Original Research

Science of Traditional Chinese Medicine

OPEN

Study on image detection and target recognition based on traditional Chinese medicine

Tianchi Mao[a], Xing Sun[a], Jiayin Zhu[a], An Liu[b], Yang Li[a,*], Jingang Ma[a,*], Cong Guo[b,*]

[a]College of Medical Information Engineering, Shandong University of Traditional Chinese Medicine, Ji'nan, China; [b]State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs, Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China

Abstract

Background: Chinese herbal pieces are an essential component of traditional Chinese medicine. Accurate identification and classification of these materials are crucial in clinical practice.

Objective: This study aims to enhance the recognition efficiency of Chinese herbal pieces using deep learning technology, while addressing the limitations of traditional manual classification methods in terms of both quality and efficiency.

Methods: A comprehensive dataset containing 201 types of Chinese herbal pieces was established. Based on Real-time Detection Transformer (RT-DETR), we designed and integrated a Feature-focused Diffusion Network (FDN), resulting in an improved model termed RT-DETR-FDN. The proposed FDN includes a Feature-focus Module and a feature diffusion mechanism, enabling the model to capture more extensive feature information from Chinese herbal pieces and diffuse it across multiple detection scales.

Results: Experimental results show that RT-DETR-FDN achieved a precision of 0.925, a recall of 0.943, and an mAP50–95 of 0.851. In addition, the model was compared with representative You Only Look Once series models commonly used in object detection. Compared with these models, RT-DETR-FDN achieved higher recognition accuracy while maintaining a lightweight architecture.

Conclusion: This study integrates deep learning with traditional Chinese medicine, providing a more effective solution for the recognition of Chinese herbal pieces.

Key words: Deep learning; Image detection and target recognition; Real-time Detection Transformer (RT-DETR); Traditional Chinese medicine

1. Introduction

Traditional Chinese medicine (TCM) is a collective term encompassing the medical practices of various ethnic groups in China, including both Han and minority traditions. With its long history, unique theoretical framework, and specialized techniques, TCM represents a vital component of China's cultural heritage and is the most well-preserved traditional medical system in the world today.[1] Before clinical application, Chinese medicinal materials must undergo processing, which may include cleaning, cutting, frying, boiling, steaming, and other methods.[2] The processed products are known as Chinese herbal pieces.[3] Processing alters the morphology of the raw materials, and many types of Chinese herbal pieces exhibit similar colors and textures. Therefore, accurate identification and classification of these pieces are particularly important.[4]

Traditional manual classification methods are subjective, cumbersome, time-consuming, and costly, making it difficult to ensure both the quality and efficiency of classification. With the continuous expansion of the TCM market and the growing clinical demand, there is an urgent need to establish accurate and efficient methods for identifying Chinese herbal pieces. In recent years, computer vision, a branch of artificial intelligence, has developed rapidly, showing wide applicability and excellent performance in object detection and image classification tasks. Therefore, integrating artificial intelligence into the field of TCM represents a promising direction for future research and development.[5]

Deep learning employs complex neural networks to simulate the processing power of the human brain, enabling learning and decision-making from large volumes of data and greatly enhancing the effectiveness of artificial intelligence applications across various fields. The development of deep learning technology for object detection has also advanced rapidly. Object detection refers to the computer's ability to accurately identify target objects within complex backgrounds. Early approaches relied on manual feature extraction and traditional machine learning methods such as sliding windows, Histogram of Oriented Gradients (HOG), and Scale-invariant Feature Transform (SIFT).[6,7] However, these methods were limited in recognition accuracy and speed. A major milestone in object detection was the introduction of Regions with Convolutional Neural Networks (R-CNN), which was the first to integrate CNNs into

Corresponding Author: Yang Li, College of Medical Information Engineering, Shandong University of Traditional Chinese Medicine, 4655 University Road, Changqing District, Ji'nan 250355, China. E-mail: liyang_1222@126.com (Y. Li). Jingang Ma, College of Medical Information Engineering, Shandong University of Traditional Chinese Medicine, 4655 University Road, Changqing District, Ji'nan 250355, China. E-mail: ma_jingang@126.com (J. Ma). Cong Guo, State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs, Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, No. 16, Dongzhimennei Nanxiaojie, Dongcheng District, Beijing 100070, China. E-mail: cguo@icmm.ac.cn (C. Guo).

Science of Traditional Chinese Medicine, (2026) 4, 1, 73–80

Supplemental digital content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's Web site (www.stcm.com).

Received January 25, 2025; Accepted August 14, 2025

http://dx.doi.org/10.1097/st9.0000000000000096

object detection. R-CNN first generates candidate regions and then classifies and regresses the boundaries of each region. This not only reduces computational complexity but also significantly improves detection accuracy and speed.[8] Another influential model is You Only Look Once (YOLO), which reframes object detection as a regression problem by directly mapping images to bounding boxes and category probabilities.[9] This design makes YOLO much faster than traditional methods such as the R-CNN series. Additional advantages of the YOLO series include end-to-end training and inference, simplifying the object detection pipeline.[10] However, YOLO models typically require Non-Maximum Suppression (NMS) for postprocessing, which reduces inference speed and introduces hyperparameters that can lead to unstable performance in both speed and accuracy. Moreover, since different applications place varying emphasis on recall and precision, selecting appropriate NMS thresholds remains a challenge, hindering the development of real-time detectors. Currently, one of the latest algorithms in object detection is the Real-Time Detection Transformer (RT-DETR). This model leverages Transformer architecture to capture global contextual information in images, thereby simplifying the object detection framework while improving both detection accuracy and speed.[11]

With the rapid development of deep learning technology, many researchers have applied it to the classification and detection of Chinese herbal pieces. For example, Dong et al.[12] introduced a lightweight GhostBottleneck module and an attention mechanism into the YOLOv5 algorithm. They also replaced the original convolutional layers with depthwise separable convolutions, thereby reducing the number of model parameters. Although the improved model achieved a lightweight structure, its mAP50–95 was only 72.08%, and such relatively low evaluation metrics limit its practical application. Han et al.[13] improved the DenseNet-201 model using the Keras framework, and their experiments showed that the model achieved over 90% accuracy on 43 types of Chinese herbal pieces. Similarly, Miao et al.[14] incorporated an ACmix module into the ConvNeXt model to enhance feature extraction and added a stacked FFN to improve specificity in identifying Chinese herbal pieces. This improved model achieved 90% accuracy in the classification and recognition of 6 types of Chinese herbal pieces. Although the latter 2 studies reported relatively high evaluation metrics, the limited number of herbal types in their datasets makes it difficult to generalize the results to the wide range of herbal pieces commonly used in clinical practice.

In light of the contributions and limitations of previous studies, to address the issues of limited herbal types and insufficient accuracy in existing datasets, this study established a dataset containing 201 types of Chinese medicines. The images were preprocessed and manually annotated. Both RT-DETR and YOLO models were used to identify the Chinese herbal pieces. To further enhance recognition accuracy, we designed a Feature-Focused Diffusion Network (FDN) module and integrated it into the RT-DETR model. The FDN incorporates a specialized Feature-Focus Module (FFM) and a feature diffusion mechanism, providing detailed contextual information across multiple feature scales and effectively improving the performance metrics of detection and recognition tasks for TCM.

2. Materials and Methods

2.1. Dataset

The herbal pieces used in this study were all sourced from the pharmacy of the Outpatient Department of the China Academy of Chinese Medical Sciences. To ensure clinical relevance, 201 types of herbal pieces were selected as experimental subjects. All herbs adhered to the standards established by the Outpatient Department of the China Academy of Chinese Medical Sciences. Image sampling was conducted in a single session to ensure accuracy of the information for each piece, and different species within the same genus were excluded. The processed herbs were procured through bidding from the pharmacy, with manufacturers complying with the 2020 edition of the *Chinese Pharmacopeia*, and products were distinguished according to pharmacopeial standards. The herbs used in this study were obtained from the following manufacturers: Beijing Bencao Fangyuan Pharmaceutical Group Co., Ltd., Beijing Chunfeng Yifang Pharmaceutical Co., Ltd., Harbin Pufang Pharmaceutical Co., Ltd., Beijing Renwei Traditional Chinese Medicine Decoction Pieces Co., Ltd., and Beijing Shuangqiao Yanjing Traditional Chinese Medicine Decoction Pieces Factory. The dataset contains 201 types of Chinese medicines, including 23 segmented rhizome slices, 78 sheet-like rhizome slices, 17 barks, 36 fruits or seeds, 9 leaves, 32 animal products, and 6 mineral substances. Each image contains one or multiple herbal samples, with representative images shown in Figure 1. All images were captured using a Canon EOS 80D camera, with white paper as the background and sufficient lighting provided by white LED lights. Multiple samples of each type were photographed in clustered arrangements, including parallel, stacked, and covered configurations, followed by images of individual representative samples. For granular or extremely small herbal pieces that are difficult to distinguish individually, photographs were taken in a stacked state. Ultimately, 3909 high-definition images were collected to form the dataset for this study.

The images in the dataset were preprocessed, and the sample size was augmented through flipping and rotation. Each image was manually labeled with the corresponding Chinese herbal piece name using the LabelImg tool. After completing the annotation, the 3909 images were divided into a training set (70%) and a validation set (30%) using a random sampling method, resulting in 2667 images for training and 1242 images for validation. To ensure the validity of model performance during evaluation, the images selected for the validation set were manually adjusted to maximize differences in shape and size compared with the training images. This procedure ensured that the model was trained on diverse data and could be effectively evaluated on the validation set.

2.2. RT-DETR model

The RT-DETR model is a real-time object detection model based on the transformer architecture, proposed in 2023.[11] The Transformer is a deep learning architecture designed for processing sequential data, with its core mechanism based on self-attention, which allows parallel processing and effective capture of long-range dependencies. This architecture has demonstrated strong performance across various sequence modeling tasks. Notably, RT-DETR does not require NMS when processing images, enabling faster predictions and reduced output latency.[15]

The RT-DETR model consists of 3 main modules: the backbone network, encoder, and decoder (Fig. 2). The backbone network first performs feature extraction, and the outputs of its last 3 stages (S3, S4, S5) are fed into the encoder. Within the encoder, 3 specialized modules are employed, including Attention-based Intra-scale Feature Interaction, CNN-based Cross-scale Feature Fusion Module (CCFM), and Intersection over Union (IoU)-aware Query Selection. The Attention-based Intra-scale Feature Interaction module reduces computational redundancy by performing intra-scale interactions only on the S5 layer of the backbone network.[16] CCFM incorporates several fusion blocks composed of convolutional layers into the fusion path, converting multiscale features into a series of unified image features. In this process, CCFM fuses adjacent feature maps into new features. The IoU-aware query selection module selects a fixed

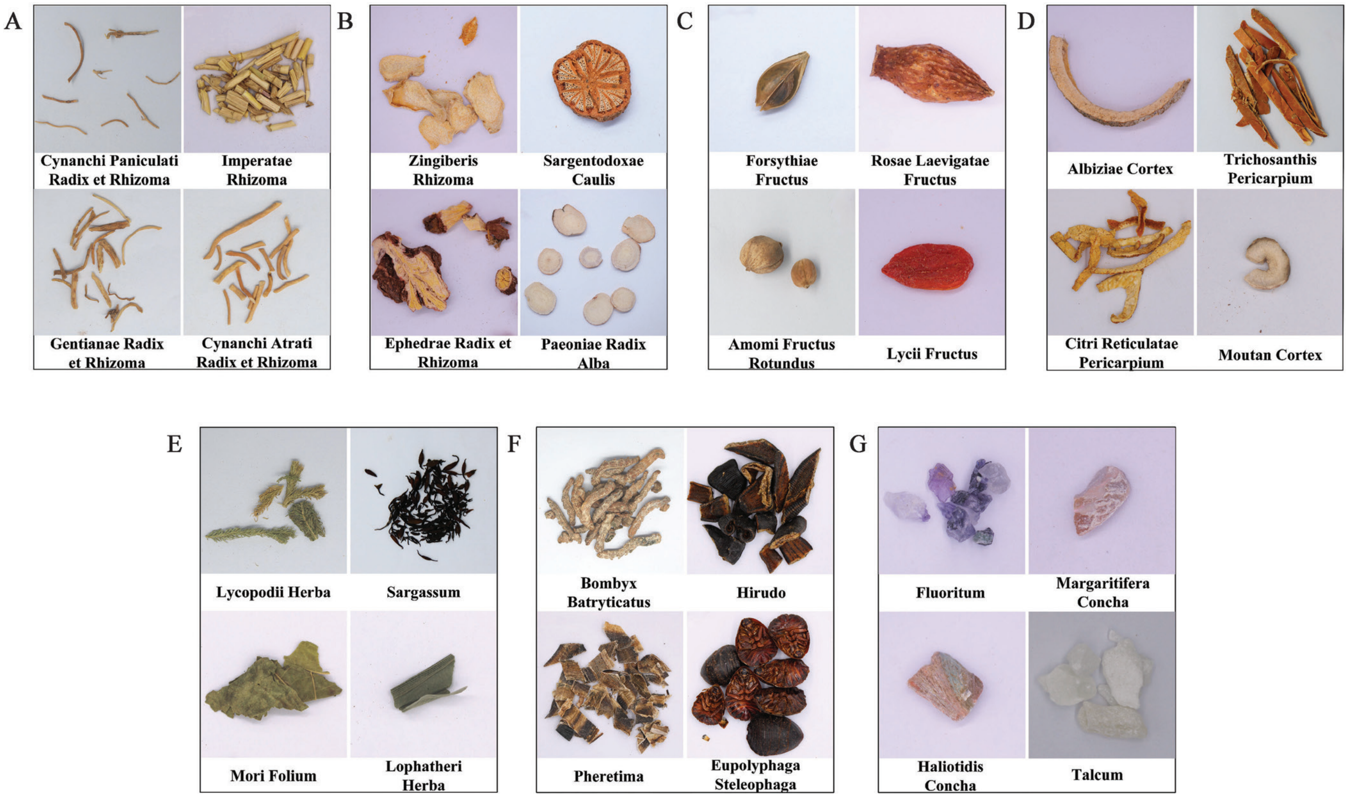

Figure 1. Representative images of Chinese medicines. (A) segmented rhizome slices; (B) sheet-like rhizome slices; (C) fruits or seeds; (D) barks; (E) leaves; (F) animals; (G) minerals.

number of image features from the encoder to serve as the initial queries for the decoder. During training, this module constrains the model to assign high classification scores to features with high IoU scores and low scores to features with low IoU scores. Consequently, the model selects prediction boxes corresponding to the top K encoder features, which have both high classification scores and high IoU scores.[17–19] Finally, the decoder iteratively refines the object queries through auxiliary prediction heads, generating bounding boxes and confidence scores for detected objects.

Regarding the loss function, the RT-DETR model employs a combination of Generalized Intersection over Union, L1 Loss, and Variational Focal Loss. Generalized Intersection over Union is an enhanced bounding box regression loss that better handles non-overlapping boxes by considering the spatial relationship between the target and the predicted box. L1 Loss evaluates model performance by calculating the absolute differences between predicted values and ground truth. Variational Focal Loss, an extension of Focal Loss, enhances the model's attention to difficult samples by incorporating variational techniques.

The original RT-DETR model supports various backbone networks, including the HGNet[20] series and the ResNet series. Among these, ResNet18 has the smallest computational and parameter requirements. ResNet18 is an 18-layer Residual Network, a type of CNN. Its structure primarily consists of an input layer, convolutional layers, residual blocks, a global average pooling layer, and a fully connected layer.[21] The initial convolutional layer of ResNet18 uses 64

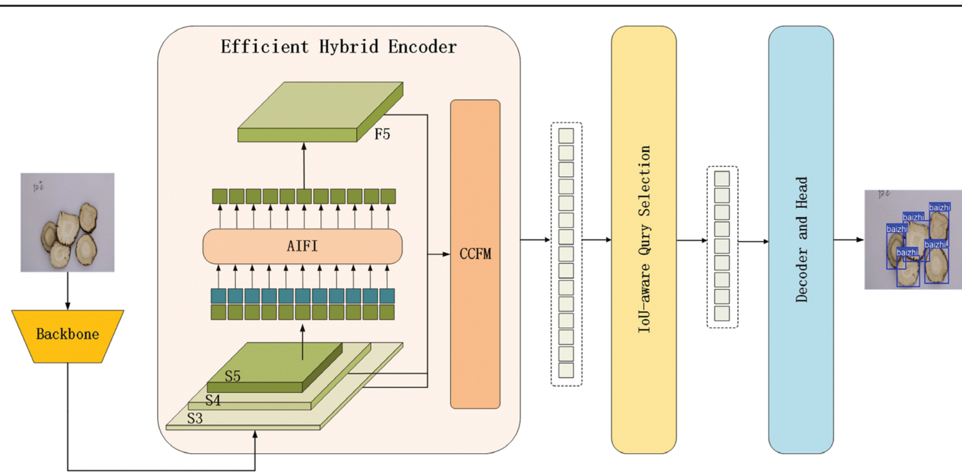

Figure 2. Structure of Real-time Detection Transformer (RT-DETR) model.

filters of size 7×7, with a stride of 2 and padding of 3, followed by a batch normalization layer[22] and a ReLU activation function,[23] and then a 3×3 max pooling layer with a stride of 2. ResNet18 consists of 4 stages, each composed of several residual blocks.[21] Each residual block includes 2 convolutional layers and a shortcut connection. The first stage consists of 2 residual blocks, each with 64 kernels, and each block comprises 2 3×3 convolutional layers.[21] The second, third, and fourth stages each contain 2 residual blocks with 128, 256, and 512 convolutional kernels, respectively.[21] In the first residual block of each stage, the first convolutional layer uses a stride of 2 for downsampling, while other convolutional layers use a stride of 1.[21] The global average pooling layer following the residual blocks converts each feature map into a single value by averaging all elements, resulting in an output size of 512.[21] Finally, the fully connected layer is used for classification tasks.

2.3. FDN network

The sample size of Chinese medicine in the dataset images is irregular, and their textures are complex. To address this, we designed an FDN, which incorporates a specialized FFM and a feature diffusion mechanism, enabling each feature scale to retain detailed contextual information. The FFM accepts inputs at 3 scales and utilizes a set of parallel deep convolutions to capture rich multiscale information. The diffusion mechanism distributes these context-rich features across various detection scales. The structure of the FFM is shown in Figure 3.

The FFM first receives 3 scales of input from the backbone network (S5, S4, S3). It then performs the following operations: upsampling and a convolution with stride 2 on S5, a convolution with stride 2 on S4 to adjust the channel dimensions, and a downsampling operation (ADown) on S3. The ADown module, proposed by Wang et al.,[24] is designed to perform downsampling efficiently while preserving more information compared with standard stride-2 convolutions, making it suitable for object detection tasks. After these operations, the 3 outputs are concatenated and passed through a series of parallel depthwise convolutions (DWConv) to capture contextual information across multiple scales. Unlike standard convolutions, DWConv applies a single convolution kernel per input channel, producing an output feature map with the same number of channels as the input. DWConv offers the advantage of faster computation and significantly reduced parameters.[25] In this study, 4 DWConvs were used, with kernel sizes of 5×5, 7×7, 9×9, and 11×11. The outputs of these DWConvs are then combined using an Add operation, which overlays the feature maps without increasing the number of channels, unlike concatenation.[26] Finally, a 1×1 convolution is used as a channel fusion mechanism to integrate features with different receptive field sizes. This design allows the FFM to capture a wide range of contextual information while preserving the integrity of local texture features in Chinese medicine images.

The FDN employs a total of 2 FFMs. The first FFM receives multiscale inputs from the backbone network (S5, S4, S3), and after performing a series of feature-focusing operations, diffuses the extracted information back to S5, S3, and the second FFM. The second FFM then receives inputs from the outputs of the first FFM, along with the fused S5 and S3 features, and performs another feature-focusing operation before further diffusing the information backward. The final RT-DETR decoder receives input that has undergone multiple stages of feature-focusing and diffusion. The feature diffusion mechanism of FDN is illustrated in Figure 4.

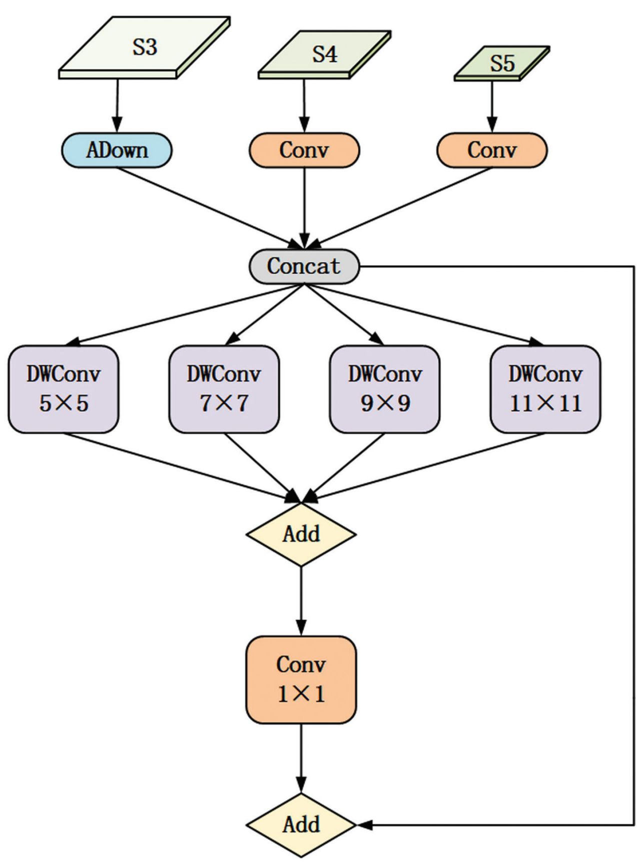

Figure 3. Structure of Feature-focus Module (FFM) in FDN.

2.4. Evaluation of model performance

To evaluate the performance of the classification models, several statistical metrics were used, including precision, recall, mAP50, mAP50–95, model parameters, and FLOPs.

Precision is defined as the proportion of detected targets that are actually positive samples, reflecting the accuracy of the model's predictions.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

Recall is defined as the proportion of actual positive samples that are correctly detected, reflecting the completeness of the model's detection results.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

Mean average precision (mAP) is the average of the average precision (AP) across all categories. AP is calculated by integrating the precision–recall curve, which is obtained by plotting precision on the vertical axis and recall on the horizontal axis. The mAP is a commonly used metric for evaluating multiclass object detection tasks. Specifically, mAP50 refers to the mAP calculated at an IoU threshold of 0.5, while mAP50–95 represents the average mAP calculated across multiple IoU thresholds, ranging from 0.5 to 0.95 in increments of 0.05.

$$\text{AP} = \int \text{Precision}(\text{Recall})\text{d}(\text{Recall}) \tag{3}$$

$$\text{mAP} = \sum_{i=1}^{\text{class}} \text{AP}(i)/\text{class} \tag{4}$$

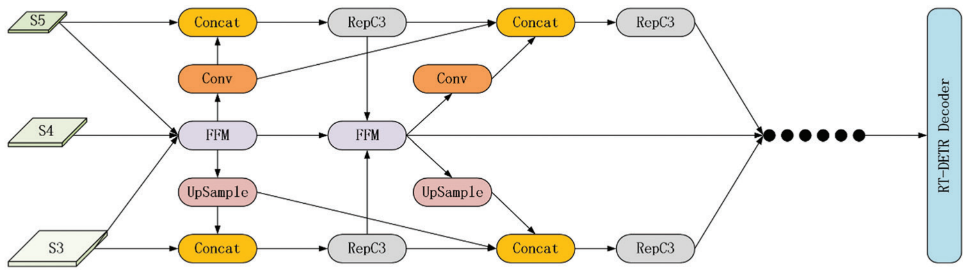

Figure 4. Feature diffusion mechanism of Feature-focused Diffusion Network (FDN).

Parameters refer to the total number of weights and bias terms that need to be learned during the training of a network model. They reflect the size and complexity of the model. Generally, a model with more parameters has greater representational capacity and can capture more complex data patterns. However, an excessive number of parameters may lead to overfitting, particularly when training data are limited. Additionally, models with more parameters require greater memory resources for both training and inference.

$$\text{Parameters} = C_{\text{in}} \times H \times W \times C_{\text{out}} + C_{\text{out}} \tag{5}$$

FLOPs represent the total number of addition and multiplication operations a model performs during forward propagation. FLOPs are commonly used to measure the computational complexity of a model. Higher FLOPs indicate greater demand for computational resources and longer inference time.

$$\text{FLOPs} = 2 \times H_{\text{out}} \times W_{\text{out}} \times C_{\text{out}} \times K_{\text{h}} \times K_{\text{w}} \times C_{\text{in}} \tag{6}$$

In the above equations, TP denotes true positives, TN denotes true negatives, FP represents false positives, and FN represents false negatives. The term class refers to the number of categories. C_{in} and C_{out} represent the number of input and output channels, respectively. H and W denote the height and width of the convolution kernel, while H_{in} and W_{in} denote the height and width of the input feature map. Similarly, H_{out} and W_{out} indicate the height and width of the output feature map, and K_h and K_w represent the height and width of the convolution kernel.

3. Results and Discussion

3.1. Model results

In this study, Ultralytics was used, which integrates object detection models such as YOLOv5, YOLOv8, YOLOv9, and RT-DETR.[27] We established 7 models: RT-DETR-R18, RT-DETR-R50, RT-DETR-L, YOLOv5m, YOLOv8m, YOLOv9c, and RT-DETR-FDN. RT-DETR-R18, RT-DETR-R50, and RT-DETR-L are RT-DETR models with ResNet18, ResNet50, and HGNetv2 as backbone networks, respectively. YOLOv5m, YOLOv8m, and YOLOv9c are built using the YOLOv5, YOLOv8, and YOLOv9 algorithms, respectively. RT-DETR-FDN was created by integrating the designed FDN module into RT-DETR-R18. The optimizer was set to AdamW with a momentum of 0.9, weight decay of 0.0001, an initial learning rate of 0.01, and 2000 warm-up iterations. The cosine learning rate scheduler was not applied during training. For data augmentation, the Albumentations library[28] was employed with the following settings: Blur (P = 0.01), MedianBlur (P = 0.01), ToGray (P = 0.01), and CLAHE (P = 0.01).

To ensure experimental fairness, all models were trained for 100 epochs under the same experimental environment and parameter settings. Each model employed identical data augmentation methods using the Albumentations library, with the same parameter settings, as well as the AdamW optimizer with consistent hyperparameters. The number of training epochs and all other parameters were kept uniform across models. The results of the experiments are presented in Table 1.

For the RT-DETR-R18 model, predictions were performed on the validation set after 100 epochs of training on the dataset. The experimental results showed that the AP, recall, mAP50, and mAP50-95 of RT-DETR-R18 across all types of Chinese medicine were 0.907, 0.936, 0.966, and 0.843, respectively. After integrating FDN and training for 100 epochs, RT-DETR-FDN achieved AP, recall, mAP50, and mAP50–95 values of 0.925, 0.943, 0.974, and 0.851, respectively. The trends of precision, recall, mAP50, and mAP50-95 with respect to the number of training epochs for the 2 models are shown in Figure 5. The evaluation metrics of both models tend to converge after 80 epochs. Compared with RT-DETR-R18, RT-DETR-FDN not only achieved higher evaluation metrics under the same number of training epochs, but also showed more pronounced improvements at earlier stages of training. This indicates that RT-DETR-FDN can deliver better performance while requiring fewer computational resources.

The mAP is commonly used as the primary evaluation metric for multiclassification tasks. However, compared with mAP, mAP50-95, which considers a wider range of IoU thresholds, provides a more comprehensive reflection of model performance in both target localization and classification. For the RT-DETR-18 model, the mAP50-95 exceeded 0.8 for 143 types of Chinese medicines, exceeded 0.9 for 73 types, and fell below 0.7 for 21 types. The improved RT-DETR-FDN model achieved an increase of 0.008 in mAP50-95. Specifically, the number of categories with mAP50-95 greater than 0.8 increased by 5 compared with the baseline experiment, reaching 148, while the number of categories with mAP50–95 less than 0.7 decreased by 9, reducing to 12. These results indicate that the RT-DETR-FDN model achieves varying degrees of improvement across most Chinese medicine categories in terms of evaluation metrics.

Compared with the RT-DETR-R18 model, the RT-DETR-FDN model showed an 11.9% increase in parameters and a 15.7% increase in FLOPs. In terms of performance, its AP across all types of Chinese medicines increased by 0.018, average recall by 0.007, and both mAP50 and mAP50-95 by 0.008. YOLOv5m, which has a scale comparable to RT-DETR-FDN, has 11.5% more parameters but 3.5% fewer FLOPs. However, its performance was lower, with a precision of 0.824, recall of 0.819, mAP50 of 0.903, and mAP50–95 of 0.735. Other models also showed increases in parameters and FLOPs, but none outperformed RT-DETR-FDN in evaluation metrics. The RT-DETR-R50 model, which uses ResNet50 as its backbone, had the highest parameter count and FLOPs, yet its precision (0.923), recall (0.923), mAP50 (0.958), and mAP50-95 (0.826) were all lower than those of RT-DETR-FDN. Similarly, the RT-DETR-L model with HGNetv2 as its

Table 1

Model results.

Model	Parameters (M)	FLOPs (G)	Precision	Recall	mAP50	mAP50–95
RT-DETR-R18	20.1	57.9	0.907	0.936	0.966	0.843
RT-DETR-FDN	22.5	67.0	0.925	0.943	0.974	0.851
RT-DETR-R50	42.3	130.5	0.923	0.923	0.958	0.826
RT-DETR-L	32.4	104.4	0.909	0.909	0.954	0.820
YOLOv5m	25.1	64.6	0.824	0.819	0.903	0.735
YOLOv8m	25.9	79.4	0.877	0.870	0.936	0.773
YOLOv9c	25.4	103.2	0.927	0.929	0.970	0.816

backbone achieves precision of 0.909, recall of 0.909, mAP50 of 0.954, and mAP50-95 of 0.820. Among the YOLO series, the YOLOv8m model achieved precision of 0.877, recall of 0.870, mAP50 of 0.936, and mAP50-95 of 0.773, while the YOLOv9c model performed better with precision of 0.927, recall of 0.929, mAP50 of 0.970, and mAP50–95 of 0.816. Overall, compared with all other models, RT-DETR-FDN achieved the best evaluation metrics while remaining lightweight. This not only highlights the efficiency of the FDN network but also demonstrates that RT-DETR-FDN provides superior accuracy in detecting and recognizing Chinese medicine.

3.2. Discussion of the results

Compared with the RT-DETR-R18 model, the RT-DETR-FDN model showed significant improvements in mAP50-95 for 9 types of Chinese medicines: Asari Radix et Rhizoma, Clematidis Radix et Rhizoma, Oroxyli Semen, Phellodendri Amurensis Cortex, Pogostemonis Herba, Ganoderma, Cuscutae Semen, Taraxaci Herba, and Asteris Radix et Rhizoma (Fig. 6). We analyzed both their morphological characteristics and evaluation indicators, with the comparative results presented in Figure 6. Specifically, the RT-DETR-FDN model improved mAP50–95 by 0.072, 0.123, 0.157, and 0.135 in identifying Asari Radix et Rhizoma, Clematidis Radix et Rhizoma, Pogostemonis Herba, and Asteris Radix et Rhizoma, respectively. For Cuscutae Semen and Taraxaci Herba, the improvements were 0.247 and 0.119, respectively. The model also achieved increases of 0.161 and 0.108 in mAP50-95 for Ganoderma and Phellodendri Amurensis Cortex, respectively, and an improvement of 0.082 for Oroxyli Semen.

In terms of morphology, Asari Radix et Rhizoma, Clematidis Radix et Rhizoma, Pogostemonis Herba, and Asteris Radix et Rhizoma are generally slender and cylindrical. Ganoderma and Phellodendri Amurensis Cortex are sheet-like, but with complex textures and diverse colors. Oroxyli Semen (commonly known as wooden butterfly) has blurred edges, the dodder is as small as particles, and the dandelion has an irregular shape like weeds. The challenge in identifying these types of herbal medicines lies in their small-target size and difficulty in accurate detection

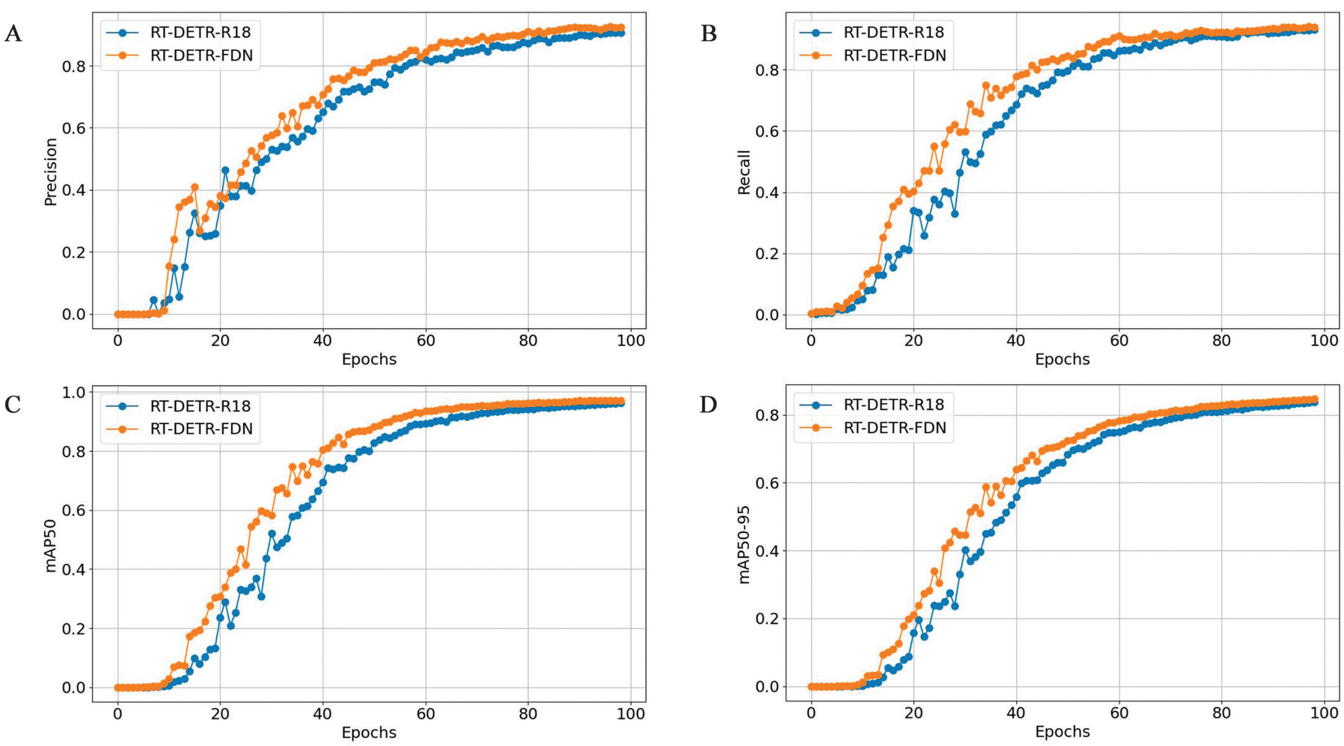

Figure 5. Variation of (A) precision, (B) recall, (C) mAP50, and (D) mAP50–95 of RT-DETR-FDN model and RT-DETR-R18 model with the number of training epochs. The horizontal axis represents the number of training epochs, the vertical axis represents various evaluation metrics, the orange line represents the RT-DETR-FDN model, and the blue line represents the RT-DETR-R18 model.

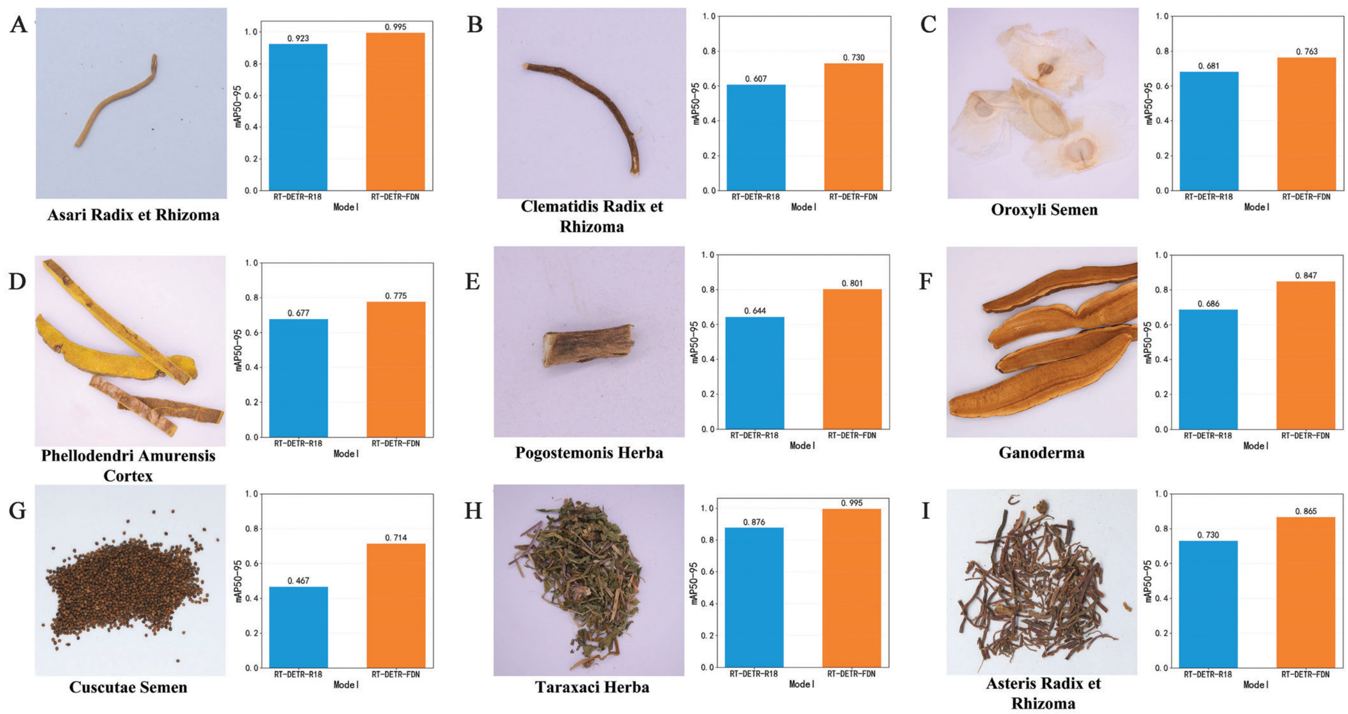

Figure 6. Images of 9 Chinese medicines, which are significantly improved in mAP50-95 of the RT-DETR-FDN model compared with the RT-DETR-R18 model. The 9 Chinese medicines are (A) Asari Radix et Rhizoma, (B) Clematidis Radix et Rhizoma, (C) Oroxyli Semen, (D) Phellodendri Amurensis Cortex, (E) Pogostemonis Herba, (F) Ganoderma, (G) Cuscutae Semen, (H) Taraxaci Herba, (I) Asteris Radix et Rhizoma.

when stacked or obstructed. RT-DETR-FDN can better address this challenge. The FDN, with its feature-focusing module and feature diffusion mechanism, can capture multiscale feature information of these small-target herbs, enabling the model to better learn and distinguish their characteristics.

The model prediction results showed that the detection and recognition performance of Puerariae Lobatae Radix, Angelicae Sinensis Radix, Atractylodis Rhizoma, Rehmanniae Radix Praeparata, Smilacis Glabrae Rhizoma, Gastrodiae Rhizoma, Asari Radix et Rhizoma, Haliotidis Concha, Sparganii Rhizoma, Coicis Semen, Sargassum, Taraxaci Herba, and Artemisiae Annuae Herba was the best, with an mAP50–95 value of 0.995. Most of these Chinese medicines have clear textures, well-defined edges, and regular shapes. However, a few species, such as Artemisiae Annuae Herba, Taraxaci Herba, and Sargassum, are herbaceous and often appear stacked or overlapped. This demonstrates that RT-DETR-FDN not only performs reliably in detecting Chinese medicines with conventional shapes but also shows advantages in recognizing small-target Chinese medicines under complex conditions. In contrast, the RT-DETR-FDN model yielded mAP50-95 values below 0.6 for 7 out of the 201 Chinese herbal pieces analyzed, namely Gentianae Macrophyllae Radix (0.473), Vaccariae Semen (0.496), Tritici Levis Fructus (0.521), Bombyx Batryticatus (0.538), Gardeniae Fructus Praeparatus (0.551), Gentianae Radix et Rhizoma (0.568), and Pseudostellariae Radix (0.597). The results of the RT-DETR-FDN model are shown in Supplemental Table S1, https://links.lww.com/STCM/A72.

We believe that the relatively low evaluation metrics for Gentianae Macrophyllae Radix and Gentianae Radix et Rhizoma are mainly due to the large variations in their length and thickness, as well as the high degree of overlap in color and texture between these 2 decoction pieces, which makes it difficult for the model to differentiate their features. Both Vaccariae Semen and Tritici Levis Fructus resemble rice grains in shape. Among them, Vaccariae Semen is smaller in size and exhibits inconsistent grain colors, further increasing the difficulty of recognition. Notably, the recall rate of Vaccariae Semen was significantly lower than that of other herbal decoction pieces. With a recall rate of only 0.5, the model demonstrated a weak predictive ability for its true positive samples. In addition, the shape and color of Bombyx Batryticatus and Pseudostellariae Radix are relatively similar, while the length of Bombyx Batryticatus varies considerably, which we consider a major factor contributing to its low evaluation index. For Gardeniae Fructus Praeparatus, the relatively dark coloration of its slices may reduce image contrast, making it difficult for the model to capture textural features. Furthermore, the dataset contained relatively few individual samples of Gardeniae Fructus Praeparatus, limiting the model's ability to learn diverse features. Beyond these factors, the small sample sizes available for these 7 decoction pieces likely played an important role in their lower mAP50–95 values. In practice, these Chinese medicines are also difficult to distinguish visually and often require additional information, such as odor and taste, for accurate identification.

4. Conclusions

To address the issues of subjectivity, complexity, time consumption, and high cost in traditional manual identification of Chinese medicines, this study proposed a recognition method based on an improved RT-DETR model. We first constructed a dataset of 201 types of Chinese herbal pieces, containing high-definition images captured with professional cameras and accurately annotated by experts. To better capture the irregular edges and complex textures of herbal pieces, we designed an FDN integrated into the RT-DETR framework. The FDN includes an FFM and a feature diffusion mechanism. Experimental results demonstrate that the RT-DETR-FDN model not only maintains a relatively small number of parameters and FLOPs but also significantly improves precision, recall, and mAP in the detection and recognition of herbal pieces. Compared with other mainstream object detection models, RT-DETR-FDN achieves superior performance while remaining lightweight. Specifically, it achieved a precision of 0.925, a recall of 0.943, and an mAP50–95 of 0.851. This study highlights the potential of combining deep learning technology with TCM to enable efficient detection and recognition, reduce manual workload, and improve the overall quality control of Chinese medicines.

Acknowledgments

The authors gratefully acknowledge the support from the Outpatient Department of the China Academy of Chinese Medical Sciences.

Statement of ethics

None declared.

Conflict of interest statement

The authors declare that they have no conflicts of interest with regard to the content of this report.

Funding source

This work was supported by the Science and Technology Innovation Project of the Chinese Academy of Traditional Chinese Medicine (CACMS Innovation Fund, CI2023E002), the Shandong Province Traditional Chinese Medicine Technology Project (M20241738), and the Shandong Province Graduate High-Quality Education and Teaching Resource Project (SDYAL2022041).

Data availability statement

The datasets generated and/or analyzed during this study are available from the corresponding author upon reasonable request.

Author contributions

Tianchi Mao wrote this manuscript. Xing Sun and Jiayin Zhu contributed to data preparation. An Liu ensured that the scientific aspect of the study is rationally valid. Cong Guo conceived the main theme on which the work is performed. Yang Li and Jingang Ma designed the project and analyzed the data. All authors have read and approved the final manuscript.

References

[1] Marshall A. Traditional Chinese medicine and clinical pharmacology. *Drug Discov Eval Methods Clin Pharmacol* 2020;23:455–482.
[2] Che CT, George V, Ijinu TP, et al. *Traditional medicine*. Pharmacognosy Academic Press 2024; 11-28.
[3] Zhou X, Li CG, Chang D, Bensoussan A. Current status and major challenges to the safety and efficacy presented by Chinese herbal medicine. *Medicines (Basel, Switzerland)* 2019;6:14.
[4] Newman DJ. Modern traditional Chinese medicine: identifying, defining and usage of TCM components. *Adv Pharmacol* 2020;87:113–158.
[5] Zhao X, Li Y, Yang C, et al. Research on dataset construction and algorithm evaluation of Chinese herbal pieces recognition. International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI). Yinchuan, NX, China, 2023; 12941:633-639.
[6] Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, 2005; 1:886-893.
[7] Lindeberg T. Scale invariant feature transform. *Scholarpedia* 2012;7:10491.
[8] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. *IEEE T Pattern Anal* 2018;42:386–397.
[9] Jiang P, Ergu D, Liu F, Cai Y, Ma B. A review of YOLO algorithm developments. *Procedia Comput Sci* 2022;199:1066–1073.
[10] Kang CH, Kim SY. Real-time object detection and segmentation technology: an analysis of the YOLO algorithm. *JMST Adv* 2023;5:69–76.
[11] Zhao Y, Lv W, Xu S, et al. Detrs beat yolos on real-time object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, 2024; 16965-16974.
[12] Dong M, Liang Y, Liu Y, Qiao Z, Niu H, Ge G. Detection and recognition of traditional Chinese medicine pieces based on improved YOLOv5. *Modern Computer* 2022;28:9–16. [in Chinese].
[13] Han Y, Lan J, Guo R, et al. Identification of Chinese herbal medicine slices based on deep learning. *Acta Agric Boreali-occident Sin* 2023;32:1859–1867. [in Chinese].
[14] Miao J, Huang Y, Wang Z, Wu Z, Lv J. Image recognition of traditional Chinese medicine based on deep learning. *Front Bioeng Biotechnol* 2023;11:1199803.
[15] Hosang J, Benenson R, Schiele B. Learning non-maximum suppression. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, 2017; 4507-4515.
[16] Zhu X, Su W, Lu L, et al. Deformable DETR: Deformable transformers for end-to-end object detection. International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020; 1-7.
[17] Cheng G, Si Y, Hong H, Yao X, Guo L. Cross-scale feature fusion for object detection in optical remote sensing images. *IEEE Geosci Remote S* 2020;18:431–435.
[18] Rezatofighi H, Tsoi N, Gwak JY, et al. Generalized intersection over union: A metric and a loss for bounding box regression. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, 2019; 658-666.
[19] Zheng D, Dong W, Hu H, et al. Less is more: focus attention for efficient detr. Proceedings of the IEEE/CVF International Conference on Computer Vision. Vancouver, Canada, 2023; 6674-6683.
[20] Xia H, Ding Z. HGNet: Hybrid generative network for zero-shot domain adaptation. European Conference on Computer Vision. Glasgow, United Kingdom, 2020; 55-70.
[21] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, 2016; 770-778.
[22] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning. Lille, France, 2015; 448-456.
[23] Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010; 807-814.
[24] Wang CY, Yeh IH, Liao HYM. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint* 2024;2402:1–21.
[25] Chen Z, He X, Ren C, Chen H, Zhang T. Enhanced separable convolution network for lightweight jpeg compression artifacts reduction. *IEEE Signal Process Lett* 2021;28:1280–1284.
[26] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, CA, 2017;31:4278-4284.
[27] Download URL of Ultralytics on Github. Available at: https://github.com/ultralytics/ultralytics. Accessed August 29, 2025.
[28] Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: fast and flexible image augmentations. *Information* 2020;11:125–120.

How to cite this article: Mao T, Sun X, Zhu J, Liu A, Li Y, Ma J, Guo C. Study on image detection and target recognition based on traditional Chinese medicine. *Sci Tradit Chin Med* 2026;4(1):73–80. doi: 10.1097/st9.0000000000000096