

Article

Segmentation of Dental Restorations on Panoramic Radiographs Using Deep Learning

Csaba Rohrer ¹, Joachim Krois ^{1,2} , Jay Patel ³, Hendrik Meyer-Lueckel ⁴ , Jonas Almeida Rodrigues ^{1,5} 
and Falk Schwendicke ^{1,2,*} 

- ¹ Oral Diagnostics, Digital Health and Health Services Research, Charité–Universitätsmedizin Berlin, 10117 Berlin, Germany; csaba.rohrer@charite.de (C.R.); joachim.krois@charite.de (J.K.); jorodrigues@ufrgs.br (J.A.R.)
² ITU/WHO Focus Group on AI for Health, Topic Group Dental Diagnostics and Digital Dentistry, 1202 Geneva, Switzerland
³ Informatics, Department of Health Services Administrations and Policy, Temple University College of Public Health, Philadelphia, PA 19140, USA; patel.jay@temple.edu
⁴ Operative, Preventive and Pediatric Dentistry, 3010 Bern, Switzerland; hendrik.meyer-lueckel@zmk.unibe.ch
⁵ Surgery and Orthopedics, UFRGS, Porto Alegre 90040-060, Brazil
* Correspondence: falk.schwendicke@charite.de

Abstract: Convolutional Neural Networks (CNNs) such as U-Net have been widely used for medical image segmentation. Dental restorations are prominent features of dental radiographs. Applying U-Net on the panoramic image is challenging, as the shape, size and frequency of different restoration types vary. We hypothesized that models trained on smaller, equally spaced rectangular image crops (tiles) of the panoramic would outperform models trained on the full image. A total of 1781 panoramic radiographs were annotated pixelwise for fillings, crowns, and root canal fillings by dental experts. We used different numbers of tiles for our experiments. Five-times-repeated three-fold cross-validation was used for model evaluation. Training with more tiles improved model performance and accelerated convergence. The F1-score for the full panoramic image was 0.7, compared to 0.83, 0.92 and 0.95 for 6, 10 and 20 tiles, respectively. For root canals fillings, which are small, cone-shaped features that appear less frequently on the radiographs, the performance improvement was even higher (+294%). Training on tiles and pooling the results thereafter improved pixelwise classification performance and reduced the time to model convergence for segmenting dental restorations. Segmentation of panoramic radiographs is biased towards more frequent and extended classes. Tiling may help to overcome this bias and increase accuracy.



Citation: Rohrer, C.; Krois, J.; Patel, J.; Meyer-Lueckel, H.; Rodrigues, J.A.; Schwendicke, F. Segmentation of Dental Restorations on Panoramic Radiographs Using Deep Learning. *Diagnostics* **2022**, *12*, 1316. <https://doi.org/10.3390/diagnostics12061316>

Academic Editor: Daniel Fried

Received: 6 May 2022

Accepted: 25 May 2022

Published: 25 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: machine learning; deep learning; image segmentation; dental restorations

1. Introduction

A wide range of applications in the field of artificial intelligence and, specifically, machine and deep learning (ML/DL) are entering health care. A particularly prolific subfield of ML/DL is computer vision, i.e., image analysis using ML/DL methods. Computer vision applications have been developed in dermatology (for analyzing skin photographs), ophthalmology (retinal imaging) and oncology [1–3]. In dental radiology, ML/DL has been increasingly adopted to analyze dental radiographs—for instance, to detect caries, apical lesions, periodontal bone loss or anatomic landmarks on a wide range of 2D and 3D imagery, including panoramic radiographs [4].

Image segmentation refers to classification tasks for pixels with semantic labels (semantic segmentation) or partitioning of individual objects (instance segmentation) [5]. Semantic segmentation performs pixel-level labeling with a set of object categories for all image pixels (e.g., human, car, tree, sky, and, as in the present case, different dental restorations). Instance segmentation extends the semantic segmentation scope further by

detecting and delineating each individual object of interest in the image (e.g., partitioning of each individual dental restorations). For medical applications, image segmentation is popular given that the model output (segmentation masks) can be interpreted by human experts and hence provides inherent explainability. The downsides of such an approach are (1) the high human expert labor costs, as each object of interest needs to be annotated by marking the area; (2) classical evaluation metrics fall short as the pixel scale as interpretation is much more difficult, hence concepts and metrics such as the intersection over union (IoU) and (weighted) mean average precision (mAP) are applied [6]; and (3) technical challenges emerge, as only a few pixels of the image actually show the area/lesion of interest in many use cases, which in turn results in an imbalanced learning task and hence a more challenging task, where many negative pixels but only a few positives need to be learnt.

While assisting practitioners in pathology detection or anatomic landmarking has obvious diagnostic value, ML/DL computer vision applications also come with the promise of increasing the efficiency and quality of reporting. To support dentists in systematic reporting of radiographic findings, a range of further ML/DL models are required—for example, for tooth detection and classification or the detection and classification of restorative materials and previous dental work experience. Having such models allows one to assign identified pathologies to teeth and to comprehensively report all findings from the radiograph in only a few seconds, which is relevant for more extended imagery such as panoramics. An increasing number of studies reported on the detection of pathologies, while fewer studies are available for detecting and classifying dental restorative materials for example [7,8].

Current ML/DL models for computer vision mainly employ Convolutional Neural Networks (CNNs). CNNs have a fixed resolution for the input image, which is smaller than the original image in most cases (especially for large imagery such as panoramics). Hence, images must be downsized before applying the CNN. It is conceivable that cropping larger images (such as panoramic radiographs) into smaller, equally spaced rectangular image crops (tiles) and training CNNs on such crops circumvents information loss due to downsizing and hence improves model performance.

Using the exemplary task of semantic segmentation of restorations and other previous dental work (such as fillings, crowns, root-canal fillings and implants) on panoramics, we hypothesized that models trained on tiled data outperform models trained on full radiographs.

2. Materials and Methods

2.1. Study Design

The goal of this study was to improve the accuracy of multi-label segmentation of dental restorations on panoramic radiographs by using a tiling approach. The original radiographs were cropped into multiple parts to increase the resolution of the images used during the training of the model. By using these subsets of the original images, we aimed to capture more details, thus increasing segmentation performance. We trained on a different number of tiles between 2 and 20 and evaluated our models using the full image. The experimental workflow is shown in Figure 1. Reporting of our study follows the respective checklists [9,10].

2.2. Dataset

Our dataset contained 1781 panoramic radiographic images. Data collection was conducted between 2015 and 2018 and was ethically approved (ethics committee of Charité Berlin, EA4/080/18). Panoramics involving at least one permanent tooth were included. There were 54.3% male and 45.7% female patients; the mean (SD, min–max) age was 54.3 (19.7, 10–95) years. Data were generated by machines from the manufacturer Dentsply Sirona (Bensheim, Germany).

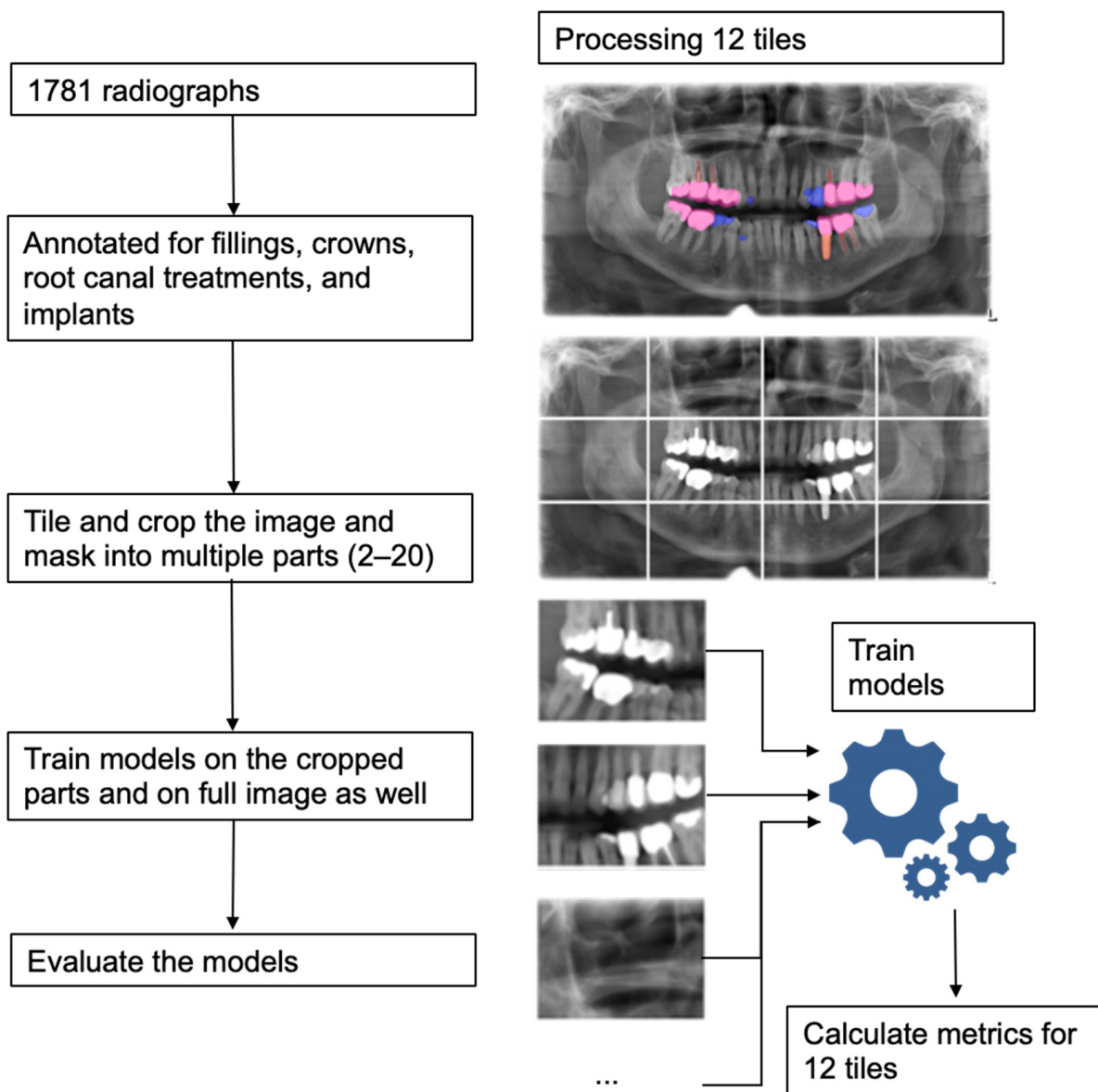


Figure 1. Experimental workflow and an example showing processing an image with 12 tiles.

The images were annotated pixelwise by dental experts using an in-house custom-built annotation tool [11] for four different restoration types, namely fillings, crowns, root canal treatments, and implants. These four annotations were stacked together to create multi-channel segmentation masks. Each of the channels represents the detections, or lack thereof, with pixelwise true and false values. The channel values are not mutually exclusive, so a pixel can belong to multiple classes in case of overlapping annotations. This means we were solving a multi-label semantic segmentation problem.

2.3. Model

The selected model for our experiments was based on U-Net [12], a Convolutional Neural Network for semantic image segmentation. It consists of an encoder and a decoder part connected with skip connections. The encoder extracts features from different spatial resolutions which are used by the decoder to define an accurate segmentation mask. The encoder is also known as the backbone since it is usually pretrained on a sizable dataset for feature extraction. For our backbone, we used the ResNeXt-50 ($32 \times 4d$), a variation of the ResNet-50 architecture with an added cardinality dimension [13]. The applied backbone

was based on the pretrained implementation of ResNext on the ImageNet dataset [14] from the PyTorch deep learning framework.

2.4. Training

We performed a five-times-repeated three-fold cross-validation (i.e., three-fold validation was repeated 5 times) to obtain more robust performance metric estimates and to reduce the bias introduced with a single training, validation, and test set partitioning. First, images were split into five groups. Each group represented a hold-out test of 20% of the dataset. We repeated a three-fold cross-validation for each test set, where the remaining four groups formed the training-validation data. These data were further split into actual training and validation images in three different splits. Each repetition resulted in three trained models, which were evaluated on their respective hold-out test set (Figure 2). Since we employed this method on all the test groups, our approach yielded fifteen models, which we used to establish mean and 95% confidence level of reported performance metrics.

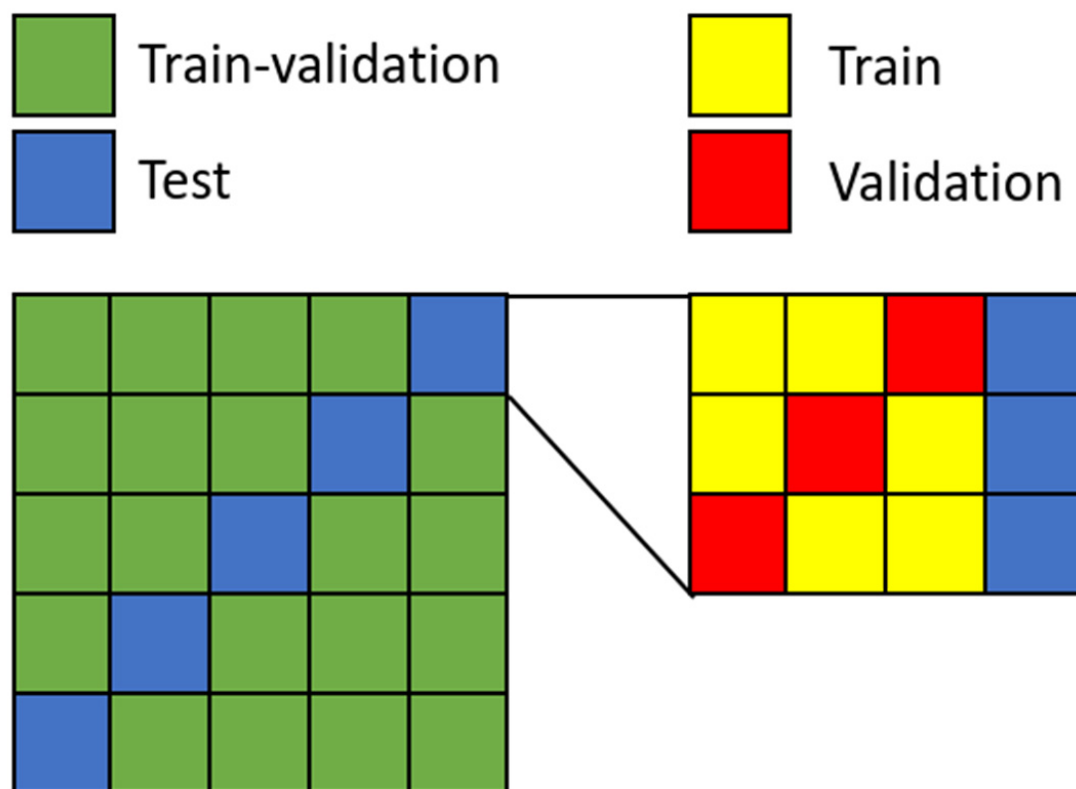


Figure 2. Visualization of the five-times-repeated three-fold cross-validation. For each hold-out test set, the remaining data is split into three different train and validation sets. Each datapoint is used as test data at some point.

We trained the decoder and fine-tuned the encoder side of our network. Before the panoramic radiograph images and their masks were fed to the model as input, we created smaller, equally sized crops (tiles) from them. We used an incremental number of tiles in our experiments from 2 up to 20. We also trained on the full images without any cropping for comparison, hence we had 11 different tile settings. The described cross-validation scheme was used for each of these settings.

The tiles would become the input to the network. The input images were resized to the resolution of 224×224 and then preprocessing functions (input range selection, normalization) were applied. The Dice loss function was used for training. The parameters were iteratively optimized using the Adam optimizer, with the initial learning rate and momentum set to 10×10^{-4} and 0.9, respectively. The batch size used was 50. We used an

adaptive learning rate based on the calculated validation loss after each epoch. If there was no improvement in the last five epochs, the learning rate was reduced to 10×10^{-5} . We employed the same strategy for early stopping as well. Once the learning rate was reduced, after no improvement in the last five epochs, we stopped the training process. This allowed us to keep optimizing the parameters further and save time by terminating the training loop. To increase the diversity of the data and to help prevent overfitting, we used data augmentation techniques during the training phase. The augmentations were applied to the original training data randomly to generate more examples. The techniques used were affine transformations, e.g., horizontal flips, rotation, brightness/contrast manipulations, image blurring and sharpening and gaussian noise application. The setup of the model architecture and optimization process was carried out using the deep learning framework PyTorch (1.4.0) and the Python (3.7.3) programming language. The models were trained using Nvidia A100 graphics cards with CUDA 11.

2.5. Metrics

We employed different metrics to compute the performance of our model. We used class-wise and overall F1-score, i.e., the harmonic mean of positive predictive value (precision) and sensitivity (recall), as well as sensitivity and, on a pixel level, intersection over union (IoU). The metric scores were evaluated on the different mask channels individually, and together as well.

2.6. Evaluation

For the evaluation, we used our trained model for inference on our test set. We cropped the images with the appropriate number of tiles, computed the segmentation masks, and calculated our evaluation metrics. Since we are dealing with a semantic segmentation task, we did not consider individual restorations separately but the whole class segmentation instead. We calculated pixelwise metrics on the image compared with the different channels of our predicted segmentation mask.

2.7. Explainability

A frequent problem with deep learning models is that the underlying neural networks are black boxes, meaning they do not provide insights into the approximation given. This is especially noticeable on classification problems, where the output of the model is a yes or no, or a simple label. In the case of segmentation, where the output is an area on the input image, the transparency of the prediction is an inherent property. For this reason, no further explainable techniques were applied to our models.

3. Results

Our experiments showed that with increasing number of tiles, model performance increased significantly. Using only one tile, the mean F1-score over all channels was 0.7 compared to 0.95 when 20 tiles were used. The most prominent increase occurred for root-canal fillings, i.e., small cone-shaped, less frequent features. In their case, the mean F1-score increased from 0.33 to 0.97. For implants and crowns, no significant improvement was detected. Sensitivity also showed a similar increase from 0.64 to 0.95 (Figure 3).

IoU on all channels also increased from 0.59 to 0.93 when using one to 20 tiles. All metrics for different restoration types are presented in the Supplementary Table S1.

While the number of input images per training epoch increases with higher number of tiles, we found tiling to nevertheless reduce convergence time, i.e., increase training efficiency (Figure S1).

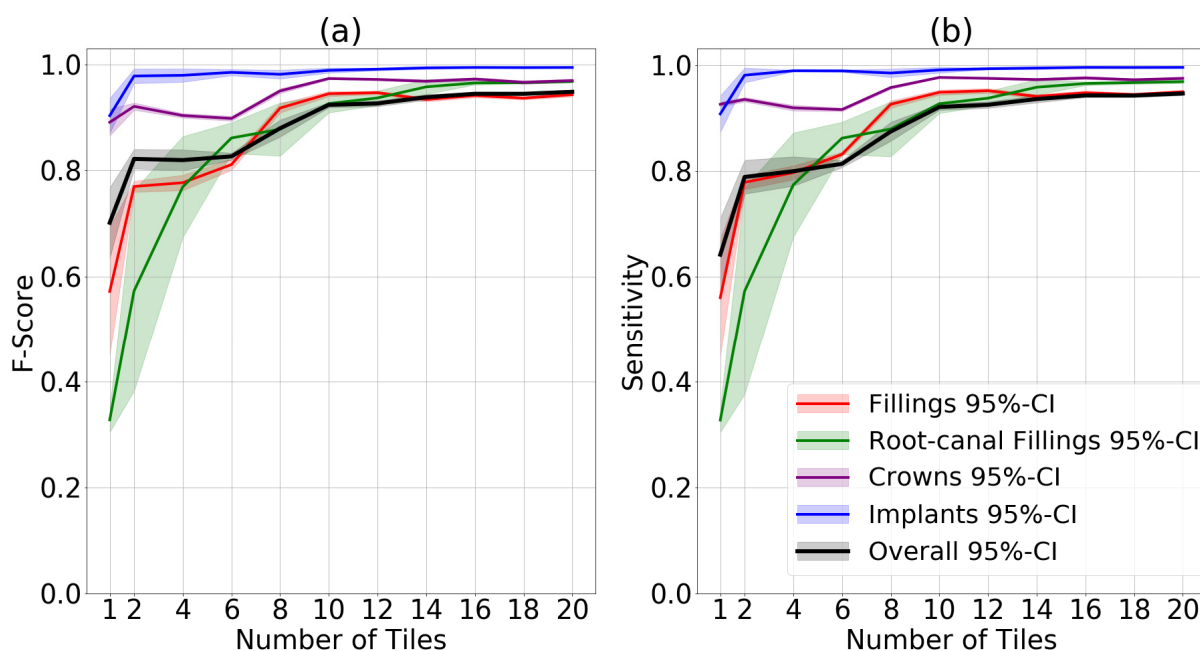


Figure 3. Mean and 95% CI F1-score (a) and sensitivity (b) for each restoration class as well as for all classes jointly.

4. Discussion

For many medical ML/DL applications involving imagery (computer vision), down-scaling image size and resolution, i.e., information loss, is required for training. This is also true for larger dental radiographs such as panoramics radiographs. Such loss of information may be well accepted for detection or segmentation tasks of larger and more prevalent entities (such as larger bone structures or, in dental imagery, teeth, crowns or bridges). For smaller and less prevalent findings such as root-canal fillings, but possibly also caries or apical lesions, this may be detrimental to the learning outcome, i.e., the diagnostic performance of the developed ML/DL application. Semantic segmentation of dental restorations and other previous dental work is a clinically relevant problem, as it allows to generate an automated and comprehensive. We hypothesized that downscaling imagery may negatively affect the segmentation accuracy of ML/DL for this task, and that tiling panoramic images may overcome this problem.

We confirmed this hypothesis; segmentation accuracy was significantly higher when considering all restoration classes (channels) for models trained on tiled than non-tiled images. A higher number of tiles showed a positive association with accuracy. Notably, when assessing the different classes, this effect was more prominent for smaller objects such as root-canal fillings or fillings than larger ones such as crowns or implants, as expected. Moreover, performance increased for both sensitivity and F1-score. Pixel-wise metrics such as the IoU also reflect that behavior. Notably, specificity and relating metrics were not estimated, as in semantic segmentation the negative class is made of the background. We further found tiling to speed up training convergence, a possible measure of efficiency.

Our findings need more detailed discussion. First, it is conceivable that the beneficial effect of tiling may be transferable to other segmentation tasks in the dental realm with class imbalances (a large number of pixels representing the background while only a few pixel represent the target), e. g., apical and caries lesions.

Second, our tiling approach reduced model convergence time, as using more tiles reduced the number of epochs needed for training.

Third, it needs highlighting that our study did not aim to develop clinically applicable segmentation models (maximizing accuracy), but that we rather aimed to systematically assess the effect of tiling to retain as much image information as possible by circumventing

the need to downscale the image. Future developers interested in clinical applicability may employ our approach and build on larger and more diverse datasets, but also leverage further instruments (such as hyperparameter tuning) to increase accuracy.

This study comes with a number of strengths and limitations. First, it was the first study to assess the effect of information retention (via tiling) for a widely employed semantic segmentation ML/DL task in dentistry. Second, and as discussed, the focus of this study was experimental, and our controlled approach with restricted margin conditions was not aimed for developing clinically applicable models. Last, and as a weakness, we employed data from one center and one x-ray manufacturer. Generalizability to other datasets or ML/DL tasks may not be given.

5. Conclusions

In this study, we explored the application of a U-Net architecture in a semantic segmentation setting to a varying number of image crops of panoramic dental radiographs. By tiling the original image and its corresponding segmentation mask, then cropping it for input to our neural network, we effectively increased the resolution. Our results showed that with a higher number of tiles, the model's performance improved. Smaller and less frequent features that might go unnoticed on the full panoramic image were especially affected. This approach could be easily tailored to other applications to raise detection performance.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/diagnostics12061316/s1>, Figure S1: Conversion of model training depending on the number of tiles; Table S1: Results for different restorations and tile numbers.

Author Contributions: Conceptualization, C.R., J.K., J.P., H.M.-L., J.A.R. and F.S.; methodology, C.R., J.K., J.P., H.M.-L., J.A.R. and F.S.; software, C.R.; validation, C.R., J.K. and F.S.; formal analysis, C.R., J.K., J.P., H.M.-L., J.A.R. and F.S.; investigation, C.R., J.K., J.P., H.M.-L., J.A.R. and F.S.; resources, C.R., J.K., J.P., H.M.-L., J.A.R. and F.S.; data curation, C.R., J.K., J.P., H.M.-L., J.A.R. and F.S.; writing—original draft preparation, C.R., J.K., J.P., H.M.-L., J.A.R. and F.S.; writing—review and editing, C.R., J.K., J.P., H.M.-L., J.A.R. and F.S.; visualization, C.R., J.K. and F.S.; supervision, J.K., F.S.; project administration, C.R., J.K., J.P., H.M.-L., J.A.R. and F.S.; funding acquisition, J.K., J.P., H.M.-L., J.A.R. and F.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft—DFG) [project number 445925495].

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Charité Berlin (EA4/080/18) approve date: 6 June 2018.

Informed Consent Statement: Given anonymized data being used, informed consent was not needed according to Berlin Krankenhausgesetz.

Data Availability Statement: The data are not publicly available due to data protection reasons.

Conflicts of Interest: F.S. and J.K. are co-founders of the dentalXrai GmbH, a startup. dentalXrai GmbH did not have any role in conceiving, conducting, or reporting this study.

References

1. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **2016**, *316*, 2402–2410. [[CrossRef](#)] [[PubMed](#)]
2. Becker, A.S.; Marcon, M.; Ghafoor, S.; Wurnig, M.C.; Frauenfelder, T.; Boss, A. deep learning in mammography: Diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Investig. Radiol.* **2017**, *52*, 434–440. [[CrossRef](#)]
3. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [[CrossRef](#)]
4. Schwendicke, F.; Golla, T.; Dreher, M.; Krois, J. Convolutional neural networks for dental image diagnostics: A scoping review. *J. Dent.* **2019**, *91*, 103226. [[CrossRef](#)] [[PubMed](#)]

5. Minaee, S.; Boykov, Y.Y.; Porikli, F.; Plaza, A.J.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *1*. [[CrossRef](#)] [[PubMed](#)]
6. Başaran, M.; Çelik, Ö.; Bayrakdar, I.S.; Bilgir, E.; Orhan, K.; Odabas, A.; Aslan, O.F.; Jagtap, R. Diagnostic charting of panoramic radiography using deep-learning artificial intelligence system. *Oral Radiol.* **2021**. [[CrossRef](#)]
7. Takahashi, T.; Nozaki, K.; Gonda, T.; Mamenno, T.; Ikebe, K. Deep learning-based detection of dental prostheses and restorations. *Sci. Rep.* **2021**, *11*, 1960. [[CrossRef](#)] [[PubMed](#)]
8. Zaidi, S.S.A.; Ansari, M.S.; Aslam, A.; Kanwal, N.; Asghar, M.; Lee, B. A survey of modern deep learning based object detection models. *Digit. Sign. Process.* **2022**, 103514. [[CrossRef](#)]
9. Mongan, J.; Moy, L.; Kahn, C.E. Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. *Radiol. Artif. Intell.* **2020**, *2*, e200029. [[CrossRef](#)]
10. Schwendicke, F.; Singh, T.; Lee, J.H.; Gaudin, R.; Chaurasia, A.; Wiegand, T.; Uribe, S.; Krois, J. Artificial intelligence in dental research: Checklist for authors, reviewers, readers. *J. Dent.* **2021**, *107*, 103610. [[CrossRef](#)] [[PubMed](#)]
11. Ekert, T.; Krois, J.; Schwendicke, F. *Building a Mass Online Annotation Tool for Dental Radiographic Imagery*; Universitätsmedizin Berlin: Berlin, Germany, 2018.
12. Ronneberger, O.; Fischer, P.; Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
13. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2017; pp. 5987–5995.
14. Deng, J.; Dong, W.; Socher, R.; Li, L.; Kai, L.; Li, F.-F. *ImageNet: A Large-Scale Hierarchical Image Database*; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.