity Mutual Learning Network (MMNet) for object re-identification. The MMNet integrates multiple modules to effectively learn distinctive features across varying visual granularities. By capturing diverse discriminative local features from multiple granularities, the MMNet demonstrated superior performance compared to

previous approaches. Wu et al. [43] presented a CNN-based image classification approach that takes advantage of multi-granularity features. The fundamental goal of this research is to incorporate the concept of hierarchical structure categorization and to investigate the incorporation of granularity computing theory in deep learning. The experimental findings revealed the enhanced model's usefulness, with higher image classification accuracy and superior generalization capabilities. Chen et al. [44] investigated the impact of label granularity on CNN classification performance. Experiments on several datasets revealed that training with fine-grained labels improved the accuracy of classifying coarse-grained classes, in contrast to training with coarse-grained labels. According to their research, while training a CNN for natural images, using fine-grained labels outperforms using coarse-grained labels from the same dataset. The utilization of fine-grained labels enables the network to learn more intricate and specific features. Zhu et al. [45] introduced a novel methodology for few-shot learning that incorporates multi-granularity techniques. The proposed approach was tested on many few-shot learning datasets, including CIFAR-FS and mini-ImageNet. The outcomes substantiated the efficacy of multi-granularity episodic contrastive learning in the context of few-shot learning.

The presented review identifies that granularity level classification leads to improvement in computational efficiency, adaptation capability (even if shallow models and the small dataset is used), and extracting fine-grained feature [43]. Additionally, the multi-granularity technique is less prone to overfitting when compared to deep networks and offers better generalization and increased classification accuracy [46] [44] [43]. However, despite these benefits, granular-level classification studies in the domain of tooth classification are seldom seen and its relevance in this domain needs to be explored.

III. Proposed Methodology

This section introduces the methodology used to achieve the main aim of this study which is to analyze the effect of teeth detection and classification at multi-granularity levels using FRCNN. The overview of the proposed method to detect teeth at multi-granularity levels using FRCNN is presented in Fig. 1. The detection pipelines as shown in Fig. 1 perform four essential steps: data collection, data pre-processing, modeling detection, and finally providing results and discussion.

The following subsections will provide the details of data collection and pre-processing criteria including inclusion and exclusion criteria, ground truth marking scheme and consequent labeling procedure, and identification and implementation of label-preserving data augmentation methods. Additionally, the proposed CNN model and model evaluation will be introduced:

A. Dataset Preparation

A significant challenge in the advancement and practical adoption of DL models lies in acquiring adequately large, curated, and representative training data, along with expert annotations. In this section, the fundamental steps for preparing a dental imaging dataset for addressing the issue of tooth classification in Intra-Oral imaging using deep learning models are described.

1) Data acquisition: With the absence of a publicly available dataset, this study proposes the GIOI dataset that offers three teeth classification granularity levels as shown in Fig. 2, i.e., Two Classes of Granularity Level (2CGL) of maxilla and mandible; Four Classes of Granularity Level (4CGL) of incisor, canine, premolar, and molar; and Seven Classes Granularity Level (7CGL) of teeth numbering In the proposed GIOI dataset development phase, the Advanced Medical and Dental Institute at University Sains Malaysia (USM) and University Technology PETRONAS (UTP) have collaborated to develop the GIOI dataset. These images represent subjects from different age groups and genders. The images are also captured at different distances and illumination levels to present rich feature diversity.

2) Data Pre-Processing: The first stage in pre-processing was to filter the dataset by setting the inclusion and exclusion criteria. The images were visually analyzed, and the images containing gum or cavity diseases are extracted. Additionally, images having missing teeth or wisdom teeth are also excluded. Table I and Table II display the inclusion and exclusion criteria that were used for the data pre-processing.

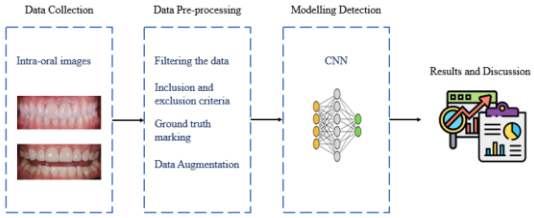

Fig. 1. Overview of the proposed tooth detection model.

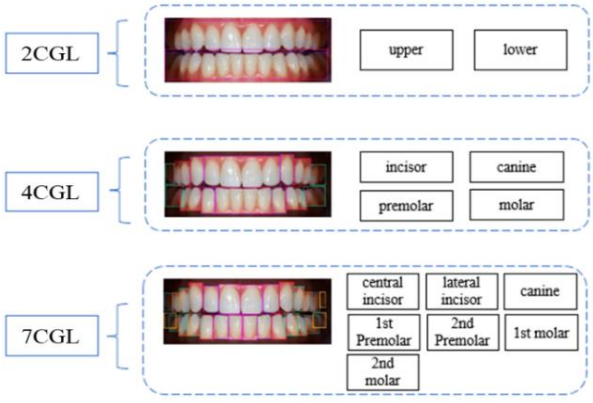

Fig. 2. Teeth classification granularity levels in the GIOI.

TABLE I. Inclusion Criteria

Inclusion Criteria
Adults between 18-50 ages were included.
Both male and female.
Stained teeth.
Different orientations (left, right, upper, lower).
The gap between teeth.

TABLE II. EXCLUSION CRITERIA

Exclusion Criteria
Crowded tooth.
Wisdom tooth.
Missing and broken tooth.
Tongue.
The tooth has braces.
Teeth have gum or cavity diseases.

The GIOI dataset contained 550 images and seven-tooth classes (Central Incisor, Lateral Incisor, Canine, 1st Premolar, 2nd Premolar, 1st molar, 2nd molar). The current study used the ISO standard tooth numbering system to identify each tooth with a unique label [7]. The VGG Image Annotator (VIA) [47] web application has been used for annotating and labeling the training set samples. VIA is an open-source software that allows human annotators to define and describe regions in an image.

3) Data augmentation: the data set contained unequal samples, that is, the number of different types of samples was different. To enhance the dataset, data augmentation was used as an option [48]. Data augmentation effectively expands the dataset size and quality. The effectiveness of data augmentation for dental image augmentation was assessed by including image mixing, geometric transformation, transforms, and kernel filters [49]. As a result, 2,260 augmented images were acquired for training. Table III contains the specifics of this assessment.

TABLE III. DATA AUGMENTATION METHODS FOR DENTAL IMAGES

Type of Augmentation	Post Augmentation Observation	Label Preservation	Selection Status
Vertical flip	Tooth visual attributes do not remain intact.	No	Rejected
Horizontal flip	Tooth visual attributes remain intact.	Yes	Selected
ChannelShuffle	Resulting in an image that is not a true representative of a real-world scenario.	No	Rejected
Brightness and contrast	Introduces a wide range of illuminations.	Yes	Selected
Noise injection	Improves the model's generalization ability	Yes	Selected
Cropping	This may result in the loss of distinguishable tooth features	No	Rejected
Motion blur	Simulates the possible sudden motion of the optical sensor/subject in a real-world scenario.	Yes	Selected
RandomGridShuffle	Tooth visual attributes do not remain intact.	No	Rejected
Histogram equalization	Improves the image's contrast level.	Yes	Selected
image compression	This keeps the resolution of an image	Yes	Selected

B. Model Architecture

In this paper, the Faster R-CNN architecture is supported by three types of ResNet [50] network: ResNet-50, ResNet-101, and ResNet-152 as backbone models. Fig. 3 shows the FR-CNN-ResNet model. The first phase of the model includes the backbone models that generate the feature map. The second phase is the region proposal network (RPN), for identifying areas of an input image that most likely contain a region of interest. The last phase includes the detection network. The RPN generates region proposals (bounding boxes) for potential objects in an image, while the detection network classifies the proposals and refines their bounding boxes. The RPN is a fully convolutional network that is trained to predict abjectness scores and bounding box offsets at each position in an image. It uses a sliding window approach to generate region proposals, which are then passed to the detection network. When the RPN generates a set of candidate regions, each region is represented by a fixed-size feature map, which can be of different sizes depending on the size of the input image and the region proposal. However, the detection network that processes these regions requires a fixed-size input to apply convolutional layers.

ROI (Region of Interest) pooling addresses this discrepancy by dividing the fixed-size feature map for each region proposal into a fixed number of equally sized sub-windows and then applying a max pooling operation to each sub-window to produce a fixed-size output. The output of the ROI pooling operation is a feature map of fixed size that can be fed into the detection network. The detection network in Faster R-CNN is based on the Fast R-CNN [51] architecture, which consists of two main components: a convolutional feature extractor and a set of fully connected layers for object classification and bounding box regression. It takes the region proposals generated by the region proposal network (RPN) as input and produces the final object detection results.

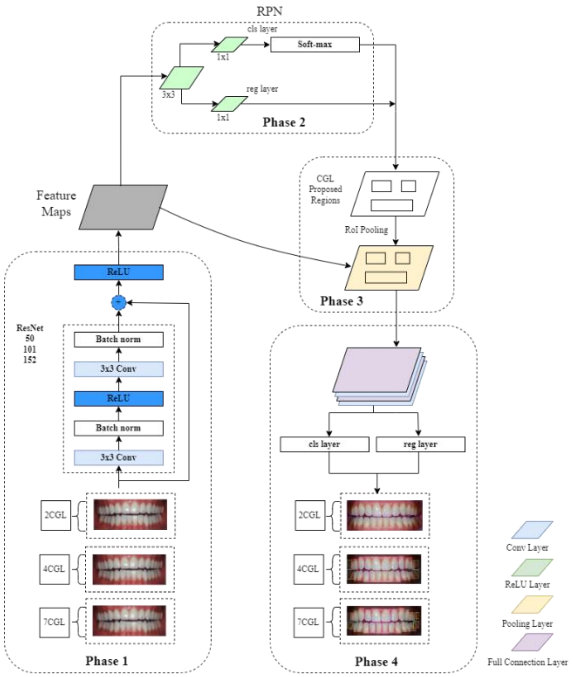

Fig. 3. FR-CNN-ResNet model.

The main activities of the FR-CNN-ResNet algorithm are presented in the following nine steps:

Step 1: The system fed the images to the backbone ResNet50, ResNet101, or ResNet101 models.

Step 2: The backbone models extract the features from the images.

Step 3: The RPN takes the feature maps as input and generates a set of object proposals, which are regions in the image that are likely to contain objects.

Step 4: The proposed regions generated by the RPN in Step 3 are passed through ROI pooling, which divides the fixed-size feature map for each region proposal into a fixed number of equally sized sub-windows.

Step 5: The detection network takes the fixed-size feature maps generated in Step 4 by the ROI pooling layer as input and produces the final object detection results.

Step 6: The final output is obtained by applying non-maximum suppression to remove duplicate predictions and keep only the most confident detections.

C. Model Evaluation and Performance Measures

Average Precision (AP) [50] is a popular evaluation metric for object detection tasks that measures the accuracy of the predicted object bounding boxes. AP is calculated based on a precision-recall curve that summarizes the trade-off between precision and recalls for different object detection thresholds. The average precision value is computed for recall values ranging from 0 to 1.

Precision [51] is a performance metric used in object detection to measure the proportion of correct positive detections out of all the positive detections made by the network. Precision measures how accurate the algorithm is in detecting objects. The given equation was used to calculate precision:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

Recall [50] is a performance metric used in object detection to measure the proportion of actual positive detections out of all the positive instances present in the dataset. In other words, recall measures how well the algorithm can detect all the objects present in the image. The following equation used for recall calculation:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

The F1 score is a performance metric used in object detection that combines precision and recalls into a single score. The F1 score provides a balanced view of the network's accuracy by considering both the number of correct detections and the number of missed detections. It is defined by the following equation:

$$F1\ score = \frac{2 \times (precision \times recall)}{(precision + recall)} \quad (3)$$

IV. Results and Discussion

The presented study evaluates the effect of granularity on tooth detection and classification using FR-CNN-ResNet models. The GIOI dataset, consisting of three teeth classification granularity levels, is considered in testing the proposed FR-CNN-ResNet model. The total number of epochs was set as 100 for all different backbone ResNet models. The batch size for all Faster R-CNN backbone models was set as 2. In addition, two learning rate values were used, i.e., 0.001 and 0.0001. This section presents the experimental results for all three classification granularity levels separately.

A. Case 1: Two Classes Granularity Level (2CGL)

1) The two classes' granularity level (2CGL) consists of two tooth classes i.e.: upper and lower. A total of 2,260 images containing 2078 upper and 1956 lower were used to train the models in seven different experiments. A total of 107 images are used to test the models. It has been identified that the lower learning rate during training of Faster RCNN variants resulted in lower mean average precision during the testing of all such models. This indicates the unsuitability of a smaller learning rate for 2CGL tooth classification. For all F-RCNN variants, the optimal accuracy was achieved using a constant learning rate of 0.001.

As depicted in Fig. 4, the highest average precision of 0.95 and 0.93 for the upper and lower tooth, respectively, is achieved by the FR-CNN-ResNet-50. With FR-CNN-ResNet-101, the average precision for upper and lower teeth is observed to be the lowest among all types of FR-CNN models. With a deeper backbone, i.e., ResNet-152, no significant improvement is observed by the FR-CNN-ResNet-152 model in the average precision of target classes. This performance indicates that at 2CGL, the FR-CNN-ResNet-50 model is the best choice. Similarly, the highest mAP of 0.94 was achieved by FR-CNN-ResNet-50. The lowest mAP of 0.742 is yielded by the FR-CNN-ResNet-101 model trained on a lower learning rate. This confirms that a lower learning rate and deeper backbones are not optimal for optimal classification at the 2CGL level.

The model exhibited FR-CNN-ResNet-50 achieves a competitive and high mAP of 0.94. As presented in Table IV, the model also exhibited perfect or near-to-perfect recalls for upper and lower teeth classification results. Additionally, the best F1 scores for upper and lower teeth classification are equal to 0.96 and 0.94 for the FR-CNN-ResNet-50 model. This performance indicates that this model for 2CGL is ideal as it is trained quickly and generates very competitive results compared to other models.

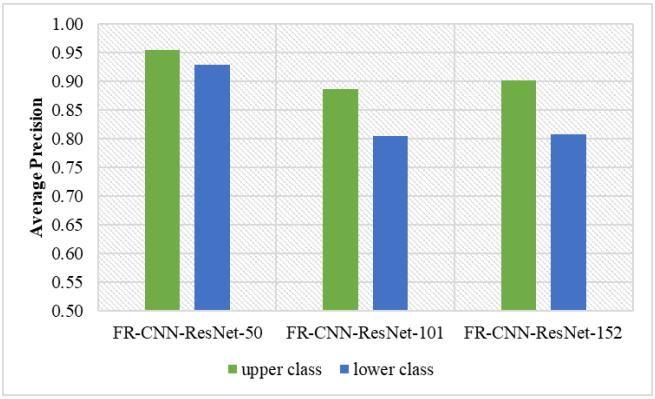

Fig. 4. 2CGL average precision comparison of the models.

TABLE IV. 2CGL AVERAGE PRECISION, RECALL, AND F1 SCORES FOR EACH MODEL

	AP			Recall			F1-Score		
2CGL Classes	***FR-CNN-ResNet-50***	***FR-CNN-ResNet-101***	***FR-CNN-ResNet-152***	***FR-CNN-ResNet-50***	***FR-CNN-ResNet-101***	***FR-CNN-ResNet-152***	***FR-CNN-ResNet-50***	***FR-CNN-ResNet-101***	***FR-CNN-ResNet-152***
Upper	**0.95**	0.89	0.90	0.97	0.92	0.93	**0.96**	0.90	0.92
Lower	**0.93**	0.81	0.81	0.96	0.85	0.86	**0.94**	0.83	0.83

Overall, the average precision of upper teeth remains higher as compared to lower teeth. This can be attributed to labeling precision as naturally lower teeth are occluded by upper teeth. For this reason, the bounded boxes for upper teeth are more precise as compared to lower teeth as it contains some part of upper teeth. FR-CNN-ResNet is generally good at detecting large objects because it uses region proposals to identify potential object locations and then applies a classifier to each region proposal to determine the presence and location of an object. The RPN in FR-CNN-ResNet generates region proposals by sliding a small network over the convolutional feature map output by the backbone network. The size of the sliding window is fixed, and the stride can be adjusted to control the region proposal density. Because of this mechanism, FR-CNN-ResNet can effectively detect large objects but may struggle with detecting small objects due to the limitations of the region proposal mechanism.

B. Case 2: Four Classes Granularity Level (4CGL)

1) This section presents the results of the granular level two (4CGL) classification, which consists of four classes, i.e., Incisor, Canine, Premolar, and Molar. A total of 2,260 images containing 4091 incisors, 8138 canines, 7940 premolars, and 6564 molars were used to train the models in seven different experiments and 107 images were used for testing. Within FR-CNN-ResNet models, the learning rate again played an important role. With a lower learning rate, i.e., 0.0001, mAP remained low, as compared to the mAP of the model trained with a higher learning rate of 0.001.

As presented in Fig. 5, the highest average precision (AP) of 0.849 is produced by the FR-CNN-ResNet-50 model for the incisor tooth class, followed by the Canine, Premolar, and Molar tooth class which achieved an AP of 0.82, 0.73 and 0.58 respectively. The following factors contribute to higher average precision for incisor class, (i) no occlusion, (ii) large size, and (iii) high illumination. As discussed in Table V, the FR-CNN-ResNet-50 model also has the highest recall and F1 values for all classes. This result also concludes that FR-CNN-ResNet-50 is less sensitive to occlusion, object size, and low illumination.

As shown in Table V, the highest mAP of 0.74 was observed by FR-CNN-ResNet-50. The other models are significantly behind where FR-CNN-ResNet-101 and FR-CNN-ResNet-152 achieved mAP of 0.71 and 0.63, respectively. This result indicates that for 4CGL, FR-CNN-ResNet-50 is the best model among the three for teeth classification and detection.

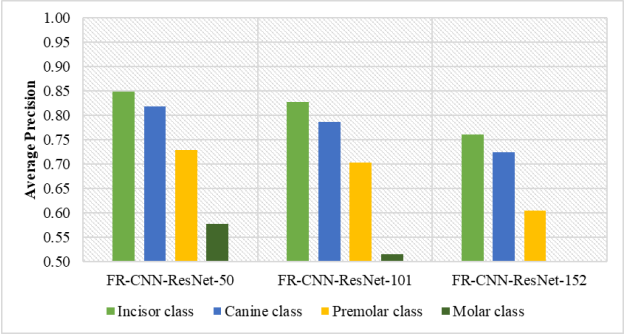

Fig. 5. 4CGL average precision comparison of the models.

TABLE V. 4CGL AVERAGE PRECISION, RECALL, AND F1 SCORES FOR EACH MODEL

	AP			Recall			F1		
4CGL Classes	***FR-CNN-ResNet-50***	***FR-CNN-ResNet-101***	***FR-CNN-ResNet-152***	***FR-CNN-ResNet-50***	***FR-CNN-ResNet-101***	***FR-CNN-ResNet-152***	***FR-CNN-ResNet-50***	***FR-CNN-ResNet-101***	***FR-CNN-ResNet-152***
Incisor	**0.85**	0.83	0.76	**0.88**	0.87	0.80	**0.87**	0.85	0.78
Canine	**0.82**	0.79	0.73	**0.86**	0.83	0.78	**0.84**	0.81	0.75
Premolar	**0.73**	0.70	0.60	**0.78**	0.75	0.67	**0.75**	0.73	0.63
Molar	**0.58**	0.52	0.41	**0.66**	0.59	0.52	**0.62**	0.55	0.46

These results conclude that by using a pre-trained ResNet-50 as the backbone network, the Faster R-CNN model can leverage the high-level features learned by ResNet-50 to accurately classify medical images. Moreover, the ResNet-50 architecture has a deep network structure that allows it to learn complex features in medical images, including subtle differences between images that may be indicative of different conditions or diseases. This makes it particularly effective in medical image classification tasks where subtle differences can be critical in diagnosing a disease. However, the choice of backbone architecture depends on the specific task and dataset, and other backbones such as ResNet-101 or ResNet-152 may perform better in some scenarios.

C. Case 3: Seven Classes Granularity Level (7CGL)

This section presents the results of the Seven Classes Granularity Level (7CGL) classification, which consists of seven classes, i.e., Central Incisor, Lateral Incisor, Canine, 1st Premolar, 2nd Premolar,1st molar, and 2nd molar. This level of granularity creates three major issues, (i) objects with low illumination conditions, (ii) large variation in object size, and (iii) class imbalance. A total of 2,260 images were used to train the models in seven different experiments, and 107 images were used for testing. Within FR-CNN models, the learning rate again played an important role. With a lower learning rate, i.e., 0.0001, mAP remained low, compared to the mAP of the model trained on a higher learning rate of 0.001.

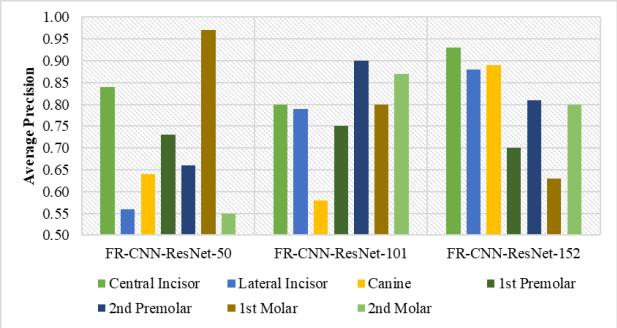

Fig. 6. 7CGL average precision comparison of the models.

A comparative performance analysis is presented in Fig. 6 which highlights that the performance of all models decreases as the target tooth is further away from the central position. However, in all performance measures, FR-CNN-ResNet-50 remains the best-performing model.

As presented in Table VI, the average precision for all models gradually decreased as tooth location moved from front to behind. Overall, FR-CNN-ResNet-101 yielded the lowest average precision score, while FR-CNN-ResNet-50 again emerged as the top-performing model. FR-CNN-ResNet-50 recall values also remained high for the central incisor and lateral incisor. The model produced a perfect recall value. Considering the F1 scores of all three models, it is again evident that at the 7CGL level, FR-CNN-ResNet-50 has the highest F1 scores for all seven classes. However, the model's performance significantly decreased for smaller and occluded teeth such as 2nd Premolar, 1st Molar, and 2nd Molar.

Overall mAP of FR-CNN-ResNet-50 remained highest at 0.69, followed by FR-CNN ResNet-101 and FR-CNN ResNet-50 with mAP of 0.55 and 0.62, respectively. One possible reason for the low mAP of FRCNN can be attributed to its limitation with detecting small objects, especially if large objects surround them, as the region proposal network may overlook.

FR-CNN-ResNet, like many object detection models, can struggle to detect small objects as the size of the RPN anchors, which are the pre-defined boxes used to search for objects in an image, may be too large relative to the size of the small objects being searched for. This means that the RPN may fail to generate proposals that accurately localize small objects. In the case of occluded objects, the RPN may still generate proposals that partially or completely overlap with the occluded object, allowing the CNN to classify and localize the object within the proposal. However, the accuracy of object detection for occluded objects may still be affected by the extent of occlusion and the quality of the proposals generated by the RPN. In the case of objects with low illumination, the features extracted from the image may be less informative due to reduced contrast and detail in the image. This can make it more difficult for the model to distinguish the object from the background or other objects in the scene.

TABLE VI. 7CGL AVERAGE PRECISION, RECALL, AND F1 SCORES FOR EACH MODEL

	AP			Recall			F1-Score		
7CGL Classes	*FR-CNN-ResNet-50*	*FR-CNN-ResNet-101*	*FR-CNN-ResNet-152*	*FR-CNN-ResNet-50*	*FR-CNN-ResNet-101*	*FR-CNN-ResNet-152*	*FR-CNN-ResNet-50*	*FR-CNN-ResNet-101*	*FR-CNN-ResNet-152*
Central Incisor	**0.84**	0.77	0.82	**0.87**	0.83	0.85	**0.85**	0.80	0.84
Lateral Incisor	0.81	0.73	0.76	0.86	0.81	0.82	0.84	0.77	0.79
Canine	0.80	0.72	0.76	0.85	0.79	0.81	0.82	0.75	0.78
1st Premolar	0.72	0.59	0.65	0.79	0.69	0.73	0.75	0.63	0.69
2nd Premolar	0.64	0.39	0.51	0.73	0.54	0.62	0.68	0.45	0.56
1st Molar	0.57	0.36	0.45	0.69	0.47	0.57	0.63	0.41	0.50
2nd Molar	**0.49**	0.30	0.38	0.61	0.42	0.50	0.54	0.35	0.43

V. ANALYSIS OF THE EFFECT OF GRANULARITY LEVELS ON TOOTH CLASSIFICATION TASK

In this study, three different models of FR-CNN ResNet were implemented for three granularity level cases named 2CGL, 4CGL, and 7CGL to demonstrate the influence of using different granularities in tooth classification. For all FR-CNN variants, the optimal performance is achieved using a constant learning rate of 0.001. Within the FR-CNN-ResNet models, the learning rate played an important role in which, with a lower learning rate, i.e., 0.0001, mAP remained low, as compared to the mAP of models trained on a higher learning rate of 0.001. This result confirms that a lower learning rate and deeper backbones are not optimal for classification at 2CGL, 4CGL, and 7CGL cases.

For an individual granularity level, the first granularity level achieves the best classification accuracy while the third is the least accurate. The best improvement can be observed in the 2CGL with the FR-CNN-ResNet-50 model, where the mAP result is 0.94 better than FR-CNN-ResNet-101, which achieved the lowest mAP of 0.85. And FR-CNN-ResNet-50 model remained significantly higher than other models in 4CGL, which achieved an mAP of 0.74. For 7CGL, the performance of all models decreases as the target tooth is further away from the central position. Overall, FR-CNN-ResNet-101 yielded the lowest average precision score, while FR-CNN-ResNet-50 again emerged as the top-performing model by achieving an mAP of 0.69.

These results indicate that with the largest granularity level as shown in 2CGL and 4CGL, the tooth structure and the tooth features are clear. Therefore, FR-CNN-ResNet has the strong ability to exploit features such as shape and texture features. The following factors contribute to higher average precision

for 2CGL and 4CGL, (i) no occlusion, (ii) large size, and (iii) high illumination. As the level of drowsiness becomes more detailed in 7CGL, it becomes increasingly difficult to achieve a high level of precision in detecting and classifying teeth due to the intricate structure of teeth [52] [53]. Furthermore, it is difficult for FR-CNN-ResNet to identify objects from low-resolution images as the features extracted from the image may be less informative due to reduced contrast and detail in the image [53]. This can make it more difficult for the model to distinguish the object from the background or other objects in the scene [21]. FR-CNN-ResNet can learn and recognize features of objects even when they are partially occluded, due to the use of shared convolutional layers that extract features from different parts of the image [54] [55]. However, the accuracy of object detection for occluded objects may still be affected by the extent of occlusion and the quality of the proposals generated by the RPN [54].

VI. Conclusion

The automatic detection and classification of teeth in intra-oral dental images are crucial for medical treatment and forensic identification. However, due to the complexity of the problem and limitations in the size of available data, this task remains challenging. To overcome such challenges, this paper investigates the intriguing problem that how granularity impacts the performance of CNN-based object detection and classification models. A Faster Region-Convolutional Neural Network based on ResNet models is proposed for teeth detection and classification at multi-granularity levels from the GIOI dataset. Three different ResNet backbones, i.e., ResNet-50, Res-Net101, and ResNet-152 were evaluated. The evaluation results showed that the proposed FR-CNN-ResNet model is appropriate for teeth classification at three granular levels named, 2CGL, 4CGL, and 7CGL. Additionally, it was revealed that the FR-CNN-ResNet-50 performed better than the FR-CNN-ResNet-101 and FR-CNN-ResNet-152 at each of the three granular levels, where the FR-CNN-ResNet-50 achieved mAP of 0.94, 0.74 and 0.69 at 2CGL, 4CGL, and 7CGL respectively. Overall, it is concluded that multi-granular approaches in intra-oral dental image analysis have the potential to capture significant details and improve the accuracy of automated detection and classification tasks, which can aid in medical treatment and forensic identification.

As a practical implementation, the integration of Faster R-CNN with additional networks will extend its capabilities beyond tooth detection and numbering. It will enable predictions regarding the presence of various dental conditions, including orthodontic issues, tooth fillings, and the overall assessment of dental health to facilitate the patient and dentist.

This study has two known limitations which will be addressed in future work. Firstly, for deep learning methods, large-curated datasets will be used to further improve the performance parameters. Secondly, only a few cases of the 3rd molar tooth class were identified during the dataset generation procedure, thus resulting in removing the 3rd molar class. Further in the future, a yolo-based model will be proposed to preserve topological information and the precise spatial location of pixels for each tooth.

Acknowledgment

The authors would like to acknowledge the Ministry of Higher Education Malaysia through the Dental Simulation and Virtual Learning Research Excellence Consortium (KKP Programme) JPT(BPKI)1000/016/018/25 Jld. 2(2).

References

[1] Z. Cui et al., “A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images,” Nat. Commun., vol. 13, no. 1, Art. no. 1, Apr. 2022, doi: 10.1038/s41467-022-29637-2.

[2] H. Benzian, R. Watt, Y. Makino, N. Stauf, and B. Varenne, “WHO calls to end the global crisis of oral health,” Lancet Lond. Engl., vol. 400, no. 10367, pp. 1909–1910, Dec. 2022, doi: 10.1016/S0140-6736(22)02322-4.

[3] J. Kühnisch, O. Meyer, M. Hesenius, R. Hickel, and V. Gruhn, “Caries Detection on Intraoral Images Using Artificial Intelligence,” J. Dent. Res., vol. 101, no. 2, pp. 158–165, Feb. 2022, doi: 10.1177/00220345211032524.

[4] J. Zhang, X. Li, Z. Gao, and J. Chen, “IMAGE DETECTION OF DENTAL DISEASES BASED ON DEEP TRANSFER LEARNING,” in 2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE), Nov. 2021, pp. 774–778. doi: 10.1109/ICAICE54393.2021.00151.

[5] T. Dhake and N. Ansari, “A Survey on Dental Disease Detection Based on Deep Learning Algorithm Performance using Various Radiographs,” in 2022 5th International Conference on Advances in Science and Technology (ICAST), Dec. 2022, pp. 291–296. doi: 10.1109/ICAST55766.2022.10039566.

[6] D. Verma, S. Puri, S. Prabhu, and K. Smriti, “Anomaly detection in panoramic dental x-rays using a hybrid Deep Learning and Machine Learning approach,” in 2020 IEEE REGION 10 CONFERENCE (TENCON), Nov. 2020, pp. 263–268. doi: 10.1109/TENCON50793.2020.9293765.

[7] Z. Cui, C. Li, and W. Wang, “ToothNet: Automatic Tooth Instance Segmentation and Identification From Cone Beam CT Images,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019, pp. 6361–6370. doi: 10.1109/CVPR.2019.00653.

[8] G. H. Kwak et al., “Automatic mandibular canal detection using a deep convolutional neural network,” Sci. Rep., vol. 10, no. 1, Art. no. 1, Mar. 2020, doi: 10.1038/s41598-020-62586-8.

[9] E. D. Fadhillah, P. C. Bramastagiri, R. Sigit, S. Sukaridhoto, A. Brahmanta, and B. S. B. Dewantara, “Smart Odontogram: Dental Diagnosis of Patients Using Deep Learning,” in 2021 International Electronics Symposium (IES), Sep. 2021, pp. 532–537. doi: 10.1109/IES53407.2021.9594027.

[10] S. AbuSalim, N. Zakaria, N. Mokhtar, S. A. Mostafa, and S. J. Abdulkadir, “Data Augmentation on Intra-Oral Images Using Image Manipulation Techniques,” in 2022 International Conference on Digital Transformation and Intelligence (ICDI), Dec. 2022, pp. 117–120. doi: 10.1109/ICDI57181.2022.10007158.

[11] C. Wu et al., “Model-based teeth reconstruction,” ACM Trans. Graph., vol. 35, no. 6, p. 220:1-220:13, Dec. 2016, doi: 10.1145/2980179.2980233.

[12] S. Tian et al., “A Dual Discriminator Adversarial Learning Approach for Dental Occlusal Surface Reconstruction,” J. Healthc. Eng., vol. 2022, p. e1933617, Apr. 2022, doi: 10.1155/2022/1933617.

[13] Y. Miki et al., “Classification of teeth in cone-beam CT using deep convolutional neural network,” Comput. Biol. Med., vol. 80, pp. 24–29, Jan. 2017, doi: 10.1016/j.compbiomed.2016.11.003.

[14] A. Betul Oktay, “Tooth detection with Convolutional Neural Networks,” in 2017 Medical Technologies National Congress (TIPTEKNO), Oct. 2017, pp. 1–4. doi: 10.1109/TIPTEKNO.2017.8238075.

[15] R. Ragodos et al., “Dental anomaly detection using intraoral photos via deep learning,” Sci. Rep., vol. 12, no. 1, Art. no. 1, Jul. 2022, doi: 10.1038/s41598-022-15788-1.

[16] S. AbuSalim, N. Zakaria, M. R. Islam, G. Kumar, N. Mokhtar, and S. J. Abdulkadir, “Analysis of Deep Learning Techniques for Dental

Informatics: A Systematic Literature Review," Healthcare, vol. 10, no. 10, Art. no. 10, Oct. 2022, doi: 10.3390/healthcare10101892.

[17] L. Alzubaidi et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," J. Big Data, vol. 8, no. 1, p. 53, 2021, doi: 10.1186/s40537-021-00444-8.

[18] K. Zhang, J. Wu, H. Chen, and P. Lyu, "An effective teeth recognition method using label tree with cascade network structure," Comput. Med. Imaging Graph., vol. 68, pp. 61–70, Sep. 2018, doi: 10.1016/j.compmedimag.2018.07.001.

[19] A. Betul Oktay, "Tooth detection with Convolutional Neural Networks," in 2017 Medical Technologies National Congress (TIPTEKNO), Oct. 2017, pp. 1–4. doi: 10.1109/TIPTEKNO.2017.8238075.

[20] H. Chen et al., "A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films," Sci. Rep., vol. 9, no. 1, Art. no. 1, Mar. 2019, doi: 10.1038/s41598-019-40414-y.

[21] C. Muramatsu et al., "Tooth detection and classification on panoramic radiographs for automatic dental chart filing: improved classification by multi-sized input data," Oral Radiol., vol. 37, no. 1, pp. 13–19, Jan. 2021, doi: 10.1007/s11282-019-00418-w.

[22] C. Görürgöz et al., "Performance of a convolutional neural network algorithm for tooth detection and numbering on periapical radiographs," Dentomaxillofacial Radiol., vol. 51, no. 3, p. 20210246, Mar. 2022, doi: 10.1259/dmfr.20210246.

[23] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," Insights Imaging, vol. 9, no. 4, Art. no. 4, Aug. 2018, doi: 10.1007/s13244-018-0639-9.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." arXiv, Jan. 06, 2016. doi: 10.48550/arXiv.1506.01497.

[25] D. V. Tuzoff et al., "Tooth detection and numbering in panoramic radiographs using convolutional neural networks," Dentomaxillofacial Radiol., vol. 48, no. 4, p. 20180051, Mar. 2019, doi: 10.1259/dmfr.20180051.

[26] F. P. Mahdi, K. Motoki, and S. Kobashi, "Optimization technique combined with deep learning method for teeth recognition in dental panoramic radiographs," Sci. Rep., vol. 10, no. 1, Art. no. 1, Nov. 2020, doi: 10.1038/s41598-020-75887-9.

[27] E. Bilgir et al., "An artificial ıntelligence approach to automatic tooth detection and numbering in panoramic radiographs," BMC Med. Imaging, vol. 21, p. 124, Aug. 2021, doi: 10.1186/s12880-021-00656-7.

[28] M. Estai et al., "Deep learning for automated detection and numbering of permanent teeth on panoramic images," Dentomaxillofacial Radiol., vol. 51, no. 2, p. 20210296, Feb. 2022, doi: 10.1259/dmfr.20210296.

[29] B. Thanathornwong and S. Suebnukarn, "Automatic detection of periodontal compromised teeth in digital panoramic radiographs using faster regional convolutional neural networks," Imaging Sci. Dent., vol. 50, no. 2, pp. 169–174, Jun. 2020, doi: 10.5624/isd.2020.50.2.169.

[30] M. Du, X. Wu, Y. Ye, S. Fang, H. Zhang, and M. Chen, "A Combined Approach for Accurate and Accelerated Teeth Detection on Cone Beam CT Images," Diagnostics, vol. 12, no. 7, p. 1679, Jul. 2022, doi: 10.3390/diagnostics12071679.

[31] A. Kumar, H. S. Bhadauria, and A. Singh, "Descriptive analysis of dental X-ray images using various practical methods: A review," PeerJ Comput. Sci., vol. 7, p. e620, Sep. 2021, doi: 10.7717/peerj-cs.620.

[32] S. Yilmaz, M. Tasyurek, M. Amuk, M. Celik, and E. M. Canger, "'Developing Deep Learning Methods for Classification of Teeth in Dental Panoramic Radiography,'" Oral Surg. Oral Med. Oral Pathol. Oral Radiol., Mar. 2023, doi: 10.1016/j.oooo.2023.02.021.

[33] Y. Yasa et al., "An artificial intelligence proposal to automatic teeth detection and numbering in dental bite-wing radiographs," Acta Odontol. Scand., vol. 79, no. 4, pp. 275–281, May 2021, doi: 10.1080/00016357.2020.1840624.

[34] F. Schwendicke, T. Golla, M. Dreher, and J. Krois, "Convolutional neural networks for dental image diagnostics: A scoping review," J. Dent., vol. 91, p. 103226, Dec. 2019, doi: 10.1016/j.jdent.2019.103226.

[35] W. Guo et al., "Remote Sensing Image Scene Classification by Multiple Granularity Semantic Learning," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 15, pp. 2546–2562, 2022, doi: 10.1109/JSTARS.2022.3158703.

[36] G. Wang, M. Jean Bosco, and Y. Hategekimana, Multi-Granularity Neural Network Encoding Method for Land Cover and Land Use Image Classification. 2021. doi: 10.20944/preprints202108.0325.v1.

[37] A. Djamanakova et al., "Tools for multiple granularity analysis of brain MRI data for individualized image analysis," NeuroImage, vol. 101, pp. 168–176, Nov. 2014, doi: 10.1016/j.neuroimage.2014.06.046.

[38] K. Wu, B. Peng, and D. Zhai, "Multi-Granularity Dilated Transformer for Lung Nodule Classification via Local Focus Scheme," Appl. Sci., vol. 13, no. 1, Art. no. 1, Jan. 2023, doi: 10.3390/app13010377.

[39] F. Wang, Y. Zhou, S. Wang, V. Vardhanabhuti, and L. Yu, "Multi-Granularity Cross-modal Alignment for Generalized Medical Visual Representation Learning." arXiv, Oct. 12, 2022. doi: 10.48550/arXiv.2210.06044.

[40] X. Zuo, J. Chu, J. Shen, and J. Sun, "Multi-Granularity Feature Aggregation with Self-Attention and Spatial Reasoning for Fine-Grained Crop Disease Classification," Agriculture, vol. 12, no. 9, Art. no. 9, Sep. 2022, doi: 10.3390/agriculture12091499.

[41] E. Ouyang, B. Li, W. Hu, G. Zhang, L. Zhao, and J. Wu, "When Multigranularity Meets Spatial–Spectral Attention: A Hybrid Transformer for Hyperspectral Image Classification," IEEE Trans. Geosci. Remote Sens., vol. 61, pp. 1–18, 2023, doi: 10.1109/TGRS.2023.3242978.

[42] M. Tu et al., "Multi-Granularity Mutual Learning Network for Object Re-Identification," IEEE Trans. Intell. Transp. Syst., vol. 23, no. 9, pp. 15178–15189, Sep. 2022, doi: 10.1109/TITS.2021.3137954.

[43] X. Wu, T. Tanprasert, and W. Jing, "Image classification based on multi-granularity convolutional Neural network model," in 2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE), Jun. 2022, pp. 1–4. doi: 10.1109/JCSSE54890.2022.9836281.

[44] Z. Chen, R. Ding, T.-W. Chin, and D. Marculescu, "Understanding the Impact of Label Granularity on CNN-based Image Classification." arXiv, Jan. 21, 2019. doi: 10.48550/arXiv.1901.07012.

[45] P. Zhu, Z. Zhu, Y. Wang, J. Zhang, and S. Zhao, "Multi-granularity episodic contrastive learning for few-shot learning," Pattern Recognit., vol. 131, p. 108820, Nov. 2022, doi: 10.1016/j.patcog.2022.108820.

[46] D. Yu, Q. Xu, H. Guo, C. Zhao, Y. Lin, and D. Li, "An Efficient and Lightweight Convolutional Neural Network for Remote Sensing Image Scene Classification," Sensors, vol. 20, no. 7, Art. no. 7, Jan. 2020, doi: 10.3390/s20071999.

[47] A. Dutta and A. Zisserman, "The VIA Annotation Software for Images, Audio and Video," in Proceedings of the 27th ACM International Conference on Multimedia, Oct. 2019, pp. 2276–2279. doi: 10.1145/3343031.3350535.

[48] S. AbuSalim, N. Zakaria, N. Mokhtar, S. A. Mostafa, and S. J. Abdulkadir, "Data Augmentation on Intra-Oral Images Using Image Manipulation Techniques," in 2022 International Conference on Digital Transformation and Intelligence (ICDI), Dec. 2022, pp. 117–120. doi: 10.1109/ICDI57181.2022.10007158.

[49] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," J. Big Data, vol. 6, no. 1, p. 60, Jul. 2019, doi: 10.1186/s40537-019-0197-0.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." arXiv, Dec. 10, 2015. doi: 10.48550/arXiv.1512.03385.

[51] R. Girshick, "Fast R-CNN." arXiv, Sep. 27, 2015. Accessed: Jun. 10, 2023. [Online]. Available: http://arxiv.org/abs/1504.08083

[52] F. Saeed, M. J. Ahmed, M. J. Gul, K. J. Hong, A. Paul, and M. S. Kavitha, "A robust approach for industrial small-object detection using an improved faster regional convolutional neural network," Sci. Rep., vol. 11, no. 1, Art. no. 1, Dec. 2021, doi: 10.1038/s41598-021-02805-y.

[53] C. Cao et al., "An Improved Faster R-CNN for Small Object Detection," IEEE Access, vol. 7, pp. 106838–106846, 2019, doi: 10.1109/ACCESS.2019.2932731.

[54] Y. Xiao, X. Wang, P. Zhang, F. Meng, and F. Shao, "Object Detection Based on Faster R-CNN Algorithm with Skip Pooling and Fusion of Contextual Information," Sensors, vol. 20, no. 19, p. 5490, Sep. 2020, doi: 10.3390/s20195490.

[55] Q. Xu, X. Zhang, R. Cheng, Y. Song, and N. Wang, "Occlusion Problem-Oriented Adversarial Faster-RCNN Scheme," IEEE Access, vol. 7, pp. 170362–170373, 2019, doi: 10.1109/ACCESS.2019.2955685.