A Two Stage Pipeline for Automated Caries Detection on Single Tooth Images from Panoramic Radiographs

Christopher J. Hansen[1][0009−0001−8978−036X], Paula Kloehn[2], Anna-Louisa Kollster[2], Toni Gehrmann[2][0009−0007−3543−7420], Jonas Conrad[2][0000−0003−2516−3351], Niklas Christoph Koser[1][0009−0006−0599−2901], Christian Graetz[2][0000−0002−8316−0565], Christof Dörfer[2], Claus-C. Glüer[1], Jan-Bernd Hövener[1][0000−0001−7255−7252], and Coenraad Mouton[1][0000−0001−8610−2478]

[1] Section Biomedical Imaging, Dept. of Radiology and Neuroradiology, University Medical Center Schleswig-Holstein (UKSH), Campus Kiel
[2] Clinic of Conservative Dentistry and Periodontology, University of Kiel, Kiel, Germany

Abstract. Panoramic dental radiographs (OPG) are the only imaging modality that captures the entire dentition in a single exposure. To support dentists with diagnosing caries it is essential to find indications for cavities on those images. While recent deep learning methods show strong results on in-distribution test sets, the generalization on out-of-distribution datasets is mostly untested. In this study, we suggest a two-stage deep learning pipeline for caries detection on single-tooth images extracted from OPGs: (1) image-level classification using a DINO-based transformer backbone and (2) instance-level segmentation using Mask-R-CNN. We perform experiments on data from the University Medical Center Schleswig-Holstein (UKSH). To study generalization, we test models on an out-of-distribution set from the Federal University of Bahia (UFBA) and also evaluate a mixed-domain setting including both UKSH and UFBA data. Further we investigate the influence of strong augmentation techniques. Results show that classification performance is high on in-distribution data but significantly drops when applied to out-of-distribution samples. Segmentation performance is moderate across all settings, with limited robustness under domain shift. These findings suggest that in-distribution results overestimate real-world performance and underscore the importance of evaluating domain shifts in dental AI pipelines.

Keywords: caries · segmentation · classification

1 Introduction

Dental caries is a progressive demineralization of tooth structure caused by bacterial activity, leading to the formation of cavities and potential tooth loss if left

untreated and is usually visible as darker spots on dental radiographs. These lesions represents the most widespread oral health condition globally, affecting approximately 2.3 billion individuals each year [15], and remains one of the most prevalent diseases worldwide. Panoramic dental radiographs (orthopantomograms, OPGs) are among the most commonly used imaging modalities in dentistry and represent the only routine radiographic technique that captures the entire dentition in a single image. In some countries, clinicians are legally obligated to perform a comprehensive assessment of each OPG when acquired, regardless of its primary indication. Although bitewing radiographs are more commonly used for caries detection due to their higher image resolution, these are typically taken only when carious lesions are already suspected. In contrast, OPGs are often acquired during routine examinations, making them a promising modality for opportunistic screening of dental caries. An AI-based system that highlights potential carious regions in OPGs could support dentists in systematically reviewing the full dentition and reduce the likelihood of overlooked lesions.

Prior work has reported strong performance in automated carious tooth classification [3, 21, 5, 16] and in segmentation of carious lesions on full panoramic radiographs [2, 8, 25, 14, 7, 1, 19], however to the best of our knowledge, no existing study integrates both tasks into a unified pipeline that performs classification and subsequent segmentation at the tooth level. Moreover, none of the models achieving high accuracy in previous studies have been evaluated on out-of-distribution test sets. This makes it unclear as to whether these models are generally applicable tools, or only solutions that perform well in a highly specific setting (e.g. a specific dental clinic). Finally, most of the datasets used in prior work are not publicly available (especially the annotations of the data), which limits reproducibility and makes comparison between different methods difficult (if not impossible).

In this study, we propose a novel two-stage pipeline for diagnosing dental caries at the tooth level. In the first stage, we employ a DINO-based transformer model to classify individual teeth as either carious or healthy. In the second stage, we perform lesion segmentation using a Mask-R-CNN exclusively on teeth identified as carious by the classification model. Given that one would expect a classification model to yield higher classification performance than a segmentation model, this approach helps reduce false-positive segmentations caused by image artifacts that may resemble carious regions. To further enhance clinical plausibility, we also incorporate an a-priori exclusion of clinically implausible regions after segmentation and before the final output. Additionally, we propose an intensive data augmentation pipeline and compare the performance of models trained with and without these augmentations applied. Finally, we investigate performance under different conditions and evaluate on both in-distribution and out-of-distribution test sets.

The layout is as follows. We start with an overview of related work in Section 2, before describing our methods in Section 3. The results are then presented in Section 4, before we proceed with a discussion in Section 5.

2 Related Work

Deep learning has shown strong potential in the automated detection and segmentation of carious lesions from dental radiographs. Existing approaches typically address either classification of carious teeth or segmentation of carious regions, most commonly using intraoral images such as bitewing or periapical radiographs [18, 22, 2, 5]. While these high-resolution modalities are well-suited for visualizing occlusal and interproximal lesions, they are usually acquired only when caries is suspected, limiting their utility for screening.

Panoramic radiographs, in contrast, are standard in general dental practice and capture the full dentition in a single image. Several studies have leveraged OPGs for classification-based detection of caries, using various CNN backbones or ensemble strategies. Bui et al. [3] proposed a fusion-based CNN approach for caries classification. PaxNet [8] combines transfer learning and capsule networks to detect dental caries. Lian et al. [16] and Vinayahalingam et al. [24] also presented classification frameworks that operate at the tooth level. Grad-CAM has been applied for lesion visualization in such classification settings [21], though its interpretability remains limited by its indirect activation-based localization [23].

Segmentation of carious lesions on full OPGs remains less explored due to challenges like low contrast, anatomical variation, and the fine-grained nature of lesions. Ying et al. [25] used a standard U-Net for semantic segmentation of caries on tooth-level images. Alharbi et al. [1] applied a nested U-Net architecture (U-Net++) to panoramic images. Dayı et al. [7] proposed a multi-stage deep learning pipeline incorporating detection and segmentation modules. Zhu et al. introduced CariesNet [27], a multi-branch architecture tailored for multi-stage caries lesion segmentation. Chen et al. [6] combined classification and severity grading using a deep segmentation network. Kawazu et al. [13] used a domain-specific transfer learning approach, but focused more on feasibility than generalization. In addition, Khan et al. [14] evaluated deep segmentation for various dental pathologies on periapical radiographs.

However, most segmentation studies report only in-domain performance and do not explicitly evaluate cross-institutional or cross-dataset generalization. In many cases, the annotations are made on a per-image basis, rather than at the tooth instance level, which is more relevant for clinical workflows. Furthermore strong augmentation techniques are also widely unexplored.

3 Methods

In this section we describe our method for automated caries classification and lesion segmentation. The high-level workflow is shown in Figure 1; we elaborate on each individual step in the following sections. Firstly, in Section 3.1, we describe the two datasets we use and the annotation process. Thereafter, in Section 3.2, we expand on the process of generating single tooth images. This is then followed by a description of our augmentation pipeline in Section 3.3. The

classification and segmentation models are then explained in Section 3.4 and Section 3.5, respectively. Finally, we describe the filtering with dental specific a-priori knowledge we use in Section 3.6.

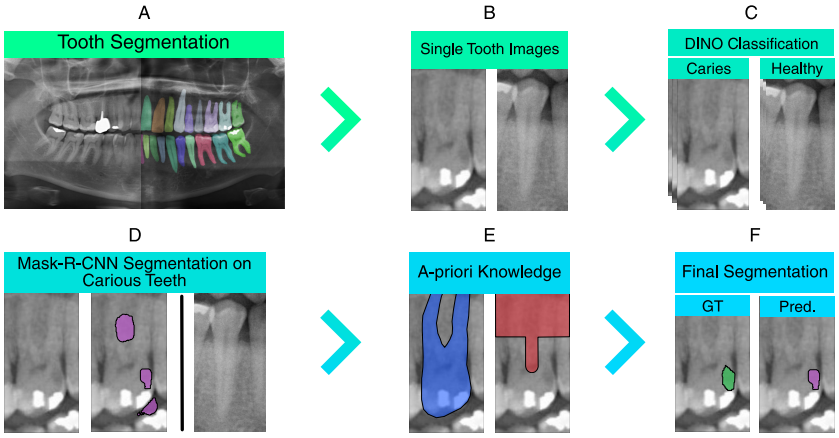

Fig. 1. Overview of the proposed pipeline from full panoramic radiographs to final single-tooth caries segmentation. (A) Tooth segmentations are extracted from panoramic x-rays using the method proposed in [9]. (B) Individual tooth regions are stored as separate images. (C) A DINO-based classification model is trained on the extracted single-tooth images to identify carious cases. (D) A Mask R-CNN segmentation model, trained on the same dataset, is applied selectively to images classified as carious. (E) Post-processing incorporates a-priori knowledge from dentists to remove clinically implausible regions. (F) The final segmentation prediction is compared against the ground truth (GT) annotations.

3.1 Data

We collected unannotated panoramic dental radiographs from two institutions: the University Medical Center Schleswig-Holstein (UKSH) and the Federal University of Bahia (UFBA). Both collections consist of full panoramic dental x-rays (OPGs). The UKSH dataset is private, while the UFBA dataset can be requested from the corresponding authors in [12]. All images across both datasets were then reviewed and annotated by experienced dentists affiliated with the UKSH.

The annotation process followed a multi-stage protocol to ensure high-quality and consistent labels. Initially, each radiograph was independently reviewed by four dentists. This was followed by a second round of individual re-evaluation to address intra-rater variability. Finally, a consensus review was conducted in which two dentists jointly assessed and reconciled all annotations across the dataset, including their own, to resolve any discrepancies. This process was employed to mitigate the inherent ambiguity in diagnosing caries from OPG images,

which are not typically used as the primary modality for caries detection due to limited contrast and overlapping structures. Nevertheless, OPGs remain among the most commonly available radiographs in clinical workflows, making them a practical target for decision-support applications.

Table 1. Composition the annotated datasets from UKSH and UFBA. 'Total images' correspond to the number of single tooth images.

Dataset	Total Images	Carious	Healthy
UKSH	1455	658	797
UFBA	460	226	234
Mixed	1915	884	1031

The number of single tooth images and the Carious/Healthy distribution for these datasets, as well as their combined numbers, are shown in Table 1. We use these three datasets in three different training/evaluation configurations to assess the performance of our classification and segmentation models:

- **Test on UKSH, Train on UKSH**: We train on a subset of the UKSH data and also evaluate the performance on a different, held out subset. This serves as an in-distribution evaluation for data gathered from a single centre.
- **Test on UFBA, Train on UKSH**: We evaluate the performance of the model(s) trained on the UKSH dataset above on the entire UFBA dataset. This serves as a test of out-of-distribution performance, where the distribution shift is characterized by the differences between centers.
- **Test on Mixed, Train on Mixed**: In this setting, we combine the two datasets and generate a train and a test subset. While this can also be considered a test of in-distribution performance, its main purpose is to indicate whether mixing training data from different centers is a viable approach.

For both the UKSH and combined datasets, we split the dataset such that 85% is used for training and validation purposes while the remaining 15% serves as the in-domain test set. The training data is further split using a five-fold cross validation strategy, where 80% is used for training and 20% for validation in each fold. Splits were stratified by both patient ID and tooth number. This stratification was essential not only to maintain a balanced distribution of carious and healthy samples but also to ensure an even representation of individual tooth types across the splits, given that caries prevalence varies significantly between different teeth. Stratifying by patient further ensured that no data leakage occurred between training, validation, and test sets.

3.2 Single Tooth Images

To extract single-tooth images, we use the pipeline proposed by Hansen et al. [9], which performs tooth instance segmentation on panoramic radiographs and

assigns tooth identification labels (11-18, 21-28, 31-38, 41-48) to each tooth according to the World Dental Federation (FDI) standard. The pipeline then generates individual tooth-level image crops based on the corresponding segmentation masks and the respective bounding boxes (see Figure 1 (A) and (B)). Note that this segmentation pipeline only identifies and crops the individual teeth - it does not identify which are healthy or carious.

3.3 Augmentations

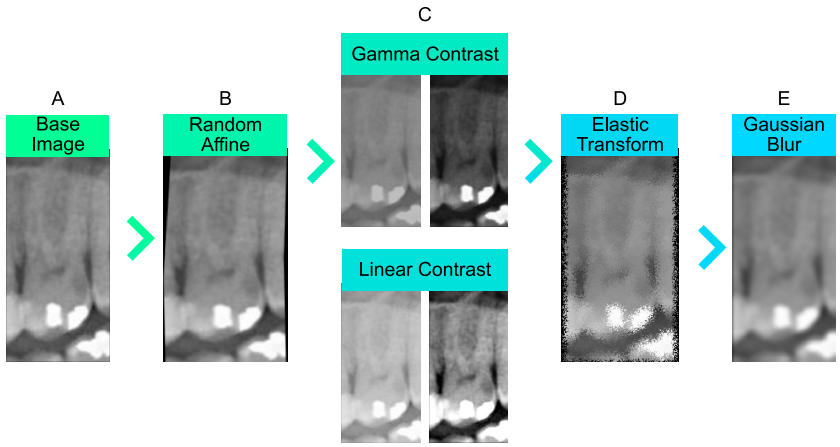

Fig. 2. Visualization of selected augmentation techniques applied to a base image (A) used in the proposed augmentation pipeline. (B) Random affine transformation introduces spatial distortions. (C) Gamma contrast adjustment enhances perceptual contrast by mimicking human brightness perception. Linear contrast is shown for comparison to illustrate the limitations in highlighting regions and visualize the difference compared to gamma contrast. (D) Elastic transformation introduces localized deformations. (E) Gaussian blur simulates acquisition artifacts or image quality issues. Horizontal and vertical flips, while part of the pipeline, are omitted for clarity.

To enhance the various models' generalization capabilities across both in-distribution and out-of-distribution test scenarios, we implement a comprehensive data augmentation pipeline. The proposed strong augmentation strategy comprises of a diverse set of transformations, including horizontal and vertical flips, Gaussian blur, gamma contrast adjustment, elastic deformation and random affine transformation operations (see Figure 2). This pipeline builds upon the augmentation approach in [9], but it is substantially modified to meet the specific challenges of caries detection.

A critical aspect in radiographic diagnosis by clinicians is the perception of contrast, particularly adjusted via gamma correction, which modulates image brightness non-linearly to better reflect human visual sensitivity [4], opposed to

linear contrast adjustments. To simulate this diagnostic process, gamma contrast adjustments are explicitly integrated into the augmentation pipeline, allowing the model to learn to identify carious lesions under varying contrast conditions (see Figure 2 (C)).

During training, the strong augmentation pipeline is applied with a probability of 50% for each tooth image. When activated, each individual augmentation operation also has an independent probability, ranging between 50% and 100%, of being applied. Furthermore, the strength of the transformation is also randomly sampled from a predefined range specific to each technique.

We evaluate the efficacy of this strong augmentation pipeline by comparing model performance when training with this pipeline and when training with a more trivial set of augmentations. The trivial augmentation pipeline is composed of only horizontal flips, vertical flips, and random rotations, following the standard augmentation practices reported in recent studies [8, 1, 17]. Note that the strong augmentation pipeline has primarily been designed for the purpose of improving segmentation (as opposed to classification) performance.

3.4 DINO classification model

For binary classification of caries presence, we finetune a pretrained DINOv2 Vision Transformer[3] [20] on single tooth images, as shown in Figure 1 (C). The model is used as a partially fine-tuned feature extractor and a single-layer classification head is added on top of this backbone to generate probability scores for caries presence. To find the best configuration for the model we conduct an extensive hyperparameter search, consisting of different learning rates, learning rate scheduling, different learning rates per layer (referred to as 'split learning rate'), freezing and unfreezing the backbone at different epochs ('unfreeze epoch'), and with and without (trivial) data augmentation. Note that the proposed augmentation strategy from Section 3.3 did not outperform the trivial augmentation in this setting. We search over the following values for each hyperparameter:

- **Learning rate:** {0.001, 0.0001, 0.00001}
- **Maximum training epochs:** {200, 400}
- **Scheduler:** {cosine annealing, step decay}
- **Unfreeze epoch:** {50, 100, 150}
- **Split learning rates:** {True, False}
- **Data augmentations:** {enabled, disabled}

As optimizer, we use AdamW with either cosine annealing or step-based learning rate scheduling. When split learning rates is enabled, the classification head uses a learning rate 10× higher than the backbone. Training uses a weighted binary cross-entropy loss, with the positive class (caries) weighted according to the ratio of negative to positive samples in each training fold. As previously mentioned (Section 3.1), we make use of 5-fold patient-wise cross-validation. For

[3] Specifically, we use the 'base' variant from hugging face: https://huggingface.co/facebook/dinov2-base

each fold, we select the best-performing model based on validation AUC and then evaluate and report the performance on the held-out test sets. Finally, we determine the classification threshold by the Youden's Index to balance sensitivity and specificity during classification [26].

For a baseline comparison, we also perform the same hyperparameter search and report results when using a (ImageNet) pretrained ResNet50 [11], as it is a popular choice for image classification tasks.

3.5 Mask-R-CNN segmentation model

For the segmentation of carious lesions, we use a Mask R-CNN architecture [10], initialized with ImageNet-pretrained weights, and train it for 2450 epochs. The model is trained on single-tooth images, each resized to 256×256 pixels (see Figure 1D). To improve generalization, we apply the strong augmentation strategy described in Section 3.3. Training is conducted in three stages to gradually unfreeze the network layers: Initially, only the network heads are trained for 25 epochs. Subsequently, layers from stage 4 onward are unfrozen and trained for an additional 150 epochs. Finally, the entire network was fine-tuned for the remaining epochs. We use a fixed learning rate of 0.0001 and a weight decay of 0.0001 throughout all training stages. The loss components, classification, bounding box regression, and mask loss, are used with equal weighting. The model is trained using stochastic gradient descent (SGD) with a momentum of 0.9 and gradient clipping, as implemented in the Matterport Mask R-CNN framework[4]. Due to the computational demands of the training process, an extensive hyperparameter search is not feasible. Instead, training hyperparameters are selected based on established best practices in the literature. We follow the same protocol as for the classification model and train one network for each data fold.

3.6 A-priori knowledge

To improve the clinical plausibility of the segmentation results during inference, we apply a two-stage postprocessing pipeline based on anatomical constraints and expert dental knowledge. In the first stage, we use the binary tooth mask, obtained from the first tooth segmentation step (recall Section 3.2, shown in Figure 1 (A)), to suppress predicted lesions located outside the visible crown. Predicted carious lesion masks extending more than 50% beyond the tooth boundary are trimmed or removed, particularly to suppress common false positives in the interdental space. In the second stage, we implement spatial filters to exclude lesion predictions within clinically irrelevant areas. Specifically, we compute the geometric center of the tooth mask and discard all predicted lesion segments within a 20-pixel radius of this center. Additionally, we remove predictions along a 20-pixel wide vertical corridor extending from the center toward the root, corresponding to the root canal region, which is radiolucent and not considered for caries diagnosis. Predictions located more than 60 pixels below the center along

[4] https://github.com/matterport/Mask_RCNN

the vertical axis, typically beneath the gumline, are also discarded. This postprocessing step is illustrated in Figure 1 (E), with the resulting refined segmentation shown in Figure 1 (F).

4 Results

In this section we present our experimental results for the various models under different training/evaluation configurations. First, we assess the performance of the classification models in isolation in Section 4.1, followed by the segmentation results in Section 4.2. Finally, we do a more fine-grained evaluation of the segmentation results in Section 4.3.

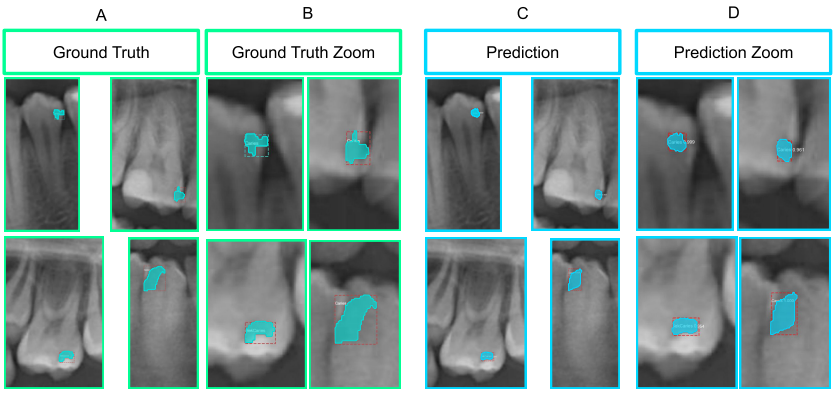

Fig. 3. A selection of ground truth segmentations (A) and the zoom in (B) together with their corresponding predictions (C) and the zoom in predictions (D), from the Mask-R-CNN illustrating the visual similarity between true positive segmentation masks and the ground truth masks.

4.1 Classification

We now evaluate the performance of our two binary classification models, namely the finetuned DINOv2 transformer as well as the baseline ResNet50 model (both explained in Section 3.4). This is done in all three training/evaluation settings: 1) UKSH Test, UKSH Train, 2) UFBA Test, UKSH Train, and 3) Mixed Test, Mixed Train, as explained in Section 3.1. The results of this evaluation are presented in Table 2. We use the area under the ROC curve (AUC), as well as Sensitivity (Sens) and Specificity (Spec), calculated on a per tooth basis, as evaluation metrics. Results are reported on the test sets as mean $\pm$ standard deviation across the five data folds. We now elaborate on the different observations that can be made from these results.

Table 2. Classification performance (mean ± standard deviation) on the UKSH, UFBA, and mixed test sets for DINOv2 and ResNet50 classification models. Results are reported for AUC, Sensitivity (Sens), and Specificity (Spec).

Test / Train	Model	AUC	Sens	Spec
UKSH / UKSH	DINOv2	0.8126 ± 0.013	0.7356 ± 0.097	0.7294 ± 0.096
	ResNet50	0.7911 ± 0.016	0.7022 ± 0.092	0.7445 ± 0.090
UFBA / UKSH	DINOv2	0.6854 ± 0.030	0.7593 ± 0.133	0.4726 ± 0.157
	ResNet50	0.6427 ± 0.032	0.7000 ± 0.170	0.4530 ± 0.184
Mixed / Mixed	DINOv2	**0.8726** ± 0.017	**0.8518** ± 0.037	0.7346 ± 0.024
	ResNet50	0.8340 ± 0.019	0.7904 ± 0.034	**0.7447** ± 0.033

Firstly, we compare the performance achieved on the UKSH (first configuration) and UFBA (second configuration) test sets for the models trained on the UKSH train set. Recall that in this setting the UFBA test set serves as an indicator of out-of-distribution performance while the UKSH test set reflects in-distribution performance.

For the DinoV2 model, we observe that the mean AUC drops from 0.8126 ± 0.013 to 0.6854 ± 0.030 for the UKSH and UFBA test sets, respectively. When considering the other metrics, it can be observed that this performance decrease can be explained by a significant reduction in the mean specificity (0.7294 ± 0.096 to 0.4726 ± 0.157). For the ResNet50 model, we observe similar results, and note a drop from 0.7911 ± 0.016 to 0.6427 ± 0.032 mean AUC, again explained by a drop in mean specificity (0.7445 ± 0.090 to 0.4530 ± 0.184).

Secondly, when considering the mixed results (third configuration, which should be considered an in distribution test), we find the best overall performance: 0.8726 ± 0.017 and 0.8340 ± 0.019 mean AUC for the DinoV2 model and ResNet50 baseline, respectively.

These results demonstrate notable differences between in-domain and out-of-domain performance, which indicates a substantial distribution shift between the two datasets. However, the combined results indicate that this shift is not troublesome when mixing two datasets, and is therefore a suitable approach for training a caries classifier.

Finally, when considering the DINOv2 model in comparison to the ResNet50 model, we observe that the DINOv2 offers superior performance in the case of all three datasets.

4.2 Segmentation

In this section we evaluate the performance of our segmentation models under various conditions. Our aim is to compare our proposed augmentation and classify-then-segment pipeline against two baselines. To this end, we evaluate the performance of three model configurations:

1. **Seg (no Aug)** - only the segmentation model applied to the single tooth images and trained without the strong augmentations (only the trivial aug-

mentations, as explained in Section 3.3). This serves as a weak baseline for comparison.
2. **Seg + Aug** - The same as (1) but now the segmentation model is trained with strong augmentations.
3. **Seg + Aug + Class** - our proposed two stage pipeline where the segmentation model is trained with strong augmentations and only applied to teeth classified as carious by the classification model. We use DINOv2 classification model from the previous section for this configuration.

This is done for all three training/evaluation configurations (as in the previous section). The results of this evaluation are shown in Table 3. We rely on the mean Intersection over Union (IoU), as well as the mean Dice score to evaluate performance. These are calculated individually for each tooth, and then averaged across all teeth. To avoid inflation from the overwhelming background, we report foreground-only scores, i.e., IoU/Dice computed on the caries class while ignoring background/true-negative pixels. Including background would yield deceptively high values because non-carious tissue occupies the vast majority of each image. [5] Furthermore, the scores we report are the mean IoU/Dice across the five data folds. We make the following observations:

Table 3. Segmentation performance (mean ± standard deviation) across different configurations on the UKSH, UFBA, and mixed test sets. The models were evaluated using a two-stage classification–segmentation pipeline (Seg + Aug + Class) or just the segmentation model with augmentations (Seg + Aug). As a weak baseline we include the results of the segmentation only without strong augmentations (Seg (No Aug)).

Test / Train	Configuration	IoU	Dice
UKSH / UKSH	Seg + Aug + Class	**0.1573** ± 0.010	**0.2285** ± 0.014
	Seg + Aug	0.0964 ± 0.005	0.1404 ± 0.005
	Seg (no Aug)	0.0329 ± 0.007	0.0483 ± 0.011
UFBA / UKSH	Seg + Aug + Class	**0.1731** ± 0.013	**0.2513** ± 0.019
	Seg + Aug	0.1466 ± 0.002	0.2144 ± 0.005
	Seg (no Aug)	0.0435 ± 0.013	0.0612 ± 0.018
Mixed / Mixed	Seg + Aug + Class	**0.1865** ± 0.011	**0.2719** ± 0.015
	Seg + Aug	0.1541 ± 0.007	0.2245 ± 0.010

Two-stage vs. Segmentation only The two-stage pipeline, in which classification is used to identify carious cases prior to segmentation, consistently yields higher segmentation performance than the two segmentation only approaches, across all test sets.

On the UKSH test set, the classification aided configuration (Seg + Aug + Class) achieves an IoU of 0.1573 ± 0.010 and a Dice score of 0.2285 ± 0.014. In

[5] Tooth images where both the predictions and ground tooth are all 0 (only true negatives are present) are disregarded from the average.

comparison, the best segmentation only model (Seg + Aug) reaches an IoU of 0.0964 ± 0.005 and Dice of 0.1404 ± 0.018. We see similar results when comparing performance on the other two test sets: 0.1731 ± 0.013 versus 0.1466 ± 0.002 IoU for the UFBA test set, and 0.1865 ± 0.011 versus 0.1541 ± 0.007 for the mixed set, where the Dice scores show similar trends.

These results indicate that restricting segmentation to teeth identified as carious by a preceding classifier leads to improved performance compared to applying the segmentation model on all images.

In-distribution vs. Out-of-distribution When comparing the UKSH and UFBA test sets, we observe that, interestingly, performance appears to be better on the out-of-distribution set than the in-distribution one for all three model configurations. For the classification aided configuration, we find an IoU of 0.1573 ± 0.010 versus 0.1731 ± 0.013, and a Dice score of 0.2285 ± 0.014 versus 0.2513 ± 0.019, for the UKSH and UFBA test sets, respectively. We see similar results when considering the segmentation only models. It is not clear why this is the case, and why we do not observe the same trends for the classification models in the previous section.

When considering the mixed dataset, we find the best performance, with the classification aided configuration we observe the highest segmentation scores among all configurations, with an IoU of 0.1731 ± 0.013 and Dice of 0.2513 ± 0.019. Similarly, the segmentation model with augmentations also reports the best performance on this dataset: an IoU of 0.1541 ± 0.007 and Dice score of 0.2245 ± 0.010. This is further evidence that mixing data from multiple centers is a good approach, and illustrates the benefits of training with diverse data.

Augmentation vs. Trivial To assess the impact of the proposed augmentation pipeline on the segmentation, we compare models trained with and without it. The segmentation model without augmentations (Seg (no Aug)) performed substantially worse on the UKSH test set with 0.0329 ± 0.007 IoU and 0.0483 ± 0.011 Dice compared to the model trained with strong image augmentation. The same trend was observed on the UFBA test set, where the Seg (no Aug) model achieved 0.0435 ± 0.013 IoU and 0.0612 ± 0.018 Dice. These findings indicate that the augmentation pipeline considerably boosts segmentation performance. Due to the substantially worse performance of the Seg (no Aug) model and the high computational cost of training the model we omitted this test for the mixed dataset.

4.3 Additional segmentation metrics

To further investigate the performance of our proposed two-stage model (The 'Seg + Aug + Class' variant from Table 3) we compute additional metrics on a *per lesion* basis. Specifically, we consider all of the model's individual mask predictions and calculate the rate of True Positives (TP), False Positives (FP), and False Negatives (FN). Note that there is no concept of True Negatives (TN)

in this setting. This provides a more fine-grained evaluation than the single scalar values provided by the per tooth level IoU or Dice score. In combination with this, we then again leverage the IoU and Dice score to assess the quality of only the TP segmentations - i.e. cases where our predicted mask overlaps with the ground truth mask. This allows us to assess the segmentation quality independently of FP cases. These additional evaluations are performed to determine if the model struggles to predict accurate masks in general, or if the dice score is influenced more by FP predictions (since these quickly degrade performance).

The rate of the different prediction cases (TP/FP/FN) for each test set is shown in Table 4, while the segmentation performance for the TP cases are shown in Table 5, as well as example TP predictions in Figure 3.

When considering Table 4, we observe that FP cases are indeed very common for all three test sets, and consist of 30.19% to 41.01% of all cases, where the most are encountered on the UFBA test set and the least on the mixed test set. Similarly, the TP rate is lower for the out-of-distribution UFBA test set (21.87%), and highest for the mixed set (28.18%). We believe that this a reflection of the classification results previously observed in Section 4.1. Interestingly, when considering the FN rate, we find that it is highest for the mixed dataset (41.63%). This is an indication that the mixed dataset encourages more conservative prediction masks.

Considering Table 5, we observe that the IoU and Dice scores are substantially better for the TP cases than the global metrics reported earlier (Table 3). This shows that the TP segmentation masks are, on average, of a more usable quality. We note an IoU range of 0.3041 ± 0.014 to 0.3177 ± 0.014, and a Dice score range of 0.4695 ± 0.022 to 0.4750 ± 0.016. Interestingly, the performance across all three datasets appear to be more similar than those for the global metrics. This is an indication that the quality of the different models is predominantly distinguished by how well they can avoid FP and FN lesion predictions.

Table 4. Number of true positives (TP), false positives (FP) and false negatives (FN) on a per lesion basis for the two-stage model on the UKSH, UFBA, and Mixed test sets.

Test / Train	Configuration	TP (%)	FP (%)	FN (%)
UKSH / UKSH	Seg + Aug + Class	27.02 ± 1.87	37.86 ± 4.51	**35.12** ± 5.23
UFBA / UKSH	Seg + Aug + Class	21.87 ± 2.47	41.01 ± 6.28	37.12 ± 7.59
Mixed / Mixed	Seg + Aug + Class	**28.18** ± 2.45	**30.19** ± 7.68	41.63 ± 2.32

5 Discussion

In this study, we investigated carious lesion segmentation on single tooth images extracted from full panoramic radiographs. More specifically, we 1) investigate whether a two-stage classification and segmentation pipeline improves perfor-

Table 5. Segmentation metrics considering only TP predictions,i.e. mask predictions where the dice score is larger than zero, for the two-stage model on the UKSH, UFBA, and Mixed test sets.

Test / Train	Configuration	IoU	Dice
UKSH / UKSH	Seg + Aug + Class	0.3143 ± 0.016	**0.4750** ± 0.016
UFBA / UKSH	Seg + Aug + Class	**0.3177** ± 0.014	0.4748 ± 0.016
Mixed / Mixed	Seg + Aug + Class	0.3041 ± 0.014	0.4695 ± 0.022

mance compared to a standalone segmentation model; 2) we propose and investigate the performance benefits of a novel data augmentation pipeline; and 3) we conduct comparisons between in-distribution and out-of-distribution test scenarios. In this section we address and discuss the results from these three investigations, and additionally contrast our results to those found in the literature.

Firstly, our results show that filtering tooth images using a classification model prior to segmentation increases segmentation accuracy compared to a direct segmentation approach without classification. Specifically, we observe that preselecting carious images using a classifier leads to improved segmentation performance, particularly by reducing false positive detections. We find that this performance increase holds in both in-distribution and out-of-distribution test scenarios. Regarding the classification component, we find that a finetuned DinoV2 model provides the best performance in comparison to a finetuned ResNet50 model.

Secondly, when considering our proposed data augmentation pipeline, we find that it significantly enhances segmentation performance, particularly in settings without classification-based preselection. This emphasizes the importance of data diversity for learning robust segmentation features. That said, our augmentation strategy did not improve classification performance (we omitted these results in this study). This suggests that the DINO transformer backbone may already encode robust and generalizable features, rendering extensive augmentation less beneficial or even counterproductive in this context.

Thirdly, when considering the different test environments, our classification experiments revealed a notable performance drop when evaluating the model on an out-of-distribution (OOD) test set compared to the in-distribution test set. This domain shift highlights the sensitivity of caries classification models to changes in image characteristics. Conversely, segmentation results on the OOD test set were slightly superior to those on the in-domain set. This counterintuitive finding may be the result of different factors: the OOD test set (UFBA) is substantially larger and contains more homogeneous images than the in-domain test set (UKSH), potentially facilitating more consistent segmentation.

We further observe that training on a mixed-domain dataset leads to improved performance compared to single-domain training for both classification and segmentation. This indicates that incorporating data from multiple domains enhances the robustness of the models and suggests that such diversity can signif-

icantly improve overall performance. Additionally, the larger number of training samples in the mixed domain setting likely contributes to more stable training and better performance.

Finally, when considering prior work, our classification performance aligns well with values reported in existing literature on panoramic dental x-rays [8], including studies using higher-resolution periapical or bitewing radiographs [21, 5]. Moreover, related work has explored classification tasks on specific subsets of teeth, such as third molars, using panoramic radiographs [24]. Across these studies, reported AUC for classification-related tasks typically range between 0.73 and 0.98. While direct comparisons are limited by the absence of a standardized test set and differences in image quality, tooth types, and caries severity, our results fall within this range.

In the context of segmentation, previous work has primarily focused on segmentation at the level of full panoramic images or multi-teeth regions, with some using methods like GradCAM to highlight decision regions [21, 7, 1]. However, segmentation on *single-tooth images* remains rather unexplored. Therefore, in addition to the lack of public test sets, direct comparisons of absolute segmentation performance remain challenging. Nevertheless, reported Dice scores for caries segmentation in the literature span a broad range, from 0.15 to 0.65 on full panoramic radiographs [7, 1], and up to 0.75 on much higher resolution periapical images [25, 14], which suggests that our results are within the expected range for this task. Furthermore, the additional segmentation metrics focusing solely on TP, lesion-level mask predictions provide a clearer picture of the models' actual segmentation capability, independent of false positives. These refined metrics demonstrate that, when the model detects a lesion, the segmentation quality is considerably more reliable than global metrics suggest. This distinction is especially relevant for clinical applications, where overlooking a carious lesion is generally considered more critical than additionally identifying an uncertain or borderline region.

6 Conclusion

In conclusion, we show that caries detection on panoramic radiographs benefits from three elements: (1) a DINO-based classifier that filters non-carious teeth before segmentation and reduces false positives; (2) a strong augmentation pipeline that increases segmentation accuracy; and (3) training with data from multiple centers to increase data diversity during training. While segmentation of carious lesions remains a challenging task, with performance reflecting the complexity and ambiguity of the underlying imaging and annotation, our results show that each component of our pipeline contributes meaningfully to improving segmentation quality. Notably, our classification results are consistently strong, highlighting the clinical utility of our approach for automated screening.

References

1. Alharbi, S.S., AlRugaibah, A.A., Alhasson, H.F., Khan, R.U.: Detection of cavities from dental panoramic x-ray images using nested u-net models. Applied Sciences **13**(23) (2023)
2. Bayrakdar, I.S., Orhan, K., Akarsu, S., Çelik, Ö., Atasoy, S., Pekince, A., Yasa, Y., Bilgir, E., Sağlam, H., Aslan, A.F., et al.: Deep-learning approach for caries detection and segmentation on dental bitewing radiographs. Oral Radiology pp. 1–12 (2022)
3. Bui, T.H., Hamamoto, K., Paing, M.P.: Deep fusion feature extraction for caries detection on dental panoramic radiographs. Applied Sciences **11**(5), 2005 (2021)
4. Charles, P., et al.: Digital video and hdtv algorithms and interfaces. Morgan Kaufmann Publishers, San Francisco **260**, 630 (2003)
5. Chen, I.D.S., Yang, C.M., Chen, M.J., Chen, M.C., Weng, R.M., Yeh, C.H.: Deep learning-based recognition of periodontitis and dental caries in dental x-ray images. Bioengineering **10**(8), 911 (2023)
6. Chen, Q., Huang, J., Zhu, H., Lian, L., Wei, K., Lai, X.: Automatic and visualized grading of dental caries using deep learning on panoramic radiographs. Multimedia Tools and Applications **82**(15), 23709–23734 (2023)
7. Dayı, B., Üzen, H., Çiçek, B., Duman, B.: A novel deep learning-based approach for segmentation of different type caries lesions on panoramic radiographs. Diagnostics **13**(2) (2023)
8. Haghanifar, A., Majdabadi, M.M., Ko, S.: Paxnet: Dental caries detection in panoramic x-ray using ensemble transfer learning and capsule classifier. CoRR **abs/2012.13666** (2020), https://arxiv.org/abs/2012.13666
9. Hansen, C.J., Conrad, J., Seidel, R., Krekiehn, N.R., Yilmaz, E., Koser, N., Goetze, M., Gehrmann, T., Lauterbach, S., Graetz, Christian Doerfer, C., Glueer, C.C.: Automated tooth instance segmentation and pathology annotation pipeline for panoramic radiographs. In: BVM 2024: Proceedings, German Conference on Medical Image Computing, Erlangen, March 10-12, 2024. p. 237. Springer-Verlag (2024)
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. IvisionLab: dnspanoramicimagesv2. https://github.com/IvisionLab/dns-panoramic-images-v2 (Jan 2021), data set comprising 450 panoramic dental X-rays, split into six folds for segmentation and numbering tasks
13. Kawazu, T., Takeshita, Y., Fujikura, M., Okada, S., Hisatomi, M., Asaumi, J.: Preliminary study of dental caries detection by deep neural network applying domain-specific transfer learning. Journal of Medical and Biological Engineering **44**(1), 43–48 (2024)
14. Khan, H.A., Haider, M.A., Ansari, H.A., Ishaq, H., Kiyani, A., Sohail, K., Muhammad, M., Khurram, S.A.: Automated feature detection in dental periapical radiographs by using deep learning. Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology **131**(6), 711–720 (2021)
15. Li, Z., Yu, C., Chen, H.: Global, regional, and national caries of permanent teeth incidence, prevalence, and disability-adjusted life years, 1990–2021: analysis for the global burden of disease study. BMC Oral Health **25**(1), 715 (May 13 2025)

16. Lian, L., Zhu, T., Zhu, F., Zhu, H.: Deep learning for caries detection and classification. Diagnostics **11**(9) (2021), https://www.mdpi.com/2075-4418/11/9/1672
17. Liu, Y., Xia, K., Cen, Y., Ying, S., Zhao, Z.: Artificial intelligence for caries detection: a novel diagnostic tool using deep learning algorithms. Oral Radiology pp. 1–10 (2024)
18. Mohammad-Rahimi, H., Motamedian, S.R., Rohban, M.H., Krois, J., Uribe, S.E., Mahmoudinia, E., Rokhshad, R., Nadimi, M., Schwendicke, F.: Deep learning for caries detection: A systematic review. Journal of Dentistry **122**, 104115 (Jul 2022)
19. Mărginean, A.C., Mureşanu, S., Hedeşiu, M., Dioşan, L.: Teeth segmentation and carious lesions segmentation in panoramic x-ray images using cariseg, a networks' ensemble. Heliyon **10**(10), e30836 (2024)
20. Oquab, M., Darcet, T., Moutakanni, T., et al.: Dinov2: Learning robust visual features without supervision (2024), https://arxiv.org/abs/2304.07193
21. Oztekin, F., Katar, O., Sadak, F., Yildirim, M., Cakar, H., Aydogan, M., Ozpolat, Z., Talo Yildirim, T., Yildirim, O., Faust, O., Acharya, U.R.: An explainable deep learning model to prediction dental caries using panoramic radiograph images. Diagnostics **13**(2) (2023), https://www.mdpi.com/2075-4418/13/2/226
22. Prados-Privado, M., García Villalón, J., Martínez-Martínez, C.H., Ivorra, C., Prados-Frutos, J.C.: Dental caries diagnosis and detection using neural networks: A systematic review. Journal of Clinical Medicine **9**(11), 3579 (2020)
23. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. CoRR **abs/1610.02391** (2016), http://arxiv.org/abs/1610.02391
24. Vinayahalingam, S., Kempers, S., Limon, L., Deibel, D., Maal, T., Hanisch, M., Bergé, S., Xi, T.: Classification of caries in third molars on panoramic radiographs using deep learning. Scientific Reports **11**(1), 12609 (2021)
25. Ying, S., Wang, B., Zhu, H., Liu, W., Huang, F.: Caries segmentation on tooth x-ray images with a deep network. Journal of Dentistry **119**, 104076 (2022)
26. Youden, W.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950)
27. Zhu, H., Cao, Z., Lian, L., Ye, G., Gao, H., Wu, J.: Cariesnet: a deep learning approach for segmentation of multi-stage caries lesion from oral panoramic x-ray image. Neural Computing and Applications **35**(22), 16051–16059 (2023)