2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

DiMNet: Multi-Label Detection Algorithm for Panoramic Radiographs

Yanfu Li[1], Ruijie Huang[2], Pei Zhou[3], Zhengzhong Zhu[3] and Jiangping Zhu[3] (✉)

[1] National key laboratory of fundamental science on synthetic vision, Sichuan University, Chengdu, China
[2] West China School/Hospital of Stomatology, Sichuan University, Chengdu, China
[3] College of Computer Science, Sichuan University, Chengdu, China
zjp16@scu.edu.cn

Abstract. In recent years, deep learning has been widely applied to single-target detection tasks in dental images, achieving promising results. Existing methods aiming to achieve multi-label detection rely heavily on fully annotated data. However, due to the difficulty in obtaining such fully annotated data, the detection accuracy remains low, failing to meet the requirements of clinical diagnosis. To address this limitation, we propose DiMNet, a end-to-end multi-label object detection model based on an improved DiffusionDet, which incorporates multi-stage training, weight transfer, and cross-stage guidance to enable the model to be trained on partially annotated data, thereby improving detection accuracy. Additionally, we enhance the feature extraction backbone by integrating the Mamba model, leveraging its linear-time sequence modeling approach to maintain high accuracy while significantly improving inference speed. The model is capable of identifying dental pathologies in panoramic X-ray images while simultaneously providing the quadrant and tooth number of the affected tooth, maintaining high accuracy and fast inference speed, thereby meeting the requirements of fully automated diagnosis. During the experiments, we utilized DENTEX2023, which features a multi-level structure, enabling a comprehensive evaluation of the effectiveness of the proposed improvements in DiMNet. Experimental results demonstrate that DiMNet achieves AR scores of 71.7% for quadrant detection, 66.8% for enumeration, and 69.1% for dental pathology detection on the test dataset, accurately detecting all three targets in dental images simultaneously.

Keywords: Multi-Label Detection, Diffusion, MambaVision.

1 Introduction

In recent years, increasing economic development and improved living standards have heightened awareness of oral health [1]. However, in practice, limited public knowledge of dental care poses challenges for maintaining oral health. As the most prevalent dental disease, caries often exhibits subtle early-stage symptoms that are difficult for non-specialists to identify. Additionally, variability in clinicians' experience and subjective diagnostic methods increase risks of misdiagnosis or missed diagnosis,

potentially leading to wasted medical resources, delayed treatment, and disease progression. Furthermore, manual diagnosis is time-consuming [2] and inefficient when processing large volumes of dental images, with susceptibility to environmental interference. These issues underscore the critical need for automated detection and computer-aided diagnosis systems to improve early caries identification and optimize clinical workflows.

Panoramic X-rays, characterized by standardized imaging angles and low noise levels, significantly reduce tooth localization complexity and provide essential support for fully automated diagnostic systems [3]. Recent advances in deep learning have yielded models for specific tasks such as quadrant [4], enumeration [5] and diagnosis [6] in panoramic X-rays. While these studies achieve task-specific success, they lack end-to-end multi-label detection capabilities, unable to simultaneously localize teeth, assign numbers, and diagnose pathologies. Existing methods employ multiple steps to achieve multi-label detection. For instance, YOLOrtho [7] utilizes separate models to detect different types of targets and then merges the results through post-processing. Other methods [8] adopt end-to-end models but rely heavily on fully annotated data for training. However, such comprehensive annotations demand substantial medical expertise and time, making them scarce [9]. Notably, annotation complexity increases hierarchically (quadrant-enumeration-diagnosis), where higher-level labels depend on lower-level ones. Consequently, conventional object detection algorithms struggle to adapt to multi-stage training requirements and multi-level data characteristics [10].

To address the limitations of existing methods and achieve fully automated diagnosis based on panoramic X-rays, we propose DiMNet, an end-to-end multi-label model based on an improved DiffusionDet. This model can simultaneously provide the quadrant and numbering of teeth while performing diagnosis. Existing datasets are typically multi-level and often only partially annotated. To fully utilize these data, we adopt a multi-stage training approach based on weight transfer. The training progresses from low-level to high-level tasks, such as first training on data annotated only with quadrant information and then fine-tuning the pre-trained model on data annotated with enumeration and diagnosis labels, thereby improving the model's detection accuracy. Additionally, considering the architectural characteristics of DiffusionDet, we introduce a cross-stage guidance strategy, where the prediction boxes from the previous stage model are concatenated with the noise boxes in the current stage to assist in the noise box restoration process. Furthermore, to meet the real-time requirements of clinical diagnosis, we enhance the feature extraction backbone by integrating the Mamba model, leveraging its linear-time sequence modeling approach to maintain high accuracy while significantly improving inference speed.Ultimately, the DiMNet model achieves fully automated diagnosis, with both accuracy and speed meeting clinical requirements. The main contributions are:

- To enhance image feature extraction capabilities, we adopt the MambaVision [12] model as the backbone, leveraging the respective strengths of Convolutional Neural Network (CNN), Mamba, and Self-attention [13] for feature extraction, thereby improving both detection accuracy and model inference speed.

- To achieve multi-label detection, we redesign the detection head to simultaneously output bounding box information and classify the detected boxes into quadrant, enumeration, and diagnosis, providing a foundational framework for fully automated diagnostic systems.
- To accommodate the multi-level annotation characteristics of the dataset, we propose a multi-stage training strategy, incorporating weight transfer and cross-stage guidance to fully utilize data at different levels, thereby enhancing the overall detection accuracy of the model.

2 Related Works

2.1 Tooth Detection and Diagnosis

In recent years, the widespread application of deep learning in computer vision tasks has brought revolutionary advancements to the field of medical image analysis, particularly in the domain of oral medicine. Numerous research efforts have successfully achieved automated detection and analysis on panoramic X-ray images.

To address the challenges of dental diagnosis, Chen et al. [14] explored the development of a CNN-based auxiliary system tailored for periapical radiographs. Their research utilized a dataset of 2900 digital periapical radiographs and implemented diverse training strategies to optimize model performance. The system demonstrated promising results, with precision and recall metrics across various disease categories consistently falling within the 0.5 to 0.6 range. These findings underscore the potential of CNN-driven approaches as effective tools for enhancing diagnostic accuracy in periapical radiograph analysis. Additionally, U-Net was specifically designed for bitewing radiographs [15]. This model demonstrated robust performance, achieving a precision of 63.29%, recall of 65.02%, and an F1-score of 64.14%, underscoring its utility as a supportive tool for clinicians in identifying dental caries.

Although many of the aforementioned deep learning models have been developed for analyzing panoramic X-ray images and have achieved certain results in specific tasks such as quadrant, enumeration and diagnosis detection, they have not achieved end-to-end multi-label detection. Consequently, they are unable to accurately diagnose dental pathologies while simultaneously performing tooth localization and numbering.

2.2 Diffusion and Mamba

With the continuous advancement of deep learning algorithms, innovative models have been developed, each demonstrating remarkable capabilities in specialized domains. For instance, the Diffusion [16] algorithm, grounded in probabilistic frameworks, has achieved groundbreaking results in generative tasks by iteratively refining data distributions to produce high-quality outputs. Similarly, the Mamba [17] algorithm, leveraging state-space models, has shown exceptional efficiency in sequential data processing, significantly reducing computational overhead while maintaining high accuracy.

Currently, numerous algorithms combining Diffusion and Mamba have been applied in the field of computer vision. Zhou [18] et al. proposed the MaDiNet model for SAR target detection, defining the coordinates and dimensions of bounding boxes as a generative task. They also designed a MambaSAR model to capture intricate spatial structural information of targets and enhance the model's capability to differentiate between targets and complex backgrounds. The experimental results on extensive SAR target detection datasets achieve state-of-the-art (SOTA) performance, proving the effectiveness of the proposed network. Additionally, Ju [19] et al. introduced the VM-DDPM model based on Diffusion and Mamba for medical image synthesis. Their experimental evaluation on three datasets of different scales, i.e., ACDC, BraTS2018, and ChestXRay, as well as qualitative evaluation by radiologists, demonstrates that VM-DDPM achieves state-of-the-art performance. The outstanding results of the aforementioned studies in their respective fields validate the feasibility of Diffusion and Mamba models in computer vision tasks, providing a foundation for their application in medical image detection in this paper.

3 Methodology

3.1 DiMNet Architecture

The overall network architecture of our proposed DiMNet is illustrated in **Fig. 1**, which primarily consists of four components: Encoder, Decoder, Detection Head, and Noise Bounding Box Generation. The Encoder processes the input image, extracting multi-scale visual features from the raw image to generate a set of feature maps rich in semantic information. These feature maps provide essential contextual and local detail information for Decoder.

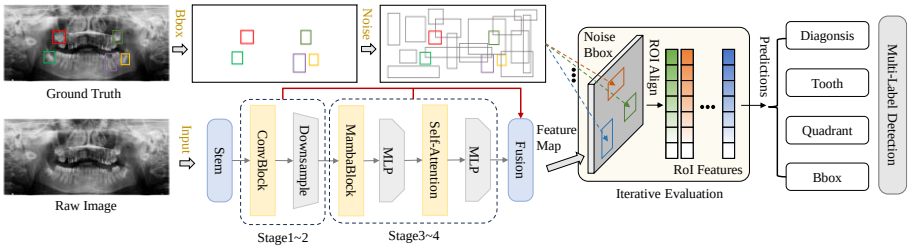

Fig. 1. DiMNet overall network architecture

During the Noise Bounding Box Generation phase, the model gradually adds Gaussian noise to the ground truth bounding boxes during training, forming a set of noisy initial bounding boxes. During inference, the initial noise bounding boxes are directly sampled from a random distribution. The Decoder takes both the encoder outputs and the noise bounding boxes as inputs, performing iterative denoising on the noisy bounding boxes.

The Detection Head receives the bounding box and feature information output by the decoder and performs final classification and regression predictions for each candidate bounding box. During training, the category heads not included in the dataset format are frozen. During inference, since each bounding box in this task has three labels, each candidate box outputs four sets of information: the detection bounding box, quadrant, enumeration and diagnosis, achieving end-to-end detection.

Mamba Encoder. The Encoder takes the raw image as input and extracts high-level feature information for the Decoder. DiMNet employs the MambaVision model as the backbone of the Encoder, leveraging a collaborative architecture of CNN, Mamba, and Transformer to enhance both the accuracy and efficiency of feature extraction. As shown in **Fig. 2**, MambaVision consists of four stages. The first two stages are composed of stacked convolutional modules and downsampling modules, enabling rapid feature extraction from high-resolution images. The latter two stages utilize MambaBlock and Self-Attention to further extract semantic information from the feature maps. DiMNet retains the Feature Pyramid Network (FPN) [20] architecture to fuse multi-scale feature maps extracted by the backbone network, enriching the semantic information passed to the Decoder.

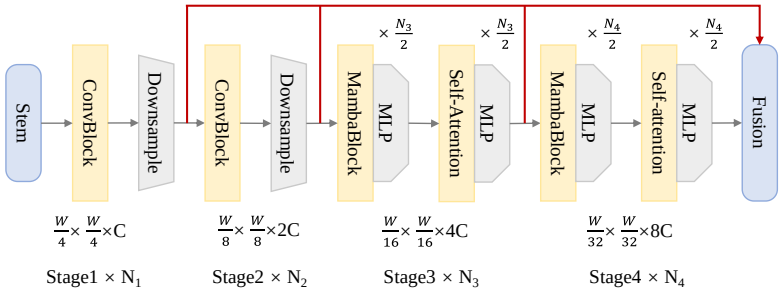

Fig. 2. Encoder detailed architecture

Specifically, given an input image of size $H \times W \times 3$, it first passes through a stem structure similar to the Inception Network [21], which consists of two consecutive 3×3 convolutional blocks with a stride of 2, outputting a feature map of size $\frac{W}{4} \times \frac{H}{4} \times 3$. The downsampling modules in Stages 1-2 are composed of two 3×3 convolutional blocks with a stride of 2, followed by batch normalization. The operation of the ConvBlock is as follows:

$$z = BN\left(\mathrm{Conv}_{3\times3}\left(GELU\left(BN\left(\mathrm{Conv}_{3\times3}(z)\right)\right)\right)\right) + z \tag{1}$$

where GELU represents the Gaussian Error Linear Unit activation function, and BN denotes batch normalization.

Stages 3-4 are composed of stacked structures of MambaBlock+MLP and Self-Attention+MLP. The MambaBlock structure, as shown in **Fig. 3**, is more suitable for vision tasks compared to the original Mamba. First, the original causal convolution is

limited to unidirectional feature extraction, making it unsuitable for vision tasks, so it is replaced with regular convolution. Second, an additional path without SSM (State Space Model) is introduced, consisting of Conv1d and SiLU activation, to compensate for the information loss caused by SSM sequence compression. The outputs of the two paths are concatenated and fed into a linear layer, ensuring that the final feature map contains both sequence and spatial information, thereby efficiently leveraging the feature extraction capabilities of both paths. Given an input X_{in}, the output X_{out} of the MambaBlock can be expressed as follows:

$$\begin{aligned} X_1 &= \mathrm{Scan}\left(\sigma\left(\mathrm{Conv}\left(\mathrm{Linear}\left(C, \frac{C}{2}\right)(X_{in})\right)\right)\right) \\ X_2 &= \sigma\left(\mathrm{Conv}\left(\mathrm{Linear}\left(C, \frac{C}{2}\right)(X_{in})\right)\right) \\ X_{out} &= \mathrm{Linear}\left(\frac{C}{2}, C\right)\left(\mathrm{Concat}(X_1, X_2)\right), \end{aligned} \tag{2}$$

here, $Linear(C_{in}, C_{out})(\cdot)$ denotes a linear layer with input dimension C_{in} and output dimension C_{out}, $Scan$ represents the state space model, σ refers to the activation function Sigmoid Linear Unit(SiLU), and $Conv$ and $Concat$ represent the 1D convolution and concatenation operations, respectively.

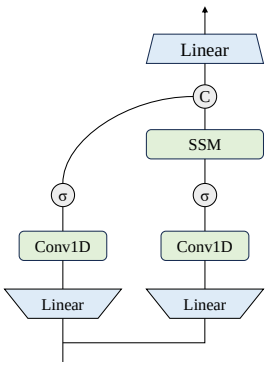

Fig. 3. MambaBlock detailed structure

Additionally, we adopt a generic multi-head self-attention mechanism in accordance with:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right)V \tag{3}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_h is the dimensionality of the attention heads. The $Softmax$ function is applied to the scaled dot-product of Q and K , followed by a weighted sum with to compute the attention output.

Multi-Label Decoder. During the training phase of DiMNet, Gaussian noise is added to the original annotated bounding boxes. This process is formulated as:

$$q(z_t|z_0) = \mathcal{N}\left(z_t\middle|\sqrt{\overline{\alpha_t}}z_0, (1-\overline{\alpha_t})I\right) \tag{4}$$

where z_0 denotes the input bounding box (defined by center coordinates x,y, width w, and height h), z_t represents the generated noisy bounding box, and $\overline{a_t}$ is the noise variance schedule.

The Decoder processes the noisy initial bounding boxes by extracting Region of Interest (RoI) features from the feature maps generated by the Encoder. This step typically employs RoI pooling or similar techniques to focus on regions most likely to contain targets, thereby avoiding redundant computation over the entire image. Subsequently, a neural network $f_\theta(z_t, t, x)$ predicts z_0 (denoised bounding box) and the corresponding class label c based on the image x. During training, model optimization and parameter updates are achieved by minimizing the L2 loss between the predicted f and the ground truth bounding box:

$$\mathcal{L}_{train} = \frac{1}{2}||f_\theta(z_t, t, x) - z_0||^2 \tag{5}$$

Additionally, to achieve multi-label detection, we incorporate four detection heads after the Decoder, respectively predicting bounding box, quadrant, enumeration, and diagnosis.

3.2 Training Strategy

To fully leverage the multi-level characteristics of the dataset, the model is trained using a three-stage progressive strategy, where detection heads for unannotated categories are frozen in each stage. The stage-specific prediction function $f_\theta(\cdot)$ is defined as:

$$f_\theta(z_t, t, x, h_q, h_e, h_d) = \begin{cases} (y_q^i, b^i), & h_q = 1, h_e = 0, h_d = 0 \quad (a) \\ (y_q^i, y_e^i, b^i), & h_q = 1, h_e = 1, h_d = 0 \quad (b) \\ (y_q^i, y_e^i, y_d^i, b^i), & h_q = 1, h_e = 1, h_d = 1 \quad (c) \end{cases} \tag{6}$$

here, h_q h_e h_d are binary flags indicating whether quadrant, enumeration or diagnosis labels are available.

It is noteworthy that these three stages are not trained independently but rather employ a **weight transfer** approach, where the model trained in the previous stage is used as the foundation for continued training on the next level of data. The advantage of weight transfer lies in its ability to fully leverage the characteristics of partially annotated data.

Additionally, the three types of annotations in the dataset are not independent; the annotations at the next level depend on those at the previous level. For example, enumeration labels are confined within quadrant bounding boxes, while diagnosis labels are further classified based on the regions defined by enumeration. Based on this, DiMNet adopts a **cross-stage guidance** strategy during training. Specifically, the input to the Decoder is not entirely composed of noisy boxes but rather combines the detection

boxes inferred by the model trained in the previous stage on the current stage's data with the noisy boxes generated for the current stage. The Decoder then extracts RoI features to achieve efficient training. For instance, when training model (b), we first use model (a) to perform inference on the current stage's data to obtain a set of predicted boxes. By filtering out predicted boxes with confidence scores greater than 0.5, we obtain valid boxes, which are then concatenated with the noisy boxes generated for stage (b). However, during the inference phase, only noisy boxes are used.

4 Experiment Design and Result Analysis

4.1 DENTEX2023 Dataset

We utilize the DENTEX2023[22] dataset for model training and validation. The dataset consists of panoramic dental X-rays collected from three institutions under standardized clinical conditions, with variations in equipment and imaging protocols. These variations reflect diverse clinical practices, thereby enhancing the model's robustness in real-world applications.

As shown in Fig. 4, DENTEX2023 provides three hierarchically annotated data types: (a) Quadrant (693 images) annotated with quadrant numbers based on the FDI system (a globally recognized dental notation standard dividing the oral cavity into quadrants 1-4); (b) Enumeration (634 images) annotated with quadrant and tooth enumeration (e.g., the third molar on the lower left is labeled "48"); (c) Diagnosis (1005 images, including a training set of 705, a validation set of 50, and a test set of 250) annotated with quadrant, tooth number, and four diagnostic categories (caries, deep caries, periapical lesions, and impacted teeth). The hierarchical annotations enable phased validation of the model's capabilities in quadrant localization, tooth counting, and pathological identification during development.

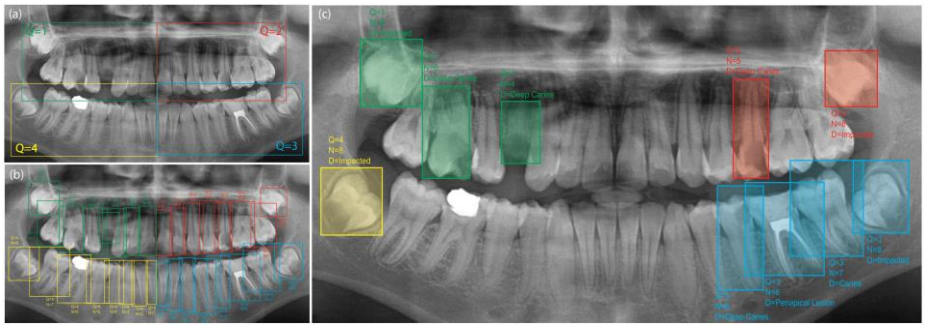

Fig. 4. DENTEX2023 dataset examples

4.2 Experimental Setup

The model training was conducted on a deep learning platform built with PyTorch framework, utilizing hardware configurations including an Intel Core i9-13900K CPU

and dual NVIDIA RTX-4090 GPUs. The training strategy incorporated dynamic learning rate adjustment: AdamW optimizer was selected with an initial learning rate of 0.003, combined with a linear warmup mechanism (warmup factor=0.01, warmup iterations=1000) to stabilize gradient updates during the initial phase. Weight decay coefficient was set to 0.0001 for model complexity control. The dataset was partitioned into training and validation sets at a 7:3 ratio, with model parameters optimized over 300 training epochs. The software environment was configured with Ubuntu 24.04 LTS operating system, accelerated by CUDA 12.3 and cuDNN 8.9.5 libraries to maximize computational efficiency.

4.3 Comparison Experiment

To quantitatively validate the superiority of the proposed diffusion-based multi-label object detection algorithm DiMNet in this chapter, we compare it with mainstream detection models including YOLOv10 [23], RetinaNet [24], Faster R-CNN [25], DETR [26], and DiffusionDet [11].

To comprehensively validate the superiority of the proposed DiMNet model in this chapter, we conduct a quantitative comparison between DiMNet and six other models (DiffusionDet, Faster R-CNN, YOLOv10, DETR, and RetinaNet) across three label categories, AR (Average Recall), AP (mean Average Precision at IoU [50:95]), FPS (Frames Per Second), and parameter count. The results are summarized in Table 1. Regarding AR and AP, DiMNet achieves optimal performance across all three label categories (quadrant, tooth number, and dental pathology). For quadrant detection, DiMNet attains AR=71.7% and AP=42.6%, establishing a robust foundation for subsequent tooth numbering. In tooth numbering detection, it achieves AR=66.8% and AP=37.3%, enabling precise identification of individual teeth and their numerical designations. Compared to single-stage object detection models (YOLOv10 and RetinaNet), DiMNet demonstrates significant improvement in detection accuracy while maintaining comparable FPS and parameter efficiency. When evaluated against two-stage models, DiMNet surpasses all competitors across all metrics.

Table 1. Comparison of experimental results of different models

Model	Quadrant		Enumeration		Diagnosis		Params	FPS
	AR	AP	AR	AP	AR	AP		
YOLOv10	55.3	32.9	49.4	27.4	51.2	25.3	35.5	103
RetinaNet	63.2	41.4	54.3	35.7	59.2	33.7	**24.4**	**138**
Faster R-CNN	59.9	39.9	48.2	25.6	50.9	24.2	44.5	83
DETR	68.3	41.2	60.6	34.8	65.7	33.1	41.3	90
DiffusionDET	62.4	39.1	59.4	31.8	58.9	29.5	32.0	107
DiMNet	**71.3**	**42.6**	**66.8**	**37.3**	**69.1**	**35.1**	33.5	105

4.4 Visual Qualitative Analysis

Fig. 5 displays partial detection results of our model and other models on the DENTEX2023 validation set. Due to the patient's misaligned teeth and concentrated dental pathologies, this case presents high detection difficulty, which clearly reflects the performance differences among models. The ground truth labels indicate caries on teeth at quadrant 4 number 1, quadrant 3 number 1, and quadrant 3 number 2, with deep caries at quadrant 4 number 2. In DiMNet's results, the model generates four prediction boxes with confidence scores all above 50%. These boxes closely align with the ground truth boxes, and all category labels are correct. Faster R-CNN produces three prediction boxes with accurate locations and labels, but misses one detection. DETR yields results close to the ground truth, yet with lower confidence scores. DiffusionDet generates three prediction boxes, two of which have correct locations and labels, while the third shows incorrect pathology classification. Both YOLOv10 and RetinaNet detect only two pathological teeth, with prediction boxes deviating significantly from the ground truth and exhibiting lower confidence scores.

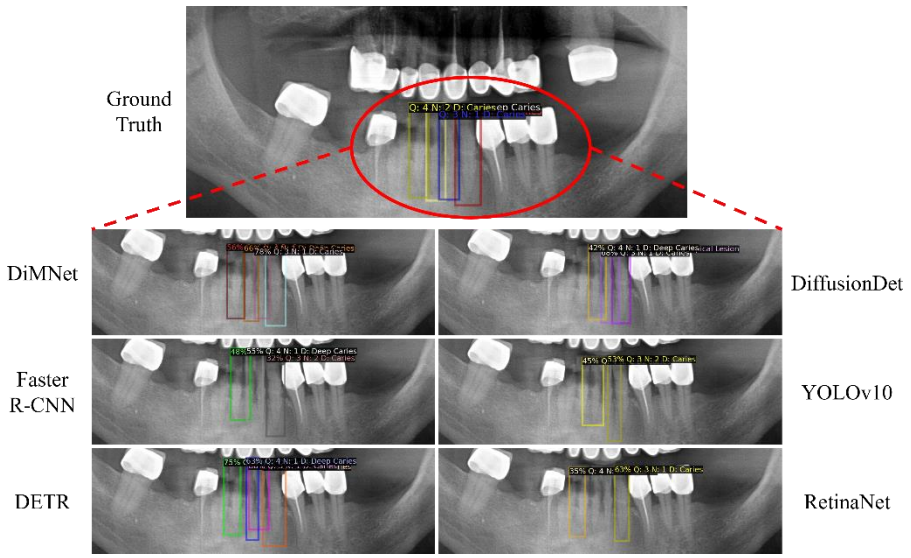

Fig. 5. Visual comparison of the results of each model on validation dataset

4.5 Ablation Experiment

To investigate the impact of different improvement strategies proposed in this chapter on model performance, we conduct ablation experiments on DENTEX2023 using two evaluation metrics: AR and AP. The experimental setup remains consistent with the comparative experiments, including six ablation groups: (1) baseline model, (2) baseline + weight transfer, (3) baseline + weight transfer + prediction box assistance, and three additional groups (4-6) with baseline + weight transfer + prediction box assistance + different encoders.

Table 2 Ablation experiment results

Groups	Model	Quadrant		Enumeration		Diagnosis	
		AR	AP	AR	AP	AR	AP
1	Baseline	55.2	32.7	48.2	25.4	48.5	21.7
2	1+Weight Transfer	60.7	40.3	52.1	32.4	56.8	30.3
3	2+Cross-Stage guidance	67.9	41.4	62.7	35.7	67.7	33.7
4	3+ResNet50	56.3	37.4	57.2	31.4	61.1	31.0
5	3+ResNet101	61.2	39.3	62.2	34.8	61.2	31.0
6	3+MambaVision	**71.7**	**42.6**	**66.8**	**37.3**	**69.1**	**35.1**

As shown in Table 2 Ablation experiment results, the baseline model (a modified DiffusionDet with a multi-label detection head) exhibits low performance across quadrant, tooth number, and pathology labels. In the second group (baseline + weight transfer), performance improvements are observed, confirming that pre-trained weights from earlier stages enhance subsequent detection capabilities. The third group (baseline + weight transfer + cross-stage guidance) further boosts detection accuracy through auxiliary bounding box prediction. The last three groups replace the encoder with ResNet50, ResNet101, and MambaVision respectively, with MambaVision achieving the best performance (improving AP by 1.2%, 1.6%, and 1.4% across labels).

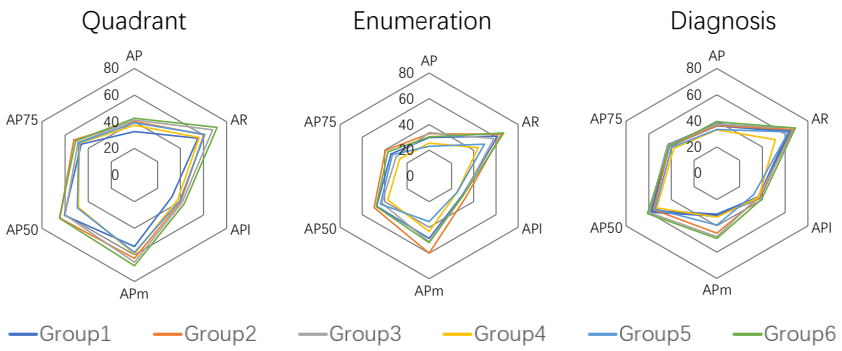

Fig. 6. Visualization of Ablation Study Results

As illustrated in Fig. 6., to provide a more intuitive comparison of evaluation metrics across ablation study groups, this section presents a radar chart based on the experimental data. In addition to the AR and AP metrics shown in Table 2, the chart includes APm (medium objects), APl (large objects), AP50 (IoU threshold of 50%), and AP75 (IoU threshold of 75%) metrics, enabling a more comprehensive and visual comparison. The results demonstrate that the transfer learning and cross-stage guidance strategies significantly enhance the model's performance in quadrant and tooth position detection, confirming that these strategies enable the model to effectively learn from low-

level annotated data. Improvements are also observed in dental disease detection, as the low-level features learned during early training stages provide foundational knowledge for subsequent training with complete, high-level annotated data. Furthermore, using MambaVision as the backbone network for the encoder leads to notable performance gains across all tasks, validating its superior feature extraction capabilities.

4.6 Training Strategy Validation

To validate the effectiveness of the multi-stage strategy, this section conducts experiments by selecting the same image from the test set for detection validation and visualizing the results. The figures clearly demonstrate that the model effectively focuses on the target task of the current stage during each training phase while exhibiting excellent detection performance.

Specifically, the first training phase uses subset (a) of the dataset, where only the positions and numbering of the four quadrants are annotated on the images. During training, the tooth position and caries detection heads are frozen, and optimization is performed using Eq. (8a). Fig. 7 shows the detection results after training, demonstrating that the model accurately identifies the four quadrants. While quadrant detection is relatively simple, it serves as the foundation for subsequent tooth numbering and disease detection, playing a crucial role in overall performance.

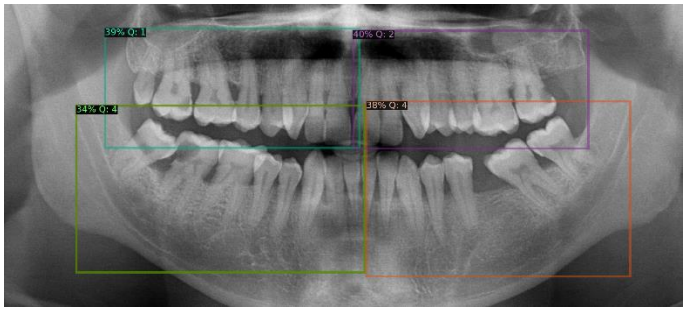

Fig. 7. Quadrant training phase

The second training phase employs subset (b) of the dataset, which includes tooth position annotations—bounding boxes for each tooth along with quadrant and tooth numbering. Accordingly, the caries detection head remains frozen, and optimization follows Eq. (8b). Notably, the initial model weights for this stage are not randomly initialized; instead, weight transfer is applied by loading the pre-trained weights from the first stage. Furthermore, the cross-stage guidance strategy first uses the model trained in the first stage to detect the current training data. High-confidence predicted boxes (confidence > 0.5) are selected and concatenated with the noisy boxes of the current stage to guide the denoising process.

Fig. 8 illustrates the model's performance in tooth numbering, showing accurate identification of each tooth along with its quadrant and numbering. The third stage can be interpreted as filtering teeth with diseases from all numbered teeth in the second

stage and diagnosing the disease type. Thus, the effectiveness of tooth numbering in the second stage is critical for the overall diagnostic system.

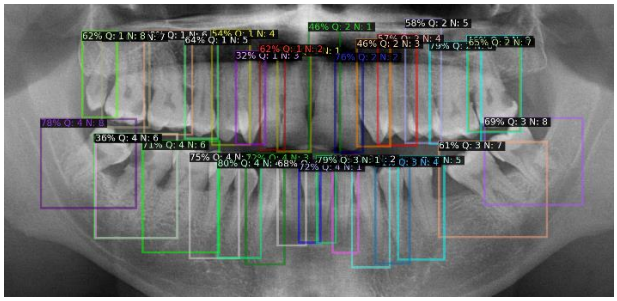

Fig. 8. Enumeration training phase

The third training phase uses subset (c) of the dataset, which contains complete annotations. Training is optimized via Eq. (8c), incorporating weight transfer and cross-stage guidance strategies. As shown in Fig. 9, the model in this stage can localize diseased teeth while simultaneously providing quadrant, numbering, and disease type information, achieving an end-to-end, multi-label detection framework.

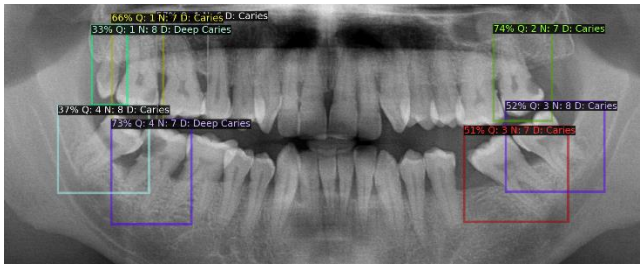

Fig. 9. Diagnosis training phase

The experimental results clearly demonstrate that the model successfully maintains focus on the stage-specific target tasks while delivering consistently excellent detection performance. These findings confirm that DiMNet's multi-head architecture significantly improves system scalability, enabling effective handling of tasks with different annotation granularities across training phases. Furthermore, our proposed multi-stage training approach not only enhances detection accuracy but also improves the model's adaptability to hierarchical tasks, thereby establishing a robust foundation for real-world clinical applications.

5 Conclusion

In this study, we proposed DiMNet, a multi-label object detection model based on an improved DiffusionDet, combined with the feature extraction capabilities of MambaVision. The model is designed to simultaneously predict the quadrant, enumeration, and diagnosis of each abnormal tooth in panoramic X-ray images. Taking into account

the structured and progressive characteristics of the dataset, this chapter adopts multi-stage training strategy, complemented by auxiliary training methods such as weight transfer and cross-stage guidance, to fully leverage the advantages of the dataset and achieve multi-label detection while improving the model's accuracy.

Through comparative experiments and result visualizations on the DENTEX2023 dataset, we demonstrate that the proposed model exhibits excellent detection performance and strong generalization capabilities. Additionally, ablation experiments confirm that the proposed improvement strategies play a significant role in enhancing the model's performance.

References

1. Tonetti, M.S., Bottenberg, P., Conrads, G., Eickholz, P., Heasman, P., Huysmans, M.-C., López, R., Madianos, P., Müller, F., Needleman, I., et al.: Dental caries and periodontal diseases in the ageing population: Call to action to protect and enhance oral health and well-being as an essential component of healthy ageing--Consensus report of group 4 of the joint EFP/ORCA workshop on the boundaries between caries and periodontal diseases. Journal of Clinical Periodontology 44, S135–S144 (2017)
2. Bruno, M.A., Walker, E.A., Abujudeh, H.H.: Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. Radiographics 35(6), 1668–1676 (2015)
3. Razzak, M.I., Naz, S., Zaib, A.: Deep learning for medical image processing: Overview, challenges and the future. Classification in BioApps: Automation of Decision Making, 323–350 (2017)
4. Pati, S., Thakur, S.P., Bhalerao, M., Thermos, S., Baid, U., Gotkowski, K., Gonzalez, C., Guley, O., Hamamci, I.E., Er, S., et al.: GaNDLF: A Generally Nuanced Deep Learning Framework for Scalable End-to-End Clinical Workflows in Medical Imaging. arXiv preprint arXiv:2103.01006 (2021)
5. Tuzoff, D.V., Tuzova, L.N., Bornstein, M.M., Krasnov, A.S., Kharchenko, M.A., Nikolenko, S.I., Sveshnikov, M.M., Bednenko, G.B.: Tooth detection and numbering in panoramic radiographs using convolutional neural networks. Dentomaxillofacial Radiology 48(4), 20180051 (2019)
6. Zhu, H., Cao, Z., Lian, L., Ye, G., Gao, H., Wu, J.: CariesNet: A deep learning approach for segmentation of multi-stage caries lesion from oral panoramic X-ray image. Neural Computing and Applications, 1–9 (2023)
7. Mei, S., Ma, C., Shen, F., Wu, H.: YOLOrtho--A Unified Framework for Teeth Enumeration and Dental Disease Detection. arXiv preprint arXiv:2308.05967 (2023).
8. Kumar, A., Bhadauria, H.S., Singh, A.: Descriptive analysis of dental X-ray images using various practical methods: A review.PeerJ Computer Science 7, e620 (2021)
9. Willemink, M.J., Koszek, W.A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L.R., Summers, R.M., Rubin, D.L., Lungren, M.P.: Preparing medical imaging data for machine learning. Radiology 295(1), 4–15 (2020)
10. Shin, S.-J., Kim, S., Kim, Y., Kim, S.: Hierarchical multi-label object detection framework for remote sensing images. Remote Sensing 12(17), 2734 (2020)
11. Chen, S., Sun, P., Song, Y., Luo, P.: DiffusionDet: Diffusion model for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 19830–19843 (2023)

12. Hatamizadeh, A., Kautz, J.: MambaVision: A hybrid mamba-transformer vision backbone. arXiv preprint arXiv:2407.08083 (2024)
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems 30 (2017)
14. Chen, H., Li, H., Zhao, Y., Zhao, J., Wang, Y.: Dental disease detection on periapical radiographs based on deep convolutional neural networks. International Journal of Computer Assisted Radiology and Surgery 16, 649–661 (2021)
15. Lee, S., Oh, S.-i., Jo, J., Kang, S., Shin, Y., Park, J.-w.: Deep learning for early dental caries detection in bitewing radiographs. Scientific Reports 11(1), 16807 (2021)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851 (2020).
17. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023).
18. Ju, Z., Zhou, W.: Vm-ddpm: Vision mamba diffusion for medical image synthesis. arXiv preprint arXiv:2405.05667 (2024).
19. Zhou, J., Xiao, C., Peng, B., Liu, T., Liu, Z., Liu, Y., Liu, L.: MaDiNet: Mamba Diffusion Network for SAR Target Detection. arXiv preprint arXiv:2411.07500 (2024).
20. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
21. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
22. Hamamci, Ibrahim Ethem, et al. "Dentex: An abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic x-rays." arXiv preprint arXiv:2305.19112 (2023).
23. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al.: Yolov10: Real-time end-to-end object detection. In: Advances in Neural Information Processing Systems, vol. 37, pp. 107984–108011 (2024).
24. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017).
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149 (2016).
26. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229 (2020).