# Design of an Intelligent Detection System for Dental Health Status Based on Improved YOLOv8

**Zhengju Guo, Yu Pu, Ting Huang, Jiarui Luo, Shengni Fu, Guangyan Wang***
*School of Information Engineering, Tianjin University of Commerce, Tianjin, China*
**Corresponding Author*

**Abstract: To address the issues of low efficiency and high error rates associated with traditional manual diagnosis in oral health examinations, this paper proposes a real-time dental health detection system based on an improved YOLOv8 algorithm. By integrating the SE attention mechanism to enhance the recognition capability for subtle lesions and optimizing the CIoU loss function to improve localization accuracy, the system is trained and validated on a constructed dataset comprising nearly 3,000 multi-modal oral images. The proposed system achieves an average precision (AP) of 84.26% for ulcers and a detection accuracy of 63.64% for caries in oral endoscopic images. For dental panoramic radiographs, the system enables accurate localization of metallic restorations, and the detection results exhibit strong robustness under complex backgrounds. This system provides a viable technical solution for the intelligent screening and computer-aided diagnosis of dental diseases.**

**Keywords: Dental Health Detection; YOLOv8 Algorithm; Deep Learning; Dental Medical Imaging; Object Detection**

## 1. Introduction

With the sustained development of social economy and the continuous improvement of national health awareness, oral health has attracted increasing attention, and promoting the innovation of dental health detection technology has become a key driver of industrial development. At present, the field is confronted with several prominent challenges: traditional manual diagnosis is inefficient and prone to errors, making it difficult to meet the demands of large-scale screening; the value of massive imaging data in medical institutions remains insufficiently explored and rationally utilized; although artificial intelligence technology has opened up new development avenues, how to deeply adapt advanced deep learning algorithms--especially the YOLO series of object detection algorithms renowned for high speed and precision [1]--to the actual clinical scenarios of stomatology still requires overcoming technical bottlenecks such as the difficulty in extracting features of small lesions and severe interference from complex backgrounds. These challenges have restricted the popularization and intelligent transformation of detection technology, particularly in primary medical settings.

To address the above issues, this study focuses on the targeted optimization of the YOLOv8 algorithm [2] to construct a high-efficiency detection system suitable for oral multi-modal images. Its core work consists of two main aspects: first, at the algorithm level, the SE attention mechanism is introduced to enhance the model's ability to identify subtle lesions in dental images, and the CIoU loss function is optimized to improve the positioning accuracy of lesions, thereby effectively resolving the adaptability problems of small object detection and complex background interference [3,4]; second, at the data level, a practical strategy for constructing medical small sample datasets is formed by integrating public datasets with a self-built dataset of nearly 3,000 clinical endoscopic images. This system is designed to realize the rapid and accurate identification of metal restorations in panoramic dental radiographs and common dental diseases such as dental caries and ulcers in endoscopic images [5]. We aim to reduce the analysis time of a single image to within one minute and assist in lowering the misdiagnosis rate, thus promoting the translation of the technology from laboratory research to practical clinical application.

The research and application of this system hold multidimensional value: for clinicians, it serves as a high-efficiency auxiliary tool, facilitating rapid preliminary screening for primary care physicians and providing support for specialists in identifying early-stage lesions, thereby improving the standardization of diagnosis; for patients, it brings more efficient and precise screening services, contributing to early intervention and an enhanced treatment experience; for the industry, the popularization of the system can drive the transformation of oral healthcare toward an AI-assisted intelligent model, help optimize resource allocation, alleviate the regional shortage of professional dentists, and narrow the gap in oral health services. Overall, this study not only provides specific technical approaches and theoretical references for the interdisciplinary application of stomatology and computer vision but also points out the direction for the subsequent development of lightweight and robust clinical-oriented systems. Ultimately, it is committed to contributing a feasible technical solution to the improvement of national oral health management.

## 2. Key Technological Theories for the Dental Health Detection System

Dental health serves as a crucial window reflecting an individual's overall well-being. It not only directly impacts fundamental physiological functions and socio-psychological states such as chewing, speech, and aesthetics but is also closely and bidirectionally linked to systemic health conditions (e.g., cardiovascular diseases, diabetes, pregnancy outcomes). The World Health Organization (WHO) has identified oral diseases as a major global public health challenge. With the increasing maturity and widespread application of artificial intelligence technologies, particularly deep learning-based object detection and image segmentation algorithms, in the field of medical imaging, building an automated, intelligent, and high-precision dental health detection system has become technically feasible. This chapter will systematically elucidate the key technological theories for constructing such a system from four dimensions: medical evaluation standards, multimodal data characteristics, core algorithm principles, and targeted optimization strategies. This provides a rigorous and comprehensive theoretical foundation for subsequent system architecture design, model training, and experimental validation [6]. A tooth is divided into three parts: the crown, the neck, and the root. Its internal structure, from the outside inward, consists of enamel, dentin, the pulp cavity, and cementum. These structures serve functions such as protecting deep tissues, resisting masticatory pressure, sensing external stimuli, providing nutritional support, and facilitating sensory conduction, making teeth highly specialized organs adapted for chewing function through biological evolution. Common oral diseases include dental caries, dental plaque, and gingival bleeding, as shown in Figure 1.

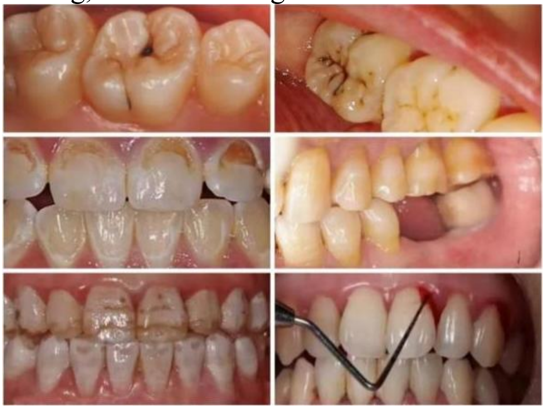

**Figure 1. Common Oral Diseases**

### 2.1 Disease Spectrum for Multi-Target Collaborative Detection

A practical clinical decision support system must possess the capability to handle complex, co-existing diseases. Its detection targets should cover a broad spectrum of oral conditions:

Hard Tissue Diseases: Dental caries are a core target. Their imaging manifestations present a continuum, from early enamel surface demineralization (faint radiolucency on X-rays, chalky white spots under endoscopy) to the formation of distinct cavities. Non-carious tooth defects such as abrasion, wedge-shaped defects, and dental erosion exhibit varied morphological characteristics, requiring the model to learn different texture and contour patterns.

Periodontal and Periapical Diseases: Alveolar bone resorption is the core radiographic manifestation of periodontitis. The model needs to assess the morphology (horizontal, vertical) and extent of resorption. Periapical lesions (e.g., granulomas, cysts) appear as round or oval radiolucent areas at the root apex on X-rays, with boundary clarity being a key indicator of their nature.

Restorations and Abnormal Structures: Identifying various restorations (fillings, crowns, bridges, implants) is crucial for tracing treatment history and diagnosing peri-restorative diseases. Locating impacted teeth (especially third molars) and assessing their relationship with adjacent teeth and the mandibular nerve canal are fundamental for surgical extraction planning.

Emerging Detection Targets: Some cutting-edge research has begun exploring the screening of mucosal abnormalities associated with early-stage oral cancers (e.g., erythroplakia, leukoplakia, textural changes). This typically requires the fusion of high-resolution color endoscopic images with more advanced feature learning networks.

The current research paradigm is rapidly evolving from early single-disease classification to multi-label, multi-task collaborative detection frameworks. For instance, an advanced system can simultaneously localize and classify caries, restorations, alveolar bone height, and impacted teeth on a single panoramic radiograph. The advantage lies in sharing general features extracted by a backbone network, improving overall efficiency, and better aligning with real clinical image-reading workflows.

The current research paradigm is evolving from single-disease classification towards multi-label collaborative detection frameworks. For an input image, the system output can be formalized as a set:

$$D = \{b_k, c_k, s_k\}_{k=1}^{k} \quad (1)$$

where K is the number of detected objects, $b_k = [x_k, y_k, w_k, h_k]^T$ represents the bounding box (center coordinates, width, height) of the k-th target, $C_k$ is a C-dimensional vector representing the probability distribution of this target belonging to C disease categories (e.g., caries, calculus, restoration), and $s_k$ is the corresponding confidence score. This joint output framework enables the system to perform a panoramic diagnosis on a single image.

## 2.2 Core Assessment Indicators for Dental Health and Common Diseases

The primary prerequisite for establishing an automated detection system is to strictly anchor its task objectives within the clinical medical evaluation framework. This system is not merely about "finding differences"; it needs to map pixel-level image features to diagnostic indicators or pathological categories with clear clinical significance.

Traditional oral epidemiology and clinical diagnosis heavily rely on a series of quantitative indices. The Decayed, Missing, and Filled Teeth (DMFT) Index is central to assessing caries burden, defined as:

$$DMFT = D + M + F \quad (2)$$

where D represents the number of decayed teeth, M represents the number of teeth missing due to caries, and F represents the number of filled teeth due to caries.

For more granular assessment, the Decayed, Missing, and Filled Surfaces (DMFS) Index is used:

$$DMFS = D_S + M_S + F_S \quad (3)$$

In an AI system, the model needs to perform multi-class classification of tooth status (healthy, decayed, filled, missing). Its output can be represented as a class probability vectorck $c_k$ . The final DMFT index calculation is then transformed into a statistical aggregation of the classification results for all teeth.

Periodontal health assessment relies on precise measurement of key anatomical points. Clinical Attachment Loss (CAL) is the gold standard, defined as:

$$CAL = PD + GR \quad (4)$$

where PD is the probing depth and GR is the gingival recession distance.

In imaging, this is equivalent to locating the Cemento-Enamel Junction (CEJ) point and the Alveolar Crest (AC) point and calculating their Euclidean distance. Assuming the CEJ point coordinates are $(x_C, y_C)$ and the AC point coordinates are $(x_a, y_a)$ CAL can be obtained by converting the pixel distance using a calibration coefficient α for the actual scale:

$$CAL = \alpha \cdot \sqrt{(x_a - x_c)^2 + (y_a - y_c)^2} \quad (5)$$

Recent segmentation networks (e.g., U-Net variants) learn to locate these points by optimizing the following objective function:

$$\mathcal{L}_{\text{seg}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{C=1}^{C} y_{i,c}\log\left(\hat{y}_{i,c}\right) \quad (6)$$

where N is the total number of pixels, C is the number of classes (e.g., background, CEJ, AC), $y_{i,c}$ is the ground truth label, and $\hat{y}_{i,c}$ is the predicted probability.

Quantitative Clinical Assessment Index System: Traditional oral epidemiology and clinical diagnosis heavily rely on a series of internationally recognized quantitative indices, which provide "standard answers" for AI models to learn.

Caries Assessment: Primarily employs the DMFT and DMFS indices. The calculation of DMFT/DMFS directly reflects the cumulative disease experience and treatment history, serving as the cornerstone for assessing population caries burden and individual risk. In an AI system, this means the model not only needs to detect "caries" but also must differentiate between "filled" and "missing due to caries" states, imposing higher demands on the algorithm's fine-grained classification capability.

Periodontal Health Assessment: This is a multi-parameter comprehensive evaluation system. Probing Depth (PD) assesses pocket depth; Clinical Attachment Loss (CAL) is the gold standard for determining the degree of periodontal support tissue destruction; the Gingival Bleeding Index (GBI) directly reflects inflammatory activity. Notably, CAL calculation relies on two anatomical landmarks: the Cemento-Enamel Junction (CEJ) and the Alveolar Crest (AC). Recent research shows that deep learning-based models (e.g., U-Net variants) can automatically segment and locate these key points from images such as oral ultrasound, achieving sub-millimeter measurement errors, demonstrating consistency and repeatability surpassing manual measurements. This marks AI's role in advancing periodontal assessment from a "semi-quantitative" stage reliant on clinician tactile sensation and experience towards an objective, precise "fully quantitative" era.

Oral Hygiene Assessment: The Plaque Index (PI) and Calculus Index (CI) are used to quantify oral cleanliness. Traditional assessment relies on visual inspection after disclosing agent staining, which is highly subjective. Computer vision-based automated assessment systems, particularly those using object detection algorithms like YOLO to identify and calculate the area of stained plaque, have become a research hotspot. They can also achieve "staging" assessment of plaque (e.g., newly formed vs. mature plaque).

### 2.3 Mainstream Detection Methods for Dental Health and Multimodal Data Fusion

Oral medical diagnosis is inherently a process of synthesizing multimodal information. Different imaging techniques, based on distinct physical principles, reveal information about tissues at different levels, collectively forming the "perceptive organs" for both clinicians and AI systems.

The are four common imaging modalities for dental health detection: X-ray imaging (including CBCT) is the gold standard for hard tissue assessment but involves radiation and artifacts; optical imaging is safe and non-invasive, suitable for early surface lesions, but has weak penetration depth and is susceptible to environmental interference; 3D surface imaging provides high-precision geometric models, aiding digital diagnosis and treatment, but only captures visible surfaces; ultrasound imaging is radiation-free and allows real-time soft tissue assessment, but suffers from high image noise and strong operator dependence. These methods complement each other, and AI technology is focusing on addressing the image analysis challenges specific to each modality.

The limitations of any single modality necessitate that multimodal data fusion is the path toward comprehensive and precise diagnosis. [7]

Fusion can occur at different levels:

Data-Level Fusion: Merging raw data from different sources after precise spatial registration. For example, fusing a high-precision crown surface model from intraoral scanning with a jawbone model from CBCT for virtual implant surgery planning. This requires solving rigid/non-rigid registration challenges across different coordinate systems.

Feature-Level Fusion: Extracting high-level features separately from different modality data, followed by concatenation or attention-based weighting. For example, extracting "bone density" features from X-rays and "gingival color and texture" features from color endoscopic photos, then feeding them together into a classifier to determine periodontitis activity. This more flexible approach is currently mainstream in research. [8]

Decision-Level Fusion: Each modality's data independently produces a preliminary diagnostic result via a sub-model, with final comprehensive decision-making through voting, weighted averaging, or a meta-learner. For example, models based on panoramic X-rays and intraoral photos respectively make judgments on "presence of interproximal caries," and the two results are integrated for a final conclusion, enhancing system robustness.

The core challenge in achieving effective fusion lies in handling data heterogeneity (e.g., images vs. point clouds, 2D vs. 3D) and misalignment. Deep learning methods, particularly cross-modal attention mechanisms and knowledge distillation techniques, are widely used to learn shared semantic spaces between modalities, allowing features like "bone structure" from CBCT and "crown morphology" from intraoral scans to align and complement each other at an abstract level, thereby driving more reliable collaborative diagnosis.

Feature-level fusion is the mainstream method. Suppose there are MM modalities, and the deep feature for the mm-th modality is $\mathrm{F}^{(m)} \in \mathrm{R}^{H\times W\times C_m}$ .Fusion can be achieved via channel concatenation and convolution:

$$F_{\text{fused}} = \mathrm{Conv}_{1\times 1}\left(\mathrm{Concat}\left(F^{(1)}, F^{(2)}, ..., F^{(M)}\right)\right) \quad (7)$$

A more advanced approach employs cross-modal attention mechanisms. Taking a bimodal example (X-ray $\mathrm{F_x}$, color photo $\mathrm{F_c}$), attention weights from modality aa to modality bb can be computed:

$$A_{a\to b} = \mathrm{Softmax}\left(\frac{(W_Q F_a)(W_K F_b)^T}{\sqrt{d_k}}\right) \quad (8)$$

$$F_a^{\text{enhanced}} = F_a + A_{a\to b}(W_V F_b) \quad (9)$$

Where $\mathrm{W_Q}, \mathrm{W}_K, \mathrm{W}_V$ are learnable linear projection matrices. This mechanism allows X-ray features to "attend to" corresponding areas of soft tissue color abnormalities in the color photo, achieving information complementarity.

### 2.4 Deep Learning and the YOLO Algorithm

Deep learning, particularly Convolutional Neural Networks (CNNs), with their hierarchical "end-to-end" feature learning capability, has become a powerful tool for parsing complex medical images. In the specific task of dental health detection, its main application forms are object detection and image segmentation. [9]

The object detection task can be formalized as: given an image $\mathbf{I} \in$ , learn a mapping $f: \mathbf{I} \to \mathcal{D}$ . YOLO divides the image into an S×S grid. Each grid cell predicts BB bounding boxes. Each box contains 5 core parameters: center coordinate offsets $\left(t_x, t_y\right)$ , width-height scaling factors $(t_w, t_h)$ , and a confidence score $t_o$ .The final predicted box parameters are calculated as:

$$\begin{aligned} b_x &= \sigma(t_x) + c_x \\ b_y &= \sigma\left(t_y\right) + c_y \\ b_w &= p_w \cdot e^{t_w} \\ b_h &= p_h \cdot e^{t_h} \end{aligned} \quad (10)$$

Where $\left(c_x, c_y\right)$are the coordinates of the top-left corner of the grid cell, $(p_w, p_h)$ are the preset anchor box dimensions for that cell, and $\sigma(\cdot)$is the Sigmoid function, which compresses the offsets to (0,1), ensuring the predicted box center lies within the current grid cell.

Evolution and Mathematical Formulation of YOLO Loss Functions:

YOLO training is achieved by minimizing a multi-task loss function. Taking the basic YOLOv1 loss as an example:

$$\begin{aligned} \mathcal{L} &= \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2\right] \\ &+ \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2\right] \\ &+ \sum_{i=0}^{s^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} (C_i - \hat{C}_i)^2 \\ &+ \lambda_{\text{noobj}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{noobj}} (C_i - \hat{C}_i)^2 \\ &+ \sum_{i=0}^{s^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c\in\text{classes}} (p_i(c) - \hat{p}_i(c))^2 \end{aligned} \quad (11)$$

Here,$\mathbb{1}_{ij}^{\text{obj}}$ is an indicator function that equals 1 if the jj-th anchor box in grid cell ii is responsible for a ground truth object. This loss comprises coordinate loss, confidence loss for boxes containing objects, confidence loss for boxes with no object, and classification loss. In subsequent evolutions, the coordinate loss was replaced by the IoU series of losses. The standard IoU and its improved forms are defined as:

$$\mathrm{IoU} = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \quad (12)$$

$$\mathrm{GIoU} = \mathrm{IoU} - \frac{|C \setminus (B \cup B^{gt})|}{|C|} \quad (13)$$

$$\mathrm{DIoU} = \mathrm{IoU} - \frac{\rho^2(b, b^{gt})}{c^2} \quad (14)$$

$$\mathrm{CIoU} = \mathrm{IoU} - \frac{\rho^2(b, b^{gt})}{c^2} - \alpha v \quad (15)$$

$$v = \frac{4}{\pi^2}\left(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h}\right)^2 \quad (16)$$

$$\alpha = \frac{v}{(1 - \mathrm{IoU}) + v} \quad (17)$$

Where C is the smallest enclosing box covering both B and$\mathrm{B^{gt}}$ , ρ is the Euclidean distance, and cc is the diagonal length of the enclosing box. The CIoU loss simultaneously considers overlap area, center point distance, and aspect ratio, making it particularly suitable for targets like teeth that require precise shape matching. Its loss is:

$$\mathcal{L}_{\mathrm{CIoU}} = 1 - \mathrm{CIoU} \quad (18)$$

Classification loss often employs Binary Cross-Entropy or Focal Loss to address class imbalance [10].

Focal Loss is defined as:

$$\mathcal{L}_{\mathrm{FL}} =-\alpha_t(1-p_t)^{\gamma}\log(p_t) \qquad (19)$$

Where $\boldsymbol{p_t}$s the model's estimated probability for the true class,$\boldsymbol{\alpha_t}$is a balancing factor, and $\boldsymbol{\gamma}$is a focusing parameter. By down-weighting the contribution of easy-to-classify samples, the model focuses more on hard-to-classify samples (e.g., early caries).

Two-Stage Detectors: Represented by Faster R-CNN, they follow the "region proposal -> region classification & regression" paradigm. A Region Proposal Network (RPN) first generates a large number of candidate bounding boxes (anchors) on feature maps, which are then refined and classified by subsequent networks. Their advantage is high accuracy, especially for dense and small objects; the drawback is a complex pipeline with high computational cost, making it difficult to meet real-time requirements.

One-Stage Detectors: Represented by the YOLO series and SSD, they reframe detection as a single regression problem. YOLO divides the input image into an S×S grid. Each grid cell directly predicts coordinates, confidence scores for multiple bounding boxes, and conditional probabilities for all classes. Its core advantage is extremely fast speed and a concise structure, making it highly suitable for rapid preliminary screening of large volumes of images in clinical settings or deployment on edge devices with limited computing resources (e.g., portable oral scanners).

The basic YOLO loss function typically consists of three parts:

Coordinate Loss (measuring error between predicted and ground truth box centers and dimensions, now commonly replaced by IoU-based losses [3].

This work adopts confidence loss (binary cross-entropy for object presence) and classification loss (cross-entropy or binary cross-entropy for category prediction). For dental image analysis, YOLO is customized with targeted optimizations:

(1) Small object detection: Shallow or extra high-resolution detection heads are added to locate small lesions (e.g., caries), while attention modules (SE, CBAM, Coord Attention) highlight local lesion textures and suppress backgrounds.

(2) Class imbalance & fine-grained classification: Weighted cross-entropy or focal loss alleviates the imbalance between healthy and lesion samples. Refined classification branches or metric learning are introduced for tasks such as plaque staging and caries depth grading.

(3) Detection-to-segmentation extension: YOLOv8's instance segmentation head enables simultaneous detection and pixel-level mask prediction, which is essential for quantifying caries area and alveolar bone loss. Dice loss or its combination with cross-entropy is used to refine segmentation contours.

(4) Light weighting and Deployment Optimization [11]: To promote adoption in clinical and mobile settings, techniques like network pruning, knowledge distillation, and quantization are used to compress models. For example, distilling knowledge from a complex teacher model with an attention mechanism into a more lightweight student model can significantly increase inference speed with minimal accuracy loss.

Recent studies validate the efficacy of these tailored strategies. A YOLOv8-Dental model optimized for dental X-rays achieved 89.8% mAP on seven-class detection. For dental plaque staging, YOLOv11m yielded 71.3% mAP@50, showing state-of-the-art performance in complex dental image analysis.

## 3. Research Proposal Design

This project adopts the collaborative innovation model of "industry-university-research-medical" to build a modular research structure. The research design deeply integrates clinical needs and technical implementation, covering a complete closed loop from data collection to clinical application. We have established a six-stage R&D system that includes data preparation, algorithm development, model training, quality control, clinical validation and product implementation. The research team has the ability to develop both MATLAB and Python platforms, and has completed the preliminary construction and loss curve analysis of the U-Net convolutional neural network in the early stage, laying a solid technical foundation for the project. Through in-depth cooperation with Tianjin Medical University Stomatological Hospital, the clinical practicability and technical forward-looking research direction are ensured.

### 3.1 Research Scope and Questions

3.1.1 Scope of research

Image data: Nearly 3,000 oral image data were selected, including MICCAI2023s [12] public panoramic dental piece dataset and self-built intraoral endoscopic image dataset (covering typical lesions such as dental caries, oral ulcers, and metal restorations).

Relevant personnel: grassroots dental clinicians (such as community dental clinic practitioners), medical AI algorithm R&D personnel (such as algorithm engineers of medical technology companies), and oral disease patients (covering different age groups and different types of oral diseases).

Algorithm model: Based on YOLOv8, it is limited to the improved YOLOv8 algorithm that introduces SE attention mechanism, multi-scale feature fusion (FPN+PAN), and CIoU loss function optimization.

Detection tasks: limited to three core tasks: panoramic metal restoration positioning, endoscopic image caries detection, and endoscopic image oral ulcer detection.

3.1.2 Research questions

Core technical issues

(1) Is there a significant difference in the detection accuracy of different types of lesions (metal restorations, dental caries, oral ulcers) in oral multimodal images (panoramic dental chips, endoscopic images) of the dental health detection system based on YOLOv8? What are the core influencing factors of difference?

(2) After the introduction of the SE attention mechanism and the optimization loss function, how much performance improvement does the improved YOLOv8 algorithm in the detection of small lesions (early microcaries and micro-ulcers) compared with the original algorithm?

(3) How to solve the problem of balancing detection speed and accuracy of the lightweight and improved YOLOv8 model on edge devices? Can it meet the needs of real-time diagnosis in primary dental clinics?

3.1.3 Clinical application questions

(1) What are the operational thresholds and trust barriers faced by grassroots dentists when using AI detection systems? Is there a difference in the acceptance of AI-assisted diagnosis among doctors with different years of practice?

(2) What are the common characteristics of the YOLOv8-based detection system in terms of lesion type and image features compared to manual diagnosis of misdiagnosis/missed diagnosis cases in clinical practice?

(3) What is the patient's acceptance of AI-assisted oral diagnosis? Does age, educational background, and type of oral disease affect patient trust in AI diagnostic results?

### 3.2 Dataset Construction

Using publicly available intraoral images and MICCAI2023 Dental Image Challenge panoramic dental slice images, it contains approximately 300 X-ray images and nearly 2000 intraoral images with a resolution of 512×512 or 1024×1024. Oral endoscopic images include caries, ulcers and other diseases, covering a total of more than 2,000 pictures; MICCAI2023 selected from the panoramic dental image images of the Dental Image Challenge and inspected the panoramic dental slices with obvious restorations, covering a total of more than 100 different images. The advantage of public datasets is that they are labeled and suitable for preliminary model training, but their scale is small and the coverage of lesion types is limited. Some screenshots of the dataset are shown in Figure 2 and 3, respectively.

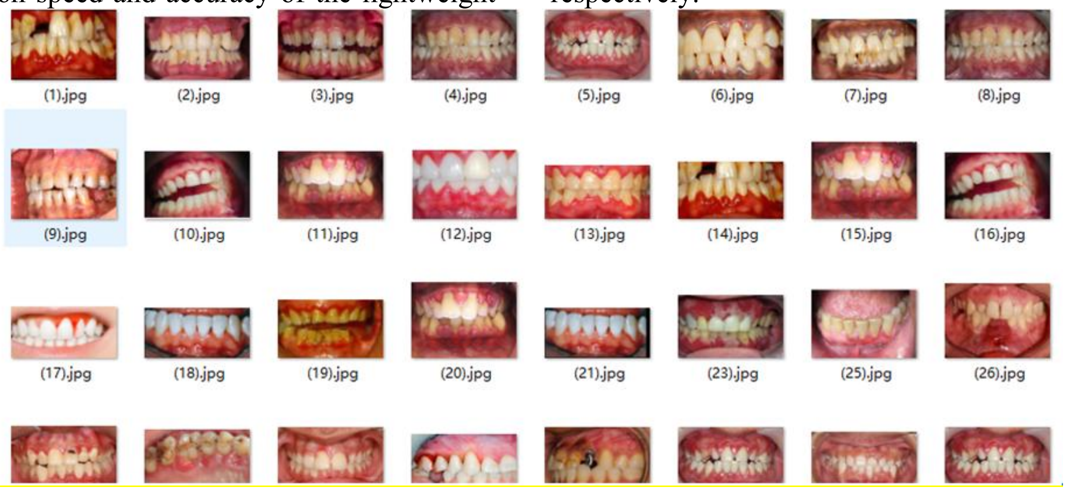

**Figure 2. Public Dataset of Oral Endoscopy**

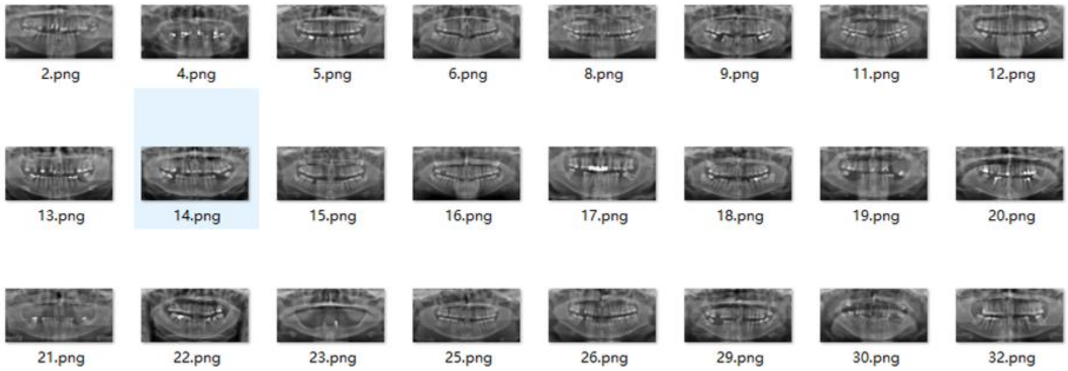

**Figure 3. Panoramic Dental Sheet Public Dataset**

Self-built dataset: Establish a strategic cooperation with Tianjin Medical University Stomatological Hospital to collect 3D endoscopic images of real cases, covering different ages, genders, and oral health conditions (such as health, dental caries, periodontal disease, etc.) to ensure their representativeness, and use high-precision oral endoscopes to collect 201 visible light tooth images with a resolution of 1280×720. These images cover different lighting conditions, dental conditions, and shooting angles. Endoscopic images complement surface features that cannot be captured by X-ray images, enriching the diversity of the dataset. Some of the data are shown in Figure 4.

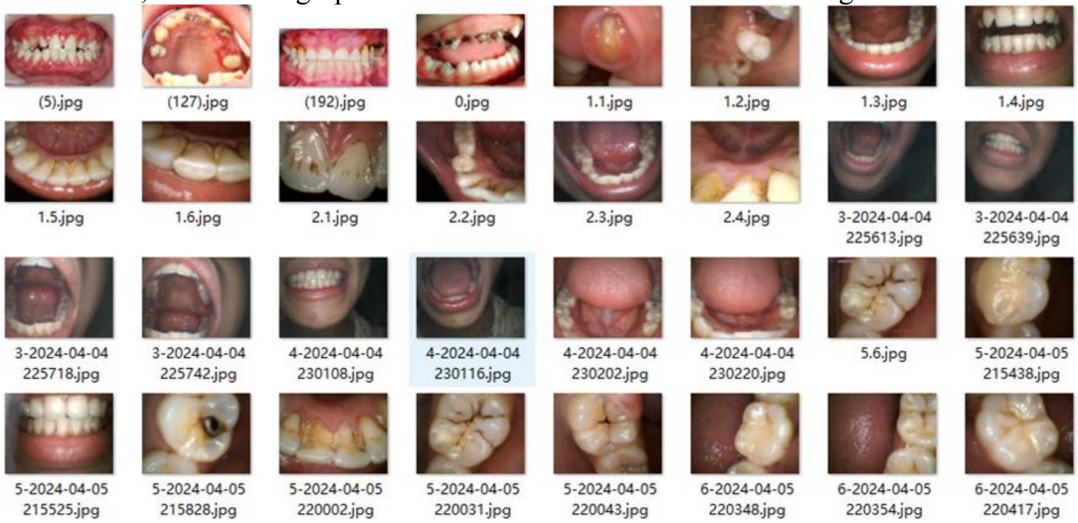

**Figure 4. Self-built Dental Endoscope Dataset**

### 3.3 Data Processing and Process

The project is based on a modular design, covering six stages: data acquisition, preprocessing, model training, weight optimization, result visualization and performance analysis, and combines MATLAB and Python platforms to realize intelligent diagnosis of oral 3D endoscopic images. The overall technical route ensures the high accuracy and real-time performance of the system through multi-stage collaboration. The process is shown in Figure 5.

Step 1: Data collection.

The project uses high-resolution 3D intraoral endoscopes (such as Sony IMX585 sensors with LED cold light sources) to acquire DICOM format images and convert them to PNG standard format through ImageJ. Data sources include a clinical case bank in cooperation with the Stomatological Hospital of Tianjin Medical University, covering common lesions such as caries and periodontal disease, and plans to collect 10,000 samples (including children, adults, and the elderly) in the first year. Data is stored in a distributed system in the cloud, supporting multi-center collaboration and federated learning frameworks to ensure data privacy and traceability.

Step 2: Data preprocessing and label generation.

In the pretreatment stage, CycleGAN and physical model are used to jointly denoise to eliminate metal artifacts. Enhance tooth tissue contrast with Adaptive Histogram Equalization (CLAHE). The label is generated based on the ICDAS grading standard (level 0-6), double-blind labeling by 5 practitioners, and the divergence rate > 10% by experts to ensure that the labeling accuracy ⩾ 95%. The script automatically generates VOC format labels and converts them into YOLO-adapted 2007_train.txt and 2007_val.txt files to support subsequent end-to-end training.

Step 3: Model training.

A dual-model collaborative architecture for caries diagnosis consisting of segmentation and detection is constructed based on the PyTorch 2.0 deep learning framework.

First, the segmentation model employs an improved U-Net++ [13] embedded with the CBAM attention module to strengthen the capture of subtle features such as caries margins, with a target Dice coefficient⩾0.85.

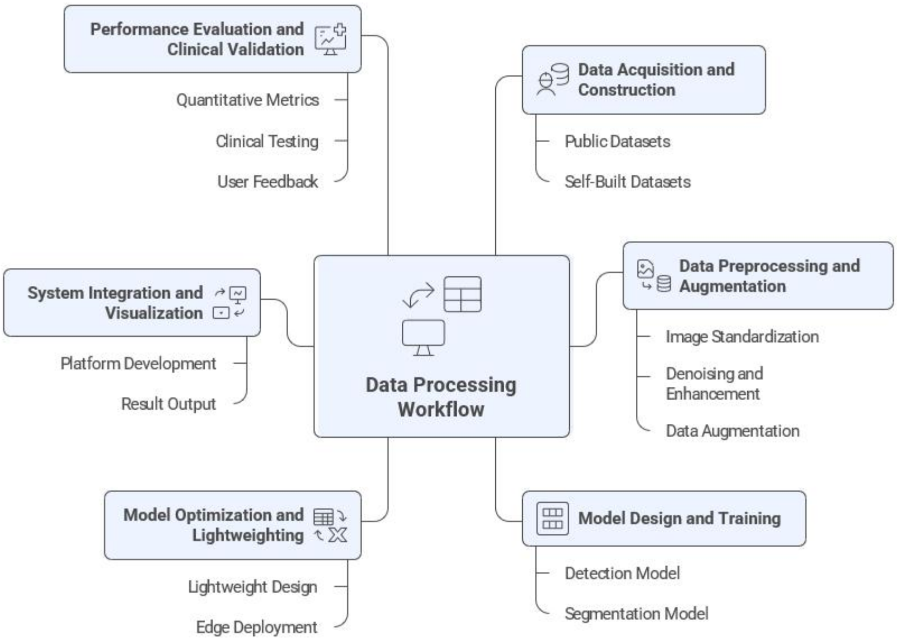

**Figure 5. Data Processing Flow Chart**

Second, the detection model is developed on the basis of YOLOv8, where depthwise separable convolution is introduced for lightweight improvement and the FPN+PAF structure is integrated to enhance multi-scale detection performance, achieving a target mAP⩾0.92.

In the training phase, the SGD optimizer (initial learning rate 0.001, momentum 0.9, weight decay 0.0005) is used for 676 training epochs. Weights are saved and metrics are recorded every 10 epochs. A dynamic learning rate adjustment strategy is adopted: the learning rate is decayed to 0.1 times if the validation set accuracy does not improve for 15 consecutive epochs.

Meanwhile, FP16 mixed-precision training is enabled to accelerate training speed and reduce memory usage.

For the loss function design, the segmentation model fuses Dice Loss (weight 0.6) and cross-entropy loss (weight 0.4), while the detection model adopts CIoU Loss to improve bounding box regression accuracy.

Lightweight model clipping is conducted according to the hardware characteristics of the Jetson Nano edge computing device, ensuring that the trained model achieves an inference speed higher than 15 frames per second on the device to meet the requirements of clinical real-time diagnosis [14].

Step 4: Weight selection.

After model training, the optimal weight file is selected from saved checkpoints to maximize output accuracy. First, all weight files in the logs folder are evaluated on the validation set, with detection mAP, segmentation Dice coefficient, and inference speed recorded. Weights satisfying mAP ⩾ 0.92 and Dice ⩾ 0.85 are retained as candidates. Next, these candidates are further validated using 300 independent clinical test samples covering various lesion types and age groups; lesion identification accuracy, missed diagnosis rate, and misdiagnosis rate are measured to determine the best-performing weight file (denoted as best.py). Finally, in the YOLO.py configuration file, the model_path parameter is set to best.py, with input image size 1920×1080 pixels, confidence threshold 0.5 (adjustable), and IOU threshold 0.45 configured. The model automatically loads the optimal weights at runtime to ensure output accuracy and stability.

Step 5: Visualize the results.

After completing the above configuration, the 3D reconstruction function (VTK rendering) is integrated through the MATLAB GUI interface, which supports multi-angle rotation of the tooth point cloud and the annotation of the thermal icon of the lesion (such as red highlighting of the enamel demineralization area). After running the predict.py script, the system outputs a structured report with lesion location, ICDAS grading and treatment recommendations, and supports PDF export and cloud synchronization.

Step 6: Performance analysis.

The system performance was verified via quantitative assessment and clinical validation to meet clinical application requirements.

First, quantitative performance analysis was conducted on a 500-sample test set, with mAP, Dice coefficient, FPS, and accuracy recorded. Models failing to meet the threshold (mAP<0.92) were retrained and adjusted.

Second, clinical applicability was blindly evaluated by 10 stomatologists, focusing on identification accuracy and grading rationality; the Kappa coefficient was calculated to assess diagnostic consistency (Kappa ⩾ 0.92 as excellent), and diagnostic efficiency improvement was documented.

Finally, robustness testing was performed under varying imaging conditions (illumination, angle, equipment) to validate stability in complex clinical scenarios. Only systems meeting all standards proceeded to clinical pilot trials.

## 4. Analysis of Experimental Results

### 4.1 Oral Endoscopy Results

The detection and identification results from oral endoscopy are shown in Figure 6.

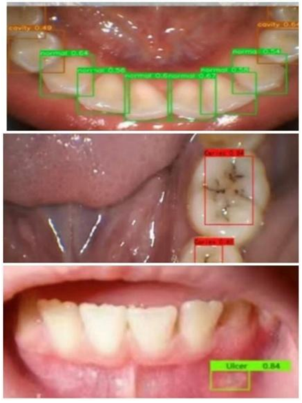

**Figure 6. Detection and Identification Results**

Figure 6 shows the occlusal surface of a tooth bounded by a box, labeled "Caries 0.84", indicating a high prediction probability from the model for caries in this region. Within the box, multiple dark spots are visible, consistent with the clinical features of dental caries. Another boxed tooth is labeled "Caries 0.61", with a lower confidence score of 0.61. The dark area within this box is smaller, likely indicating early-stage caries. Normal teeth are annotated with green boxes. The difference in confidence scores reflects the model's sensitivity to lesion severity. An ulcer region is also boxed and labeled "Ulcer 0.84". The high confidence score of 0.84 indicates the model's high confidence in identifying this region as an ulcer, and the symptoms within the box match the characteristics of an ulcer. The accurate bounding demonstrates the robustness of the YOLO algorithm for detecting soft tissue lesions. [15]

The confidence scores in the figure all fall within the range of 0.61 to 0.84, indicating the YOLO model's high reliability in detecting dental health issues. The repeated appearance of the confidence score 0.84 (for both Caries and Ulcer) reflects the model's stable predictive capability for obvious lesions. Conversely, the lower score of 0.61 suggests insufficient sensitivity to early-stage or borderline lesions, possibly due to a limited number of early lesion samples in the training data. The complex background in the image (saliva reflections, mucosa) did not significantly interfere with the accuracy of the detection boxes, and its performance in a simpler background further validates the model's adaptability to different scenarios. This is consistent with the described YOLOv8 multi-scale feature fusion and SE attention mechanism optimization in the paper, which enhances the ability to characterize health conditions in complex environments. The detection results demonstrate the potential of the YOLO algorithm in characterizing dental health status.

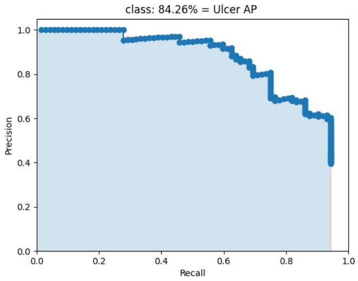

**Figure 7. Ulcer AP Rate Curve**

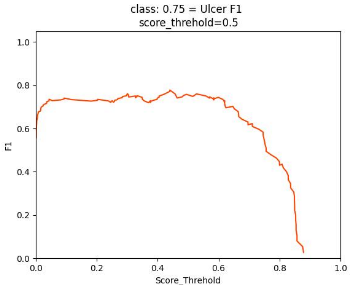

**Figure 8. Ulcer F1 Score Curve**

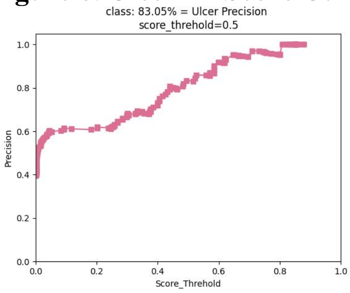

**Figure 9. Ulcer Precision Curve**

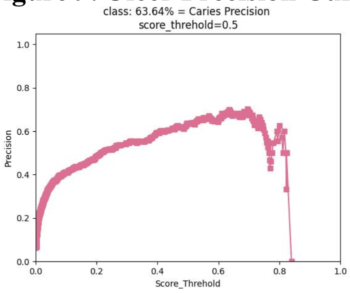

**Figure 10. Caries Precision Curve**

The system performance effects are shown in Figures 7-10. In Figure 7, the initial Precision is approximately 1.0, maintaining a high value up to a threshold of 0.6. Recall slowly declines between 0.0 and 0.6. The Average Precision (AP), representing the area under the Precision-Recall curve, is 84.26%, demonstrating the overall performance for "Ulcer" and reflecting the high stability of the model for "Ulcer" detection. Figure 8 depicts the F1 curve for ulcers. The F1 score is the harmonic mean of precision and recall, reflecting the balanced performance of the model in "Ulcer" detection. An initial value of 0.75 indicates the model has high accuracy and coverage at low thresholds. Figure 9 shows the Precision curve for ulcers. Precision represents the model's accuracy. The initial Precision is about 0.4, rising rapidly to 1.0. An average value of 83.05% indicates a high correct prediction rate for "Ulcer" by the model. The rapid rise between thresholds 0.2 and 0.8 reflects the effectiveness of confidence score filtering. Figure 10 shows the Precision curve for caries. The initial Precision is about 0.2, rising rapidly to 1.0 between thresholds 0.2 and 0.6, with an average of 63.64%. This indicates a higher false positive rate in "Caries" detection. The larger curve fluctuations suggest the model has limited discriminative ability for "Caries".

An F1 score of 0.75 and an AP of 84.26% indicate that the model has high accuracy and stability in detecting "Ulcer". A Precision of 83.05% approaching 1.0 at high thresholds reflects the model's reliable screening of confidence scores for "Ulcer". The Precision of 63.64% for caries is lower than for "Ulcer". Although F1 and AP data are missing for caries, the declining trend of the curve suggests a lower recall rate. The low precision rate may be related to the varying severity of "Caries" lesions.

### 4.2 Analysis of Panoramic Radiograph Results

The training loss curves for panoramic radiographs are shown in Figure 11. The results indicate that the current model performs well on the detection task (mAP50(B)=0.7), but its segmentation performance still needs improvement, manifested as inaccurate masks and missing targets. Considering the original images and segmentation results, the issues may stem from preprocessing, limited data volume, and model capacity constraints. Significant improvements in segmentation effectiveness are expected by optimizing preprocessing parameters, increasing data, and enhancing model complexity [16].

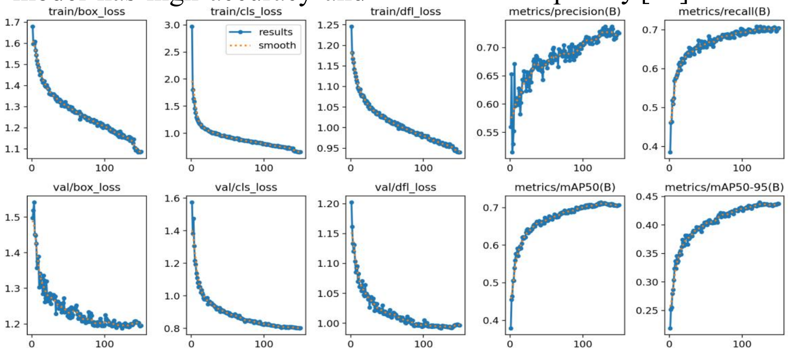

**Figure 11. System Training Loss Graph**

**Table 1. Performance Analysis of the Object Detection and Segmentation Model for Panoramic Radiographs**

| Indicator category | Specific indicators | Initial value/Range | Final value/range | Change trends and explanations |
|---|---|---|---|---|
| Training loss | train/box_loss | 1.7 | ≈1.1 | It declines continuously, dropping rapidly early and leveling off later, showing gradual model convergence. |
| | train/cls_loss | 3.0 | ≈1.0 | Same as above. |
| | train/dfl_loss | 1.25 | ≈1.0 | Same as above. |
| Validation loss | val/box_loss | Stable (1.2–1.5) | stable (1.2–1.5) | The fluctuations are relatively small, and the performance is relatively stable. |
| | val/cls_loss | Fluctuate greatly | Fluctuate greatly | It has strong volatility and may be affected by data distribution or model complexity. |
| | val/dfl_loss | stable (1.05–1.2) | stable (1.05-1.2) | perform steadily |
| Performance indicators | metrics/precision(B) | 0.5 | 0.7 | Gradually rising, indicating an improvement in the prediction accuracy of the model. |
| | metrics/recall(B) | 0.4 | 0.6 | Gradually increase, and the detection coverage rate improves. |
| | metrics/mAP50(B) | 0.4 | 0.7 | Significant improvement indicates good detection performance under the condition of IoU = 0.5. |
| | metrics/mAP50-95(B) | 0.25 | 0.35 | There has been some improvement, but there is still room for improvement within a stricter IoU range. |
| Overall performance evaluation | Detection performance | mAP50(B)=0.7 | - | The detection task performed well. |
| | Segmentation performance | Inaccurate mask, missing target | - | It is necessary to further optimize the preprocessing, increase the data, and improve the model capacity. |
| | Convergence situation | It tends to be flat after 50–70 epochs | - | The model may have approached the performance ceiling of the current dataset. |

It is recommended to adjust according to the above directions, retrain the model, and provide updated results for further analysis. The train/box_loss decreased from 1.7 to approximately 1.1, train/cls_loss decreased from 3.0 to 1.0, and train/dfl_loss decreased from 1.25 to 1.0. The overall downward trend of the loss curves indicates that the model gradually converged on the training set, learning features for object detection and classification. The descent rate was faster in the early stages and stabilized later, suggesting the model was approaching convergence, albeit with minor fluctuations. The val/box_loss stabilized between 1.2 and 1.5, val/cls_loss showed larger fluctuations, and val/dfl_loss stabilized between 1.05 and 1.2. The validation losses exhibited strong volatility, particularly in val/cls_loss. metrics/precision(B) increased from 0.5 to 0.7, metrics/recall(B) from 0.4 to 0.6, metrics/mAP50(B) from 0.4 to 0.7, and metrics/mAP50-95(B) from 0.25 to 0.35. The mAP metrics indicate good detection performance under the IoU=0.5 condition. The performance curves flattened after 50-70 epochs, suggesting the model may have reached the performance ceiling of the current dataset. Detailed distinctions are shown in Table 1.

The design and implementation of this system possess the following salient advantages, providing significant contributions to the intelligent development of oral healthcare:

(1) Robust Model Architecture. The system adopts the YOLOv8-s model architecture [17] constructed with YOLOBody as the backbone network. Its design balances lightweight characteristics with powerful functionality, achieving an exquisite equilibrium between detection speed and accuracy. This backbone network outputs crucial information including bounding box regression, class probabilities, and dimensions, specifically optimized for a small-scale dataset of approximately 2000 images.

During initialization, generalizable pre-trained weights are loaded onto the backbone network. Compared to random initialization, this approach significantly reduces training time and effectively avoids convergence issues, which is crucial for scenarios with limited data. In the weight loading phase, the system employs a partial weight loading strategy and meticulously records mismatched keys, ensuring correct application of backbone weights and preventing errors in the head layers. Furthermore, this architecture possesses multi-scale feature extraction capabilities, outputting feature maps at three different scales. This characteristic makes it excel in medical imaging tasks, enabling precise localization of both small objects like carious lesions and larger objects like ulcer regions.

(2) This study adopts an advanced two-stage training pipeline to improve model convergence and resource efficiency. In the initial freezing stage (0-50 epochs, batch size 32), the backbone network is frozen to reduce memory consumption and stabilize early-stage training while fine-tuning the detection head.

In the subsequent unfreezing stage (50-676 epochs, batch size 16), all layers are trained to enable the network to better adapt to dental image characteristics; a reduced batch size is used to accommodate higher memory requirements. This phased training scheme effectively balances computational resource utilization and domain-specific feature adaptation, facilitating full model convergence.

(3) To alleviate the limitation of the small dataset, sophisticated data augmentation and standardized preprocessing strategies were adopted. MixUp augmentation (mixup=True, mixup_prob=0.5, effective probability 25%) combined with Mosaic augmentation was applied to smooth inter-sample differences and enhance the model's adaptability to variations in oral image appearance, thus improving the average precision for targets such as ulcers. During preprocessing, images were uniformly resized to 640×640 in YOLO Dataset, bounding boxes were normalized to the range [0, 1], and grayscale images were automatically converted to RGB to ensure format consistency. These strategies effectively enriched training sample diversity. Even with a dataset of only approximately 2000 images, the model obtained strong generalization ability.

(4) This work introduces an optimized composite loss function to boost detection performance. For bounding box regression, CIoU box loss (box_gain = 0.05) is adopted, which integrates overlap, distance, and aspect ratio to generate tighter bounding boxes. For classification, BCE loss (cls_gain = 0.5) with class weights [1.0, 2.0] is used to emphasize the ulcer class and improve its accuracy. For objectness, an implicit loss (obj_gain = 1.0) computed from maximum classification confidence simplifies computation while preserving robustness. This combined loss design balances high precision and reliability in object detection.

## 5. Conclusions

This study targets the low efficiency and experience dependence of traditional oral diagnosis, responding to the demand for oral health digital detection. By integrating deep learning and computer vision, a real-time dental health detection system based on an improved YOLOv8 model is established, forming a complete technical route from theory to clinical application.

To overcome the challenges of low lesion contrast, large scale variation, and artifact interference in oral images, the proposed model is improved on three key aspects based on YOLOv8: introducing the SE channel attention mechanism to enhance lesion feature representation; optimizing the feature pyramid structure to strengthen multi-scale feature fusion and small lesion detection; and adopting a two-phase freeze-thaw training strategy with CIoU loss to refine bounding box regression.

This study constructs a multimodal oral dataset of over 3,000 images across 12 categories, and improves model robustness using MixUp augmentation and transfer learning. INT8 quantization and network pruning produce a lightweight model with 1.2M parameters and 20 FPS real-time inference on embedded devices. The deployed system supports mobile APP and cloud platforms with dual offline-online diagnosis.

The key innovations include an attention mechanism and multi-scale fusion for oral imaging, as well as a reusable technical paradigm covering data, model, optimization and deployment. Limitations involve insufficient rare-case samples and the lack of lesion segmentation and quantitative analysis. Future work will expand the multi-center dataset to over 10,000 samples, integrate U-Net for precise lesion segmentation, and explore hybrid YOLO-Transformer architectures to enhance detection performance and generalization.

**Acknowledgments**

This paper is supported by Tianjin College Students' Innovation Training Program Project (No.202510069067).

## References

[1] M. Lin, L. Zou, R. Gao and R. Peng, "Research on visual image processing based on YOLO algorithm," 2025 2nd International Conference on Digital Image Processing and Computer Applications (DIPCA), Xi'an, China, 2025, pp. 10-14.

[2] L. Hu, "An Improved YOLOv5 Algorithm of Target Recognition,"2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), Changchun, China, 2023, pp.1373-1377.

[3] Li F, Sun H, Wang F, et al. Lightweight optimization of YOLOv8m for robotic vision-based snack cucumber sorting and palletizing. Journal of Agriculture and Food Research, 2025, 23102248-102248.

[4] G. -H. Li, T. -C. Hsung, W. -K. Ling, W. Yu-Hang Lam, G. Pelekos and C. McGrath, "Automatic Site-Specific Multiple Level Gum Disease Detection Based on Deep Neural Network," 2021 15th International Symposium on Medical Information and Communication Technology (ISMICT), Xiamen, China, 2021, pp. 201-205.

[5] Qi S, Fu Y, Zhang Q, et al. Localisation and classification of multi-stage caries on CBCT images with a 3D convolutional neural network. Clinical Oral Investigations, 2025, 29: 246.

[6] Atni M H M, Rosdy M N M M N, Tajudin M A A M, et al. Development and evaluation of a multi-model stacking approach for caries risk assessment in adults using supervised machine learning. British Dental Journal, 2025, (prepublish):1-7.

[7] Lin W C, Chiang C P. Artificial intelligence measurement of multi-layer tooth structures using semantic segmentation and computer vision. Journal of Dental Sciences, 2024, 20(1): 723-725.

[8] Ghorbani M, Nozari S, Nekooei M, et al. Artificial Intelligence for Root Canal Segmentation on Radiographic Images: A Scoping Review. Research Square, 2025-03-19.

[9] Edik M, Çelebi F, Çukurluoğlu A. Deep- learning- based detection of open- apex teeth on panoramic radiographs using YOLO models. Oral radiology, 2025: 1-11.

[10]Mendes C A, Quintanilha P B D, Pessoa P C A, et al. Automated Tooth Detection and Numbering in Panoramic Radiographs Using YOLO. Procedia Computer Science, 2025, 2561318-1325.

[11]Mao Z, Li X, Hu S, et al. A GPU accelerated mixed-precision Smoothed Particle Hydrodynamics framework with cell-based relative coordinates. Engineering Analysis with Boundary Elements, 2024, 161113-125.

[12]Zhuo Y, Kirkpatrick A W, Couperus K, et al. The Trauma THOMPSON Challenge Report MICCAI 2023. Lecture Notes in Computer Science, 2025: 61-71.

[13]Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention. MICCAI 2015. Springer International Publishing, 2015: 234-241.

[14]Liu J, Zhang H, Chen J, et al. Automated detection and segmentation of dental caries using a novel cascaded learning approach. Biomedical Signal Processing and Control, 2025, 96: 106498.

[15]Yeh J-Y, Wu S-Y. Probability-Based Two-Stage CNN for Pulmonary Embolism Detection in Computed Tomography Pulmonary Angiography. International Journal on Engineering, Science and Technology, 2023, 4(4): 344 – 367..

[16]Casalegno F, Newton T, Daher R, et al. Caries detection with near -infrared transillumination using deep learning. J Dent Res, 2019, 98(11): 1227-1233.24

[17]Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Advances in neural information processing systems, MIT Press, 2014, 27: 2672 - 2680.