scientific reports

Check for updates

OPEN

Lightweight YOLOv8 for tongue teeth marks and fissures detection based on C2f_DCNv3

Chunyang Jin, Delong Zhang, Xiyuan Cao✉, Zhidong Zhang, Chenyang Xue & Yanjun Zhang

This paper propose a significantly enhanced YOLOv8 model specifically designed for detecting tongue fissures and teeth marks in Traditional Chinese Medicine (TCM) diagnostic images. By integrating the C2f_DCNv3 module, which incorporates Deformable Convolutions (DCN), replace the original C2f module, enabling the model to exhibit exceptional adaptability to intricate and irregular features, such as fine fissures and teeth marks. Furthermore, the introduction of the Squeeze-and-Excitation (SE) attention mechanism optimizes feature weighting, allowing the model to focus more accurately on key regions of the image, even in the presence of complex backgrounds. The proposed model demonstrates a significant performance improvement, achieving an average precision (mAP) of 92.77%, which marks a substantial enhancement over the original YOLOv8. Additionally, the model reduces computational cost by approximately one-third in terms of FLOPS, maintaining high accuracy while greatly decreasing the number of parameters, thus offering a more robust and resource-efficient solution. For tongue crack detection, the mAP increases to 91.34%, with notable improvements in F1 score, precision, and recall. Teeth mark detection also sees a significant boost, achieving an mAP of 94.21%. These advancements underscore the model's outstanding performance in TCM tongue image analysis, providing a more accurate, efficient, and reliable tool for clinical diagnostic applications.

Tongue diagnosis is one of the most important diagnostic methods in Traditional Chinese Medicine (TCM) and serves as a non-invasive and convenient assessment of human health[1]. TCM posits that the tongue is closely linked to an individual's health. When there are abnormalities or diseases in the body, the characteristics of the tongue undergo significant changes. Doctors evaluate the severity of the condition by analyzing changes in the patient's tongue features and making corresponding diagnoses[2]. The tip of the tongue corresponds to the lungs, shoulders, heart, and brain, while the body of the tongue reflects the spleen and stomach; the sides relate to the liver and gallbladder, and the back corresponds to the kidneys and intestines[3]. An essential step in TCM diagnosis is examining the tongue and tongue coating. Different tongue characteristics correspond to varying states of the body. For instance, a red tongue with pronounced teeth marks may indicate excessive dampness within the body. If the tongue surface exhibits fissures of varying depth and shape, it is termed a "fissure tongue," often a symptom of yin deficiency and excessive heat.

Tongue fissures and teeth marks are two common tongue features in TCM diagnosis[4], holding significant clinical diagnostic value. Tongue fissures typically refer to longitudinal or transverse fissures on the tongue surface, often related to insufficient qi and blood or dysfunction of internal organs. Based on the shape and distribution of these fissures, TCM practitioners can infer the patient's organ conditions and overall health. Teeth marks appear as indentations along the tongue's edges, usually associated with dysfunction of the spleen and stomach or qi deficiency[5]. By analyzing tongue fissures and teeth marks, TCM practitioners can make more accurate assessments of diseases and formulate tailored treatment plans, thus enhancing the comprehensiveness and personalization of diagnoses[6].

From the above, it is evident that effectively detecting the status of tongue teeth marks and fissures can help anticipate a patient's constitution, thereby improving the prediction of potential diseases. Utilizing computer vision and neural networks to automatically extract features related to tongue teeth marks and fissures can assist doctors in accurately assessing patients' physiological conditions. Providing precise feature predictions can furnish doctors with comprehensive reference information. Additionally, such tools can be integrated into applications or mini-programs to assist patients in self-assessment, thus allowing for proactive avoidance of health risks.

Tang et al.[7] proposed using a cascaded Convolutional Neural Network (CNN) for tongue image segmentation and landmark prediction, combining Deformable Convolutions (DCN) for fine classification. Although this

Key Laboratory of Instrumentation Science, Dynamic Measurement of Ministry of Education, North University of China, Taiyuan 030051, Shanxi, China. ✉email: caoxiyuan@nuc.edu.cn

improved detection accuracy, it increased model complexity and made hyperparameter tuning sensitive. Lo Lun-chien et al.[8] introduced a system using HIS color space transformation, polar coordinate transformation, and active contour models for tongue image segmentation and analysis of tongue surface features. This system can accurately detect features such as fissures and teeth marks but is limited by lighting conditions and dataset size. Wu et al.[9] proposed a multimodal intelligent diagnostic model (MMTV) based on EEG and tongue images for assisting depression diagnosis, incorporating a dual-stream input mechanism and self-attention into EEGNet, as well as a multi-step pre-training approach for the ViT model, all of which enhance feature extraction capabilities. However, this requires substantial computational resources, which may limit its application in resource-constrained environments. Ni et al.[10] presented an improved capsule network model (TongueCaps) for tongue color classification in TCM, though capsule networks are generally considered less interpretable than CNNs, potentially affecting their transparency and reliability in clinical applications. Wang et al.[11] achieved over 90% accuracy in identifying teeth-marked tongues using ResNet34, although the study noted a relatively small sample size and potential class imbalance, which could affect model performance. Li et al.[12] employed multi-instance learning and CNN feature recognition for teeth-marked tongues, with experimental results indicating reasonable accuracy, but the specifics of dataset size, diversity, and experimental settings warrant further consideration.

In recent years, with the development of object detection algorithms, the YOLO series has gained popularity in various fields due to its lightweight, fast, and efficient attributes. The introduction of deep learning has significantly enhanced object detection performance, particularly through the excellent image feature extraction capabilities of Convolutional Neural Networks (CNNs). Deep learning-based object detection methods can be categorized into two main types: two-stage methods and single-stage methods. Two-stage methods, represented by the R-CNN[13] series-including R-CNN, Fast R-CNN[14], and Mask R-CNN[15]-rely on stepwise processing for candidate region generation and classification, offering high accuracy but with greater computational overhead. In contrast, single-stage methods, represented by YOLO and SSD[16], such as the YOLOv1[17] to YOLOv8[18] series, directly output bounding boxes and classes through a single network prediction, balancing speed and accuracy, making them suitable for real-time detection scenarios. Among single-stage methods, the YOLO (You Only Look Once) series stands out for its simplified structure and efficient detection speed. YOLOv1 introduced an innovative framework that simplifies the object detection task into a single regression problem; however, while it is fast, its accuracy is somewhat lacking. With successive iterations, YOLOv2, YOLOv3, and YOLOv4 gradually improved accuracy and robustness while maintaining high-speed detection advantages. The latest YOLOv5[19], YOLOv7, and YOLOv8 further optimized the model architecture, achieving a better balance between speed and accuracy, making them applicable to a wider range of scenarios. Another single-stage method, SSD (Single Shot Multibox Detector), predicts through multi-scale feature mapping layers, balancing speed and accuracy, particularly suited for real-time applications. Additionally, RetinaNet introduces Focal Loss to address the imbalance between positive and negative samples, effectively enhancing detection performance for small objects, making it suitable for large-scale datasets.

In summary, for tongue feature detection, deploying a lightweight and fast model on hardware is essential. Although two-stage object detection algorithms offer high accuracy, they require more time and resources. Single-stage object detection algorithms still have room for improvement in accuracy. This paper utilizes an improved YOLOv8 to recognize tongue teeth marks and fissures, ensuring enhanced accuracy while maintaining high speed to meet the demands of the application scenario.

The main contributions of this paper are as follows:

1. Development of an Improved Tongue Diagnosis Image Detection Model Based on YOLOv8: Building upon the traditional YOLOv8 model, the paper propose a novel improved model specifically designed for tongue diagnosis image detection. By optimizing the model architecture, particularly in handling the subtle features of the tongue and complex backgrounds, enhance the detection capability of the model.

2. C2f_DCNv3 and SE Attention Mechanism:This study innovatively integrates Deformable Convolution (DCN)[20] and the Squeeze-and-Excitation (SE)[21] Attention Mechanism into the YOLOv8 model. Deformable Convolution allows for adaptive adjustment of the convolution kernel's position, enabling more precise capture of irregular features such as tongue fissures and teeth marks. The C2f_DCNv3 module, through weighted fusion within the module, not only retains the feature-capturing capability of DCNv3 but also significantly reduces the model's parameter count. The SE Attention Mechanism[15] dynamically adjusts the weights of feature channels, enhancing the model's focus on critical features. The combination of C2f_DCNv3 and SE leads to a substantial improvement in detection accuracy and efficiency.

3. Dataset and Model Performance: The dataset used for this model was annotated by professional practitioners at the Shanxi University of Traditional Chinese Medicine Hospital. Experimental results show that the improved model reduces the FLOPS from the original YOLOv8's 28.8G to 21G, achieving a one-third reduction in computational load while maintaining high performance and significantly enhancing computational efficiency. Compared to the original YOLOv8 model, our proposed improved model demonstrates a significant performance increase in detecting tongue fissures and teeth marks, achieving a total mean Average Precision (mAP) of 92.77%. Specifically, the mAP for tongue fissures reached 91.34%, a 5.24 percentage point improvement over the original model, with enhancements in F1 score, precision, and recall. The mAP for tongue teeth marks increased to 94.21%. These experiments validate the effectiveness of the improved model in complex tongue feature detection.

YOLOv8 algorithm

The YOLOv8 architecture continues the design philosophy of YOLOv5 and incorporates optimizations and improvements in various aspects. Despite its impressive performance in many applications, YOLOv8 still faces limitations when dealing with complex backgrounds and irregular features. Traditional convolution

operations extract features within a fixed receptive field, which may lack the adaptability needed for complex or deformed targets like tongue fissures and teeth marks. Consequently, YOLOv8 may experience feature loss or false detections in such cases, leading to decreased accuracy. Furthermore, when processing highly textured or complex background medical images, background interference can adversely affect precise target recognition.

Therefore, improving the YOLOv8 model to enhance its detection capabilities for irregular and complex targets has become a crucial research direction for enhancing its performance in medical imaging applications. This paper further optimizes the multi-scale feature fusion in the Neck module, enabling the model to better handle targets of varying sizes, particularly small pathological features in complex backgrounds. Through these improvements, YOLOv8 significantly enhances its ability to identify fine-grained targets while maintaining efficient detection speeds.

Improved YOLOv8 network

YOLOv8 iimprovements

In the improved YOLOv8 model, as shown in the Fig. 1, the original C2f module has been replaced with the lightweight C2f_DCNv3 module. Compared to the original C2f module, C2f_DCNv3 significantly reduces both the number of parameters and computational costs while expanding the receptive field for spatial information, further enhancing the model's detection capabilities. The C2f_DCNv3 module incorporates Deformable Convolution (DCN), which strengthens the model's ability to capture complex and irregular fine features.

Additionally, the Squeeze-and-Excitation (SE) attention mechanism has been integrated between the Backbone and Neck frameworks. This not only retains the Backbone's effective feature extraction but also enhances the model's focus on critical features within the original image, improving feature weighting capabilities. The SE mechanism adaptively adjusts the feature weights across different channels, enabling the model to maintain high-precision detection of targets even in complex backgrounds.

This improvement, combining the spatial deformation capability of DCNv3 with the channel attention of the SE mechanism, significantly enhances the accuracy and robustness of the YOLOv8 model in tongue image detection tasks. It provides stronger support for detecting complex morphological features such as tongue indentations and fissures, which are crucial in Traditional Chinese Medicine (TCM) tongue diagnosis.

By optimizing the network architecture, the improved YOLOv8 model ensures high precision in specific detection tasks. Particularly in the complex scenario of tongue diagnosis, YOLOv8's efficient feature extraction and target localization capabilities enable accurate identification of fine features on the tongue's surface. The enhanced model not only offers rapid and precise detection capabilities but also meets the real-time detection requirements in TCM diagnostics, providing robust support for intelligent tongue diagnosis.

Deformable convolution DCNv3

Traditional convolution uses fixed convolutional kernels, as shown in Fig. 2a which limits the ability to model geometric transformations, such as object deformations. Since such deformations are common in tasks like object detection and medical image segmentation, introduce deformable convolution, as illustrated in figures

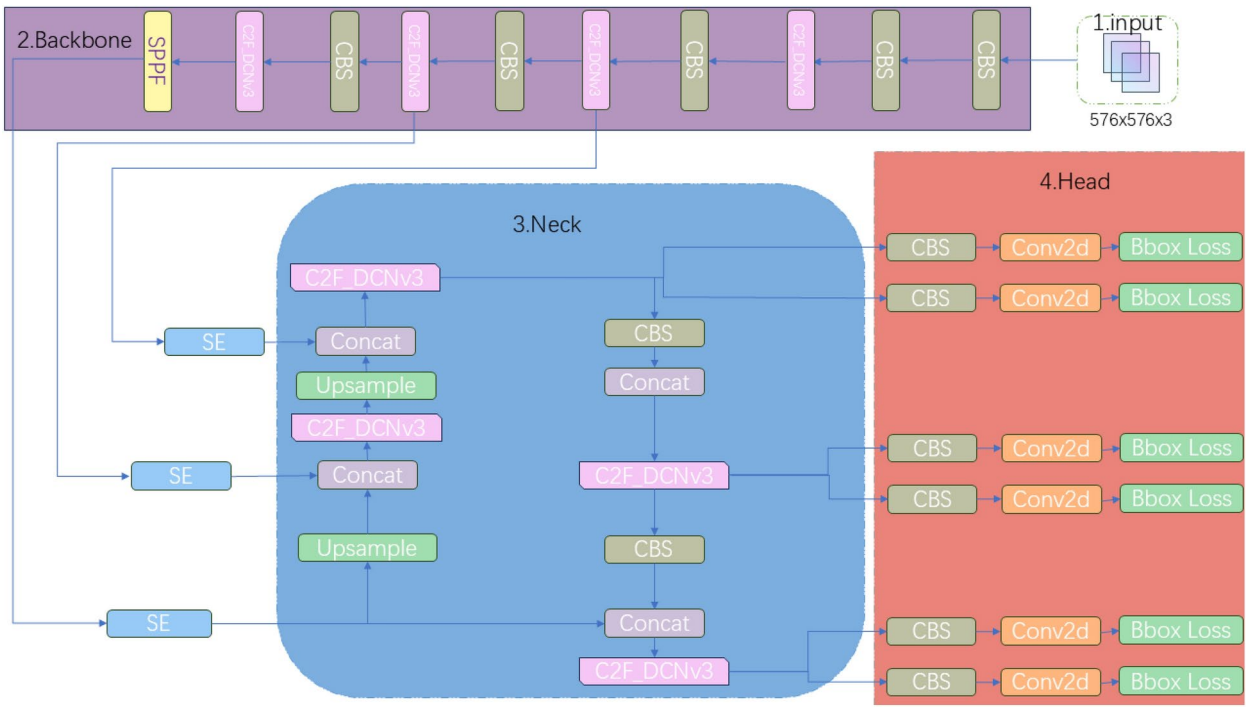

Figure 1. yolov8-our.

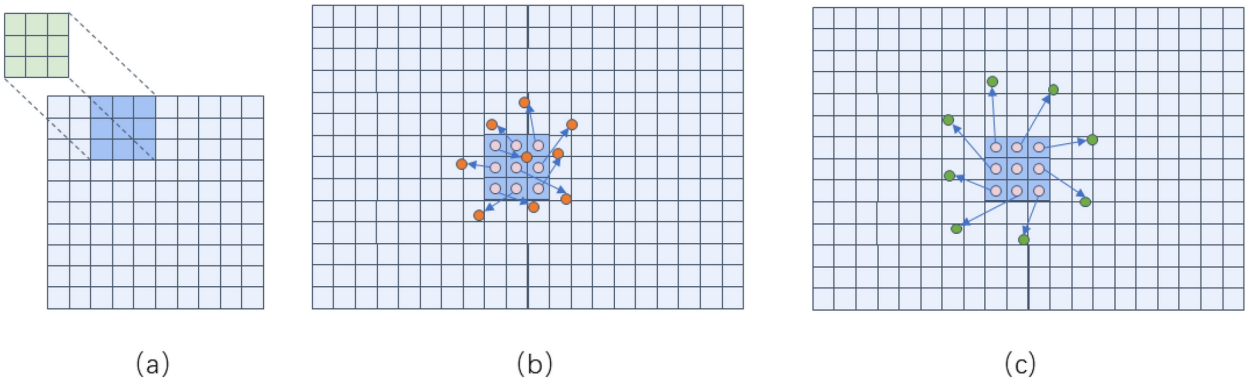

Figure 2. Dcnv3.

(b) and (c). This enhancement aims to improve the accuracy of detecting features like tongue indentations and fissures. By allowing the convolutional kernel to adaptively adjust its shape and size, DCNv3 effectively captures irregular patterns and complex structures, leading to better performance in identifying subtle features in tongue images.

Standard convolution samples input features at fixed positions, as shown in Eq. (1). In contrast, Deformable Convolution Networks (DCN) introduce learnable offsets, as described in Eq. (2), allowing the network to dynamically adapt to the spatial structures of objects. These offsets,as described in Eqs. (3) and (4), are predicted from the input feature map and applied during the convolution operation, enabling the model to focus on regions of interest that may not align with the grid. By incorporating spatial flexibility, DCNv3 can learn more complex geometric variations and object shapes, making it particularly effective for tasks involving irregular or deformable objects, such as indentations and fissures. Compared to DCNv1[22] and DCNv2[23], DCNv3 has been further optimized for accurate prediction and computational efficiency, enhancing accuracy. Thus, incorporating DCNv3 into the YOLOv8 model for detecting tongue features like indentations and fissures can significantly improve the model's performance. Given that these features often exhibit substantial morphological variability, using DCNv3 allows for more flexible capture of these changes, enhancing the model's generalization ability and diagnostic accuracy.

Regular conv

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \tag{1}$$

Deformable conv

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \bigtriangleup p_n) \tag{2}$$

$$x(p) = \sum_{q} G(q, p) \cdot x(q) \tag{3}$$

$$p = p_0 + p_n + \bigtriangleup p_n \tag{4}$$

C2f_DCNv3 model

As shown in Fig. 3, the C2f_DCNv3 module addresses the challenges associated with increased computational complexity and parameter count as the network depth increases, which can limit the model's applicability on resource-constrained devices. This module optimizes the network structure by introducing the Bottleneck_DCNv3.The input image first passes through a convolutional layer with a kernel size of 1x1. This layer not only alters the dimensionality but also facilitates cross-channel information exchange, simplifying the model. The output from this convolutional layer is then split into two equal parts, each with dimensions of Eq. (5).

$$h \times w \times \frac{C_{out}}{2} \tag{5}$$

Each partition is processed through a Bottleneck_DCNv3 structure, which further reduces the parameter count and computational complexity while maintaining the model's expressive power. The outputs of the two Bottleneck_DCNv3 structures are merged along the channel dimension, resulting in a combined feature map with dimensions of Eq. (6),where n is the expansion coefficient of the Bottleneck layer. Finally, the merged feature map is processed by another convolutional layer, which also employs a 1x1 kernel with a stride of 1, ultimately yielding an output with CoutC_outCout channels.

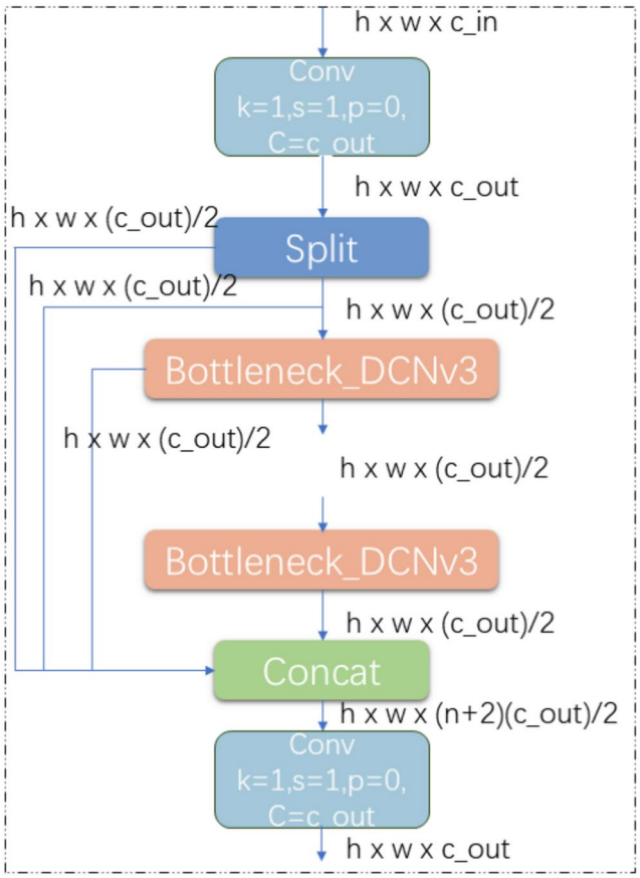

Figure 3. C2f_Dcnv3.

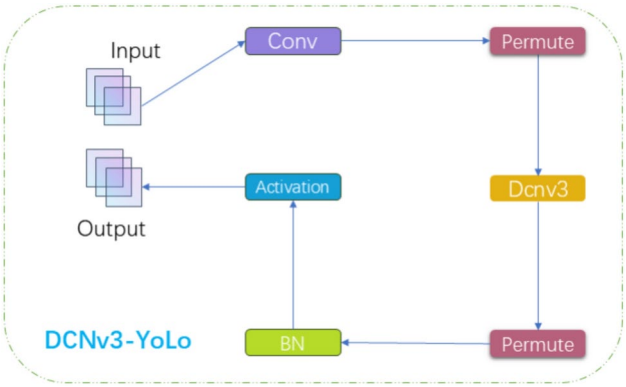

Figure 4. Dcnv3_YoLo.

$$h \times w \times (n+2)\frac{C_{out}}{2} \tag{6}$$

As shown in Fig. 4, the left side depicts the DCNv3-YOLO module. The input features are first reshaped using a Permute operation before being fed into the DCNv3 module, allowing it to learn from a broader receptive field. The features are then adjusted back to their original shape using Permute, followed by batch normalization (BN) and an activation function to produce the output.

As shown in Fig. 5, the Bottleneck_DCNv3 structure is the core of this model, utilizing depthwise separable convolutions to reduce both the parameter count and computational complexity. The input feature map first passes through a depthwise separable convolution layer (with K = 3, s = 1, p = 1), which halves the number of output channels to h×w×c/2 . Subsequently, batch normalization is applied to reduce internal covariate shift, enhancing the model's generalization capability. The output of the batch normalization is then passed through

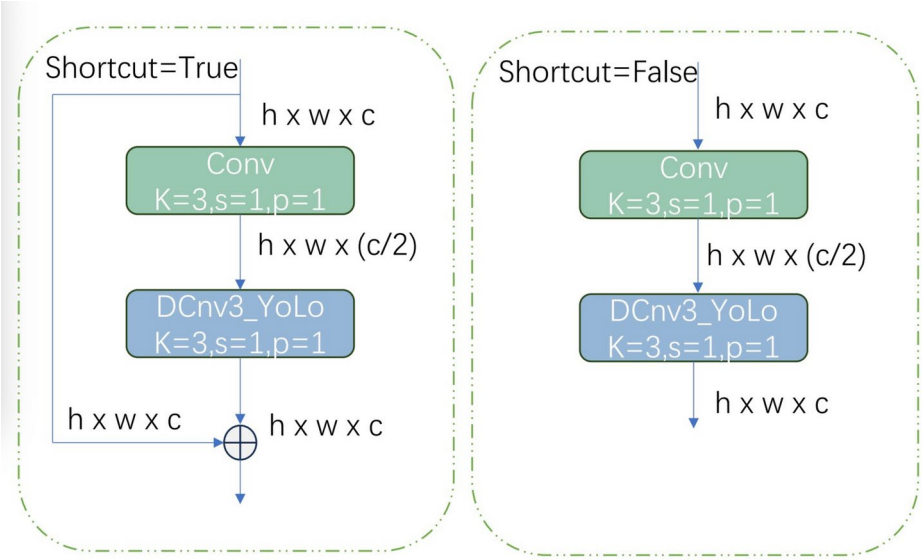

Figure 5. Bootleneck-DCNv3.

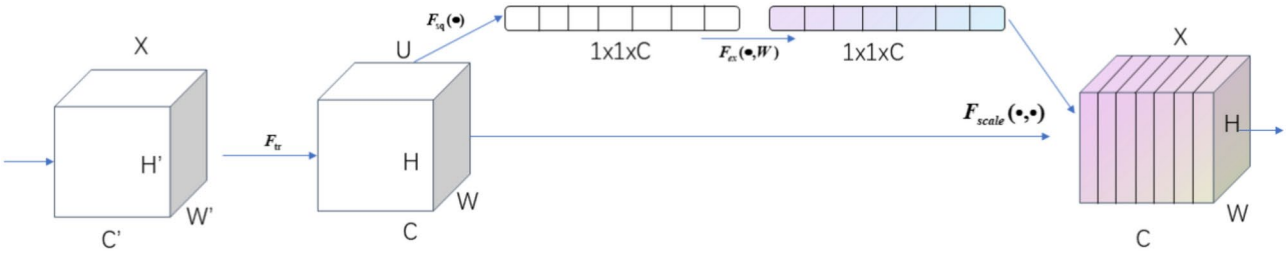

Figure 6. SE_attention.

an activation function (ReLU) to increase the model's non-linear expressiveness.The second component is the deformable convolution, which enhances the model's ability to compute the receptive fields for irregular features. Finally, a check on whether the Shortcut is set to true determines if a residual connection is applied, resulting in a final output feature map with dimensions of h×w×c.

The C2f_DCNv3 module significantly reduces the number of model parameters and computational load through the use of depthwise separable convolutions and deformable convolutions, thus enhancing the operational efficiency of the model. Following this, batch normalization and activation functions help maintain the model's expressive capability. Compared to the original C2f module, this structure effectively compresses and enhances the input features by integrating depthwise separable and deformable convolutions.

SE module

As shown in Fig. 6, the SE module effectively enhances the representation power of the network by allowing it to focus on important features, thereby improving overall model performance.

Squeeze Operation : The squeeze operation involves applying global average pooling (denoted as Eq. (7)) to the output feature map U, generating a feature map fg of size 1×1×C1. This process effectively compresses the spatial dimensions, resulting in a representation that summarizes the global information of each channel, allowing the model to capture the overall significance of features across the input.

$$f_s = F_{sq}(U) \tag{7}$$

Excitation Operation: As shown in Eq. (8), in the excitation operation, the feature map fg is fed into a fully connected layer (FC) followed by a ReLU[24] activation function. This is then passed through another fully connected layer with a sigmoid[25] activation function, resulting in a scaling vector fs. This vector captures the importance of each channel by dynamically adjusting their weights, effectively enhancing the model's focus on the most relevant features for improved performance.

$$f_s = \sigma(FC(ReLU(FC(f_g)))) \tag{8}$$

Scale Operation: As shown in Eq. (9), the scale operation involves element-wise multiplication of the original feature map U with the scaling vector fs. This operation, denoted as Fscale, produces the final output feature map X. By applying this scaling, the model emphasizes important channels while suppressing less relevant ones, thereby enhancing the overall feature representation and improving detection accuracy.

$$X = F_{scale}(U, f_s) \tag{9}$$

The structure of the SE attention mechanism is in the Fig. 7. After passing through the C2f module, the feature maps enter a global pooling layer for compression. Subsequently, the outputs are processed through ReLU and Sigmoid functions to obtain the scaling factors. This process effectively adjusts the weights for each channel, enhancing the model's focus on important channels and improving overall feature representation.

Results

Datasets and environments

Image quality[26] is critical for effective training,use dedicated imaging equipment to collect pictures.As shown in the Fig. 9, the dataset used in this study was provided by the Affiliated Hospital of Shanxi University of Traditional Chinese Medicine, with tongue diagnosis features annotated by experienced physicians from the hospital.

As shown in the Fig. 8, all images were divided into training, validation, and test sets. Since each image may contain multiple annotations for teeth marks and fissures, the actual number of annotations is higher than indicated in the dataset description. During training, all images were automatically resized to 576x576. The optimizer used in this study was the SGD[27]optimizer, which is more suitable for detecting fine features such as teeth marks and fissures compared to the Adam[28] optimizer. The batch size was set to 12, and an adaptive learning rate was applied to train the model. The experimental environment included an Ubuntu 20.04 system, three NVIDIA GeForce RTX 4090 GPUs, and the PyTorch framework.

Experimental process

Figures 10 and 11 illustrates the training and validation processes for different models. All models show a gradual convergence as the number of epochs increases, with the loss values rapidly decreasing. This indicates that the models effectively learned during the training process.

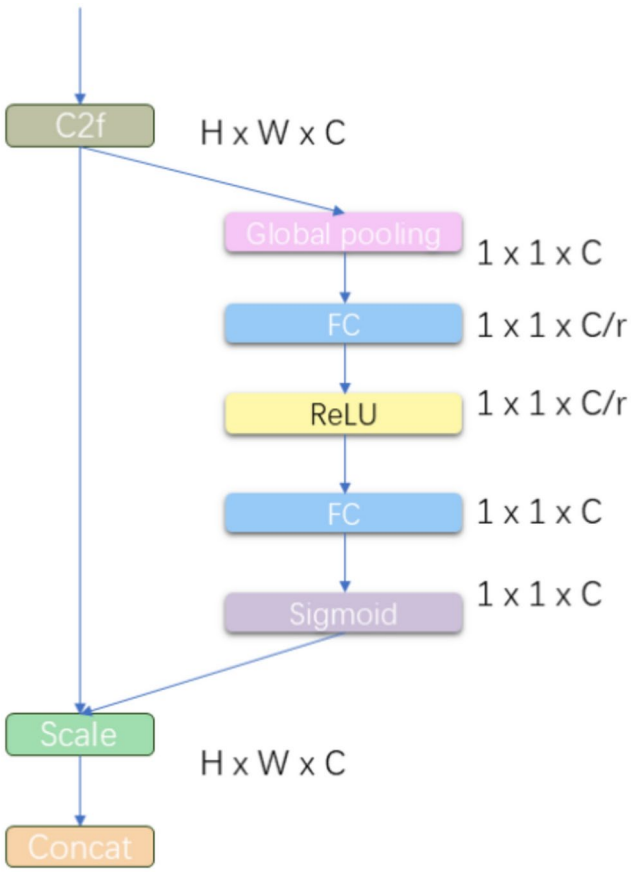

Figure 7. SE_attention progress.

Train val and test	Class	number
Train	Toothmark	1124
	Fissure	1239
Val	Toothmark	1161
	Fissure	1024
test	Toothmark	117
	fissure	129

Figure 8. Dataset.

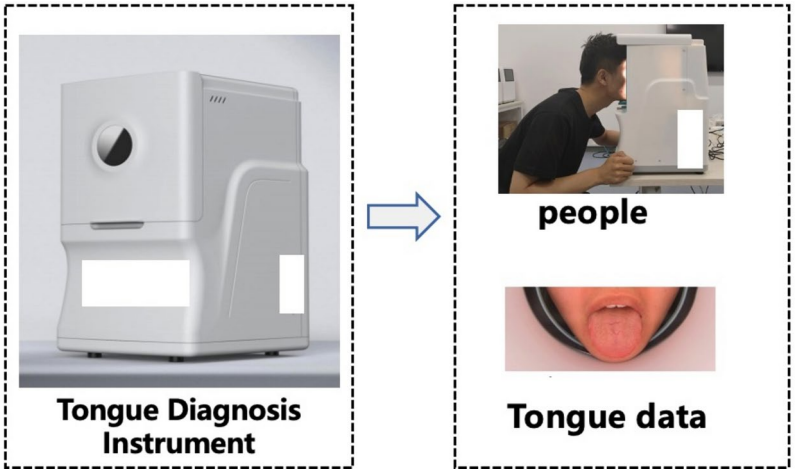

Figure 9. Collect tongue data.

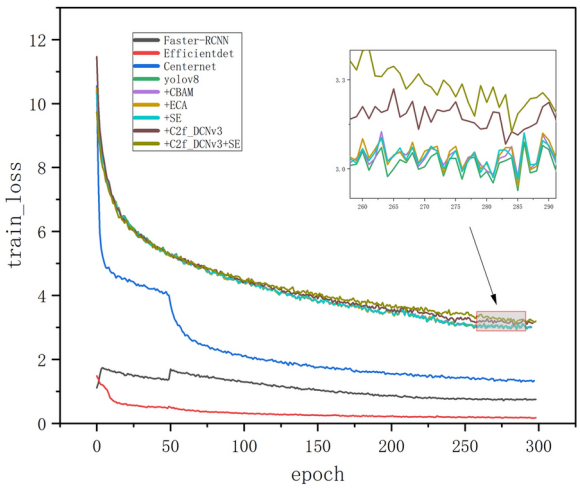

Figure 10. Train_loss.

During the training process, the mean Average Precision (mAP) is also a crucial metric. mAP is widely used in object detection tasks to assess performance and measures the average precision of the model in identifying objects across different categories. Specifically, mAP@0.5 refers to the mAP calculated when the IoU (Intersection over Union) threshold is set to 0.5.

Figure 12 demonstrates the mAP changes throughout the training process. As seen in the figure, with an increasing number of training epochs, the YOLOv8+C2f_DCNv3+SE (our model) exhibits excellent performance. In comparison, models incorporating other attention mechanisms such as CBAM, ECA, or those using only SE or C2f_DCNv3 alone, show lower mAP@0.5 values under the same training conditions. Moreover,

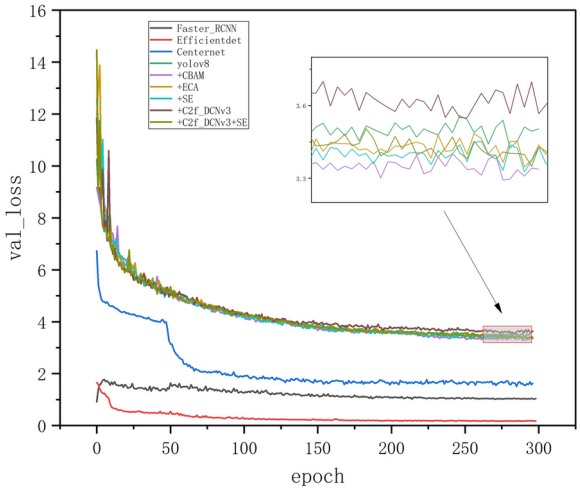

Figure 11. Val_loss.

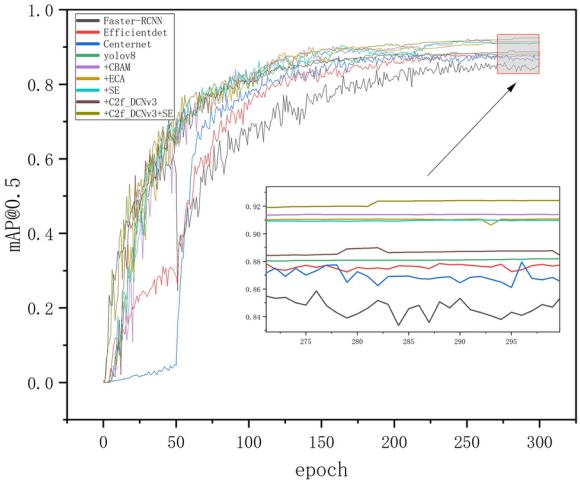

Figure 12. mAP@0.5.

when compared to non-YOLO models like the Faster R-CNN model, our model demonstrates a significant advantage.

The final mAP@0.5 value for the C2f_DCNv3+SE model is the highest, reaching 92.4%. This further validates the effectiveness of the C2f_DCNv3 and SE modules in enhancing object detection accuracy and efficiency.

Ablation study

In this study, the original YOLOv8 model was used as the baseline for experiments, and various attention mechanisms such as CBAM[29] and ECA[30] were incorporated for comparison. Additionally, conducted experiments with C2f_DCNv3 and SE separately to demonstrate the superiority of our proposed model. Precision was used to evaluate the accuracy of the model's predictions for tongue teeth marks and tongue fissures, reflecting the proportion of true cases among the correctly predicted cases. Recall measures whether all instances of tongue teeth marks and fissures were detected by the model. The F1 score represents the harmonic mean of Precision and Recall. Meanwhile, mAP (mean Average Precision) indicates the overall performance, with higher values suggesting better model effectiveness. The following two tables show the prediction results of tongue teeth marks and fissures under different models in the ablation experiments. As shown in Table 1 ,it is evident that after incorporating the C2f_DCNv3 and SE attention mechanisms, the model's performance significantly surpassed that of other models. Specifically, in the prediction of fissures, the mAP reached 91.34%, which was much higher than other models, and the F1 score was 0.87, positioning it as the best among similar models. Although the Recall score was 84.94%, slightly lower than that of Faster-R-CNN, the overall performance of our model remained at a clear advantage.

Table 2 shows the prediction performance for Toothmark. The mAP and F1 score reached the highest values, demonstrating the superior generalization ability of the model compared to other models. This highlights the overall performance advantage of our approach in accurately detecting toothmarks.

Network	mAP	F1	Precision	Recall
Fater-RNN	86.00%	0.75	65.71%	88.46%
Efficientdet	87.56%	0.85	88.43%	82.31%
Centernet	91.26%	0.83	92.12%	73.85%
Yolov8	86.10%	0.85	89.66%	80.00%
Yolov8+CBAM	87.20%	0.86	88.11%	81.54%
Yolov8+ECA	87.30%	0.85	87.70%	82.31%
Yolov8+SE	89.23%	0.86	88.62%	83.85%
Yolov8+C2f_DCNv3	87.79%	0.83	86.36%	80.69%
Yolov8+C2f_DCNv3+SE	**91.34%**	0.87	88.35%	84.94%

Table 1. Fissures.

Network	mAP	F1	Precision	Recall
Fater-RNN	81.52%	0.69	55.94%	88.98%
Efficientdet	86.41%	0.78	84.55%	73.23%
Centernet	90.78%	0.79	92.63%	69.29%
Yolov8	93.62%	0.88	94.59%	82.68%
Yolov8+CBAM	93.83%	0.86	92.49%	83.46%
Yolov8+ECA	93.50%	0.88	92.98%	83.46%
Yolov8+SE	92.81%	0.85	92.59%	78.74%
Yolov8+C2f_DCNv3	93.78%	0.87	93.64%	81.00%
Yolov8+C2f_DCNv3+SE	**94.21%**	0.88	93.33%	84.00%

Table 2. Toothmark.

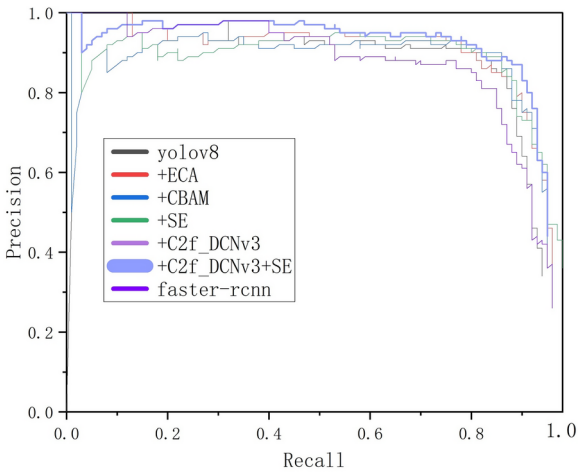

Figure 13. Tongue fissures-PR.

The Precision-Recall (PR) curve represents the relationship between Precision and Recall. The closer the PR curve is to the top right corner, the better the model's training performance. As shown in Figs. 13 and 14, with the incorporation of the C2f_DCNv3 and SE modules, the model demonstrates significant advantages in predicting toothmarks and tongue fissures. This indicates that our model excels at detecting such small and subtle features. It is clearly visible from the figure that the PR curve of the C2f_DCNv3 and SE model combination is the closest to the top right corner, further validating the effectiveness of the model.Since Faster-RCNN[31], EfficientDet[32], and CenterNet[33] are non-YOLO models, the PR curve only presents Faster-RCNN to represent the comparison with other models.

Table 3 presents the mAP parameters for predictions made on the test set. Unlike the training and validation sets, the test set better reflects the model's superiority. From the table, it is evident that our model not only achieves a significantly higher mAP compared to other models but also has a noticeably lower GFLOPS. This demonstrates that our model can significantly reduce computational complexity while maintaining high accuracy, achieving efficient performance.

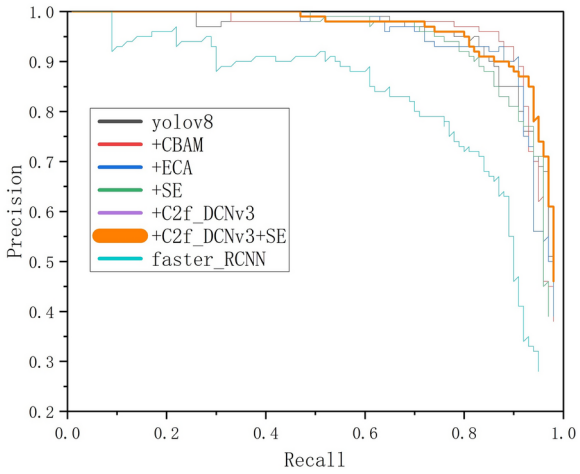

Figure 14. PR-toothmark.

Network	mAP	GFLOPS
Fater-RNN	83.76%	970.21G
Efficientdet	86.98%	26.17G
Centernet	91.20%	70.22G
Yolov8	90.89%	28.81G
+CBAM	90.74%	28.82G
+ECA	91.22%	28.82G
+SE	91.02%	28.82G
+C2f_DCNv3	90.78%	**21.01*G***
+C2f_DCNv3+SE	**92.77%**	**21.01*G***

Table 3. mAP@0.5 and GFLOPS.

To enhance the computational efficiency of our model, applied several lightweight techniques to reduce the FLOPS (Floating Point Operations Per Second). Through optimizations such as fusing convolution and batch normalization layers and introducing the C2f_DCNV3 module with reduced channel widths, the model's complexity was effectively reduced. As a result, the total computational cost was reduced from 28.8 GFLOPS to 22.91 GFLOPS, representing a significant 20.5% reduction in FLOPS. This reduction was achieved while maintaining model accuracy, demonstrating the efficiency of the proposed optimizations.

Figure 15 shows the prediction results, where (a) represents a tongue with teeth marks, (c) represents a tongue with fissures, and (b) represents a tongue with both teeth marks and fissures. Panels (d), (e), and (f) correspond to the prediction results for the aforementioned images. It can be observed that the model is able to accurately predict these fine pathological features, such as teeth marks and fissures, with high precision, which meets the clinical needs of Traditional Chinese Medicine (TCM).

Conclusion and limitations

In this study, proposed an improved YOLOv8 model for the efficient detection of tongue fissures and teeth marks, utilizing the C2f_DCNv3 module. By introducing deformable convolution and the SE attention mechanism, the model demonstrated superior adaptability and accuracy in handling complex backgrounds and irregular features. Experimental results showed that the improved model achieved a mean Average Precision (mAP) of 92.77% in detecting tongue fissures and teeth marks, significantly enhancing detection performance while optimizing computational efficiency. This achievement not only provides more reliable technical support for Traditional Chinese Medicine (TCM) tongue diagnosis but also opens new application prospects for medical image analysis in related fields.At the present time, the model cannot be applied to video recognition.In the future, the model can be further adjusted to accommodate a wider range of scenarios.

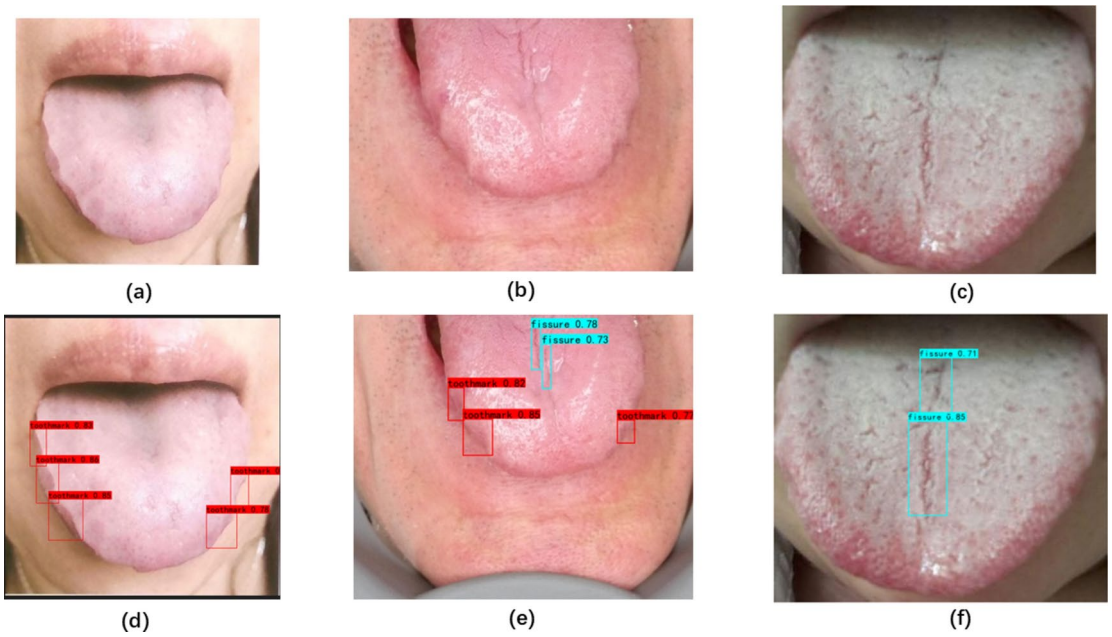

Figure 15. Tongue detection result.

Data availability

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to this data being supplied by Key Laboratory of Instrumentation Science& Dynamic Measurement (North University of China)and Affiliated Hospital of Shanxi University of Traditional Chinese Medicine, Ministry of Education and so cannot be made freely available.

Received: 8 October 2024; Accepted: 7 January 2025
Published online: 10 January 2025

References

1. Shao, Q., Li, X. & Fu, Z. Recognition of teeth-marked tongue based on gradient of concave region. In *2014 International Conference on Audio, Language and Image Processing*, 968–972 (IEEE, 2014).
2. Tania, M. H., Lwin, K. & Hossain, M. A. Advances in automated tongue diagnosis techniques. *Integr. Med. Res.* **8**, 42–56 (2019).
3. Kim, S.-R., Choi, W., Yeo, I. & Nam, D.-H. Comparative analysis of tongue indices between patients with and without a self-reported yin deficiency: A cross-sectional study. *Evid. Based Complement. Altern. Med.* **2017**, 1279052 (2017).
4. Kim, J. et al. Tongue diagnosis system for quantitative assessment of tongue coating in patients with functional dyspepsia: a clinical trial. *J. Ethnopharmacol.* **155**, 709–713 (2014).
5. Ma, J., Wen, G., Wang, C. & Jiang, L. Complexity perception classification method for tongue constitution recognition. *Artif. Intell. Med.* **96**, 123–133 (2019).
6. Cui, Y., Liao, S. & Wang, H. Roc-boosting: A feature selection method for health identification using tongue image. *Comput. Math. Methods Med.* **2015**, 362806 (2015).
7. Tang, W. et al. An automatic recognition of tooth-marked tongue based on tongue region detection and tongue landmark detection via deep learning. *Ieee Access* **8**, 153470–153478 (2020).
8. Lo, L.-C., Hou, M. C.-C., Chen, Y.-l., Chiang, J. Y. & Hsu, C.-C. Automatic tongue diagnosis system. In *2009 2nd international conference on biomedical engineering and informatics*, 1–5 (IEEE, 2009).
9. Wu, K. et al. Integrating transient and long-term physical states for depression intelligent diagnosis. In *BMVC*, 119–122 (2023).
10. Ni, J., Yan, Z. & Jiang, J. Tonguecaps: An improved capsule network model for multi-classification of tongue color. *Diagnostics* **12**, 653 (2022).
11. Wang, X. et al. Artificial intelligence in tongue diagnosis: Using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark. *Comput. Struct. Biotechnol. J.* **18**, 973–980 (2020).
12. Li, X., Zhang, Y., Cui, Q., Yi, X. & Zhang, Y. Tooth-marked tongue recognition using multiple instance learning and cnn features. *IEEE Trans. Cybern.* **49**, 380–387 (2018).
13. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587 (2014).
14. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149 (2016).
15. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969 (2017).
16. Liu, W. et al. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 21–37 (Springer, 2016).
17. Redmon, J. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016).
18. Varghese, R. & Sambath, M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 1–6 (IEEE, 2024).
19. Zhang, Y. et al. Real-time vehicle detection based on improved yolo v5. *Sustainability* **14**, 12274 (2022).

20. Wang, W. et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14408–14419 (2023).
21. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141 (2018).
22. Dai, J. et al. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773 (2017).
23. Zhu, X., Hu, H., Lin, S. & Dai, J. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9308–9316 (2019).
24. Glorot, X., Bordes, A. & Bengio, Y. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323 (JMLR Workshop and Conference Proceedings, 2011).
25. Elfwing, S., Uchibe, E. & Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Netw.* **107**, 3–11 (2018).
26. Muksimova, S., Umirzakova, S., Mardieva, S. & Cho, Y.-I. Enhancing medical image denoising with innovative teacher-student model-based approaches for precision diagnostics. *Sensors* **23**, 9502 (2023).
27. Ruder, S. An overview of gradient descent optimization algorithms. arXiv:1609.04747 (2016).
28. Diederik, P. K. Adam: A method for stochastic optimization. *(No Title)* (2014).
29. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19 (2018).
30. Wang, Q. et al. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11534–11542 (2020).
31. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149 (2016).
32. Tan, M., Pang, R. & Le, Q. V. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10781–10790 (2020).
33. Duan, K. et al. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6569–6578 (2019).

Acknowledgements

The authors gratefully acknowledge the medical and technical support of Jiuzhang Men and his team at the Shanxi University of Chinese Medicine's Affiliated Hospital.

Author contributions

C.J. participated in algorithm development, programming, manuscript writing, and image acquisition. D.Z. was involved in image collection and annotation. X.C. and Z.Z. provided revisions and suggestions for the manuscript and wrote the comments and editorial. X.C., C.X., and Y.Z. assisted with manuscript formatting review and editing. All the authors contributed extensively to the manuscript.

Funding

This work was supported by the Key Research and Development Program of Shanxi Province[No. 202102130501011]and State Key Laboratory of Dynamic Measurement Technology, School of Instrument and Electronics, North University of China, Taiyuan, 030051, China the National Natural Science Foundation of China [No. 62204231].

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.C.