ADVANCEMENTS IN ARTIFICIAL INTELLIGENCE AND COMPUTER VISION FOR DENTAL IMAGING ANALYSIS: SELF-SUPERVISED LEARNING INNOVATIONS

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

by
Amani Hasan Almalki
August 2024

Examining Committee Members:

Longin Jan Latecki, Advisory Chair, Computer and Information Sciences
Xinghua Mindy Shi, Computer and Information Sciences
Stephen MacNeil, Computer and Information Sciences
Louis DiPede, External Reader, Restorative Dentistry

ABSTRACT

This dissertation explores the application of self-supervised learning methods in dental radiology to address the challenges posed by limited data availability for training deep learning models. The overarching goal is to enhance the efficiency and accuracy of automated systems for various dental diagnostic tasks, including teeth numbering, detection of dental restorations, orthodontic appliances, implant systems, marginal bone level, and dental caries from panoramic radiographs, CBCT images, intra-oral 3D scans, and dental radiographs.

Key contributions include the development of several novel approaches:

- **Self-supervised Learning for Dental Panoramic Radiographs**: Utilizing SimMIM and UM-MAE with Swin Transformer, we achieved significant improvements in teeth detection and instance segmentation, increasing the average precision by 13.4% and 12.8%, respectively, over baseline methods.

- **Self-Distillation Enhanced Self-supervised Learning (SD-SimMIM)**: Enhancing SimMIM with self-distillation loss, we improved performance on teeth numbering, dental restoration detection, and orthodontic appliance detection tasks, demonstrating superior outcomes compared to other methods.

- **DentalMAE for Intra-oral 3D Scans**: Extending the mesh masked autoencoder (MeshMAE), DentalMAE evaluates predicted deep embeddings of masked mesh triangles, yielding better generalization and higher accuracy in teeth segmentation tasks.

- **DEMAE for Dental CBCT Images**: Proposing the Deep Embedding MAE (DEMAE), which measures the closeness of predicted deep embeddings of masked patches to their originals, we achieved significant accuracy improvements in teeth segmentation from CBCT images.

- **Masked Deep Embedding (MDE) for Implant Detection**: By leveraging MIM, we developed MDE to enhance dental implant detection, creating a comprehensive Implant Design Dataset (IDD) with expert annotations, significantly boosting detection performance.

- **Deep Embedding of Patches (DEP) for Bone Loss Assessment**: An extension of MAE, DEP improved the accuracy of marginal bone level detection, supported by the creation of a Bone Loss Assessment Dataset (BLAD) with detailed annotations.

- **Masked Deep Embedding of Patches (MDEP) for Caries Detection**: This method enhanced dental caries detection performance, validated on the CariesXrays dataset, demonstrating higher precision and recall rates compared to traditional baselines.

Through these innovations, the dissertation establishes the efficacy of self-supervised learning in overcoming data scarcity in dental imaging, offering promising AI-driven solutions for improved diagnostics and patient care in dentistry.

To my beloved father Hasan, my mother Khadijah, my husband Abdulrahman, my baby Faisal, my brothers Naif, Ahmed, and Khalid, and my sister Alaa.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. Longin Jan Lateki, for his unwavering support and guidance throughout my journey. This work would not have been possible without his dedicated mentorship. I also extend my heartfelt thanks to all the committee members, Dr. Xinghua Mindy Shi, Dr. Stephen MacNeil, and Dr. Louis DiPede, for their invaluable comments and suggestions, which significantly improved this project and elevated it to new heights. Finally, I am profoundly grateful to my family and my newborn baby for their patience and understanding, and for their constant support and encouragement throughout this demanding process.

TABLE OF CONTENTS

Page

ABSTRACT iii

ACKNOWLEDGMENTS vi

LIST OF TABLES xi

LIST OF FIGURES xiii

CHAPTER

1 IMPROVING DENTAL DIAGNOSTICS WITH SELF-SUPERVISED LEARNING 1

1.1 Introduction 1

1.2 Teeth numbering 4

1.3 Methods 4

1.4 Experiments 7

1.4.1 Dataset augmentation and correction 7

1.4.2 Evaluation metric 9

1.4.3 Implementation details 9

1.5 Results and analysis 10

1.5.1 Quantitative results 10

1.5.2 Qualitative results 12

1.5.3 Pre-training time and memory consumption 15

1.6 Conclusions 16

2 ENHANCED MASKED IMAGE MODELING FOR DENTAL PANORAMIC RADIOGRAPHS 17

2.1 Introduction 17

2.2 Methods 19

2.2.1 SimMIM 19

Page

2.2.2 Self-distillation 21

2.3 Experiments 22

2.3.1 Dataset 22

2.3.2 Evaluation metric 23

2.3.3 Implementation details 23

2.3.4 Quantitative results 24

2.3.5 Qualitative results 25

2.4 Conclusions 27

3 TOOTH SEGMENTATION FROM INTRA-ORAL 3D SCANS 29

3.1 Introduction 29

3.2 Related work 31

3.2.1 Handcrafted features-based approaches 31

3.2.2 Learning-based approaches 32

3.3 Methods 35

3.3.1 Mesh Transformer 37

3.3.2 Mesh Pre-training Task 39

3.4 Experiments 40

3.4.1 Dataset 40

3.4.2 Evaluation metric 40

3.4.3 Implementation details 41

3.5 Results and analysis 44

3.5.1 Quantitative results 44

3.5.2 Qualitative results 47

3.6 Conclusions 47

4 TEETH SEGMENTATION FROM CONE-BEAM CT IMAGES 48

4.1 Introduction 48

4.2 Methods 50

4.2.1 Vision Transformer 50

4.2.2 Self-Supervised Pre-training with Masked Autoencoders 51

Page
4.2.3 Architectures for Downstream Tasks 53
4.3 Experiments and Results 53
4.3.1 Datasets and Implementation Details 53
4.3.2 Evaluation metric 54
4.3.3 Results 55
4.4 Conclusions 57
5 DENTAL IMPLANT IDENTIFICATION 59
5.1 Introduction 59
5.2 Dental Implant Design 61
5.3 Methods 64
5.3.1 Two-Stage Implant Detection Methodology 64
5.3.2 Self-Supervised Pre-training with Masked Autoencoders 66
5.3.3 Architectures for Downstream Tasks 67
5.4 Experiments 68
5.4.1 Dataset 68
5.4.2 Evaluation metric 68
5.4.3 Implementation details 68
5.5 Results and analysis 69
5.5.1 MDE reconstruction 69
5.5.2 Dental implant classification and identification. . . . 70
5.5.3 Qualitative Results 73
5.5.4 Parameter setting 74
5.6 Conclusions 74
6 ANALYZING DENTAL PANORAMIC RADIOGRAPHS FOR ASSESSMENT OF PERI-IMPLANT BONE LOSS 76
6.1 Introduction 76
6.2 Related Work 79
6.3 Methods 81
6.3.1 Self-supervised Learning 81

Page

6.3.2 Architectures for Downstream Tasks 83

6.3.3 Evaluation Methods 86

6.4 Experiments 88

6.4.1 Dataset 88

6.4.2 Evaluation Metrics 88

6.4.3 Details on Implementation 89

6.5 Results and analysis 89

6.5.1 DEP reconstruction 89

6.5.2 Detection of dental implants and marginal bone levels 91

6.5.3 Parameter setting 95

6.6 Conclusions 96

7 DENTAL CARIES DETECTION 97

7.1 Introduction 97

7.2 Related Work 100

7.2.1 Dental X-rays 100

7.2.2 Self-supervised Learning 102

7.2.3 Object Detection 104

7.3 Methods 105

7.3.1 Masked Autoencoders 105

7.3.2 Architectures for Downstream Tasks 106

7.3.3 Training and Data Augmentation 107

7.4 Experimental Evaluation 107

7.4.1 Data Sets and Evaluation Metrics 107

7.4.2 Quantitative Results 108

7.4.3 Qualitative Results 111

7.4.4 Mask Ratio Analysis 111

7.4.5 Implementation Details 114

7.5 Conclusion 115

REFERENCES CITED 116

LIST OF TABLES

Table Page

1.1 Results of teeth detection and instance segmentation only. 11

1.2 Results after augmenting dental restorations. 12

1.3 The influence of Mask Ratios on teeth detection and instance segmentation tasks. 13

1.4 Correction of teeth segmentation. 13

1.5 The comparison of pre-training time and memory consumption. 16

2.1 Results of teeth numbering, detection of dental restorations, and instance segmentation only. * denotes L_1 loss is computed on the whole image. AP^{box} and AP^{mask} indicate Average Precision for object detection and instance segmentation, respectively. 25

2.2 Results after including orthodontics appliances. * denotes L_1 loss is computed on the whole image. 27

3.1 The tooth segmentation results from different methods in terms of the label-wise Dice Score. 42

3.2 The tooth segmentation results from different methods in terms of the Overall Accuracy, the Dice Score, the Sensitivity, and the Positive Predictive Value. 45

3.3 The influence of Mask Ratios/strategies on teeth segmentation of our DentalMAE. 46

4.1 Tooth Segmentation on CBCT scans. UNETR+DEMAE self pre-training improves upon the UNETR baseline, ImageNet supervised pre-training, and MAE self-supervised learning. 56

4.2 The influence of Mask Ratios and Pre-training Epochs on teeth segmentation of our UNETR+DEMAE. 56

5.1 Details of classes based on the classification of implant design. 63

5.2 Results of dental implant classification and identification before employing dental implant design labeling. 72

5.3 Results of dental implant classification and identification after employing dental implant design labeling. 72

Table Page

5.4 Impact of Mask Ratios on dental implant detection. 74

6.1 Results of dental implant detection. 92

6.2 Results of keypoints detection. 93

6.3 Impact of mask ratios and pretraining epochs on dental implant detection. 95

7.1 Results of dental caries detection on periapical X-rays dataset [159] in terms of precision (P), recall (R), average precision (AP), and average precision at IoU thresholds of 50% (AP@50) and 75% (AP@75). 108

7.2 Results of dental caries detection on CariesXrays dataset [158] in terms of Average Precision (AP) metrics. 109

7.3 Comparison with state-of-the-art on CariesXrays Dataset [158]. 110

7.4 Impact of Mask Ratios on dental caries detection. 112

LIST OF FIGURES

Figure Page

1.1 Pipeline for teeth detection, detection of dental restorations, and instance segmentation with MIM Self Pre-training. (a) A Swin Transformer is first pre-trained by MIM methods on the target dataset. (b) The pre-trained Swin Transformer is used as the backbone in Cascade Mask R-CNN with FPN for the detection and segmentation tasks. 5

1.2 Illustration of the architecture for object detection. 6

1.3 Comparison between the old and new dataset annotations. (a) Dataset old annotations. (b) Dataset new annotations. The blue arrow donates the inclusion of instance overlapping, while the green arrow indicates the correction of systematic errors, for example, unsegmented molar roots. 8

1.4 Samples of segmentation of dental restorations. Red arrows show an example of a) indirect restoration, b) direct restoration, and c) root canal therapy. 8

1.5 SimMIM and UM-MAE reconstruction results. The first column is the original image, and the second and fourth columns are the masked image where the masked region is denoted by gray patches. The third and fifth columns are the reconstruction of MIM from the unmasked patches. . . 11

1.6 Qualitative results of teeth detection and instance segmentation only. Note that teeth detection and instance segmentation are missing (white arrows) when created by the baseline Swin Transformer approach compared to the segmentation produced by Swin Transformer pre-trained with SimMIM architecture (orange arrows). 14

1.7 Qualitative results of detecting teeth and dental restorations and instance segmentation using SimMIM. 15

2.1 Our SD-SimMIM framework. Alongside the original SimMIM, we benefit from decoded visible patches (as the teacher) and transfer knowledge to their peers after encoding. (Best viewed in color) 20

2.2 Samples of segmentation of orthodontics appliances, a) shows examples of bands (yellow arrows) and brackets (green arrows), and b) a retainer (orange arrow). (Best viewed in color) 23

Figure Page

2.3 Images reconstructed by SimMIM and SD-SimMIM. SD-SimMIM shows a clearly better reconstruction than SimMIM. The color boxes highlight their details. (Best viewed in color) 26

2.4 Qualitative results of detection and instance segmentation. Note that teeth detection and instance segmentation are missing (red arrows) when Swin Transformer is pre-trained with SimMIM compared to the ones produced by Swin Transformer pre-trained with SD-SimMIM architecture (green arrows). (Best viewed in color.) 28

3.1 The teeth segmentation pipeline for DentalMAE self-pre-training. Initially, the input mesh is divided into non-overlap patches. These patches are then embedded using an MLP. During the pre-training phase, the patch embeddings are randomly masked, and only the visible embeddings are utilized by the transformer. Subsequently, the masked embeddings are combined with the encoded embeddings and sent to the decoder. The objective of the decoder is to reconstruct the vertices and face features of the masked patches, followed by the prediction of the patch embeddings of the masked patches. The L_2 loss is used to compare the masked patch embeddings. After the completion of pre-training, the decoder is discarded, and the encoder is employed for segmentation. 36

3.2 The remeshing operation involves several steps. Initially, the input mesh undergoes a simplification process. Subsequently, a mapping is established between the original mesh and the base mesh. The base mesh is then subdivided three times, and the newly generated vertices are projected back onto the input mesh. 37

3.3 Tooth segmentation and the corresponding color coding. 41

3.4 Comparison of teeth segmentation of DentalMAE and baselines. The first three rows show samples of the lower jaw, while the last two rows show the upper jaw. 43

4.1 Segmentation Pipeline with MAE Self Pre-training. Left: A ViT encoder is first pre-trained with MAE. A random subset of patches is input to the encoder and a transformer decoder reconstructs the full image. Right: The pre-trained ViT weights are transferred to initialize the segmentation encoder. Then the whole segmentation network, e.g., UNETR [100], is finetuned for segmentation. 52

4.2 Comparison of teeth segmentation of UNETR+DEMAE and baseline. . 57

Figure Page

5.1 Categorization of (a) the coronal portion's design based on bone or tissue level, the existence of microthreads, and the specific thread design, (b) the middle section's design based on the shape of the body and the specific thread design, and (c) the design for the apical portion based on the shape of the apical hole, the configuration of the apical body, the presence of a groove, and the shape of the apex. 62

5.2 Image of a sample for categorizing and labeling implant design. 64

5.3 Dental Implant Detection Pipeline with MDE Self Pre-training. The initial step in the dental implant detection pipeline for MDE self-pre-training involves dividing the input into non-overlapping patches. These patches undergo embedding using an MLP. Throughout the pre-training phase, the patch embeddings undergo random masking, and only the visible embeddings are employed by the transformer. Subsequently, the masked embeddings are merged with the encoded embeddings and directed to the decoder. The decoder's role is to reconstruct the masked patches, followed by predicting the patch embeddings of these masked patches. The L_1 loss is employed to assess the similarity between the masked patch embeddings. Once pre-training is complete, the decoder is omitted, and the encoder is utilized as the backbone in Mask R-CNN with FPN for the detection. . 66

5.4 Results of MDE reconstruction. The first column displays the original images, while the second column shows the masked images, with gray patches indicating the masked regions. The third and fourth columns exhibit the reconstructions achieved through MAE and our MDE, respectively, from the unmasked patches. 70

5.5 Qualitative results of dental implant detection and identification. The ViT pre-trained with the MAE approach exhibits missing or incorrect detections, whereas the ViT pre-trained with the MDE approach demonstrates accurate detection. Blue indicates the Bego dental implant system, yellow indicates the Bicon dental implant system, and red indicates the ITI dental implant system. 73

Figure Page

6.1 Dental Implant and keypoints Detection Pipeline with DEP Self pretraining. The initial step in the dental implant and keypoints detection pipeline for DEP self-pretraining involves dividing the input into non-overlapping patches. These patches undergo embedding using an MLP. Throughout the pretraining phase, the patch embeddings undergo random masking, and only the visible embeddings are employed by the transformer. Subsequently, the masked embeddings are merged with the encoded embeddings and directed to the decoder. The decoder's role is to reconstruct the masked patches, followed by predicting the patch embeddings of these masked patches. The L_1 loss is employed to assess the similarity between the masked patch embeddings. Once pretraining is complete, the decoder is omitted, and the encoder is utilized as the backbone in Mask R-CNN with FPN for the detection. 82

6.2 Depiction of anticipated bounding boxes and keypoints includes: (a) bounding boxes, indicated by blue arrows; (b) keypoints, represented as red dots positioned at the center of the radiographs; (c) calculation of radiographic bone loss percentage, based on the locations of the keypoints. 85

6.3 Results of DEP reconstruction. The first column displays the original images, while the second column shows the masked images, with blue patches indicating the masked regions. The third and fourth columns exhibit the reconstructions achieved through MAE and our DEP, respectively, from the unmasked patches. 90

6.4 Illustration of anticipated outcomes: Each implant is identified using a bounding box, and the predicted keypoints are displayed within the box. The ratio of radiographic bone loss is determined from the positions of the keypoints. Additionally, confidence scores for both implant and keypoint detection are provided. 94

7.1 Illustration of dental X-ray images, displaying the various forms and appearances of dental caries. 98

7.2 Dental Caries Detection Pipeline with MDEP Self Pre-training. The first step in this pipeline involves splitting the input into non-overlapping patches, which are then embedded using a multi-layer perceptron (MLP). During the pre-training phase, random masking is applied to the patch embeddings, with only the visible embeddings processed by the transformer. The masked embeddings are combined with the encoded embeddings and sent to the decoder. The decoder's task is to predict the embeddings of the masked patches. The similarity between the predicted and actual masked patch embeddings is evaluated using the L_1 loss. After pre-training, the decoder is discarded, and the encoder serves as the backbone in the Mask R-CNN with FPN for detection. 101

Figure Page

7.3 Comparison of dental caries detection on periapical radiographs dataset [159]. Bounding boxes (BB) are used to highlight detected caries. Red BBs indicate detected cavities with confidence scores, while green BBs indicate the ground truth cavities. 113

7.4 Comparison of dental caries detection on the CariesXrays dataset [158]. 113

CHAPTER 1
IMPROVING DENTAL DIAGNOSTICS WITH SELF-SUPERVISED LEARNING

1.1 Introduction

The need for computer-assisted decisions is rising to facilitate diagnosis and treatment planning for dental care providers. Dental imaging is a valuable diagnostic tool for diagnosis and treatment plans, which is not possible solely through clinical exams and patient history [1]. A dental panoramic X-ray is a comprehensive tool that screens the teeth, surrounding alveolar bone and upper and lower jaws [2].

Moreover, dental restoration is a biocompatible synthetic material used to restore missing tooth structures. The missing tooth structure can be restored with full and partial coverage depending on the extension and intensity of the missing structure to restore the tooth's coronal (top) part. Furthermore, root canal filling is a restorative procedure used to fill the space inside the tooth structure (root portion) with biocompatible restorative materials. Various dental restorative materials are available in the dental world; each has its indication, advantages, disadvantages, and clinician preferences. Most dental restorative materials appear radiopaque in the x-ray, and they can be identified by dental care providers [3, 4].

However, manual intervention for teeth numbering and identification of tooth restorations is time-consuming and may overlook significant data. Thus, the interest in computer vision and computer science for automated processes was aroused. Few studies have attempted to apply computer vision algorithms in dental radiograph analysis. They include convolutional neural networks (CNNs) for teeth numbering and instance segmentation [5], two-stage network [6], Faster R-CNN [7–11], PANet [12], Mask R-CNN [13–16], and U-Net network [17–19]. Recently, CNNs have enormous emerging applications in analyzing medical images with the advent of computation hardware/algorithm and expansion in the amount of data [5]. However, CNNs are limited in overall capability because of inherent inductive biases [20].

In this study, we propose to use a recently introduced Swin Transformer [21] to analyze dental panoramic radiographs. However, Swin Transformer requires large data for training, but there is only a very limited number of available dental radiographs. To alleviate this problem, we propose to use self-supervised learning. To the best of our knowledge, this is the first study that applied self-supervised learning methods to Swin Transformer on dental panoramic radiographs.

Recently, the self-supervised learning methods, SimMIM [22], UM-MAE [23], BEiT [24], MAE [25], SplitMask [26], MoCo v3 [27], and DINO [28], are effective in pre-training Transformers [20,21] for learning visual representation. However, only UM-MAE and SimMIM pre-training methods are enabled for Pyramid-based ViTs with locality (Swin Transformer). Generally, the Masked Image Modeling (MIM) methods mask some image patches before they are fed into the transformer to predict the original patches in the masked area. This feature of aggregating information from the context helps many vision tasks. Although both UM-MAE and SimMIM provide a simple and efficient pre-training strategy for the Swin transformer encoder [21], the process of the input to the encoder is dissimilar. MAE discards the masked tokens and inputs only visible patches to the lightweight decoder. However, MAE also breaks the two-dimensional structure of the input image. Therefore, it is not applicable to the Swin transformer without the Uniform Masking (UM) introduced in [23] to bridge the gap between the MAE and Swin transformer. SimMIM includes the masked tokens in the encoder and uses them as a direct prediction mechanism. Using the randomly masked patches for SimMIM is a reasonable reconstruction target, and a lightweight prediction head is sufficient for pre-training. In addition, the location of the patches is essential in dental radiographs for a predictable outcome. SimMIM maintains the location of the patches known to both encoder and decoder, while MAE drops the location information, which may induce inaccuracy, as we demonstrate in this paper.

As there is no standard dental image dataset for pre-training (unlike ImageNet for natural images), SimMIM and UM-MAE are trained on the same dataset as the downstream tasks (excluding the test dataset). We conduct experiments on dental image tasks, including teeth numbering, detection of dental restorations, and instance segmentation on the dental panoramic X-rays dataset [12]. For these tasks, we use the base Swin Transformer (Swin-B) [21] as the backbone of Cascade Mask R-CNN [29]. We compare four Swin Transformer initializations, including SimMIM and UM-MAE, supervised initialization, and random initialization baseline. Our results show that SimMIM self pre-training can significantly improve object detection and instance segmentation performance on dental images.

Although previous studies have investigated teeth segmentation, we still address many gaps in this work. First, there is no comprehensive instance segmentation data set for teeth numbering. Previous work on the matter [12] used modified versions of binary semantic segmentation masks, which leads to a lack of instance overlapping and low-resolution outputs, resulting in inaccurate predictions, especially on the boundaries of the teeth. Second, there is a considerable amount of systematic errors because of the absence of dental expert supervision. Third, no prior work has simultaneously considered dental restoration segmentation besides tooth segmentation. The inclusion of teeth restorations increases the complexity of the computer vision problem because of class quantity and class imbalance.

To solve the data set issues, we augment and correct the existing dataset introduced in [12]. In addition to correcting the manual segmentation errors under expert supervision, we further expand the dataset by developing annotations for dental restorations, including direct restorations, indirect restorations, and root canal therapy. The labeling procedure resulted in a unique high quality, augmented dataset. Our data is available, upon request, under the name TNDRS (Teeth Numbering, Detection of Restorations, and Segmentation) annotations.

Our main contributions are twofold:

- We utilize self-supervised learning with SimMIM and UM-MAE to alleviate the problem of small data for panoramic radiographs.

- The corrected dataset leads to a significant increase in performance, while added labeling of dental restorations extends the horizon of possible dental applications.

1.2 Teeth numbering

In dentistry, various dental numbering systems are available for teeth numbering for adults and children. These numbering systems are universally accepted for better communication between dental care providers. The Universal Numbering System, Palmer Notation Numbering System, and Federation Dentiaure International numbering system (FDI) are the most commonly used system across the globe among dental professionals. The FDI system is the most widely used international system. In this system, every single tooth is assigned two-digit numbers; the first digit number represents each quadrant. The maxillary right and left quadrants are identified by the numbers 1 and 2, while the mandibular left and right quadrants are the numbers 3 and 4, respectively. The second digit numbers represent each tooth based on its location in the jaw from the middle. The central incisor is assigned to number 1, whereas the third molar is set to number 8 [9, 30].

1.3 Methods

The methods include two stages: the MIM pre-training and the downstream tasks, as illustrated in Fig. 7.2.

In the first stage, Swin Transformer is pre-trained with MIM self-supervised learning methods as the encoder. SimMIM divides the image into patches, replacing some random patches with mask tokens. Then, these patches, along with mask tokens, are input to the Swin encoder. Hence the positional encoding of both visible and masked patches is preserved, while UM-MAE drops those mask positions entirely. UM-MAE samples three random patches from each two-by-two grid, dropping 25% of the entire image. Then it randomly masks 25% of the already sampled areas as shared learnable tokens. Finally, the sampled patches and the masked tokens are reorganized as

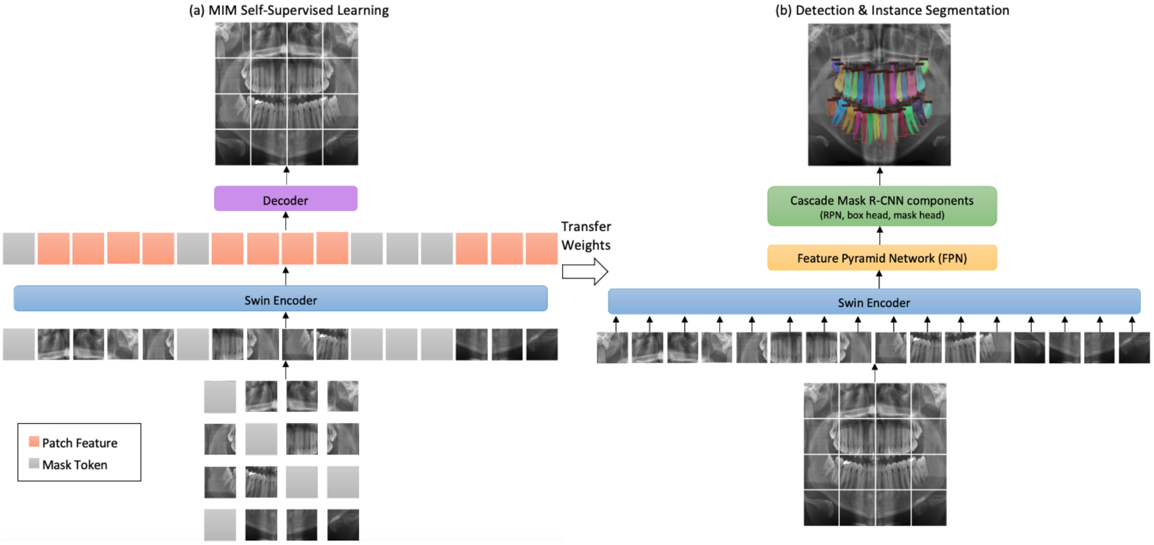

(a) MIM Self-Supervised Learning
Decoder
Swin Encoder
Patch Feature
Mask Token
Transfer Weights
(b) Detection & Instance Segmentation
Cascade Mask R-CNN components
(RPN, box head, mask head)
Feature Pyramid Network (FPN)
Swin Encoder

Figure 1.1. Pipeline for teeth detection, detection of dental restorations, and instance segmentation with MIM Self Pre-training. (a) A Swin Transformer is first pre-trained by MIM methods on the target dataset. (b) The pre-trained Swin Transformer is used as the backbone in Cascade Mask R-CNN with FPN for the detection and segmentation tasks.

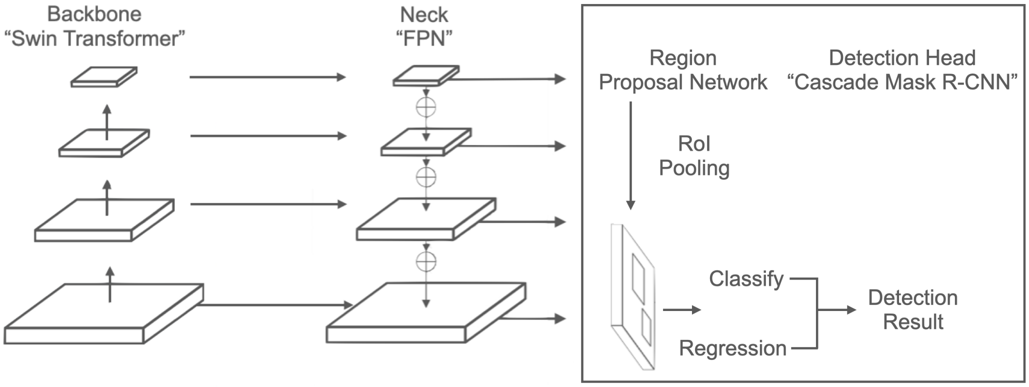


Figure 1.2. Illustration of the architecture for object detection.

a compact two-dimensional input under a quarter of the original image resolution to feed via the Swin encoder.

Then a decoder is appended to reconstruct the original patches at the masked area for both methods. In the second stage, the pre-trained Swin weights are transferred to initialize the detection and segmentation encoder. The features of the Swin Transformer backbone are fed to the neck (FPN [31]) and detection head (Cascade Mask R-CNN) for bounding box regression and classification as illustrated in Fig. 1.2. We select the Cascade Mask R-CNN [29] framework due to its ubiquitous presence in object detection and instance segmentation research. Then, the whole network is fine-tuned to perform the detection and segmentation tasks.

We use the base Swin Transformer backbone (Swin-B) and compare the effectiveness of four configurations as follows:

Random. The network is trained from scratch with randomly initialized weights, and no self-supervised methods are used. The Swin backbone configuration follows the code of [21], and the Cascade Mask R-CNN configuration uses the defaults in MMDetection [32].

Supervised. The Swin backbone is pre-trained for supervised object detection and instance segmentation using ImageNet-1K [33] images with their labels. We use the weights from [21] for Swin-B. Swin-B was pre-trained for 300 epochs.

SimMIM. We use the Swin-B weights pre-trained on self-supervised ImageNet-1K from [22]. This model was pre-trained for 100 epochs.

UM-MAE. Since ImageNet-1K pre-trained weights are not available; we use the official UM-MAE code release [23] to train Swin-B ourselves for 800 epochs (the default training length used in [23]) on unsupervised ImageNet-1K.

1.4 Experiments

1.4.1 Dataset augmentation and correction

TNDRS dental panoramic radiographs dataset. Detection, Numbering, and Segmentation (DNS) [12] is a dental panoramic X-rays dataset consisting of 543 annotated images with ground truth segmentation labels, including numbering information based on the FDI teeth numbering system. The image size is 1991x1127 pixels. The dataset annotations have some limitations as follows: 1) lack of instance overlapping; 2) some systematic errors because of the absence of dental expert supervision; 3) no segmentation of dental restorations. To overcome these issues, we modify and correct teeth instance segmentation and overlapping in all images. In addition, we contribute to further expanding the dataset by developing segmentation for dental restorations, including direct restorations, indirect restorations, and root canal therapy. This process was under a supervision of a dentist using the COCO-Annotator tool [34]. We attended weekly meetings where related issues, such as numbering, dental restorations, and segmentation questions, were discussed. In the end, the annotations were reviewed to assure quality and avoid systematic and random errors. Fig. 1.3 shows a sample comparing the old and new versions of the dataset annotations, highlighting both the instance overlapping (blue arrow) and the correction of systematic errors (green arrow). Fig. 1.4 presents samples of segmentation of dental restorations.

We believe this is the most inclusive dataset for segmenting teeth and dental restorations in dental panoramic radiographs. We are providing our data, upon re-

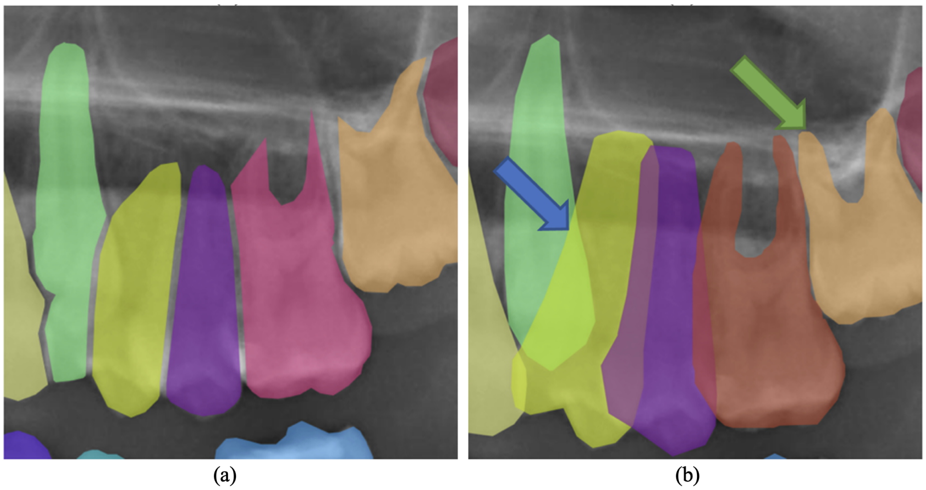


Figure 1.3. Comparison between the old and new dataset annotations. (a) Dataset old annotations. (b) Dataset new annotations. The blue arrow donates the inclusion of instance overlapping, while the green arrow indicates the correction of systematic errors, for example, unsegmented molar roots.

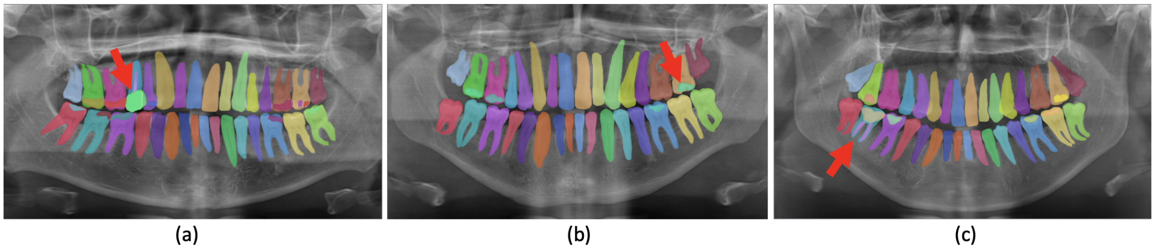


Figure 1.4. Samples of segmentation of dental restorations. Red arrows show an example of a) indirect restoration, b) direct restoration, and c) root canal therapy.

quest, under the name TNDRS (Teeth Numbering, Detection of Restorations, and Segmentation) annotations.

1.4.2 Evaluation metric

For all our experiments, we split the data into five folds, each containing approximately 20% of the images. One of these folds is fixed as the test dataset (consisting of 111 images), and the other four folds (consisting of 108 images each) compose the training and validation datasets in a cross-validation manner. This process is repeated five times. The evaluation metric we adopt is the Average Precision for object detection and instance segmentation models.

1.4.3 Implementation details

Our experiments are implemented based on the PyTorch [35] framework and trained with NVIDIA Tesla Volta V100 GPUs. In all experiments, the batch size equals the total number of the training sample, which is 432. The input images are all resized to 800×600 pixels. We utilize the AdamW [36] optimizer in all experiments.

Data augmentation. We apply noise addition and horizontal flipping, which changes teeth numbers to their equivalent new values (left teeth numbers turned into the right numbers and vice-versa).

SimMIM pre-training. The base learning rate is set to 8e-4, weight decay is 0.05, $\beta 1 = 0.9$, $\beta 2 = 0.999$, with a cosine learning rate scheduler with warm-up for 10 epochs. We use a random MIM with a patch size of 16×16 and a mask ratio of 20%. We employ a linear prediction head with a target image size of 800×600 and use L1 loss to compute the loss for masked pixel prediction.

UM-MAE pre-training. The base learning rate is set to 1.5e-4, weight decay is 0.05, $\beta 1 = 0.9$, $\beta 2 = 0.95$, with a cosine decay learning rate scheduler with warm-up for 10 epochs. We use a random MIM with a patch size of 16×16 and a mask ratio of 25%. We employ a linear prediction head with a target image size of 800×600 and adopt mean squared error (MSE) to compute the loss for masked pixel prediction.

Task fine-tuning. For downstream tasks, we utilize single-scale training. The initial learning rate is 0.0001, and the weight decay is 0.05.

1.5 Results and analysis

SimMIM and UM-MAE reconstruction. The reconstruction results of SimMIM and UM-MAE are shown in Fig. 6.3. The five columns show the original images, the UM-MAE masked images, the UM-MAE reconstructed images, the SimMIM masked images, and the SimMIM reconstructed images. The results show that both MIM methods can restore lost information from the random context. It is worth noting that the ultimate goal of the MIM is to benefit the downstream tasks instead of generating high-quality reconstructions.

1.5.1 Quantitative results

Comparing initializations. Table 3.2 shows the results of teeth detection and instance segmentation only and compares them to the previously published article from Silva et al. [12]. We present TNDRS fine-tuning results using the pre-trained models and random configurations described in Section 1.3. We make several observations.

(1) All four Swin Transformer initializations surpass the CNN-based SOTA of PANet with ResNet-50 backbone using ImageNet pre-training from Silva et al. [12].

(2) Fine-tuning from supervised IN-1K pre-training yields 3.4 higher AP^{box} than training from scratch (79.1 vs. 75.7) and 3.5 higher AP^{mask} (78.3 vs. 74.8).

(3) UM-MAE substantially outperforms supervised initialization by 5.4 AP^{box} (84.5 vs. 79.1), and 4.9 AP^{mask} (83.2 vs. 78.3).

(4) SimMIM outperforms UM-MAE by 1.6 AP^{box} (86.1 vs. 84.5), and 1.4 AP^{mask} (84.6 vs. 83.2).

Table 1.2 compares the four Swin Transformer initializations after data augmentation of dental restorations. Our results prove that the SimMIM method achieved the

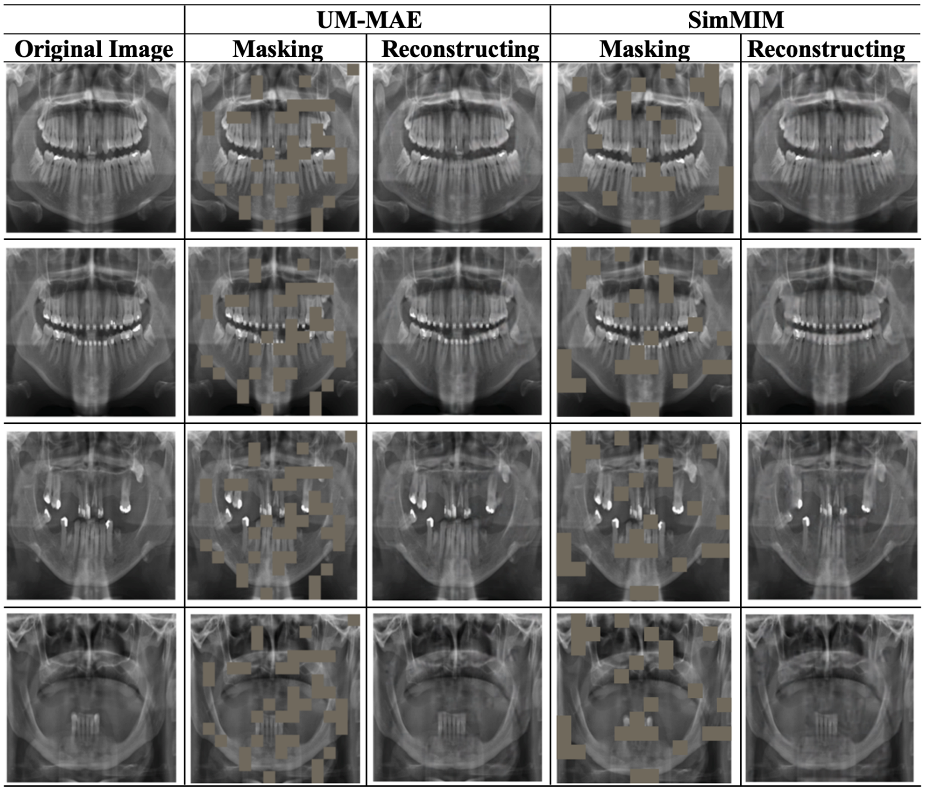


Figure 1.5. SimMIM and UM-MAE reconstruction results. The first column is the original image, and the second and fourth columns are the masked image where the masked region is denoted by gray patches. The third and fifth columns are the reconstruction of MIM from the unmasked patches.

Initialization	Backbone	Pre-training Data	AP^{box}	AP^{mask}
PANet [12]	ResNet-50	IN-1K w/ Labels	75.4	73.9
Random	Swin-B	None	75.7	74.8
Supervised	Swin-B	IN-1K w/ Labels	79.1	78.3
UM-MAE	Swin-B	IN-1K	84.5	83.2
SimMIM	Swin-B	IN-1K	**86.1**	**84.6**

Table 1.1.
Results of teeth detection and instance segmentation only.

Initialization	Backbone	Pre-training Data	AP^{box}	AP^{mask}
Random	Swin-B	None	77.0	76.1
Supervised	Swin-B	IN-1K w/ Labels	80.3	79.2
UM-MAE	Swin-B	IN-1K	88.3	85.7
SimMIM	Swin-B	IN-1K	**90.4**	**88.9**

Table 1.2.
Results after augmenting dental restorations.

highest performance of 90.4% and 88.9% on detecting teeth and dental restorations and instance segmentation, respectively.

Parameter setting. In Table 7.4, we conduct experiments on teeth detection and instance segmentation tasks with different SimMIM pre-training epochs and mask ratios. First, the performance of SimMIM does not benefit from longer training. Second, unlike the high mask ratio [22] adopted in natural images, the downstream tasks show different preferences for the mask ratio. Both tasks are consistently improved with a decrease in mask ratio from 60% to 10%. The reason why this decrease facilitates the training may be attributed to the fact that the relevant features are small on panoramic X-rays.

Dataset correction. After we correct teeth segmentation on DNS discussed in Section 3.4.1, teeth detection and instance segmentation performance are remarkably improved by 5.9 AP^{box} and 6.4 AP^{mask} as shown in Table 1.4.

1.5.2 Qualitative results

In Fig. 5.5, the displayed results for four different images demonstrate qualitative samples of improved performance when Swin Transformer is pre-trained with SimMIM for teeth detection and segmentation only. These improvements in detection and segmentation agree with the quantitative results in Section 5.5.2.

Mask ratio	Pre-training Epochs	AP^{box}	AP^{mask}
60%	100	84.3	83.2
50%	100	84.7	83.6
50%	800	83.1	83.0
40%	100	85.5	83.9
30%	100	85.9	84.1
20%	100	**86.1**	**84.6**
10%	100	85.8	84.3

Table 1.3.
The influence of Mask Ratios on teeth detection and instance segmentation tasks.

DNS Annotations	AP^{box}	AP^{mask}
Before Correction	80.2	78.2
After Correction	**86.1**	**84.6**

Table 1.4.
Correction of teeth segmentation.

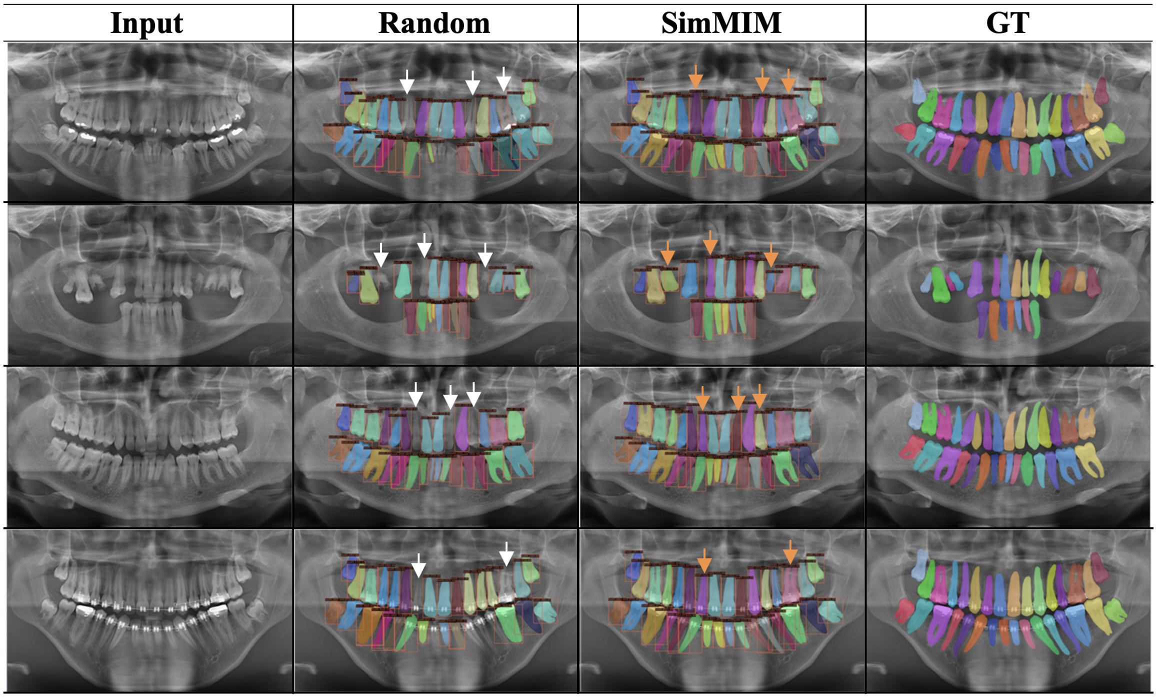


Figure 1.6. Qualitative results of teeth detection and instance segmentation only. Note that teeth detection and instance segmentation are missing (white arrows) when created by the baseline Swin Transformer approach compared to the segmentation produced by Swin Transformer pre-trained with SimMIM architecture (orange arrows).

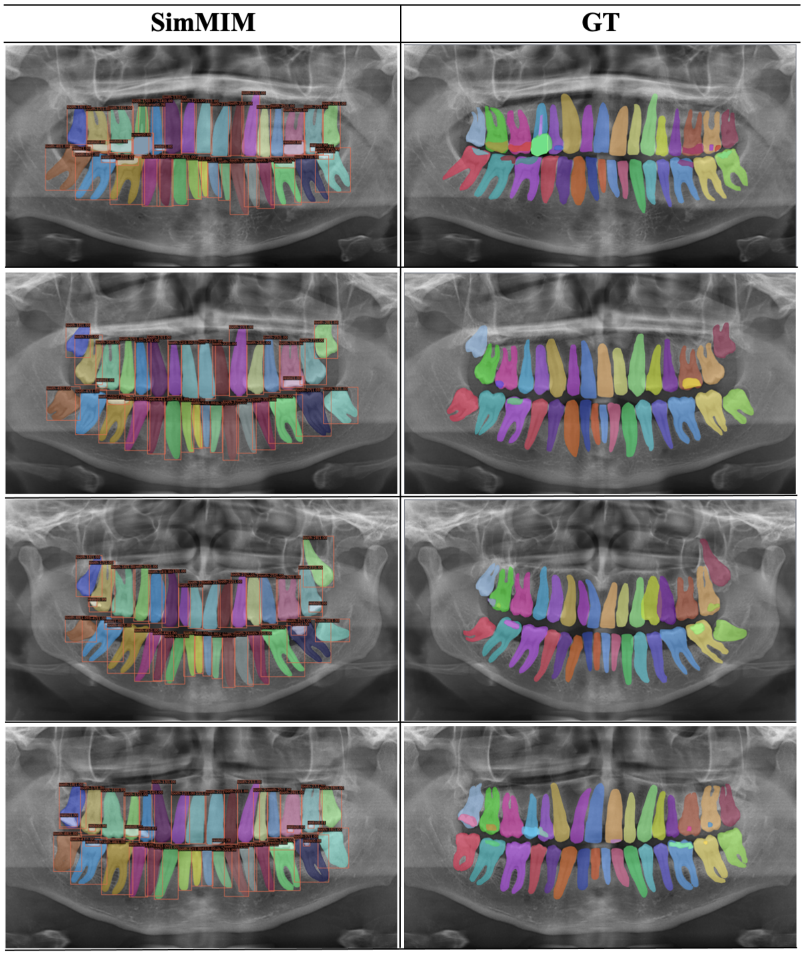


Figure 1.7. Qualitative results of detecting teeth and dental restorations and instance segmentation using SimMIM.

Fig. 1.7 displays qualitative results after augmenting dental restorations when Swin Transformer is pre-trained with SimMIM.

1.5.3 Pre-training time and memory consumption

Comparing UM-MAE to the SimMIM framework, the core advantage of UM-MAE is the memory and runtime efficiency. In Table 1.5, we show their clear comparisons based on Swin-B. It is observed that UM-MAE speeds up by about 2× and reduces

Method	Time	Memory
SimMIM	24.6 h	18.7 GB
UM-MAE	**12.5 h**	**6.7 GB**

Table 1.5.
The comparison of pre-training time and memory consumption.

the memory by at least 2× against SimMIM, where their performances under the downstream tasks show the opposite.

1.6 Conclusions

Two self-supervised learning methods were applied to Swin Transformer on dental panoramic radiographs: SimMIM and UM-MAE. The results of the masking-based method, SimMIM, obtained superior performance than UM-MAE, supervised and random initialization for detection of teeth, dental restorations, and instance segmentation. Based on this experiment, we can conclude that adjusting parameters, including mask ratio and pre-training epochs, is useful when applying SimMIM pre-training to the dental imaging domain for reliable outcomes. In addition, correcting the dataset annotations lead to further improvements that significantly surpass the available state-of-the-art results. Our plan for future work is to examine the efficacy of SimMIM pre-training in prognosis and outcome prediction tasks.

CHAPTER 2
ENHANCED MASKED IMAGE MODELING FOR DENTAL PANORAMIC RADIOGRAPHS

2.1 Introduction

The computer-assisted decisions are essential in dental practice to help dentists diagnose and plan for treatments. Dental imaging is a valuable tool that facilitates diagnosis and treatment plans, which is impossible through clinical examination and patient history only [1]. A Dental X-ray is a two-dimensional radiograph that captures the patient's entire mouth from ear to ear in a single image, including the upper and lower jaws and surrounding alveolar bone [2].

In dentistry, many teeth numbering systems provide a specific code for each tooth. Specifically, in this study, we utilize The Federation Dentiaure International numbering system (FDI), which is internationally known among dental care providers. It is a two-digit code where the first digit is given for each quadrant from 1 to 4 for permanent adult teeth. And the second digit is assigned for each tooth number based on its location in the jaw, starting from the middle front teeth (number 1) and moving back up to the third molar (number 8) [30].

Furthermore, dental restorations are used to restore the tooth's missing structure resulting from caries or trauma with full or partial coverage. Moreover, root canal fillings are utilized to fill the space of the root portion inside the tooth structure because of decay or other damage. In addition, orthodontic appliances apply force onto the teeth to be moved into the correct position; such appliances include but are not limited to bands, brackets, and retainers. The restorative materials and orthodontic appliances appear radiopaque in the X-rays and can be identified by dental practitioners [4].

Deep learning models are successful when trained with a large amount of data, however, a very limited number of dental radiographs is available for training. To mitigate this problem, we propose a new self-distillation and self-supervised learning combination for training a Swin Transformer [21] for dental panoramic X-rays analysis.

Recently, self-supervised learning methods with masked image modeling (MIM) such as SimMIM [22], MAE [25] and UM-MAE [23] are shown to be effective in pre-training deep learning models, like Transformers [21,37]. However, only SimMIM and UM-MAE are applicable to Swin Transformer. Generally, the idea of MIM methods is to mask some patches before they are fed into the Swin encoder and predict the original patches to gain more understanding of the images. However, the patches' location is important in dental panoramic X-rays for a predictable outcome. SimMIM maintains the patches location known to both the encoder and decoder, while UM-MAE drops the location information unknown to the encoder, which may induce inaccuracy. Therefore, SimMIM pre-training is selected in this study.

Inspired by [22,25,38], we hypothesize that the Swin encoder can be improved by transferring knowledge obtained by decoded visible patches to their encoded peers through self-distillation. We believe that the visible patches in the decoder contain more knowledge than the ones in the encoder. Moreover, similar to [22, 25] and unlike [38], we found out that predicting the masked area only outperforms predicting all image pixels.

The proposed SD-SimMIM is trained on the same dataset as the downstream tasks, excluding the test dataset. We apply SD-SimMIM on dental panoramic X-rays for teeth numbering, detection of dental restorations and orthodontic appliances, and instance segmentation tasks. It is shown that SD-SimMIM performs better than other self-supervised learning methods.

Although previous studies investigated teeth numbering [12] and segmentation of dental restorations [39], there is no comprehensive dataset that simultaneously studied orthodontics appliance segmentation. We believe that the inclusion of segmentation of orthodontics appliances increases the complexity of the computer vision problem because of class quantity and class imbalance. Therefore, we augment the existing dataset introduced in [12] under dental expert supervision. We further expand the dataset by developing annotations for orthodontics appliances, including bands, brackets, and retainers. The labeling process led to a unique high-quality augmented dataset. Our data will be available, upon request, under the name **Dent**al an**alysis** (Dentalysis) annotations. Our main contributions are twofold:

- We introduce SD-SimMIM, a self-distillation enhanced SimMIM. It aims to boost the feature representation on top of SimMIM to alleviate the demands on large data for dental panoramic radiographs, and further help downstream tasks.
- The augmented dataset increases performance, while added labeling of orthodontics appliances extends the horizon of possible dental applications.

2.2 Methods

Fig. 7.2 illustrates our SD-SimMIM framework. It includes two modules, masked image modeling (MIM) and visible image modeling (VIM). MIM generates self-supervised learning on unlabeled data by masking some image patches, while VIM imposes self-distillation constraints on visible patches for better and more powerful encoder learning. Hence, VIM enhances the original SimMIM, particularly for dental panoramic radiographs.

2.2.1 SimMIM

SimMIM framework includes four components: patchifying and masking, encoder, decoder, and prediction target.

Patchifying and masking designs how to select the area to mask, and how to implement masking of the selected area. The Patchifying first divides the input image x into N patches. Then, it flattens each patch to a token (a one-dimensional vector of visual features) with length D. Hence, the formulation of the representation of

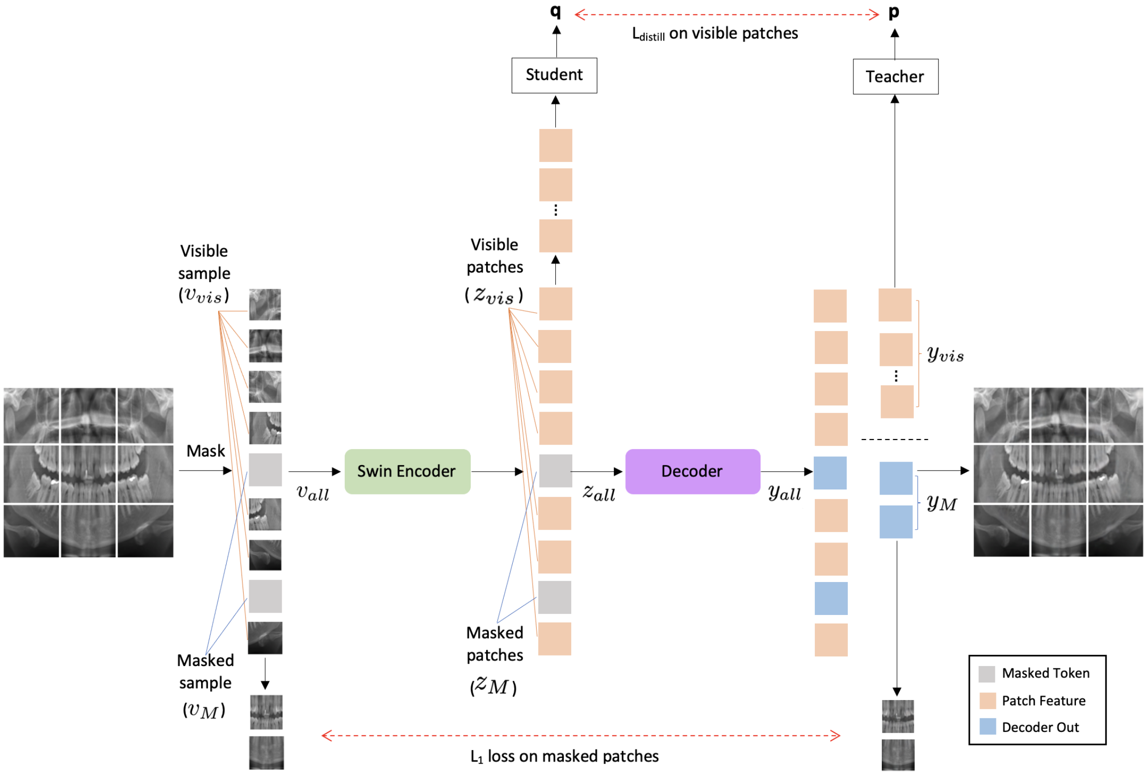


Figure 2.1. Our SD-SimMIM framework. Alongside the original SimMIM, we benefit from decoded visible patches (as the teacher) and transfer knowledge to their peers after encoding. (Best viewed in color)

all patches is $v_{all} \subseteq \mathbb{R}^{N \times D}$. Next, the masking randomly divides the patches into two sets with respect to a masking ratio M, more precisely $v_{all} \subseteq \mathbb{R}^{N \times D} \rightarrow v_{vis} \subseteq \mathbb{R}^{N' \times D}, v_M \subseteq \mathbb{R}^{\tilde{N}' \times D}$ where $N' = N \times (1 - M), \tilde{N}' = N \times M$. v_{all} will be the input to the encoder and v_M are the labels.

Encoder takes v_{all} as input, and extracts latent feature from visible patches. First, it maps D dimensions of tokens to D' with a linear projection, and then these patch tokens are processed via Swin Transformer blocks to get latent representation vectors of patches $z_{vis} \subseteq \mathbb{R}^{N' \times D'}$ and masked tokens $z_M \subseteq \mathbb{R}^{\tilde{N}' \times D'}$.

Decoder takes $z_{all} \subseteq \mathbb{R}^{N \times D'}$ as input, and learns low-level representation from visible patches for image reconstruction. Hence, the decoder output $y_{all} \subseteq \mathbb{R}^{N \times D'}$ will divided into y_{vis} and y_M, as visible and masked tokens, respectively.

Prediction target defines the form of original signals to predict. First, we consider the original masked tokens after normalizing $Y_M = Norm(v_M)$ as our prediction target. The decoder applies a linear layer to align y_M and Y, i.e. $y_M \rightarrow y'_M$. The L_1 loss is computed between the predicted masked tokens y'_M and the original masked tokens after normalization Y_M as described in Eq. 2.1.

$$L_1 = \ell_1(y'_M, Y_M), \quad y'_M, Y_M \subseteq \mathbb{R}^{\tilde{N}' \times D} \tag{2.1}$$

2.2.2 Self-distillation

Knowledge Distillation is the process of transferring knowledge from a large model to a smaller one [40]. Previous studies apply it to the vectors at various depths within the same network, either a convolutional neural network (CNN) [41] or a Vision Transformer (ViT) [38]. Hence, knowledge is distilled from deep layers to shallow layers, augmenting the feature representation of shallow layers. Considering the imbalance of knowledge, we found that this is exactly how knowledge in the visible tokens can be transferred from the decoder to the encoder through this distillation paradigm. Particularly, there are two types of latent representation vectors for visible tokens in SimMIM, i.e. z_{vis} outputted from the encoder and y_{vis} from the decoder.

We treat z_{vis} as shallow features and y_{vis} as deeper features in the self-distillation framework [41]. We use a 3-layer MLP over these two vectors, resulting in probability distributions over K dimensional feature denoted by q and p, respectively. Each of them is normalized with a $Softmax$ over the feature dimension. Thus, we learn to match these distributions by minimizing the cross-entropy loss as shown in Eq. (2.2).

$$\begin{aligned} &q = MLP(z_{vis}), \quad p = MLP(y_{vis}) \\ &q' = Softmax(q), \quad p' = Softmax(p) \\ &L_{distill} = -p' log(q') \end{aligned} \tag{2.2}$$

The total loss is formulated as shown in Eq.(2.3).

$$L = \alpha L_1 + (1 - \alpha) L_{distill} \tag{2.3}$$

where α is the empirically defined scaling factor (in this study, α is equal to 0.2).

2.3 Experiments

2.3.1 Dataset

Detection, Numbering, and Segmentation (DNS) [12] is a dental panoramic X-rays dataset consisting of 543 annotated images with ground truth segmentation labels, including numbering information based on the FDI teeth numbering system. Each image size is 1991x1127 pixels. The dataset annotations from [39] do not contain any segmentation of orthodontic appliances. Therefore, we contribute to expanding the dataset by developing segmentation for orthodontic appliances and introducing three more classes, namely bands, brackets, and retainers. This process was under a supervision of a dentist using the COCO-Annotator tool [34]. We attended weekly meetings where related issues and questions were discussed. In the end, the annotations were reviewed to assure quality and avoid systematic and random errors. Fig. 2.2 presents samples of segmentation of orthodontics appliances. We believe this is the most inclusive dataset for segmenting teeth, dental restorations, and orthodontic appliances

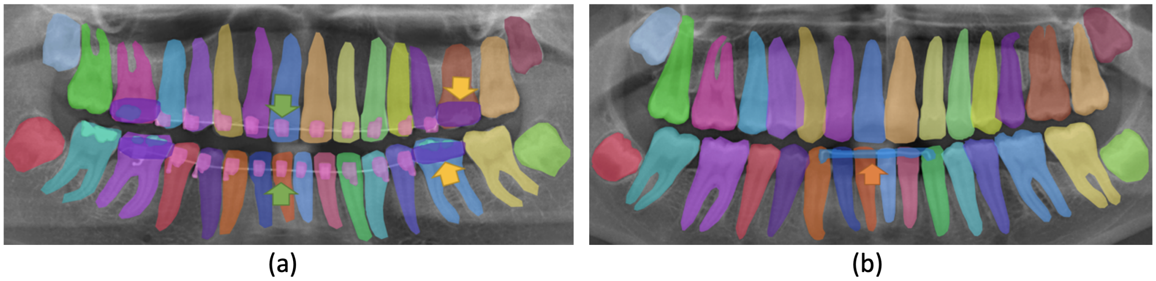


Figure 2.2. Samples of segmentation of orthodontics appliances, a) shows examples of bands (yellow arrows) and brackets (green arrows), and b) a retainer (orange arrow). (Best viewed in color)

in dental panoramic radiographs. Our data will be available upon request, namely Dentalysis annotations.

2.3.2 Evaluation metric

For all our experiments, we split the data into five folds, each containing about 20% of the images. One fold is fixed as the test set (111 images), and the other four folds (108 images each) compose the training and validation datasets in a cross-validation manner. This process is repeated five times. The evaluation metric we adopt is the Average Precision for object detection and instance segmentation models.

2.3.3 Implementation details

Our experiments are implemented based on the PyTorch [35] framework and trained with NVIDIA Tesla Volta V100 GPUs. In all experiments, the batch size equals the total number of training samples, which is 432. The input images are all resized to 800×600 pixels. We utilize the AdamW [36] optimizer in all experiments.

Data augmentation. We apply noise addition and horizontal flipping, which turns left teeth numbers into right teeth numbers and vice-versa.

SD-SimMIM pre-training. We follow a similar protocol to SimMIM [22] to train our SD-SimMIM. We use Swin-B [21] as the encoder and a lightweight decoder with a linear projection. The base learning rate is set to 8e-4, weight decay is 0.05, $\beta 1 = 0.9$, $\beta 2 = 0.999$, with a cosine learning rate scheduler. We use a random MIM with a patch size of 16×16 and a mask ratio of 20%. We apply the L2-normalization bottleneck [28] (dimension 256 for the bottleneck and K dimensions equals 4096) as the projection head in self-distillation. This model was pre-trained for 100 epochs with a warm-up for 10 epochs. The target image size is 800×600.

Task fine-tuning. We utilize single-scale training. The initial learning rate is 0.0001, and the weight decay is 0.05.

2.3.4 Quantitative results

Table 3.2 shows the results of different methods on the dataset for teeth numbering, detection of dental restorations, and instance segmentation only. As a baseline (the first row, called Supervised), Swin-B [21] is trained using the dataset without self pre-training to demonstrate the improvement obtained by self-supervised learning. The original Swin-B was trained on the Image Net dataset with 1000 classes denoted as (IN-1K). The CNN-based network, PANet [12], reports a result that is worse than Swin-B. This can be explained as the difference in the network capacity, where ResNet-50 is used as the backbone in PANet. As a comparison, the way SimMIM uses image reconstruction is obviously more suitable than UM-MAE for dental images. The reason may be attributed to the fact that the location of the patches is essential in dental radiographs for a predictable outcome. SimMIM maintains the location of the patches known to both the encoder and decoder, while UM-MAE drops the location information, which may induce inaccuracy. The proposed SD-SimMIM shows steady improvements over SimMIM and yields the best performance. Hence transferring decoder information to the encoder with self-distillation improves the outcomes of self-learning. We also observe that similar to [22,25] and unlike [38], our results show that predicting the masked area only outperforms predicting all image pixels for both SimMIM and our SD-SimMIM.

Table 2.1.
Results of teeth numbering, detection of dental restorations, and instance segmentation only. * denotes L_1 loss is computed on the whole image. AP^{box} and AP^{mask} indicate Average Precision for object detection and instance segmentation, respectively.

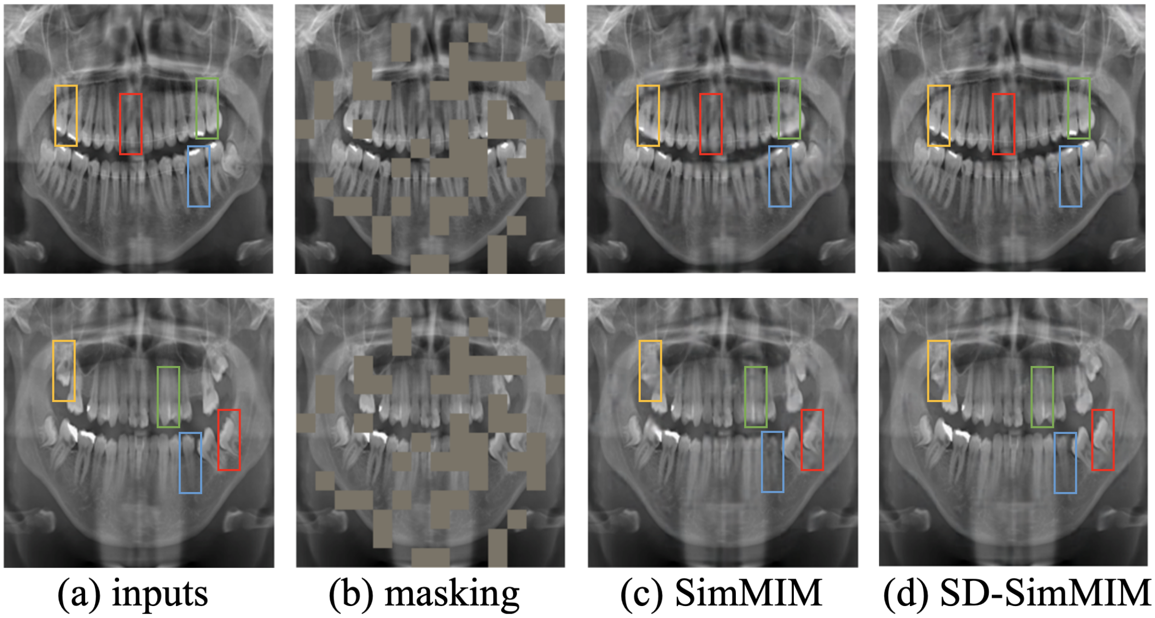
Initialization	Backbone	Pre-train Data	AP^{box}	AP^{mask}
Supervised	Swin-B	IN-1K w/ Labels	80.3	79.2
PANet [12]	ResNet-50	IN-1K w/ Labels	76.8	75.1
UM-MAE [39]	Swin-B	IN-1K	88.3	85.7
SimMIM*	Swin-B	IN-1K	89.9	88.5
SimMIM [39]	Swin-B	IN-1K	90.4	88.9
SD-SimMIM*	Swin-B	IN-1K	90.7	89.6
SD-SimMIM	Swin-B	IN-1K	**92.4**	**90.2**

Table 2.2 shows results after including the annotations of orthodontics appliances. The proposed SD-SimMIM method achieves the highest performance of 92.7% and 90.8% on detecting teeth, dental restorations and orthodontics appliances, and instance segmentation, respectively. Again it is worth noting that the best performance is gained when computing the loss on the masked areas only.

2.3.5 Qualitative results

To illustrate the effectiveness of adding self-distillation to simMIM, we provide some visualization examples. Firstly, we are curious about the results of image reconstruction. Fig. 2.3 presents two reconstruction examples using our SD-SimMIM. As shown, SD-SimMIM obtains a slightly better reconstruction than SimMIM. It proves that self-distillation reinforces the learning capability of the SimMIM encoder.

Secondly, Fig. 2.4 displays four different qualitative samples of improved performance when the Swin Transformer is pre-trained with SD-SimMIM for teeth number-



(a) inputs

(b) masking

(c) SimMIM

(d) SD-SimMIM

Figure 2.3. Images reconstructed by SimMIM and SD-SimMIM. SD-SimMIM shows a clearly better reconstruction than SimMIM. The color boxes highlight their details. (Best viewed in color)

Table 2.2.
Results after including orthodontics appliances. * denotes L_1 loss is computed on the whole image.

Initialization	Backbone	Pre-train Data	AP^{box}	AP^{mask}
Supervised	Swin-B	IN-1K w/ Labels	81.9	80.1
SimMIM*	Swin-B	IN-1K	90.3	88.8
SimMIM [39]	Swin-B	IN-1K	90.8	89.4
SD-SimMIM*	Swin-B	IN-1K	91.2	90.0
SD-SimMIM	Swin-B	IN-1K	**92.7**	**90.8**

ing, detecting dental restorations, orthodontic appliances, and instance segmentation. Those improvements in detection and segmentation agree with the quantitative results in Section 5.5.2.

2.4 Conclusions

We propose SD-SimMIM, a novel self-distillation scheme that transfers knowledge from the decoder to the encoder to guide a more effective visual pre-training. The quantitative and qualitative results present the benefits of our SD-SimMIM, which is a promising tool for the analysis of dental radiographs. For future work, we will evaluate our SD-SimMIM on different downstream tasks such as detecting dental disease on dental bitewing radiographs.

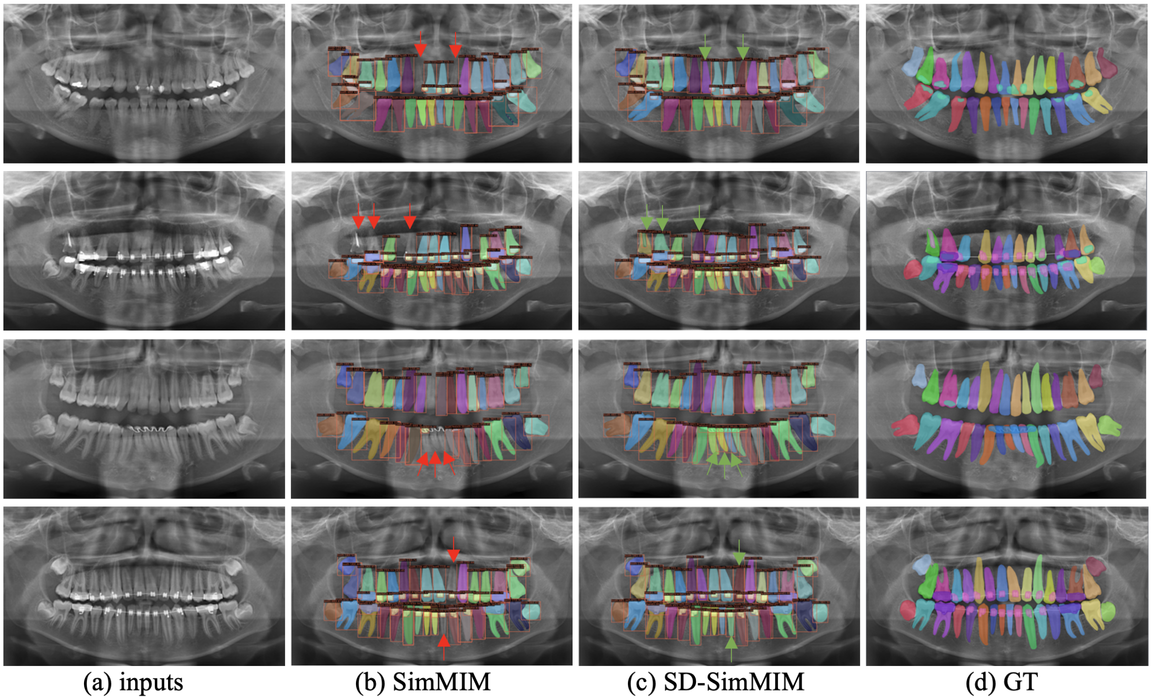


Figure 2.4. Qualitative results of detection and instance segmentation. Note that teeth detection and instance segmentation are missing (red arrows) when Swin Transformer is pre-trained with SimMIM compared to the ones produced by Swin Transformer pre-trained with SD-SimMIM architecture (green arrows). (Best viewed in color.)

CHAPTER 3
TOOTH SEGMENTATION FROM INTRA-ORAL 3D SCANS

3.1 Introduction

Computer-aided design (CAD) tools have gained significant popularity in modern dentistry, especially in orthodontic or prosthetic CAD systems, for accurate treatment planning. Advanced intra-oral scanners (IOS) are widely used to obtain precise digital surface models of dentition. The IOSs produce 3D surface reconstructions of the teeth either in the form of a point cloud or in a mesh format, or both. These models are invaluable in simulating teeth extraction, movement, deletion, and rearrangement, enabling dentists to predict treatment outcomes with greater ease. Consequently, digital teeth models have the potential to alleviate dentists' time-consuming and tedious tasks.

Tooth segmentation from intra-oral scans is a key step in computer-aided dentistry. It can help in recognizing and classifying different dental/oral conditions like gingivitis, caries, and white lesions. While tooth segmentation and labeling is a first step in digital dentistry, it is difficult due to the inherent similarities between teeth shapes and the ambiguity surrounding their positions on jaws. Furthermore, variations in teeth position and shape across different individuals present additional challenges in this process. Other challenges involved in tooth mesh segmentation, such as - crowded teeth, misaligned teeth, and missing teeth. The size of teeth can also vary widely across meshes. The second and third molars may evade capturing due to their being in the deep intra-oral regions. Or the second/third molar might not be fully formed. Different teeth and gum conditions, like recession, enamel loss, etc, can also alter the appearance of the teeth significantly.

Furthermore, the manual process of segmenting and labeling teeth is a time-consuming task that can potentially miss important data. This has led to a growing interest in leveraging computer vision and computer science to automate these processes. Multiple automatic tooth mesh segmentation algorithms have been proposed [42–44]. They include convolutional neural networks (CNNs) for teeth segmentation from 3D intra-oral scans [45–50]. Recently, the use of CNNs in the analysis of medical images has experienced significant growth due to advancements in computational hardware, algorithms, and expansion in the amount of data [5]. However, CNNs are constrained in their overall capability due to the inherent inductive biases they possess [37].

Recent advancements in self-supervised learning have demonstrated the effectiveness of masked image modeling (MIM) [24, 25, 51] as a pre-training strategy for the Vision Transformer (ViT) [37] and the hierarchical Vision Transformer using shifted windows (Swin) [21,39,52]. MIM involves the masking and subsequent reconstruction of image patches, allowing the network to infer the masked regions by leveraging contextual information. We believe that the ability to aggregate contextual information is crucial in the context of 3D dental scan analysis. Among various MIM frameworks, the Masked Autoencoder (MAE) [25] stands out as a simple yet effective approach. MAE employs an encoder-decoder architecture, with a ViT encoder that receives only visible tokens and a lightweight decoder that reconstructs the masked patches using the encoder's patchwise output and trainable mask tokens.

This paper introduces a novel approach to teeth segmentation in 3D dental scans called Dental Masked Autoencoder (DentalMAE) based self pre-training, which works for 3D dental meshes analysis. We apply DentalMAE pre-training on the same dataset, referred to as the train set, which is used for the downstream task. We term this approach self pre-training, which is particularly advantageous in scenarios where acquiring suitable pre-training data is challenging. Additionally, self pre-training eliminates the domain discrepancy between the pre-training and fine-tuning stages by unifying the training data. Our experiments focus on teeth segmentation in 3D intra-oral scans [53].

Specifically, We extend the self-supervised learning framework of the mesh masked autoencoder (MeshMAE) transformer [54]. While the MeshMAE loss measures the quality of reconstructed masked mesh triangles, the loss of the proposed DentalMAE evaluates the predicted deep embeddings of masked mesh triangles. After pre-training, the decoder is discarded, and the encoder is applied to the downstream task, i.e., teeth segmentation. We compare three ViT Transformer initializations, including our proposed DentalMAE, MeshMAE [54], and a mesh transformer without any self-pre-training. The experimental results demonstrate that DentalMAE self-pre-training significantly enhances dental scan segmentation performance compared to the baselines. Our main contributions are threefold:

- We utilize self-supervised learning with masked autoencoders to alleviate the problem of small data for 3D intra-oral scans.
- We replace the MeshMAE reconstruction of masked mesh patches with the reconstruction of mesh patch embeddings. Hence our loss is simply the L_2 distance between the predicted and computed embeddings over the masked patches, which is much simpler than the loss used by MeshMAE.
- Our proposed method leads to a significant performance improvement. DentalMAE outperforms all state-of-the-art methods on the tooth mesh segmentation task.

3.2 Related work

Most of the existing research in this field can be categorized into two groups: approaches based on handcrafted features and approaches based on learning.

3.2.1 Handcrafted features-based approaches

Previous methods primarily focused on extracting manually designed geometric features to segment 3D dental scans. These methods can be classified into three types: surface curvature-based methods, contour line-based methods, and harmonic field-based methods. Surface curvature is particularly useful for describing tooth surfaces and identifying tooth/gum boundaries in IOS. Zhao et al. [42] proposed a semi-automatic teeth segmentation method based on curvature thresholding, followed by gum separation and identification of 3D teeth boundary curves. Another approach by Yuan et al. [55] used minimum surface curvature calculation to extract individual teeth regions and separate them. Wu et al. [44] presented a morphological skeleton-based method for teeth segmentation in IOS, utilizing area growing operations. Similarly, Kronfeld et al. [56] introduced a system that detects tooth-gingiva boundaries using active contour models. Contour line-based methods involve manual selection of tooth boundary landmarks, followed by contour line generation based on geodesic information, as demonstrated in studies such as Sinthanayothin et al. and Yaqi et al. [57,58]. Harmonic field methods require less user interaction, as they allow a limited number of surface points to be selected prior to the segmentation process, as seen in studies by Zou et al. [59] and Liao et al. [60].

However, these approaches have limitations in achieving robust and fully automated segmentation of dental 3D scans. Setting the optimal threshold for surface curvature-based methods is challenging, and they are sensitive to noise. Incorrect threshold selection can significantly impact segmentation accuracy, leading to over- or under-segmentation. Moreover, the manual threshold selection makes these methods unsuitable for fully automatic segmentation. Contour line-based methods are time-consuming, difficult to use, and rely heavily on human interaction. Harmonic field techniques involve complex and computationally intensive preprocessing steps.

3.2.2 Learning-based approaches

Recent advancements in deep learning techniques have shifted the focus of teeth segmentation from handcrafted features to learned features. It is now widely recognized that data-driven feature extraction, using techniques like convolutional neural networks (CNNs), outperforms handcrafted features in various computer vision tasks, including object detection [61] and image classification [62]. The same applies to 3D teeth segmentation and labeling. Learning-based approaches can be divided into two main categories based on the input data: 2D image segmentation and 3D mesh segmentation.

For 2D image segmentation, CNNs have been extensively used to extract relevant features. Cui et al. [63] introduced a two-stage deep supervised neural network architecture for tooth segmentation and identification in Cone-Beam Computed Tomography (CBCT) images. They employed an autoencoder CNN to extract edge maps from CBCT slices, which were then fed into a Mask R-CNN network for tooth segmentation and recognition. Similarly, Miki et al. [64] fine-tuned a pre-trained AlexNet network on CBCT dental slices for automatic teeth classification. Rao et al. [65] proposed a symmetric fully convolutional residual neural network for tooth segmentation in CBCT images. They incorporated dense conditional random field techniques and a deep bottleneck architecture for teeth boundary smoothing and segmentation enhancement, respectively. Zhang et al. [66] isomorphically mapped 3D dental scans into a 2D harmonic parameter space and used a CNN based on the U-Net architecture for tooth image segmentation.

Learning-based methods applied directly to 3D dental meshes have also been explored. Sun et al. [67] used a graph CNN-based architecture called FeaStNet for automated tooth segmentation and labeling from 3D dental scans. They extended this architecture to propose an end-to-end graph convolutional network-based model that achieved tooth segmentation and dense correspondence in 3D dental scans. Xu et al. [68] introduced a multi-stage framework based on a deep CNN architecture for 3D dental mesh segmentation. They employed two independent CNNs for teeth-gingiva and inter-teeth labeling. Zanjani et al. [69] proposed an end-to-end deep learning system based on the PointNet network architecture for semantic segmentation of individual teeth and gingiva from point clouds. They also used a secondary neural network as a discriminator in an adversarial learning setting to refine teeth labeling. Lian et al. [47] modified the PointNet architecture by incorporating graph-constrained learning modules to extract multi-scale local contextual features for teeth segmentation and labeling in 3D intra-oral scans. Tian et al. [70] introduced a preprocessing step that encoded input 3D scans using sparse voxel octree partitioning. They then employed three-level hierarchical CNNs for the segmentation process and another two-level hierarchical CNNs for teeth recognition. Other studies, such as Cui et al. [71] and Zanjani et al. [72], proposed pipeline-based architectures combining multiple CNNs for teeth localization, segmentation, and labeling. Ma et al. [73] suggested a deep neural network architecture for pre-detected teeth classification based on adjacency similarity and relative position feature vectors, explicitly modeling spatial relationships between adjacent teeth.

Zhao et al. [74] proposed an end-to-end network utilizing graph attentional convolution layers and a global structure branch for fine-grained local geometric feature extraction and global feature learning from raw mesh data. These features were fused to perform segmentation and labeling tasks. In another study, Zhao et al. [75] introduced a two-stream graph convolutional network (TSGCN). The first stream captured coarse structures of teeth from 3D coordinate information, while the second stream extracted distinctive structural details from normal vectors. To address the reliance on expensive point-wise annotations in current learning-based methods, Qiu et al. [76] presented the Dental Arch (DArch) method for 3D tooth segmentation using weak low-cost annotated data. The DArch consists of two stages: tooth centroid detection and segmentation. It generates the dental arch using Bezier curve regression and refines it using a graph-based convolutional network (GCN).

To the best of our knowledge, there have been no studies in the literature that specifically employ transformer models, such as the Vision Transformer (ViT) [37], for 3D dental scan analysis. Additionally, the application of self-supervised learning techniques to ViT on intra-oral scans is also unprecedented.

Transformer models, originally introduced in natural language processing tasks [77], have shown remarkable success in various computer vision domains, including image classification, object detection, and image segmentation. The ViT architecture, in particular, has gained attention for its ability to effectively process 2D images by leveraging self-attention mechanisms.

However, the application of transformer models to 3D dental scans and the use of self-supervised learning techniques on intra-oral scans have not been explored in the existing literature. This indicates a research gap and an opportunity to investigate the potential benefits and challenges of utilizing ViT and self-supervised learning in the context of 3D dental scan analysis.

By applying self-supervised learning to ViT on intra-oral scans, it becomes possible to mitigate the limited number of available intra-oral scans. This can help overcome the limitations of traditional supervised learning approaches, which rely heavily on large data for training. Self-supervised learning enables the model to learn from the inherent structure and properties present in the data, leading to improved generalization and potentially reducing the need for extensive manual labeling.

The application of transformer models and self-supervised learning techniques to 3D dental scans, specifically intra-oral scans, has the potential to advance the field by providing new insights and improved performance in tasks such as segmentation, labeling, and analysis of dental structures. Further research in this direction could pave the way for more accurate and efficient automated dental scan analysis, benefiting various clinical applications and oral healthcare practices.

3.3 Methods

In this paper, we use the Mesh Transformer framework for tooth mesh segmentation, which extends the Vision Transformer to mesh analysis. We propose a novel self-supervised learning pre-training strategy, which is based on mesh masked autoencoding. Fig. 7.2 illustrates the DentalMAE framework. DentalMAE divides the input mesh into non-overlap patches, these patches are embedded using an MLP, and certain random patches are replaced with mask tokens. Only the visible patches are utilized by the ViT encoder. Subsequently, the mask tokens are combined with the encoded embeddings and are input to the decoder. The primary objective of the decoder is to reconstruct the vertices and face features of the masked patches, followed by the prediction of the patch embeddings of the masked patches. We do the two-stage pro-

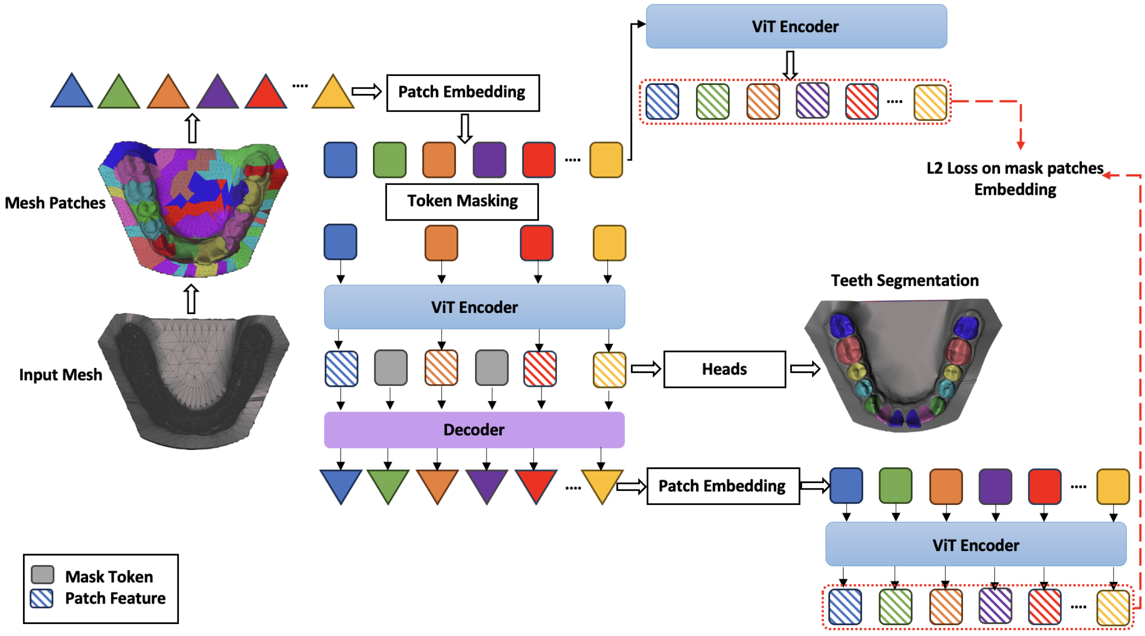


Figure 3.1. The teeth segmentation pipeline for DentalMAE self-pre-training. Initially, the input mesh is divided into non-overlap patches. These patches are then embedded using an MLP. During the pre-training phase, the patch embeddings are randomly masked, and only the visible embeddings are utilized by the transformer. Subsequently, the masked embeddings are combined with the encoded embeddings and sent to the decoder. The objective of the decoder is to reconstruct the vertices and face features of the masked patches, followed by the prediction of the patch embeddings of the masked patches. The L_2 loss is used to compare the masked patch embeddings. After the completion of pre-training, the decoder is discarded, and the encoder is employed for segmentation.

cess of reconstructing vertices and face features followed by computing embeddings because it performs better than directly predicting the embeddings as shown in the supplementary materials. Compared to MeshMAE [54], its loss measures the quality of reconstructed masked mesh triangles, while the loss of the proposed DentalMAE evaluates the predicted deep embeddings of masked mesh triangles. Following the pre-training phase, the decoder is discarded, and the encoder is employed for the specific task of tooth segmentation.

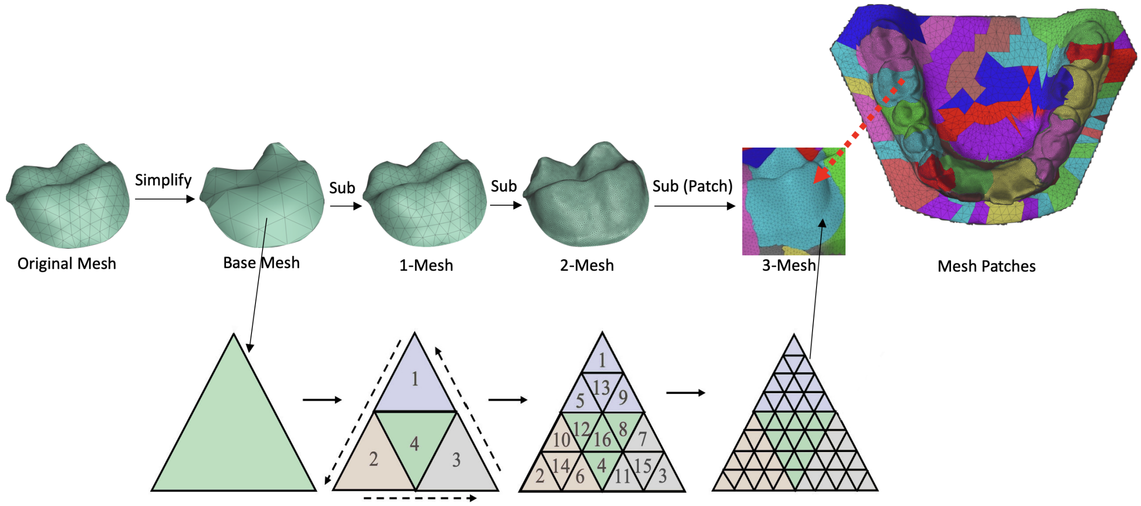


Figure 3.2. The remeshing operation involves several steps. Initially, the input mesh undergoes a simplification process. Subsequently, a mapping is established between the original mesh and the base mesh. The base mesh is then subdivided three times, and the newly generated vertices are projected back onto the input mesh.

3.3.1 Mesh Transformer

Mesh Patch Split. The faces of a 3D mesh establish connections between vertices, allowing us to utilize geometric information from each face to represent their features. Similar to SubdivNet [78], we define a 10-dimensional vector for each face f_i comprising the face area (1-dim), three interior angles of the triangle (3-dim), face normal (3-dim), and three inner products between the face normal and three vertex normals (3-dim).

Transformers, with their self-attention-based architectures, simplify the process of designing feature aggregation operations for 3D meshes. However, applying self-attention to all faces incurs a prohibitively high computational cost due to quadratic complexity. To overcome this, the faces are grouped into non-overlapping patches before applying transformers. Unlike regular image data that can be divided into grid-like patches, mesh data is irregular, and faces are typically unordered.

To address this challenge, we utilize a "re-meshing" step to regularize and hierarchically structure the original mesh. We employ the MAPS algorithm [79] to simplify the mesh into a coarser base mesh with a varying number of faces N faces within a specific range ($96 \leq N \leq 256$ in our experiments). Although less accurate in shape representation, the resulting base mesh serves as a foundation. To refine it, we further subdivide all faces in the base mesh t times in a 1-to-4 manner, resulting in a more detailed mesh called $t-$mesh. By grouping the faces of the $t-$mesh corresponding to the same face in the base mesh, we create non-overlapping patches. In our implementation, we perform three subdivisions, yielding patches consisting of 64 faces each. The process is illustrated in Fig. 3.2.

Transformer Backbone. The transformer serves as the backbone network for the Mesh Transformer. It consists of multi-headed self-attention layers and feed-forward network (FFN) blocks. To represent each patch, we concatenate the feature vectors of the constituent faces belonging to that patch. The order of concatenation is determined by the re-meshing process, which guarantees a consistent and predictable face order. Consequently, an MLP is employed to project the feature vector of each patch into a representation denoted as $\{e_i\}_{i=1}^{g}$, where g denotes the number of patches. These representations serve as inputs to the transformer.

In addition to shape information captured by the input features, transformer-based methods often rely on positional embeddings to provide spatial information. Since mesh data contains 3D spatial coordinates for each face, we leverage the center 3D coordinates of the faces to compute the positional embeddings. To accomplish this, we calculate the center point coordinates $\{c_i\}_{i=1}^{g}$ for each patch and apply an MLP to obtain the positional embedding $\{p_i\}_{i=1}^{g}$ associated with each patch.

Formally, the input embeddings $X = \{x_i\}_{i=1}^{g}$ are defined as the combination of the patch embeddings $E = \{e_i\}_{i=1}^{g}$ and positional embeddings $P = \{p_i\}_{i=1}^{g}$. This results in an overall input sequence denoted as $H^0 = x_1, x_2, ..., x_g$. The encoder network consists of L layers of transformer blocks, and the output of the last layer $H^L = h_1^L, ..., h_g^L$ represents the encoded representations of the input patches.

3.3.2 Mesh Pre-training Task

In this section, we provide a detailed description of the mesh pre-training task, which employs a masked modeling strategy based on the Mesh Transformer architecture. The task aims to predict deep embeddings of masked mesh triangles from embeddings of visible mesh triangles. We outline the components of the pre-training task, including the encoder and decoder networks, masked sequence generation, and prediction.

Encoder and Decoder. The encoder and decoder networks used in the pre-training task are composed of several transformer blocks. The Mesh Transformer serves as the encoder, consisting of 12 layers, while a lightweight decoder with 6 layers is employed. During pre-training, a predefined masking ratio is applied to randomly mask a subset of patches in the input mesh. The visible patches are fed into the encoder, and a shared mask embedding is used to replace the masked embeddings in the input before feeding them into the decoder. The positional embeddings are added to both the masked and visible patches to provide location information. It is important to note that the decoder is only used during pre-training for mesh reconstruction tasks, while the encoder is utilized in downstream tasks.

Masked Sequence Generation. Mesh embeddings, represented by E, have corresponding indices denoted as I. Following the MAE approach, we randomly mask a subset of patches by sampling indices I_m from I with a ratio r. Masked embeddings are represented as E_m, while unmasked embeddings are denoted as E_{um}. We replace the masked embeddings E_m with a shared learnable mask embedding E_{mask} without altering their positional embeddings. Finally, the corrupted mesh embeddings E_c are formed by combining E_{um} with the sum of E_{mask} and positional embeddings p_i for each index i in I_m. These corrupted embeddings are then inputted into the encoder for further processing.

Prediction. MeshMAE [54] recovers the shape of the masked patches as the reconstruction target. It predicts 3D relative coordinates of vertices to match the ground truth positions, where the reconstruction loss is calculated using the Chamfer distance [80] between the predicted relative coordinates and the ground truth relative coordinates. It also predicts the face-wise features using a linear layer behind the decoder. It uses face-wise mean squared error (MSE) loss to evaluate the reconstruction effect of the features.

The overall optimization objective of MeshMAE combines the Chamfer distance loss $\mathcal{L}_{CD}$ and the MSE loss $\mathcal{L}_{MSE}$ to $\mathcal{L} = \mathcal{L}_{MSE} + \lambda \cdot \mathcal{L}_{CD}$, where λ is the loss weight. In contrast, our loss is simpler in that it does not require any meta parameter λ. We simply compute the L_2 loss between the original and predicted embeddings of the mask triangle patches.

3.4 Experiments

3.4.1 Dataset

We use the public dataset 3D Teeth Seg Challenge 2022 [53]. There are a total of 1800 3D intra-oral scans collected for 900 patients covering their upper and lower jaws separately. They are separated into training (1200 scans, 16004 teeth) and test data (600 scans, 7995 teeth). The task is tooth segmentation from the 3D dental model. Throughout the paper, we use the color coding shown in Fig. 3.3 to visualize the teeth labels. There are 8 different semantic parts, indicating the central incisor (T7), lateral incisor (T6), canine/cuspid (T5), 1st premolar/bicuspid (T4), 2nd premolar/bicuspid (T3), 1st molar (T2), 2nd molar (T1), and background/gingiva (BG).

3.4.2 Evaluation metric

We use Dice Score(DSC), Overall Accuracy (OA), sensitivity (SEN), and Positive Predictive Value (PPV) to evaluate the performance of our model.

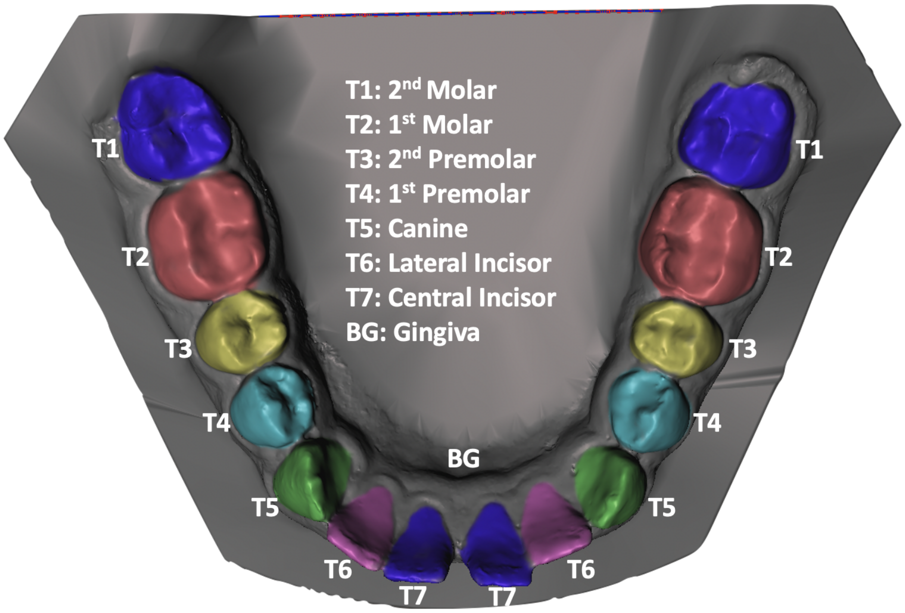


Figure 3.3. Tooth segmentation and the corresponding color coding.

3.4.3 Implementation details

Data Pre-processing. The dataset is processed by the re-meshing operation, and the face labels are obtained from the mapping between the re-meshed data and the raw meshes using the nearest face strategy.

Data Augmentation. We employ three data augmentation techniques: 1) random rotation, 2) random translation, and 3) random rescaling. By applying these techniques, we generate 40 augmented versions for each data point, resulting in the creation of 40 additional samples for every jaw scan.

Training Details. For pre-training, We utilize ViT-Base [37] as the encoder network with very slight modification, e.g., the number of input features' channels. And following [25], we set a lightweight decoder, which has 6 layers. We employ an AdamW optimizer, using an initial learning rate of 1e-4 with a cosine learning schedule. The weight decay is set as 0.05, and the batch size is set as 32. We set the same encoder network of pre-training in the downstream task. For our segmentation

Method	BG	T1	T2	T3	T4	T5	T6	T7
PointNet [81]	0.947	0.793	0.920	0.895	0.925	0.903	0.909	0.933
PointNet++ [82]	0.924	0.780	0.903	0.876	0.883	0.837	0.782	0.837
DGCNN [83]	0.968	0.847	0.944	0.936	0.945	0.941	0.939	0.947
MeshSegNet [47]	0.922	0.712	0.799	0.775	0.860	0.831	0.684	0.794
MeshSegNet+GCO [47]	0.957	0.850	0.904	0.902	0.926	0.879	0.778	0.906
TSGCNet [49]	0.962	0.642	0.915	0.916	0.945	0.937	0.916	0.926
GAC [48]	0.909	0.643	0.819	0.759	0.828	0.846	0.823	0.845
BAAFNet [84]	0.511	0.465	0.677	0.639	0.673	0.655	0.586	0.682
pointMLP [85]	0.975	0.865	0.959	0.950	0.969	0.959	0.945	0.953
PCT [86]	0.789	0.307	0.524	0.459	0.330	0.375	0.459	0.588
MBESegNet [50]	0.818	0.420	0.708	0.695	0.739	0.661	0.556	0.535
CurveNet [87]	0.964	0.783	0.923	0.917	0.939	0.922	0.918	0.939
Point-MAE [88]	0.971	0.802	0.956	0.924	0.949	0.943	0.942	0.948
Point-BERT [89]	0.976	0.835	0.962	0.939	0.952	0.951	0.951	0.957
ViT	0.985	0.885	0.971	0.966	0.959	0.969	0.959	0.968
ViT+MeshMAE	0.990	0.908	0.982	0.976	0.978	0.985	0.961	0.983
Ours	**0.995**	**0.921**	**0.989**	**0.988**	**0.986**	**0.992**	**0.974**	**0.990**

Table 3.1.
The tooth segmentation results from different methods in terms of the label-wise Dice Score.

task, we utilize two segmentation heads to provide a two-level feature aggregation. Specifically, we concatenate the output of the encoder with the feature embedding of each face to provide a fine-grained embedding. We set the batch size as 32 and employed an AdamW optimizer with an initial learning rate of 1e-4. The learning rate is decayed by a factor of 0.1 at 80 and 160 epochs.

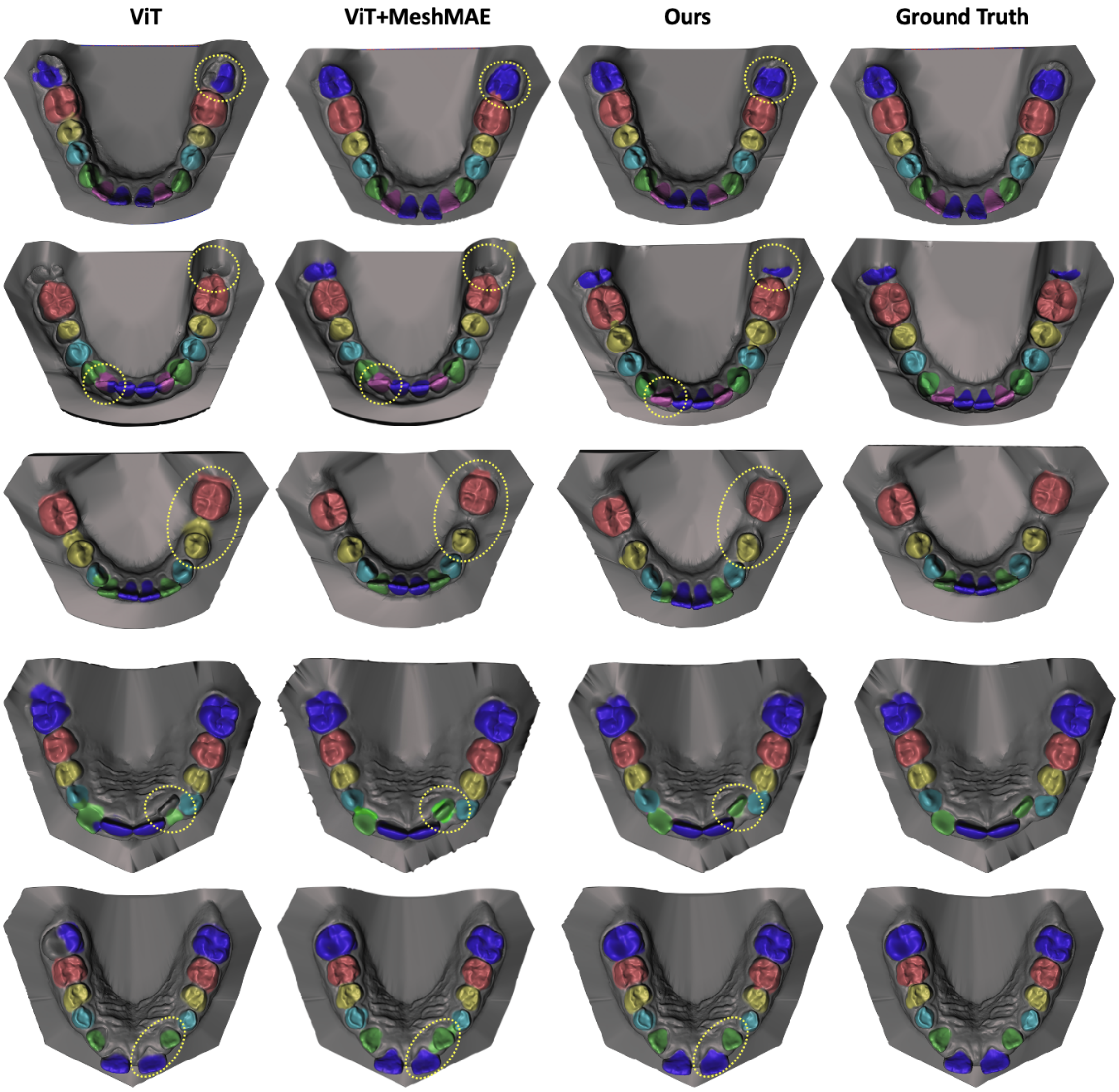


Figure 3.4. Comparison of teeth segmentation of DentalMAE and baselines. The first three rows show samples of the lower jaw, while the last two rows show the upper jaw.

3.5 Results and analysis

3.5.1 Quantitative results

Table 3.1 presents the quantitative results of tooth segmentation using various methods, and it clearly shows that DentalMAE outperforms other state-of-the-art methods.

Comparing the Dice Scores of ViT with the other methods, it is evident that ViT achieves higher scores on almost all tooth labels (T1-T7) and the background (BG). ViT achieves Dice Scores ranging from 0.885 to 0.985, indicating its effectiveness in accurately segmenting tooth structures. This demonstrates the capability of the Vision Transformer to capture relevant features and contextual information, leading to improved segmentation results.

The results of ViT+MeshMAE outperform the standard ViT, indicating further improvements. The combination of ViT and MeshMAE enhances the segmentation accuracy and ensures more precise delineation of tooth boundaries.

Our method, DentalMAE, surpasses not only the other methods but also the standalone ViT and its enhanced version MeshMAE. It is evident that our method consistently achieves the highest Dice Scores across all tooth labels (T1-T7) and the background (BG). The Dice Scores range from 0.921 to 0.995, highlighting the effectiveness of incorporating the loss on mask patches embedding for tooth structure reconstruction.

All ViT variants outperform traditional methods like PointNet [81], PointNet++ [82], DGCNN [83], and MeshSegNet [47], as well as advanced methods such as MeshSegNet+GCO [47], TSGCNet [49], GAC [48], BAAFNet [84], pointMLP [85], PCT [86], MBESegNet [50], and CurveNet [87]. It also performs better than state-of-the-art self-supervised learning methods, Point-MAE [88] and Point-BERT [89]. This indicates the superiority of our proposed methods in accurately segmenting tooth structures and surpassing the performance of existing state-of-the-art approaches.

Table 3.2 presents additional quantitative results for tooth segmentation, evaluating various methods based on Overall Accuracy (OA), Dice Score (DSC), Sensitivity

(SEN), and Positive Predictive Value (PPV). The results further confirm the superior performance of our proposed method, DentalMAE, compared to other state-of-the-art techniques.

Method	OA	DSC	SEN	PPV
PointNet [81]	0.926	0.903	0.913	0.912
PointNet++ [82]	0.892	0.853	0.864	0.865
DGCNN [83]	0.933	0.915	0.923	0.923
MeshSegNet [47]	0.901	0.873	0.888	0.879
MeshSegNet+GCO [47]	0.931	0.918	0.929	0.911
TSGCNet [49]	0.936	0.895	0.924	0.902
GAC [48]	0.855	0.809	0.818	0.844
BAAFNet [84]	0.601	0.611	0.755	0.594
pointMLP [85]	0.943	0.927	0.936	0.931
PCT [86]	0.629	0.479	0.509	0.586
MBESegNet [50]	0.716	0.642	0.710	0.644
CurveNet [87]	0.939	0.912	0.922	0.923
Point-MAE [88]	0.945	0.927	0.942	0.936
Point-BERT [89]	0.949	0.935	0.948	0.944
ViT	0.955	0.945	0.950	0.957
ViT+MeshMAE	0.971	0.954	0.966	0.983
Ours	**0.983**	**0.970**	**0.977**	**0.989**

Table 3.2.
The tooth segmentation results from different methods in terms of the Overall Accuracy, the Dice Score, the Sensitivity, and the Positive Predictive Value.

Our method, DentalMAE, achieves an OA value of 0.983. This score indicates the overall accuracy of the tooth segmentation results obtained by our method. It is evident that DentalMAE outperforms all other SOTA methods.

The Dice Score measures the similarity between the predicted and ground truth tooth segmentations. In terms of DSC, our method, DentalMAE, achieves a score of

0.970. These scores demonstrate the accuracy and overlap of the segmented tooth structures compared to the ground truth. Notably, our method consistently outperforms all other methods, including the top-performing MeshMAE method.

SEN and PPV evaluate the ability of the segmentation methods to correctly identify tooth structures (SEN) and the precision of the predicted tooth segmentations (PPV). Our method exhibits high SEN and PPV scores, with a SEN value of 0.977, and a PPV value of 0.989. These results indicate the robustness and accuracy of our method in identifying tooth structures while minimizing false positives and false negatives.

Parameter Setting and Masking Strategies. The experiments conducted in Table 3.3 explore the effects of different masking strategies and ratios on teeth segmentation. In contrast to the high mask ratios commonly used in 3D natural models [54], the segmentation task for teeth exhibits distinct preferences regarding the mask ratio. Notably, we consistently observe performance improvements as the mask ratio decreases from 50% to 20%. This finding suggests that reducing the mask ratio is beneficial for training the model, potentially because relevant features in 3D intra-oral models tend to be smaller in scale.

Mask ratio	strategy	OA	DSC
50%	random	0.947	0.936
50%	block	0.931	0.930
50%	grid	0.943	0.932
40%	random	0.955	0.939
30%	random	0.959	0.941
20%	random	**0.971**	**0.954**
10%	random	0.958	0.943

Table 3.3.
The influence of Mask Ratios/strategies on teeth segmentation of our DentalMAE.

Additionally, the random masking strategy outperforms the block and grid strategies, emphasizing its effectiveness in generating masks during the training process. These findings contribute to our understanding of optimal parameter settings for teeth segmentation and inform the development of more accurate and efficient segmentation models in this domain.

3.5.2 Qualitative results

Figure 5.5 presents qualitative examples that showcase the enhanced performance achieved through pre-training the ViT mesh transformer with DentalMAE for teeth segmentation. The observed improvements in segmentation align with the quantitative findings discussed in Section 5.5.2.

3.6 Conclusions

We have demonstrated that DentalMAE pre-training improves SOTA segmentation performance on 3D dental scan analysis. Importantly, DentalMAE self-pre-training outperforms existing methods on a small dataset, something that has not previously been explored. Our results also suggest that parameters, including mask ratio and strategy, should be tailored when applying masked autoencoders pre-training to the 3D dental scan domain. Together, these observations suggest that DentalMAE can further improve the already impressive performance of mesh ViTs in intra-oral scan analysis. In future work, we will test the efficacy of DentalMAE pretraining in prognosis and outcome prediction tasks.

CHAPTER 4
TEETH SEGMENTATION FROM CONE-BEAM CT IMAGES

4.1 Introduction

In the last decade, digital dentistry has rapidly evolved, emphasizing the acquisition and division of complete three-dimensional (3D) tooth models. These models are crucial for defining the intended arrangement and movements of individual teeth, particularly for orthodontic diagnosis and treatment planning. Obtaining these comprehensive 3D tooth models presents a challenge. Currently, two main technologies for acquiring these models are intraoral or desktop scanning and cone beam computed tomography (CBCT) [90]. Intraoral or desktop scanning is convenient for capturing the surface geometry of tooth crowns but lacks information about tooth roots, essential for precise diagnoses and treatments. Conversely, CBCT provides comprehensive 3D volumetric data for all oral tissues, including teeth, and due to its high spatial resolution, it is widely used in oral surgery and digital orthodontics. This paper focuses on 3D tooth segmentation and identification from CBCT images, which crucial for digital orthodontics applications.

Segmenting teeth from CBCT images presents significant challenges due to several reasons. Firstly, in natural occlusion conditions where upper and lower teeth touch, it is difficult to differentiate and separate lower teeth from the opposing upper teeth along their occlusal surface due to a lack of variations in gray values [91, 92]. Similarly, distinguishing teeth from their surrounding alveolar bone is challenging due to their similar densities. Additionally, adjacent teeth with similar appearances pose confusion in identifying different teeth. Consequently, relying solely on the intensity variation of CT images, as attempted in previous tooth segmentation methods, has proven insufficient.

Prior attempts to address these issues involved using either the level-set method [91–94] or template-based fitting methods [95] for tooth segmentation. The former methods necessitate a suitable initialization, often requiring laborious user annotations and yielding unsatisfactory results in natural occlusion conditions. The latter methods lack robustness when confronted with significant shape variations among different patients. While deep learning methods for medical image analysis [96–98] have shown promise in various tasks, their application to tooth segmentation has been limited.

Recent advancements in self-supervised learning have demonstrated the effectiveness of masked image modeling (MIM) [24, 25, 51] as a pre-training strategy for the Vision Transformer (ViT) [37] and the hierarchical Vision Transformer using shifted windows (Swin) [21,39,52]. MIM involves the masking and subsequent reconstruction of image patches, allowing the network to infer the masked regions by leveraging contextual information. We believe that the ability to aggregate contextual information is crucial in the context of CBCT image analysis. Among various MIM frameworks, the Masked Autoencoder (MAE) [25] stands out as a simple yet effective approach. MAE employs an encoder-decoder architecture, with a ViT encoder that receives only visible tokens and a lightweight decoder that reconstructs the masked patches using the encoder's patchwise output and trainable mask tokens.

We propose to use self pre-training since it is particularly advantageous in scenarios where acquiring suitable pre-training data is challenging. Additionally, self pre-training eliminates the domain discrepancy between the pre-training and fine-tuning stages by unifying the training data. Our experiments focus on teeth segmentation in 3D CT scans [99]. As our base model, we use UNEt TRansformer (UNTER) introduced in [100] for 3D CT scan analysis. Therefore, we call the proposed method UNETR+DEMAE. We apply UNETR+DEMAE pre-training on the same dataset that is used for the downstream task, i.e., to the training dataset.

Specifically, we propose to extend the self-supervised learning framework of the masked autoencoder (MAE) transformer [25]. While the MAE loss measures the quality of reconstructed masked patches, the loss of the proposed UNETR+DEMAE evaluates the predicted deep embeddings of masked patches. After pre-training, the decoder is discarded, and the encoder is applied to the downstream task, i.e., teeth segmentation. We compare three ViT Transformer initializations, including our proposed UNETR+DEMAE, MAE [25], and a transformer without any self-pre-training. The experimental results demonstrate that UNETR+DEMAE self-pre-training significantly enhances CBCT segmentation performance compared to the baselines. Our main contributions are threefold:

- We utilize self-supervised learning with masked autoencoders to alleviate the problem of small data for 3D CT scans.
- We replace the MAE reconstruction of masked patches with the reconstruction of patch embeddings. Hence, our loss is simply the L_2 distance between the predicted and computed embeddings over the masked patches.
- Our proposed method leads to a significant performance improvement. UNETR+DEMAE outperforms all state-of-the-art methods on the tooth segmentation task.

4.2 Methods

4.2.1 Vision Transformer

Our framework utilizes the Vision Transformer (ViT) as the foundational architecture for both pre-training and subsequent tasks. The ViT comprises a patch embedding layer, position embedding, and Transformer blocks.

Patch Embedding: The patch embedding layer within the ViT is responsible for transforming data into sequences. Initially, 3D volumes $\mathbf{x} \in \mathbb{R}^{H \times W \times D \times C}$ are reshaped into a sequence of flattened 3D patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^3 \cdot C)}$. The parameters (H, W, D) represent the image resolution, (P, P, P) denotes the patch resolution, C signifies the input channel, and $N = HWD/P^3$ stands for the number of patches or the sequence length fed into the Transformer. These patches are then mapped to patch embeddings via a trainable linear projection.

Position Embedding: To retain positional information, the patch embeddings are supplemented with position embeddings. While the standard ViT utilizes 1D learnable position embeddings, our experiments led us to employ sine-cosine [25, 27] position embeddings during the pre-training stage. Sine-cosine functions provide a fixed pattern that is not learned during training. This can be advantageous to the model to learn more generalizable features and avoid overfitting to the specifics of the training data, which is very scarce. Subsequently, for downstream tasks, we initialize the learnable position embeddings with the sine-cosine embedding values.

Transformer Block: The ViT architecture involves layers comprising multi-headed self-attention (MSA) [77] and MLP blocks.

4.2.2 Self-Supervised Pre-training with Masked Autoencoders

This section delineates the constituents of the Masked Autoencoder (MAE): the encoder, the decoder, and the associated loss function.

Encoder. As illustrated in Fig. 7.2(Left), the ViT encoder is responsible for reconstructing the complete input data from partially masked patches. The input undergoes partitioning into non-overlapping patches, which are then randomly divided into visible and masked groups. The MAE encoder operates solely on visible patches, incorporating position embeddings to retain positional information. The resulting representation serves the purpose of reconstructing the masked input, urging the encoder to derive a comprehensive representation from partial observations.

Decoder. The MAE decoder is fed with a complete set of tokens, encompassing patch-wise representations from the encoder, alongside learnable mask tokens placed in the positions of masked patches. By integrating positional embeddings with all input tokens, the decoder aims to restore each specific patch within its masked posi-

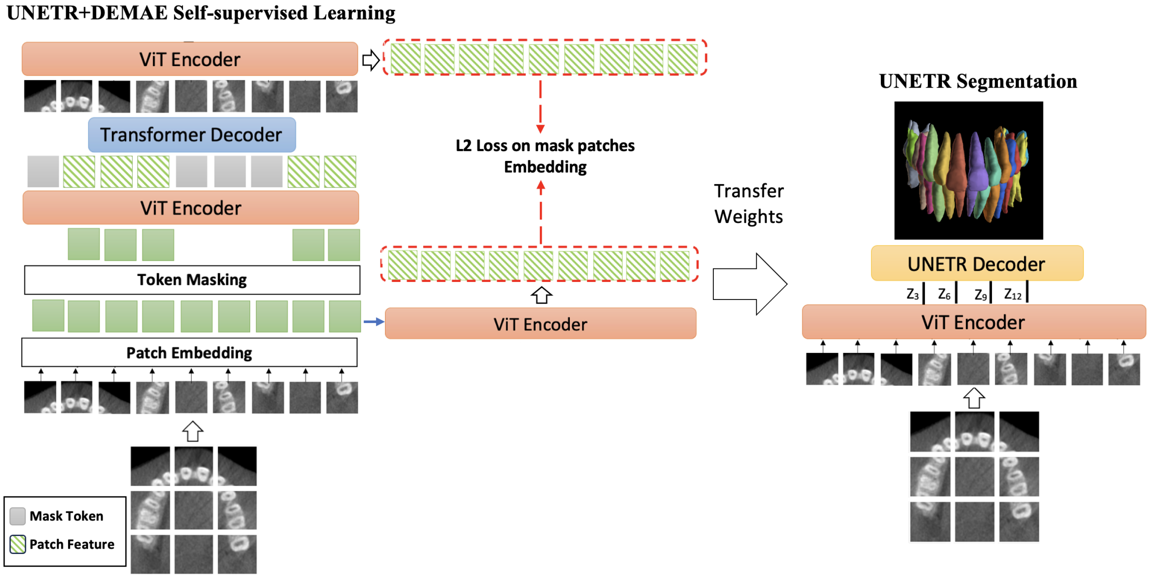


Figure 4.1. Segmentation Pipeline with MAE Self Pre-training. Left: A ViT encoder is first pre-trained with MAE. A random subset of patches is input to the encoder and a transformer decoder reconstructs the full image. Right: The pre-trained ViT weights are transferred to initialize the segmentation encoder. Then the whole segmentation network, e.g., UNETR [100], is finetuned for segmentation.

tion. It's noteworthy that the decoder serves as an auxiliary module exclusively for pre-training and is not utilized in downstream tasks.

Masked Sequence Generation. Patch embeddings are represented by a set E. Following the MAE approach, we randomly mask a subset of patches, represented as E_m, while unmasked embeddings are denoted as E_{um}. We replace the masked embeddings E_m with a shared learnable mask embedding E_{mask} without altering their positional embeddings. Finally, the corrupted embeddings E_c are formed by combining E_{um} with the sum of E_{mask} and a set of positional embeddings p. These corrupted embeddings are then inputted into the encoder for further processing.

Prediction. MAE [25] reconstructs the input by predicting the pixel values for each masked patch. Its loss function computes the mean squared error (MSE) between the reconstructed and original images in the pixel space. In contrast, We propose to compute the L_2 loss between the original and predicted embeddings of the mask patches. As our experimental results in Sec. 4.3.3 demonstrate, this leads to a performance increase.

4.2.3 Architectures for Downstream Tasks

Following MAE self-pre-training, we append task-specific head for the downstream task, i.e., tooth segmentation.

We employ the UNETR [100] built upon the pre-trained ViT encoder via MAE, in conjunction with a convolutional decoder initialized randomly. UNETR, designed for 3D image segmentation tasks, mirrors the concept of U-Net [17]. It involves skip connections between features from various encoder resolutions and the decoder. The input to the UNETR decoder constitutes a sequence of representations from the encoder. Each representation is reshaped to restore spatial dimensions, followed by iterative upsampling and concatenation with shallower features to enhance segmentation resolution.

4.3 Experiments and Results

4.3.1 Datasets and Implementation Details

Tooth segmentation on CBCT images. We use the public dataset 3D CT scans [99]. There are a total of 150 CBCT images with a resolution varied from 0.25 mm to 0.35 mm. We randomly split the dataset into 80% for training and 20% for validation. The task is tooth segmentation from the 3D CT scans. Next, we normalize the intensity of the CBCT image to fit within the range of [0, 1]. For the creation of training data, we randomly extract 150 sub-volumes measuring $128 \times 128 \times 128$ around the alveolar bone ridge in the CT scan, resulting in approximately 18,000 sub-volumes for training. The dataset's ground truth includes annotations with tooth-level bounding boxes, masks, and labels.

During the testing phase, we employ the overlapped sliding window method to crop sub-volumes of size $128 \times 128 \times 128$ with a stride of $32 \times 32 \times 32$. Subsequently, in the scenario where two teeth segments overlap, we select the one with the highest value of $P_{cls} \times P_{id}$ as the final tooth prediction if the Intersection over Union (IoU) of their teeth segmentation results is greater than 0.2. Here, P_{cls} and P_{id} represent the probabilities for tooth classification and identification, respectively.

We conduct our experiments using PyTorch [35] and MONAI [101]. ViT-B/16 serves as the backbone, and we utilize AdamW as the optimizer across all experiments. The patch size for 3D volumes is set at $16 \times 16 \times 16$.

4.3.2 Evaluation metric

We use Dice similarity coefficient (DSC) to evaluate the performance of our model as follows:

$$DSC = \frac{2 \times |Y \cap Z|}{|Y| + |Z|}, \tag{4.1}$$

where Y and Z represent the voxelized predicted outcomes and the ground truth masks, respectively.

Additionally, we establish the accuracy of detection and identification as follows: assuming G represents the entirety of teeth within the ground truth data, and D indicates the set of teeth detected by our network, where within D there are L correctly labeled teeth. The detection accuracy (DA) and identification accuracy (FA) are determined through the following calculations:

$$DA = \frac{|D|}{|D \cup G|} \text{ and } FA = \frac{|L|}{|D \cup G|} \tag{4.2}$$

UNETR+DEMAE Self Pre-training. The starting learning rate (lr) remains at 1.5e-3, and the weight decay is set at 0.05. The learning rate decays to zero using a cosine schedule that includes warm-up periods. The pre-training for UNETR+DEMAE lasts for 100 epochs, utilizing training batch sizes of 256.

Finetuning for Teeth Segmentation. We apply a layer-wise learning rate decay (with a layer decay ratio of 0.75) to ensure the stability of UNETR training, along with implementing random DropPath with a 10% probability. The learning rate is set at 8e-3, and the training batch size is maintained at 256. Additionally, the learning rate during the fine-tuning phase also follows a cosine decay schedule.

4.3.3 Results

Teeth segmentation on CBCT images. Table 4.1 presents the quantitative results of tooth segmentation using various methods, and it clearly shows that UNETR+DEMAE outperforms other state-of-the-art methods.

Comparing the scores of DSC, PA and FA of UNETR with the other methods, it is evident that it achieves the highest performance, indicating its effectiveness in accurately segmenting tooth structures. This demonstrates the capability of the Transformers to capture relevant features and contextual information, leading to improved segmentation results.

The result of UNETR+MAE is superior to the standard UNETR, indicating further improvements. It also outperforms the ImageNet pre-training paradigm (UNETR+ImageNet). The combination of UNETR and MAE enhances the segmentation accuracy and ensures more precise delineation of tooth boundaries.

Our method, UNETR+DEMAE, surpasses the other methods and the standalone UNETR and its enhanced version MAE. Our method consistently achieves the highest results, highlighting the effectiveness of incorporating the loss on mask patch embeddings for tooth structure reconstruction.

Parameter Setting. We perform experiments involving various UNETR+DEMAE pre-training epochs and mask ratios, as detailed in Table 7.4. Firstly, we note that the performance of UNETR+DEMAE does not improve with longer training periods. Secondly, unlike the high mask ratio commonly used in natural images [25], the segmentation task demonstrates varied preferences for different mask ratios. The most optimal segmentation outcomes are attained with a mask ratio of 25%.

Qualitative results. Figure 5.5 presents qualitative examples that showcase the enhanced performance achieved on teeth segmentation through our UNETR+DEMAE

Table 4.1.
Tooth Segmentation on CBCT scans. UNETR+DEMAE self pre-training improves upon the UNETR baseline, ImageNet supervised pre-training, and MAE self-supervised learning.

Framework	DSC	DA	FA
U-Net(R50) [17]	84.18	82.84	79.19
AttnUNet(R50) [102]	85.92	63.91	79.20
TransUNet [103]	87.23	83.13	81.87
DSTUNet [104]	88.16	87.40	87.46
nnFormer [105]	86.07	80.17	86.57
nnUNet [106]	88.92	81.77	85.57
UNETR	89.46	90.88	88.03
UNETR+ImageNet	92.04	94.29	93.44
UNETR+MAE	93.01	95.25	94.37
UNETR+DEMAE	**94.20**	**99.65**	**97.57**

Table 4.2.
The influence of Mask Ratios and Pre-training Epochs on teeth segmentation of our UNETR+DEMAE.

Mask ratio	Pre-training Epochs	DSC	DA	FA
85%	100	91.32	96.43	95.59
75%	100	91.73	97.85	95.74
75%	800	90.14	97.32	94.89
50%	100	93.56	98.93	95.98
25%	100	**94.20**	**99.65**	**97.57**
10%	100	93.10	97.82	97.09

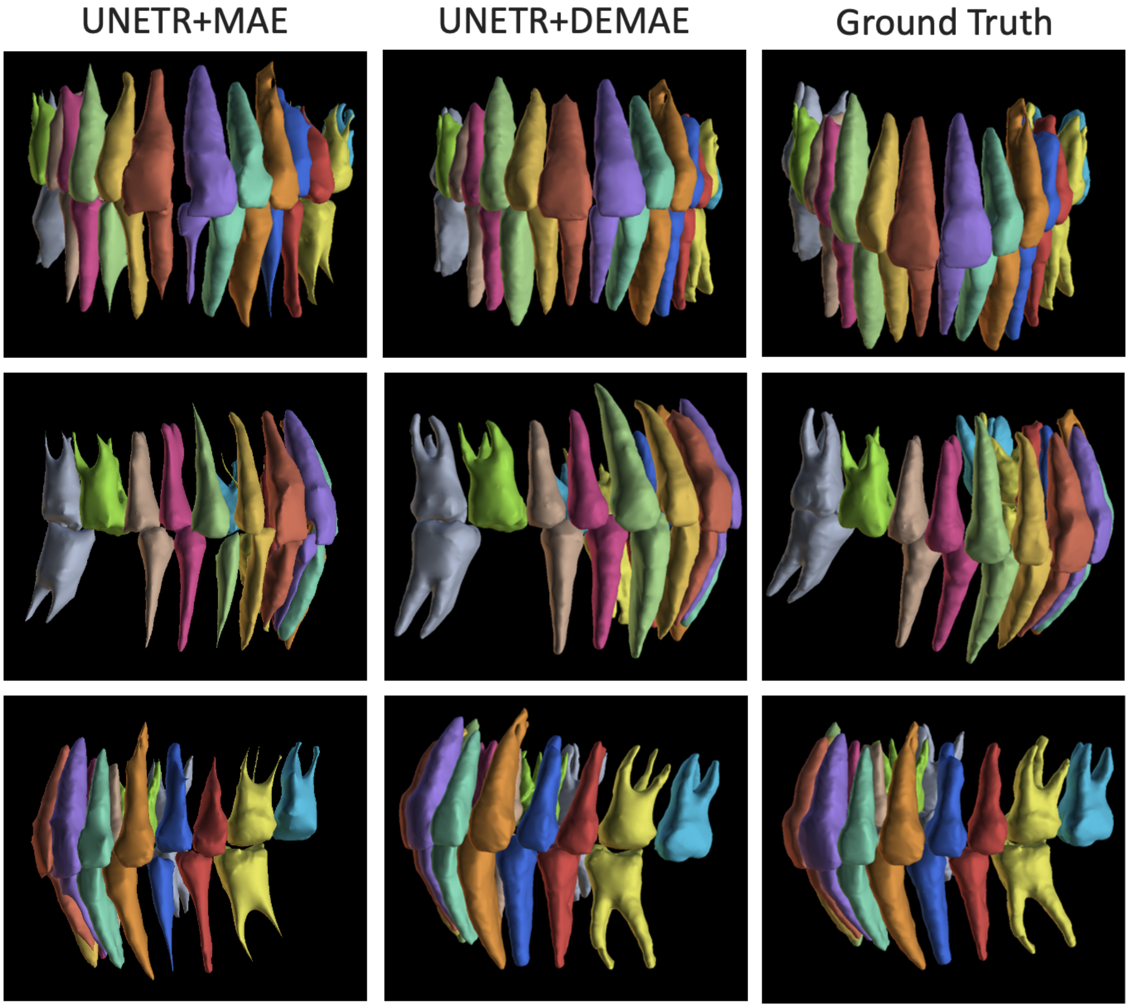


Figure 4.2. Comparison of teeth segmentation of UNETR+DEMAE and baseline.

pre-training in comparison to UNETR+MAE. The observed improvements in segmentation align with the quantitative findings shown in Table 4.1.

4.4 Conclusions

We have demonstrated that UNETR+DEMAE pre-training improves SOTA segmentation performance on 3D dental CT scan analysis. Importantly, UNETR+DEMAE self-pre-training outperforms existing methods on a small dataset, something that has not previously been explored. Our results also suggest that parameters, including mask ratio and strategy, should be tailored when applying masked autoencoders pre-training to the 3D dental scan domain. Together, these observations suggest that UNETR+DEMAE can further improve the already impressive performance of ViTs in CBCT scan analysis. In future work, we will test the efficacy of UNETR+DEMAE pretraining in prognosis and outcome prediction tasks.

CHAPTER 5
DENTAL IMPLANT IDENTIFICATION

5.1 Introduction

In the 1980s, dental implants were introduced and have since become a global solution for patients with missing teeth [107]. Their impact on dental care has been significant, contributing to improved quality of life [108, 109]. While implant treatments are now common, over decades of clinical use have brought forth challenges, including complications in superstructures or implants and peri-implantitis [110, 111]. Addressing these issues often requires additional prosthodontic, periodontic, or surgical interventions, necessitating detailed information about the intra-oral implant.

Accessing such information is straightforward when patients were previously treated at the same clinic, but complications arise when patients seek care elsewhere due to relocation or clinic closures. Dentists faced with limited data, such as oral photographs and radiographs, must identify crucial implant details, particularly the implant system, to proceed with treatments. While experienced dentists can navigate this process, those lacking sufficient knowledge face difficulties. Consequently, there is a demand for a system that can identify implant systems from limited data, irrespective of a dentist's expertise.

Artificial intelligence (AI) technology, widely utilized in various fields, offers promising solutions. In medicine, AI has already proven valuable in robotics, medical diagnosis, statistics, and human biology [112]. Deep learning, a subset of AI, excels in tasks like prediction, object detection, and classification. Dentistry has seen the application of deep learning in diagnosing dental diseases from images, predicting treatments, classification, and statistical analysis [113–120]. Notably, deep learning-based object detection algorithms have enhanced diagnostic systems [121], often matching or surpassing human capabilities.

This study aims to develop an automated system using a deep learning-based object detection method to identify implant systems. The hypothesis is that this system can effectively detect and identify implants, offering a valuable tool for dentists and patients grappling with implant-related issues.

Recent progress in self-supervised learning has highlighted the efficacy of masked image modeling (MIM) [22,24,25,122,123] as a pre-training strategy for Vision Transformer (ViT) [37,124] and the hierarchical Vision Transformer using shifted windows (Swin) [21, 39, 52]. MIM involves masking image patches and reconstructing them, allowing the network to deduce masked regions by utilizing contextual information. The capacity to aggregate contextual information is deemed crucial in the context of dental radiograph analysis. Among various MIM frameworks, the Masked Autoencoder (MAE) [25] stands out as a straightforward yet effective approach. MAE utilizes an encoder-decoder architecture, incorporating a ViT encoder that receives visible tokens and a lightweight decoder reconstructing masked patches using the encoder's patchwise output and trainable mask tokens.

We advocate for self-pre-training, particularly advantageous when obtaining suitable pre-training data is challenging. Self-pre-training also eradicates domain discrepancies between pre-training and fine-tuning by unifying the training data [123]. Our experiments center on dental implant identification and classification in panoramic and periapical radiographs [125]. We apply our proposed method, Masked Deep Embedding (MDE) pre-training, on the same dataset used for the downstream task, i.e., the training dataset.

Specifically, we extend the self-supervised learning framework of the masked autoencoder (MAE) transformer [25]. While the MAE loss gauges the reconstructed masked patches' quality, the MDE loss assesses predicted deep embeddings of masked patches. Post pre-training, the decoder is discarded, and the encoder is applied to the downstream task, i.e., dental implant detection. We compare three ViT Transformer initializations, including our proposed MDE, MAE [25], and a transformer without any self-pre-training. Experimental results demonstrate that MDE self-pre-training significantly enhances dental implant detection performance compared to the baselines.

Moreover, a comprehensive dataset addressing the simultaneous examination of implant design is currently unavailable. We contend that incorporating implant design amplifies the complexity of the computer vision problem, given the increased number of classes and potential class imbalances. Consequently, we enhance the existing dataset introduced in [125] under the supervision of a dental expert. We further enrich the dataset by creating annotations specifically for implant design, encompassing the classification of coronal, middle, and apical parts. The meticulous labeling process has resulted in a distinctive, high-quality augmented dataset. Interested parties can access our data, named Implant Design Dataset (IDD), upon request.

Our contributions are threefold:

- We replace MAE's reconstruction of masked patches with the reconstruction of patch embeddings. Consequently, our loss is the simple L_1 distance between predicted and computed embeddings over masked patches.
- Our proposed method yields substantial performance improvement, surpassing all state-of-the-art methods in the dental implant detection task.
- The labeling of implant design extends the horizon of possible dental applications.

5.2 Dental Implant Design

The categorization of implant design in the images was carried out, as detailed in Table 5.1. The coronal one-third of the implant underwent classification based on bone level, tissue level, microthread, and thread design (see Fig.5.1(a)). The middle one-third was categorized concerning body shape (straight or tapered) and thread design (see Fig.5.1(b)). The apical part was classified based on criteria such as the presence of a groove in the apical part, the shape of the apical hole, the shape of the apical body, and the apex shape (see Fig.5.1(c)). An experienced prosthodontist

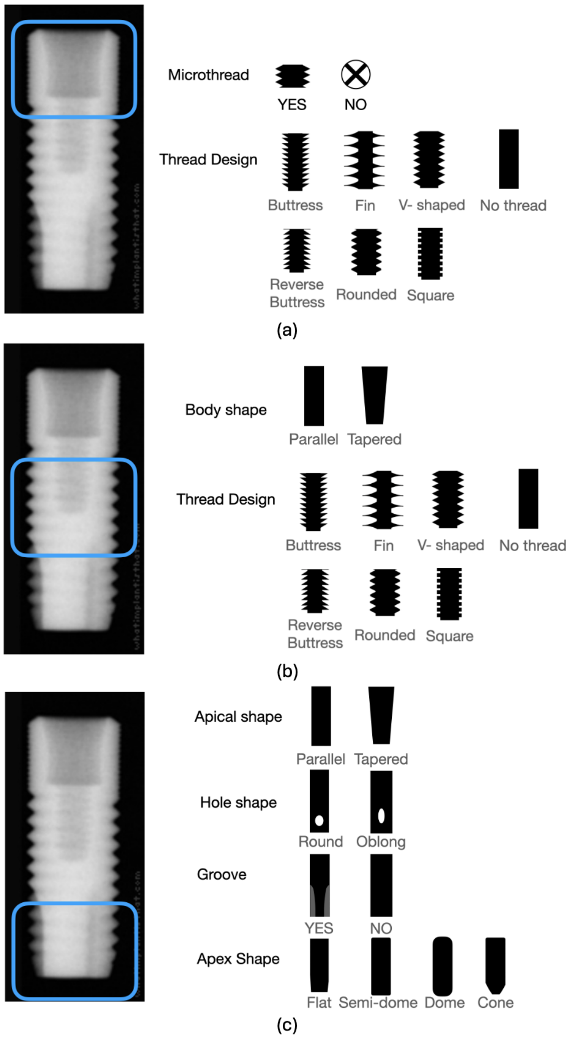


Figure 5.1. Categorization of (a) the coronal portion's design based on bone or tissue level, the existence of microthreads, and the specific thread design, (b) the middle section's design based on the shape of the body and the specific thread design, and (c) the design for the apical portion based on the shape of the apical hole, the configuration of the apical body, the presence of a groove, and the shape of the apex.

classified each group by referencing the manufacturer's catalog and radiographs using the COCO-Annotator tool [34]. Subsequently, implant images were labeled according to the design classifications (see Fig. 5.2).

Table 5.1.
Details of classes based on the classification of implant design.

Coronal	**Middle**	**Apical**
Bone level	Parallel fin	Hole round
Tissue level	Tapered fin	Hole oblong
Microthread	Parallel square	Parallel groove
Fin	Tapered square	Tapered groove
Square	Parallel no threads	Parallel no groove
No threads	Tapered no threads	Tapered no groove
V-shaped	Parallel V-shaped	Apex shape flat
Rounded	Tapered V-shaped	Apex shape cone
Buttress	Parallel rounded	Apex shape dome
Reverse buttress	Tapered rounded	Apex shape semi-dome
	Parallel buttress	
	Tapered buttress	
	Parallel reverse buttress	
	Tapered reverse buttress	

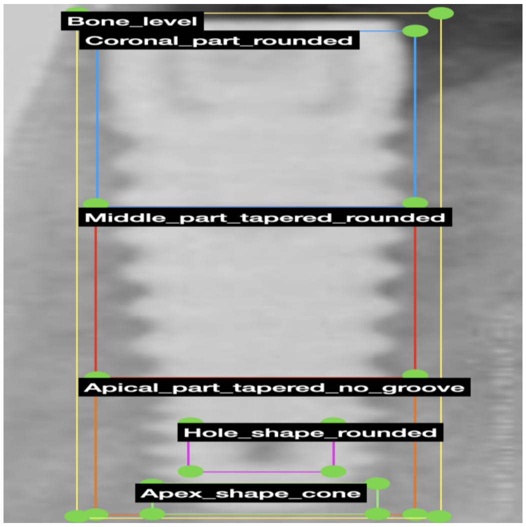


Figure 5.2. Image of a sample for categorizing and labeling implant design.

5.3 Methods

5.3.1 Two-Stage Implant Detection Methodology

To address the task of dental implant detection, we propose a two-stage detection approach comprising the identification of individual implant design parts in the first stage and the subsequent inference of implant bounding boxes in the second stage. This method aims to enhance detection accuracy by breaking down the complex implant structure into distinct components before consolidating them into a comprehensive bounding box representation.

Implant Design Parts Detection (First Stage). In the initial stage of our methodology, we employ a dedicated object detection algorithm trained to recognize specific implant design parts. The chosen algorithm, in this case, is Mask R-CNN, which has been trained on an annotated dataset containing diverse dental implant images. The annotations include bounding box coordinates for each implant part such as the body, threads, and head.

During the inference process, the trained model scans input images, identifying the presence and localization of individual implant design parts. The output consists of bounding boxes for each detected component, providing a detailed spatial representation of the identified implant parts.

Implant Bounding Box Inference (Second Stage). Building upon the results of the first stage, the subsequent step involves inferring bounding boxes that encapsulate the entire dental implant structure. Post-processing techniques are applied to consolidate the detected implant design parts into a cohesive representation of the complete implant.

This involves:

- Handling Missing Implant Parts by developing post-processing strategies to infer or estimate missing parts based on the detected components. We implement techniques such as predictive models [126], spline interpolation [127], and adaptive thresholds [128] to enhance robustness in the presence of incomplete information.
- Analyzing spatial relationships between detected parts to refine the assembly process and improve the accuracy of the final representation.
- Employing clustering algorithms, such as K-Means Clustering [129], to group related implant design parts, adapting to variations in implant geometry, and aiding in the identification of missing components.
- Implementing heuristics based on known implant geometries to guide the assembly process, especially when dealing with missing parts.

Upon successful grouping of individual implant parts, a bounding box is inferred to encapsulate the entire implant structure. This final bounding box serves as a holistic representation of the detected dental implant in the input image.

By dividing the detection process into these two stages and incorporating strategies to handle missing implant parts, our methodology aims to enhance the accuracy and robustness of dental implant detection, particularly in scenarios involving complex implant geometries and variations in image quality. The proposed approach

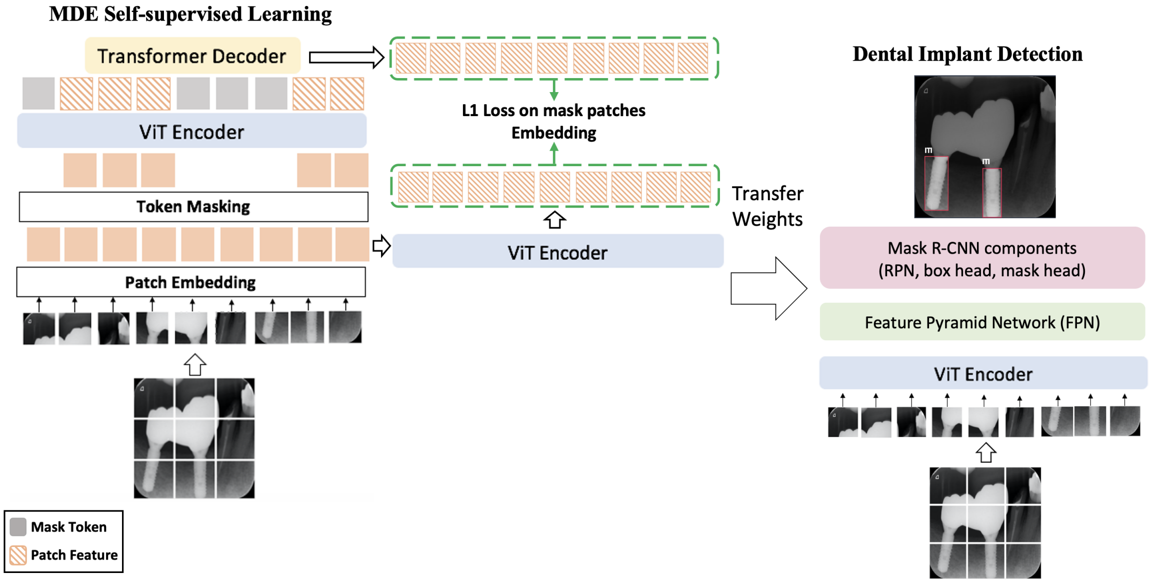


Figure 5.3. Dental Implant Detection Pipeline with MDE Self Pre-training. The initial step in the dental implant detection pipeline for MDE self-pre-training involves dividing the input into non-overlapping patches. These patches undergo embedding using an MLP. Throughout the pre-training phase, the patch embeddings undergo random masking, and only the visible embeddings are employed by the transformer. Subsequently, the masked embeddings are merged with the encoded embeddings and directed to the decoder. The decoder's role is to reconstruct the masked patches, followed by predicting the patch embeddings of these masked patches. The L_1 loss is employed to assess the similarity between the masked patch embeddings. Once pre-training is complete, the decoder is omitted, and the encoder is utilized as the backbone in Mask R-CNN with FPN for the detection.

provides a structured and systematic means of addressing the challenges associated with implant detection in diverse clinical contexts.

5.3.2 Self-Supervised Pre-training with Masked Autoencoders

This section details the constituents of the Masked Autoencoder (MAE): the encoder, the decoder, and the associated loss function.

Encoder. As illustrated in Fig. 7.2(Left), the input undergoes partitioning into non-overlapping patches, randomly divided into visible and masked groups. The MAE encoder operates solely on visible patches, incorporating position embeddings to retain positional information. The resulting representation serves to reconstruct the masked input.

Masked Sequence Generation. Patch embeddings E are represented by a set. Following the MAE approach, a subset of patches is randomly masked, represented as E_m, and unmasked embeddings as E_{um}. Masked embeddings E_m are replaced with a shared learnable mask embedding E_{mask}. Corrupted embeddings E_c are formed by combining E_{um} with the sum of E_{mask} and positional embeddings p, inputted into the encoder.

Decoder. The MAE decoder is fed with a complete set of tokens, including patch-wise representations from the encoder and learnable mask tokens. Integrating positional embeddings with input tokens, the decoder aims to restore each patch embedding within its masked position, serving as an auxiliary module exclusively for pre-training.

Loss computation. We propose computing the L_1 loss between original and predicted embeddings of masked patches, deviating from MAE's mean squared error (MSE) in pixel space. As our experimental results demonstrate, this change leads to performance improvement. This is in accord with observations in [130] that predict deep embedding of patches instead of pixel values yields better generalization and performance improvements.

5.3.3 Architectures for Downstream Tasks

After completing self-pre-training with MAE, we attach a task-specific head for the subsequent task, namely, the detection of dental implants.

The pre-trained ViT weights are utilized to initialize the encoder for detection. The features from the ViT backbone are conveyed to both the neck (FPN [31]) and the detection head (Mask R-CNN) to facilitate bounding box regression and classification.

We opt for the Mask R-CNN [131] framework, given its widespread use in object detection research. Subsequently, the entire network undergoes fine-tuning to execute the detection task.

5.4 Experiments

5.4.1 Dataset

Implants Image Dataset [125] is a dental panoramic and periapical X-rays dataset consisting of 5572 annotated images with ground truth detection labels of dental implants. Each image size is 416x416 pixels.

We contribute to further expanding the dataset by developing bounding boxes for dental implant design parts, including the thread design, body shape, apical shape, hole shape, and apex shape. This process was done by a prosthodontist using the COCO-Annotator tool [34].

We believe this is the most inclusive dataset for dental implant identification and classification in dental radiographs. We are providing our data, upon request, under the name Implant Design Dataset (IDD).

5.4.2 Evaluation metric

In all our experiments, we divided the data into five sets, each comprising around 20% of the images. Among these, one set remains constant as the test dataset, containing 1116 images, while the remaining four sets, each with 1114 images, form the training and validation datasets using cross-validation. This procedure is iterated five times. We use the Average Precision metric to evaluate object detection models.

5.4.3 Implementation details

We conducted our experiments using the PyTorch framework [35] and trained them on Nvidia Tesla V100 GPUs. Throughout all experiments, the batch size remains consistent at 4456, which corresponds to the total number of training samples. The AdamW optimizer [36] is employed in all instances.

Data augmentation. We apply noise addition up to 6% of pixels, horizontal and vertical flipping, and 90° rotation of clockwise, counter-clockwise.

MDE pre-training. The base learning rate is established at 1.5e-4, weight decay is set to 0.05, $\beta 1$ is 0.9, and $\beta 2$ is 0.95. A cosine decay learning rate scheduler with a warm-up period of 10 epochs is applied. We utilize a random Masked Image Modeling approach with a patch size of 16×16 and a mask ratio of 25%. Additionally, we employ a linear prediction head, targeting an image size of 416x416.

Task fine-tuning. For downstream tasks, we employ single-scale training. The starting learning rate is 0.0001, and the weight decay is set at 0.05.

5.5 Results and analysis

5.5.1 MDE reconstruction

The reconstruction outcomes of MDE are depicted in Fig. 6.3. The figure comprises four columns illustrating the original images, masked images, images reconstructed using MAE, and images reconstructed using MDE. The results indicate that our approach excels in recovering missing information from the random context. It is important to emphasize that the primary objective of MIM is to enhance downstream tasks rather than produce reconstructions of the highest quality. It is worth noting that the process involves a reconstruction step. Instead of directly restoring pixel values from patches, the deep embeddings guide a generative model or decoding process to produce image samples. This generative model uses the high-level information encoded in the embeddings to create new pixel values, providing a reconstructed representation of the original images.

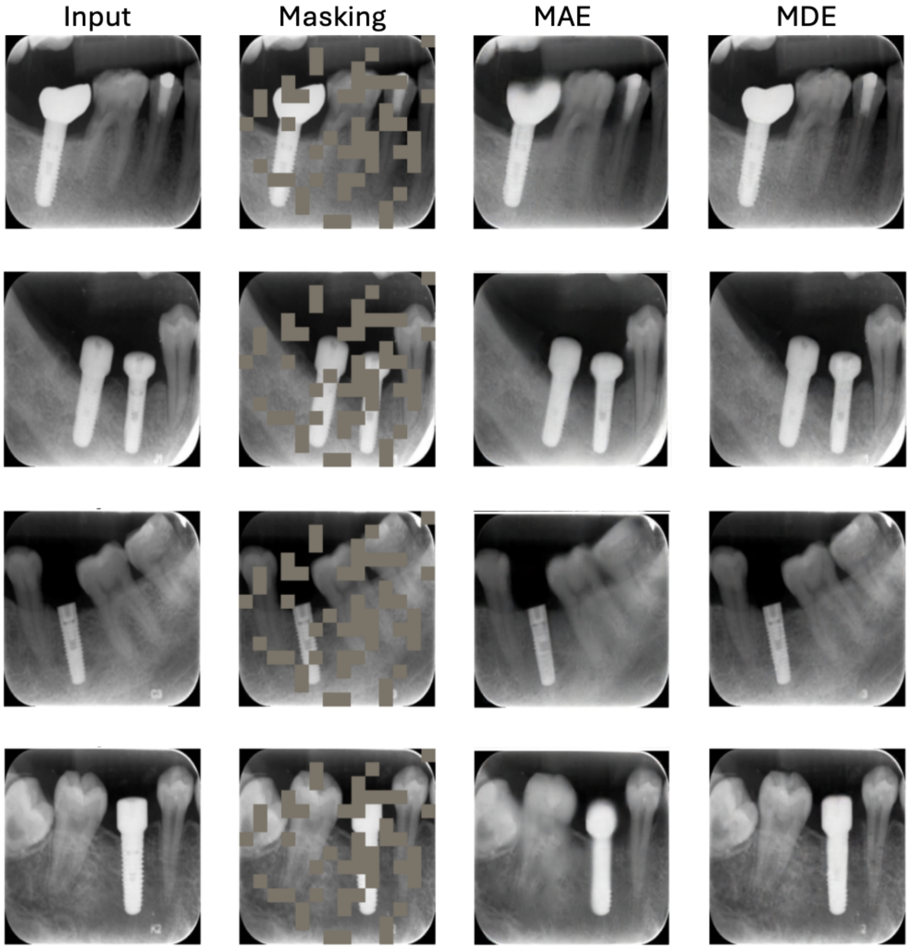


Figure 5.4. Results of MDE reconstruction. The first column displays the original images, while the second column shows the masked images, with gray patches indicating the masked regions. The third and fourth columns exhibit the reconstructions achieved through MAE and our MDE, respectively, from the unmasked patches.

5.5.2 Dental implant classification and identification.

The presented tables provide a comprehensive overview of the results obtained in dental implant classification and identification tasks before and after the implementation of dental implant design labeling. The evaluation metric utilized is the AP^{box} (average precision for bounding box detection), and various initialization strategies and backbones are compared.

In Table 7.1 (before the application of dental implant design labeling), the YOLOv5 model with CSPDarknet53 backbone achieved an AP^{box} of 91.5, serving as a baseline for comparison. Notably, the ViT-B model with Random initialization, despite having no pre-training data, demonstrated a competitive performance with an AP^{box} of 91.9, showcasing the ViT's ability to learn meaningful features even in the absence of specific pre-training. The Supervised ViT-B model, pre-trained on ImageNet-1K with labels, improved the performance further to 92.6. The introduction of novel pre-training approaches, MAE and our MDE, both based on ViT-B and pre-trained on ImageNet-1K without labels, yielded impressive results of 93.2 and 94.9, respectively, with our MDE standing out as the most effective approach.

Moving to Table 7.2, where dental implant design labeling is employed, we observe consistent improvements across all models. The Random ViT-B model achieved an AP^{box} of 92.4, showcasing the impact of incorporating dental implant design information. The Supervised ViT-B model experienced an increase to 93.2, emphasizing the value of labeled implant design data for pre-training. The MAE and our MDE models, both pre-trained on ImageNet-1K without labels, demonstrated substantial improvements, reaching AP^{box} values of 94.0 and an impressive 96.1, respectively, with our MDE once again outperforming the other methods.

The results suggest that dental implant design labeling significantly enhances the performance of dental implant classification and identification models, regardless of the initialization strategy. Furthermore, the effectiveness of our MDE as a pre-training criterion is reinforced, indicating its robustness and suitability for the specific task at hand. These findings have important implications for the field of medical image analysis, underscoring the importance of domain-specific pre-training and the potential benefits of incorporating design information for more accurate and reliable dental implant detection. The combination of advanced pre-training techniques and domain-specific data augmentation can contribute to further advancements in the development of robust and precise models for dental implant recognition.

Table 5.2.
Results of dental implant classification and identification before employing dental implant design labeling.

Initialization	**Backbone**	**Pre-training Data**	$\boldsymbol{AP^{box}}$
YOLOv5 [132]	CSPDarknet53	IN-1K w/ Labels	91.5
Random	ViT-B	None	91.9
Supervised	ViT-B	IN-1K w/ Labels	92.6
MAE	ViT-B	IN-1K	93.2
MDE (ours)	ViT-B	IN-1K	**94.9**

Table 5.3.
Results of dental implant classification and identification after employing dental implant design labeling.

Initialization	**Backbone**	**Pre-training Data**	$\boldsymbol{AP^{box}}$
Random	ViT-B	None	92.4
Supervised	ViT-B	IN-1K w/ Labels	93.2
MAE	ViT-B	IN-1K	94.0
MDE (ours)	ViT-B	IN-1K	**96.1**

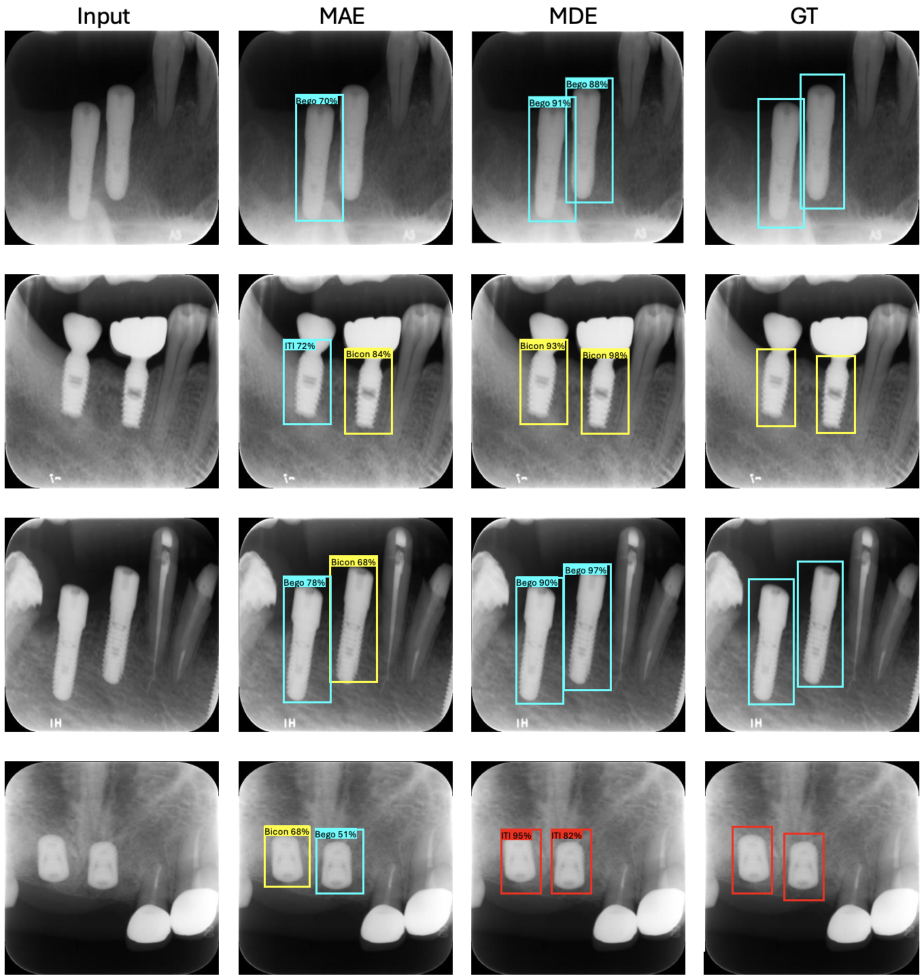


Figure 5.5. Qualitative results of dental implant detection and identification. The ViT pre-trained with the MAE approach exhibits missing or incorrect detections, whereas the ViT pre-trained with the MDE approach demonstrates accurate detection. Blue indicates the Bego dental implant system, yellow indicates the Bicon dental implant system, and red indicates the ITI dental implant system.

5.5.3 Qualitative Results

Figure 5.5 showcases qualitative samples highlighting the improved performance of dental implant detection and identification when using ViT pre-trained with MDE. The visual improvements align with the quantitative findings discussed in Section 5.5.2.

Table 5.4.
Impact of Mask Ratios on dental implant detection.

Mask Ratio	**Pre-training Epochs**	AP^{box}
65%	100	92.5
55%	100	93.2
55%	800	91.6
45%	100	94.0
35%	100	94.4
25%	100	**94.9**
15%	100	94.3

5.5.4 Parameter setting

In Table 7.4, we conduct experiments focusing on dental implant detection with varying pre-training epochs and mask ratios for our MDE method. Firstly, we observe that extending the training duration does not lead to improved performance for MDE. Secondly, in contrast to the high mask ratio used in natural images [25], we find distinct preferences for mask ratios in downstream tasks related to dental implant detection. Notably, both tasks consistently exhibit enhancements as the mask ratio decreases from 65% to 25%. This improvement may be attributed to the fact that relevant features on dental X-rays tend to be smaller in size.

5.6 Conclusions

We have illustrated that the proposed MDE pre-training enhances state-of-the-art detection performance in the analysis of dental X-rays. Notably, MDE self-pre-training surpasses the performance of existing methods, particularly on a limited dataset, an aspect not previously explored. Our findings also indicate that parameters, such as mask ratio and pre-training epochs, should be customized when applying masked autoencoders pre-training to the domain of dental radiographs. These insights suggest that MDE has the potential to further enhance the already remarkable performance of ViTs in the analysis of dental X-rays. In our future work, we aim to assess the effectiveness of MDE pretraining in tasks related to prognosis and outcome prediction. We have also demonstrated that two-stage object detection with the first stage focused on domain-specific object parts like implant design parts can enhance object detection results.

CHAPTER 6
ANALYZING DENTAL PANORAMIC RADIOGRAPHS FOR ASSESSMENT OF PERI-IMPLANT BONE LOSS

6.1 Introduction

Since the advent of dental implants in the 1980s, they have significantly revolutionized the landscape of restorative dentistry, offering a durable solution for missing teeth. This breakthrough has not only improved the aesthetic outcomes for patients but has also had a profound impact on their overall quality of life by restoring functionality and comfort. Dental implants have been lauded for their ability to provide a foundation for replacement teeth that look, feel, and function like natural teeth [133, 134]. Despite the widespread success and adoption of dental implants, their longevity and stability are critically dependent on the maintenance of peri-implant bone health. The advent of peri-implant diseases, especially peri-implantitis, highlights the urgent need for accurate and early detection of marginal bone loss to prevent implant failure and ensure the sustainability of the treatment [135].

The diagnostic challenge of accurately assessing bone levels around dental implants is compounded when detailed patient histories or specific implant details are lacking. This scenario is common when patients transfer their care due to relocation or when the original dental practice is no longer accessible. The nuanced interpretation of radiographic evidence of bone loss around implants requires a high degree of expertise and experience, underscoring the necessity for an automated system that can consistently and reliably assess peri-implant bone health [136].

Artificial Intelligence (AI), and specifically the domain of deep learning within AI, has emerged as a beacon of innovation across various medical fields, including dentistry. Deep learning's ability to decipher and learn from complex patterns in voluminous datasets has been a game-changer, offering new horizons in diagnostic precision and patient care [137]. In dental radiography, the application of deep learning models holds the promise of substantially enhancing the diagnostic process for detecting marginal bone levels around dental implants, a task that remains challenging even for seasoned practitioners [138].

In comparison to the realm of medical image classification, there has been a notable dearth of research investigating the detection of specific lesions in radiographs [139, 140] or magnetic resonance imaging (MRI) [141] through semantic segmentation. While certain studies have endeavored to segment lesions using techniques like U-Net [17], the exploration of instance segmentation in radiographs or MRI, where each distinct object requires identification, remains limited. Additionally, the research landscape concerning Vision Transformers (ViT) in dentistry is notably less extensive compared to the field of medicine [123], particularly in the context of identifying individual objects in radiographs. Despite being a driving force in advancing diagnostic imaging in dentistry, the utilization of these machine learning approaches is currently underrepresented [142].

Recent advancements in self-supervised learning, particularly masked image modeling (MIM), have presented an exciting avenue for training more effective and efficient deep learning models in contexts where annotated data is scarce or labor-intensive to produce. This approach, which involves the masking of parts of the input data to compel the model to predict the missing information, has proven to be particularly effective in improving model performance by leveraging the rich unlabeled data available in medical imaging [25].

Building on these technological advancements, this study introduces an innovative automated system utilizing deep learning-based object detection to precisely identify key landmarks such as the marginal bone level, top, and apex of dental implants from radiographs. Our approach is grounded in the principles of self-supervised learning, harnessing the potential of our newly proposed Deep Embedding of Patches (DEP) pretraining method—an advancement on the traditional masked autoencoder (MAE) transformer to markedly enhance the detection accuracy of peri-implant bone changes.

Acknowledging the challenge presented by the scarcity of comprehensive, high-quality datasets in dental radiography, we undertook the meticulous task of enriching an existing dataset with expert annotations of peri-implant landmarks, thereby creating the Bone Loss Assessment Dataset (BLAD).

To our knowledge, the detection of individual dental implants and the precise localization of significant landmarks, such as the marginal bone level, utilizing fully end-to-end deep learning methods, remains an area yet to be thoroughly explored. The current study aims to address this research gap by evaluating a deep learning model for the accurate localization of implants and the identification of key points within the detected implant site. Furthermore, this investigation includes the calculation of the marginal bone loss ratio and subsequent classification, offering potential assistance to dentists in the analysis of periapical radiographs.

Our contributions are threefold:

- We develop a cutting-edge deep learning methodology that leverages self-supervised learning for the detailed and accurate detection of marginal bone levels, tops, and apexes of implants in dental radiographs, paving the way for earlier and more precise interventions in the management of peri-implant diseases.

- We present the Deep Embedding of Patches (DEP) pretraining method, which represents a significant advancement over traditional baseline approaches in enhancing the detection accuracy of peri-implant bone loss. Rather than utilizing the mean squared error (MSE) metric in the pixel domain as the loss function, we advocate calculating the L_1 loss between the original and predicted embeddings of masked patches. Our experimental findings demonstrate that this modification results in enhanced performance. This is consistent with the observations reported in [130], which indicate that predicting deep embeddings instead of pixel values leads to better generalization and performance improvements.

- We introduce the Bone Loss Assessment Dataset (BLAD), a meticulously annotated dataset designed specifically for the training and evaluation of AI models in the detection of peri-implant bone loss. This dataset sets a new benchmark for data quality and specificity in the realm of dental radiography research.

Through these contributions, this study aims to bridge the gap in diagnostic capabilities within implant dentistry, democratizing access to high-level diagnostic tools for dental professionals worldwide. By enhancing the precision and reliability of peri-implant bone loss detection, we anticipate facilitating improved patient care outcomes and advancing the field of implant dentistry into a new era of AI-enabled diagnostics.

6.2 Related Work

The application of artificial intelligence (AI) and machine learning (ML) in medical imaging has seen substantial growth over the past decade. Particularly, deep learning (DL) approaches have gained prominence due to their ability to automatically learn and extract features from large datasets, which is critical in the domain of medical diagnostics. This section reviews the pertinent literature on deep learning applications in dental radiography, with a specific focus on the detection and analysis of peri-implant bone health.

Deep learning models have been widely applied to various tasks in medical image analysis, such as classification, segmentation, and object detection. Convolutional neural networks (CNNs) have been particularly effective in handling image data, thanks to their ability to capture spatial hierarchies and patterns within images. For example, CNNs have been successfully employed in the diagnosis of dental caries, the detection of periodontal diseases, and the identification of oral cancers from radiographic images [**?**, **?**, 113].

In the context of dental implantology, accurate detection of peri-implant bone levels is crucial for assessing the health and stability of dental implants. Traditional methods rely heavily on the expertise of clinicians to interpret radiographic images, which can be subjective and prone to variability. To address this challenge, researchers have explored the use of deep learning techniques to automate the detection and measurement of peri-implant bone levels. One notable study utilized a CNN-based approach to evaluate peri-implant bone loss, demonstrating significant improvements in diagnostic accuracy compared to traditional methods [?].

Despite these advances, there remains a significant gap in the literature regarding the use of deep learning for instance segmentation in dental radiographs. While U-Net and its variants have been commonly used for medical image segmentation [17], the application of these models to the segmentation of individual dental implants and the precise localization of peri-implant landmarks has been limited.

Vision Transformers (ViTs) have recently emerged as a powerful alternative to CNNs, offering a different mechanism for capturing image features through self-attention mechanisms. ViTs have shown promising results in various image classification tasks and are gaining traction in medical imaging research [20]. However, their application in dental radiography, particularly for tasks such as instance segmentation and landmark detection, is still in its infancy. The potential of ViTs to handle the complex and detailed nature of dental radiographs presents an exciting avenue for future research.

Self-supervised learning, and specifically masked image modeling (MIM), has introduced new possibilities for training deep learning models in scenarios with limited labeled data. MIM techniques, such as the Masked Autoencoder (MAE), involve masking parts of the input image and training the model to predict the missing regions. This approach has been shown to improve the robustness and generalization of models, making it particularly useful in medical imaging where annotated data is often scarce [25].

The proposed Deep Embedding of Patches (DEP) pretraining method builds on these advancements by incorporating self-supervised learning into the training process. By calculating the L_1 loss between the original and predicted embeddings of masked patches, rather than relying on pixel-level reconstruction, the DEP method aims to enhance the model's ability to capture meaningful features and improve detection accuracy. This approach aligns with recent findings that suggest predicting deep embeddings can lead to better performance than traditional pixel-based methods [130].

While significant progress has been made in the application of deep learning to dental radiography, there remains a considerable need for further research, particularly in the areas of instance segmentation and landmark detection. The introduction of advanced techniques such as ViTs and self-supervised learning holds promise for addressing these challenges and advancing the field. This study seeks to contribute to this evolving landscape by proposing a novel deep learning-based system for the detection and analysis of peri-implant bone levels, leveraging the latest advancements in self-supervised learning and vision transformer architectures.

6.3 Methods

6.3.1 Self-supervised Learning

Our main contribution is replacing patch reconstruction with deep embedding prediction within the MAE framework. This shift introduces a new dimension to self-supervised learning, specifically tailored for dental radiographs, resulting in substantial performance improvements. By adopting this approach, our model more effectively captures the nuances of dental radiographic imagery, enhancing the accuracy of key landmark detections, such as the marginal bone level, top, and apex of implants. These contributions are clearly articulated in this section. Here, we outline the components of the Deep Embedding of Patches (DEP), including the encoder, decoder, and the associated loss function.

Tokenizer. The input is partitioned into non-overlapping patches, randomly categorized into visible and masked groups, as depicted in Fig. 6.1(Left).

Masked Sequence Generation. Patch embeddings E are represented as a set. Following the MAE approach, a subset of patches is randomly masked, denoted as E_m, while unmasked embeddings are labeled as E_{um}. Masked embeddings E_m are substituted with a shared learnable mask embedding E_{mask}. Corrupted embeddings

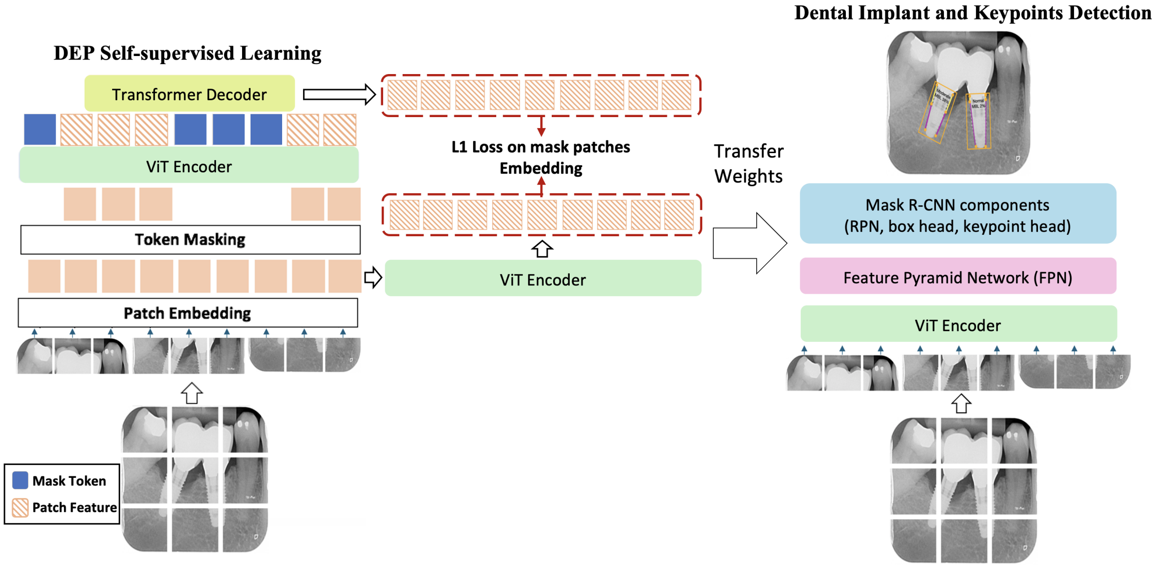

DEP Self-supervised Learning
Transformer Decoder
ViT Encoder
Token Masking
Patch Embedding
Mask Token
Patch Feature
L1 Loss on mask patches Embedding
ViT Encoder
Transfer Weights
Dental Implant and Keypoints Detection
Mask R-CNN components (RPN, box head, keypoint head)
Feature Pyramid Network (FPN)
ViT Encoder

Figure 6.1. Dental Implant and keypoints Detection Pipeline with DEP Self pretraining. The initial step in the dental implant and keypoints detection pipeline for DEP self-pretraining involves dividing the input into non-overlapping patches. These patches undergo embedding using an MLP. Throughout the pretraining phase, the patch embeddings undergo random masking, and only the visible embeddings are employed by the transformer. Subsequently, the masked embeddings are merged with the encoded embeddings and directed to the decoder. The decoder's role is to reconstruct the masked patches, followed by predicting the patch embeddings of these masked patches. The L_1 loss is employed to assess the similarity between the masked patch embeddings. Once pretraining is complete, the decoder is omitted, and the encoder is utilized as the backbone in Mask R-CNN with FPN for the detection.

E_c are generated by combining E_{um} with the sum of E_{mask} and positional embeddings p.

Encoder. The DEP encoder exclusively processes visible patches, incorporating position embeddings to preserve positional information. The resulting representation is used to reconstruct the masked input.

Decoder. The DEP decoder receives a complete set of tokens, including patch-wise representations from the encoder and learnable mask tokens. By integrating positional embeddings with input tokens, the decoder aims to restore each patch embedding within its masked position, serving as an auxiliary module exclusively for pretraining.

As stated above the original MAE restores the original patches, not their embeddings.

Loss computation. Instead of employing mean squared error (MSE) in pixel space, we propose computing the L_1 loss between original and predicted embeddings of masked patches. This change, as evidenced by our experimental results, leads to performance improvements.

6.3.2 Architectures for Downstream Tasks

After completing self-pretraining with DEP, we attach task-specific heads for the subsequent tasks, namely, the detection of dental implants and the detection of key-points.

The pre-trained ViT weights are utilized to initialize the encoder for detection. The features from the ViT backbone are conveyed to both the neck (FPN [31]) and the detection head (Mask R-CNN with Keypoint Head) to facilitate bounding box regression and classification. We opt for the Mask R-CNN [131] framework, given its widespread use in object detection research. Subsequently, the entire network undergoes fine-tuning to execute the detection task.

Mask R-CNN with Keypoint Head

In this phase, individual implants are detected with bounding boxes. Based on the detected region, the six keypoints, including mesial and distal marginal bone level, are predicted.

For this procedure, a modified R-CNN architecture was used, Mask R-CNN. Mask R-CNN, the latest descendant of the R-CNN model, comprised a “backbone” and “heads” [131]. The backbone network is Vision Transformer (ViT) which outputs feature maps from the original input image. It can be of various types, but the feature pyramid network (FPN) [31] based on ViT is known for robust results when used for Mask R-CNN, and, thus, it was adopted in this study.

Using the feature maps from the backbone network, the box head performed object classification and bounding box regression, and the mask head performed the object segmentation task. By attaching a keypoint detection head and properly training the network, the model can predict specific keypoints on the objects that were detected by the box head. As shown in the previous study, this method with a keypoint head can be used for human pose estimation, wherein the model picks some keypoints of the human body, such as eyes, elbows, and knees [131]. In the present study, we adopted this architecture, the Mask R-CNN based on ViT backbone with a keypoint detection head. The scheme of the model is shown in Fig. 6.1.

The model was trained and tested for detecting implants and locating the six keypoints on each detected implant in dental periapical radiographs. The six keypoints were peri-implant bone level, the implant apex, and the implant top, which all have right and left sides as shown in Fig. 6.2. To cover various types of implants, the most coronal thread was annotated as the top of the implant. We refer to these six positions as ”keypoints” since it is a widely used terminology in point detecting tasks, such as human pose estimation [143, 144] or facial keypoint detection [145].

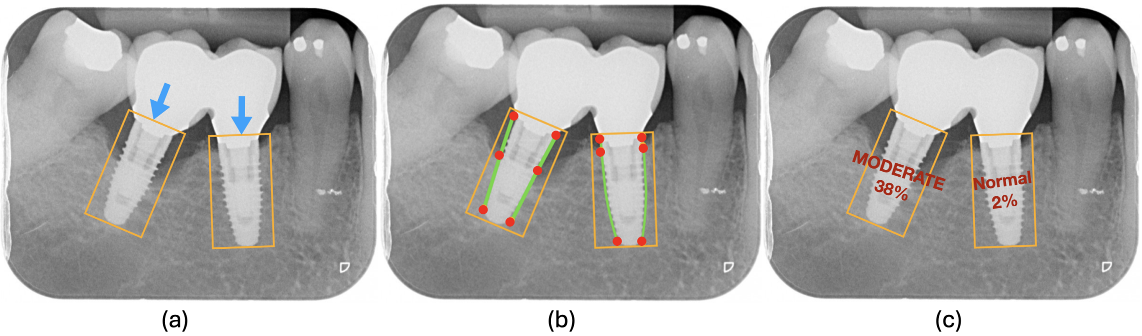


Figure 6.2. Depiction of anticipated bounding boxes and keypoints includes: (a) bounding boxes, indicated by blue arrows; (b) keypoints, represented as red dots positioned at the center of the radiographs; (c) calculation of radiographic bone loss percentage, based on the locations of the keypoints.

Bone Loss Ratio and Classification

Some studies exist on classification systems for peri-implantitis that use radiographic bone loss together with clinical indicators, such as bleeding/suppuration on probing or probing depth [146–148]. These studies use the ratio of the radiographic bone loss over the total implant length to classify the peri-implantitis. Based on the criteria suggested by these studies, we calculated and classified the bone loss ratio so that dental practitioners can easily refer to it.

Using the coordinates of the six keypoints that resulted from the prediction, the total length of the implant and the implant length that are not surrounded by sound bone can be calculated. The total length was measured from the center of the apex to the center of the implant top, and the length corresponding to the radiographic bone loss was measured from the center of the implant top to the center of the two marginal bone level keypoints. From these values, the percentage of the implant length in the bone defect site over the total length was calculated.

Based on this percentage, the severity of the bone loss around the implant was classified. As suggested by previous studies [146–148], the severity was categorized into four groups: normal, if the percentage is $\leq$10%, early, if the percentage is ¿10% and $\leq$25%, moderate, if the percentage is ¿25% and $\leq$50%, and severe, if the percentage is ¿50%.

6.3.3 Evaluation Methods

The evaluation framework consists of two distinct stages in the prediction workflow: implant bounding box detection and implant keypoint identification. Accordingly, two separate evaluation metrics are employed for each phase.

Intersection over Union (IoU)

To gauge the accuracy of the model in identifying implants, an effective metric is required to ascertain the precision of the model-generated bounding boxes in comparison to the actual bounding boxes. The Intersection over Union (IoU), also recognized as the Jaccard index, serves this purpose. The IoU metric is derived by dividing the area of intersection between the ground truth bounding box (A) and the model's predicted bounding box (B) by the union area of both boxes.

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} \tag{6.1}$$

By adjusting the IoU threshold values, it is possible to calculate the model's average precision (AP) and average recall (AR).

Object Keypoint Similarity (OKS)

For assessing the accuracy of the model's keypoint detection, the Object Keypoint Similarity (OKS) metric is utilized, functioning similarly to the IoU for object detection evaluations. OKS, which varies between 0 and 1, measures how closely the model's keypoint predictions align with the actual ground truth, with values approaching 1 indicating higher accuracy. OKS is computed for each detected implant as follows.

$$\text{OKS}_j = \frac{\sum_i \left[\exp\left(\frac{-d_{ji}^2}{2s_j^2k_i^2}\right)\delta\left(v_{ji} > 0\right)\right]}{\sum_i \delta\left(v_{ji} > 0\right)} \tag{6.2}$$

In the equation, j denotes each individual implant, and i corresponds to each keypoint type. In this study, i represents various anatomical points such as the left bone level, right bone level, left apex, right apex, left implant top, and right implant top. Furthermore, v are the visibility flags for the ground truth keypoints, where $v = 0$ indicates not labeled, $v = 1$ signifies labeled but not visible, and $v = 2$ means labeled and visible. The keypoints for each implant are represented as $[x_1, y_1, v_1, \ldots, x_6, y_6, v_6]$, with x, y indicating the keypoint locations, and v the visibility flag.

Consider a vector $\vec{d_{ji}}$ that extends from a ground truth keypoint to a detected keypoint, where d_{ji} denotes the distance between these two keypoints. Additionally, s_j is defined as the scale of implant j, calculated as the square root of the ground truth segmented area of the implant.

Moreover, k_i is considered as the per-keypoint standard deviation multiplied by a constant factor, chosen here as 2 ($k_i = 2\sigma_i$). The per-keypoint standard deviation σ_i, standardized to the scale s of the implant, is derived from redundantly annotated images in the validation dataset, using the formula $\sigma_i^2 = E(j)[d_{ji}^2/s_j^2]$, where $E(j)$ signifies an average over j. As the mean of $\vec{d_{ji}}/s_j$ over j approaches a zero vector, σ_i can be calculated by averaging d_{ji}^2/s_j^2 over j.

OKS is then employed as a threshold to determine precision and recall based on keypoint detection. Among the detected keypoints of implants, only those with an OKS value exceeding the OKS threshold are deemed true positives. By adjusting the OKS thresholds, the AP, and AR values are derived.

6.4 Experiments

6.4.1 Dataset

The Implants Image Dataset [125], a collection of dental panoramic and periapical X-rays, includes 5572 images annotated with precise detection labels for dental implants, each measuring 416x416 pixels.

Our contribution extends this dataset by adding annotations for critical peri-implant landmarks, such as the levels of peri-implant bone, both apices of the implants, and the tops of the implants, marked for both the right and left sides. This annotation was carried out by a skilled prosthodontist using the COCO-Annotator tool [34].

We regard this dataset as the most comprehensive for the detection of dental implants and marginal bone levels in dental radiographs. The dataset, named Bone Loss Assessment Dataset (BLAD), is available upon request.

6.4.2 Evaluation Metrics

For our experiments, the dataset was segmented into five equal parts, with each part representing approximately 20% of the total images. A fixed set, containing

1116 images, was used as the consistent test dataset across all iterations, while the other four sets, each comprising 1114 images, were utilized for training and validation through a cross-validation approach. This method was repeated five times. The performance of the detection models was assessed using both the Average Precision and Average Recall metrics.

6.4.3 Details on Implementation

Our experiments were carried out using the PyTorch library [35] and executed on Nvidia Tesla V100 graphics processing units. The batch size was fixed at 4456 for all experiments, matching the total number of training instances. We utilized the AdamW optimizer [36] in every case.

Data Augmentation Techniques. We incorporated noise in up to 6% of image pixels and performed both horizontal and vertical flips, along with rotations of 90 degrees in both directions.

DEP pretraining Settings. We set an initial learning rate of 1.5e-4 and a weight decay of 0.05, with $\beta 1$ at 0.9 and $\beta 2$ at 0.95. The learning rate followed a cosine decay schedule, complemented by a 10-epoch warm-up phase. Our pretraining involved a random Masked Image Modeling technique using 16x16 patches, with a masking proportion of 25%. A linear prediction head was used for predicting images of 416x416 pixels.

Fine-Tuning for Specific Tasks. In the fine-tuning phase for particular tasks, we trained at a single scale. The learning rate was initiated at 0.0001, with a weight decay parameter of 0.05.

6.5 Results and analysis

6.5.1 DEP reconstruction

The reconstruction outcomes of MDE are depicted in Fig. 6.3. The figure comprises four columns illustrating the original images, masked images, images recon-

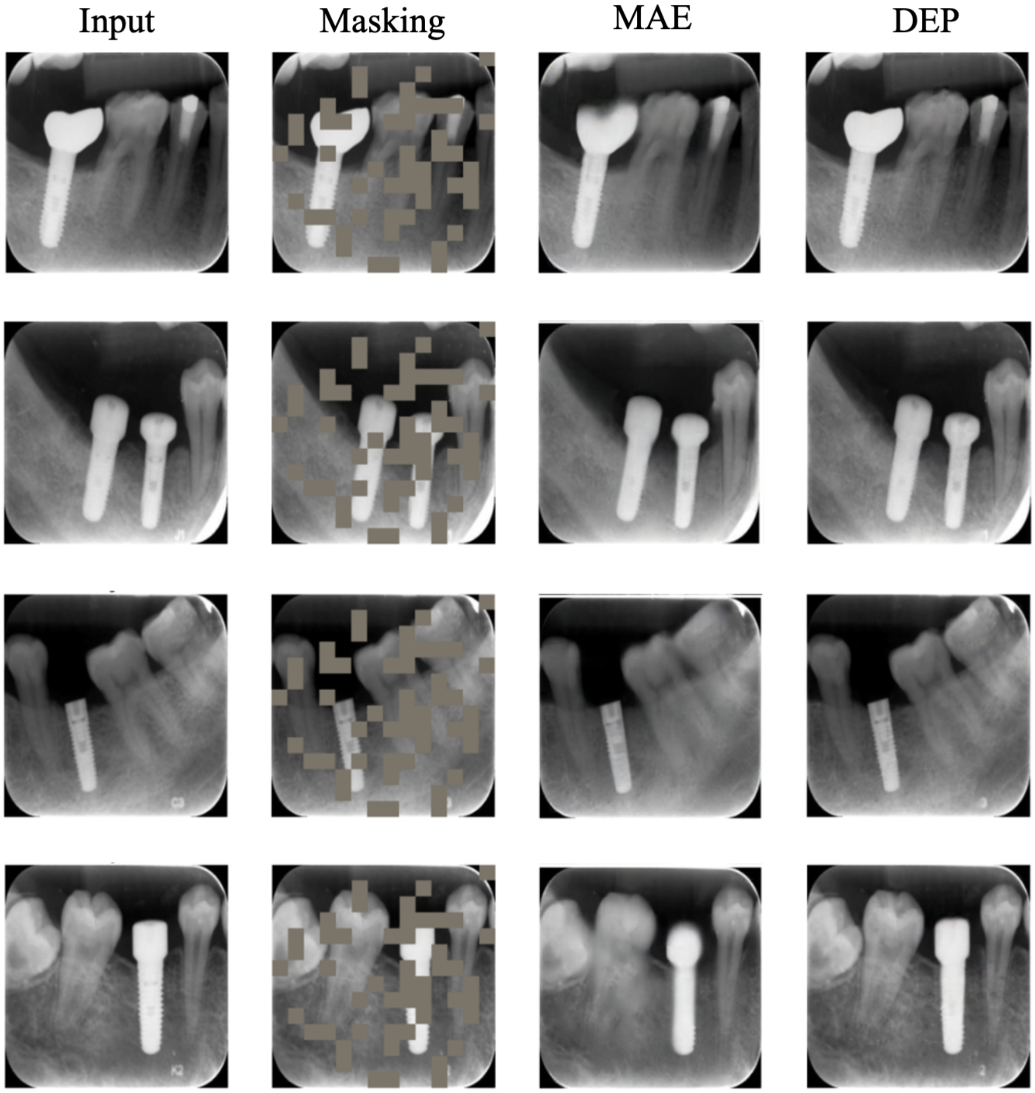


Figure 6.3. Results of DEP reconstruction. The first column displays the original images, while the second column shows the masked images, with blue patches indicating the masked regions. The third and fourth columns exhibit the reconstructions achieved through MAE and our DEP, respectively, from the unmasked patches.

structed using MAE, and images reconstructed using DEP. The results indicate that our approach excels in recovering missing information from the random context. It is important to emphasize that the primary objective of MIM is to enhance downstream tasks rather than produce reconstructions of the highest quality. We stress that our approach predicts patch embedding and does not predict pixel values. So we added for this visualization a process to perform the pixel reconstruction step, where the predicted patch embeddings guide a generative model (i.e., the decoding process) to produce image samples. This generative model uses the high-level information encoded in the embeddings to create new pixel values, providing a reconstructed representation of the original images.

6.5.2 Detection of dental implants and marginal bone levels

In this study, we evaluated the performance of various initialization and pretraining strategies for dental implant detection. The outcomes, summarized in Table 7.1, showcase the comparative analysis of different models and their effectiveness in accurately identifying dental implants from X-ray images.

The YOLOv5 model with a CSPDarknet backbone, pre-trained on the ImageNet-1K (IN-1K) dataset with labels, achieved an Average Precision (AP) of 91.5% and an Average Recall (AR) of 96.4%. The YOLOv8 and YOLOx models, also with CSPDarknet backbones and pre-trained on IN-1K with labels, showed slight improvements, achieving AP/AR scores of 91.7%/96.6% and 91.8%/96.7% respectively.

Switching to a Vision Transformer (ViT-B) architecture initialized with random weights and without any pretraining data resulted in an AP of 91.9% and an AR of 96.9%. This marginal improvement suggests that the ViT-B architecture has potential even without domain-specific pretraining.

Further experimentation involved supervised fine-tuning of the ViT-B on the dental X-rays dataset with labels, leading to an AP of 92.6% and an AR of 97.2%. This indicates the benefit of supervised fine-tuning for the task of dental implant detection. Using self-supervised learning methods, we observed more substantial improvements:

Table 6.1.
Results of dental implant detection.

Initialization	**Backbone**	**Pretraining Data**	$\boldsymbol{AP^{all}}$	$\boldsymbol{AR^{all}}$
YOLOv5 [132]	CSPDarknet	IN-1K w/ Labels	91.5	96.4
YOLOv8 [149]	CSPDarknet	IN-1K w/ Labels	91.7	96.6
YOLOx [150]	CSPDarknet	IN-1K w/ Labels	91.8	96.7
Random	ViT-B	None	91.9	96.9
Supervised Fine-tuning	ViT-B	IN-1K w/ Labels	92.6	97.2
DINOv2 [?]	ViT-B	IN-1K	92.8	97.4
SimMIM [22]	ViT-B	IN-1K	93.0	97.5
MAE [25]	ViT-B	IN-1K	93.2	97.8
DEP (ours)	ViT-B	IN-1K	**94.9**	**98.3**

the DINOv2 model achieved an AP of 92.8% and an AR of 97.4%, while SimMIM attained an AP of 93.0% and an AR of 97.5%.

The Masked Autoencoder (MAE) pretraining method yielded even higher performance, with an AP of 93.2% and an AR of 97.8%, demonstrating the effectiveness of MAE in enhancing the ViT-B model's detection capabilities.

The best performance was achieved with our proposed Deep Embedding of Patches (DEP), which, when applied to the ViT-B model pre-trained on the IN-1K dataset, resulted in the highest performance metrics—an AP of 94.9% and an AR of 98.3%. This demonstrates the effectiveness of our DEP approach in enhancing the model's ability to accurately detect dental implants, surpassing both traditional and alternative pretraining methods.

These results underline the potential of specialized pretraining strategies, such as DEP, in improving the performance of deep learning models for specific tasks like dental implant detection in radiographic imagery. The success of DEP highlights the

Table 6.2.
Results of keypoints detection.

Initialization	**Backbone**	**pretraining Data**	$\boldsymbol{AP^{all}}$	$\boldsymbol{AR^{all}}$
Random	ViT-B	None	93.5	97.1
Supervised Fine-tuning	ViT-B	IN-1K w/ Labels	94.0	97.8
MAE [25]	ViT-B	IN-1K	94.8	98.6
DEP (ours)	ViT-B	IN-1K	**95.7**	**98.9**

importance of tailoring pretraining approaches to the nuances of the target domain, in this case, dental radiographs, to achieve optimal results.

In the pursuit of enhancing detection accuracy for marginal bone levels in dental radiographs, we extended our investigation to evaluate the performance of different initialization and pretraining strategies tailored for this specific task. The summarized results in Table 7.2 reflect a rigorous comparative analysis, highlighting the advancements in detection precision facilitated through innovative pretraining approaches.

Commencing with a baseline model, a Vision Transformer (ViT-B) initialized with random weights and devoid of any pretraining, demonstrated a commendable Average Precision (AP) of 93.5% and an Average Recall (AR) of 97.1%. This set a foundational benchmark indicating the inherent capabilities of the ViT-B architecture in processing complex dental radiographic imagery.

Progressing to supervised pretraining utilizing the ImageNet-1K (IN-1K) dataset with labels, we observe an enhancement in performance metrics, with the AP increasing to 94.0% and the AR to 97.8%. This increment underscores the benefit of leveraging a large-scale labeled dataset to imbue the model with a richer understanding of visual features, which is pertinent to the nuanced task of marginal bone level detection.

The application of Masked Autoencoder (MAE) for unsupervised pretraining on the IN-1K dataset yielded further improvements, pushing the AP to 94.8% and the AR

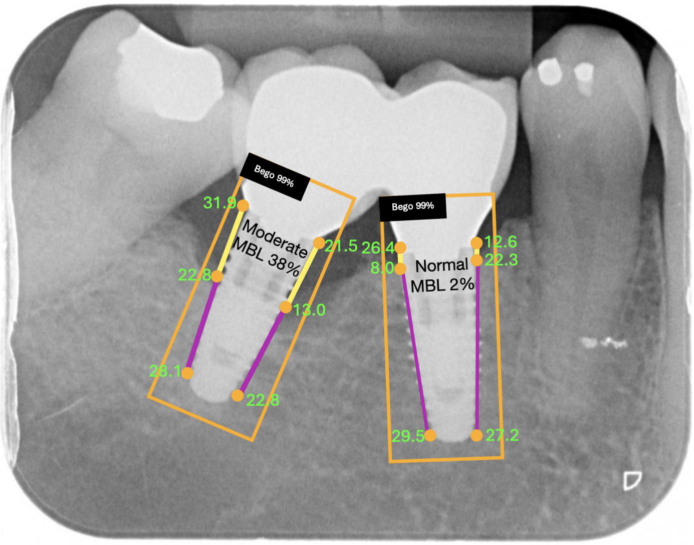


Figure 6.4. Illustration of anticipated outcomes: Each implant is identified using a bounding box, and the predicted keypoints are displayed within the box. The ratio of radiographic bone loss is determined from the positions of the keypoints. Additionally, confidence scores for both implant and keypoint detection are provided.

to 98.6%. This leap forward illuminates the effectiveness of self-supervised learning strategies in extracting and leveraging latent representations that are highly relevant to the task at hand, even in the absence of explicit labels.

Our proposed Deep Embedding of Patches (DEP) strategy, when applied to the ViT-B model alongside pretraining on the IN-1K dataset, achieved the highest performance metrics, with an AP of 95.7% and an AR of 98.9%. This remarkable outcome not only substantiates the superior efficacy of the DEP approach in detecting marginal bone levels but also signifies a notable advancement over conventional and alternative pretraining methods. An example of the predicted results are shown in Fig. 6.4.

These findings illuminate the pivotal role of specialized pretraining strategies, such as DEP, in pushing the boundaries of model performance for specific dental radiographic analysis. The notable success of DEP in both dental implant detection and marginal bone level identification underscores the strategy's potential to significantly

Table 6.3.
Impact of mask ratios and pretraining epochs on dental implant detection.

Mask Ratio	**pretraining Epochs**	$\boldsymbol{AP^{box}}$
65%	100	92.5
55%	100	93.2
55%	800	91.6
45%	100	94.0
35%	100	94.4
25%	100	**94.9**
15%	100	94.3

elevate the precision of radiographic interpretations in dental diagnostics, offering promising prospects for future research and clinical applications.

6.5.3 Parameter setting

In Table 7.4, we conduct experiments focusing on dental implant detection with varying pretraining epochs and mask ratios for our DEP method. Firstly, we observe that extending the training duration does not lead to improved performance for DEP. Secondly, in contrast to the high mask ratio used in natural images [25], we find distinct preferences for mask ratios in downstream tasks related to dental implant detection. Notably, both tasks consistently exhibit enhancements as the mask ratio decreases from 65% to 25%. This improvement may be attributed to the fact that relevant features on dental X-rays tend to be smaller in size.

6.6 Conclusions

This study demonstrates the significant improvement in detection accuracy for dental X-ray analysis provided by the innovative DEP pretraining method. Remarkably, DEP's self-supervised pretraining method outperforms existing techniques, especially in scenarios involving limited datasets, a challenge not adequately addressed before. Our research highlights the importance of tailoring specific parameters, such as the mask ratio and the duration of pretraining epochs, when implementing masked autoencoders for pretraining in dental radiography. These findings underscore the potential of DEP to further enhance the already impressive capabilities of Vision Transformers (ViTs) in dental X-ray analysis. Looking ahead, our next goal is to explore the impact of DEP pretraining on prognostic and outcome prediction tasks within dentistry.

CHAPTER 7
DENTAL CARIES DETECTION

7.1 Introduction

Dental caries, also known as tooth decay or cavities (illustrated in red circles in Figure 7.1), is a prevalent oral health issue with significant global health implications. According to the World Health Organization, nearly 60-90% of school children and almost 100% of adults worldwide have dental cavities, making it one of the most common non-communicable diseases. Early detection and intervention are crucial to preventing the progression of caries, which can lead to more severe dental complications such as pulpitis, abscesses, and tooth loss if left untreated. The economic burden associated with dental caries is substantial, impacting healthcare systems and individuals due to the costs of treatment and loss of productivity [151, 152].

Traditional methods of caries detection, such as visual-tactile examinations and radiography, have limitations in sensitivity and specificity, often resulting in delayed diagnosis and treatment [153–157]. Visual-tactile examinations rely heavily on the practitioner's experience and can be subjective, leading to variability in diagnosis. Radiographic methods, while more reliable, still face challenges such as overlapping structures in the images that can obscure caries, and the radiation exposure risk associated with frequent use. These conventional approaches are prone to human error and often require follow-up visits, increasing patient inconvenience and healthcare costs. These limitations highlight the need for more reliable and automated diagnostic tools that can provide consistent and accurate results, minimizing human error and ensuring timely treatment.

In recent years, the advent of computer vision and machine learning technologies has opened new avenues for enhancing diagnostic accuracy in various medical fields, including dentistry. Automated dental caries detection systems, leveraging these tech-

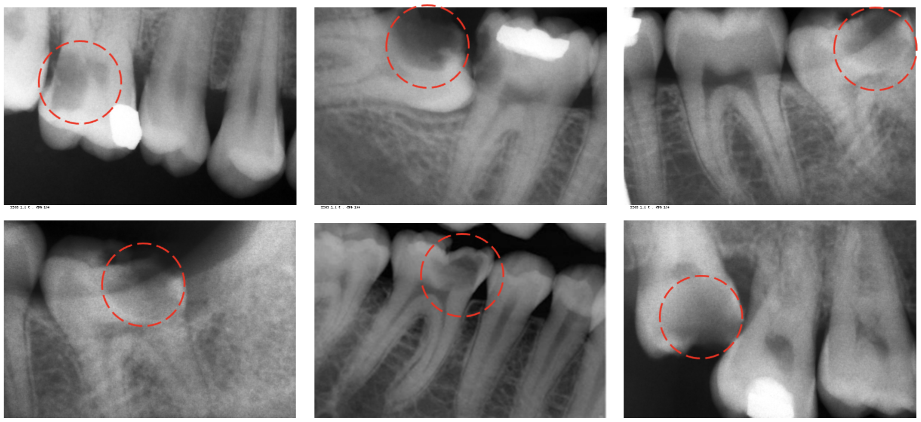


Figure 7.1. Illustration of dental X-ray images, displaying the various forms and appearances of dental caries.

nologies, have the potential to provide timely, accurate, and non-invasive assessments, thus improving patient outcomes and reducing healthcare costs. These advancements promise not only to streamline the diagnostic process but also to mitigate the subjectivity inherent in traditional diagnostic methods. Machine learning algorithms, particularly deep learning, have shown remarkable performance in image analysis tasks, enabling the development of systems that can analyze dental radiographs with high precision and consistency. These systems can assist dentists by highlighting potential carious lesions, thereby serving as a second opinion and reducing diagnostic errors.

Recent progress in self-supervised learning has underscored the potential of masked image modeling (MIM) [22, 24, 25, 122, 123] as an effective pre-training strategy for Vision Transformers (ViT) [37] and hierarchical Vision Transformers using shifted windows (Swin) [21, 39, 52, 124]. MIM, which involves masking image patches and reconstructing them, enables the network to infer masked regions by leveraging contextual information. This capability is particularly relevant for the analysis of dental radiographs, where integrating contextual information can significantly improve diagnostic accuracy. Given the subtle and complex nature of dental caries, an enhanced ability to discern these details through sophisticated image analysis is crucial. Self-supervised learning methods do not require labeled data for pre-training, which is advantageous in medical fields where acquiring large labeled datasets can be challenging and expensive.

In this study, we propose a novel approach to dental caries detection using self-supervised learning with Masked Deep Embeddings of Patches (MDEP) for dental radiographs. Our method builds upon the framework of the Masked Autoencoder (MAE) [25] but focuses on enhancing representation learning by predicting deep embeddings of masked patches instead of reconstructing them. This approach leverages the Vision Transformer (ViT) architecture [37], where the encoder processes visible patches, and the MDEP loss function evaluates the predicted embeddings of the masked patches. To further validate our approach, we experiment with an additional dataset called CariesXrays [158], demonstrating the robustness and adaptability of our method across different datasets. By utilizing the power of deep learning, our technique extracts and utilizes intricate details from radiographic images, thereby improving the overall diagnostic process for dental caries detection.

Self-pre-training is especially beneficial when domain-specific pre-training data is scarce, as it mitigates domain discrepancies between pre-training and fine-tuning stages by utilizing the same dataset. We apply our MDEP pre-training method on the same dataset used for the downstream task of dental caries detection, ensuring consistency and relevancy in the learned representations. This approach addresses a significant challenge in medical image analysis, where labeled data is often limited and expensive to obtain. Furthermore, the MDEP method leverages the advantages of both self-supervised learning and the powerful ViT architecture to create a robust model capable of high performance even with limited data. Our main contributions are threefold:

- We introduce a self-supervised learning framework with masked deep embeddings, specifically tailored for dental radiograph analysis. This framework replaces the traditional MAE reconstruction of masked patches with the reconstruction of patch embeddings.

- We demonstrate that MDEP self-pre-training significantly improves performance in dental caries detection, outperforming other state-of-the-art methods.
- We conduct an extensive evaluation of our method on two different types of radiographs, periapical [159] and panoramic (CariesXrays dataset [158]), showcasing its robustness and generalizability across different patient populations and imaging conditions.

The proposed method addresses the challenge of limited data availability in dental radiographs and presents a robust solution for automated dental caries detection, potentially transforming diagnostic practices and improving patient outcomes. By advancing the capabilities of dental diagnostics through cutting-edge machine learning techniques, this research contributes to the broader goal of enhancing healthcare delivery and patient care in the field of dentistry. The implications of this study extend beyond dental caries detection, suggesting that self-supervised learning can be effectively applied to other areas of medical imaging where data scarcity is a significant challenge.

By leveraging the strengths of modern machine learning techniques and addressing the limitations of traditional diagnostic methods, this research offers a pathway toward more accurate and efficient dental care. The integration of automated systems in dental diagnostics not only improves the accuracy of caries detection but also has the potential to revolutionize the field of dentistry by enabling early intervention, personalized treatment plans, and ultimately better patient outcomes.

7.2 Related Work

7.2.1 Dental X-rays

Recent advancements in artificial intelligence and deep learning have significantly reduced the effort and error rates in identifying dental caries [160]. Various machine learning methods have been integrated into clinical practices to enhance the diagnosis of dental caries [161]. Techniques for localizing dental caries employ image

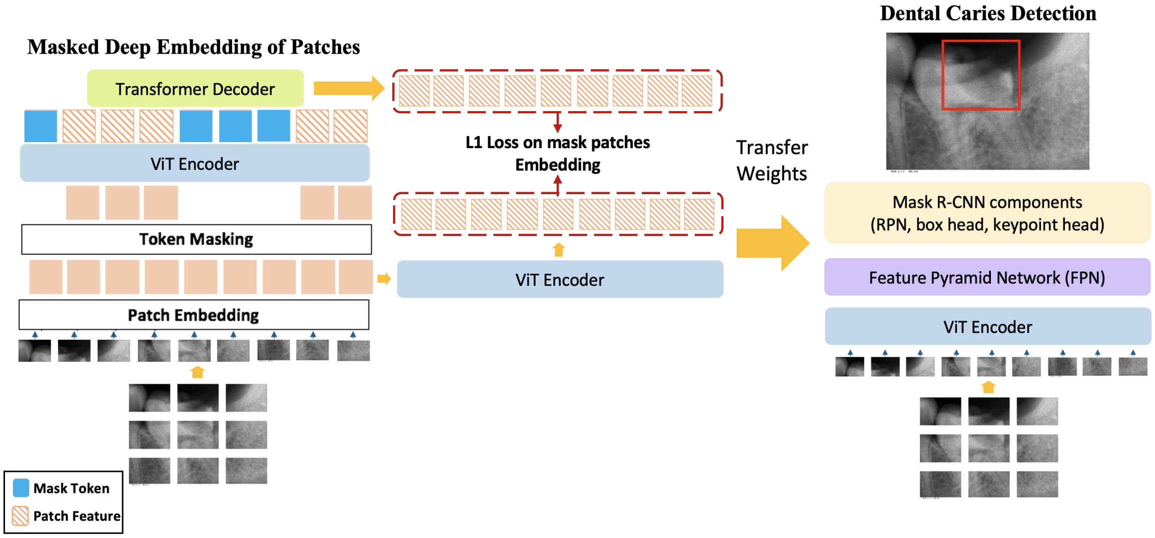


Figure 7.2. Dental Caries Detection Pipeline with MDEP Self Pre-training. The first step in this pipeline involves splitting the input into non-overlapping patches, which are then embedded using a multi-layer perceptron (MLP). During the pre-training phase, random masking is applied to the patch embeddings, with only the visible embeddings processed by the transformer. The masked embeddings are combined with the encoded embeddings and sent to the decoder. The decoder's task is to predict the embeddings of the masked patches. The similarity between the predicted and actual masked patch embeddings is evaluated using the L_1 loss. After pre-training, the decoder is discarded, and the encoder serves as the backbone in the Mask R-CNN with FPN for detection.

processing, classifiers, and neural networks to detect affected dental regions [13, 162]. Additionally, several tools have been developed for dental caries localization. This section reviews the approaches and tools used in dental caries localization and discusses state-of-the-art detection and treatment techniques within a clinical context.

Different types of images are used in clinical settings to diagnose carious regions, including panoramic X-rays, periapical X-rays, bitewing X-rays, and ultrasound [163]. Each imaging modality has its own set of challenges and advantages in dental caries detection.

Panoramic X-ray images provide a comprehensive view of the entire dentition and surrounding structures but can be challenging to interpret due to the inclusion of surrounding areas such as the chin, spine, and jaws [113]. This can complicate the localization of dental cavities, often requiring multiple images to achieve accurate results [164]. The complexity of these images can lead to less satisfactory results in detecting dental cavities and increase the risk of misdiagnosis, potentially resulting in improper treatment [165].

Periapical X-rays, on the other hand, offer detailed views of specific teeth and their surrounding bone structures, demonstrating good accuracy in detecting carious regions even with small datasets [166]. These images are particularly effective for identifying issues at the root level and are less prone to the complexities seen in panoramic X-rays.

Bitewing radiographs, although useful for detecting proximal and occlusal lesions, have shown lower sensitivity for identifying non-cavitated lesions [167]. Ultrasound imaging for cavity localization via deep learning also presents challenges, requiring extensive expertise for accurate interpretation [168].

In our research, we tested our model using both periapical and panoramic X-rays to leverage the strengths of each modality. Periapical X-rays provided high-resolution, localized views that facilitated precise caries detection, while panoramic X-rays offered a broad overview of the dental structures, enabling comprehensive analysis despite the challenges in localizing specific lesions. This dual approach allowed us to validate the robustness and versatility of our model in different clinical scenarios, ultimately enhancing its diagnostic accuracy and reliability.

7.2.2 Self-supervised Learning

Masked language modeling and its autoregressive variants, such as BERT [169] and GPT [170–172], have proven to be exceptionally effective for pre-training in natural language processing (NLP). These techniques involve withholding a segment of the input sequence and training models to predict the missing parts. It has been demonstrated that these methods scale efficiently [172], and extensive evidence shows that these pre-trained representations generalize effectively across various downstream tasks.

Autoencoding is a traditional approach for learning representations, involving an encoder that maps an input to a latent representation and a decoder that reconstructs the input. Examples of autoencoders include PCA and k-means [173]. Denoising autoencoders (DAE) [174] are a subclass of autoencoders that introduce noise to the input signal and train the model to reconstruct the original, uncorrupted signal. Several methods can be considered generalized DAEs under different corruption strategies, such as masking pixels [175–177] or removing color channels [178]. Our Masked Deep Embedding of Patches (MDEP) is a type of denoising autoencoder, differing from traditional DAEs in several aspects.

Masked image encoding methods focus on learning representations from images corrupted by masking. The foundational work of Vincent et al. [175] introduces masking as a type of noise in DAEs. Context Encoder [176] utilizes convolutional networks to inpaint large missing regions. Inspired by the success in NLP, recent methods [24, 37, 177] employ Transformers [77]. For instance, iGPT [177] operates on pixel sequences and predicts unknown pixels, while the ViT paper [37] investigates masked patch prediction for self-supervised learning. Recently, BEiT [24] has suggested predicting discrete tokens [179, 180].

Self-supervised learning methods have garnered considerable interest in computer vision, often centered on various pretext tasks for pre-training [178, 181–185]. Recently, contrastive learning [186, 187] has gained popularity, exemplified by methods such as [188–191], which model image similarity and dissimilarity (or solely similarity [192, 193]) between multiple views. Contrastive and related methods heavily rely on data augmentation [191–193]. In contrast, autoencoding follows a conceptually distinct path and exhibits different behaviors, as we will demonstrate.

7.2.3 Object Detection

Previous research on object detection has significantly advanced computer vision. The introduction of R-CNN [194] was a groundbreaking milestone, revolutionizing object detection by proposing the concept of region proposals, which led to accurate localization and classification of objects within images. Inspired by R-CNN's success, subsequent studies aimed to enhance the speed and efficiency of object detection algorithms, resulting in developments like Fast R-CNN [195], Faster R-CNN [61], and SSD [196]. Additionally, the inclusion of two-stage detectors, such as Mask R-CNN [131] and Cascade R-CNN [197], has further amplified object detection systems' capabilities. Furthermore, the YOLO family [150,198,199] has made substantial contributions to computer vision by providing real-time and efficient object detection solutions. DERT [200] employs a transformer encoder-decoder architecture, capturing long-range dependencies and modeling global contextual information in object detection. Despite advancements in conventional object detection algorithms, dental caries detection remains challenging due to the complex and multifaceted nature of caries presentations.

The CariesXrays dataset, introduced by Chen et al. [158], represents a significant advancement in dental caries detection. This hospital-scale panoramic dental X-ray benchmark comprises 6,000 panoramic dental X-ray images with 13,783 instances of dental caries meticulously annotated by dental professionals. The authors proposed a novel Feature Pyramid Contrastive Learning (FPCL) framework that incorporates a dual-directional feature pyramid network (D2D-FPN) and a proposals-prototype contrastive regularization learning (P2P-CRL) mechanism to enhance the generalization ability and detection accuracy of dental caries. Their extensive experiments on the CariesXrays dataset demonstrated the potential of FPCL to make a significant social impact on caries diagnosis by providing a robust and efficient diagnostic tool for dental practitioners.

7.3 Methods

7.3.1 Masked Autoencoders

Encoding. As depicted in Figure 7.2 (Left), the input image is divided into non-overlapping patches. These patches are then randomly separated into two groups: visible and masked. The encoder of the Masked Autoencoder (MAE) processes only the visible patches, incorporating positional embeddings to retain spatial information. This step ensures that the model learns the spatial dependencies between different parts of the image, which is essential for understanding the context of the patches during the pre-training phase.

Masked Sequence Generation. Patch embeddings E are treated as a set. Following the MAE method, a subset of patches is randomly masked, denoted as E_m, while unmasked embeddings are denoted as E_{um}. The masked embeddings E_m are replaced with a shared learnable mask embedding E_{mask}. Corrupted embeddings E_c are created by combining E_{um} with the sum of E_{mask} and positional embeddings p, which are then inputted into the encoder. This approach allows the model to learn robust feature representations by predicting the embeddings of masked patches based on their surrounding context, which is crucial for capturing the intricate details necessary for accurate dental caries detection.

Decoding. The MAE decoder is not involved in this process since our focus is on obtaining deep embeddings rather than reconstructing the original patches. The embeddings generated by the encoder are directly used for further tasks, ensuring that the learned representations capture the essential features of the dental radiographs.

Loss Function. Instead of using the mean squared error (MSE) in pixel space as in traditional MAEs, we propose computing the L_1 loss between the original and predicted embeddings of masked patches. Our experimental results demonstrate that this approach leads to improved performance. This is consistent with findings in [130], which suggest that predicting deep embeddings of patches rather than pixel values results in better generalization and performance enhancements. The L_1 loss is more robust to outliers and can lead to sharper and more accurate embeddings, which is particularly beneficial for medical image analysis where fine details are crucial.

7.3.2 Architectures for Downstream Tasks

Upon completing self-pre-training using MAE, we incorporate a task-specific head for the subsequent task, which in this case is the detection of dental caries. The integration of a task-specific head allows the model to fine-tune its learned representations for the specific requirements of dental caries detection, ensuring optimal performance.

The pre-trained Vision Transformer (ViT) weights are employed to initialize the encoder for the detection task. The features extracted by the ViT backbone are then passed to both the neck (Feature Pyramid Network, FPN) and the detection head (Mask R-CNN) to facilitate bounding box regression and classification. We choose the Mask R-CNN framework [131] due to its widespread adoption in object detection research. The FPN helps in building high-level semantic feature maps at different scales, which improves the model's ability to detect objects of various sizes. The Mask R-CNN adds an additional mask output, which provides pixel-level segmentation information, making it a suitable choice for precise localization tasks such as dental caries detection.

The entire network undergoes fine-tuning to perform the detection task. During this phase, the pre-trained weights serve as a strong initialization, allowing the network to converge faster and perform better with limited training data. Fine-tuning involves adjusting the weights of the entire network, including the pre-trained encoder and the newly added layers, to optimize for the specific task of dental caries detection. This process ensures the model can accurately identify and localize carious lesions in dental radiographs, leveraging the robust features learned during the self-supervised pre-training phase.

7.3.3 Training and Data Augmentation

Effective data augmentation techniques are employed to increase the diversity of the training dataset and improve the model's robustness. These techniques include random rotations, flipping, and adding noise to the images. Data augmentation helps in mitigating overfitting and enhances the model's generalizability to new, unseen data. Additionally, all images are normalized to a standard scale and resolution to ensure consistency during training.

The training process involves a two-stage approach. First, the self-supervised pre-training phase using the proposed MDEP framework is conducted. This phase focuses on learning meaningful representations from the unlabelled dental radiographs. Second, the fine-tuning phase adapts these learned representations to the specific task of dental caries detection using labeled data.

7.4 Experimental Evaluation

7.4.1 Data Sets and Evaluation Metrics

We use two datasets in this study. The first dataset [159] comprises dental periapical X-rays with annotations for dental caries detection. This dataset contains 936 images, each sized at 748×512 pixels. We partitioned the dataset into five subsets for cross-validation purposes. One subset, comprising 188 images, was used as the test dataset, while the remaining four subsets, each containing 187 images, formed the training and validation datasets. This process was iterated five times.

To further validate the robustness and generalizability of the proposed method, we conducted additional experiments using the CariesXrays dataset [158]. This dataset consists of 6,000 annotated panoramic radiographs, each sized at 1333×800 pixels. Similar to the previous dataset, we partitioned the CariesXrays dataset into five subsets for cross-validation. One subset, comprising 1,200 images, was used as the test dataset, while the remaining four subsets, each containing 1,200 images, formed the training and validation datasets.

Initialization	**Backbone**	**Pre-training Data**	***P***	***R***	***AP***	***AP*@50**	***AP*@75**
M-RCNN [159]	ResNet-50	IN-1K w/ Labels	95.8	96.2	-	-	-
Random	ViT-B	None	96.9	96.8	95.1	95.9	94.4
Supervised	ViT-B	IN-1K w/ Labels	97.3	97.3	96.2	96.9	95.7
MAE	ViT-B	IN-1K	97.9	97.9	96.7	97.5	96.2
MDEP (ours)	ViT-B	IN-1K	**98.1**	**99.0**	**97.3**	**98.1**	**97.1**

Table 7.1.
Results of dental caries detection on periapical X-rays dataset [159] in terms of precision (P), recall (R), average precision (AP), and average precision at IoU thresholds of 50% (AP@50) and 75% (AP@75).

The evaluation metrics for object detection models include precision (P), recall (R), and Average Precision (AP). Precision is the ratio of true positives to total predicted positives, while recall is the ratio of true positives to all actual positives. Average Precision (AP) summarizes the precision-recall curve, with AP@50 and AP@75 measured at IoU thresholds of 0.50 and 0.75, respectively.

7.4.2 Quantitative Results

The proposed approach, Masked Deep Embeddings of Patches (MDEP), demonstrates superior performance across multiple metrics on two dental caries detection datasets: the periapical X-rays dataset and the CariesXrays dataset.

For the periapical X-rays dataset (Table 7.1), our method outperforms the Mask R-CNN baseline and other ViT-B based models. Specifically, MDEP achieves a precision (P) of 98.1%, a recall (R) of 99.0%, and an average precision (AP) of 97.3%. These results indicate a significant improvement in both precision and recall, with the highest AP values at different IoU thresholds (AP@50 and AP@75) compared to other methods. The results of the M-RCNN method are taken from [159]. It is

Initialization	**Backbone**	**Pre-training Data**	***AP***	***AP*@50**	***AP*@75**
FPCL [158]	ResNet-50	IN-1K	48.2	84.1	50.6
Random	ViT-B	None	46.3	82.4	49.1
Supervised	ViT-B	IN-1K w/ Labels	49.8	85.3	52.2
MAE	ViT-B	IN-1K	51.6	86.7	54.3
MDEP (ours)	ViT-B	IN-1K	**54.4**	**87.6**	**60.9**

Table 7.2.
Results of dental caries detection on CariesXrays dataset [158] in terms of Average Precision (AP) metrics.

worth noting that [159] did not report AP results, which is why they are not included in Table 7.1. The improvement can be attributed to the effective utilization of masked autoencoders for pre-training, which enhances the model's ability to capture fine-grained features in dental X-ray images, thus leading to better detection performance.

On the CariesXrays dataset (Table 7.2), our MDEP method also sets new benchmarks. With an AP of 54.4%, AP@50 of 87.6%, and AP@75 of 60.9%, MDEP surpasses previous state-of-the-art methods, including FPCL, which achieved an AP of 48.2%. The substantial margin by which MDEP outperforms existing methods highlights the robustness and generalization capability of our approach. The masked deep embeddings allow the model to learn more generalized and transferable features. In particular, the significant improvement over MAE shows the importance of predicting deep embeddings instead of actual pixel values of image patches.

When comparing our method with other state-of-the-art techniques on the CariesXrays dataset (Table 7.3), MDEP maintains a consistent lead. Traditional methods like SSD [196] and RetinaNet [31] show considerably lower performance, while even advanced models like YOLOv8 [149] and Conditional-DETR [203] fall short of MDEP's results. The highest scores achieved by MDEP can be linked to its transformer back-

Initialization	**Backbone**	***AP***	***AP*@50**	***AP*@75**
SSD [196]	VGG	12.7	36.2	5.90
RetinaNet [31]	ResNet-50	13.0	30.5	10.2
DETR [200]	Transformer	25.7	64.5	13.7
EfficientDet [201]	EfficientNet	34.1	52.5	36.0
FCOS [202]	ResNet-50	35.9	75.6	29.5
YOLOv7 [199]	CSPDarkNet	39.3	79.8	34.3
Faster R-CNN [61]	ResNet-50	39.9	78.0	37.8
YOLOv8 [149]	CSPDarkNet	40.3	80.7	35.5
YOLOx [150]	CSPDarkNet	40.5	81.3	36.1
Conditional-DETR [203]	Transformer	42.2	80.6	40.4
FPCL [158]	ResNet-50	48.2	84.1	50.6
MDEP (ours)	Transformer	**54.4**	**87.6**	**60.9**

Table 7.3.
Comparison with state-of-the-art on CariesXrays Dataset [158].

bone, which provides a more effective architecture for handling the complex patterns present in dental images. The use of self-supervised learning further enhances the model's ability to discern subtle caries features, resulting in higher detection accuracy. The results of other methods in Table 7.3 are taken from [158].

The significant performance gains achieved by MDEP across various metrics underscore the effectiveness of our approach in dental caries detection. The superior results demonstrate the advantage of leveraging masked autoencoders and transformers in medical image analysis, paving the way for more accurate and reliable diagnostic tools in dental care.

7.4.3 Qualitative Results

Figures 7.3 and 7.4 compare dental caries detection using the state-of-the-art methods and our proposed Masked Deep Embeddings of Patches (MDEP) on two datasets: the periapical radiographs dataset [159] and the CariesXrays dataset [158].

7.4.4 Mask Ratio Analysis

The impact of different mask ratios on the precision (P) of dental caries detection was evaluated, as presented in Table 7.4. Various mask ratios were tested, each pre-trained for 100 epochs, except for one configuration which was pre-trained for 800 epochs. The results highlight the relationship between mask ratio, pre-training epochs, and detection precision.

The results show a clear trend in precision performance with varying mask ratios:

At higher mask ratios, such as 65% and 55%, the precision values are 95.5% and 95.8%, respectively. The model pre-trained for 800 epochs at a 55% mask ratio yielded a slightly lower precision of 95.6%. This suggests that increasing the number of pre-training epochs does not significantly enhance performance at higher mask ratios.

Mask Ratio	**Pre-training Epochs**	***P***
65%	100	95.5
55%	100	95.8
55%	800	95.6
45%	100	96.1
35%	100	97.4
25%	100	**98.1**
15%	100	98.0

Table 7.4.
Impact of Mask Ratios on dental caries detection.

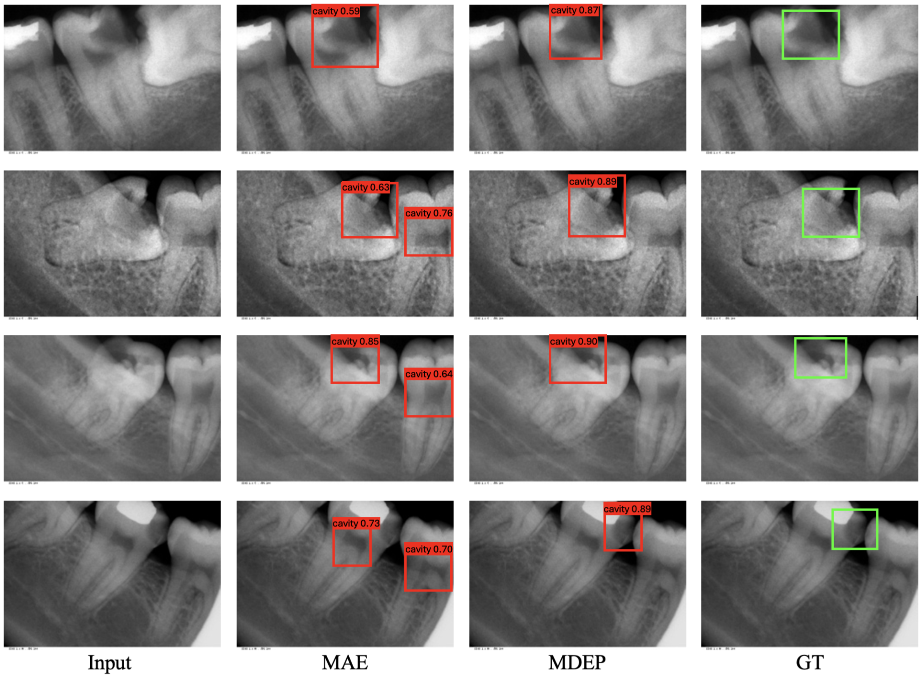


Figure 7.3. Comparison of dental caries detection on periapical radiographs dataset [159]. Bounding boxes (BB) are used to highlight detected caries. Red BBs indicate detected cavities with confidence scores, while green BBs indicate the ground truth cavities.

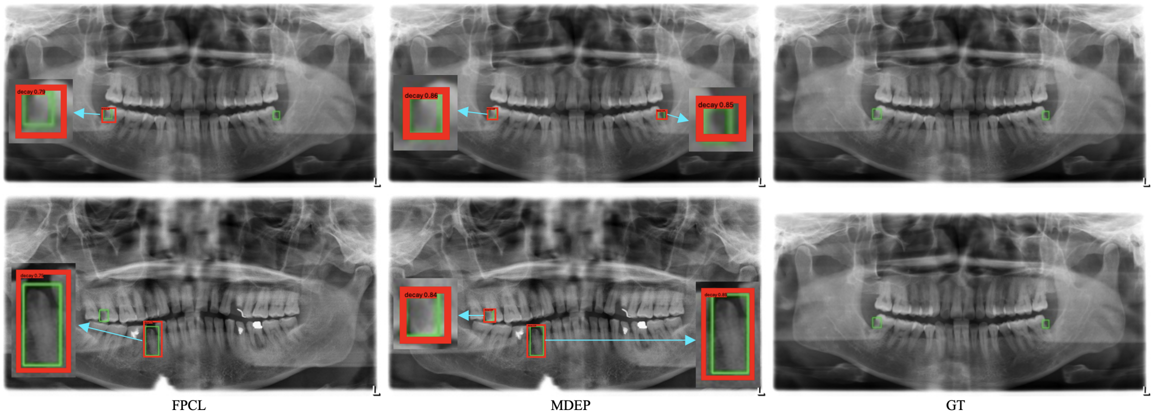


Figure 7.4. Comparison of dental caries detection on the CariesXrays dataset [158].

Reducing the mask ratio to 45% resulted in a noticeable improvement, with a precision of 96.1%. A further decrease to 35% produced a significant jump in precision to 97.4%.

The highest precision was achieved at a 25% mask ratio, with a precision of 98.1%, marking the peak performance among all tested configurations. Interestingly, a slight increase in mask ratio to 15% resulted in a marginally lower precision of 98.0%.

The findings indicate that lower mask ratios generally lead to higher precision in dental caries detection. The optimal mask ratio identified in this study is 25%, which produced the highest precision score. This suggests that a moderate amount of masking allows the model to learn more effectively from the available data, enhancing its detection capabilities.

Additionally, the results show that the number of pre-training epochs plays a less significant role in improving precision when compared to the choice of mask ratio. The model with 800 pre-training epochs at a 55% mask ratio did not perform better than the models with fewer epochs but lower mask ratios.

These insights highlight the importance of selecting an appropriate mask ratio to maximize the performance of dental caries detection models. Future work could explore even finer granularity in mask ratios and the potential benefits of combining different pre-training strategies to further enhance detection accuracy.

7.4.5 Implementation Details

Our experiments were conducted using the PyTorch framework [35] and trained on Nvidia Tesla V100 GPUs. We used a batch size of 748 for [159] and 4,800 for [158], corresponding to the total training samples. The AdamW optimizer [36] was employed for all experiments.

Data Augmentation: We applied various augmentation techniques, including noise addition up to 6% of pixels, horizontal and vertical flipping, and 90° rotation in both clockwise and counterclockwise directions.

MDEP Pre-training: The pre-training phase involved setting the base learning rate to 1.5e-4, weight decay to 0.05, β_1 to 0.9, and β_2 to 0.95. We utilized a cosine decay learning rate scheduler with a warm-up period of 10 epochs. A random Masked Image Modeling approach with a patch size of 16x16 and a mask ratio ranging from 15% to 65% was employed. Additionally, we utilized a linear prediction head targeting an image size of 400x300 for [159] and 800x600 for [158].

Task Fine-tuning: For downstream tasks, we employed single-scale training with a starting learning rate of 0.0001 and weight decay set at 0.05.

The implementation details remained consistent with those used for the original dataset [159]. The same data augmentation techniques, pre-training, and fine-tuning procedures were applied to ensure a fair comparison between the two datasets.

7.5 Conclusion

In this study, we introduced the Masked Deep Embedding of Patches (MDEP) pre-training approach, which significantly enhanced dental caries detection performance in dental X-ray analysis. Our findings demonstrate the efficacy of MDEP, particularly in scenarios with limited training data, by leveraging self-supervised learning to achieve superior results compared to traditional methods. Additionally, our experimentation with two different imaging modalities, periapical and panoramic X-rays, confirmed the robustness and adaptability of our method across different datasets and imaging conditions. The consistent outperformance of MDEP underscores its potential in practical applications where labeled data is scarce or expensive to obtain. Future research will focus on extending the application of MDEP to other medical imaging tasks, such as prognosis and outcome prediction, to further validate its effectiveness and broaden its impact in the field of dental diagnostics and beyond.

REFERENCES CITED

[1] S. C. White, E. W. Heslop, L. G. Hollender, K. M. Mosier, A. Ruprecht, M. K. Shrout *et al.*, "Parameters of radiologic care: An official report of the american academy of oral and maxillofacial radiology," *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, vol. 91, no. 5, pp. 498–511, 2001.

[2] V. Rushton and K. Horner, "The use of panoramic radiology in dental practice," *Journal of dentistry*, vol. 24, no. 3, pp. 185–201, 1996.

[3] R. Abdalla-Aslan, T. Yeshua, D. Kabla, I. Leichter, and C. Nadler, "An artificial intelligence system using machine-learning for automatic detection and classification of dental restorations in panoramic radiography," *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 130, no. 5, pp. 593–602, 2020.

[4] B. Molander, "Panoramic radiography in dental diagnostics." *Swedish Dental journal. Supplement*, vol. 119, pp. 1–26, 1996.

[5] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[6] Y. Zhao, P. Li, C. Gao, Y. Liu, Q. Chen, F. Yang, and D. Meng, "Tsasnet: Tooth segmentation on dental panoramic x-ray images by two-stage attention segmentation network," *Knowledge-Based Systems*, vol. 206, p. 106338, 2020.

[7] H. Chen, K. Zhang, P. Lyu, H. Li, L. Zhang, J. Wu, and C.-H. Lee, "A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[8] C. Kim, D. Kim, H. Jeong, S.-J. Yoon, and S. Youm, "Automatic tooth detection and numbering using a combination of a cnn and heuristic algorithm," *Applied Sciences*, vol. 10, no. 16, p. 5624, 2020.

[9] D. V. Tuzoff, L. N. Tuzova, M. M. Bornstein, A. S. Krasnov, M. A. Kharchenko, S. I. Nikolenko, M. M. Sveshnikov, and G. B. Bednenko, "Tooth detection and numbering in panoramic radiographs using convolutional neural networks," *Dentomaxillofacial Radiology*, vol. 48, no. 4, p. 20180051, 2019.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] K. Zhang, J. Wu, H. Chen, and P. Lyu, "An effective teeth recognition method using label tree with cascade network structure," *Computerized Medical Imaging and Graphics*, vol. 68, pp. 61–70, 2018.

[12] B. Silva, L. Pinheiro, L. Oliveira, and M. Pithon, "A study on tooth segmentation and numbering using end-to-end deep neural networks," in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2020, pp. 164–171.

[13] G. Jader, J. Fontineli, M. Ruiz, K. Abdalla, M. Pithon, and L. Oliveira, "Deep instance segmentation of teeth in panoramic x-ray images," in *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2018, pp. 400–407.

[14] J.-H. Lee, S.-S. Han, Y. H. Kim, C. Lee, and I. Kim, "Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs," *Oral surgery, oral medicine, oral pathology and oral radiology*, vol. 129, no. 6, pp. 635–642, 2020.

[15] A. F. Leite, A. V. Gerven, H. Willems, T. Beznik, P. Lahoud, H. Gaêta-Araujo, M. Vranckx, and R. Jacobs, "Artificial intelligence-driven novel tool for tooth detection and segmentation on panoramic radiographs," *Clinical oral investigations*, vol. 25, no. 4, pp. 2257–2267, 2021.

[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[18] ——, "Dental x-ray image segmentation using a u-shaped deep convolutional network," in *International Symposium on Biomedical Imaging*, vol. 1, 2015, pp. 1–13.

[19] T. L. Koch, M. Perslev, C. Igel, and S. S. Brandt, "Accurate segmentation of dental panoramic radiographs with u-nets," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 15–19.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[22] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9653–9663.

[23] X. Li, W. Wang, L. Yang, and J. Yang, "Uniform masking: Enabling mae pretraining for pyramid-based vision transformers with locality," *arXiv preprint arXiv:2205.10063*, 2022.

[24] H. Bao, L. Dong, S. Piao, and F. Wei, “Beit: Bert pre-training of image transformers,” *arXiv preprint arXiv:2106.08254*, 2021.

[25] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[26] A. El-Nouby, G. Izacard, H. Touvron, I. Laptev, H. Jegou, and E. Grave, “Are large-scale datasets necessary for self-supervised pre-training?” *arXiv preprint arXiv:2112.10740*, 2021.

[27] X. Chen, S. Xie, and K. He, “An empirical study of training self-supervised vision transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9640–9649.

[28] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[29] Z. Cai and N. Vasconcelos, “Cascade r-cnn: high quality object detection and instance segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 5, pp. 1483–1498, 2019.

[30] D. Smith, “The numbering of teeth,” *New Zealand School Dental Service gazette*, vol. 37, no. 4, p. 56, 1976.

[31] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[32] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, “Mmdetection: Open mmlab detection toolbox and benchmark,” *arXiv preprint arXiv:1906.07155*, 2019.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[34] J. Brooks, “COCO Annotator,” https://github.com/jsbroks/coco-annotator/, 2019.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.

[36] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *ICLR*, 2021.

[38] Y. Luo, Z. Chen, and X. Gao, "Self-distillation augmented masked autoencoders for histopathological image classification," *arXiv preprint arXiv:2203.16983*, 2022.

[39] A. Almalki and L. J. Latecki, "Self-supervised learning with masked image modeling for teeth numbering, detection of dental restorations, and instance segmentation in dental panoramic radiographs," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5594–5603.

[40] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *NIPS Deep Learning and Representation Learning Workshop*, 2014.

[41] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3713–3722.

[42] M. Zhao, L. Ma, W. Tan, and D. Nie, "Interactive tooth segmentation of dental models," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference.* IEEE, 2006, pp. 654–657.

[43] T. Yuan, W. Liao, N. Dai, X. Cheng, and Q. Yu, "Single-tooth modeling for 3d dental model," *International journal of biomedical imaging*, vol. 2010, 2010.

[44] K. Wu, L. Chen, J. Li, and Y. Zhou, "Tooth segmentation on dental meshes using morphologic skeleton," *Computers & Graphics*, vol. 38, pp. 199–211, 2014.

[45] X. Xu, C. Liu, and Y. Zheng, "3d tooth segmentation and labeling using deep convolutional neural networks," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 7, pp. 2336–2348, 2018.

[46] C. Lian, L. Wang, T.-H. Wu, M. Liu, F. Durán, C.-C. Ko, and D. Shen, "Meshsnet: Deep multi-scale mesh feature learning for end-to-end tooth labeling on 3d dental surfaces," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22.* Springer, 2019, pp. 837–845.

[47] C. Lian, L. Wang, T.-H. Wu, F. Wang, P.-T. Yap, C.-C. Ko, and D. Shen, "Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3d intraoral scanners," *IEEE transactions on medical imaging*, vol. 39, no. 7, pp. 2440–2450, 2020.

[48] Y. Zhao, L. Zhang, C. Yang, Y. Tan, Y. Liu, P. Li, T. Huang, and C. Gao, "3D dental model segmentation with graph attentional convolution network," *Pattern Rec. Letters*, vol. 152, 2021.

[49] L. Zhang, Y. Zhao, D. Meng, Z. Cui, C. Gao, X. Gao, C. Lian, and D. Shen, "Tsgcnet: Discriminative geometric feature learning with two-stream graph convolutional network for 3d dental model segmentation," in *CVPR*, 2021.

[50] Z. Li, T. Liu, J. Wang, C. Zhang, and X. Jia, "Multi-scale bidirectional enhancement network for 3d dental model segmentation," in *IEEE 19th Int. Symposium on Biomedical Imaging (ISBI).* IEEE, 2022, pp. 1–5.

[51] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[52] A. Almalki and L. J. Latecki, "Enhanced masked image modeling for analysis of dental panoramic radiographs," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023.

[53] A. Ben-Hamadou, O. Smaoui, H. Chaabouni-Chouayakh, A. Rekik, S. Pujades, E. Boyer, J. Strippoli, A. Thollot, H. Setbon, C. Trosset *et al.*, "Teeth3ds: a benchmark for teeth segmentation and labeling from intra-oral 3d scans," *arXiv preprint arXiv:2210.06094*, 2022.

[54] Y. Liang, S. Zhao, B. Yu, J. Zhang, and F. He, "Meshmae: Masked autoencoders for 3d mesh data analysis," in *ECCV*, 2022.

[55] T. Yuan, W. Liao, N. Dai, X. Cheng, and Q. Yu, "Single-tooth modeling for 3d dental model," *International journal of biomedical imaging*, vol. 2010, 01 2010.

[56] T. Kronfeld, D. Brunner, and G. Brunnett, "Snake-based segmentation of teeth from virtual dental casts," *Computer-Aided Design and Applications*, vol. 7, no. 2, pp. 221–233, 2010.

[57] C. Sinthanayothin and W. Tharanont, "Orthodontics treatment simulation by teeth segmentation and setup," in *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTICON'08)*, vol. 1. IEEE, 2008, pp. 81–84.

[58] M. Yaqi and L. Zhongke, "Computer aided orthodontics treatment by virtual segmentation and adjustment," in *2010 International Conference on Image Analysis and Signal Processing*. IEEE, 2010, pp. 336–339.

[59] B.-j. Zou, S.-j. Liu, S.-h. Liao, X. Ding, and Y. Liang, "Interactive tooth partition of dental mesh base on tooth-target harmonic field," *Computers in biology and medicine*, vol. 56, pp. 132–144, 2015.

[60] S.-h. Liao, S.-j. Liu, B.-j. Zou, X. Ding, Y. Liang, and J.-h. Huang, "Automatic tooth segmentation of dental mesh based on harmonic fields," *BioMed research international*, vol. 2015, 2015.

[61] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.

[62] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 2285–2294.

[63] Z. Cui, C. Li, and W. Wang, "Toothnet: Automatic tooth instance segmentation and identification from cone beam ct images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6368–6377.

[64] Y. Miki, C. Muramatsu, T. Hayashi, X. Zhou, T. Hara, A. Katsumata, and H. Fujita, "Classification of teeth in cone-beam ct using deep convolutional neural network," *Computers in biology and medicine*, vol. 80, pp. 24–29, 2017.

[65] Y. Rao, Y. Wang, F. Meng, J. Pu, J. Sun, and Q. Wang, "A symmetric fully convolutional residual network with dcrf for accurate tooth segmentation," *IEEE Access*, vol. 8, pp. 92 028–92 038, 2020.

[66] J. Zhang, C. Li, Q. Song, L. Gao, and Y.-K. Lai, "Automatic 3D Tooth Segmentation using Convolutional Neural Networks in Harmonic Parameter Space," *Elsevier Graphical Models*, vol. 39, p. 101071, 2020.

[67] D. Sun, Y. Pei, G. Song, Y. Guo, G. Ma, T. Xu, and H. Zha, "Tooth segmentation and labeling from digital dental casts," in *IEEE International Symposium on Biomedical Imaging (ISBI'20)*, April, Iowa City, IA, USA 2020.

[68] X. Xu, C. Liu, and Y. Zheng, "3d tooth segmentation and labeling using deep convolutional neural networks," *IEEE transactions on visualization and computer graphics*, vol. 25, pp. 2336–2348, 2018.

[69] F. G. Zanjani, D. A. Moin, B. Verheij, F. Claessen, T. Cherici, T. Tan *et al.*, "Deep learning approach to semantic segmentation in 3d point cloud intra-oral scans of teeth," in *International Conference on Medical Imaging with Deep Learning.* PMLR, 2019, pp. 557–571.

[70] S. Tian, N. Dai, B. Zhang, F. Yuan, Q. Yu, and X. Cheng, "Automatic classification and segmentation of teeth on 3d dental model using hierarchical deep learning networks," *IEEE Access*, vol. 7, pp. 84 817–84 828, 2019.

[71] Z. Cui, C. Li, N. Chen, G. Wei, R. Chen, Y. Zhou, and W. Wang, "Tsegnet: an efficient and accurate tooth segmentation network on 3d dental model," *Medical Image Analysis*, vol. 69, p. 101949, 2020.

[72] F. G. Zanjani, D. A. Moin, F. Claessen, T. Cherici, S. Parinussa, A. Pourtaherian, and S. Zinger, "Mask-mcnet: Instance segmentation in 3d point cloud of intra-oral scans," pp. 128–136, 2019.

[73] Q. Ma, G. Wei, Y. Zhou, X. Pan, S. Xin, and W. Wang, "Srf-net: Spatial relationship feature network for tooth point cloud classification," in *Computer Graphics Forum*, vol. 39, no. 7. Wiley Online Library, 2020, pp. 267–277.

[74] Y. Zhao, L. Zhang, C. Yang, Y. Tan, Y. Liu, P. Li, T. Huang, and C. Gao, "3d dental model segmentation with graph attentional convolution network," *Pattern Rec. Letters*, vol. 152, 2021.

[75] Y. Zhao, L. Zhang, Y. Liu, D. Meng, Z. Cui, C. Gao, X. Gao, C. Lian, and D. Shen, "Two-stream graph convolutional network for intra-oral scanner image segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, 2022.

[76] L. Qiu, C. Ye, P. Chen, Y. Liu, X. Han, and S. Cui, "Darch: Dental arch prior-assisted 3d tooth instance segmentation with weak annotations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 752–20 761.

[77] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[78] S.-M. Hu, Z.-N. Liu, M.-H. Guo, J.-X. Cai, J. Huang, T.-J. Mu, and R. R. Martin, "Subdivision-based mesh convolution networks," *ACM Transactions on Graphics (TOG)*, 2021.

[79] A. W. Lee, W. Sweldens, P. Schröder, L. Cowsar, and D. Dobkin, "Maps: Multiresolution adaptive parameterization of surfaces," in *ACM SIGGRAPH*, 1998, pp. 95–104.

[80] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 605–613.

[81] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[82] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[83] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *Acm Transactions On Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[84] S. Qiu, S. Anwar, and N. Barnes, "Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1757–1767.

[85] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu, "Rethinking network design and local geometry in point cloud: A simple residual mlp framework," *arXiv preprint arXiv:2202.07123*, 2022.

[86] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "Pct: Point cloud transformer," *Computational Visual Media*, vol. 7, no. 2, pp. 187–199, 2021.

[87] T. Xiang, C. Zhang, Y. Song, J. Yu, and W. Cai, "Walk in the cloud: Learning curves for point clouds shape analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 915–924.

[88] Y. Pang, W. Wang, F. E. Tay, W. Liu, Y. Tian, and L. Yuan, "Masked autoencoders for point cloud self-supervised learning," in *ECCV*, 2022.

[89] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-bert: Pre-training 3d point cloud transformers with masked point modeling," in *CVPR*, 2022.

[90] L. Lechuga *et al.*, "Cone beam ct vs. fan beam ct: a comparison of image quality and dose delivered between two differing ct imaging modalities," *Cureus*, vol. 8, no. 9, 2016.

[91] M. Hosntalab *et al.*, "Segmentation of teeth in ct volumetric dataset by panoramic projection and variational level set," *International Journal of Computer Assisted Radiology and Surgery*, vol. 3, pp. 257–265, 2008.

[92] D. X. Ji *et al.*, "A level-set based approach for anterior teeth segmentation in cone beam computed tomography images," *Computers in biology and medicine*, vol. 50, pp. 116–128, 2014.

[93] Y. Gan *et al.*, "Tooth and alveolar bone segmentation from dental computed tomography images," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 196–204, 2017.

[94] H. Gao *et al.*, "Individual tooth segmentation from ct images using level set method with shape and intensity prior," *Pattern Recognition*, vol. 43, no. 7, pp. 2406–2417, 2010.

[95] S. Barone *et al.*, "Ct segmentation of dental shapes by anatomy-driven reformation imaging and b-spline modelling," *International journal for numerical methods in biomedical engineering*, vol. 32, no. 6, p. e02747, 2016.

[96] Q. Yu *et al.*, "Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation," in *CVPR*, 2018, pp. 8280–8289.

[97] Z. Zhang *et al.*, "Mdnet: A semantically and visually interpretable medical image diagnosis network," in *CVPR*, 2017, pp. 6428–6436.

[98] ——, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *CVPR*, 2018, pp. 9242–9251.

[99] Z. Cui, B. Zhang, C. Lian, C. Li, L. Yang, W. Wang, M. Zhu, and D. Shen, "Hierarchical morphology-guided tooth instance segmentation from cbct images," in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 150–162.

[100] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[101] MONAI Consortium, "MONAI: Medical Open Network for AI," 3 2020. [Online]. Available: https://github.com/Project-MONAI/MONAI

[102] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[103] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[104] Z. Cai *et al.*, "Dstunet: Unet with efficient dense swin transformer pathway for medical image segmentation," in *ISBI*, 2022.

[105] H.-Y. Zhou *et al.*, "nnformer: Interleaved transformer for volumetric segmentation," *arXiv preprint arXiv:2109.03201*, 2021.

[106] F. Isensee *et al.*, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, 2021.

[107] L. Shulman and T. Driskell, "Dental implants: a historical perspective," *Implants in dentistry. Philadelphia: WB Saunders*, p. 6, 1997.

[108] G. Boven, G. Raghoebar, A. Vissink, and H. Meijer, "Improving masticatory performance, bite force, nutritional state and patient's satisfaction with implant overdentures: a systematic review of the literature," *Journal of oral rehabilitation*, vol. 42, no. 3, pp. 220–233, 2015.

[109] Y. Kanehira, K. Arai, T. Kanehira, K. Nagahisa, and S. Baba, "Oral health-related quality of life in patients with implant treatment," *The Journal of Advanced Prosthodontics*, vol. 9, no. 6, pp. 476–481, 2017.

[110] I. J. De Kok, I. S. Duqum, L. H. Katz, and L. F. Cooper, "Management of implant/prosthodontic complications," *Dental Clinics*, vol. 63, no. 2, pp. 217–231, 2019.

[111] D. Hashim, N. Cionca, C. Combescure, and A. Mombelli, "The diagnosis of peri-implantitis: A systematic review on the predictive value of bleeding on probing," *Clinical oral implants research*, vol. 29, pp. 276–293, 2018.

[112] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism*, vol. 69, pp. S36–S40, 2017.

[113] J.-H. Lee, D.-H. Kim, S.-N. Jeong, and S.-H. Choi, "Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm," *Journal of dentistry*, vol. 77, pp. 106–111, 2018.

[114] S. Yamaguchi, C. Lee, O. Karaer, S. Ban, A. Mine, and S. Imazato, "Predicting the debonding of cad/cam composite resin crowns with ai," *Journal of Dental Research*, vol. 98, no. 11, pp. 1234–1238, 2019.

[115] T. Takahashi, K. Nozaki, T. Gonda, and K. Ikebe, "A system for designing removable partial dentures using artificial intelligence. part 1. classification of partially edentulous arches using a convolutional neural network," *Journal of prosthodontic research*, vol. 65, no. 1, pp. 115–118, 2021.

[116] T. Joda, T. Waltimo, N. Probst-Hensch, C. Pauli-Magnus, and N. U. Zitzmann, "Health data in dentistry: an attempt to master the digital challenge," *Public Health Genomics*, vol. 22, no. 1-2, pp. 1–7, 2019.

[117] I. E. Hamamci, S. Er, E. Simsar, A. Sekuboyina, M. Gundogar, B. Stadlinger, A. Mehl, and B. Menze, "Diffusion-based hierarchical multi-label object detection to analyze panoramic dental x-rays," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Springer Nature Switzerland, 2023, pp. 389–399.

[118] J. Shi, B. Sun, X. Ye, Z. Wang, X. Luo, J. Liu, H. Gao, and H. Li, "Semantic decomposition network with contrastive and structural constraints for dental plaque segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 4, pp. 935–946, 2022.

[119] Z. Liu, X. He, H. Wang, H. Xiong, Y. Zhang, G. Wang, J. Hao, Y. Feng, F. Zhu, and H. Hu, "Hierarchical self-supervised learning for 3d tooth segmentation in intra-oral mesh scans," *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 467–480, 2022.

[120] G. Chen, J. Qin, B. B. Amor, W. Zhou, H. Dai, T. Zhou, H. Huang, and L. Shao, "Automatic detection of tooth-gingiva trim lines on dental surfaces," *IEEE Transactions on Medical Imaging*, 2023.

[121] J.-J. Hwang, Y.-H. Jung, B.-H. Cho, and M.-S. Heo, "An overview of deep learning in the field of dentistry," *Imaging science in dentistry*, vol. 49, no. 1, pp. 1–7, 2019.

[122] B. Bozorgtabar, D. Mahapatra, and J.-P. Thiran, "Amae: Adaptation of pre-trained masked autoencoder for dual-distribution anomaly detection in chest x-rays," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Springer Nature Switzerland, 2023, pp. 195–205.

[123] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, and P. Prasanna, "Self pre-training with masked autoencoders for medical image classification and segmentation," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–6.

[124] A. Almalki and L. J. Latecki, "Self-supervised learning with masked autoencoders for teeth segmentation from intra-oral 3d scans," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7820–7830.

[125] al, "implants dataset," https://universe.roboflow.com/al-xfpta/implants-meckw, aug 2023, visited on 2024-02-20. [Online]. Available: https://universe.roboflow.com/al-xfpta/implants-meckw

[126] W. Yao and L. Li, "A new regression model: modal linear regression," *Scandinavian Journal of Statistics*, vol. 41, no. 3, pp. 656–671, 2014.

[127] C. Habermann and F. Kindermann, "Multidimensional spline interpolation: Theory and applications," *Computational Economics*, vol. 30, pp. 153–169, 2007.

[128] B. R. Siqueira, F. C. Ferrari, K. E. Souza, V. V. Camargo, and R. de Lemos, "Testing of adaptive and context-aware systems: approaches and challenges," *Software Testing, Verification and Reliability*, vol. 31, no. 7, p. e1772, 2021.

[129] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[130] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *IEEE/CVF CVPR*, 2023.

[131] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[132] Ultralytics, "YOLOv5: A state-of-the-art real-time object detection system," https://docs.ultralytics.com, 2021, visited on 2024-02-20.

[133] R. Adell, U. Lekholm, B. Rockler, and P.-I. Brånemark, "A 15-year study of osseointegrated implants in the treatment of the edentulous jaw," *International journal of oral surgery*, vol. 10, no. 6, pp. 387–416, 1981.

[134] D. Buser, R. Mericske-stern, J. P. Pierre Bernard, A. Behneke, N. Behneke, H. P. Hirt, U. C. Belser, and N. P. Lang, "Long-term evaluation of non-submerged iti implants. part 1: 8-year life table analysis of a prospective multi-center study with 2359 implants." *Clinical oral implants research*, vol. 8, no. 3, pp. 161–172, 1997.

[135] A. Mombelli and N. P. Lang, "The diagnosis and treatment of peri-implantitis," *Periodontology 2000*, vol. 17, no. 1, pp. 63–76, 1998.

[136] S. Renvert, G. R. Persson, F. Q. Pirih, and P. M. Camargo, "Peri-implant health, peri-implant mucositis, and peri-implantitis: Case definitions and diagnostic considerations," *Journal of clinical periodontology*, vol. 45, pp. S278–S285, 2018.

[137] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[138] T. Shan, F. Tay, and L. Gu, "Application of artificial intelligence in dentistry," *Journal of dental research*, vol. 100, no. 3, pp. 232–244, 2021.

[139] A. G. Cantu, S. Gehrung, J. Krois, A. Chaurasia, J. G. Rossi, R. Gaudin, K. Elhennawy, and F. Schwendicke, "Detecting caries lesions of different radiographic extension on bitewings using deep learning," *Journal of dentistry*, vol. 100, p. 103425, 2020.

[140] A. A. Novikov *et al.*, "Fully convolutional architectures for multiclass segmentation in chest radiographs," *IEEE transactions on medical imaging*, vol. 37, no. 8, pp. 1865–1876, 2018.

[141] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[142] T. Joda, M. M. Bornstein, R. E. Jung, M. Ferrari, T. Waltimo, and N. U. Zitzmann, "Recent trends and future direction of dental research in the digital era," *International journal of environmental research and public health*, vol. 17, no. 6, p. 1987, 2020.

[143] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[144] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13.* Springer, 2014, pp. 740–755.

[145] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3677–3685.

[146] G.-H. Lin, Y. Kapila, and H.-L. Wang, "Parameters to define peri-implantitis: A review and a proposed multi-domain scale," *Journal of Oral Implantology*, vol. 43, no. 6, pp. 491–496, 2017.

[147] A. M. Decker, R. Sheridan, G.-H. Lin, P. Sutthiboonyapan, W. Carroll, and H.-L. Wang, "A prognosis system for periimplant diseases," *Implant dentistry*, vol. 24, no. 4, pp. 416–421, 2015.

[148] S. J. Froum and P. S. Rosen, "A proposed classification for peri-implantitis," *International Journal of Periodontics and Restorative Dentistry*, vol. 32, no. 5, p. 533, 2012.

[149] A. Aboah, B. Wang, U. Bagci, and Y. Adu-Gyamfi, "Real-time multi-class helmet violation detection using few-shot data sampling technique and yolov8," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5349–5357.

[150] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[151] N. B. Pitts, D. T. Zero, P. D. Marsh, K. Ekstrand, J. A. Weintraub, F. Ramos-Gomez, J. Tagami, S. Twetman, G. Tsakos, and A. Ismail, "Dental caries," *Nature reviews Disease primers*, vol. 3, no. 1, pp. 1–16, 2017.

[152] P. Wen, M. Chen, Y. Zhong, Q. Dong, and H. Wong, "Global burden and inequality of dental caries, 1990 to 2019," *Journal of dental research*, vol. 101, no. 4, pp. 392–399, 2022.

[153] R. H. Selwitz, A. I. Ismail, and N. B. Pitts, "Dental caries," *The Lancet*, vol. 369, no. 9555, pp. 51–59, 2007.

[154] N. Kassebaum, E. Bernabé, M. Dahiya, B. Bhandari, C. Murray, and W. Marcenes, "Global burden of untreated caries: a systematic review and metaregression," *Journal of dental research*, vol. 94, no. 5, pp. 650–658, 2015.

[155] J. D. Bader, D. A. Shugars, and A. J. Bonito, "A systematic review of the performance of methods for identifying carious lesions," *Journal of public health dentistry*, vol. 62, no. 4, pp. 201–213, 2002.

[156] A. Wenzel, "Digital radiography and caries diagnosis," *Dentomaxillofacial Radiology*, vol. 27, no. 1, pp. 3–11, 1998.

[157] I. A. Pretty, "Caries detection and diagnosis: novel technologies," *Journal of dentistry*, vol. 34, no. 10, pp. 727–739, 2006.

[158] B. Chen, S. Fu, Y. Liu, J. Pan, G. Lu, and Z. Zhang, "Cariesxrays: Enhancing caries detection in hospital-scale panoramic dental x-rays via feature pyramid contrastive learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 20, 2024, pp. 21 940–21 948.

[159] U. Rashid, A. Javid, A. R. Khan, L. Liu, A. Ahmed, O. Khalid, K. Saleem, S. Meraj, U. Iqbal, and R. Nawaz, "A hybrid mask rcnn-based tool to localize dental cavities from real-time mixed photographic images," *PeerJ Computer Science*, vol. 8, p. e888, 2022.

[160] C. Serban, D. Lungeanu, S.-D. Bota, C. C. Cotca, M. L. Negrutiu, V.-F. Duma, C. Sinescu, and E. L. Craciunescu, "Emerging technologies for dentin caries detection—a systematic review and meta-analysis," *Journal of Clinical Medicine*, vol. 11, no. 3, p. 674, 2022.

[161] M. Prados-Privado, J. García Villalón, C. H. Martínez-Martínez, C. Ivorra, and J. C. Prados-Frutos, "Dental caries diagnosis and detection using neural networks: a systematic review," *Journal of clinical medicine*, vol. 9, no. 11, p. 3579, 2020.

[162] M. Ezhov, M. Gusarev, M. Golitsyna, J. M. Yates, E. Kushnerev, D. Tamimi, S. Aksoy, E. Shumilov, A. Sanders, and K. Orhan, "Clinically applicable artificial intelligence system for dental diagnosis with cbct," *Scientific reports*, vol. 11, no. 1, p. 15006, 2021.

[163] A. Kumar, H. S. Bhadauria, and A. Singh, "Descriptive analysis of dental x-ray images using various practical methods: A review," *PeerJ Computer Science*, vol. 7, p. e620, 2021.

[164] Y. Liang, W. Song, J. Yang, L. Qiu, K. Wang, and L. He, "X2teeth: 3d teeth reconstruction from a single panoramic radiograph," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*. Springer, 2020, pp. 400–409.

[165] M. G. Endres, F. Hillen, M. Salloumis, A. R. Sedaghat, S. M. Niehues, O. Quatela, H. Hanken, R. Smeets, B. Beck-Broichsitter, C. Rendenbach *et al.*, "Development of a deep learning algorithm for periapical disease detection in dental radiographs," *Diagnostics*, vol. 10, no. 6, p. 430, 2020.

[166] A. E. Rad, M. S. M. Rahim, H. Kolivand, and A. Norouzi, "Automatic computer-aided caries detection from dental x-ray images using intelligent level set," *Multimedia Tools and Applications*, vol. 77, no. 21, pp. 28 843–28 862, 2018.

[167] Y.-w. Chen, K. Stanley, W. Att *et al.*, "Artificial intelligence in dentistry: current applications and future perspectives," *Quintessence Int*, vol. 51, no. 3, pp. 248–57, 2020.

[168] K.-C. T. Nguyen, B. M. Le, M. Li, F. T. Almeida, P. W. Major, N. R. Kaipatur, E. H. Lou, K. Punithakumar, and L. H. Le, "Localization of cementoenamel junction in intraoral ultrasonographs with machine learning," *Journal of Dentistry*, vol. 112, p. 103752, 2021.

[169] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[170] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[171] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[172] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[173] G. E. Hinton and R. Zemel, "Autoencoders, minimum description length and helmholtz free energy," *Advances in neural information processing systems*, vol. 6, 1993.

[174] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[175] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *Journal of machine learning research*, vol. 11, no. 12, 2010.

[176] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016, pp. 2536–2544.

[177] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *International conference on machine learning.* PMLR, 2020, pp. 1691–1703.

[178] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14.* Springer, 2016, pp. 649–666.

[179] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[180] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning.* PMLR, 2021, pp. 8821–8831.

[181] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.

[182] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2794–2802.

[183] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision.* Springer, 2016, pp. 69–84.

[184] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan, "Learning features by watching objects move," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2701–2710.

[185] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[186] S. Becker and G. E. Hinton, "Self-organizing neural network that discovers surfaces in random-dot stereograms," *Nature*, vol. 355, no. 6356, pp. 161–163, 1992.

[187] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[188] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *CVPR*, 2018, pp. 3733–3742.

[189] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[190] K. He *et al.*, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020.

[191] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning.* PMLR, 2020, pp. 1597–1607.

[192] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *arXiv preprint arXiv:2006.07733*, 2020.

[193] X. Chen and K. He, "Exploring simple siamese representation learning," in *CVPR*, 2021, pp. 15 750–15 758.

[194] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[195] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[196] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[197] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[198] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[199] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7464–7475.

[200] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[201] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[202] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: A simple and strong anchor-free object detector," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 1922–1933, 2020.

[203] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang, "Conditional detr for fast training convergence," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3651–3660.