Research on Intelligent Detection Algorithm of Dental Diseases Based on YOLO Algorithm

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

Abstract: Amid the digital transformation of stomatology, deep learning-driven intelligent dental disease diagnosis is a research hotspot, yet traditional YOLOv5 has high small lesion miss detection and insufficient feature fusion in oral endoscopic image analysis. This study proposes an improved YOLOv5 algorithm, integrating CBAM and BiFPN for enhanced tiny lesion feature extraction and structured pruning with INT8 quantization for model lightweighting. We built a high-quality annotated dataset with tertiary hospital clinical images and MICCAI2023 public dataset (split 7:2:1 for training, validation and testing), and developed a PyQt5-based cross-platform system for clinical chairside use. The improved model achieved 89.5% precision, 88.1% recall, 88.8% F1-score, 92.3% mAP@0.5 (4.7 percentage points higher than original YOLOv5s) and 78.6% mAP@0.5:0.95 on the independent test set. Ablation experiments confirmed CBAM and BiFPN significantly boosted detection performance with only slight parameter growth. This algorithm resolves YOLOv5's application limitations, providing an accurate and efficient intelligent auxiliary diagnosis scheme for clinical chairside scenarios, improving early dental disease diagnosis in primary care and advancing the digital and intelligent development of oral healthcare to support the Healthy China strategy.

Keywords: YOLOv5; Dental Disease Detection; Oral Endoscopy; CBAM Attention Mechanism; Bifpn Feature Fusion

1. Introduction

With the continuous improvement of national living standards and the profound awakening of national health awareness, the attention paid to oral health by all sectors of society has risen significantly, and oral medical services are transforming from the traditional disease treatment and functional restoration to a modern diagnosis and treatment model featuring precision, efficiency and intellectualization. The in-depth integration of artificial intelligence technology and stomatology has become the core engine driving the digital transformation of oral healthcare. Among them, the intelligent diagnosis technology for dental diseases based on deep learning has emerged as a research hotspot in the field of oral medical artificial intelligence due to its technical advantages in lesion recognition and quantitative analysis [1].

Early and accurate diagnosis of dental diseases is a key prerequisite for oral health management. However, the diagnosis of common diseases such as dental caries and periodontal disease still highly relies on clinicians' experience. Tiny lesions like early dental caries have obscure textures and low contrast in oral endoscopic images, leading to easy missed diagnosis and misdiagnosis in manual examination. In addition, the shortage of professional clinicians in primary medical institutions further restricts the popularization of early diagnosis [2]. Deep learning object detection models provide a feasible solution to this pain point. The YOLOv5 series models have shown great potential in medical image analysis due to their lightweight architecture, high inference speed and excellent small object detection performance [3]. Nevertheless, when directly applied to oral endoscopic images, they still face technical bottlenecks such as high missed detection rate of small lesions caused by dense tooth arrangement, mucosal texture interference and large lesion scale differences, insufficient feature fusion efficiency, redundant model parameters with slow inference speed, and poor generalization ability due to the scarcity of dedicated datasets [4], which urgently require targeted improvement.

In recent years, scholars at home and abroad have carried out numerous studies in this direction, but obvious limitations still exist.

Domestically, Xiang Jinfan et al. (2025) improved YOLOv5 to realize caries detection in oral periapical films, but did not involve endoscopic images; Liu Feng et al. (2022) verified the feasibility of YOLOv5 in lesion detection of dental films without optimizing for small lesions and lightweight design; Ding Baichen et al. (2021) realized caries detection in oral photos based on YOLOv3, with performance failing to meet the requirements of clinical chairside diagnosis. Internationally, Fahad et al. (2024) developed a YOLOv5s edge-side caries detection application with prominent problems of missed detection of small lesions; the Oral-Mamba architecture proposed by Kim et al. (2024) only focused on lesion segmentation without achieving object detection and quantitative analysis. Existing studies have not yet formed a complete solution that balances detection accuracy, inference efficiency and adaptability to clinical chairside scenarios.

Based on clinical demands, this study takes overcoming the application bottlenecks of YOLOv5 in the intelligent diagnosis of dental diseases as the core goal. We reconstruct the Backbone and Neck modules of the model by introducing the CBAM attention mechanism [5] and BiFPN multi-scale feature fusion architecture [6] to enhance the feature extraction capability for tiny lesions, then realize lightweight optimization through structured pruning and INT8 quantization. Combined with the self-constructed annotated dataset and the MICCAI2023 dataset, we complete model training and verification, and develop a cross-platform intelligent diagnosis platform based on PyQt5, ultimately forming a set of solutions adapted to clinical chairside scenarios.

This research can not only fill the theoretical gap of YOLOv5 in the subdivision field of oral healthcare and provide a theoretical paradigm for the improvement of deep learning models, but also assist primary clinicians in making accurate diagnoses by improving the recall rate and inference speed of early caries detection, lower the clinical application threshold of intelligent diagnosis technology, drive the digital transformation of the oral medical industry [7], and contribute to the realization of the strategic goal of "Healthy China". This study is strictly limited to the diagnosis of dental caries and hard tissue defects of teeth under oral endoscopic images, focusing on the real-time detection and platform deployment in chairside scenarios. It aims to reveal the technical shortcomings of the original YOLOv5 model, verify the actual effect of the improved modules [8], quantify the impact of lightweight technologies, and test the clinical adaptability of the intelligent diagnosis platform.

2. Theoretical Foundation for Intelligent Diagnosis of Dental Diseases

The digitization and intelligent development of stomatology have increasingly highlighted the core role of medical imaging in disease screening. Unlike traditional radiographic imaging that relies on radiation, oral endoscopy can directly acquire color optical images of tooth surfaces and soft tissues under radiation-free conditions. It clearly reveals early caries, enamel cracks, and gingival status, featuring real-time imaging, operational flexibility, and patient-friendliness, making it suitable for primary healthcare and routine screening.

The realization of an intelligent diagnosis system for dental diseases relies on the interdisciplinary integration of multiple theoretical fields. This paper aims to construct a systematic and in-depth theoretical framework to provide comprehensive theoretical support for the YOLO-based disease detection method in oral endoscopic images. This section will elucidate the complete theoretical underpinnings by analyzing the unique challenges of oral endoscopic images, explaining the principles of single-stage object detection centered on YOLOv5, demonstrating the optimization logic behind introducing attention mechanisms and advanced feature fusion networks, and establishing a quantitative evaluation system for model performance.

2.1 Pathological Features and Classification Criteria of Common Dental Diseases

Oral diseases are diverse and complex in classification. Dental caries (see Figure 1 (a)) is one of the most prevalent oral health issues. Its pathological process can generally be divided into three stages: early, intermediate, and advanced. Pulpitis (as shown in Figure 1 (b)) often develops from caries. Its primary pathological feature is an inflammatory reaction within the pulp tissue, clinically characterized by spontaneous pain, which typically intensifies at night. Periodontitis is another classic oral disease (as shown in Figure 1 (c)). According to the classification criteria of the American Academy of Periodontology, it can be categorized into subtypes such as chronic periodontitis,

aggressive periodontitis, and periodontitis as a manifestation of systemic diseases. Its typical pathological features include gingival inflammation, periodontal pocket formation, and alveolar bone resorption.

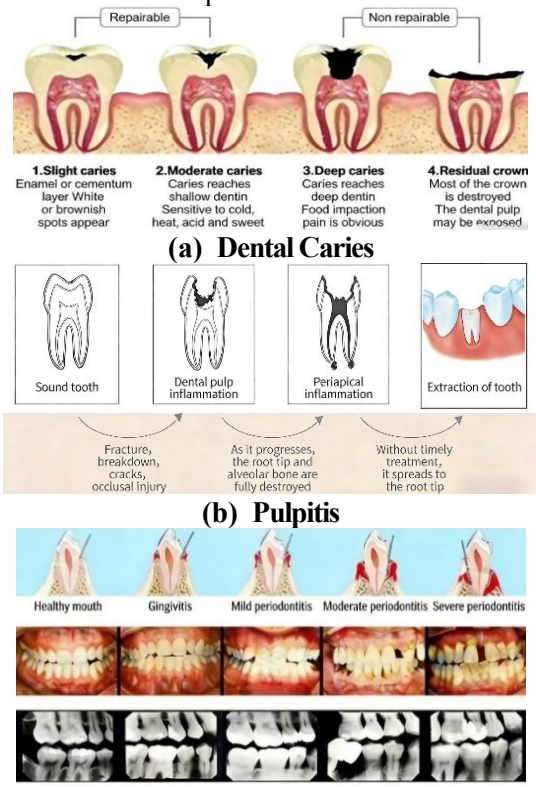


Figure 1. Common Dental Diseases

2.2 Oral Endoscopic Image Enhancement Techniques

To address the aforementioned challenges, systematic data augmentation is a crucial theoretical step for improving model generalization capability. By applying a series of controllable transformations to training images to simulate real-world imaging variations, it expands the training data distribution. Its core includes two major categories: geometric augmentation and photometric augmentation.

Geometric Augmentation: Aims to make the model invariant to changes in target shape, position, and viewpoint through spatial transformations. The core mathematical operation is affine transformation, which can be represented by a transformation matrix in homogeneous coordinates:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} s_x \cos\theta & -s_y \sin\theta & t_x \\ s_x \sin\theta & s_y \cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (1)$$

where (x, y) are the original pixel coordinates, and (x', y') are the transformed coordinates. Parameters include: rotation angle θ, scaling factors s_x and s_y, and translation amounts t_x and t_y. By randomly sampling these parameters, training samples with random rotation, scaling, translation, and shearing can be generated, thereby enhancing the model's spatial invariance.

Photometric Augmentation: Modifies pixel values to simulate complex lighting conditions and color variations, enhancing the model's robustness to illumination fluctuations. Main operations include:

Brightness and Contrast Adjustment: Can be expressed as a linear transformation:

$$G(x) = a * f(x) + b \quad (2)$$

where $f(x)$ represents the original image pixel, $G(x)$ represents the output image pixel. The parameter a (requiring a > 0) is referred to as the gain, used to control contrast adjustment. The parameter b is typically called the bias, used to control brightness adjustment.

Color Saturation and Hue Adjustment:In HSV or HSL color space, randomly scaling or shifting the saturation channel (S) and hue channel (H) can simulate variations in light source color temperature and object surface color perception. The formulas are:

$$H' = (H + \Delta H) \bmod 1 \quad (3)$$

$$S' = clip(\alpha_s \bullet S, 0, 1) \quad (4)$$

Where ΔH is randomly sampled within a limited interval, and $\bmod 1$ ensures the cyclic continuity of hue values. α_s is randomly sampled within a reasonable range, and the $clip$ operation ensures the result stays within the valid range [9].

Adding Random Noise: Injecting Gaussian noise can make the model more robust to image sensor noise.

$$G(x, y) = \frac{1}{2\pi\sigma^2} exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (5)$$

where σ controls the noise level. For impulse noise that may be present in endoscopic images, median filtering is more effective in removing noise while preserving edges:

$$I_{out}(x, y) = median\{I(x + i, y + j) | (i, j) \in W\} \quad (6)$$

2.3 YOLOv5 Algorithm Principle

YOLOv5, an excellent and well-balanced engineering implementation within the YOLO series, was selected as the baseline model for this study. It consists of four modules connected in series: the Input, Backbone, Neck, and Head.

The specific connections and functional divisions of these modules can be clearly illustrated by its network framework diagram (as shown in Figure 2):

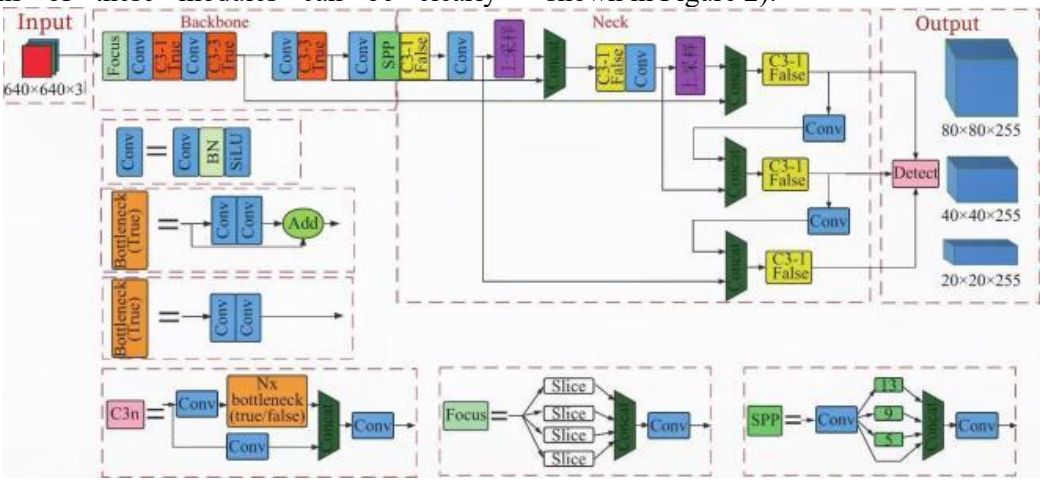


Figure 2. YOLOv5 Network Framework Diagram

Input: Employs Mosaic data augmentation, which randomly stitches four training images together, providing richer contextual information and small object samples within a single training iteration. Its mathematical essence is the expansion of the joint distribution of samples within a batch. Additionally, adaptive anchor box calculation (AutoAnchor) is used, dynamically computing the most appropriate prior anchor box sizes on the training set annotations via the K-means clustering algorithm, leading to better model initialization.

Backbone: CSPDarknet. Its core innovation is the Cross-Stage Partial Network (CSPNet) structure. This structure splits the feature map channels into two parts: one undergoes deep feature transformation, while the other bypasses via a shortcut connection, and the two are merged later. For an input feature tensor X, the operation of a CSP module can be summarized as: split X along the channel dimension; pass one part through multiple stacked Bottleneck residual blocks for transformation; concatenate the output with the other part; and finally pass through a convolutional layer. This design significantly reduces computational cost and mitigates gradient redundancy.

Neck: Path Aggregation Network (PANet). It combines top-down (Feature Pyramid Network, FPN) and bottom-up bidirectional fusion paths. First, the FPN path generates features through upsampling and fusion. Subsequently, PANet adds a bottom-up path, transmitting high-resolution shallow features (rich in localization information) back to deeper layers, enhancing the localization capability of the entire feature pyramid. This structure ensures sufficient interaction between deep semantic information and shallow detail information.

Head: Corresponds to the "Output" module in Figure 2, outputting feature maps at three different scales responsible for detecting small, medium, and large objects, respectively. At each spatial location (or grid cell) of each scale, the model predicts multiple anchor boxes. For each prediction, it outputs a vector containing: bounding box coordinate offsets, object confidence score, and conditional probabilities for each class.

The model is trained by optimizing a multi-task loss function, which is a weighted sum of bounding box regression loss, confidence loss, and classification loss:

$$L_{total} = \lambda_{box}L_{box} + \lambda_{obj}L_{obj} + \lambda_{cls}L_{cls} \quad (7)$$

The object detection task can be formally defined as: for an input image I, the model needs to output a set of bounding boxes $B = (b_k)$ and their corresponding class labels $C = (c_k)$, where $b_k = (x_k, y_k, w_k, h_k)$ represents the center coordinates, width, and height of a box. In the field of deep learning, mainstream detectors are divided into two-stage paradigms represented by Faster R-CNN and single-stage paradigms represented by YOLO.

Two-stage detectors adopt a cascaded "proposal + detection" structure, first generating region proposals and then classifying and regressing bounding boxes for each region, offering high accuracy but slower speed.

Single-stage detectors reframe the detection task as a unified, dense regression and classification problem, enabling end-to-end prediction and holding a significant advantage in speed.

2.4 Synergistic Optimization Theory of Attention Mechanism and Feature Fusion

To enable the model to focus more precisely on lesion areas and effectively fuse multi-scale information, this research primarily focuses on integrating the Convolutional Block Attention Module (CBAM) and the Weighted Bidirectional Feature Pyramid Network (BiFPN) into the YOLOv5 baseline. The synergistic effect of these two mechanisms theoretically provides a guarantee for improving detection performance in complex scenarios.

Convolutional Block Attention Module (CBAM): A lightweight, general-purpose attention module that sequentially computes attention weights along two independent dimensions: channel and space, forming a "channel attention map" and a "spatial attention map" [10]. Its computational process is as follows:

First, the channel attention module uses both Global Average Pooling (GAP) and Global Max Pooling (GMP) to aggregate spatial information, generating two different context descriptors. These descriptors are fed through a shared Multi-Layer Perceptron (MLP, typically a fully connected network with one hidden layer), then summed, and finally passed through a Sigmoid activation function to generate channel attention weights. This process allows the model to learn and emphasize feature channels most relevant to specific pathologies (e.g., the texture of demineralization in caries, the red channel in gingivitis).

Next, the features weighted by channel attention are fed into the spatial attention module. This module performs average pooling and max pooling along the channel dimension separately, resulting in two feature maps. These are concatenated and fused through a convolutional layer, then passed through a Sigmoid function to generate spatial attention weights. This enables the model to focus on lesion areas in the spatial domain of the feature map, suppressing irrelevant background (e.g., saliva, healthy gingiva). Embedding CBAM at key positions in the backbone network can guide the model's resources towards more discriminative features.

Weighted Bidirectional Feature Pyramid Network (BiFPN): Introduces significant improvements over the standard PANet's feature fusion approach [11]. The core innovations of BiFPN lie in two aspects:

Simplified and Optimized Network Structure: It removes nodes in PANet that have only one input edge and adds extra skip connections between original input and output nodes at the same level, forming a more efficient and rapid bidirectional (top-down + bottom-up) information flow.

Learnable Feature Weights: When fusing features from different resolutions, traditional methods use direct summation or concatenation. BiFPN employs fast normalized fusion:

$$P_l^{out} = \sum_i \frac{w_i}{\epsilon + \sum_j w_j} \cdot P_{l,i}^{in} \quad (8)$$

Where i iterates over input features (e.g., from the previous layer, next layer, and same-level skip connection), w_i are learnable scalar weights (typically ensured to be non-negative via ReLU), and ϵ is a small constant for numerical stability. This mechanism allows the network to automatically learn and assign greater weights to more important input features during training (e.g., higher-resolution feature maps should be assigned higher weights for detecting tiny lesions), thereby achieving adaptive, non-equal multi-scale feature fusion. Theoretically, this can more effectively aggregate information and improve detection accuracy, especially for small objects.

2.5 Quantitative Evaluation Theory for Model Performance

To scientifically and objectively measure and compare the performance of intelligent diagnostic models, it is essential to establish a rigorous, quantifiable evaluation framework. The core of this framework lies in systematically comparing the model's predictions (predicted bounding boxes and their classes) with expert-annotated "gold standard" (ground truth boxes) and digitizing detection capability through a series of mathematical metrics. These metrics are key for determining a model's potential for clinical application and for comparing the advantages and disadvantages of different algorithms.

The evaluation begins with a fundamental geometric measure—Intersection over Union (IoU)—which measures the overlap between a predicted bounding box A and a ground truth bounding box B. The calculation formula is:

$$\text{IoU} = \frac{A \cap B}{A \cup B} \quad (9)$$

Based on IoU comparison, each prediction can be classified into three basic types: True Positive (TP, correct prediction), False Positive (FP, false alarm), and False Negative (FN, missed

detection). This leads to four core evaluation metrics [12], whose definitions, calculation formulas, and clinical significance are shown in the table 1.

3. Research Protocol and Dataset Construction

3.1 Research Protocol

This study employs YOLOv5 as the foundational framework, integrating the Focus module, CSP Darknet backbone network, and FPN/PAN multi-scale feature fusion mechanism to significantly enhance the accuracy and reliability of the dental disease intelligent diagnosis system. The Focus module optimizes the feature extraction process through slice manipulation and channel recombination techniques, effectively preserving fine textures and core lesion information in dental images [13]. The CSP Darknet backbone network combines staged cross-layer connections and deep separable convolution strategies, markedly improving the models representation capability for complex dental structures while achieving efficient computational resource utilization. The FPN/PAN architecture, leveraging multi-level feature aggregation mechanisms, accurately identifies typical pathological manifestations such as caries, dental fissures, and periapical periodontitis across different scales. By integrating the Mosaic data enhancement method with the GIoU loss function, the feature acquisition capability is significantly improved, and bounding box prediction precision is enhanced, thereby substantially improving the models generalization ability and diagnostic accuracy in oral disease classification and localization tasks.

Table 1. Definitions and Significance of Core Object Detection Evaluation Metrics

Metric	Formula	Physical Meaning and Clinical Correspondence
Precision	$P = \frac{TP}{TP + FP}$	Measures the reliability of detection results. High precision indicates a low false alarm rate for lesion areas highlighted by the system, helping to build clinician trust in the results.
Recall	$R = \frac{TP}{TP + FN}$	Measures the coverage of detection results. High recall indicates a low missed diagnosis rate, which is crucial for early disease screening and comprehensive assessment.
F1-Score	$F_1 = \frac{2 * (P * R)}{(P + R)}$	Harmonic mean of Precision and Recall. Used for a balanced overall assessment, especially more informative when positive and negative samples are imbalanced (e.g., lesion areas are much smaller than background).
Average Precision (AP)	$AP = \int_0^1 P(r)\, dr$	Comprehensively measures the performance of a single class across all decision thresholds. Its value is the area under the Precision-Recall curve, reflecting the model's stable detection capability for a specific disease (e.g., dental caries).

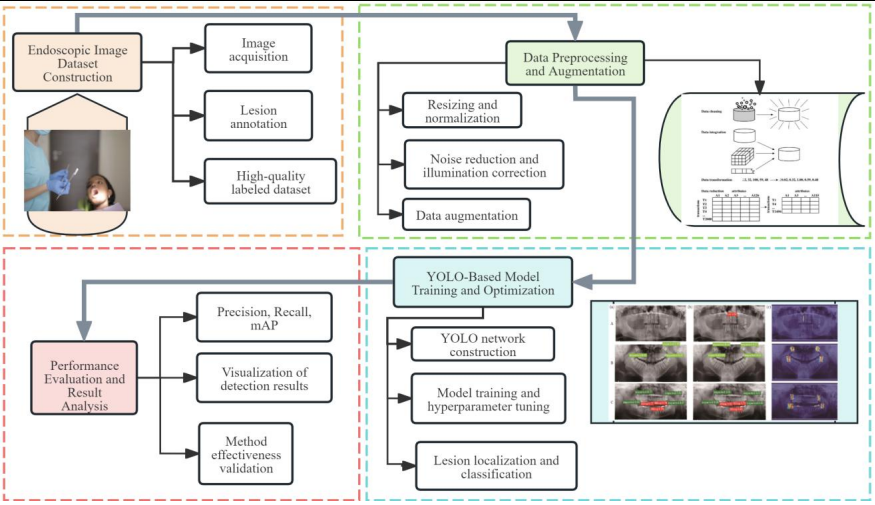


Figure 3. Flowchart of the Study Protocol

The research protocol of this study is illustrated in Figure 3, systematically encompassing the following steps: First, constructing a high-quality annotated endoscopic image dataset; Second, implementing targeted data preprocessing and enhancement to improve model robustness; Third, training and optimizing the model based on the YOLO network to achieve automatic lesion localization and classification; Finally, validating the methods effectiveness through

quantitative metrics and visual analysis. [14]

3.2 Image Dataset and Preprocessing Methods

The data in this study were derived from clinical endoscopic examinations, which exhibit significant non-standardization due to variations in imaging angles, light sources, and patient anatomical differences. To construct a high-quality dedicated dataset, this study adhered to a standardized workflow encompassing data acquisition, screening, annotation, preprocessing, and enhancement.

During the data collection phase, we selected clinical endoscopic examination cases from the Department of Stomatology of a tertiary hospital in China from January 2023 to June 2024, systematically collecting 286 original images of 34 patients.[15] These images encompassed various typical clinical scenarios, including normal teeth, gingivitis, periodontal disease of varying degrees, and caries, ensuring the representativeness and broad coverage of the data.

To ensure model reliability, we conducted rigorous cleaning and screening of the original images. First, visual evaluation was performed to exclude samples with blurred images, abnormal exposure, lesion occlusion, or incomplete key information. Subsequently, based on clinical diagnostic criteria, invalid data without clear lesion annotations or resolution below 1080×1080 pixels were removed. Ultimately, 203 valid images were retained (as shown in Figure 4), comprehensively demonstrating the typical characteristics of common oral diseases. All valid images were uniformly resized to 640×640 pixels prior to model input to meet the input size requirements of the improved YOLOv5 model [16].

Figure 4. Endoscopic Dataset

The data annotation adopts the rectangular bounding box method commonly used in object detection. The annotation content includes the precise bounding box coordinates of the lesion area and the corresponding category labels, which are classified according to clinical standards such as caries severity and periodontal disease severity. To ensure consistency and accuracy in annotation, two dentists with over 5 years of clinical experience jointly completed the annotation. Prior to annotation, we organized a unified training session to clarify the bounding box dimensions and category definition criteria. After completion, the results were cross-validated and reviewed, with divergent samples subject to expert deliberation to reach consensus. The ultimately formed high-quality annotated dataset provides a reliable basis for model training and strengthens the technical support for clinical diagnosis. Prior to model input, all images were uniformly resized and normalized.

To further enhance the models generalization capability, we introduced a multi-type data augmentation strategy. Without altering the essential features of lesions, the training set images underwent random horizontal flipping, random rotation within a ±15° range, and perturbations in brightness and contrast. These operations effectively simulated the diversity caused by variations in imaging angles and light intensity in clinical settings, thereby improving the models adaptability and robustness to complex environments.

3.3 Experimental Procedure

To systematically evaluate the tooth disease detection model based on improved YOLOv5, this study designed a comprehensive experimental protocol encompassing

environmental setup, data preparation, model training, performance evaluation, and ablation analysis.

The experimental hardware consists of a server equipped with an NVIDIA RTX 4090 GPU (24GB video memory). Software development was based on the PyTorch 2.1.0 framework, with CUDA 11.8 for acceleration. The dataset integrates self-collected clinical endoscopic images and publicly available data (partial data from MICCAI 2023), which were cleaned and standardized, resulting in over 5,000 images of various teeth and lesions. Following the data partitioning method used in similar studies, the dataset was randomly divided into training, validation, and test sets in a 7:2:1 ratio, ensuring balanced distribution across categories to evaluate the models generalization ability.

This study employed YOLOv5s as the baseline model. Input images were uniformly resized to 640×640 pixels. Training was conducted using the stochastic gradient descent (SGD) optimizer, with an initial learning rate of 0.01, momentum of 0.937, weight decay of 5×10^{-4}, and batch size of 16 [17]. The total number of training epochs was set to 300, and a cosine annealing scheduling strategy was employed to dynamically adjust the learning rate to promote model convergence. To enhance generalization, data augmentation techniques such as random horizontal flipping, ±15° rotation, and brightness/contrast perturbations (±20%) were applied during training. The loss function was composed of weighted classification loss, bounding box regression loss, and object confidence loss.

4. Analysis Of Experimental Results

4.1 Oral Disease Test Results

Figure 5 shows the comparison of the results before and after the detection of some intraoral endoscopic images, and Figure 6 shows the detection results of panoramic dental pieces.

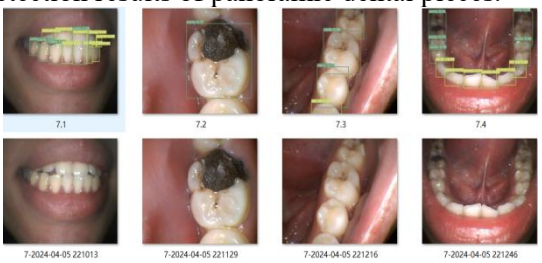


Figure 5. Intraoral Endoscopic Image Detection Results

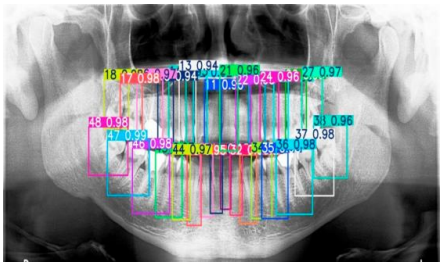


Figure 6. Panoramic Tooth Chip Detection Results

4.2 Data Analysis

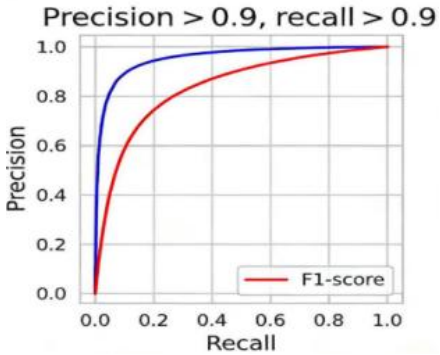
The dental disease detection model based on the improved YOLOv5 constructed in this study showed excellent detection performance on the independent test set, as shown in Table 2, and the model achieved a true positive rate (TPR) of 90% on the main target caries, indicating that the model can effectively identify the vast majority of caries lesions. The accuracy rate reached 89.5%, indicating that the false positive rate of the model is low, which can effectively avoid misjudging healthy teeth as diseased teeth [18].

Table 2. Comprehensive Performance Evaluation Results of the Model

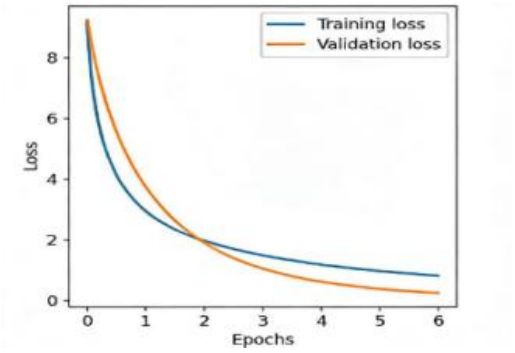
Evaluate the indicators	Value (%)	Description
Precision	89.5	The model predicts the proportion of true positive in a positive sample
Recall	88.1	The proportion of the true positive sample that is correctly detected by the model
F1-Score	88.8	The harmonized average of accuracy and recall
mAP@0.5	92.3	The average accuracy of the intersection and merge ratio threshold is 0.5
mAP@0.5:0.95	78.6	The average accuracy of the intersection and merge ratio threshold ranges from 0.5 to 0.95 (step size 0.05).

In terms of classification performance, the F1 score curve in Figure 7(a) further validates the overall performance of the model, with the F1 score remaining above 0.85 in the area where the recall is greater than 0.8, showing a good balance between accuracy and recall. From the loss curve shown in Figure 7(b), it can be observed that both training loss and validation loss continue to decrease with the increase of the number of training rounds, the training basically converges after the 6th round, the loss value tends to be stable, and the gap between training loss and validation loss is small, indicating that the model does not have obvious overfitting, and the final

training loss is about 1.2 and the validation loss is about 0.8, both of which are at a low level and remain stable. This indicates that the model achieves good performance on both the training set and the validation set.



(a)F1-Confidence



(b)Training and Validation Loss

Figure 7 Simulation Result Curves

4.3 Ablation Experimental Analysis

In order to quantitatively evaluate the contribution of the CBAM attention mechanism and BiFPN feature fusion network to the model performance [19], we conducted systematic ablation experiments, and the results are shown in Table 3.

Table 3 Comparison of Ablation Experimental Results

Experimental group	Model configuration	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameter quantity (M)
A	YOLOv5s (Baseline)	87.6	73.2	7.2
B	Baseline + CBAM	89.7 (+2.1)	75.8 (+2.6)	7.4
C	Baseline + CBAM + BiFPN	92.3 (+4.7)	78.6 (+5.4)	7.9

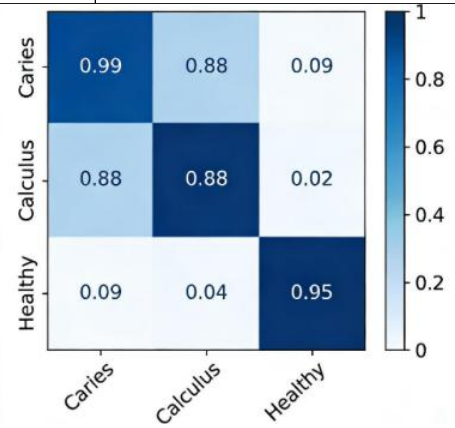


Figure 8. Confusion Matrix

From the confusion matrix in Figure 8, it can be seen that the true positive rate of its caries category reaches 90%, the false positive rate is only 10%, the detection accuracy of the tartar category is 87%, the identification accuracy of healthy teeth is the highest at 95%, and the model has relatively little confusion between caries and tartar, with a cross-false positive rate of less than 5%.

The experimental results show that the introduction of CBAM attention mechanism (experimental group B) alone increases mAP@0.5 by 2.1 percentage points, which proves that it can effectively guide the model to focus on the characteristic channels and spatial regions related to the lesion and suppress background interference. After further introduction of BiFPN structure (experimental group C), mAP@0.5 was significantly improved, reaching 92.3%, which was 4.7 percentage points higher than that of the baseline model. This confirms that BiFPN significantly enhances the detection of lesions with large scale differences in oral endoscopic images, especially early microcaries lesions, by efficiently fusing multi-scale features. The continued improvement in model performance resulted in only a small parameter increase (from 7.2M to 7.9M), reflecting a good balance between efficiency and performance in the improved scheme. Similarly, a study has observed a stable increase in mAP after introducing an attention module in YOLOv8 and adjusting the feature pyramid.

5. Conclusion

This paper selects the YOLOv5 network model. Through systematic design, parameter adjustment and experimental verification, an intelligent diagnosis system for dental diseases integrating multi-modal features is constructed. The core advantages of the YOLO algorithm lie in its one-stage and end-to-end design. Its network is usually composed of a feature

extraction backbone, a feature fusion module (such as FPN) and a detection head. Aiming at the challenges of small lesion size and blurred boundary in endoscopic images, YOLO fuses shallow detailed features and deep semantic information through the feature pyramid structure to achieve efficient multi-scale object detection.

By using deep convolutional neural networks to efficiently extract key discriminative information from images, the system realizes accurate location of periodontal regions and precise classification of a variety of typical oral diseases. Experimental results show that this method has good performance in detection accuracy, highlights the full-process automation, improves the intelligence level of the whole diagnosis and treatment of dental diseases, and further expands the application scope of medical and health services.

Acknowledgments

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

References:

[1] Li, Y., Zhang, H., & Wang, X. Artificial intelligence tools in dentistry: A systematic review on their application and outcomes. Healthcare, 2025, 13(10), 1452.

[2] Kandaswamy, D., Priya, S., & Kumar, P. Developing an AI-based application for caries index detection on intraoral photographs. Computers in Biology and Medicine, 2024, 174, 107968.

[3] Abidin, N. Z., Rahim, N. A., & Yusoff, N. Evaluating YOLO for dental caries diagnosis: a systematic review and meta-analysis. Journal of Dentistry, 2025, 154, 104542.

[4] Shaji, A., Thomas, S., & Mathew, J. Visual diagnostics of dental caries through deep learning of non-standardised photographs using a hybrid YOLO ensemble and transfer learning model. International Journal of Environmental Research and Public Health, 2023, 20(7), 5351.

[5] Tan, M., & Le, Q. V. EfficientDet: Scalable and efficient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11), 2020, 3713-3727.

[6] Dogan, S., Kaya, M., & Yildirim, O. Diagnosis and classification of dental caries with deep learning methods in dental radiography. International Journal of Theoretical and Applied Physics, 2025, 17(2), 12-20.

[7] Alom, M. Z., Hasan, M., & Yakopcic, C. U-NET model based on CBAM attention mechanism for coronary angiography segmentation. Current Medical Imaging, 2024, 20(10), 1089-1100.

[8] SHORTEN C, KHOSHGOFTAAR T M. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 2019, 6: 60.

[9] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional Block Attention Module[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.

[10] PADILLA R, NETTO S L, DA SILVA E A B. A Survey on Performance Metrics for Object-Detection Algorithms. 2020 International Conference on Systems, Signals and Image Processing (IWSSIP). 2020: 237-242.

[11] Ramirez-Pedraza A, Salazar-Colores S, Cardenas-Valle C, et al. Deep Learning in Oral Hygiene: Automated Dental Plaque Detection via YOLO Frameworks and Quantification Using the O'Leary Index. Diagnostics. 2025, 15(2): 231.

[12] Evaluation of the Performance of a YOLOv10-Based Deep Learning Model for Tooth Detection and Numbering on Panoramic Radiographs of Patients in the Mixed Dentition Period. Diagnostics. 2025, 15(4): 405.

[13] N. N W, M. J, X. F. Simulation-Based Robustness Evaluation of Coordination under Stochastic Demand. International Journal of Simulation Modeling,2025,24(4):730-741.

[14] Weiming Z, Patricia B, Allison L, et al. Real-World Positive Predictive Value and False Positive Rates of Laboratory-Based HIV Antigen/Antibody Testing by Age and Sex — United States, 2019–2024.JAIDS Journal of Acquired Immune Deficiency Syndromes,2025.

[15] Zhou J, Zou J, Qiu Y, et al. Polarization image fusion via analytical attention heads: A multi-scale feature integration framework. Optics and Lasers in Engineering, 2026, 2011, 09628-109628.

[16] Redmon J, Farhadi A. YOLOv3: An

Incremental Improvement. arXiv preprint arXiv:1804.02767, 2018.

[17]Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117-2125.

[18]Li A, Qiu B, Goettsch M, et al. Convolutional neural networks combined with classification algorithms for the diagnosis of periodontitis. Journal of Clinical Periodontology, 2023, 50(5): 591-603.

[19]Ozsunkar P S, Özen D Ç, Abdelkarim A Z, et al. Detecting white spot lesions on post-orthodontic oral photographs using deep learning based on the YOLOv5x algorithm: a pilot study. BMC Oral Health, 2025, 25(1): 1128.

