TeethDE-GCNet: accurate tooth numbering in dental panoramic radiographs via detail-enhanced and context-aware detection

Yongming Xie[a], Xin Huang[b,d,*], Wei Deng[c], Weixia Xu[c], Xiao Li[b], Zhongming Liu[b]
[a]Digital Industry College, Jiangxi Normal University, Shangrao, Jiangxi, 334000, China
[b]Software College, Jiangxi Normal University, Nanchang, Jiangxi, 330001, China
[c]Department of Stomatology, The Third Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, 330006 China
[d]Jiangxi Provincial Engineering Research Center of Blockchain Data Security and Governance, Nanchang, Jiangxi, 330031, China
*xinhuang@jxnu.edu.cn

ABSTRACT

Automatic labeling of teeth in oral panoramic films is crucial for clinical diagnosis and treatment. In this paper, we propose TeethDE-GCNet model, introduce detail enhancement and context-aware module based on YOLOv8, and optimize the effect of small-target detection by combining Focaler-IoU loss and small-target detection layer. The experiments are conducted on Tufts public dataset and self-constructed clinical dataset with FDI standard annotation. The results show that the proposed method outperforms the existing methods in metrics such as mAP50 and mAP50:95, and exhibits strong practicality and robustness.

Keywords: Automatic Tooth Numbering, Dental Panoramic Radiograph, Small Object Detection

1. INTRODUCTION

Oral health is an important component of overall health, and the accuracy of its diagnosis and treatment directly affects patients' quality of life and disease progression [1]. According to the 2019 Global Burden of Disease Study report, over 3.5 billion people worldwide suffer from oral diseases, and dental infections can even increase the risk of systemic diseases by 2.3 to 4.7 times [2][3]. The precise localization and numbering of teeth are not only key steps in disease diagnosis and treatment plan formulation, but also play a significant role in improving treatment efficiency, standardizing medical procedures, and reducing human error.

Panoramic X-ray images are commonly used as dental imaging tools. However, factors such as complex dental structures, significant individual differences, and image artifacts can lead to misinterpretation when relying on manual analysis, severely limiting the quality of diagnosis and treatment. In recent years, the widespread application of artificial intelligence technologies such as deep learning in medical imaging has provided new avenues for the task of automatic tooth numbering. [4][5].

To address the limitations of existing methods in terms of performance on small-scale structures, blurred boundaries, and complex backgrounds, this paper proposes an improved object detection network, TeethDE-GCNet, which incorporates multi-scale feature enhancement and context modeling strategies based on the YOLOv8 framework to enhance the accuracy and robustness of tooth detection. We conducted a systematic evaluation on the Tufts public dataset [6] and real clinical panoramic image data. The experimental results validate the significant superiority of the proposed method across multiple metrics.

2. RELATED WORK

In recent years, deep learning technology has been widely applied in the field of medical image analysis, particularly demonstrating outstanding performance in tasks such as object detection, image segmentation, and image classification [7][8]. In the field of dental image analysis, researchers have proposed various methods for tooth number identification [5]. Traditional image processing methods primarily rely on techniques such as image enhancement, edge detection, and region

Fifth International Conference on Biomedicine and Bioinformatics Engineering (ICBBE 2025),
edited by Muhammad Zubair Iqbal, Lun Hu, Yunfeng Qiu, Proc. of SPIE Vol. 13971,
139710E · © 2025 SPIE · 0277-786X · doi: 10.1117/12.3095504

growing. However, when dealing with panoramic X-ray images characterized by complex structures and severe artifact interference, these methods often lack sufficient robustness and generalization capabilities.

With the development of the YOLO series of algorithms, their efficient end-to-end detection capabilities have been widely applied to medical object detection tasks. Kaya et al. [9] were the first to propose the use of the YOLOv4 model for the identification and numbering of deciduous and permanent teeth in panoramic images, analyzing 4,545 panoramic images to validate the model's advantages in detection accuracy and inference speed. However, the original YOLO architecture still faces challenges in detecting complex cases such as abnormal tooth structures and small targets (e.g., root remnants, cavities), resulting in insufficient detection accuracy. Subsequently, Beşer et al. [10] employed the YOLOv5 network for automatic detection, segmentation, and numbering of panoramic images of children's mixed dentition, achieving precise identification of deciduous and permanent teeth, and verifying the model's high sensitivity and accuracy. Due to the variability of tooth morphology (such as missing teeth, cavities, implants, and fixed bridges) [11][12], existing models still struggle to achieve robust and precise automatic tooth numbering.

In addition, the conventional IoU loss function is difficult to effectively optimize for hard-to-detect targets in scenarios where the difficulty of samples varies greatly, further limiting the improvement of model performance. Therefore, improving the robustness of the model in complex structure recognition, small target localization, and sample imbalance conditions remains a key issue in current automatic tooth numbering research.

3. METHOD

This paper proposes an automatic tooth numbering model based on the YOLOv8 object detection framework, which integrates multi-scale differential convolutions and dynamic IoU optimization strategies. The overall structure of the model is shown in Figure 1. To address the issues of complex tooth structures and difficult recognition of small objects in panoramic oral images, several improvements have been made to the model structure to enhance detection accuracy and robustness.

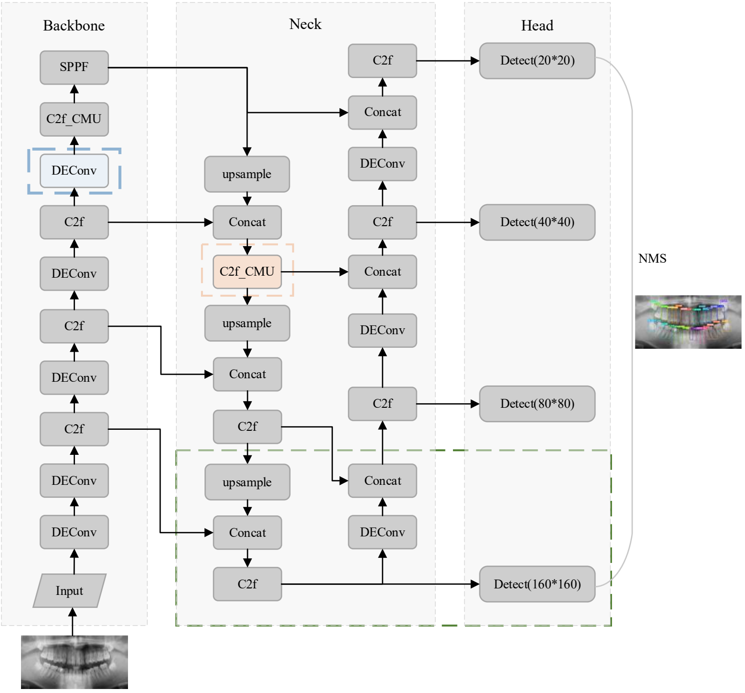

Figure 1. Overall Architecture of the TeethDE-GCNet Network.

First, the basic convolutional modules in the original YOLOv8 are replaced with the proposed Detail Enhancement Convolution (DEConv) to strengthen the representation of critical details such as tooth edges and contours, thereby improving the perception of low-level targets. Second, a context fusion module, C2f_CMU, constructed using large kernel convolutions and an inverse bottleneck structure, is introduced to model long-range spatial dependencies and enhance the understanding of complex anatomical structures.

To improve the detection of small-scale targets such as caries and residual roots, an additional detection branch with a resolution of 160×160 is incorporated into the detection head. This branch fuses shallow positional features with deep semantic information to improve detection accuracy and reduce the missed detection rate. To improve the detection of small-scale targets such as caries and residual roots, an additional detection branch with a resolution of 160×160 is incorporated into the detection head. This branch fuses shallow positional features with deep semantic information to improve detection accuracy and reduce the missed detection rate.

3.1. Detail-enhanced convolution module

In traditional image denoising tasks, existing methods typically employ Vanilla Convolution (VC) layers for feature extraction and learning [13][14]. In traditional image denoising tasks, existing methods typically employ Vanilla Convolution (VC) layers for feature extraction and learning.

To this end, this study introduces a Detail Enhancement Convolutional Module (DEConv) [15], as illustrated in Figure 2. This module incorporates pixel-level gradient difference modeling into the convolution process, embedding prior knowledge to enhance the model's sensitivity to fine-grained features. Five convolutional branches are deployed in parallel, including one vanilla convolution branch and four differential convolution branches: Angular Difference Convolution (ADC), Central Difference Convolution (CDC), Vertical Difference Convolution (VDC), and Horizontal Difference Convolution (HDC). The vanilla convolution captures intensity-level information, while the differential convolutions are designed to enhance gradient-level features. The formulation is as follows:

$$F_{out} = \mathrm{DEConv}(F_{in}) = \sum_{i=1}^{5} F_{in} * K_i = F_{in} * \left(\sum_{i=1}^{5} K_i\right) = F_{in} * K_{cvt} \tag{1}$$

$DEConv(*)$ denotes the operation of the DEConv module; $K_{i=1:5}$ represents the kernels of VC, ADC, CDC, HDC, and VDC, respectively; $*$denotes the convolution operation; and K_{cvt} denotes the transformed kernel that integrates the outputs of the parallel convolutions into a unified representation.

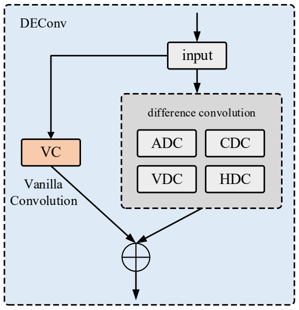

Figure 2. DEConv Module Architecture.

3.2. Global context-aware module

To enhance the model's capability in representing complex dental structures, the CMUNeXtBlock module is introduced to replace the original C2f structure in YOLOv8. As illustrated in Figure 3, this module is built upon large-kernel depthwise separable convolution, which enables the modeling of global spatial dependencies while maintaining a lightweight architecture. This design effectively mitigates the limitations of conventional convolutional networks in capturing long-range contextual information [16].

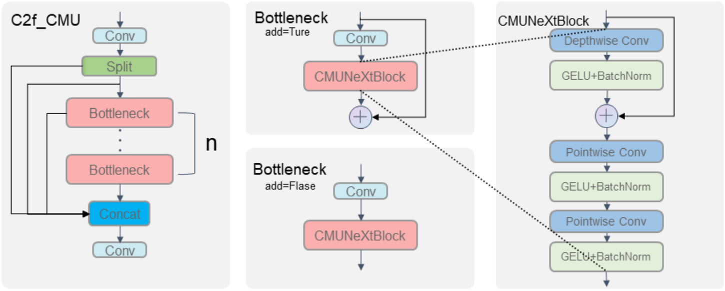

Figure 3. C2f_CMU Module Architecture.

The CMUNeXtBlock is embedded within the Bottleneck module of YOLOv8, which consists of a standard convolutional layer followed by the CMUNeXtBlock. The presence of a residual connection is determined by the add parameter: when set to True, a residual branch is incorporated; otherwise, feature concatenation is applied [17].

3.3. Focaler-IoU loss function

The baseline model adopts Complete IoU (CIoU) as the bounding box regression loss function. However, CIoU exhibits ambiguity when handling aspect ratios and fails to account for the varying difficulty of samples. To address this limitation, this study introduces Focaler-IoU in conjunction with CIoU to enhance regression performance [18]. Considering the issue of sample imbalance in object detection tasks, samples can be categorized into easily detectable and hard-to-detect cases. Conventional objects generally fall into the former category, whereas small objects, due to localization challenges, are typically hard-to-detect samples. In tasks dominated by small targets, Focaler-IoU effectively increases focus on difficult samples, thereby improving overall detection accuracy. The loss is reformulated using a linear interval mapping strategy to better reflect bounding box quality. The formulation is as follows:

$$IoU^{focaler} = \{ \begin{matrix} 0, & IoU < d \\ \frac{IoU-d}{u-d}, & d \ll IoU \ll u \\ 1, & IoU > u \end{matrix} \tag{2}$$

The Focaler-IoU loss is incorporated into the IoU-based bounding box regression framework, and the Focaler-CIoU loss, denoted as L_"Focaler-CIoU", is defined as follows:

$$L_{Focaler\text{-}CIoU} = L_{CIoU} + IoU - IoU^{fuculer} \tag{3}$$

3.4. Optimization of small object detection head

The baseline model suffers from the loss of small object information during the sampling process, as well as inaccuracies in the localization and recognition of small objects within deep feature maps. To address these issues, this study introduces an additional small object detection layer with a resolution of 160×160 at the model's head, dedicated specifically to small object detection. This newly added layer enables more refined target discrimination and allocates greater focus to small objects. By integrating shallow positional features with deep semantic information, the layer achieves more precise localization and identification of small objects, thereby effectively reducing missed detections. The architecture of the small object detection layer is illustrated in Figure 4.

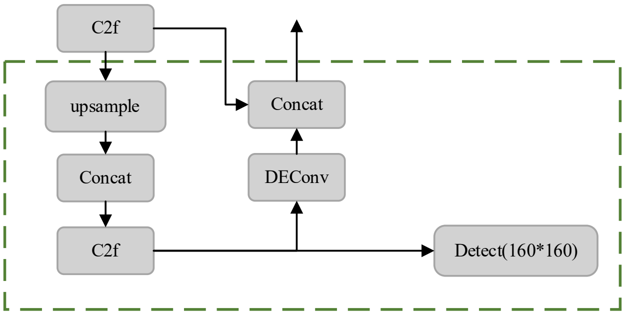

Figure 4. small object detection layer.

4. EXPERIMENTAL DATA AND RESULTS

4.1. Data and basic settings

To comprehensively validate the effectiveness of the proposed method, this study not only conducted experiments on a publicly available dental segmentation dataset but also constructed and introduced a novel oral imaging dataset.

4.1.1. The tufts dental database

This multimodal panoramic oral X-ray dataset is designed to serve as a benchmark for dental anomaly detection and segmentation tasks. The dataset includes expert annotations of both normal and abnormal teeth. The original images have a resolution of 1615 × 840 pixels. To accommodate model input requirements and reduce computational overhead, the images were resized and cropped uniformly to 720 × 360 pixels. Dental position labels were automatically extracted via scripting. The entire dataset comprises 1,000 images, which are partitioned into training and testing sets at a 9:1 ratio, containing 900 and 100 images respectively [6].

4.1.2. Our database

This study, conducted in collaboration with the Department of Stomatology at a tertiary hospital in Nanchang, has established a novel clinical panoramic oral image dataset tailored for tooth detection and segmentation tasks. The data were collected from 500 patients aged between 18 and 65 years. Images are stored in BMP format, and annotations were meticulously performed by a professional dental team using the LabelImg tool, applying the FDI tooth numbering system to precisely label each tooth region. To reduce computational complexity and mitigate interference from metal artifacts, all images were uniformly resized and cropped to 720 × 360 pixels. The final dataset comprises 500 annotated images, partitioned into training and testing sets at a 9:1 ratio, with 450 images for training and 50 for testing.

4.1.3. Experimental setup

The hardware environment for this experiment is configured as follows: the central processing unit (CPU) is an Intel® Core™ i9-14900K, equipped with an Nvidia GeForce RTX 4080s GPU featuring 16GB of video memory. The Python version employed is 3.9. The hyperparameters for training the TeethDE-GCNet model are set as follows: a batch size of 8, Adam optimizer, an initial learning rate of 5×10^{-3}, weight decay of 5×10^{-4}, a warm-up period of 10 epochs, and a total training duration of 100 epochs. The input image resolution for the network is 640×640 pixels.

4.2. Results and analysis

4.2.1. Ablation study on basic convolution module

To validate the effectiveness of the proposed Detail-Enhanced (DE) convolution in improving feature extraction capability and suppressing irrelevant information, five sets of comparative experiments were conducted on two datasets. The specific experimental results are shown in Table 1.

Table 1. Comparative experiments on convolution modules.

Method	Pre (%) ↑		Recall (%) ↑		mAP50 (%) ↑		mAP50:95 (%) ↑	
	Tufts	**Ours**	**Tufts**	**Ours**	**Tufts**	**Ours**	**Tufts**	**Ours**
Baseline	91.61	91.19	89.41	91.60	95.12	95.83	62.96	67.22
ECA	95.27	96.51	94.25	96.71	97.77	98.25	69.79	73.86
DA	93.17	95.73	91.60	96.30	96.54	98.16	67.27	72.79
GE	95.48	96.32	94.34	96.16	97.49	98.23	68.66	73.5
CBAM	95.57	96.42	94.00	96.60	97.63	98.18	69.48	74.71
Our	**95.66**	**96.30**	**94.99**	**97.46**	**97.67**	**98.35**	**70.53**	**75.17**

4.2.2. Ablation study

To evaluate the effectiveness of the proposed Detail-Enhanced (DE) module, CMUNeXtBlock, the Focaler-IoU loss for bounding box regression, and the small-object detection head, five sets of ablation experiments were conducted on two datasets under otherwise identical conditions. As each component was progressively introduced, the model exhibited continuous improvements in key metrics such as Recall, mAP@50, and mAP@50:95. Although Precision decreased slightly by 0.07 and 0.08 percentage points, the overall performance improved significantly. In particular, the increase in Recall is especially valuable for medical image analysis tasks, as it helps to reduce the risk of missed detections. The detailed experimental results are shown in Table 2.

Table 2. Ablation study.

Method	Pre (%) ↑		Recall (%) ↑		mAP50 (%) ↑		mAP50:95 (%) ↑	
	Tufts	**Ours**	**Tufts**	**Ours**	**Tufts**	**Ours**	**Tufts**	**Ours**
Baseline	91.61	91.19	89.41	91.60	95.12	95.83	62.96	67.22
FIoU	95.51	97.03	95.58	97.62	97.35	98.23	70.22	75.91
FIoU+CMU	95.81	97.61	95.00	97.47	97.52	98.51	70.22	76.18
FIoU+DE	95.93	97.25	94.77	97.71	97.54	98.49	70.32	76.39
FIoU+CMU+DE	**96.09**	**97.69**	95.15	97.09	97.61	98.51	70.40	76.85
Our	96.02	97.61	**95.2**	**98.07**	**97.87**	**98.91**	**71.17**	**77.52**

To intuitively demonstrate the effectiveness of each module, the detection results were visualized, as shown in Figure 5. It can be observed that the baseline model exhibits notable discrepancies compared to the ground truth annotations. For instance, in Figure 5 a, the baseline model incorrectly identifies the edentulous region as a normal tooth in a missing-tooth scenario. In Figure 5 b, it fails to accurately detect subtle residual roots. Figure 5 c illustrates that the baseline model does not recognize a horizontally impacted wisdom tooth. As the proposed enhancement modules are gradually integrated, the consistency between the model's predictions and the ground truth significantly improves.

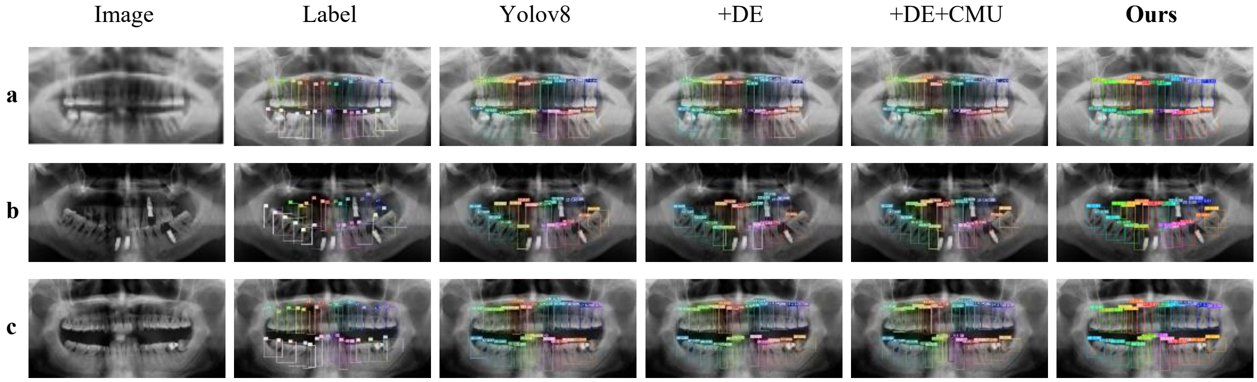

Figure 5. Visualization results of tooth identification on our dataset, a Missing tooth, b Subtle residual root, c Horizontally impacted wisdom tooth

4.2.3. Comparison experiments with public models

To further validate the effectiveness of the proposed method, we conducted comparative experiments on both the public dataset and the self-constructed clinical dataset against a range of state-of-the-art object detection networks. The quantitative detection results on the two datasets are presented in the Table 3. The results demonstrate that the proposed method achieves significant improvements across all evaluation metrics compared to other advanced detection models.

Table 3. Quantitative detection results of comparative experiments.

Method	Pre (%) ↑		Recall (%) ↑		mAP50(%) ↑		mAP50:95(%) ↑	
	Tufts	**Ours**	**Tufts**	**Ours**	**Tufts**	**Ours**	**Tufts**	**Ours**
rtdetr	86.94	83.89	95.71	85.56	92.69	86.61	59.19	56.29
Yolov3	90.73	90.65	91.97	91.57	95.60	95.81	62.39	64.34
Yolov5	88.17	88.64	86.99	87.56	92.31	94.2	59.59	72.2
Yolov6	87.45	90.23	85.16	91.73	92.47	95.5	60.82	64.7
Yolov8	91.61	91.19	89.41	91.60	95.12	95.83	62.96	67.22
Yolov11	91.44	90.95	89.46	89.91	94.85	95.02	61.03	66.76
Our	**96.02**	**97.61**	**95.2**	**98.07**	**97.87**	**98.91**	**71.17**	**77.52**

5. CONCLUSION

This paper proposes TeethDE-GCNet, an automatic tooth numbering framework for panoramic dental radiographs. By integrating Detail-Enhanced Convolution (DEConv), the CMUNeXtBlock module, a small-object detection head, and the Focaler-IoU loss function, the framework significantly enhances performance in recognizing complex dental structures and detecting small targets. Experimental results demonstrate that the proposed method achieves outstanding performance on both public datasets and real clinical data. This study provides a robust technical foundation for intelligent analysis and assisted diagnosis in dental medical imaging. Future work will focus on multimodal information fusion and model lightweighting to further improve deployment efficiency and applicability in clinical settings.

ACKNOWLEDGEMENTS

This study was supported by the National Natural Science Foundation of China (No. 62262029), the Natural Science Foundation of Jiangxi Province (No. 20212BAB202016), the Science and Technology Research Project of Jiangxi Provincial Department of Education (Nos. GJJ200318 and GJJ210520), and Jiangxi Provincial Health Commission Science and Technology Program Project (No. 202410538).

REFERENCES

[1] C.R. Janes, et al., Global health. A Companion to Medical Anthropology, 2022, 109–125; https://doi.org/10.1002/9781119718963.ch6.

[2] T. Vos et al., “Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019”, The Lancet, vol. 396, no. 10258, pp. 1204–1222, Oct. 2020; doi: 10.1016/S0140-6736(20)30925-9.

[3] N. Jain, U. Dutt, I. Radenkov, and S. Jain, “WHO’s global oral health status report 2022: Actions, discussion and implementation”, Oral Diseases, vol. 30, no. 2, pp. 73–79, 2024; doi: 10.1111/odi.14516.

[4] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. P. Campbell, “Introduction to Machine Learning, Neural Networks, and Deep Learning”, Transl Vis Sci Technol, vol. 9, no. 2, p. 14, Feb. 2020, doi: 10.1167/tvst.9.2.14.

[5] P. C. Maganur et al., “Development of Artificial Intelligence Models for Tooth Numbering and Detection: A Systematic Review”, International Dental Journal, vol. 74, no. 5, pp. 917–929, Oct. 2024; doi: 10.1016/j.identj.2024.04.021.

[6] K. Panetta, R. Rajendran, A. Ramesh, S. Rao, and S. Agaian, “Tufts Dental Database: A Multimodal Panoramic X-Ray Dataset for Benchmarking Diagnostic Systems”, IEEE J. Biomed. Health Inform., vol. 26, no. 4, pp. 1650–1659, Apr. 2022; doi: 10.1109/JBHI.2021.3117575.

[7] Huang X, Zhong B, Cao Y, et al. Chest X-ray lung Chinese description generation based on semantic labels and hierarchical LSTM[C]//2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2020: 1020-1023; doi: 10.1109/BIBM49941.2020.9313293.

[8] Yi Y, Jiang Y, Zhou B, et al. C2FTFNet: Coarse-to-fine transformer network for joint optic disc and cup segmentation[J]. Computers in Biology and Medicine, 2023, 164: 107215; https://doi.org/10.1016/j.compbiomed.2023.107215.

[9] E. Kaya, H. G. Gunec, S. S. Gokyay, S. Kutal, S. Gulum, and H. F. Ates, “Proposing a CNN Method for Primary and Permanent Tooth Detection and Enumeration on Pediatric Dental Radiographs”, J Clin Pediatr Dent, vol. 46, no. 4, pp. 293–298, Jul. 2022; doi: 10.22514/1053-4625-46.4.6.

[10] B. Beser et al., “YOLO-V5 based deep learning approach for tooth detection and segmentation on pediatric panoramic radiographs in mixed dentition”, BMC Medical Imaging, vol. 24, no. 1, p. 172, Jul. 2024; doi: 10.1186/s12880-024-01338-w.

[11] X. Xu, C. Liu, and Y. Zheng, “3D Tooth Segmentation and Labeling Using Deep Convolutional Neural Networks”, IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 7, pp. 2336–2348, Jul. 2019; doi: 10.1109/TVCG.2018.2839685.

[12] Y.-R. Van Eycke, A. Foucart, and C. Decaestecker, “Strategies to Reduce the Expert Supervision Required for Deep Learning-Based Segmentation of Histopathological Images”, Front Med (Lausanne), vol. 6, p. 222, Oct. 2019; doi: 10.3389/fmed.2019.00222.

[13] H. Wu et al., “Contrastive Learning for Compact Single Image Dehazing”, in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2021; doi: 10.1109/CVPR46437.2021.01041.

[14] H. Wu, J. Liu, Y. Xie, Y. Qu, and L. Ma, “Knowledge Transfer Dehazing Network for NonHomogeneous Dehazing”, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Jun. 2020; doi: 10.1109/CVPRW50498.2020.00247.

[15] Z. Chen, Z. He, and Z.-M. Lu, “DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention”. arXiv, 2023; doi: 10.48550/ARXIV.2301.04805.

[16] F. Tang, J. Ding, L. Wang, C. Ning, and S. K. Zhou, “CMUNeXt: An Efficient Medical Image Segmentation Network based on Large Kernel and Skip Fusion”. arXiv, Aug. 03, 2023; doi: 10.48550/arXiv.2308.01239.

[17] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s”. arXiv, Mar. 02, 2022; doi: 10.48550/arXiv.2201.03545.

[18] H. Zhang and S. Zhang, “Focaler-IoU: More Focused Intersection over Union Loss”. arXiv, Jan. 19, 2024; doi: 10.48550/arXiv.2401.10525.