Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(6): 3655–3677.
DOI: 10.3934/era.2026165
Received: 26 February 2026
Revised: 09 April 2026
Accepted: 16 April 2026
Published: 29 April 2026

Research article

Dual attention enhancement network for Chinese herbal medicine classification

Min Fu[1,*] **and Hanyu Hong**[2]

[1] School of Mathematics and Physics, Wuhan Institute of Technology, LiuFang Campus, No.206, Guanggu 1st road, Donghu New & High Technology Development Zone, Hubei 430205, China

[2] School of Electrical and Information Engineering, Wuhan Institute of Technology, LiuFang Campus, No.206, Guanggu 1st road, Donghu New & High Technology Development Zone, Hubei 430205, China

* **Correspondence:** Email: minfujoyful@163.com.

Abstract: Chinese herbal medicine (CHM) classification is an important and emerging topic in intelligent medicine. However, due to the limited local receptive field, existing Convolutional Neural Network-based methods are susceptible to background interference, which fails to capture discriminative visual cues and hampers the accurate recognition of visually similar herbal categories. To address these limitations, we propose a dual attention enhancement network for Chinese herbal medicine classification. Specifically, we introduce an object localization module to suppress background interference by accurately localizing the target region, thus guiding the network to focus on discriminative regions. Subsequently, we introduce a fused attention module to integrate horizontal and vertical directional information at different feature scales to capture long-range spatial dependencies and enhance the global contextual perception. Moreover, we propose a dual attention module composed of self-attention and cross-attention mechanisms to achieve explicit fusion and the semantic alignment of multi-scale features. Finally, we build a semantic feature enhancement module to further strengthen the inter-layer complementarity through adaptive semantic fusion, thereby improving the discriminative ability and robustness of feature representations. Extensive experimental results on two CHM datasets demonstrate that the proposed method outperforms existing state-of-the-art methods.

Keywords: dual attention enhancement; object localization; Chinese herbal medicine classification

1. Introduction

Chinese herbal medicine (CHM) is an essential component of traditional Chinese medicine (TCM), which plays an irreplaceable role in the prevention [1–3], diagnosis, and treatment [4, 5] of diseases. However, due to the high visual similarity among certain CHMs, non-experts often find it difficult to accurately distinguish these categories, and the accidental ingestion of visually similar CHMs with opposite properties may pose serious health risks [6]. Therefore, achieving an accurate classification of CHMs has become a major challenge in the field of intelligent traditional Chinese medicine research, particularly in leveraging the essential information contained in CHM images to accurately identify their categories and perform precise classification [7–9].

To address the problem of CHM classification, traditional machine learning algorithms such as K-Nearest Neighbor (KNN) [10], Support Vector Machine (SVM) [11], Decision Tree (DT) [12], and Bayesian methods [13] have been widely used in CHM classification tasks. For example, Mallah et al. [14] first extracted three types of leaf features, then used KNN to classify each type of feature or the combined features, and finally employed a probabilistic method to integrate the classification results from different features. Luo et al. [15] combined an electronic nose with computer vision to extract both scent and color characteristics, and then employed an SVM for classification. However, these methods heavily rely on handcrafted features and are only suitable for small-scale datasets.

In recent years, with the rapid development of deep learning [16, 17], an increasing number of CNNs, such as AlexNet [18], VGG [19], GoogleNet [20], and ResNet [21–23], owing to their strong capability in extracting local visual features, have been widely applied to Chinese herbal medicine classification and recognition tasks, which achieve remarkable performances. For instance, Liu et al. [20] employed GoogleNet to automatically extract visual features from CHM datasets and achieve end-to-end classification. Cai et al. [23] integrated a CNN with a broad learning system to perform CHM identification, thus achieving promising recognition results. However, these methods are usually constrained by the limited receptive field of CNNs, which allows the network to only focus on features within the local neighborhood of each region. Therefore, the model fails to capture global contextual relationships and long-range dependencies, thereby impairing its ability to distinguish visually similar categories under complex backgrounds.

Based on above analysis, we propose a dual attention enhancement network for CHM classification, which attempts to tackle two critical challenges in classification tasks: 1) most samples in CHM datasets are collected through web crawling, which inevitably introduces watermarks and background noise, thus making it difficult to extract discriminative features; and 2) the inherent local receptive fields of CNNs lead to a local bias that limits their ability to capture long-range dependencies and global contextual information. Specifically, to address 1), we introduce an object localization module to suppress the impact of background interference and extract discriminative features beneficial for the classification task. To address 2), we sequentially introduce a fused attention module, a dual attention module, and a semantic feature enhancement module. These modules work collaboratively, from spatial focus and structural fusion to semantic enhancement, thereby effectively alleviating the local bias problem of CNNs. Specifically, we first introduce a fused attention module to capture long-range spatial dependencies by integrating horizontal and vertical directional information. Then, a dual attention module is designed to explicitly fuse multi-scale features through self-attention and cross-attention mechanisms, thereby bridging the semantic gap

across hierarchical layers. Finally, we propose a semantic feature enhancement module to fuse complementary semantics across layers, which further strengthens the feature representation and robustness. The effectiveness of our method was validated on two constructed CHM datasets that contain 20 and 95 categories, respectively. Experimental results demonstrate that our approach consistently achieves high recognition accuracy across all categories and outperforms existing methods on both datasets.

In summary, the main contributions of this paper are listed as follows:

(1) We propose an object localization module that alleviates the impact of background information while focusing on discriminative regional features.
(2) We propose a local bias-compensated learning framework composed of a fused attention module, a dual attention module, and a semantic feature enhancement module. It alleviates the locality limitation from three complementary perspectives: spatial dependency modeling, cross-layer structural fusion, and semantic feature enhancement, which captures more comprehensive and discriminative representations.
(3) Extensive experimental results on two CHM datasets demonstrate our superiority over the state-of-the-art methods.

2. Related works

Recently, some methods have been proposed to address the CHM image classification and recognition tasks, which can be roughly divided into handcrafted methods and deep learning methods. Handcrafted methods usually employ techniques such as color and texture analysis, random forests, and SVMs, and generally focus on handcrafted features designed based on domain knowledge. For example, Liu et al. [24] extracted image texture features using the gray-level co-occurrence matrix (GLCM) and applied them to the recognition of CHMs. Luo et al. [15] proposed an electronic nose with computer vision to extract both olfactory and color features, and subsequently employed an SVM for classification. However, these methods heavily rely on manually designed features, which often limits their ability to capture complex and abstract representations.

With the rapid development of neural networks, Fine-Grained Visual Categorization (FGVC) methods can be roughly classified into CNN-based methods [25–28] and vision transformer methods [29–34]. Existing CNN-based methods can be further divided into feature encoding methods and part localization methods. Feature encoding methods [25, 28] mainly utilize higher-order feature encoding to explore the spatial relationships of high-level pixels and discover subtle differences. Part localization methods [26, 27] focus on either identifying discriminative regions or devising an effective strategy for salient object feature extraction to facilitate the FGVC task. Recently, the vision transformer (VIT [35]) adopts a sequential patch embedding strategy combined with a multi-head self-attention mechanism, thus enabling it to globally learn attention maps that highlight discriminative regions. For example, TransFG [29] utilizes the attention links among tokens as an indicator of the importance to select discriminative tokens for distinguishing subcategories. Recently, Multi-granularity part sampling attention (MPSA) [31] proposes a part sampling attention mechanism to capture variable-shaped discriminative parts for fine-grained classification. Fine-grained anttention-locating vision transformer (FAL-ViT) [33] proposes a fine-grained attention locating vision transformer to locate target information from patches.

In the field of CHM research, some studies have applied CNNs to CHM recognition task due to their powerful local feature extraction capability. Sun and Qian [7] applied the VGGNet-16 as backbone with transfer learning techniques to effectively recognize and retrieve images of 100 types of CHMs. Similarly, Li et al. [36] employed a LeNet-5 with multi-scale image inputs and data augmentation strategies to improve the accuracy of CHM classification. Cai et al. [23] integrated a VGGNet-19 with a broad learning system to identify CHMs, thus achieving superior recognition performance. Hu et al. [37] incorporated the concept of multi-task learning by employing neural networks as the primary framework and traditional features as complementary information for Chinese herbal slice recognition. Xing et al. [38] applied a ResNet-based deep transfer learning model to recognize and classify images of traditional Chinese medicines. Chen and Zou [39] proposed an intelligent screening method for Chinese medicinal pieces based on a BMFNet-WGAN model to enhance the recognition accuracy and feature generation quality. Liu et al. [40] developed a Chinese herbal plant classification approach that combines image segmentation with the GoogLeNet-based deep learning framework, thereby effectively tackling challenges such as the large number of categories, high morphological similarity, and complex backgrounds in Chinese medicine image datasets. Recently, some methods [41, 42] have been proposed to improve traditional CNNs, thereby making them more suitable for CHM classification and recognition tasks. For example, Miao et al. [41] enhanced the traditional VGG network by introducing an interleaved three-layer Conv-BN–Max Pooling structure and employing global average pooling to improve the feature extraction and generalization, thus making it more effective for CHM recognition tasks. Huang and Xu [42] enhanced the ResNet101 backbone by integrating SENet and a convolutional block attention module, and further applied Bayesian optimization, thus enabling a more accurate classification. Although these methods focus on enhancing the CNNs' ability to capture subtle visual differences among CHMs, they overlook the limitations of CNNs' local receptive fields, which may impair the acquisition of global contextual information and consequently affect the completeness and discriminative ability of the feature representation.

3. Methods

Figure 1 illustrates the whole framework. Our model mainly consists of four components; an object localization module, a fused attention module, a dual attention module, and a semantic feature enhancement module. The object localization module enables the model to effectively eliminate inaccuracies in object localization caused by background interference while filtering out irrelevant information. The fused attention module integrates multi-scale information along the horizontal and vertical directions to effectively capture long-range dependencies in the spatial domain, thus enabling the extraction of more fine-grained and discriminative feature representations. The dual attention module combines self-attention and hierarchical cross-attention mechanisms, thus enabling the model to focus on key regions within the same level while integrating cross-scale features to capture the overall semantic representations. The semantic feature enhancement module further strengthens inter-layer complementarity through adaptive semantic fusion, thereby improving the discriminative ability and robustness of feature representations.

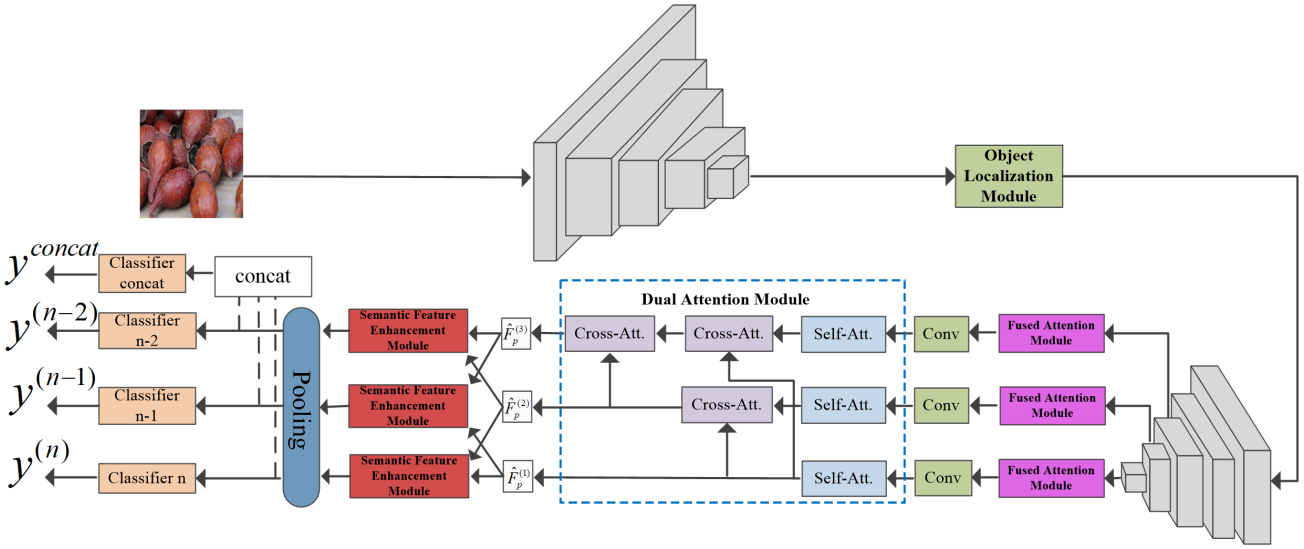

Figure 1. The overview of the network structure. First, we introduce the object localization module to alleviate the impact of background information. Then, we introduce the fused attention module to integrate horizontal and vertical directional information across layers. The dual attention module is designed to enable explicit fusion and semantic alignment across multi-scale features. Finally, we introduce the semantic feature enhancement module to further enhance inter-layer complementarity via adaptive semantic integration.

3.1. Object localization module

In CHM classification and recognition tasks, images may be affected by background noise, which may distract the model's attention from the target regions during the training phase. To solve the above problem, we introduce an object localization module to focus on the target regions within the image while ignoring the background as much as possible. Figure 2 illustrates the design of the object localization module.

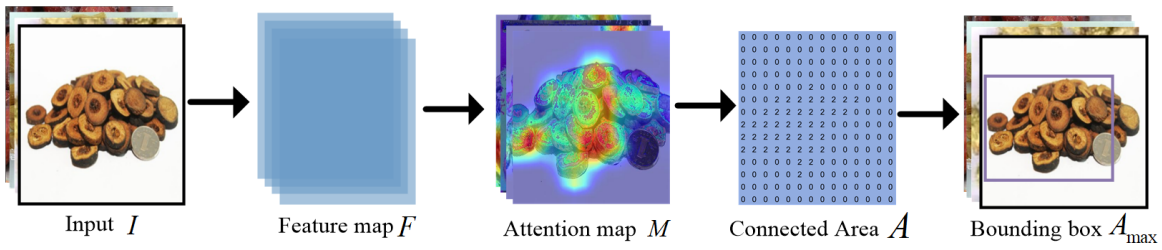

Figure 2. The pipeline of the object localization module.

Specifically, given the input I, the feature map F is obtained from the last convolutional layer of the backbone. Then, we use channel-wise average pooling to perform compression along the channel dimension and generate the attention map M; this approach can roughly reflect the spatial distribution of the target. Subsequently, to accurately identify the target regions, we further convert the attention map M into a binary mask. More specifically, we set a drop threshold α, which is a certain proportion of

the maximum value in M. If the intensity of a pixel exceeds this threshold, then the pixel is considered part of the target region and the corresponding position in the mask is set to 1, while the remaining positions are set to 0, finally obtain a coarse foreground mask M_{mask} as follows;

$$M_{\text{mask}} = \begin{cases} 1, & \text{if } M_{(i,j)} > \alpha \times \max(M_{(i,j)}) \\ 0, & \text{otherwise} \end{cases}, \tag{3.1}$$

where α denotes a hyper-parameter. However, the mask generated in this way often still contains some falsely activated regions, which correspond to non-target areas in the images. To further refine the target regions, we introduce the connected component analysis algorithm from traditional image processing. This algorithm is used to identify connected sub-regions in the mask that consist of adjacent pixels sharing the connected area A:

$$A = \omega(M_{mask}) = \{A_0, A_1, \cdots, A_n\}, \tag{3.2}$$

where ω denotes the connectivity component algorithm. We choose the maximum area A_{max}, calculate a predicted bounding box based on its boundary. Finally, the target-containing region is cropped from the original image according to the bounding box. During training, we utilize random crop and random horizontal flip, while we use center crop during testing. Then, we feed it back into the network for subsequent processing, which effectively alleviates the impact of irrelevant background information.

3.2. Fused attention module

Different layers focus on different levels of semantic information. The ResNet feature extractor consists of five stages. As the network depth increases, the spatial resolution of the feature maps gradually decreases, while the number of channels increases. Consequently, the shallow layers are more attentive to local details in the image, whereas the deeper layers capture more global semantic information. However, ResNet primarily relies on local receptive fields and poorly performs in capturing long-range spatial dependencies and global contextual information. Inspired by [43], we introduce a fused attention module (FAM). Specifically, FAM independently operates on each of the last three stages of ResNet, thereby applying directional pooling and convolution operations along the horizontal and vertical dimensions within each stage's feature map. This design enhances the spatial representation of each scale by supplementing it with direction-aware global priors, without introducing interactions across different stages. The output of each preceding stage is fed as input to the subsequent one following the original ResNet architecture, thus preserving the inherent inter-stage structure. Figure 3 illustrates the design of the fused attention module.

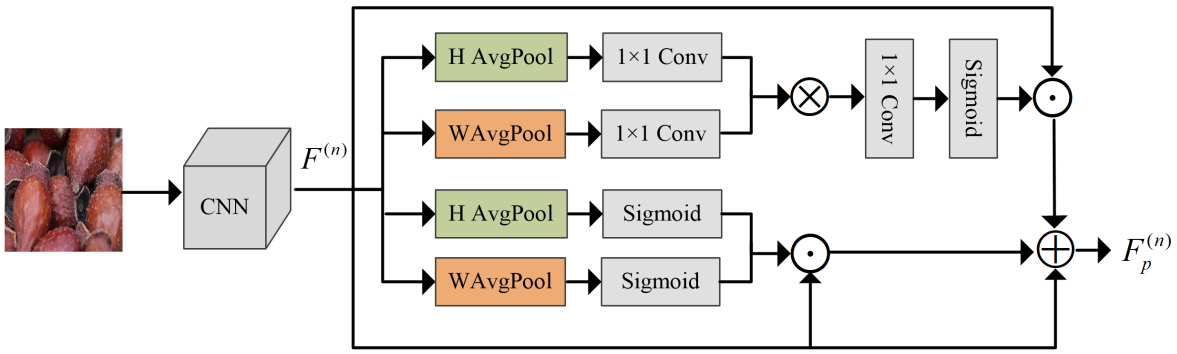

Figure 3. The pipeline of the fused attention module.

Specifically, the features refined by the object localization module are forwarded into the ResNet backbone. At the n-th stage, the corresponding feature map is represented as $F^{(n)} \in \mathbb{R}^{C_n \times H_n \times W_n}$. To capture directional contextual cues, global average pooling is independently performed along the horizontal and vertical dimensions, producing $F_h^{(n)} \in \mathbb{R}^{C_n \times H_n}$ and $F_w^{(n)} \in \mathbb{R}^{C_n \times W_n}$, respectively. This operation enables the model to establish long-range dependencies across different spatial orientations. Subsequently, both $F_h^{(n)}$ and $F_w^{(n)}$ are passed through a 1×1 convolution, followed by a ReLU activation to perform dimensionality compression, thus yielding the compact representations F_h^{conv} and F_w^{conv} as follows:

$$F_h^{conv} = \text{ReLu}(Conv(F_h^{(n)})), \quad F_w^{conv} = \text{ReLu}(Conv(F_w^{(n)})), \tag{3.3}$$

where $Conv$ denotes the 1×1 convolution. To generate features enriched with more global prior information, the features from the two directions are multiplied element-wise, followed by a 1×1 convolution, ReLU activation, and a sigmoid function to obtain the following:

$$F_l^{(n)} = \sigma(\text{ReLU}(Conv(F_h^{\text{conv}} \otimes F_w^{\text{conv}}))). \tag{3.4}$$

This process first compresses the channels to reduce the computational complexity, and then restores them to the original dimensions to preserve the representational capacity. At the same time, we encode the F_h^{conv} and F_w^{conv} into two directional attention maps, which respectively indicate whether each row and each column contains regions of interest, then combine the generated attention maps through element-wise multiplication to obtain the output $F_r^{(n)}$ as follows;

$$F_r^{(n)} = \sigma(Conv(F_h^{\text{conv}})) \times \sigma(Conv(F_w^{\text{conv}})). \tag{3.5}$$

Finally, $F_l^{(n)}$ and $F_r^{(n)}$ are used as weights applied to the original feature map $F^{(n)}$. By performing dot products followed by summation, the final output feature $F_p^{(n)}$ is obtained, thus achieving enhanced fusion of global and local information as follows;

$$F_p^{(n)} = F^{(n)} \oplus (F_l^{(n)} \odot F^{(n)}) \oplus (F_r^{(n)} \odot F^{(n)}). \tag{3.6}$$

3.3. Dual attention module

Although the fused attention module effectively enhances spatial representations within each feature scale by integrating directional information, it independently operates on each scale without establishing any interaction across different hierarchical levels. As a result, the complementary information between scales, where shallow features capture fine-grained details and deep features encode rich semantics remains largely unexploited, leads to incomplete feature fusion and a potential loss of discriminative information. To address this limitation, we design a dual attention module to establish more comprehensive multi-level feature interactions and bridge the semantic gaps across different layers. Inspired by [44], we first employ a self-attention mechanism that enables the model to adaptively learn dependencies among features at different spatial positions, thus allowing it to focus more precisely on key regions of the target object. While self-attention enhances the discriminative representation within each layer, features from different layers remain complementary in their semantic and structural information. To exploit this complementarity, we introduce a cross-attention mechanism that facilitates explicit interactions across hierarchical levels, allowing the model to effectively align and fuse multi-scale features effectively. Through this hierarchical design, the dual attention module captures both fine-grained local details and high-level thus semantic structures, thereby strengthening multi-scale representation learning and improving the classification performance. Figure 4 illustrates the details of the dual attention module.

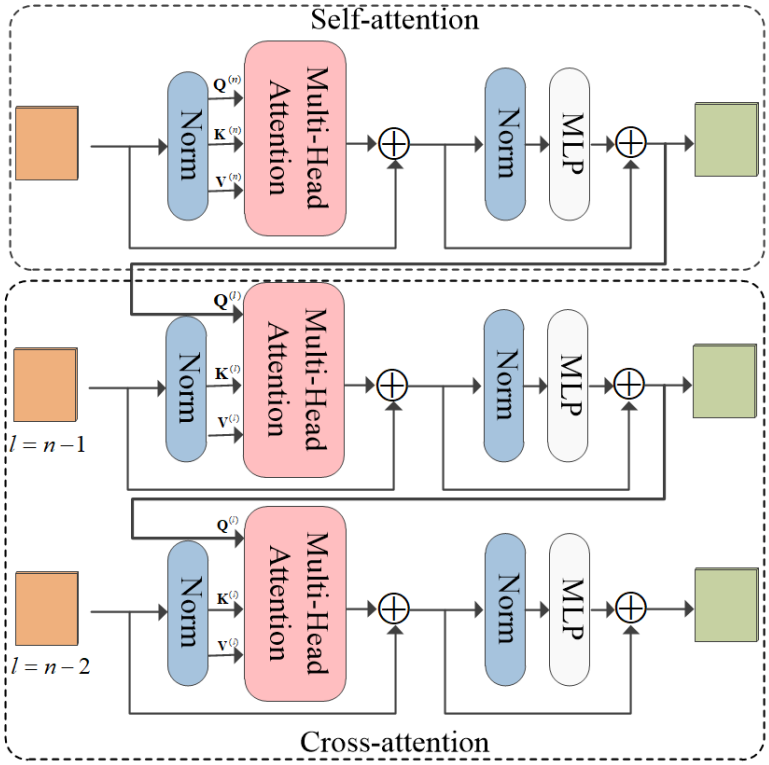

Figure 4. The pipeline of the dual attention module, which mainly consists of self-attention layers and cross-attention mechanisms.

Specifically, the self-attention module follows the standard self-attention mechanism [45], which computes attention maps within the input feature space to capture internal dependencies. The module assigns importance weights to each spatial position by measuring the similarity between that position and all other positions in the feature sequence, thereby allowing the network to emphasize more informative regions.

Concretely, given an input feature sequence $F_p^{(n)}$ processed by a fused attention module at the n-th layer, the query, key, and value projections are obtained via three learnable linear projections:

$$Q^{(n)} = F_p^{(n)} W^Q, \quad K^{(n)} = F_p^{(n)} W^K, \quad V^{(n)} = F_p^{(n)} W^V, \tag{3.7}$$

where $W^Q \in \mathbb{R}^{d \times d_q}$, $W^K \in \mathbb{R}^{d \times d_k}$, and $W^V \in \mathbb{R}^{d \times d_v}$. These projections map the input into distinct representational subspaces so that contextual correlations can be computed and information can be adaptively aggregated adaptively across spatial locations. Therefore, the self-attention output is computed as follows:

$$\text{Attn}_{self} = \text{Attention}(Q^{(n)}, K^{(n)}, V^{(n)}) = \text{softmax}(\frac{Q^{(n)} K^{(n),\text{T}}}{\sqrt{d^k}}) V^{(n)}, \tag{3.8}$$

where $\frac{1}{\sqrt{d^k}}$ is applied to avoid excessively large dot-product values, and T denote the transpose operation. After computing self-attention at each layer, the resulting representation at the n-th layer can serve as queries when computing attention with feature maps from other layers (e.g., lower or higher-resolution feature maps) by modeling attention between embeddings at different depths. For instance, between block embeddings at a layer n and higher-resolution embeddings, the module enables efficient multi-scale information integration.

The cross-attention mechanism further enhances feature representations by explicitly modeling relationships between different feature sources, which highlights target-relevant information while suppressing irrelevant content. While self-attention aggregates information within the same hierarchy by computing similarity weights among elements of a single feature map, hierarchical cross-attention systematically analyzes attention distributions across feature maps of different resolutions to construct a multi-scale attention mechanism.

Specifically, let X_1 denote the output of a self-attention block (serving as the query source), and let $F_p^{(n)}$ denotes the feature map from stage n, which provides keys and values. First, we compute the cross-attention projections as follows:

$$Q_{cross} = X_1 W^Q, \quad K^{(n)} = F_p^{(n)} W^K, \quad V^{(n)} = F_p^{(n)} W^V; \tag{3.9}$$

then, we obtain the cross-attention output as follows:

$$\text{Attn}_{cross} = \text{Attention}(X_1, F_p^{(n)}) = \text{softmax}(\frac{Q_{cross} K^{(n),\text{T}}}{\sqrt{d^k}}) V^{(n)}. \tag{3.10}$$

In our method, we iteratively apply Eq (3.10) to the last three layers of the backbone repeatedly to derive the attention maps that correspond to each layer to obtain $\hat{F}_p^{(1)}$, $\hat{F}_p^{(2)}$, and $\hat{F}_p^{(3)}$.

3.4. Semantic feature enhancement module

We introduce the semantic feature enhancement module to further integrate features from different levels and enable the model to learn more diverse representations. Figure 5 illustrates the details of the semantic feature enhancement module. Specifically, for the last three feature layers processed by the dual attention module, $\hat{F}_p^{(1)} \in \mathbb{R}^{C\times H_1\times W_1}$, $\hat{F}_p^{(2)} \in \mathbb{R}^{C\times H_2\times W_2}$, and $\hat{F}_p^{(3)} \in \mathbb{R}^{C\times H_3\times W_3}$, we reshape their dimensions and convert them into the form of two-dimensional matrices $\hat{F}_p^{(n)'} \in \mathbb{R}^{C\times N_n}$.

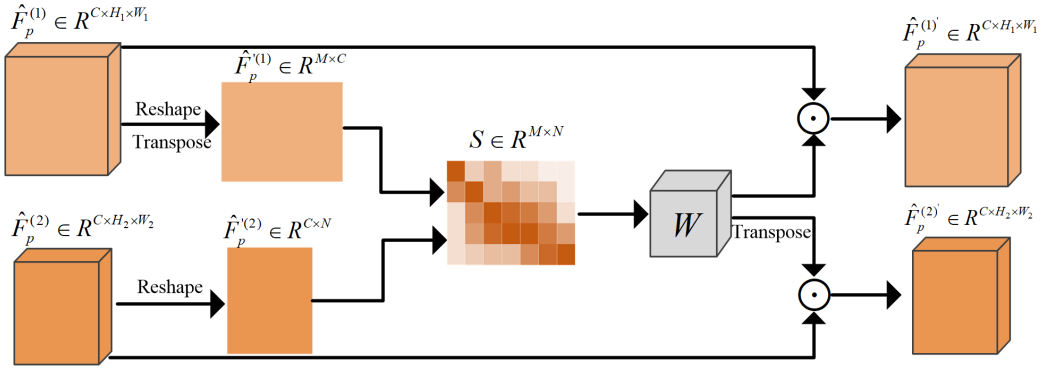

Figure 5. The pipeline of the semantic feature enhancement module. For example, we take different-scale features $\hat{F}_p^{(1)}$ and $\hat{F}_p^{(2)}$.

Subesequently, for any two feature levels i and j, we compute their similarity matrix as follows:

$$S_{ij} = \hat{F}_p^{(i)',\mathrm{T}} \times \hat{F}_p^{(j)'}. \tag{3.11}$$

Thus we can take S_{12}, S_{13}, and S_{23}. To convert similarity into attention weights, we normalize each row of $-S_{ij}$ using softmax as follows:

$$W_{ij} = \mathrm{softmax}(-S_{ij}). \tag{3.12}$$

For level i, the complementary feature contributed by level j is computed as:

$$\hat{F}_p^{(i\leftarrow j)'} = W_{ij}^T \hat{F}_p^{(j)'}. \tag{3.13}$$

The complementary features are fused with the original features to obtain the enhanced semantic feature $\hat{F}_{se}^{(i)}$ as follows;

$$\hat{F}_{se}^{(i)} = \hat{F}_p^{(i)} + \sum_{j\neq i} \hat{F}_p^{(i\leftarrow j)}. \tag{3.14}$$

Subsequently, these features are pooled to produce the final predictions.

3.5. Optimization

In CHM classification tasks, our network generates four prediction outputs from different semantic levels, including the outputs of the (n)-th stage, the (n-1)-th stage, the (n-2)-th stage, and the concatenated multi-scale feature output. For each prediction branch, we employ a classifier followed

by a softmax layer to obtain the confidence scores across all categories. Subsequently, the predicted probability distributions are denoted as $y^{(n)}$, $y^{(n-1)}$, $y^{(n-2)}$, and y^{concat}, respectively. The classification loss is computed as follows:

$$L_{cls} = L_{CE}(y^{(n)}, y) + L_{CE}(y^{(n-1)}, y) + L_{CE}(y^{(n-2)}, y) + L_{CE}(y^{concat}, y). \tag{3.15}$$

4. Experiments

In this section, we will introduce the composition of the datasets, the setting of experimental parameters, the comparison with existing fine-grained classification methods, and the performance improvement after integrating the module proposed in this paper into some of the comparison methods and visualize the results.

4.1. Datasets

To validate the effectiveness of our proposed method, we apply the model to two herbal medicine datasets for CHM classification tasks; Chinese materia medica image dataset [46] and NB-TCM-CHM [47]. Specifically, the NB-TCM-CHM dataset consists of 20 categories with a total of 3784 images, about 90% of the samples are used for training and 10% of the samples for testing. The Chinese materia medica image dataset consists of 95 categories with a total of 5518 images, where about 80% of the samples are used for training and the rest 20% of the samples are used for testing. However, due to the diversity and openness of data sources, the original images within Chinese ateria Medica Image dataset have obvious data quality issues, such as watermarks such as copyright symbols, and some of them cover the main area of the herbal medicine. Moreover, the backgrounds of the images are complex, which causes blurring of the boundary between the target area and the background. These problems pose significant challenges for subsequent classification tasks. Figure 6 shows some images in our database, and the detail composition of the datasets are shown in Table 1.

Table 1. The statistics of two herbal datasets.

Dataset	Classes	Total images	Training images	Test images
Chinese materia medica image [46]	95	5518	4375	1143
NB-TCM-CHM [47]	20	3784	3384	400

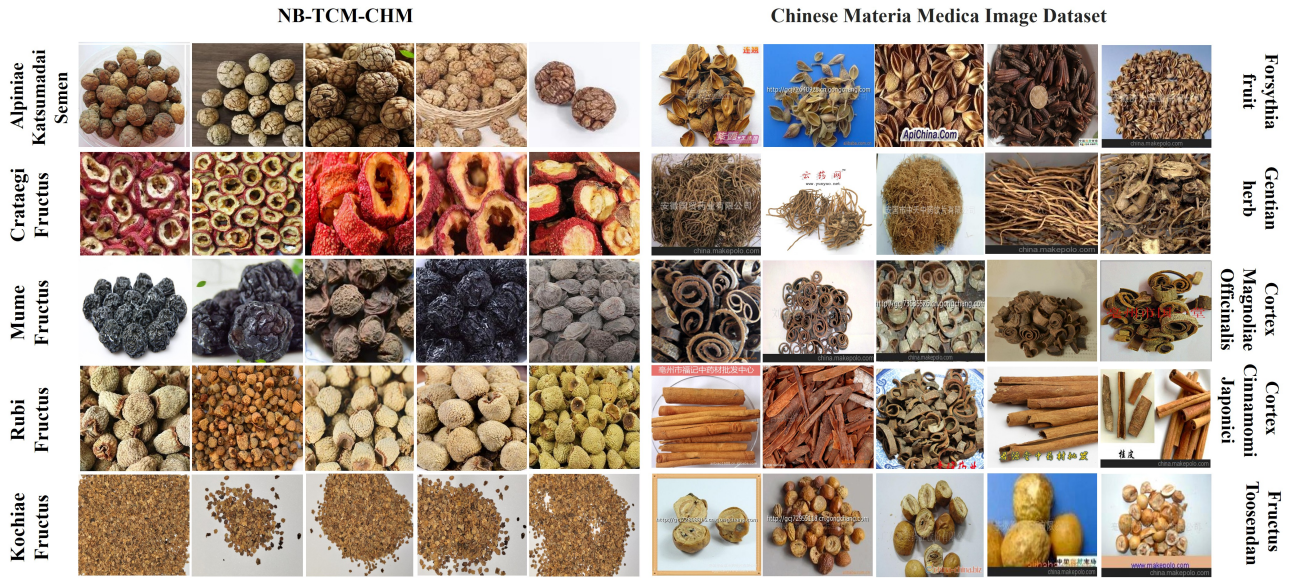

Figure 6. Some samples from two herbal datasets. Images in the left part are from NB-TCM-CHM dataset, and those in right part are from the Chinese Materia Medica Image dataset.

4.2. Evaluation metrics

We utilize accuracy (%) to evaluate the performance of the module, which is a evaluation metric widely adopted in classification tasks. Accuracy is defined as the ratio of correctly predicted samples to the total number of samples, which can straightly reflect the performance of the classification. It effectively reflects the model's ability to make correct predictions and serves as a reliable indicator to assess the effectiveness of our proposed method.

4.3. Implement details

In the experiment, we use ResNet50 as the backbone network and use the SGD optimizer to optimize our model. In the training and testing phases, we resize the input images to 550 * 550 and randomly crop them to 448 * 448. During training, we train for 100 epochs, with the momentum and weight decay set to 0.9 and 1e-5, respectively, and the batch size is set to 10. We set the $\alpha = 1.5$ in Eq (3.1) on two datasets. During training, we perform data augmentation through random cropping and random horizontal flipping. To ensure fairness and verify the effectiveness of our method, we set the parameters to be the same as the original model architecture. The comparation between our method and existing methods is shown in Table 2. Our network is implemented in Pytorch with a version higher than 1.7 over NVIDIA RTX A6000 GPU.

4.4. Quantitative comparison

Table 2. Comparison results of different methods on medicine and NCB-TCM-CHM.

Method	Base model	Medcine (%)	NCB-TCM-CHM (%)
Base CNN	ResNet50	68.2	88.5
Base ViT(2020) [35]	ViT	-	44.1
MobileViT(2021) [48]	ViT	-	**98.0**
FAL-VIT(TCSVT2025) [33]	ViT	78.2	97.8
PIM(2022) [49]	ResNet50	76.0	92.5
PMG(ECCV2020) [50]	ResNet50	76.1	89.6
UniRepLKNet54(CVPR2024) [51]	ConvNet	76.2	96.5
CMAL(PR2023) [52]	ResNet50	76.7	97.2
ACC-FGVC(TVC 2025) [53]	ResNet50	77.8	97.5
CGFF(APPL2025) [54]	ResNet50	77.2	97.0
Ours	ResNet50	**78.5** (**+ 0.7**)	**98.0** (**+ 0.5**)

We conduct a comparative analysis to evaluate the performance of our proposed method against existing approaches. As shown in Table 2, the experimental results demonstrate substantial improvements in accuracy on both the Medicine and NB-TCM-CHM datasets. In particular, we use ResNet50 as a backbone network, our method achieves gains of 0.7% and 0.5% on the two datasets, respectively. Furthermore, our approach consistently outperforms all existing methods across both datasets, thus highlighting its clear advantages in terms of accuracy. Table 3 presents the performance comparison between our model and classical network models in terms of Accuracy, Precision, Recall, and F1-score. As shown in Table 3, we can observe that our method significantly outperforms early CNN models such as AlexNet [17] and VGG16 [55]. In addition, lightweight model's such as MobileNetv2 [56], ShuffleNet [57], and GhostNet [58] significantly reduce the model complexity but at the cost of sacrificing the representational capacity fine-grained discriminative features. Moreover, MobileViT [48] achieves a competing performance compared with our method, since it can effectively employ the global self-attention mechanism to capture long-range feature dependencies.

Table 3. Comparative experimental results of proposed method on NB-TCM-CHM.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
AlexNet [17]	90.3	90.4	90.1	90.2
VGG16 [55]	92.4	92.8	92.3	92.4
MobileNetv2 [56]	85.9	86.1	85.7	85.9
ShuffleNet [57]	97.5	97.5	97.6	97.5
GhostNet [58]	93.8	93.41	93.7	93.9
MobileViT [48]	**98.0**	**98.1**	98.0	98.0
Ours	**98.0**	97.8	**98.2**	**98.1**

4.5. Visualization

To verify the effectiveness of the proposed method, we employ Grad-CAM [59] to visualize the attention maps of our approach and the comparative methods listed in Table 2. For a fair comparison, all methods use ResNet50 as the backbone network. As illustrated in Figures 7 and 8, our method produces more focused and semantically discriminative activation regions than existing approaches, thereby demonstrating its superior ability to localize key areas. To further interpret the experimental results, we apply t-Distributed Stochastic Neighbor Embedding (t-SNE) [60] to visualize the feature distributions of our network outputs. As shown in Figure 7, our method achieves clear and well-separated visualizations on both the NB-TCM-CHM and Chinese Materia Medica Image datasets. In particular, on the NB-TCM-CHM dataset, samples from the same class form compact clusters with distinct boundaries between different categories, further confirming the effectiveness of our method for CHM classification.

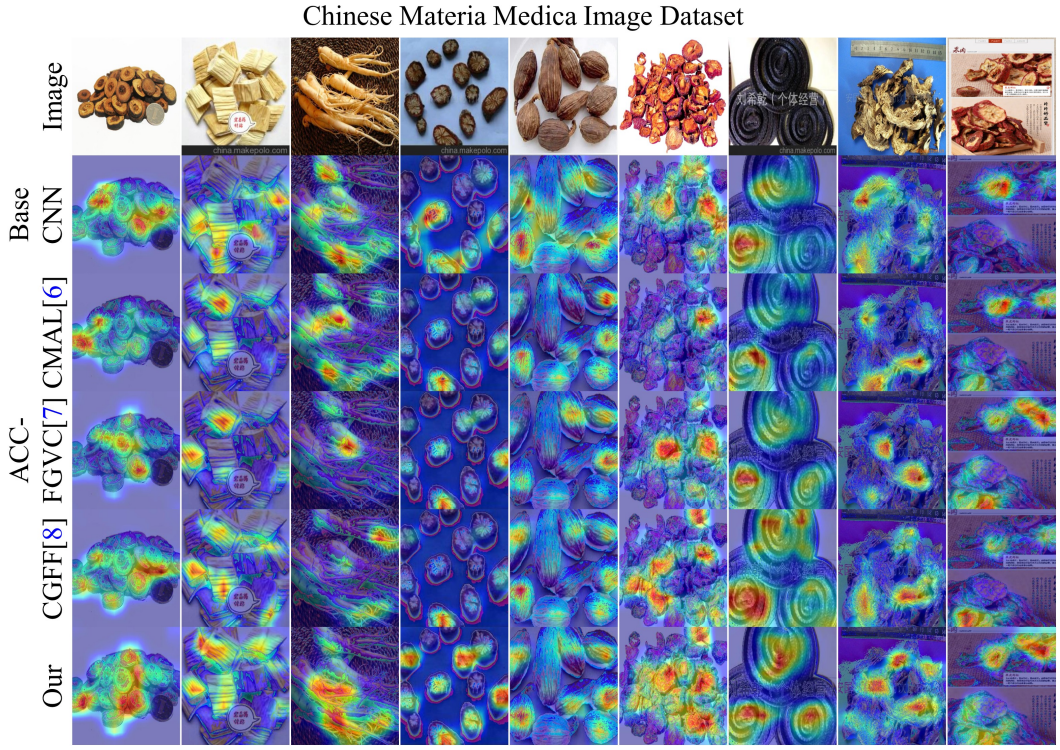

Figure 7. The visualization of Grad-CAM on the Chinese materia medica image dataset.

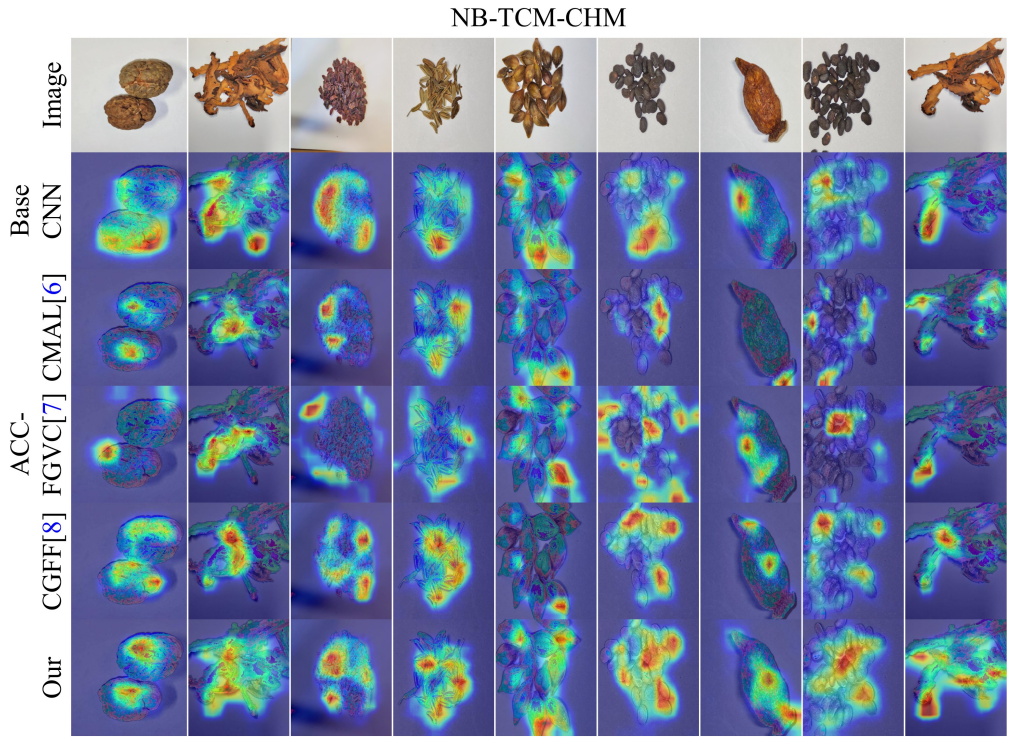

Figure 8. The visualization of Grad-CAM on the NB-TCM-CHM dataset.

4.6. Ablation study

To investigate the contribution of each component in our proposed framework, we conduct ablation experiments by systematically removing or altering key modules and evaluating their impact on the classification performance on the NB-TCM-CHM and Chinese materia medica image datasets. Specifically, the proposed method is comprised of four essential components: an object localization module (OLM), a fused attention module (FAM), a dual attention module (DAM), and a semantic feature enhancement module (SFEM). The results are summarized in Table 4, and the following insights can be drawn.

Table 4. Ablation study on NB-TCM-CHM and Chinese materia medica image datasets.

OLM	FAM	DAM	SFEM	NB-TCM-CHM	Chinese materia medica image
×	✓	✓	✓	78.0	97.5
✓	×	✓	✓	77.5	97.2
✓	✓	×	✓	77.9	97.5
✓	✓	✓	×	78.0	**98.0**
✓	✓	✓	✓	**78.5**	**98.0**

To verify the importance of OLM, we remove the OLM from the whole model, which leads to a noticeable drop in performance (from 78.5% to 78.0% on the NB-TCM-CHM and from 98.0% to 97.5% on the Chinese materia medica image dataset). It demonstrates the critical role of the OLM

in reducing background interference and enhancing the purity of the target region, thereby enabling the model to focus on more discriminative regions.

To evaluate the FAM module's performance, we conduct an ablation study by removing the FAM component. As shown in the second row of Table 4, the FAM enhances the model's ability to capture long-range dependencies and global context information by encoding directional attention across spatial dimensions.

To validate the effectiveness of the DAM, we remove the DAM from our network. As shown in the third row of Table 4, its absence results in the largest overall accuracy drop(resulted in a decrease of 0.6% on the NB-TCM-CHM dataset and 0.5% on the Chinese materia medica image dataset, respectively). It demonstrates that DAM leverages both self-attention and hierarchical cross-attention to fuse multi-scale features, especially in complex background scenarios, thus highlighting its key role in extracting and integrating cross-scale information.

To validate the effectiveness of the SFEM, we remove the SFEM from the whole model. As shown in the fourth row of Table 4, its absence caused a performance drop of 0.5% on both the NB-TCM-CHM dataset (from 78.5% to 78.0%) and the Chinese materia medica image dataset (from 98.0% to 97.5%). This result highlights the role of the SFEM in enhancing inter-layer feature interactions, thus allowing the model to better capture complementary semantics across different layers.

We added a pure ResNet50 baseline named 'Base' as the fundamental network for comparison, which only employs the ResNet50 backbone, where the features from the last three stages are extracted and solely optimized solely with the classification loss. Subsequently, we provide a structured and diagnostic ablation study in Table 5, which shows the accuracy results of the ablation experiments for each component. From Table 5, we can observe that the OLM contributes the most to performance improvement, especially on the NB-TCM-CHM dataset. It demonstrates that the OLM alleviates the impact of background information while focusing on discriminative regional features. In addition, the FAM also achieves a promising performance improvement on the two CHM datasets. This demonstrates that FAM integrates horizontal and vertical directional information across adjacent layers to capture long-range spatial dependencies and enhance the global contextual perception. Moreover, the DAM can further boost the performance across all datasets, which achieves an explicit fusion and semantic alignment of multi-scale features. Finally, the SFEM yields additional performance gains across all datasets, which demonstrates that the SFEM strengthens inter-layer complementarity via adaptive semantic fusion, thus further improving feature discrimination and robustness.

Table 5. More ablation study on NB-TCM-CHM and Chinese materia medica image datasets.

Base	OLM	FAM	DAM	SFEM	NB-TCM-CHM (%)	Chinese materia medica image (%)
✓					70.3	94.3
✓	✓				75.6	95.2
✓	✓	✓			77.0	96.9
✓	✓	✓	✓		78.0	**98.0**
✓	✓	✓	✓	✓	**78.5**	**98.0**

We conducted ablation experiments on the Chinese materia medica image and NCB-TCM-CHM

datasets to evaluate the impact of different ResNet50 stages. Since the proposed DAM and SEFM modules are designed to explore cross-stage feature interactions, we removed both modules when only one stage (S = 1) was used. As shown in Table 6, the model achieves the best performance when the last three stages (S = 3) are selected. The reasons are as follows: 1) the shallower stages produce high-resolution but low-semantic features, primarily capturing low-level information such as edges and textures, which contribute little to the discriminative semantic information required for fine-grained classification; 2) the deeper stage yield strong semantic features but suffer from excessively low spatial, which may lose the fine-grained localization details necessary for accurate recognition; and 3) the last three stages strike a favorable balance between semantic abstraction and spatial resolution, thus providing both medium-level semantic information and sufficient spatial details. Further increasing the number of stages would introduce substantially higher memory overhead, which is unsuitable for large-scale fine-grained classification tasks.

Table 6. Effect of different stages of ResNet50 on the Chinese materia medica image and NCB-TCM-CHM datasets.

S	Chinese materia medica image (%)	NCB-TCM-CHM (%)
1	75.5	95.6
2	78.0	97.5
3	**78.5**	**98.0**

We added a random-seed experiments to evaluate the robustness of the proposed method. As shown in Table 7, we repeat the training and testing process five times using five different random seeds and report the mean ± standard deviation of accuracy on the NCB-TCM-CHM dataset. From Table 7, we can observe that the variation in the performance across different random seeds is small, which verifies the robustness of our proposed method.

Table 7. Random seed results (mean ± std).

Folds	Chinese materia medica image (%)	NCB-TCM-CHM (%)
1	78.5	98.0
2	78.6	98.0
3	78.3	98.1
4	78.3	97.7
5	78.4	97.9
Mean ± std	**78.4 ± 0.2**	**97.9 ± 0.1**

4.7. Efficiency study

As shown in the Table 8, we compare our method with PIM, PMG, and CMAL interm of parameters and FLOPs on NB-TCM-CHM. From Table 8, we can observe that the proposed method increases FLOPs and the number of parameters due to the introduction of multiple attention-related modules. Nevertheless, our method outperforms other comparison methods in terms of the classification performance by a large margin, which demonstrates a favorable trade-off between effectiveness and efficiency.

Table 8. Efficiency study on NB-TCM-CHM dataset.

Method	Parameters (M)	Flops (G)	NB-TCM-CHM (%)
PIM [49]	59.93	55.24	92.5
PMG [50]	45.92	37.43	89.6
CMAL [52]	42.87	36.42	97.2
Ours	101.63	113.23	**98.0**

5. Conclusions

In this paper, we proposed a dual attention enhancement network for CHM classification. To address challenges caused by background interference and the local bias of CNNs, we designed a comprehensive framework the consists of four key modules. The object localization module effectively suppresses background noise and guides the network to focus on discriminative target regions. The fused attention module captures long-range spatial dependencies and enriches global contextual information, while the dual attention module achieves explicit multi-scale feature fusion through self-attention and cross-attention mechanisms. Additionally, the semantic feature enhancement module strengthens the inter-layer complementarity via adaptive semantic fusion, thus further improving feature discrimination and robustness. Extensive experiments on two challenging herbal image datasets—Chinese materia medica image and NB-TCM-CHM-demonstrates that our method consistently outperforms existing approaches. The results confirm that our framework effectively enhances the feature representation and classification accuracy for CHM,thus providing a promising solution for intelligent TCM image analyses.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 62171329.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. W. Peng, R. Lauche, C. Ferguson, J. Frawley, J. Adams, D. Sibbritt, Efficacy of Chinese herbal medicine for stroke modifiable risk factors: A systematic review, *Chin. Med.*, **12** (2017), 25–54. https://doi.org/10.1186/s13020-017-0146-9

2. S. Hu, D. Luo, Q. Zhu, J. Pan, B. Chen, M. Furian, et al., An updated meta-analysis of Chinese herbal medicine for the prevention of COVID-19 based on western-eastern medicine, *Front. Pharmacol.*, **14** (2023), 1–25. https://doi.org/10.3389/fphar.2023.1257345

3. Y. Yang, G. Xiao, P. Cheng, J. Zeng, Y. Liu, Protective application of Chinese herbal compounds and formulae in intestinal inflammation in humans and animals, *Molecules*, **28** (2023), 6811. https://doi.org/10.3390/molecules28196811

4. Y. Xi, X. Lu, L. Zhu, X. Sun, Y. Jiang, W. He, et al., Clinical trial for conventional medicine integrated with traditional Chinese medicine (TCM) in the treatment of patients with chronic kidney disease, *Medicine*, **99** (2020), 20234–20240. https://doi.org/10.1097/md.0000000000020234

5. Y. Gao, Z. Li, Y. Wang, H. Zhang, K. Huang, Y. Fu, et al., Analysis of clinical evidence on traditional Chinese medicine for the treatment of diabetic nephropathy: A comprehensive review with evidence mapping, *Front. Endocrinol.*, **15** (2024), 1324782. https://doi.org/10.3389/fendo.2024.1324782

6. Q. Hou, W. Yang, G. Liu, Chinese herbal medicine recognition network based on knowledge distillation and cross-attention, *Sci. Rep.*, **15** (2025), 1687–1702. https://doi.org/10.1038/s41598-025-85697-6

7. X. Sun, H. Qian, Chinese herbal medicine image recognition and retrieval by convolutional neural network, *PLoS One*, **11** (2016), 0156327. https://doi.org/10.1371/journal.pone.0156327

8. S. S. Chang, H. J. Huang, C. Y. C. Chen, Two birds with one stone? Possible dual-targeting h1n1 inhibitors from traditional Chinese medicine, *PLoS Comput. Biol.*, **7** (2011), 1002315. https://doi.org/10.1371/journal.pcbi.1002315

9. N. L. Zhang, S. Yuan, T. Chen, Y. Wang, Hierarchical latent class models and statistical foundation for traditional Chinese medicine, in *Conference on Artificial Intelligence in Medicine in Europe*, (2007), 139–143. https://doi.org/10.1007/978-3-540-73599-1_15

10. J. M. Yang, P. T. Yu, B. C. Kuo, A nonparametric feature extraction and its application to nearest neighbor classification for hyperspectral image data, *IEEE Trans. Geosci. Remote Sens.*, **48** (2010), 1279–1293. https://doi.org/10.1109/tgrs.2009.2031812

11. P. Rebentrost, M. Mohseni, S. Lloyd, Quantum support vector machine for big data classification, *Phys. Rev. Lett.*, **113** (2014), 130503. https://doi.org/10.1103/physrevlett.113.130503

12. D. M. Farid, L. Zhang, C. M. Rahman, M. A. Hossain, R. Strachan, Hybrid decision tree and naïve bayes classifiers for multi-class classification tasks, *Expert Syst. Appl.*, **41** (2014), 1937–1946. https://doi.org/10.1016/j.eswa.2013.08.089

13. Y. Chen, S. Wang, Y. Ge, A survey on the applications of image classification based on convolution neural network, in *2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, (2022), 381–384. https://doi.org/10.1109/ipec54454.2022.9777354

14. C. Mallah, J. Cope, J. Orwell, Plant leaf classification using probabilistic integration of shape, texture and margin features, *Signal Process. Pattern Recognit. Appl.*, **5** (2013), 45–54. https://doi.org/10.2316/p.2013.798-098

15. D. Luo, J. Wang, Y. Chen, G. Hamid, Classification of Chinese herbal medicines based on svm, in *2014 International Conference on Information Science, Electronics and Electrical Engineering*, (2014), 453–456. https://doi.org/10.1109/infoseee.2014.6948152

16. Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature*, **521** (2015), 436–444. https://doi.org/10.1038/nature14539

17. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Commun. ACM*, **60** (2017), 84–90. https://doi.org/10.1145/3065386

18. L. Mu, Z. Gao, Y. Cui, K. Li, H. Liu, L. Fu, Kiwifruit detection of far-view and occluded fruit based on improved alexnet (in Chinese), *Trans. Chin. Soc. Agric. Mach.*, **50** (2019), 24–34. https://doi.org/10.6041/j.issn.1000-1298.2019.10.003

19. Y. Chen, C. Gong, Y. Liu, X. Fang, Fish identification method based on ftvgg16 convolutional neural network (in Chinese), *Trans. Chin. Soc. Agric. Mach.*, **50** (2019), 223–231. https://doi.org/10.6041/j.issn.1000-1298.2019.05.026

20. S. Liu, W. Chen, X. Dong, Automatic classification of Chinese herbal based on deep learning method, in *14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, (2018), 235–238. https://doi.org/10.1109/fskd.2018.8687165

21. M. Han, J. Zhang, Y. Zeng, F. Hao, Y. Ren, A novel method of Chinese herbal medicine classification based on mutual learning, *Mathematics*, **10** (2022), 1557–1566. https://doi.org/10.3390/math10091557

22. Y. Kang, M. Wu, Y. Gao, L. Fu, C. Li, Z. Song, Deep learning-based classification of traditional Chinese medicine: A novel approach, *Quant. Imaging Med. Surg.*, **15** (2025), 7483–7496. https://doi.org/10.21037/qims-24-1354

23. C. Cai, S. Liu, L. Wang, B. Yang, M. Zhi, R. Wang, et al., Classification of Chinese herbal medicine using combination of broad learning system and convolutional neural network, in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, (2019), 3907–3912. https://doi.org/10.1109/smc.2019.8914437

24. Q. Liu, X. P. Liu, L. J. Zhang, L. M. Zhao, Image texture feature extraction & recognition of Chinese herbal medicine based on gray level co-occurrence matrix, *Adv. Mater. Res.*, **605** (2013), 2240–2244. https://doi.org/10.4028/www.scientific.net/amr.605-607.2240

25. Z. Tang, H. Yang, C. Chen, Weakly supervised posture mining for fine grained classification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 23735–23744. https://doi.org/10.1109/cvpr52729.2023.02273

26. Z. Wang, T. Li, Y. Qin, B. Sheng, Contrastive decoupling: Dynamic regularization for enhanced fine-grained image classification, *IEEE Trans. Circuits Syst. Video Technol.*, (2026), 1–15. https://doi.org/10.1109/tcsvt.2026.3661760

27. A. Bera, Z. Wharton, Y. Liu, N. Bessis, A. Behera, Sr-gnn: Spatial relation aware graph neural network for fine-grained image categorization, *IEEE Trans. Image Process.*, **31** (2022), 6017–6031. https://doi.org/10.1109/tip.2022.3205215

28. H. Yang, H. He, Y. Li, F. Cao, Revisiting fine-grained classification: A dual branch method for noise-resilient and global-local discriminative learning, *Expert Syst. Appl.*, **295** (2026), 128802. https://doi.org/10.1016/j.eswa.2025.128802
29. J. He, J. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, et al., Transfg: A transformer architecture for fine-grained recognition, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **36** (2022), 852–860. https://doi.org/10.1609/aaai.v36i1.19967
30. Q. Wang, J. Wang, H. Deng, X. Wu, Y. Wang, G. Hao, Aa-trans: Core attention aggregating transformer with information entropy selector for fine-grained visual classification, *Pattern Recognit.*, **140** (2023), 109547. https://doi.org/10.1016/j.patcog.2023.109547
31. J. Wang, Q. Xu, B. Jiang, B. Luo, J. Tang, Multi-granularity part sampling attention for fine-grained visual classification, *IEEE Trans. Image Process.*, **33** (2024), 4529–4542. https://doi.org/10.1109/tip.2024.3441813
32. X. Jiang, H. Tang, J. Gao, X. Du, S. He, Z. Li, Delving into multimodal prompting for fine-grained visual classification, in *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, **38** (2024), 2570–2578. https://doi.org/10.1609/aaai.v38i3.28034
33. Y. Huang, Z. Hechen, M. Zhou, Z. Li, S. Kwong, An attention-locating algorithm for eliminating background effects in fine-grained visual classification, *IEEE Trans. Circuits Syst. Video Technol.*, **35** (2025), 5993–6006. https://doi.org/10.1109/tcsvt.2025.3535818
34. S. Yang, J. Wen, B. Fang, Fg-moe: Heterogeneous mixture of experts model for fine-grained visual classification, *Pattern Recognit.*, **175** (2026), 113050. https://doi.org/10.1016/j.patcog.2026.113050
35. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929.
36. T. Li, F. Sun, R. Sun, L. Wang, M. Li, H. Yang, Chinese herbal medicine classification using convolutional neural network with multiscale images and data augmentation, in *2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, (2018), 109–113. https://doi.org/10.1109/spac46244.2018.8965566
37. J. Hu, Y. Wang, Z. Che, Q. Li, H. Jiang, L. Liu, Image recognition of Chinese herbal pieces based on multi-task learning model, in *Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (2020), 1555–1559. https://doi.org/10.1109/bibm49941.2020.9313412
38. C. Xing, Y. Huo, X. Huang, C. Lu, Y. Liang, A. Wang, Research on image recognition technology of traditional Chinese medicine based on deep transfer learning, in *2020 International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)*, (2020), 140–146. https://doi.org/10.1109/aiea51086.2020.00037
39. Y. Chen, L. Zou, Intelligent screening of pieces of Chinese medicine based on bmfnet-wgan (in Chinese), *Chin. J. Exp. Tradit. Med. Formulae*, **27** (2021), 107–114. https://doi.org/10.13422/j.cnki.syfjx.20210819
40. S. Liu, W. Chen, Z. Li, X. Dong, Chinese herbal classification based on image segmentation and deep learning methods, in *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery*, Springer, **89** (2022), 267–275. https://doi.org/10.1007/978-3-030-89698-0_28

41. J. Miao, Y. Huang, Z. Wang, Z. Wu, J. Lv, Image recognition of traditional Chinese medicine based on deep learning, *Front. Bioeng. Biotechnol.*, **11** (2023), 1199803. https://doi.org/10.3389/fbioe.2023.1199803

42. M. Huang, Y. Xu, Image classification of Chinese medicinal flowers based on convolutional neural network, *Math. Biosci. Eng.*, **20** (2023), 14978–14994. https://doi.org/10.3934/mbe.2023671

43. Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 13713–13722. https://doi.org/10.1109/cvpr46437.2021.01350

44. M. Ghahremani, M. Khateri, B. Jian, B. Wiestler, E. Adeli, C. Wachinger, H-vit: A hierarchical vision transformer for deformable image registration, in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2024), 11513–11523. https://doi.org/10.1109/cvpr52733.2024.01094

45. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, *Adv. Neural Inf. Process. Syst.*, **30** (2017), 5998–6008.

46. Figshare, 2025. https://doi.org/10.6084/m9.figshare.29432726

47. D. Tian, C. Zhou, Y. Wang, R. Zhang, Y. Yao, Nb-tcm-chm: Image dataset of the Chinese herbal medicine fruits and its application in classification through deep learning, *Data Brief*, **54** (2024), 110405. https://doi.org/10.1016/j.dib.2024.110405

48. S. Mehta, M. Rastegari, Mobilevit: Light-weight, general-purpose, and mobile friendly vision transformer, *J. Phys. Conf. Ser.*, **2562** (2023), 012012. https://doi.org/10.1088/1742-6596/2562/1/012012

49. P. Y. Chou, C. H. Lin, W. C. Kao, A novel plug-in module for fine-grained visual classification, preprint, arXiv:2202.03822.

50. R. Du, D. Chang, A. K. Bhunia, J. Xie, Z. Ma, Y. Z. Song, et al., Fine grained visual classification via progressive multi-granularity training of jigsaw patches, in *Lecture Notes in Computer Science*, Springer, **12370** (2020), 153–168. https://doi.org/10.1007/978-3-030-58565-5_10

51. X. Ding, Y. Zhang, Y. Ge, S. Zhao, L. Song, X. Yue, et al., Unireplknet: A universal perception large-kernel convnet for audio video point cloud time series and image recognition, in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2024), 5513–5524. https://doi.org/10.1109/cvpr52733.2024.00527

52. D. Liu, L. Zhao, Y. Wang, J. Kato, Learn from each other to classify better: Cross-layer mutual attention learning for fine-grained visual classification, *Pattern Recognit.*, **140** (2023), 109550. https://doi.org/10.1016/j.patcog.2023.109550

53. D. Yu, Z. Fang, Y. Jiang, Alleviating category confusion in fine-grained visual classification, *Visual Comput.*, **41** (2025), 7417–7432. https://doi.org/10.1007/s00371-025-03814-y

54. S. Wu, J. Hu, C. Sun, F. Zhong, Q. Zhang, G. Wang, A cross-granularity feature fusion method for fine-grained image recognition, *Appl. Intell.*, **55** (2025), 1–19. https://doi.org/10.1007/s10489-024-05891-3

55. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.

56. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 4510–4520. https://doi.org/10.1109/cvpr.2018.00474

57. N. Ma, X. Zhang, H. T. Zheng, J. Sun, Shufflenet v2: Practical guidelines for efficient cnn architecture design, in *Lecture Notes in Computer Science*, Springer, (2018), 122–138. https://doi.org/10.1007/978-3-030-01264-9_8

58. K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, Ghostnet: More features from cheap operations, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 1577–1586. https://doi.org/10.1109/cvpr42600.2020.00165

59. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 618–626. https://doi.org/10.1109/iccv.2017.74

60. L. van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.*, **9** (2008), 2579–2605.

AIMS Press