Diffusion-Based Hierarchical Multi-label Object Detection to Analyze Panoramic Dental X-Rays

Ibrahim Ethem Hamamci[1(✉)], Sezgin Er[2], Enis Simsar[3], Anjany Sekuboyina[1], Mustafa Gundogar[4], Bernd Stadlinger[5], Albert Mehl[5], and Bjoern Menze[1]

[1] Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland
ibrahim.hamamci@uzh.ch
[2] International School of Medicine, Istanbul Medipol University, Istanbul, Turkey
[3] Department of Computer Science, ETH Zurich, Zurich, Switzerland
[4] Department of Endodontics, Istanbul Medipol University, Istanbul, Turkey
[5] Center of Dental Medicine, University of Zurich, Zurich, Switzerland

Abstract. Due to the necessity for precise treatment planning, the use of panoramic X-rays to identify different dental diseases has tremendously increased. Although numerous ML models have been developed for the interpretation of panoramic X-rays, there has not been an end-to-end model developed that can identify problematic teeth with dental enumeration and associated diagnoses at the same time. To develop such a model, we structure the three distinct types of annotated data hierarchically following the FDI system, the first labeled with only quadrant, the second labeled with quadrant-enumeration, and the third fully labeled with quadrant-enumeration-diagnosis. To learn from all three hierarchies jointly, we introduce a novel diffusion-based hierarchical multi-label object detection framework by adapting a diffusion-based method that formulates object detection as a denoising diffusion process from noisy boxes to object boxes. Specifically, to take advantage of the hierarchically annotated data, our method utilizes a novel noisy box manipulation technique by adapting the denoising process in the diffusion network with the inference from the previously trained model in hierarchical order. We also utilize a multi-label object detection method to learn efficiently from partial annotations and to give all the needed information about each abnormal tooth for treatment planning. Experimental results show that our method significantly outperforms state-of-the-art object detection methods, including RetinaNet, Faster R-CNN, DETR, and DiffusionDet for the analysis of panoramic X-rays, demonstrating the great potential of our method for hierarchically and partially annotated datasets. The code and the datasets are available at https://github.com/ibrahimethemhamamci/HierarchicalDet.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43987-2_38.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 389–399, 2023.
https://doi.org/10.1007/978-3-031-43987-2_38

Keywords: Diffusion Network · Hierarchical Learning · Multi-Label Object Detection · Panoramic Dental X-ray · Transformers

1 Introduction

The use of panoramic X-rays to diagnose numerous dental diseases has increased exponentially due to the demand for precise treatment planning [11]. However, visual interpretation of panoramic X-rays may consume a significant amount of essential clinical time [2] and interpreters may not always have dedicated training in reading scans as specialized radiologists have [13]. Thus, the diagnostic process can be automatized and enhanced by getting the help of Machine Learning (ML) models. For instance, an ML model that automatically detects abnormal teeth with dental enumeration and associated diagnoses would provide a tremendous advantage for dentists in making decisions quickly and saving their time.

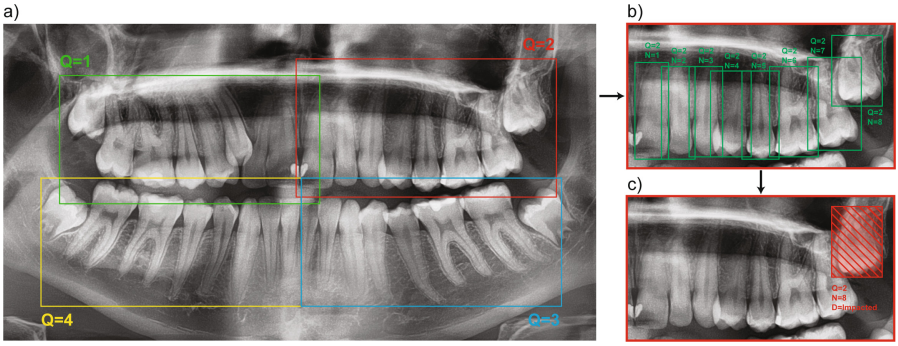

Fig. 1. The annotated datasets are organized hierarchically as (a) quadrant-only, (b) quadrant-enumeration, and (c) quadrant-enumeration-diagnosis respectively.

Many ML models to interpret panoramic X-rays have been developed specifically for individual tasks such as quadrant segmentation [19,29], tooth detection [6], dental enumeration [14,23], diagnosis of some abnormalities [12,30], as well as treatment planning [27]. Although many of these studies have achieved good results, three main issues still remain. *(1) Multi-label detection:* there has not been an end-to-end model developed that gives all the necessary information for treatment planning by detecting abnormal teeth with dental enumeration and multiple diagnoses simultaneously [1]. *(2) Data availability:* to train a model that performs this task with high accuracy, a large set of fully annotated data is needed [13]. Because labeling every tooth with all required classes may require expertise and take a long time, such kind of fully labeled large datasets do not always exist [24]. For instance, we structure three different available annotated data hierarchically shown in Fig. 1, using the Fédération Dentaire Internationale (FDI) system. The first data is partially labeled because it only included

quadrant information. The second data is also partially labeled but contains additional enumeration information along with the quadrant. The third data is fully labeled because it includes all quadrant-enumeration-diagnosis information for each abnormal tooth. Thus, conventional object detection algorithms would not be well applicable to this kind of hierarchically and partially annotated data [21]. *(3) Model performance:* to the best of our knowledge, models designed to detect multiple diagnoses on panoramic X-rays have not achieved the same high level of accuracy as those specifically designed for individual tasks, such as tooth detection, dental enumeration, or detecting single abnormalities [18].

To circumvent the limitations of the existing methods, we propose a novel diffusion-based hierarchical multi-label object detection method to point out each abnormal tooth with dental enumeration and associated diagnosis concurrently on panoramic X-rays, see Fig. 2. Due to the partial annotated and hierarchical characteristics of our data, we adapt a diffusion-based method [5] that formulates object detection as a denoising diffusion process from noisy boxes to object boxes. Compared to the previous object detection methods that utilize conventional weight transfer [3] or cropping strategies [22] for hierarchical learning, the denoising process enables us to propose a novel hierarchical diffusion network by utilizing the inference from the previously trained model in hierarchical order to manipulate the noisy bounding boxes as in Fig. 2. Besides, instead of pseudo labeling techniques [28] for partially annotated data, we develop a multi-label object detection method to learn efficiently from partial annotations and to give all the needed information about each abnormal tooth for treatment planning. Finally, we demonstrate the effectiveness of our multi-label detection method on partially annotated data and the efficacy of our proposed bounding box manipulation technique in diffusion networks for hierarchical data.

The contributions of our work are three-fold. (1) We propose a multi-label detector to learn efficiently from partial annotations and to detect the abnormal tooth with all three necessary classes, as shown in Fig 3 for treatment planning. (2) We rely on the denoising process of diffusion models [5] and frame the detection problem as a hierarchical learning task by proposing a novel bounding box manipulation technique that outperforms conventional weight transfer as shown in Fig. 4. (3) Experimental results show that our model with bounding box manipulation and multi-label detection significantly outperforms state-of-the-art object detection methods on panoramic X-ray analysis, as shown in Table 1.

2 Methods

Figure 2 illustrates our proposed framework. We utilize the DiffusionDet [5] model, which formulates object detection as a denoising diffusion process from noisy boxes to object boxes. Unlike other state-of-the-art detection models, the denoising property of the model enables us to propose a novel manipulation technique to utilize a hierarchical learning architecture by using previously inferred boxes. Besides, to learn efficiently from partial annotations, we design a multi-label detector with adaptable classification layers based on available labels. In

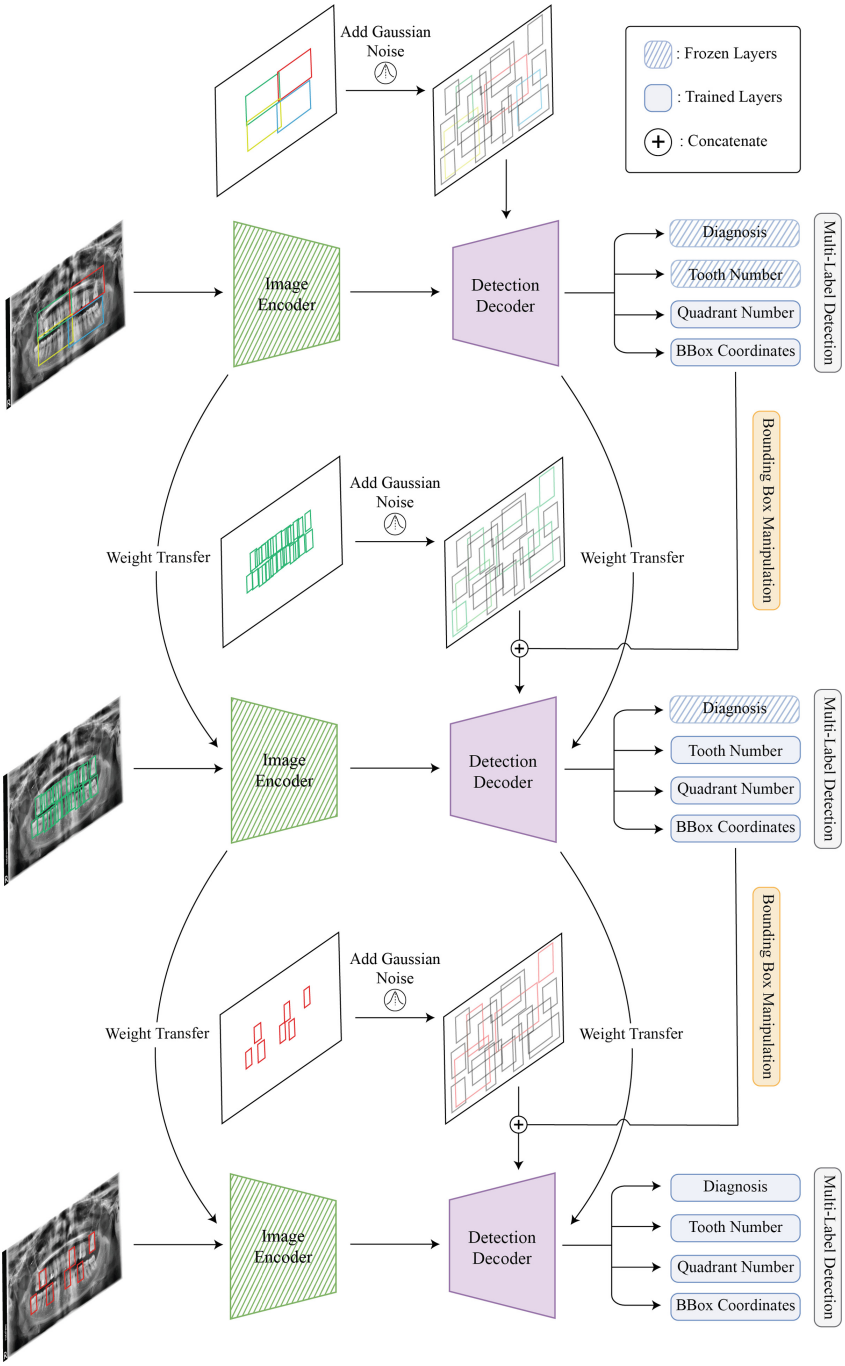

Fig. 2. Our method relies on a hierarchical learning approach utilizing a combination of multi-label detection, bounding box manipulation, and weight transfer.

addition, we designed our approach to serve as a foundational baseline for the Dental Enumeration and Diagnosis on Panoramic X-rays Challenge (DENTEX), set to take place at MICCAI 2023. Remarkably, the data and annotations we utilized for our method mirror exactly those employed for DENTEX [9].

2.1 Base Model

Our method employs the DiffusionDet [5] that comprises two essential components, an image encoder that extracts high-level features from the raw image and a detection decoder that refines the box predictions from the noisy boxes using those features. The set of initial noisy bounding boxes is defined as:

$$q(z_t|z_0) = \mathcal{N}(z_t|\sqrt{\bar{\alpha}_t}z_0, (1-\bar{\alpha}_t)I) \tag{1}$$

where z_0 represents the input bounding box b, and $b \in \mathbb{R}^{N\times 4}$ is a set of bounding boxes, z_t represents the latent noisy boxes, and $\bar{\alpha}_t$ represents the noise variance schedule. The DiffusionDet model [5] $f_\theta(z_t, t, x)$, is trained to predict the final bounding boxes defined as $b^i = (c_x^i, c_y^i, w^i, h^i)$ where (c_x^i, c_y^i) are the center coordinates of the bounding box and (w^i, h^i) are the width and height of the bounding boxes and category labels defined as y^i for objects.

2.2 Proposed Framework

To improve computational efficiency during the denoising process, DiffusionDet [5] is divided into two parts: an image encoder and a detection decoder. Iterative denoising is applied only for the detection decoder, using the outputs of the image encoder as a condition. Our method employs this approach with several adjustments, including multi-label detection and bounding box manipulation. Finally, we utilize conventional transfer learning for comparison.

Image Encoder. Our method utilizes a Swin-transformer [17] backbone pretrained on the ImageNet-22k [7] with a Feature Pyramid Network (FPN) architecture [15] as it was shown to outperform convolutional neural network-based models such as ResNet50 [10]. We also apply pre-training to the image encoder using our unlabeled data, as it is not trained during the training process. We utilize SimMIM [26] that uses masked image modeling to finetune the encoder.

Detection Decoder. Our method employs a detection decoder that inputs noisy initial boxes to extract Region of Interest (RoI) features from the encoder-generated feature map and predicts box coordinates and classifications using a detection head. However, our detection decoder has several differences from DiffusionDet [5]. Our proposed detection decoder (1) has three classification heads instead of one, which allows us to train the same model with partially annotated data by freezing the heads according to the unlabeled classes, (2) employs manipulated bounding boxes to extract RoI features, and (3) leverages transfer learning from previous training steps.

Multi-label Detection. We utilize three classification heads as quadrant-enumeration-diagnosis for each bounding box and freeze the heads for the unlabeled classes, shown in Fig. 2. Our model denoted by f_θ is trained to predict:

$$f_\theta(z_t, t, x, h_q, h_e, h_d) = \begin{cases} (y_q^i, b^i), & h_q = 1, h_e = 0, h_d = 0 \quad (a) \\ (y_q^i, y_e^i, b^i), & h_q = 1, h_e = 1, h_d = 0 \quad (b) \\ (y_q^i, y_e^i, y_d^i, b^i), & h_q = 1, h_e = 1, h_d = 1 \quad (c) \end{cases} \tag{2}$$

where y_q^i, y_e^i, and y_d^i represent the bounding box classifications for quadrant, enumeration, and diagnosis, respectively, and h_q, h_e, and h_d represent binary indicators of whether the labels are present in the training dataset. By adapting this approach, we leverage the full range of available information and improve our ability to handle partially labeled data. This stands in contrast to conventional object detection methods, which rely on a single classification head for each bounding box [25] and may not capture the full complexity of the underlying data. Besides, this approach enables the model to detect abnormal teeth with all three necessary classes for clinicians to plan the treatment, as seen in Fig. 3.

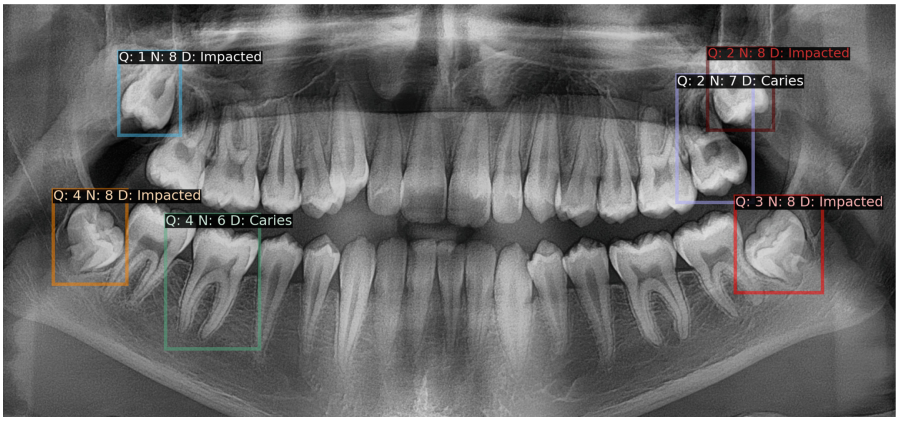

Fig. 3. Output from our final model showing well-defined boxes for diseased teeth with corresponding quadrant (Q), enumeration (N), and diagnosis (D) labels.

Bounding Box Manipulation. Instead of completely noisy boxes, we use manipulated bounding boxes to extract RoI features from the encoder-generated feature map and to learn efficiently from hierarchical annotations as shown in Fig. 2. Specifically, to train the model (b) in Eq. (2), we concatenate the noisy boxes described in Eq. (1) with the boxes inferred from the model (a) in Eq. (2) with a score greater than 0.5. Similarly, we manipulate the denoising process during the training of the model (c) in Eq. (2) by concatenating the noisy boxes

with boxes inferred from the model (b) in Eq. (2) with a score greater than 0.5. The set of manipulated boxes b_m, and $b_m \in \mathbb{R}^{N\times 4}$, can be defined as $b_m = [b_n[:-k], b_i]$, where b_n, and $b_n \in \mathbb{R}^{N\times 4}$, represents the set of noisy boxes and, b_i, and $b_i \in \mathbb{R}^{k\times 4}$, represents the set of inferred boxes from the previous training. Our framework utilizes completely noisy boxes during the inference.

3 Experiments and Results

We evaluate models' performances using a combination of Average Recall (AR) and Average Precision (AP) scores with various Intersection over Union (IoU) thresholds. This included $AP_{[0.5,0.95]}$, AP_{50}, AP_{75}, and separate AP scores for large objects (AP_l), and medium objects (AP_m).

Data. All panoramic X-rays were acquired from patients above 12 years of age using the VistaPano S X-ray unit (Durr Dental, Germany). To ensure patient privacy and confidentiality, panoramic X-rays were randomly selected from the hospital's database without considering any personal information.

To effectively utilize FDI system [8], three distinct types of data are organized hierarchically as in Fig. 1 (a) 693 X-rays labeled only for quadrant detection, (b) 634 X-rays labeled for tooth detection with both quadrant and tooth enumeration classifications, and (c) 1005 X-rays fully labeled for diseased tooth detection with quadrant, tooth enumeration, and diagnosis classifications. In the diagnosis, there are four specific classes corresponding to four different diagnoses: caries, deep caries, periapical lesions, and impacted teeth. The remaining 1571 unlabeled X-rays are used for pre-training. All necessary permissions were obtained from the ethics committee.

Experimental Design. To evaluate our proposed method, we conduct two experiments: (1) Comparison with state-of-the-art object detection models, including DETR [4], Faster R-CNN [20], RetinaNet [16], and DiffusionDet [5] in Table 1. (2) A comprehensive ablation study to assess the effect of our modifications to DiffusionDet in hierarchical detection performance in Fig. 4.

Evaluation. Fig. 3 presents the output prediction of the final trained model. As depicted in the figure, the model effectively assigns three distinct classes to each well-defined bounding box. Our approach that utilizes novel box manipulation and multi-label detection, significantly outperforms state-of-the-art methods. The box manipulation approach specifically leads to significantly higher AP and AR scores compared to other state-of-the-art methods, including RetinaNet, Faster-R-CNN, DETR, and DiffusionDet. Although the impact of conventional transfer learning on these scores can vary depending on the data, our bounding box manipulation outperforms it. Specifically, the bounding box manipulation approach is the sole factor that improves the accuracy of the model, while weight transfer does not improve the overall accuracy, as shown in Fig. 4.

Table 1. Our method outperforms state-of-the-art methods, and our bounding box manipulation approach outperforms the weight transfer. Results shown here indicate the different tasks in the test set which is multi-labeled (quadrant-enumeration-diagnosis) for abnormal tooth detection.

Method	AR	AP	AP_{50}	AP_{75}	AP_m	AP_l
Quadrant						
RetinaNet[16]	0.604	25.1	41.7	28.8	32.9	25.1
Faster R-CNN[20]	0.588	29.5	48.6	33.0	39.9	29.5
DETR[4]	0.659	39.1	60.5	47.6	55.0	39.1
Base (DiffusionDet)[5]	0.677	38.8	60.7	46.1	39.1	39.0
Ours w/o Transfer	0.699	42.7	64.7	**52.4**	50.5	42.8
Ours w/o Manipulation	**0.727**	40.0	60.7	48.2	59.3	40.0
Ours w/o Manipulation and Transfer	0.658	38.1	60.1	45.3	45.1	38.1
Ours (Manipulation+Transfer+Multilabel)	0.717	**43.2**	**65.1**	51.0	**68.3**	**43.1**
Enumeration						
RetinaNet[16]	0.560	25.4	41.5	28.5	55.1	25.2
Faster R-CNN[20]	0.496	25.6	43.7	27.0	53.3	25.2
DETR[4]	0.440	23.1	37.3	26.6	43.4	23.0
Base (DiffusionDet)[5]	0.617	29.9	47.4	34.2	48.6	29.7
Ours w/o Transfer	0.648	**32.8**	**49.4**	**39.4**	**60.1**	**32.9**
Ours w/o Manipulation	0.662	30.4	46.5	36.6	58.4	30.5
Ours w/o Manipulation and Transfer	0.557	26.8	42.4	29.5	51.4	26.5
Ours (Manipulation+Transfer+Multilabel)	**0.668**	30.5	47.6	37.1	51.8	30.4
Diagnosis						
RetinaNet[16]	0.587	32.5	54.2	35.6	41.7	32.5
Faster R-CNN[20]	0.533	33.2	54.3	38.0	24.2	33.3
DETR[4]	0.514	33.4	52.8	41.7	48.3	33.4
Base (DiffusionDet)[5]	0.644	37.0	58.1	42.6	31.8	37.2
Ours w/o Transfer	0.669	**39.4**	**61.3**	**47.9**	**49.7**	**39.5**
Ours w/o Manipulation	0.688	36.3	55.5	43.1	45.6	37.4
Ours w/o Manipulation and Transfer	0.648	37.3	59.5	42.8	33.6	36.4
Ours (Manipulation+Transfer+Multilabel)	**0.691**	37.6	60.2	44.0	36.0	37.7

Ablation Study. Our ablation study results, shown in Fig. 4 and Table 1, indicate that our approaches have a synergistic impact on the detection model's accuracy, with the highest increase seen through bounding box manipulation. We systematically remove every combination of bounding box manipulation and weight transfer, to demonstrate the efficacy of our methodology. Conventional transfer learning does not positively affect the models' performances compared to the bounding box manipulation, especially for enumeration and diagnosis.

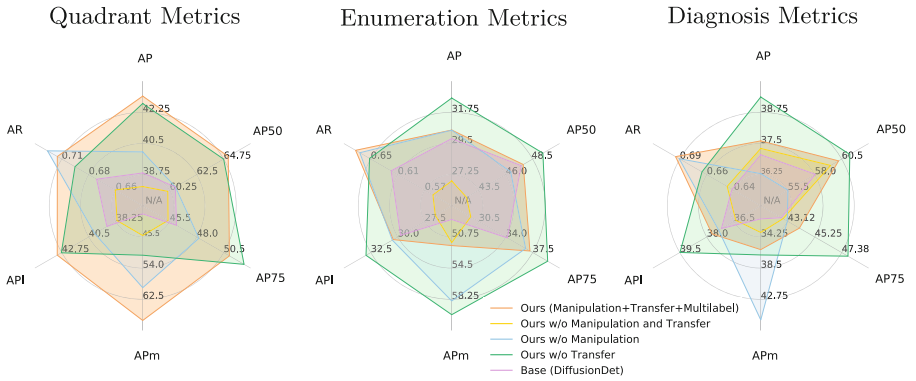

Fig. 4. The results of the ablation study reveals that our bounding box manipulation method outperforms conventional weight transfer.

4 Discussion and Conclusion

In this paper, we introduce a novel diffusion-based multi-label object detection framework to overcome one of the significant obstacles to the clinical application of ML models for medical and dental diagnosis, which is the difficulty in getting a large volume of fully labeled data. Specifically, we propose a novel bounding box manipulation technique during the denoising process of the diffusion networks with the inference from the previously trained model to take advantage of hierarchical data. Moreover, we utilize a multi-label detector to learn efficiently from partial annotations and to assign all necessary classes to each box for treatment planning. Our framework outperforms state-of-the-art object detection models for training with hierarchical and partially annotated panoramic X-ray data.

From the clinical perspective, we develop a novel framework that simultaneously points out abnormal teeth with dental enumeration and associated diagnosis on panoramic dental X-rays with the help of our novel diffusion-based hierarchical multi-label object detection method. With some limits due to partially annotated and limited amount of data, our model that provides three necessary classes for treatment planning has a wide range of applications in the real world, from being a clinical decision support system to being a guide for dentistry students.

Acknowledgements. We would like to thank the Helmut Horten Foundation for supporting our research.

References

1. AbuSalim, S., Zakaria, N., Islam, M.R., Kumar, G., Mokhtar, N., Abdulkadir, S.J.: Analysis of deep learning techniques for dental informatics: a systematic literature review. Healthcare (Basel) **10**(10), 1892 (2022)
2. Bruno, M.A., Walker, E.A., Abujudeh, H.H.: Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics **35**(6), 1668–1676 (2015)
3. Bu, X., Peng, J., Yan, J., Tan, T., Zhang, Z.: GAIA: a transfer learning system of object detection that fits your needs. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 274–283 (2021)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
5. Chen, S., Sun, P., Song, Y., Luo, P.: DiffusionDet: diffusion model for object detection. arXiv preprint arXiv:2211.09788 (2022)
6. Chung, M., et al.: Individual tooth detection and identification from dental panoramic X-ray images via point-wise localization and distance regularization. Artif. Intell. Med. **111**, 101996 (2021)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, pp. 248–255 (2009)
8. Glick, M., et al.: FDI vision 2020: shaping the future of oral health. Int. Dent. J. **62**(6), 278 (2012)
9. Hamamci, I.E., et al.: DENTEX: an abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic X-rays. arXiv preprint arXiv:2305.19112 (2023)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arxiv 2015. arXiv preprint arXiv:1512.03385 (2015)
11. Hwang, J.J., Jung, Y.H., Cho, B.H., Heo, M.S.: An overview of deep learning in the field of dentistry. Imaging Sci. Dent. **49**(1), 1–7 (2019)
12. Krois, J.: Deep learning for the radiographic detection of periodontal bone loss. Sci. Rep. **9**(1), 8495 (2019)
13. Kumar, A., Bhadauria, H.S., Singh, A.: Descriptive analysis of dental X-ray images using various practical methods: a review. PeerJ Comput. Sci. **7**, e620 (2021)
14. Lin, S.Y., Chang, H.Y.: Tooth numbering and condition recognition on dental panoramic radiograph images using CNNs. IEEE Access **9**, 166008–166026 (2021)
15. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 2117–2125 (2017)
16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017)
17. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
18. Panetta, K., Rajendran, R., Ramesh, A., Rao, S.P., Agaian, S.: Tufts dental database: a multimodal panoramic X-ray dataset for benchmarking diagnostic systems. IEEE J. Biomed. Health Inform. **26**(4), 1650–1659 (2021)

19. Pati, S., et al.: GaNDLF: a generally nuanced deep learning framework for scalable end-to-end clinical workflows in medical imaging. arXiv preprint arXiv:2103.01006 (2021)
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, vol. 28 (2015)
21. Shin, S.J., Kim, S., Kim, Y., Kim, S.: Hierarchical multi-label object detection framework for remote sensing images. Remote Sens. **12**(17), 2734 (2020)
22. Shin, S.J., Kim, S., Kim, Y., Kim, S.: Hierarchical multi-label object detection framework for remote sensing images. Remote Sens. **12**(17), 2734 (2020)
23. Tuzoff, D.V., et al.: Tooth detection and numbering in panoramic radiographs using convolutional neural networks. Dentomaxillofacial Radiol. **48**(4), 20180051 (2019)
24. Willemink, M.J., et al.: Preparing medical imaging data for machine learning. Radiology **295**(1), 4–15 (2020)
25. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2 (2019)
26. Xie, Z., et al.: SimMIM: a simple framework for masked image modeling. arXiv preprint arXiv:2111.09886 (2021)
27. Yüksel, A.E., et al.: Dental enumeration and multiple treatment detection on panoramic X-rays using deep learning. Sci. Rep. **11**(1), 1–10 (2021)
28. Zhao, X., Schulter, S., Sharma, G., Tsai, Y.-H., Chandraker, M., Wu, Y.: Object detection with a unified label space from multiple datasets. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12359, pp. 178–193. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58568-6_11
29. Zhao, Y., et al.: TsasNet: tooth segmentation on dental panoramic X-ray images by two-stage attention segmentation network. Knowl.-Based Syst. **206**, 106338 (2020)
30. 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 Comput. Appl. **35**(22), 16051–16059 (2023). https://doi.org/10.1007/s00521-021-06684-2