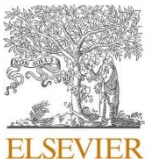
Arab Economic and Business Journal 0000 (2014) 000–000



Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/aebj

Pest Identification Model Based on Multiscale CNN and ViT

Naihao Xu[a], Hui Deng[b,*], Meijun Sun[b], and Zhiliang qin[c]

[a] *College of ..*

[b] *College of Intelligence and Computing, Tianjin University, Tianjin 300354, China*

[c] *Weihai Beiyang Electric Group Co., Ltd., Weihai 264209, China*

ARTICLEINFO

Keywords:
Cross-modal fusion
Contextual transformer
Pyramid squeeze attention mechanism
Convolutional neural network and bi-directional long short-term memory
Pest recognition

ABSTRACT

As one of the necessary cash crops in China and many other countries, wolfberry is parasitized by multiple pests, and its yield is highly susceptible to being affected. On the other hand, agricultural pest backgrounds are complex. When identifying them, single-modal models cannot utilize diverse data types across modalities, resulting in low identification accuracy and data utilization. Traditional unimodal identification models can no longer meet the needs of multimodal data development in agriculture. To overcome these challenges, the ITF-WPI cross-modal feature fusion model is proposed, which consists of CoTN and ODLS for parallel processing of images and text, respectively. We incorporate the Transformer structure (CoT), which focuses on contextual feature extraction, into CoTN to make full use of the rich static and dynamic linear fusion contexts between adjacent keys and improve the 4-stage network of CoTN using Pyramid Squeezed Attention (PSA) to improve the extraction of multi-scale feature structure information and effectively promote the interaction of in-depth features with multi-scale spatial information. The ODLS network constructed by introducing 1D convolutional and bidirectional LSTM stacking has been shown to have more robust text feature acquisition than other advanced convolutional neural network-long short-term memory (CNN-LSTM) models from experimental results, with a 30% reduction in MACCs compared to the optimal model. The results showed that ITF-WPI performed well in accuracy, F1 score, model size, and MACCs with 97.98%, 93.19%, 52.20 MB, and 7.828G compared to the classical state-of-the-art (SOTA) model, lightweight SOTA model and advanced Transformer neural network synthesis, respectively. The model has critical practical applications for promoting the development of cross-modal models in agriculture and research on wolfberry pest control and improving wolfberry yields.

1. Introduction

As a critical herbal medicine, wolfberry (Lycium barbarum L.) is widely used in traditional chinese medicine (TCM) clinical treatment and dietary therapy for its flavonoids, polysaccharides, carotenoids, phenolic compounds, and other active ingredients (Vidovićet al., 2022; Yang et al., 2022), which are known to tonic the liver and kidney, clear heat and brighten the eyes, anti-aging, anti-fatigue, anti-tumor, hypo-lipidemic, tranquilizing, and regulating the immune system. Besides, it can be used as a tea or cooked and consumed with other foods (Toh et al., 2021; Wenli et al., 2021). The wolfberry produced in Ningxia, China, has been favored by domestic and international markets due to its large fruit, beautiful shape, and high content of active ingredients. Lycium barbarum is resistant to salinity, sandy wasteland, and drought. It can grow in deep soil bank ditches, mountain slopes, and field ridges, often used for soil and water conservation and afforestation. In recent years, with the gradual expansion of planting area, its ecological and economic value began to be highlighted. Now it has become one of the critical economic crops in Ningxia and even the whole northwest arid zone of China (Yajun et al., 2019). However, Wolfberry is a multi-insect host with poor pest resistance, making it susceptible to pest infestation, which can profoundly impact yield and quality and lead to severe

* Corresponding author at: Tianjin Key Laboratory of Machine Learning, Tianjin University, 300192, Tianjin, China.
E-mail addresses: naihao0117@gmail.com (N. Xu), huidenng@126.com(H. Deng), meijunsun@sina.cn (M. Sun), beiyang0611@163.com (Z. Qin).

economic losses. Therefore, rapid and accurate identification of the multifaceted information of wolfberry pests and timely adoption of precise control measures to reduce the use of pesticides is essential to avoid the further spread of pests, improve the yield and quality of Wolfberry, and promote the development of the wolfberry industry.

Traditional pest identification methods are mainly observed and identified with the help of agricultural experts or technicians, which is time-consuming, costly, time-sensitive, and challenging to use widely (Dai et al., 2023; Ye et al., 2023). With the development of precision agriculture and smart agriculture, the successful application of deep learning in agriculture, which is currently a cutting-edge, modern, and promising technology, has gained more attention (Thakur et al., 2022). For example, Zhou et al. (2023) collected 805 images of four types ofrice leaf diseases. They proposed a residual distillation transformer architecture for rice leaf disease identification using pre-trained visual transformers and distillation transformers as residual tandem blocks and prediction with a multilayer perceptron (MLP) for more robust feature characterization, achieving a Top-1 accuracy of 92% for rice leaf disease identification. Huang et al. (2022) applied a convolutional neural network (CNN) with migration learning to extract tomato pest features. They classified the extracted pest features using three machine learning classifiers with hyperparameters optimized by Bayesian, achieving excellent classification performance on eight pest datasets with 97.12% classification accuracy. Coulibaly et al. (2022) used Inceptionv3 as the backbone for feature extraction, improved Inceptionv3 by migratory learning with replacement classification layers, and used explainable artificial intelligence (XAI) to develop clear and transparent rules for pest identification, with 67.88% classification accuracy on the IP102 pest dataset. Nigam et al. (2023) proposed a wheat disease recognition model based on fine-tuned EfficientNet architecture with 99.35% accuracy for wheat stripe, leaf, and stem rust. Bao et al. (2022) took cotton aphid infestation images in the natural environment, classified the severity of aphid infestation in cotton into four levels, and established a cotton disease image dataset based on the DenseNet network structure, embedded Coordinate attention (CA) mechanism into the feature extraction structure, and proposed CA_DenseNet_BC_40 Lightweight network model with 97.3% accuracy in classifying the degree of damage caused by cotton aphids under natural field conditions. Yu et al. (2023) introduced the Transformer structure in the convolutional architecture and designed the Inception Convolutional Vision Transformer (ICVT) for various plant disease recognition. Dynamic pattern decomposition (DMD) based on color channel information can extract different degrees of feature distinctness for each category of rice leaf images to simulate human attention (visually salient regions). In-depth features of preprocessed images using DMD can be implemented to classify rice leaf diseases. DMD has achieved better results in migration learning deep CNN (DCNN) classifiers and machine learning (ML) models (Sudhesh et al., 2023). Chodey & Noorullah Shariff (2023) proposed the use of manual feature extraction combined with sequence processing neural network for cotton bollworm and rice bug detection; images were segmented using fuzzy C-mean segmentation; manual feature extraction was performed using gray level co-occurrence matrix (GLCM), Sobel operator and multichannel feature extraction mixed collection; long short-term memory (LSTM) network and recurrent neural network (RNN) combined with hybrid classifier constitutes the sequence processing neural network.

Agricultural pest image recognition is a challenging problem, mainly due to the complex and variable background environment and the relatively small number of available pest image samples. Current research on insect pest identification involved in plant pest identification is relatively limited, and most of the work builds on plant leaf diseases with less focus on insect pests. With the diversification of agricultural information, multiple types of modal information have emerged (Yang et al., 2021). The differences in the comprehensiveness of describing semantic information due to the different objects described by various types of modal information and the difficulty of extracting all features from current neural networks have led to the absence of a single modal information description or the lack of semantic associations (Wang et al., 2021). Cross-modal feature fusion can solve the interaction between different modal information and utilize multimodal information in agriculture in a more diversified way. Therefore, introducing cross-modal feature fusion into agriculture to achieve cross-modal identification of pests and diseases can better meet the practical needs of agricultural development. However, cross-modal feature fusion is rarely reported in agriculture and deserves in-depth study.

To address the above problems, this study takes 17 types of wolfberry pest images and corresponding description texts as the research objects. It introduces cross-modal feature fusion technology into wolfberry pest recognition research for the problem of single utilization of modal information by existing methods. It proposes a cross-modal feature fusion recognition model of images and texts named ITF-WPI. ITF-WPI consists of an image encoder (CoTN) and a text encoder (ODLS), which process image and cross-modal text information in parallel, respectively; CoTN uses the Contextual Transformer (CoT) and Pyramid Squeezed Attention (PSA) mechanisms for feature extraction of image data, and ODLS uses CNN and LSTM networks built with memory capabilities that enables it to process serialized data to extract critical information. Pattern learning of image and textsalience semantic information becomes accessible. The main contributions of this study are summarized as follows:

(1) The ITF-WPI model aims to identify wolfberry pests by image and text cross-modal feature fusion techniques, which combines CoTN as an image feature extraction structure with a new neural network architecture constructed by image description text feature extraction ODLS.

(2) The CoT structure with Transformer style is adopted by the four stages of the CoTN network, which elegantly combines context mining and self-attentive learning, thus enhancing visual representation.

(3) The PSA attention mechanism of CoTN embedding can effectively extract multi-scale spatial information of wolfberry pests at a finer-grained level, enabling the model to focus on the necessary fine-grained parts of the image to obtain rich features and improve the efficiency of feature extraction.

(4) The ODLS network constructed by stacking 1D convolution and bidirectional LSTM can extract pest description text information effectively. The subsequent Dropout layer of bidirectional LSTM reduces the overfitting phenomenon and replaces the random initialization word embedding layer with Word2Vec to reduce the model training time consumption.

(5) A GradCAM explainable artificial intelligence (XAI) visualization method was used to explain whether the ITF-WPI model correctly focused on the wolfberry pest characteristics or pattern information.

The rest of this study is organized as follows: **Section 2** describes the WPIT9K dataset for wolfberry pests in detail, followed by the principles and workflow of the proposed ITF-WPI image and text cross-modal feature fusion model. **Section 3** presents the experimental working conditions, analyzes the results of ITF-WPI pest recognition, ablation validates the commonly used attention mechanisms, and compares the performance of CoTN, ODLS, and other SOTA models in wolfberry pest recognition, respectively. **Section 4** concludes the paper and discusses future work.

2. Materials and methods

2.1. Wolfberry pest datasets

The Wolfberry Pest Image and Text Cross-Modal Dataset (WPIT9K) is a study of common Wolfberry pests in Ningxia, China (105°37′21″E, 37°28′59″N). Data collection was differentiated according to pest categories. The primary data was collected from the Internet and constructed image data subsets through field research, photography, and book scanning. The textual data subset was obtained by reviewing relevant professional books, consulting Generative Artificial Intelligence (GAI), web searches, professional pest databases, and consulting experts to

obtain descriptive information for each category of wolfberry pests and prepare corresponding textual descriptions. The specific wolfberry pest images and text samples for each category are shown in Fig. 1. The text data subset contains information on the scientific name profile, source distribution, habitat, and prevention methods of each type of pest. The original sample size collected from the image and text data subset is 2235.

Because there are too few learning samples in the image and text cross-modal data sets of wolfberry pests, overfitting is prone to occur in complex networks. The data augmentation technology is used to expand the original samples for image and text data simultaneously to ensure the consistency of input and improve the model 's generalization ability. For the image data subset, 4 types of data augmentation were performed using random horizontal-vertical flip, random brightness adjustment, random crop, and random shadow transformation on the original images, and similar images were filtered out using the structural similarity (SSIM) algorithm to obtain a total of 8363 enhanced image data. For the subset of text data, each combined with text descriptions was randomly inserted, synonym substituted, randomly deleted, and randomly swapped to obtain 4 types of enhanced text data totaling 8363 items, and together with the original data samples, the image data and its corresponding text description samples were 10,598 items each. The pest images and text descriptions were divided into 17 categories each. Each category corresponds to its corresponding wolfberry pest, constituting a cross-modal dataset of wolfberry pest images and text based on the Wikipedia data structure (Table 1) called WPIT9K. In addition, the wolfberry pest images of the constructed data and the text have consistency. The preprocessing makes the pest images and texts appear in pairs before the input model, which significantly facilitates the cross-modal model to extract and encode the features, and the image and text corresponding relations in the part of the dataset are shown in Fig. 2. Then, the subsequent modeling process was divided into the training set, validation set, and test set in the ratio of 7:2:1.

Table 1
Description of the image and text dataset (WPIT9K) for 17 types of common wolfberry pests.

Latin names of wolfberry pests	Dataset serial number	Image sample volume	Text sample volume	Text word count range
Geometridae	0	999	999	24 – 142
Cicadella viridis	1	1015	1015	50–124
Crioceridae	2	615	615	41–98
Elthemidea sp	3	478	478	43–122
Membracidae	4	624	624	34–128
Mylabris speciosa Pallas	5	384	384	55–143
Tropidothorax elegans distant	6	874	874	46–103
Cerambycidae	7	806	806	28–115
Nephrotoma sp	8	579	579	31–105
Thripidae	9	916	916	51–433
Epitri abeillei	10	578	578	49–104
Bedbug	11	430	430	45–101
Tephritidae	12	388	388	46–108
Agrotis ypsilon	13	484	484	61–141
Adelgoidea	14	368	368	60–118
Plodia interpunctella	15	380	380	45–103
Carposinidae	16	680	680	34–114

2.2. Proposed ITF-WPI architecture

In this study, an image and text cross-modal feature fusion model named ITF-WPI is proposed to address the identification and classification of wolfberry pests in complex agricultural environments to promote the research development of cross-modal models in agriculture and wolfberry pest control while reducing pesticide usage rates to protect agroecosystems. The proposed ITF-WPI model consists of two main building blocks: the image encoder CoTN and the text encoder ODLS. For CoTN, it is responsible for processing the image data in the input image and text to obtain shallow and deep image features; the structure incorporates Transformer and attention mechanisms, which together solve the problem of not being able to exploit the rich contextual information between adjacent features, enabling the model to perceive and enhance visual representation globally. ODLS is responsible for the feature extraction process of the text description information in the input image and text. The main structure is implemented by 1D convolution and BiLSTM; 1d convolution extracts valuable information for the input text, BiLSTM avoids the defect of losing critical information, and the two work together to complete the memory retention of

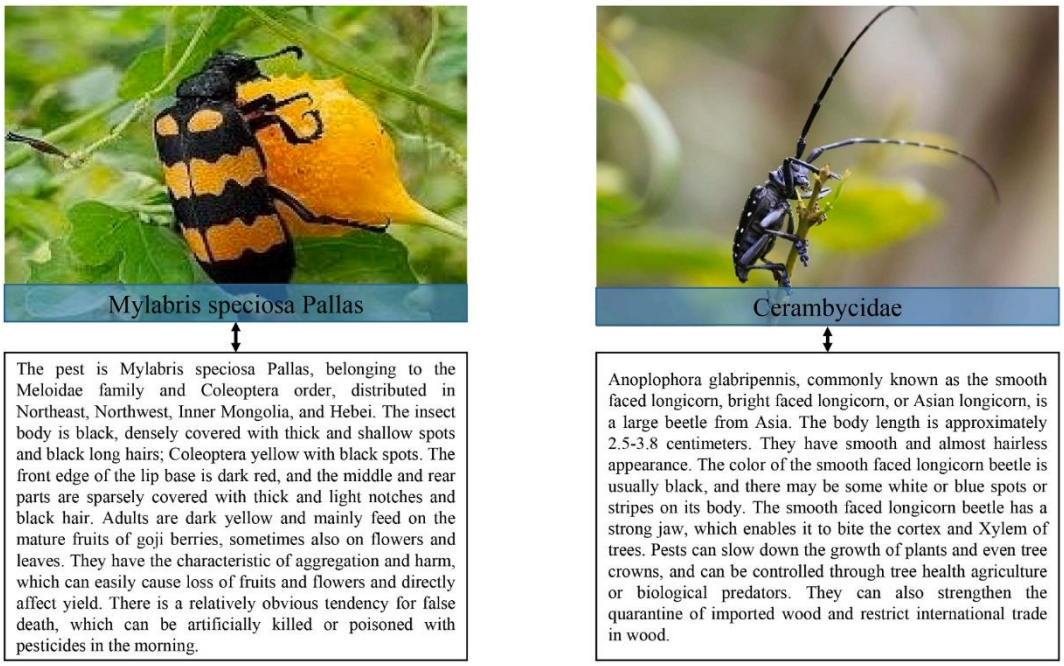


Fig. 1. Examples of two types of samples in the WPIT9K cross-modal data set of wolfberry pests.

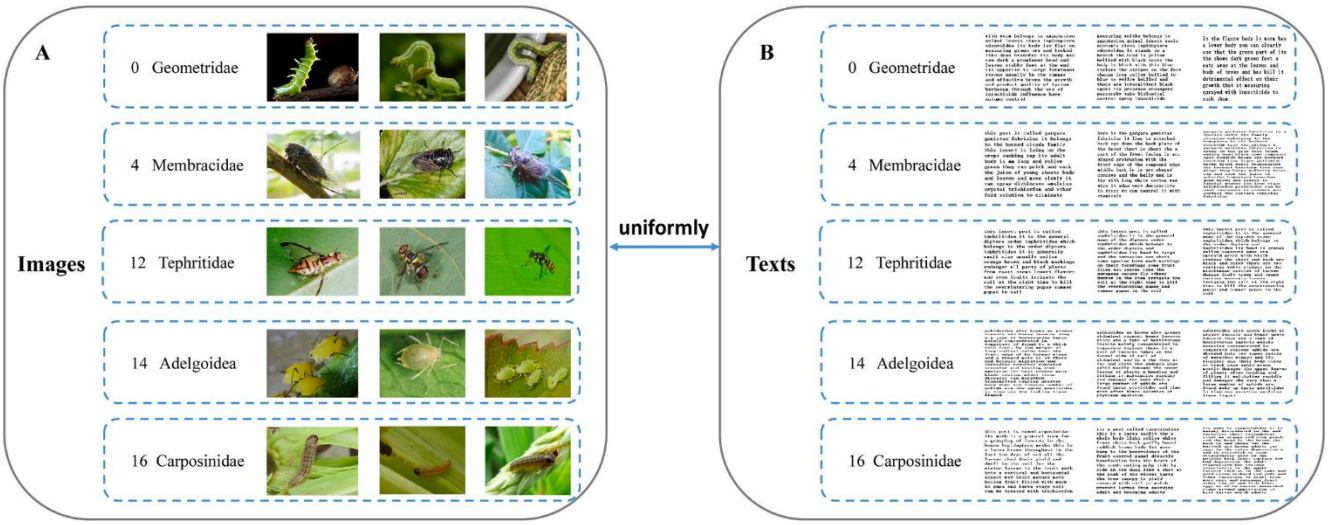


Fig. 2. Example of some category images and corresponding text of the WPIT9K dataset of Wolfberry pest. Wolfberry pest image (A), Wolfberry pest image corresponding text description (B).

crucial information. The image and text feature information jointly extracted by CoTN and ODLS are correlated and complementary. The multilayer perceptron (MLP) aims to achieve the recognition and classification of wolfberry pests after fusion feature information processing. Regarding the improvement of model convergence, ODLS uses Word2-Vec to build word vectors to complete the initialization of embedding layer weights, and CoTN keeps the weight data of the optimal model as a pre-training model. The structure and technical route of the overall model are shown in Fig. 3.

2.3. CoTN

The cross-modal feature fusion model ITF-WPI contains an image encoder CoTN structure that enables deep extraction of image features. The structure is designed with a Transformer-style architecture that fully uses contextual feature information between input keys to guide pattern extraction of the dynamic attention matrix, thus enhancing the visual representation.

The CoTN structure is mainly designed and implemented using the Contextual Transformer Network (CoT) structure (Li et al., 2023), which integrates contextual information mining and self-attentive learning into a unified structure. As shown in Fig. 4, the input X is a 2D feature map, which has a size of H × W × C, where H denotes the feature map height, W denotes the feature map width, and C denotes the number of feature map channels. The input X can be represented as $X \in \mathrm{R}^{H\times W\times C}$, followed by different aggregation operations according to the three streams of keys, queries, and values, respectively, with keys defined as $K = X$, queries defined as $Q = X$, and values defined as $V = X\, w_v$.

$$\mathrm{A} = [K^1, Q]W_\theta W_\theta \tag{1}$$

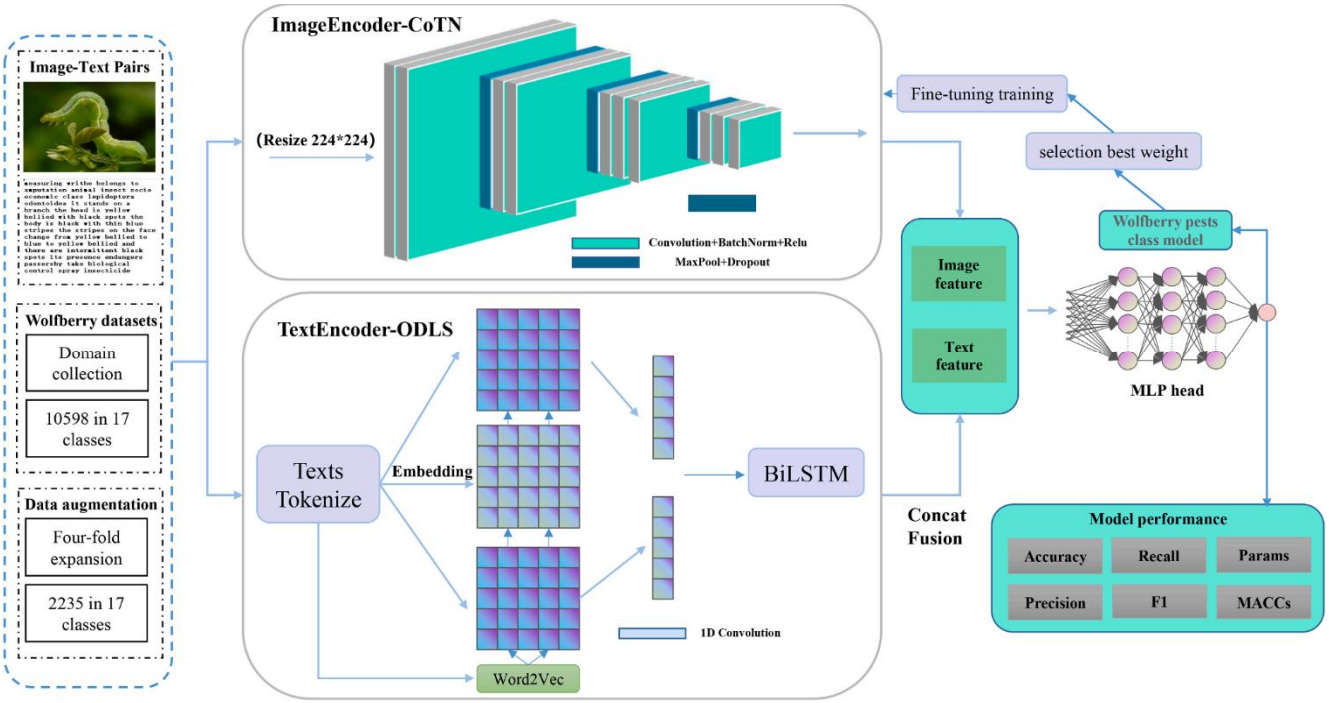


Fig. 3. Technology roadmap for ITF-WPI model architecture. image feature extraction encoder (CoTN), text feature extraction encoder (ODLS).

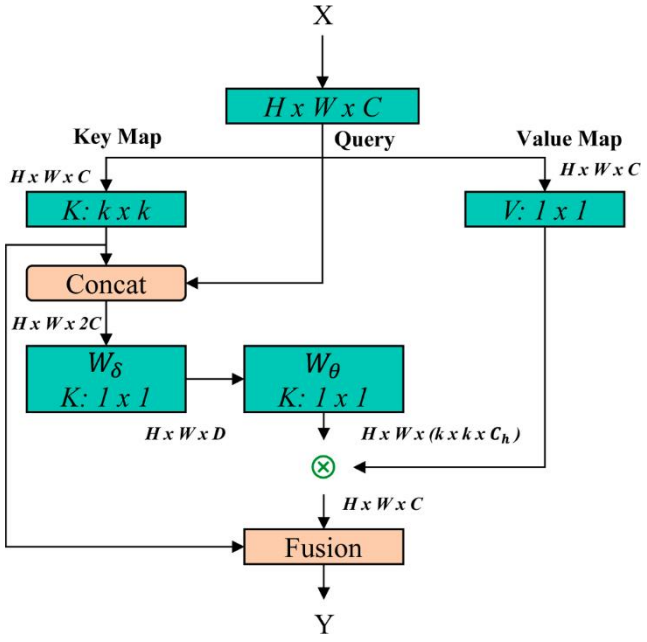


Fig. 4. Contextual Transformer (CoT) Network Structure.

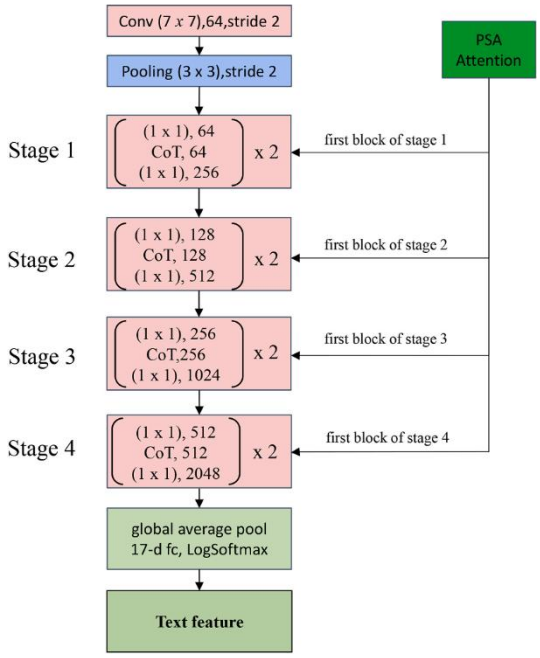


Fig. 5. The main structure of the CoTN network with 4 stages, CoT with PSA is embedded in four stage blocks.

$$K^2 = V \otimes A \tag{2}$$

First, the contextual representation of each key is achieved by performing ak × k group convolution on the space of all neighboring keys within the k × k grid of the feature map to obtain a contextual key $K^1 \in R^{H\times W\times C}$, which can reflect the static contextual information between local neighboring keys. Thus K^1 is used as the static contextual representation of the input X. Next, the query is concatenated with the contextual key K^1. The attention matrix A is realized by two successive 1 × 1 convolutions, a process that can be summarized as Equation (1). Notably, the first 1 × 1 convolution has aReLU activation function. Each spatial location of A will have a local attention matrix that is learned not based on independent query-key pairs but by combining query and vital contextual features. This approach enhances self-attention learning under the guidance of using static context K^1. Then, as shown in Equation (2), the feature vector of each spatial location is reshaped into a local attention matrix of sizeC_h × k × k according to the contextual attention matrix A. Next, the aggregation of the local attention matrix is obtained by aggregating the attention matrix A with all values V by local matrix multiplication, which is called the attended feature map K^2. Given that the participating feature maps K^2 can capture the dynamic feature interactions between the inputs, we name it the dynamic contextual representation of the inputs. This representation contains the interactions between different elements in the inputs. It can provide richer information than the original inputs while reflecting the input data's temporal relationships, interdependencies, and contextual information. This can help the model understand the input data better and make accurate predictions. Finally, the final output Y is obtained by fusing the static context K^1 and the dynamic context K^2 using the attention mechanism (Li et al., 2019), and this fusion fully considers the relationship between the static and dynamic features of the input data, which can improve the accuracy and robustness of the model. The attention mechanism allows the model to dynamically assign weights to different contextual representations, thus allowing the model to adapt better to different tasks and input data.

The main structure of the CoTN network is divided into four stages (Fig. 5), and the features extracted in each stage deepen with more filters. The initial convolution layer generates a 112 × 112 feature map using a 7 × 7 filter in stride 2, followed by a pooling operation on the feature map using a 3 × 3 max-pooling layers for further feature extraction. With this initial convolutional layer, we can extract rich low-level features from the input data and lay the foundation for subsequent computations. Similar network structures were used in stages 1 to 4, and each stage was generated as 2 networks with the same depth; the CoT structure was reused 2 times in each stage of the network, and the Swish activation function was used to improve the stability and accuracy of the model in addition to the normal ReLU activation function (Ramachandran et al., 2017). For the feature maps generated by the four-stage network, the initial convolutional feature map is processed by the first-stage network to obtain a feature map with 256 channels of 56 × 56 size; after that, stage 2 generates a feature map with 512 channels of 28 × 28 size, stage 3 generates a feature map with 1024 channels of 14 × 14 size, and stage 4 generates a feature map with 2048 channels of 7 × 7 size. This part of feature extraction processing is a cumulative process, from initial low-level features of color, texture, edge, and corner regions to high-level semantic abstract features, these high-level semantic features can represent objects, scenes, and semantic information in the image with more prosperous expressive power, and the addition of CoT structure makes it possible to focus on high-level semantic features while paying more attention to the semantic features'contextual information. In addition, to improve the model's ability to perceive helpful information while suppressing unwanted noise, a Pyramid Squeeze Attention (PSA) module (Zhang et al., 2022) is embedded in the first network of each stage to effectively improve the performance and expressiveness of the convolutional neural network. The structure of the attention mechanism about the CoTN network is incorporated into the implementation as a plug-in to facilitate a smooth testing process.

2.4. PSA attention

PSA can explore channel feature information at multiple scales, facilitating the efficient extraction of multi-scale spatial information at a finer granularity level and adaptively recalibrating cross-channel attention weights to enrich the feature space. Each stage block of the CoTN network uses the PSA module for multi-scale feature fusion, and Fig. 6 shows the specific process of the PSA attention module. The input feature map X has C channels, and by dividing the input X into S groups and calculating the weights of each channel in different groups, in this

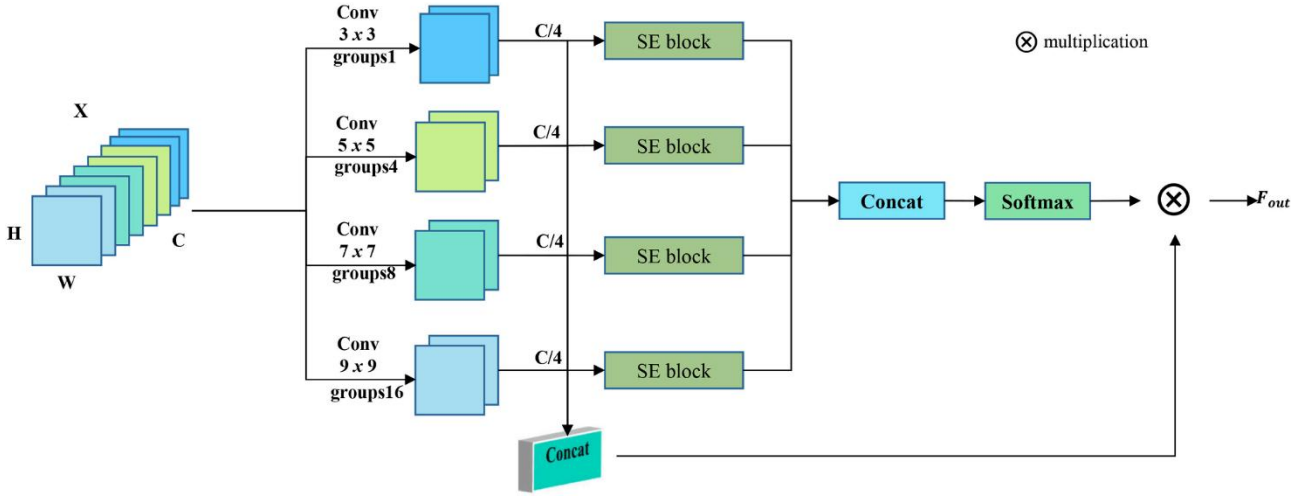


Fig. 6. Detailed description of the PSA attention module structure.

study, S is 4. The output channels are consistent for each group of Convolution and are calculated by *c/s*. The computed multi-scale features can be expressed as:

$$F_i = Conv2d(k_i \times k_i, G_i)(X)(i = 0,1,\cdots,S-1) \tag{3}$$

Where Conv2d denotes the 2-dimensional convolution operation, k_i denotes the convolution kernel size determined by $2 \times i + 3$, G_i is the parameter of the i-th group convolution, and the size of the S-group convolution kernel in this study is $k = \{3, 5, 7, 9\}$. The number of groups is $G = \{1, 4, 8, 16\}$. The entire multiscale feature map obtained by connecting the features of multiple branches is given by the following equation:

$$F = Concat([F_0, F_1, \cdots, F_{S-1}]) \tag{4}$$

where Concat denotes the feature map connection operation and the feature map $F \in R^{H \times W \times C}$. Then,the multiscale features $F_i (i = 0, 1, \cdots, S-1)$ of each channel are used to calculate the channel weights of different groups separately by the channel attention mechanismSEWeight, which can be expressed as:

$$SEW = Concat(SEWeight(F_i)), (i = 0,1,\cdots,S-1) \tag{5}$$

Concat connects different channels 'attention weights, and SEW denotes the connected multi-scale attention weight vector. Finally, the multi-scale channel attention weights are recalibrated by the Softmax function, and the multiplication operation ($\otimes$) on the channels is per- formed with the multi-scale feature map F. This method incorporates different sizes of perceptual fields and rich multi-scale spatial informa- tion, and the final output of the multi-scale feature map as:

$$F_{out} = F \otimes \text{Softmax}(\text{SEW}) \tag{6}$$

2.5. ODLS

The main structure of the ODLS network is divided into 3 layers (Fig. 7), including the word embedding layer, convolutional network layer and memory network layer. First, the word embedding layer inputs text vectors into the network, and the text vector interval is the maximum of the batch text vectors. In general, the word embedding layer weights are initialized randomly. To accelerate the convergence process, the process uses the word vector weight parameters obtained by Word2Vec instead of randomly initializing the embedding word layer. Next, the convolutional network layer is constructed by stacking the convolutional and pooling layers. 1D convolution and pooling are used to build the convolutional network layer because it is more computationally efficient than 2D convolution and pooling. Subsequently, considering that the input data is usually a sequence of words or characters, and there is a temporal sequence between these words or characters, it is necessary to consider the contextual information. The bidirectional long and short-term memory network (BiLSTM) used in the memory network layer can retain this memory capability. 2 BiLSTMs are used in the structure. The first BiLSTM is stacked with 2 LSTM units, and the stacked LSTM units can deepen this memory capability similar to human behaviour. However, overfitting may occur, and then Dropout is used to reduce the possible overfitting of the stacked units. Finally, the output features of the memory layer are flattened as the text feature input. Regarding the ITF-WPI image and text cross-modal feature fusion,

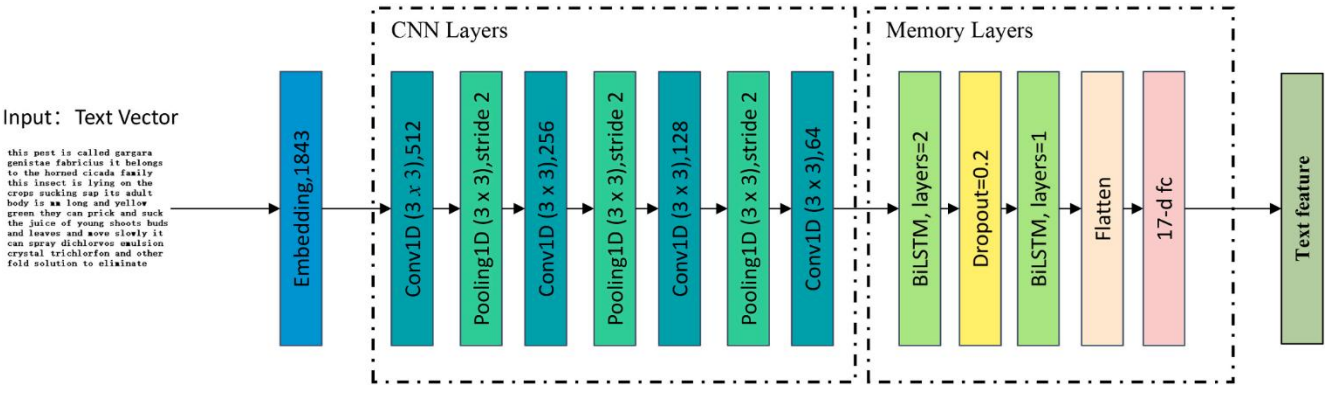


Fig. 7. Convolutional and memory layers form the ODLS network, which consists of 1D convolutional layers stacked with BiLSTM layers, respectively.

by fusing the image and text features output from CoTN and ODLS networks at the end as the input to the MLP header, the MLP hidden layer is defined as 128, and the output results as the final type of 17 pests of Wolfberry.

2.6. Model evaluation metrics

To comprehensively evaluate the model performance, four commonly used metrics: Accuracy (Equation (7)), Precision (Equation (8)), Sensitivity (Equation (9)) and F1-Score were selected for the experiment to determine the accuracy of wolfberry pest classification.

$$\text{Accuracy}=\frac{TP_{wolfberry}+TN_{wolfberry}}{TP_{wolfberry}+FP_{wolfberry}+TN_{wolfberry}+FN_{wolfberry}} \tag{7}$$

$$\text{Precision}=\frac{TP_{wolfberry}}{TP_{wolfberry}+FP_{wolfberry}} \tag{8}$$

$$\text{Sensitivity}=\frac{TP_{wolfberry}}{TP_{wolfberry}+FN_{wolfberry}} \tag{9}$$

The combination of actual and predicted categories is detected according to the model, where $TP_{wolfberry}$ indicates the number of correctly classified positive samples (true positive); $FP_{wolfberry}$ indicates the number of incorrectly classified positive samples (false positive); $FN_{wolfberry}$ indicates the number of incorrectly classified negative samples (false negative); and $TN_{wolfberry}$ indicates the number of correctly classified negative samples (true negative).

For the estimation of model complexity, the number of model floating-point computations (MACCs) and the number of parameters (Params) are chosen as essential metrics for evaluation. The number of floating-point computations reflects the complexity of the model in time, and the size of the number of parameters is directly related to the model size and also affects memory usage during model inference. In other words, the number of parameters directly determines the model size and computation consumption, and the larger the size of the model parameters, the more memory is required.

3. Experimental results and analysis

The experiments were performed on a graphics workstation; the main algorithms were executed by Microsoft VSCode and Python 3.9.13, the Pytorch deep learning framework with version 1.13.1 + cu117, TorchVision version 0.14.1 + cu117, cross-modal model construction with TorchMultimodal, by GPU use and acceleration, CUDA and cuDNN version 11.7, Matplotlib and Seaborn are chosen for image drawing, and Scikit-learn is chosen for performance evaluation. The operating system is Windows 11 Professional Workstation Edition, Intel i9-13900KF processor, memory option 4th generation DDR4 with 128.00 GB RAM, graphics card NVIDIA GeForce RTX 3090 with 24 GB video memory, and 2 T solid state drive.

In this study, although the CoTN network in the ITF-WPI model embeds a batch normalization (BN) layer, we randomly sampled the dataset to speedup the model convergence. We calculated the mean and standard deviation of the image pixels and normalized the obtained values to the training set input images. Since Word2Vec can learn the vector representation of words in a high-dimensional vector space and calculate the cosine distance between words to show theirrelevance (Ma et al., 2023a), finding the semantic relationships between words in a description document is easy. Next, by using Word2Vec was used to check the semantic relationships between words describing Wolfberry pests; the two-dimensional spatial projection of all word vectors drawn is shown in Fig. 8, which shows the visualization of the relationships in two-dimensional space for nine pest names randomly selected from 17 categories of pest names. In addition, the matrix constructed by Word2Vec is used to initialize the embedding word layer of the ODLS network to reduce the fitting time. The selection of an appropriate learning rate directly impacts the convergence speed and performance of the model. This study uses a dynamic learning rate adjustment called the Cosine Annealing Warm Restart strategy (Lee et al., 2023). For the optimization function, we used SGDR as the optimizer (Loshchilov & Hutter, 2017), which has a periodic restart mechanism and thus has the potential to make the optimization process jump out of the optimal local solution, thus obtaining better classification performance and reducing the training time. The early stop method, a simple but effective technique to prevent overfitting (Liu et al., 2022), was integrated with the ITF-WPI training process. The patience value for monitoring the validation loss was set to 3. All experiments were implemented using PyTorch, and Table 2 shows the optimized hyperparameter values for the proposed ITF-WPI model.

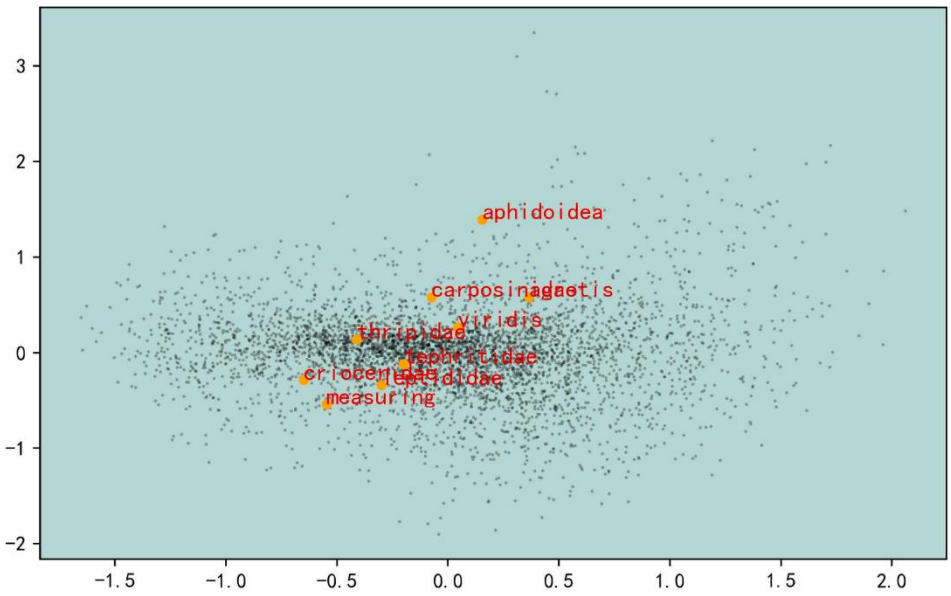


Fig. 8. WPIT9K Two-dimensional spatial visualization of text vectors.

Table 2
Optimized hyperparameter values of the ITF-WPI model.

Experimental Settings	Hyperparameter	Optimized value
Training Settings	batch size	32
	optimizer	SGDR
	momentum	0.9000
	weight decay	0.0005
	learning rate	0.0100
	loss function	CrossEntropyLoss
	epochs	30
Image Normalize	image mean	[0.4777281,0.52651316,0.3463837]
	image std	[0.2726728,0.26582387,0.2776033]
EarlyStopping	patience	6
LearningRate Scheduler	T_0	10
	T_mult	2
	eta_min	0
	min_count	20
Word2Vec Settings	epochs	10
	window	2
	vector_size	100
	negative	10

3.1. Model classification effect evaluation

This section validates the classification performance using the wolfberry pest image and text datasets. The dataset 's number of samples directly affects the model 's accuracy. The ITF-WPI model is trained on the training set and tested on the corresponding test dataset to verify the stability of the model trained on different numbers of training sets. The image and text sample sizes for training and testing are 9538 and 1060, respectively, and the sample size of the training set is increased sequentially during testing. The obtained model is validated on the test dataset. The experimental results are shown in Table 3; with the increase of the sample size of the training set, the accuracy of the model increased sequentially, and when the sample size of the training set was 7982 and 9538, the accuracy of the model was close, with an average accuracy of 97.91%. Table 4 shows the accuracy of each wolfberry pest, after randomly selected data consisting of test images with text samples, the average accuracy of all wolfberry pests was 97.98%, among which the accuracy of Adelgoidea and Carposinidae is lower; from the visual analysis, many pests in the Adelgoidea category are harder to distinguish and similar to the colour of plant leaves, and there are many pests in the Carposinidae category that are similar to They are the main reasons for the low accuracy.

We conducted an additional 3-fold cross-validation experiment using the wolfberry pest images with the text dataset to validate the model 's accuracy further. The wolfberry pest images and text dataset were divided into three parts; each subset disjoints with 3179, 3179 and 3179 images, respectively. 30% of one of the subsets was extracted as the test dataset, the remaining part was used as the training dataset, and 20% of the training set was used as the validation set. The whole process was repeated three times until each subset was used as the test set, and the remaining subsets were used as the training set. Table 5 shows the experimental results of the 3-fold cross-validation with an average accuracy of 97.39%, which is slightly less accurate compared with Tables 3 and 4, but the accuracy loss is controlled within 0.55%; therefore, we believe that the accuracy of the ITF-WPI model is not affected by the K-fold cross-validation.

Table 3
Test accuracy of ITF-WPI model on training dataset with different sample sizes (units: %).

Dataset	Sample size	Precision	Sensitivity	F1-Score	Accuracy
Wolfberry pest image and text dataset	3988	68.59	71.41	69.97	76.36
	4652	79.68	76.47	78.04	81.64
	6317	88.39	85.43	86.88	84.90
	7982	92.39	93.44	92.91	97.89
	9538	92.47	93.49	92.97	97.94

Table 4
Accuracy of ITF-WPI on WPIT9K dataset for various pests (units: %).

# Class names	Precision	Sensitivity	F1-Score	Accuracy
1 Geometridae	79.67	82.56	81.08	99.27
2 Cicadella viridis	83.33	100.00	90.91	95.47
3 Crioceridae	95.68	89.79	92.64	97.58
4 Elthemidea sp	97.64	87.41	92.24	99.64
5 Membracidae	100.00	96.34	98.13	98.54
6 Mylabris speciosa Pallas	98.78	99.54	99.15	99.37
7 Tropidothorax elegans distant	96.58	97.61	97.09	94.59
8 Cerambycidae	100.00	100.00	100.00	100.00
9 Nephrotoma sp	100.00	84.25	91.45	97.63
10 Thripidae	100.00	100.00	100.00	100.00
11 Epitri abeillei	81.44	92.91	86.79	94.79
12 Bedbug	95.47	95.69	95.58	99.29
13 Tephritidae	96.69	100.00	98.31	99.27
14 Agrotis ypsilon	100.00	100.00	100.00	100.00
15 Adelgoidea	79.98	87.49	83.56	96.59
16 Plodia interpunctella	91.69	97.48	94.49	99.47
17 Carposinidae	79.84	86.25	82.92	94.25
Average	92.75	93.96	93.19	97.98

Table 5
Method accuracy by 3-fold cross-validation on the WPIT9K dataset (units: %).

Dataset	Folds	Precision	Sensitivity	F1-Score	Accuracy
Wolfberry pest image and text dataset (ITF-WPI)	Fold 1	91.41	89.98	90.68	96.15
	Fold 2	91.24	92.86	92.04	98.34
	Fold 3	93.04	91.34	92.18	97.69
	Average	91.89	91.39	91.63	97.39

3.2. Transfer learning experiments

Transfer learning (TL) is transferring the knowledge of an already trained model to another relevant task (Morid et al., 2021). By this method, the already learned knowledge can be used to accelerate the training and improve the model 's generalisation. In this study, the branched CoTN and ODLS networks of the ITF-WPI model can participate in knowledge transfer as independent parts. The CoTN network uses the image data part of the wolfberry pest image and cross-modal text data set to generate pre-weighting and only changes the original fully connected layer when integrated into the ITF-WPI model. The embedding word layer in the initial part of the ODLS network facilitates knowledge migration, and the word vectors obtained through the Word2Vec method training are used as the initialization weights of the word embedding layer. CoTN and ODLS networks complete the initialization of their respective weights before training the ITF-WPI model, which can be regarded as an overall knowledge migration scheme for initializing the weights of the cross-modal model from both image and text aspects. The results are shown in Fig. 9, where the ITF-WPI model with knowledge migration has a lower loss value and higher accuracy than the normal training. At the beginning of training, the validation loss value of migration learning is at least 5% lower than the training loss value of normal training. The validation accuracy of migration learning is at least 21.5% higher than the training accuracy of normal training. The migration learning reaches the maximum accuracy rate of 5 Epochs earlier than normal training, and the convergence rate is significantly faster. The accuracy and loss curves generally have similar patterns, and the larger amplitude of the curve pattern is caused by the cosine annealing hot restart strategy to adjust the learning rate dynamically.

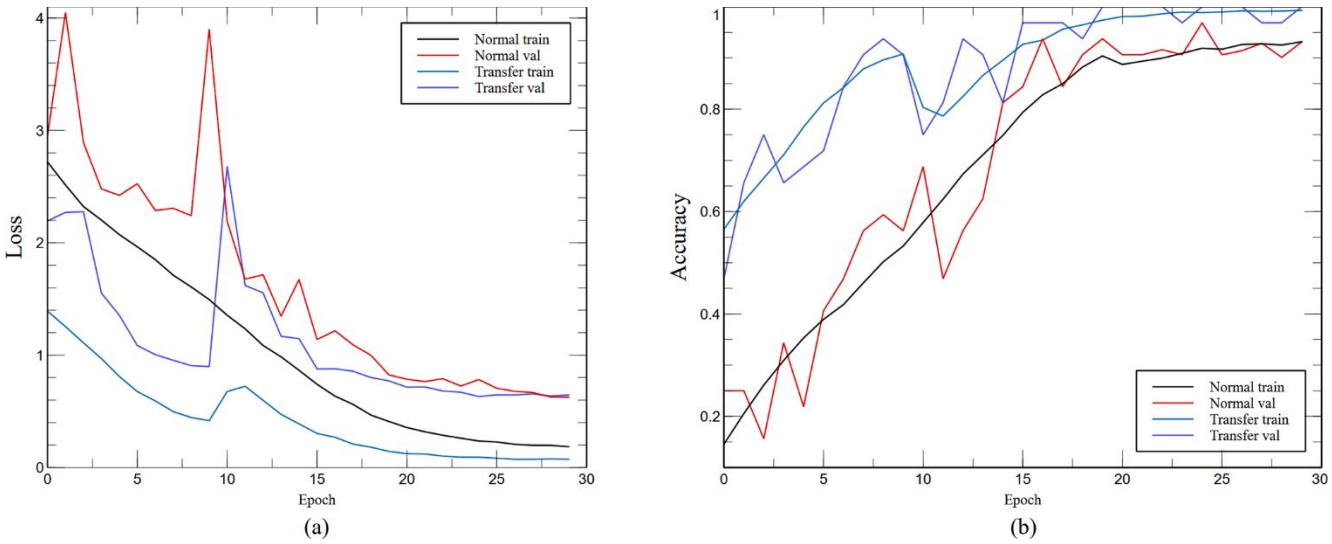


Fig. 9. Accuracy and loss of normal training vs. migration training at ITF-WPI. (a) Training vs. validation loss; (b) Training vs. validation accuracy.

3.3. Attention mechanism experiment

Attention mechanisms are widely used in deep convolutional neural network models because they can provide attention to important information and thus improve the model 's performance. We integrate the PSA mechanism in the CoTN network structure of the ITF-WPI model, and to verify the impact of other attention mechanisms on the model performance, the attention mechanism structure that can be flexibly integrated into existing network architectures and is widely used is selected for testing. Specifically, the SENet that adds the attention mechanism in the channel dimension, the core unit of SENet is the Squeeze-and-Excitation (SE) module (Zhu et al., 2023); the Effective Squeeze-Excitation (ESE) module (Chen et al., 2021), as an improved version of SENet, solves the information loss in the process of incre- mental and downscaling The Efficient Channel Attention (ECA) module can learn channel attention efficiently with low model complexity (Wang et al., 2023); the Convolutional block attention module (CBAM) uses tandem structure to fuse channel attention with spatial attention (Ijaz et al., 2023), which achieves a sequential attention structure from channel to space; ParNet is a novel attention module (Goyal et al., 2022), which consists of several parallel sub-networks, each of which is responsible for extracting feature information at different levels and fusing feature information at different scales through an attention mechanism. The above attention structures were embedded into the CoTN network structure of the ITF-WPI model for testing, and Fig. 10 shows the experimental results. The PSA attention structure achieved the highest accuracy of 97.98% with model parameters of 52.20 M. The PSA and ParNet attention structures had similar accuracy, but the PSA model parameters were 31.88% lower than those of ParNet. PSA and ParNet were at least 1.72% more accurate than CBAM, but the model parameters were, on average, 75.11% more than CBAM. Therefore, the attention mechanism structure with more parameters brings better

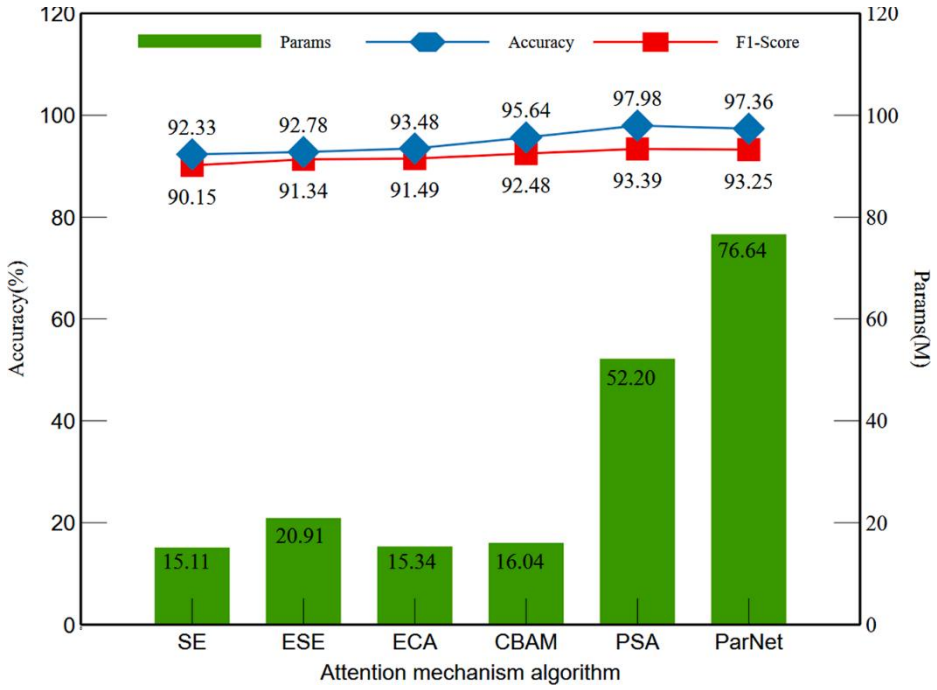


Fig. 10. Performance of SE, ESE, ECA, CBAM, PSA, ParNet attention mechanisms in ITF-WPI.

performance. However, it puts higher requirements on the training equipment, and the appropriate attention mechanism structure should be selected according to the actual situation.

3.4. Ablation experiments

In this section, ablation experiments are conducted on the main methodological structures involved in the ITF-WPI model. Two major network structures, CoTN andODLS, constitute the ITF-WPI model. The CoTN network uses a CoT structure that enables contextual information extraction between input keys, and a PSA attention mechanism extracts multi-scale spatial information. The main methods used by the ODLS network include Word2Vec extracted word vectors to initialize the embedding layer weights, the first layer BiLSTM (BiLSTM-One) and the second layer BiLSTM (BiLSTM-Two) used in the memory network layer. Considering the specificity of the input text structure, the memory network layer of the ODLS network in the experiment always ensures the presence of LSTM units. The above five techniques are combined for ablation verification, and the experimental results are shown in Table 6. The intervention of the CoT structure improved the accuracy of the ITF-WPI model by 2.92%, CoT compared to the PSA attention mechanism by 2.31%, and both CoT and PSA together improved the accuracy by 3.68%. The Transformer-style CoT-based structure achieves the most significant improvement in the performance of the ITF-WPI model. 0.84% improvement in the accuracy of BiLSTM-One compared to BiLSTM-Two is attributed to the fact that BiLSTM-One has two LSTM units and an additional Dropout layer, as demonstrated in experimental sequences 7–9. Word2Vec hardly affects the performance ofthe ITF-WPI model, and the combined five methods result in a classification accuracy close to 98%. On the other hand, to verify whether the ODLS memory network layer stacked BiLSTM is reasonable, we stacked four layers of BiLSTM on the original memory network layer for comparison (Table 7). The initial stacked BiLSTM exceeded the accuracy of the ODLS network by 0.14%, and the number of parameters and MACCs increased by 1.19 M and 0.058, respectively (Table 7, rows 1–2), and the subsequent stacked BiLSTMs made the accuracy lower and also brought more computational burden. Based on the above results, the ITF-WPI cross-modal model shows superior performance under comprehensive performance and computational reasonableness.

3.5. Feature map visualization

Visualization of neural network models is a powerful tool to explore and understand the black-box learning behaviour of deep learning models. By visualizing the feature maps, we can diagnose the feature extraction of each network layer during model training and identify potential problems with the model by the performance of the feature maps. By obtaining the feature maps extracted by the convolutional neural network filters layer by layer, we can observe the feature extraction of each network layer during training, which helps us better understand how the network layers abstract and process the data.

Randomly selected pest images of Wolfberry were used as inputs, and the image input dimensions obeyed the input conventions of the ITF-WPI network model. As shown in Fig. 11, the observable features are intercepted from different network layers of the ITF-WPI model. Fig. 11 **(A)** represents the 3D structure of the multiple overlapping convolutional feature maps used for feature extraction. Fig. 11**(B)** represents a two-dimensional feature map intercepted from multiple overlapping convolutional feature maps for representing the feature extraction process. A and B are, to some extent, corresponding processes. As can be seen in the B figure, the Stage1 block gains the ability to extract the edges of the image by learning samples, and the Stage2 block focuses on the extraction of the texture features of the berry pest, obtaining a better representation of the overall structure of the pest limb. It is worth noting that the feature objects focused on the same convolutional block structure are not consistent; some feature maps express the edges of the pest, some express the edges of the limb, in addition to a few features expressing both the edges of the pest and the texture of the limb. From Stage 3 to Stage 4 block structure, the edge and texture information of the pest becomes blurred, and the generated images become indistinguishable from naked-eye observation. Therefore, the mapping process of extracting features from the neural network model can help us to gain a deeper understanding of how the model works and thus further optimize the performance and effectiveness of the model.

Table 7
Performance experimental results of ODLS networks stacked with BiLSTM layers.

Layers	Units	Probability	Accuracy (%)	Params (M)	MACCs (G)
BiLSTM	6	_	_	_	_
Dropout	_	0.2	98.12	53.39	7.886
BiLSTM	4	_	_	_	_
Dropout	_	0.2	97.45	54.97	7.964
BiLSTM	2	_	_	_	_
Dropout	_	0.2	96.62	55.76	8.003
BiLSTM	2	_	_	_	_
Dropout	_	0.2	95.49	56.55	8.042

3.6. General comparison of SOTA models

3.6.1. Comparative analysis of CoTN and SOTA models

This section will thoroughly verify the performance and generalizability of the ITF-WPI model. Since the ITF-WPI model is composed of two major network structures, CoTN andODLS, and CoTN and ODLS are responsible for the encoding of images and text, respectively, different SOTA models will be selected as feature extraction structures for separate validation, which is carried out based on the WPIT9K dataset of wolfberry pests. For the CoTN network, nine SOTA models (AlexNet, ResNet50, ShuffleNetV2, MobileNetV3, InceptionV3, ResNeXt50, SwinTransformer V2 (SwinTV2 -Small), VisionTransformer-B/16 (ViT-B/16), ConvNeXt (ConvNeXt-Small)) were evaluated. VisionTransformer-B/16 (ViT-B/16) and ConvNeXt (ConvNeXt-Small)) are the most influential models up to now. These models ' structure and weight parameters can be obtained from the Torchvision model library,

Table 6
Ablation experiments with the ITF-WPI model.

No.	Trick and methods					Accuracy (%)	Params (M)	MACCs (G)
	CoT	PSA	Word2Vec	BiLSTM-One	BiLSTM-Two			
1					*	92.95	13.60	2.229
2	*				*	95.87	14.55	2.354
3		*			*	93.56	46.99	7.283
4	*	*			*	96.63	51.41	7.789
5	*	*		*		97.47	51.81	7.808
6	*	*		*	*	97.93	52.20	7.828
7	*	*	*	*		97.34	51.81	7.808
8	*	*	*		*	96.59	51.41	7.789
9	*	*	*	*	*	97.98	52.20	7.828

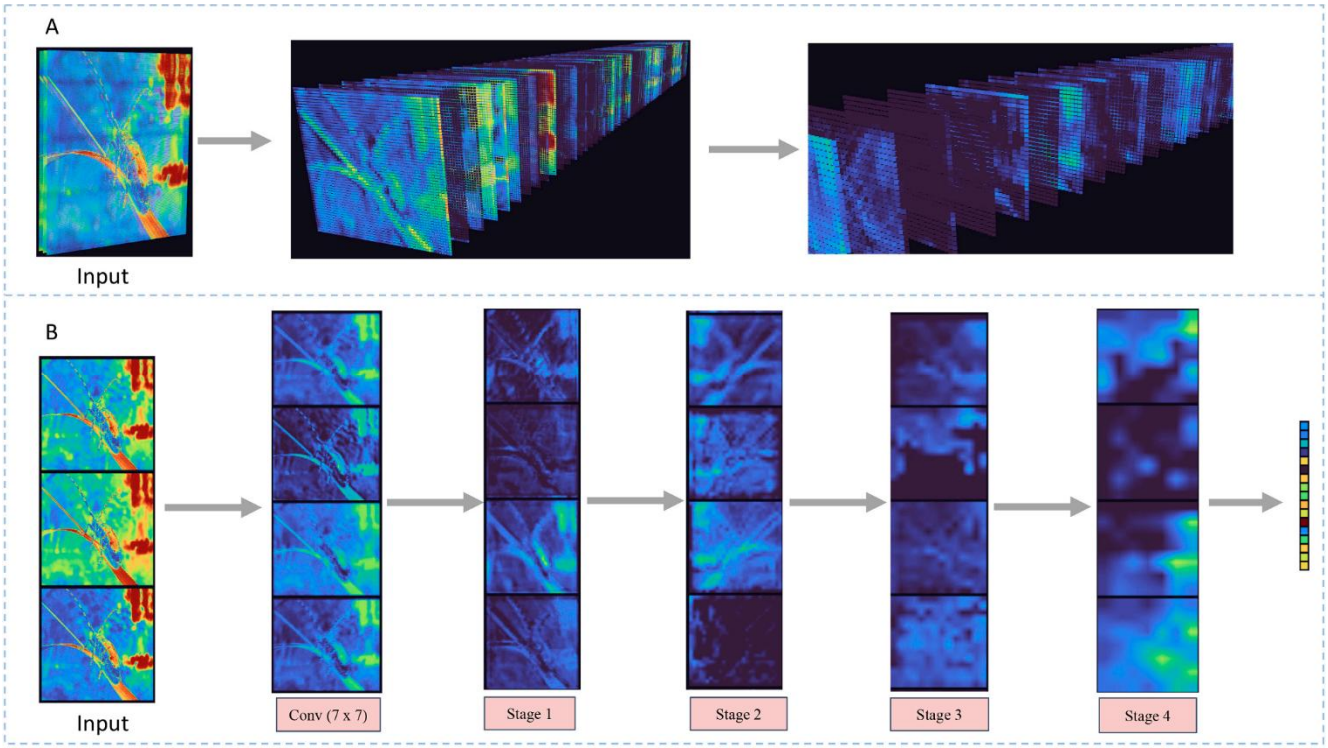


Fig. 11. The convolutional learning behavior process of the ITF-WPI model, the B-plot corresponds to the convolutional features contained in the A-plot.

except for ResNet50 and ResNeXt50, which use IMAGENET1K_V2 weights, while the rest use IMAGENET1K_V1 weights. In addition, custom models can support niche industry features and more complex application scenarios. These model structures have some comparative value; e.g., Wang et al. (2022) proposed a backbone network with an improved Swin Transformer for enhanced cucumber leaf disease identification performance. (Wu et al., 2023) proposed an improved ResNet-50 deep learning algorithm for chicken gender recognition based on the SE attention mechanism, Swish loss function and optimizer Ranger. An improved DenseNet-CNN model (CA_DenseNet_BC_40) based on the CA attention mechanism was proposed to classify damage caused by cotton aphids (Bao et al., 2022). Ma et al. (2023b) improved the VGG16 structure and proposed a lightweight CNN deep learning model (CornNet) to classify corncob seeds. In order to meet the feature fusion of the two sub-networks of the ITF-WPI model in this study, the number of neurons of the last FC of all comparison models was changed to 17. Since the structure of the backbone network of the comparison models was not changed, the training could be performed by migration learning to ensure the credibility of our experimental results, shown in Table 8.

Table 8
Experimental results of different SOTA DL models in CoTN networks.

Model	Accuracy (%)	F1-Score (%)	Params (M)	MACCs (G)
AlexNet	87.42	84.98	60.28	1.155
ResNet50	90.16	89.69	26.75	4.576
InceptionV3	87.19	86.37	28.35	3.299
ResNeXt50-32x4d	91.49	89.87	26.22	4.731
ShuffleNetV2 × 0.5	84.16	83.97	3.57	0.488
MobileNetV3-large	85.38	85.06	7.43	0.678
SwinTV2-Small	98.12	96.82	52.36	6.240
ViT-B/16	98.64	97.34	89.02	11.730
ConvNeXt-Small	97.85	96.29	52.67	9.141
Improved ResNet-50	92.68	91.04	28.49	4.859
CA_DenseNet_BC_40	83.56	82.45	3.39	0.436
CornNet	85.72	84.96	3.31	0.453
Improved SwinT	98.57	96.43	189.38	9.225
ITF-WPI (ours)	97.98	93.39	52.20	7.828

The average accuracy of the classical four SOTA neural network (AlexNet, ResNet50, InceptionV3, ResNeXt50-32x4d) models was 89.06%, with an F1 score of 87.72%, which was 4.5% lower than the accuracy of the ITF-WPI model using only the PSA attention mechanism fused. At the sametime, it was lower than the CoT structure used, with a 6.81% lower accuracy (Table 6). The improved ResNet-50, although higher than the classical SOTA neural network model, still has lower accuracy than the ITF-WPI. In addition, the ITF-WPI model using only the CoT structure is more advantageous regarding model computation and parameter size.

Lightweight SOTA neural networks (ShuffleNetV2 × 0.5 and MobileNetV3-large) and improved lightweight neural network models (CA_DenseNet_BC_40 and CornNet) have an average accuracy of 84.7% and an F1 score of 84.11%, which is lower than the ITF-WPI model, but lightweight models lower resource requirements of the computing platform and can be used for integration to cross-modal models in particular scenarios.

The average accuracy of the Transformer structure-based neural networks (SwinTV2-Small, ViT-B/16 and Improved SwinT) is 98.44%, with anF1 score of 96.86%. ConvNeXt, as an evolutionary version ofthe convolutional network after combining all the extraordinary designs of SwinT and ViT in one, upgrades the ResNet architecture to obtain comparable accuracy of Transformer structured network, ITF-WPI model has lower model parameters compared to the former, with 0.66% lower accuracy compared to ViT-B/16, but only 3/5 of model parameters and 33.27% lower MACCs. In terms of combined computational consumption and accuracy, although SwinTV2-Small achieved the optimal computational performance, comparing the ITF-WPI model incorporating the CoT structure, the accuracy increased by only 0.14%, yet the MACCs increased by 0.31%. The decisive advantage of the Transformer-style CoT-based structure is demonstrated, as well as the comprehensiveness of the CoT structure for enhanced visual representation in cross-modal modeling.

3.6.2. Comparative analysis of ODLS and SOTA models

The ODLS network mainly consists of CNN and memory network layers to complete feature extraction and transfer; a neural network (CNN-LSTM) comprises stacked CNN and LSTM layers to compare with other advanced structural neural networks and four custom-built CNN-LSTM networks are selected. These models are a model to predict water temperature based on spatiotemporal characteristics of soil temperature field (CNN-LSTM-A) (Zhang et al., 2023); Hybrid CNN-LSTM model (Hybrid CNN-LSTM) to predict new daily COVID-19 cases in India (Verma et al., 2022); a model to predict salmon freshness by mining the temperature change during salmon storage (CNN-LSTM- B) (Wu et al., 2022); Deep learning model for island detection by one-dimensional CNN implementation (CNN-LSTM-C) (Ozcanli & Baysal, 2022). We implemented the network structures proposed in these studies and integrated them into the ITF-WPI model for testing. Table 9 shows the experimental results of the four models. The CNN-LSTM-C model achieves an optimal accuracy of 98.21%, which is 0.23% more accurate than the ITF-WPI model. The ITF-WPI has the second-best accuracy and lower model parameters than the other four CNN-LSTM models. In addition, the MACCs are 30% lower than the CNN– LSTM-C. The accuracy of ITF-WPI is the second best and has lower model parameters than the other four CNN-LSTM models. From the analysis of the model structure, CNN-LSTM-C and ITF-WPI have similar structures, and they both choose to stack more 1D convolution and 1D max-pooling layers on top of the CNN structure of CNN-LSTM. It should be noted that ITF-WPI uses BiLSTM on the LSTM structure of CNN-LSTM and chooses fewer cells so that it can keep lower mode parameters and MACCs. CNN-LSTM-A, Hybrid CNN-LSTM and CNN-LSTM-B all flatten the extracted features located on the CNN structure of CNN-LSTM to one-dimensional vectors, directly affecting their accuracy. 2-dimensional convolution used in CNN-LSTM-A leads to increased computational cost and also increases the complexity of the model, and is therefore not suitable for use in the extraction of serialized data for feature extraction.

3.7. Visualization of regions of interest

The training of the cross-modal ITF-WPI model has achieved initial success, but the explanatory mechanism within the model still needs to be clarified. The effect of feature extraction by model visual interpretation can better express the mechanism inside the model. We applied the gradient-weighted class activation mapping (Grad-CAM) algorithm to the study of model interpretability (Batchuluun et al., 2023; He et al., 2023). The magnitude of the Grad-CAM activation map indicates the degree of influence of the pixel at the corresponding position in the original image on the classification result, so the most vital position of the activation map is the position where the target is located, i.e., the region of interest (ROI) of the model. Fig. 12 shows the actual Wolfberry pest identification heat map for the three types of models; the ITF-WPI model with PSA has a larger field of perception and can cover a broader range of essential regions compared to the ITF-WPI with fused CBAM, and the ITF-WPI without attention to highlight the concept of predicted targets. The ITF-WPI, without using any attention mechanism, can identify important regions based on objects (1–2 images in the third row on the right), reducing the coverage of redundant regions. At the same time, the other two models do not show ROI heatmaps for this feature. Both the ITF-WPI with CBAM and the ITF-WPI without the attention mechanism suffer from the loss of important recognition object regions (the second row on the left versus the first image in the third row). However, they both classify this image typically and therefore do not affect recognition accuracy as a classification task. The above demonstrates that both the CoT structure and PSA that ITF-WPI have played a vital role in the model.

Table 9
Experimental results of different advanced CNN-LSTM models in ODLS networks.

Model	Accuracy (%)	F1-Score (%)	Params (M)	MACCs (G)
CNN-LSTM-A	96.59	92.77	70.94	13.245
Hybrid CNN-LSTM	96.87	92.89	67.59	7.649
CNN-LSTM-B	96.49	92.64	72.92	7.629
CNN-LSTM-C	98.21	94.67	57.427	11.321
ITF-WPI (ours)	97.98	93.39	52.20	7.828

4. Conclusion and discussion

This study addresses the problems of less research related to pest recognition, complex and variable recognition background environment, and single modality of pest information recognition in agriculture. In this study, a deep learning ITF-WPI model based on cross-modal feature fusion of image and text encoders is proposed for the recognition of 17 types of pests of wolfberry using both image and text scenes. It is demonstrated that the introduced cosine annealing hot restart adjustment strategy algorithm, SGDR optimizer and early stop method help the model to obtain better classification performance and shorten the training time. The proposed pest recognition model ITF-WPI has a recognition accuracy of 97.98% with parallel image and text inputs, and the MACCs of the model are 7.828 G with only 52.20 M parameters. Convergence was significantly accelerated by migration learning, with a 5% reduction in the validation loss value and a 21.5% improvement in training accuracy. The PSA attention mechanism possessed better performance, as it improved the accuracy by at least 1.72%. The ablation experiments show that the CoT embedded in the COTN network is the main contributor to the performance of ITF-WPI, as it is based on and improves the Transformer structure and improves the accuracy by 2.31% relative to using only the PSA attention mechanism. In addition, CoTN performs best among the nine SOTA models, outperforms the current research hotspot Transformer neural network in terms of all- around performance, and is less computationally intensive in the same level of accuracy model. The ODLS model is simple in structure and performs best at the same level as state-of-the-art CNN-LSTM neural network studies. On the other hand, the validation of the ODLS memory layer stacked with more BiLSTM and Dropout makes the accuracy lower, which integrally justifies the ODLS structure. Therefore, the proposed cross-modal ITF-WPI shows excellent potential in complex background wolfberry pest identification.

Although image-based methods have been widely used for disease or pest detection, certain limitations remain (Huang et al., 2022; Coulibaly et al., 2022; Bao et al., 2022). Images alone may not always provide sufficient context or detailed information about pest species, life cycle stages, or potential damage. With the ODLS network implementation proposed in this study to process the textual description information of images for feature extraction, it is experimentally demonstrated that textual information can fill these gaps and provide additional insights to improve the accuracy and interpretability of model predictions. The best application scenario for the ITF-WPI model applies to scenarios where both pest images and text data are readily available or can be easily obtained. While it may be relatively simple to obtain pest images in agricultural application scenarios, however, in many agricultural or horticultural contexts, pest management databases, scientific literature, pest control guidelines, or expert knowledge sources often provide textual information about pests. Thus, our proposed cross-modal ITF-WPI model is advantageous when image and textdata are accessible and can be combined to achieve more accurate and comprehensive pest identification.

The study 's objective was to develop a model that effectively integrates image and text data to improve the identification of wolfberry pests. By exploiting the complementary information of the two modalities, the aim is to improve the accuracy and applicability of pest identification and provide valuable insights for pest management in wolfberry cultivation. In future work, we plan to research the diversified

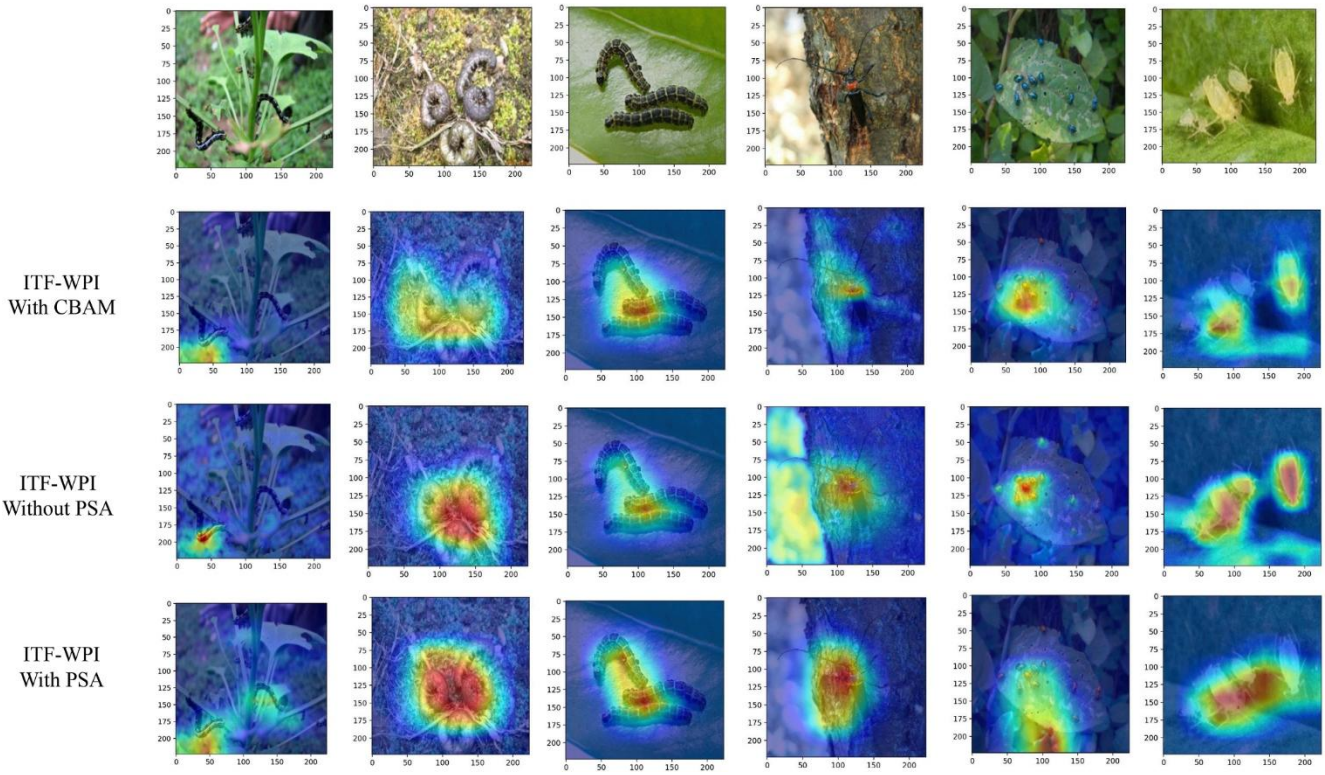


Fig. 12. Heat map comparing ITF-WPI incorporating CBAM, ITF-WPI incorporating PSA and ITF-WPI model without attention mechanism in different wolfberry pest identification.

modal information based on deep learning models in agriculture.

CRediT authorship contribution statement

Naihao Xu: Conceptualization, Methodology, Validation, Investigation, Data curation, Writing – original draft. **Hui Deng:** Writing – review & editing, Formal analysis, Project administration. **Meijun Sun:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing. **Zhiliang Qin:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Agriculture Science Data Center (NASDC2023XM00-05) and the National Natural Science Foundation of China (62076180). The authors thank the funding agency for financial support. The authors would also like to thank all authors and anonymous reviewers cited in this paper for their helpful comments and suggestions.

References

Bao, W., Cheng, T., Zhou, X.-G., Guo, W., Wang, Y., Zhang, X., Qiao, H., Zhang, D., 2022. An improved DenseNet model to classify the damage caused by cotton aphid. Comput. Electron. Agric. 203, 107485 https://doi.org/10.1016/j.compag.2022.107485.

Batchuluun, G., Choi, J., Park, K.R., 2023. CAM-CAN: Class activation map-based categorical adversarial network. Expert Syst. Appl. 222, 119809 https://doi.org/10.1016/j.eswa.2023.119809.

Chen, L., Weng, T., Xing, J., Li, Z., Yuan, Z., Pan, Z., Tan, S., Luo, R., 2021. Employing deep learning for automatic river bridge detection from SAR images based on Adaptively effective feature fusion. Int. J. Appl. Earth Obs. Geoinf. 102, 102425 https://doi.org/10.1016/j.jag.2021.102425.

Chodey, M.D., Noorullah Shariff, C., 2023. Pest detection via hybrid classification model with fuzzy C-means segmentation and proposed texture feature. Biomed. Signal Process. Control 84, 104710. https://doi.org/10.1016/j.bspc.2023.104710.

Coulibaly, S., Kamsu-Foguem, B., Kamissoko, D., Traore, D., 2022. Explainable deep convolutional neural networks for insect pest recognition. J. Clean. Prod. 371, 133638 https://doi.org/10.1016/j.jclepro.2022.133638.

Dai, G., Fan, J., Tian, Z., Wang, C., 2023. PPLC-Net: Neural network-based plant disease identification model supported by weather data augmentation and multi-level attention mechanism. J. King Saud University – Comput. Inform. Sci. 101555 https://doi.org/10.1016/j.jksuci.2023.101555.

Goyal, A., Bochkovskiy, A., Deng, J., Koltun, V., 2022. Non-deep networks. Adv. Neural Inf. Process. Sys. 35, 6789–6801.

He, C., Qiao, Y., Mao, R., Li, M., Wang, M., 2023. Enhanced LiteHRNet based sheep weight estimation using RGB-D images. Comput. Electron. Agric. 206, 107667 https://doi.org/10.1016/j.compag.2023.107667.

Huang, M.-L., Chuang, T.-C., Liao, Y.-C., 2022. Application of transfer learning and image augmentation technology for tomato pest identification. Sustainable Comput. Inf. Syst. 33, 100646 https://doi.org/10.1016/j.suscom.2021.100646.

Ijaz, A., Raza, B., Kiran, I., Waheed, A., Raza, A., Shah, H., Aftan, S., 2023. Modality specific CBAM-VGGNet model for the classification of breast histopathology images via transfer learning. IEEE Access 11, 15750–15762. https://doi.org/10.1109/ACCESS.2023.3245023.

Lee, J., Lee, P., Park, S., Byun, H., 2023. Expert-guided contrastive learning for video-text retrieval. Neurocomputing 536, 50–58. https://doi.org/10.1016/j.neucom.2023.03.022.

Li, X., Wang, W., Hu, X., Yang, J., 2019. Selective kernel networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 510–519.

Li, Y., Yao, T., Pan, Y., Mei, T., 2023. Contextual transformer networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 45 (2), 1489–1500. https://doi.org/10.1109/TPAMI.2022.3164083.

Liu, Y., Liu, S., Xu, J., Kong, X., Xie, L., Chen, K., Liao, Y., Fan, B., Wang, K., 2022. Forest pest identification based on a new dataset and convolutional neural network model with enhancement strategy. Comput. Electron. Agric. 192, 106625 https://doi.org/10.1016/j.compag.2021.106625.

Loshchilov, I., Hutter, F.. SGDR: Stochastic Gradient Descent with Warm Restarts. https://doi.org/10.48550/arXiv.1608.03983.

Ma, X., Li, Y., Wan, L., Xu, Z., Song, J., Huang, J., 2023b. Classification of seed corn ears based on custom lightweight convolutional neural network and improved training

strategies. Eng. Appl. Artif. Intel. 120, 105936 https://doi.org/10.1016/j.engappai.2023.105936.

Ma, J., Wang, L., Zhang, Y.-R., Yuan, W., Guo, W., 2023a. An integrated latent Dirichlet allocation and Word2vec method for generating the topic evolution of mental models from global to local. Expert Syst. Appl. 212, 118695 https://doi.org/10.1016/j.eswa.2022.118695.

Morid, M.A., Borjali,A., Del Fiol, G., 2021. A scoping review of transfer learning research on medical image analysis using ImageNet. Comput. Biol. Med. 128, 104115 https://doi.org/10.1016/j.compbiomed.2020.104115.

Nigam, S., Jain, R., Marwaha, S., Arora, A., Haque, M.A., Dheeraj, A., Singh, V.K., 2023. Deep transfer learning model for disease identification in wheat crop. Eco. Inform. 75, 102068 https://doi.org/10.1016/j.ecoinf.2023.102068.

Ozcanli, A.K., Baysal, M., 2022. Islanding detection in microgrid using deep learning based on 1D CNN and CNN-LSTM networks. Sustainable Energy Grids Networks 32, 100839. https://doi.org/10.1016/j.segan.2022.100839.

Sudhesh, K.M., Sowmya, V., Kurian, S., Sikha, O.K., 2023. AI based rice leaf disease identification enhanced by Dynamic Mode Decomposition. Eng. Appl. Artif. Intell. 120, 105836. https://doi.org/10.1016/j.engappai.2023.105836.

Thakur, P.S., Khanna, P., Sheorey, T., Ojha, A., 2022. Trends in vision-based machine learning techniques for plant disease identification: a systematic review. Expert Syst. Appl. 208, 118117 https://doi.org/10.1016/j.eswa.2022.118117.

Toh, D.W.K., Xia, X., Sutanto, C.N., Low, J.H.M., Poh, K.K., Wang, J.-W., Foo, R.-S.-Y., Kim, J.E., 2021. Enhancing the cardiovascular protective effects of a healthy dietary pattern with wolfberry (Lycium barbarum): A randomized controlled trial. Am. J. Clin. Nutr. 114 (1), 80–89. https://doi.org/10.1093/ajcn/nqab062.

Verma, H., Mandal, S., Gupta, A., 2022. Temporal deep learning architecture for prediction of COVID-19 cases in India. Expert Syst. Appl. 195, 116611 https://doi.org/10.1016/j.eswa.2022.116611.

Vidović, B.B., Milinčić, D.D., Marčetić, M.D., Djuriš, J.D., Ilić, T.D., Kostić, A.Ž., Pešić, M. B., 2022. Health benefits and applications of Goji berries in functional food products development. A review. Antioxidants 11 (2), Article 2. https://doi.org/10.3390/antiox11020248.

Wang, S., Qu, Z., Li, C., Gao, L., 2023. BANet: Small and multi-object detection with a bidirectional attention network for traffic scenes. Eng. Appl. Artif. Intel. 117, 105504 https://doi.org/10.1016/j.engappai.2022.105504.

Wang, F., Rao, Y., Luo, Q., Jin, X., Jiang, Z., Zhang, W., Li, S., 2022. Practical cucumber leaf disease recognition using improved Swin Transformer and small sample size. Comput. Electron. Agric. 199, 107163 https://doi.org/10.1016/j.compag.2022.107163.

Wang, C., Zhou, J., Zhao, C., Li, J., Teng, G., Wu, H., 2021. Few-shot vegetable disease recognition model based on image text collaborative representation learning. Comput. Electron. Agric. 184, 106098 https://doi.org/10.1016/j.compag.2021.106098.

Wenli, S., Shahrajabian, M.H., Qi, C., 2021. Health benefits of wolfberry (Gou Qi Zi, Fructus barbarum L.) on the basis of ancient Chineseherbalism and Western modern medicine. Avicenna Journal of Phytomedicine 11 (2), 109–119. https://doi.org/10.22038/ajp.2020.17147.

Wu, T., Lu, J., Zou, J., Chen, N., Yang, L., 2022. Accurate prediction of salmon freshness under temperature fluctuations using the convolutional neural network long short-term memory model. J. Food Eng. 334, 111171 https://doi.org/10.1016/j.jfoodeng.2022.111171.

Wu, D., Ying, Y., Zhou, M., Pan, J., Cui, D., 2023. Improved ResNet-50 deep learning algorithm for identifying chicken gender. Comput. Electron. Agric. 205, 107622 https://doi.org/10.1016/j.compag.2023.107622.

Yajun, W., Xiaojie, L., Sujuan, G., Yuekun, L., Bo, Z., Yue, Y., Wei, A., Youlong, C., Jianhua, Z., 2019. Evaluation of nutrients and related environmental factors for wolfberry (Lycium barbarum) fruits grown in the different areas of China. Biochem. Syst. Ecol. 86, 103916 https://doi.org/10.1016/j.bse.2019.103916.

Yang, T., Hu, Y., Yan, Y., Zhou, W., Chen, G., Zeng, X., Cao, Y., 2022. Characterization and Evaluation of Antioxidant and Anti-Inflammatory Activities of Flavonoids from the Fruits of Lycium barbarum. Foods 11(3), Article 3. https://doi.org/10.3390/foods11030306.

Yang, X., Shu, L., Chen, J., Ferrag, M.A., Wu, J., Nurellari, E., Huang, K., 2021. A survey on smart agriculture: development modes, technologies, and security and privacy challenges. IEEE/CAA J. Autom. Sin. 8 (2), 273–302. https://doi.org/10.1109/JAS.2020.1003536.

Ye, Y., Huang,Q., Rong, Y., Yu, X., Liang, W., Chen, Y., Xiong, S., 2023. Field detection of small pests through stochastic gradient descent with genetic algorithm. Comput. Electron. Agric. 206, 107694 https://doi.org/10.1016/j.compag.2023.107694.

Yu, S., Xie, L., Huang, Q., 2023. Inception convolutional vision transformers for plant disease identification. Internet of Things 21, 100650. https://doi.org/10.1016/j.iot.2022.100650.

Ramachandran, P., Zoph, B., Le, Q.V., 2017. Searching for Activation Functions. https://doi.org/10.48550/arXiv.1710.05941.

Zhang, H., Zu, K., Lu, J., Zou, Y., Meng, D. EPSANet: An Efficient Pyramid Squeeze Attention Block on Convolutional Neural Network 1161–1177. https://doi.org/10.48550/arXiv.2105.14447.

Zhang, W., Zhou, H., Bao, X., Cui, H., 2023. Outlet water temperature prediction of energy pile based on spatial-temporal feature extraction through CNN–LSTM hybrid model. Energy 264, 126190. https://doi.org/10.1016/j.energy.2022.126190.

Zhou, C., Zhong, Y., Zhou, S., Song, J., Xiang, W., 2023. Rice leaf disease identification by residual-distilled transformer. Eng. Appl. Artif. Intel. 121, 106020 https://doi.org/10.1016/j.engappai.2023.106020.

Zhu, H., Gu, W., Wang, L., Xu, Z., Sheng, V.S., 2023. Android malware detection based on multi-head squeeze-and-excitation residual network. Expert Syst. Appl. 212, 118705 https://doi.org/10.1016/j.eswa.2022.118705.