GPPK4PCM: pest classification model integrating growth period prior knowledge

Jianhua Zheng,[1,2,3] Junde Lu,[1] Yusha Fu,1 Ruolin Zhao,[1] JinFang Liu,[1] ZhaoXi Luo,[1] Zhijie Luo[1,2,3]

[1]College of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou; [2]Guangzhou Key Laboratory of Agricultural Products Quality & Safety Traceability Information Technology Zhongkai University of Agriculture and Engineering, Guangzhou; [3]Smart Agriculture Innovation Research Institute, Zhongkai University of Agriculture and Engineering, Guangzhou, China

Abstract

Recent advancements in computer vision technology have significantly improved pest classification. However, pests of the same species exhibit distinct morphological changes throughout different life periods. Traditional methods apply the same feature extraction techniques to all periods, limiting classification precision. In addition to its inherent visual characteristics, pest images contain implicit growth period information. To address this issue, we propose a Pest Classification Model Integrating Growth Period Prior Knowledge. The model is composed of three sub-modules where: i) a deep learning network first identifies the growth periods of pests, and this prior knowledge is then used to guide the text encoder of the CLIP pre-trained model in generating period-specific textual features; ii) a parallel deep learning network extracts visual features from pest images; iii) an efficient low-rank multimodal fusion module integrates textual and visual features through parameter-optimized tensor decomposition, significantly improving classification accuracy across pest developmental phases. To evaluate its effectiveness, a dataset containing pests at different growth periods was constructed from Sichuan Agricultural University's pest dataset. Experimental results show that GPPK4PCM outperforms well-established deep learning neural networks. Compared to other advanced models, the proposed model excels in pest and disease classification tasks, effectively handling significant morphological differences across life periods.

Correspondence: Zhijie Luo, College of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou, 510225, China. E-mail: luozhijie@zhku.edu.cn

Key words: deep learning; pest classification; prior knowledge; multimodel.

Received: 14 April 2025.
Accepted: 15 June 2025.

©*Copyright: the Author(s), 2025*
Licensee PAGEPress, Italy
Journal of Agricultural Engineering 2025; LVI:1814
doi:10.4081/jae.2025.1814

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Publisher's note: all claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Introduction

In the process of agricultural production, pests are recognized as having a significant and widespread impact on crops. Crop growth is not only directly damaged , resulting in reduced yield, but also indirectly influence agricultural product quality and potentially transmit plant diseases. Hence, it is essential to promptly classify and precisely identify pests to avert the spread of their damaging effects.

This issue has garnered considerable interest from researchers, with many recently adopting deep learning technology for pest classification and identification. Liu *et al.* (2020) proposed the DFF-ResNet for performing pest identification tasks. They developed a residual network that integrates deep feature fusion by adding branches to the original residual blocks and stacking feature fusion residual blocks with early residual groups. This network exceeds the performance of both the original ResNet and other leading techniques. Guo *et al*. (2024) introduced pest image classifiers designed for open-world scenarios that leverage features learned from previous pest categories to identify new ones. To prevent the model from collapsing, a normalized cross-entropy loss scaled by temperature was used, and a trained ResNet8 matching network was employed to assess the similarity between support for correlation of query image features and class prototyping, promoting effective pest identification. This approach reached a peak accuracy of 84.29% on D0, utilizing a 40-way 5-shot support set. Setiawan *et al*. (2022) introduced an effective training framework to improve the compact MobileNetV2 model's performance by using dynamic learning rates, CutMix augmentation, frozen layers, and sparse regularization, they achieved a top accuracy of 71.32% by integrating these techniques during training. Many of the studies mentioned use identical deep-learning networks for feature extraction from pest images. Yet, significant morphological variations among pests of the same species at different life periods hinder accurate classification.

Extensive research has shown that integrating prior knowledge beyond images into deep learning models can enhance their performance. Deng *et al*. (2022) introduced a novel model named VSGCN multi-label image classification (MLIC) by reducing redundancy in traditional word embeddings through the use of visual and semantic prototypes. VSGCN utilizes a multi-head GCN approach to build a label correlation graph and model label correlations in different subspaces, reducing inconsistent predictions across visual and semantic spaces. Extensive experiments have demonstrated the superior performance of VSGCN on multi-label image datasets. Lu *et al.* (2022) introduced an effective model for segmenting detailed root images, Regions of interest were utilized and incorporated prior knowledge into CNNs. Incorporating prior knowledge minimized background mislearning, with an average F1 Score of over 90%, resulting in effective root feature segmentation. Bai *et al*. (2024) combined the benefits

of prior knowledge-based approaches with deep learning techniques, proposing the APFS method, which combines prior knowledge of modulation tasks with feature information obtained through contrast learning. Feature extraction guided by prior knowledge accurately captures key patterns, while contrast learning reveals intrinsic differences between various modulation patterns. Experimental results show that APFS demonstrates superiority in various baseline and combined performance comparisons. These studies indicate that introducing domain knowledge or existing experience into models can enhance their understanding and thereby improve the accuracy of their predictions. Nevertheless, extracting prior information from pest images presents a significant challenge in this study.

To effectively address the challenge of pest classification across different growth periods, we proposed an innovative a priori information fusion method GPPK4PCM, namely pest classification model integrating growth period prior knowledge, as illustrated in Figure 1. GPPK4PCM is composed of three core components: the text feature extraction module (TFEM), the image feature extraction module (IFEM), and the feature fusion module (FFM). In TFEM, we first apply the classical ResNet50 to analyze pest images and estimate their corresponding growth periods. These period labels are then converted into descriptive text, which is passed through the text encoder of a pre-trained CLIP model. This process generates semantic feature vectors that reflect the developmental period of each pest, providing structured prior knowledge for the model. At the same time, IFEM focuses on extracting detailed visual features from the pest images. It uses a backbone such as Xception to capture fine-grained morphological traits that are critical for accurate classification.

Directly adding features often limits the models capability, as it lacks effective interaction between modalities. Motivated by the low-rank multimodal fusion (LMF) framework, we adopt this approach to achieve more efficient integration of visual and text features. Through this design, the model can better leverage growth period knowledge to support image recognition, thereby improving classification performance across various pest development periods.

This paper's primary contributions are:

- To address the limitations of existing pest identification methods, we proposed the GPPK4PCM, an advanced approach aimed at comprehensively leveraging the growth-period-specific morphological information of pests to enhance the recognition capability of pest classification models.
- In GPPK4PCM, we introduce an innovative scheme to extract prior information, which is subsequently integrated with visual features obtained through conventional methods using an efficient multimodal feature fusion module.
- Based on the pests dataset from Sichuan Agricultural University (Sichuan Agriculture University, 2020), we created a dataset that included the larval and adult periods of each pest. Experimental verification was conducted, and the results have shown that GPPK4PCM exhibits superior classification performance compared to methods that directly extract pest features for classification.

Related work

Pest image classification

Traditional plant disease and pest identification heavily rely on a limited number of experienced experts who physically inspect fields to search for pests or signs of their damage, and then differentiate them by characteristics such as color, shape, and size. This method is both expensive and inefficient. With recent advancements in machine learning and computer vision, researchers are now applying techniques including SVM, decision trees, and others to automatically identify pests. Kasinathan *et al*. (2020) used shape features and various machine learning techniques to conduct classification experiments on 9 and 24 classes of insects. The algorithm successfully identifies insects in complex backgrounds through foreground extraction and contour recognition, achieving classification rates of 91.5% and 90% with the CNN model. Compared with traditional methods, this algorithm has significantly improved in classification accuracy and computational efficiency, for early insect identification to enhance crop yield and quality. Tuda *et al*. (2021) applied machine learning methods to distinguish between the gender and species of stored product pests, including beetles and their parasites. They achieved classification accuracy rates ranging from 88.5% to 98.5%. This research highlighted that integrating object-level and pixel-level features notably improves classification performance, marking it as one of the pioneering studies to identify insect gender from static images. Ebrahimi *et al*. (2017) applied the SVM classification technique to identify thrips on strawberry canopy images. By utilizing innovative image processing methods and differential kernel functions, they performed classification based on area and color indices. The evaluation results showed that the SVM method was the most effective for classifying thrips, with an average percentage error of less than 2.25%. While the aforementioned methods based on machine learning have achieved impressive levels of accuracy, the image processing involved in machine learning can be cumbersome. Data annotation is a time-consuming task, and the extracted features may not be comprehensive enough, which can potentially impact the classification performance of trained models. Additionally, many of the extracted features are task-specific and dataset-specific, leading to limited compatibility and generalizability.

However, recent years have seen substantial progress in artificial intelligence and deep learning technologies, which are now extensively used across different tasks. These technologies have notably improved the accuracy of classification and recognition tasks while steadily reducing error rates. Unlike traditional approaches, deep learning technologies do not rely on manual feature extraction, but instead autonomously learn from well-labeled datasets, effectively capturing advanced features within the data. Deep learning technologies have shown outstanding effectiveness in tasks including image classification. They excel in tasks such as object and scene recognition, object detection -which involves locating and classifying multiple objects within an image- and semantic segmentation, which offers a more granular understanding by assigning a class label to each individual pixel. In the field of crop recognition, CNNs had particularly become the most widely used models. Prominent architectures of CNNs include AlexNet (Krizhevsky *et al*., 2012), VGG (Simonyan *et al.,* 2014)·ResNet (He *et al*., 2016), Inception (Szegedy *et al*., 2015), MobileNet (Howard. *et al*., 2017), GhostNet (Han *et al*., 2020), among others.

Many agricultural professionals have begun utilizing these technologies to analyze crop images. In comparison to conventional manual and mechanical methods, Artificial intelligence technology excels in identification efficiency and accuracy, offering a better approach for pest image classification in agriculture. Khanramaki *et al*. (2021) proposed an advanced approach leveraging deep learning techniques, employing an ensemble model of the CNN utilized for the identification of three citrus pests. The study was assessed using a dataset of 1,774 images of citrus leaves, applying data augmentation and 10-fold cross-validation techniques. The findings indicated that the proposed ensemble classifi-

er reached an accuracy of 99.04% across various conditions, surpassing other comparable methods. Wei *et al*. (2021) introduced a crop pest classification approach utilizing Multi-Scale Feature Fusion (MFFNet). The method extracts multi-scale and deep features from pest images using dilated convolution and integrates them to ensure comprehensive and precise classification and recognition. Evaluated on a dataset with 12 pest categories, this approach achieved a classification accuracy of 98.2% and proved highly effective. Albattah *et al*. (2023) proposed an automated system leveraging deep learning and drone technology to identify and classify crop pests. This system integrates DenseNet-100 with CornerNet and is organized into three phases: extracting regions of interest, performing deep keypoint detection, and classifying pests. Tests conducted on the IP102 dataset demonstrated that this approach significantly enhanced on-site recognition accuracy and recall rate. Wang *et al.* (2024)introduced an innovative approach named InsectMamba to tackle the challenges of pest recognition. InsectMamba integrates multi-head self-attention (MSA), convolutional neural networks (CNN), state space models (SSM), and multi-layer perceptrons (MLP) to form a Mix-SSM block for extracting comprehensive visual features. A selective module is used to adaptively aggregate these features, thereby enhancing the model's discriminative power. Evaluation results on five pest classification datasets showed that InsectMamba performed excellently and validated the importance of each model component through ablation studies.

The researchers mentioned above have focused on studying different methods of extracting features from pest images. However, they have overlooked the important aspect of temporal information in the growth cycle of pests and how their features change during various growth periods.

Contrastive language-image pretraining

The CLIP model from OpenAI is a revolutionary multimodal model designed to embed both images and text into the same representational space through contrastive learning methods, thereby achieving efficient image-text matching and understanding, demonstrating its cross-modal matching capabilities.

Recently, substantial advancements have been achieved in various downstream tasks. Yi *et al.* (2023) proposed a novel feature extraction technique named CODER to tackle the variable performance of the CLIP model in unimodal feature extraction. CODER treats text features as precise neighborhoods of image features, leveraging the distance structure between images and neighboring text to enhance the quality of feature representation. To construct high-quality CODER, this paper introduces an automatic text generator that can generate diverse texts matching images without data and training. Experimental results of CODER across different datasets and models in both zero-shot and few-shot image classification scenarios confirm its effectiveness. Li *et al*. (2023) investigated the application of the CLIP model for fine-grained image re-identification (ReID) tasks. They found that by adjusting only the visual model within the CLIP image encoder, it is possible to achieve competitive results across various ReID scenarios. The paper introduces a two-phase approach: initially, the image and text encoders of CLIP remain unchanged, focusing on optimizing the learnable text tokens; subsequently, ID-specific text tokens and encoders are fixed to refine the image encoder. This strategy has been validated and improved CLIP's performance in person and vehicle ReID tasks. Lin *et al.* (2023) introduced CLIP-ES, a weakly supervised semantic segmentation framework utilizing the CLIP pre-trained model, which operates with just image-level labels and does not require additional training. It introduces the softmax function in GradCAM, using CLIP's zero-shot capabilities to suppress background confusion, and customizes text-driven strategies. By employing the multi-head self-attention mechanism from CLIP-ViTs, and proposed the class-aware attention affinity module to enhance CAM. Additionally, a confidence-guided loss function is incorporated during training. CLIP-ES has reached top performance levels on different datasets and has shortened the time needed to produce pseudo masks. The studies mentioned above indicate that the text features produced by improving the model's comprehension, the CLIP model also reveals its broad usefulness in a range of practical applications. Moreover, the text features of CLIP enhance the model's capacity to express itself and adapt, thus allowing it to excel in various complex tasks.

Multimodal learning

In the real world, information manifests in multiple formats, including images, text, audio, and video. Although unimodal learning has been successful in many tasks. However, challenges remain due to insufficient information for certain tasks. For instance, image data provides visual information but lacks semantic and contextual relationships. On the other hand, text data offers in-depth semantic information but lacks visual features.

Multimodal learning is a method that combines two or more modalities of information for joint learning and analysis. By aligning modalities, fusing them, and generating cross-modal representations, different types of data can be utilized to accomplish specific tasks. This fusion approach exploits the strengths of each modality, compensates for their shortcomings, and enables models to obtain more comprehensive information, thereby enhancing generalization performance and effectiveness on complex tasks. Consequently, multimodal learning has emerged as a prominent area of research. Dai *et al.* (2023) proposed the ITF-WPI, integrating both image and text data processing. The model includes CoTN and ODLS components, which process images and text respectively. By integrating transformer structures and pyramid squeeze attention (PSA), CoTN improves the capability to capture multi-scale features. ODLS employs 1D convolutions and bidirectional LSTM stacking to bolster text feature extraction. The experimental results confirm that the ITF-WPI model has surpassed other advanced models in terms of accuracy, achieving a high accuracy rate of 97.98%. Zhou *et al*. (2021) developed a multimodal identification technique for diseases, ITK-Net, that employs semantic embeddings of visual and textual data for joint representation learning, directed by a high-level domain knowledge graph. The research subjects are common infectious diseases of tomatoes and cucumbers. The 'image-text' dataset results for ITK-Net are impressive, reporting an accuracy rate of 99.63%, along with precision, sensitivity, and specificity of 99%, 99.07%, and 99.78%, respectively. This method improves the credibility and interpretability of disease identification, providing an intelligent solution for crop disease identification. In their research, Zhang *et al.* (2023) developed the MMFGT mode for identifying pests. The model leverages self-supervised and contrastive learning to refine the transformer framework, thus minimizing the dependency on large-scale data. Additionally, the model integrates fine-grained recognition features, focusing on the nuances of image variations, and amalgamates multimodal data from visuals and natural language descriptions to boost the precision of recognition. Experimental results indicate that MMFGT excelled in pest identification tasks, achieving an identification accuracy rate of 98.12%, which is a 5.92% improvement over the leading DINO method.

The above studies suggest that multimodal learning effectively compensates for the limitations of models that rely on a single

modality by integrating information from diverse modalities. Among these methods, integrating multimodal data from images and text, with the guidance of text to acquire features, improves the model's classification abilities. However, these methods often involve introducing external, independent text information to directly guide the feature acquisition process of the model.

Materials and Methods

In this section, we provide a detailed description of the proposed GPPK4PCM method. This approach is designed to leverage the implicit prior knowledge of pest growth periods in images, with the aim of enhancing the model's classification accuracy. The architecture and workflow of the proposed model, detailing of the image feature extraction module (IFEM) and the text feature extraction module (TFEM), followed by a description of the feature fusion module (FFM) are described in the following chapters. An overview of the GPPK4PCM framework is illustrated in Figure 1.

GPPK4PCM: pest classification model integrating growth period prior knowledge

This study presents a pest classification model that integrates prior knowledge of the pest growth period, as illustrated in Figure 1. The model, which relies solely on pest imagery for data input, is composed of three essential modules: the IFEM for capturing insect visual features, the TFEM for textual pest's growth phase features, and the FFM for integrating features. In this approach, each pest image is processed using two distinct feature extraction methods. Firstly, the IFEM is utilized to extract high-dimensional feature vectors from the input images through a deep learning model. In this study, Xception (Chollet *et al*., 2017) is employed for image feature extraction. Secondly, the TFEM first utilizes any deep learning model to identify the growth period of the pest. In this study, the ResNet50 model is adopted. Subsequently, the text encoder of the CLIP pre-trained model is employed to obtain the textual feature vector representing the pest's growth period. Following this, the FFM is employed to effectively integrate the image feature and text feature vectors to generate a comprehensive vector that encompasses information on the pest's category and period. Finally, precise classification of pests at various growth periods is achieved through a fully connected layer.

Pest image feature extraction module

The IFEM primarily utilizes deep learning models to extract feature vectors from pest images. It is not restricted to a specific network architecture. It can employ various convolutional neural network architectures, such as Xception, ResNet50, EfficientNet (Tan *et al*., 2019), GhostNet, InceptionV4 (Szegedy *et al*., 2017), or the Transformer-based Vision Transformer (Dosovitskiy *et al.,* 2020). Each of these networks has unique characteristics and demonstrates strong feature extraction performance in different application scenarios. Any of these networks can be applied within the IFEM module. The selection of the most suitable network is determined based on specific task requirements and the characteristics of the input data. This ensures efficient and accurate feature extraction, providing precise image features for subsequent multimodal feature fusion and pest classification.

To conveniently represent the process of extracting image features using IFEM, the deep learning model is employed for feature extraction:

$$z = f_{Network}(X) \quad \text{(Eq. 1)}$$

The Xception network introduces depthwise separable convolution technology, which efficiently extracts subtle differences and complex textures in images while reducing computation and parameters. In the following sections, Xception is selected as the backbone network for extracting image feature vectors within the IFEM module. The process of extracting image features of pests using the Xception network is represented:

$$z_{Image} = f_{xception}(X) \quad \text{(Eq. 2)}$$

Pest period text feature extraction model

At the moment, standard approaches to pest image classification primarily rely on identifying visual characteristics -such as form, color, and physical structure- from photographs. However, semantic information in the images, such as the growth period of the pests (e.g., larval, adult, or other developmental periods), is often overlooked. To address this issue, the TFEM utilizes the text encoder of the CLIP pre-trained model to extract prior knowledge about the growth periods of pests. This approach significantly improves the accuracy of subsequent pest classification. At the outset, the TFEM leverages a deep learning framework to determine the pest growth phase. The account of the pest's growth phase is

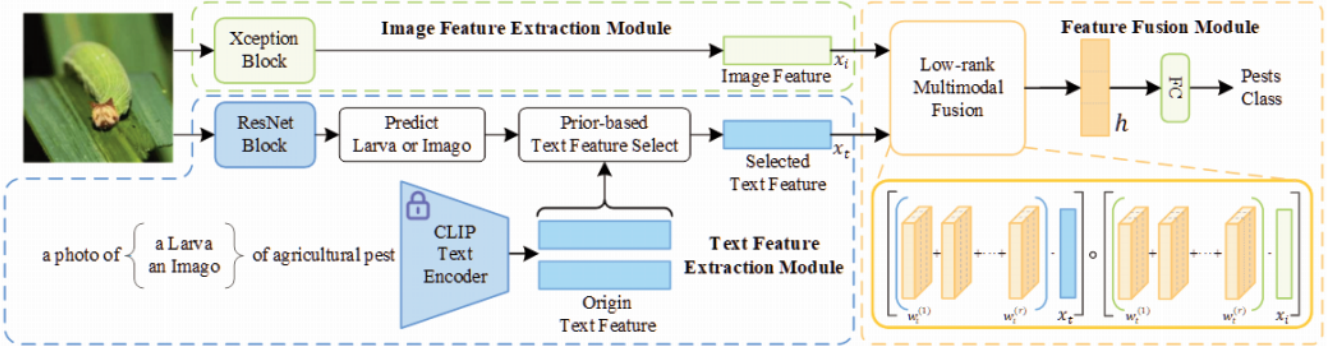

Figure 1. The overall architecture of our GPPK4PCM model.

subsequently encoded by the text encoder of the CLIP model, pre-trained to extract a feature vector indicative of the pest's growth period. In this method, any deep learning network model can be used for identifying the growth period of pests. Considering the exceptional performance of ResNet50 in terms of image feature extraction efficiency and classification accuracy, this paper chooses the ResNet50 network as the model for pest growth period identification. After ResNet50 has isolated the temporal features indicative of the pests, the conclusive period of the pests' development is ascertained by employing a fully connected layer and concluding with a Softmax layer. The procedure can be summarized in the subsequent steps:

$$Z_{Stage} = f_{ResNet50}{}^{(X)} \quad \text{(Eq. 3)}$$

$$\hat{c} = arg\,max\left(softmax\left(\boldsymbol{W}_{fc} \cdot z_{Stage} + \boldsymbol{b}_{fc}\right)\right) \quad \text{(Eq. 4)}$$

The CLIP model, a pre-trained model, accomplishes joint learning of image and text features using a contrastive learning approach. This approach ensures that both image and text features are highly similar in the same vector space, thus enabling cross-modal feature extraction. We specify the prompt template for CLIP as "A photo of {object} of an agricultural pest", where 'object' refers to both the larva and adult periods of the pest. Subsequently, we employ the CLIP model's text encoder, which is pre-trained based on ViT-B-16, to extract text features. This process ultimately yields a 1024-dimensional text feature vector for each 'object' representing the pest period:

$$Z_{Text} = f_{CLIP}{}^{(T)} \quad \text{(Eq. 5)}$$

where T represents the prompt template, $\mathcal{Z}_{\mathbf{T}} \in \mathbb{R}^{C \times d}$ represents the feature matrix of text T, and C is the number of categories, with each category having a d-dimensional feature vector.

The ResNet50 model identifies the period of the pest in the image and selects one out of the 1024 different pest period text feature vectors. Ultimately, it obtains a single-dimensional 1024 feature vector that describes the period. The vector functions as a text feature vector that holds data pertaining to the pest's growth period, corresponding to the provided image. The following steps detail the process:

$$z_{Text,\hat{c}} = z_{Text}[\hat{c}, :] \quad \text{(Eq. 6)}$$

Low-rank multimodal feature fusion module

The aim of multi-modal fusion is to combine various modalities in order to exploit the complementary nature of the data, thereby providing more powerful predictions. In the IFEM and TFEM modules, we obtain the image feature vector and text feature vector of the input pest image, capturing the key information of their respective modalities. After extracting the features of each modality, it is necessary to effectively fuse these features. Modal fusion can be achieved through various methods, with direct element-wise addition of features from various modules being one of the common approaches. The calculation method for element-wise addition of the extracted image and text features is as follows:

$$\mathcal{Z} = z_{Image} + z_{Text,\hat{c}} \quad \text{(Eq. 7)}$$

Element-wise addition for feature fusion is a method that has low computational complexity and is easy to implement. However, despite its simplicity in calculation, this method has limited feature representation capability as it overlooks the more complex interactive relationships between modalities. The tensor fusion network (TFN) (Zadeh *et al*., 2017) is a network designed for multi-modal data integration, predominantly used in the domain of sentiment analysis for combining diverse data types. TFN introduces the concept of tensor fusion in an innovative manner by appending an additional dimension to the uni-modal representations before performing the tensor product, which is represented as. By conducting the vector outer product of each modality's tensor, it generates a Cartesian product space that effectively represents the multi-modality:

$$\mathcal{Z} = \bigotimes_{m=1}^{M} z_m, z_m \in \mathbb{R}^{d_m} \quad \text{(Eq. 8)}$$

where: M represents the number of different modalities, m is a specified modality, and denotes the outer product between vectors. Subsequently, a multi-modal representation is generated through a linear layer:

$$h = g(\mathcal{Z}; \mathcal{W}, b) = \mathcal{W} \cdot \mathcal{Z} + b \quad \text{(Eq. 9)}$$

where: W is a d_h tensor of order M+1, $\mathcal{W}_k \in \mathbb{R}^{d_1 \times \ldots \times d_M},\ k = 1, \ldots, d_h$. TFN calculates the correlation between two modalities, generating a higher-order tensor to capture the interactive information between the modalities while also preserving the information of each modality. This approach, when compared to simple concatenation or weighted averaging, demonstrates significant advantages in capturing complex interactive relationships between multiple modalities. However, TFN faces issues with computational efficiency and increased memory consumption due to the higher-order tensors and calculations. These issues become more prominent as the feature dimensionality increases.

To overcome the computational efficiency issues of TFN, Liu *et al.* (2018) introduced the Low-rank Multimodal Fusion network (LMF). In LMF, a fixed rank r and r decomposition factor parameters $\{\{w_{m,k}^{(i)}\}_{m=1}^{M}\}_{i=1}^{r}, k = 1, \ldots, d_h$ are set. For each modality m, its corresponding decomposition factor is $\{w_m^{(i)}\}_{i=1}^{r}$. Similar to TFN, we represents the additional dimension appended to the representation. For ease of representation, let it be denoted as $w_m^{(i)} =$ $[w_{m,1}^{(i)}, w_{m,2}^{(i)}, \ldots, w_{m,d_h}^{(i)}]$, and represented by the following formula for the low-rank weight tensor:

$$\mathcal{W} = \sum_{i=1}^{r} \bigotimes_{m=1}^{M} w_m^{(i)} \quad \text{(Eq. 10)}$$

the key to LMF lies in the fusion of parallel decomposition. Equation (10) decomposes W into M groups of specific modal factor matrices, which allows for parallel computation between the low-rank factors and the tensor Z. By inputting different modal tensors, multiple modalities can be derived through parallel decomposition calculations to obtain the multi-modal representation h:

$$h = \left(\sum_{i=1}^{r} \bigotimes_{m=1}^{M} w_m^{(i)}\right) \cdot z = \prod_{m=1}^{M} \left[\sum_{i=1}^{r} w_m^{(i)} \cdot z_m\right] \quad \text{(Eq. 11)}$$

In contrast to TFN, LMF employs low-rank parallel decomposition to project high-dimensional multi-modal information into a lower-dimensional space, thereby avoiding the direct computation of high-dimensional tensors. This reduces the computational complexity of tensor fusion from $O(d_y \prod_{m=1}^{M} d_m)$ to $O(d_y \times r \times \sum_{m=1}^{M} d_m)$, retaining the main information of the data while decreasing computational complexity.

In this paper, after inputting the extracted image and text feature vectors into the LMF module, we obtain the multi-modal representation:

$$h = \left(\sum_{i=1}^{r} w_{Image}^{(i)} \otimes w_{Text}^{(i)}\right) \cdot z = \left(\sum_{i=1}^{r} w_{Image}^{(i)} \cdot z_{Image}\right) \circ \left(\sum_{i=1}^{r} w_{Text}^{(i)} \cdot z_{Text,\hat{c}}\right) \quad \text{(Eq. 12)}$$

in this context, W_{Image} and W_{Text} represent the low-rank weight tensors corresponding to the image and text features, respectively. Given the image feature Z_{Image} from IFEM (Eq.2) and the text feature $z_{Text,\hat{c}}$ from TFEM (Eq. 6), the LMF module fuses them *via* parallel decomposition factor $w_{Image}^{(i)}$ and $w_{Text}^{(i)}$.

By utilizing the LMF module for feature fusion, the resulting feature vector encompasses both image information and the multi-modal representation of text information. This form of multi-modal representation enables a more comprehensive capture of pest characteristics, subsequently improving the accuracy and robustness of the classification process. In the final period, the feature vector is processed through tensor fusion and low-rank decomposition, subsequently advancing to a fully connected layer for categorizing pests into various developmental periods.

Loss function

Accurately quantifying the discrepancy between predicted results and ground-truth labels is essential for training the GPPK4PCM model, particularly its TFEM module. To address the dual tasks of pest classification and growth stage recognition, we designed task-specific loss functions that account for class imbalance. While the standard cross-entropy loss performs well under balanced class distributions, agricultural pest datasets often exhibit significant class imbalance, which is a common issue in deep learning-based image processing. Such imbalance can hinder the model's performance in recognizing categories with fewer samples. In real-world agricultural data collection, pest species vary greatly in population size due to environmental and seasonal factors. Some pests appear in large numbers during specific seasons, while others are nearly absent outside their peak periods. This seasonal fluctuation, coupled with the challenges of manual data collection, leads to extreme scarcity of certain categories in the dataset. To mitigate this issue and enhance the model's ability to learn from underrepresented classes, we adopt a weighted cross-entropy loss. Specifically, the class weight for the *i*-th pest category is computed using the following formula:

$$\omega_i = \frac{N}{n_{class} \cdot n_i} \quad \text{(Eq. 13)}$$

where: N is the total number of images in the dataset, n_{class} is the total number of classes, and n_j indicates the count of images within the *i*-th class. In addition, TFEM is designed to identify the growth stages of pests and extract corresponding textual features. To address the imbalance in growth stage data during the training of the ResNet50-based growth stage classifier, we adopt a similar class weighting strategy. The weight ω_j' for the *j*-th growth stage is defined as:

$$\omega_i' = \frac{N}{n_{stage} \cdot n_j} \quad \text{(Eq. 14)}$$

where: N denotes the total number of samples in the dataset, n_{stage} is the total number of growth stages, and n_j represents the number of samples belonging to the *j*-th stage. The corresponding weighted cross-entropy loss function is given by:

$$\mathcal{L}_{stage} = -\sum_{j=1}^{N} \omega_j' \cdot y_j' \cdot \log(q_j) \quad \text{(Eq. 15)}$$

where: y'_j denotes the ground-truth label, q_j is the predicted probability for *j*-th stage, and is the weight assigned to *j*-th stage.

Results

Dataset

This study is conducted based on the Sichuan Agricultural University Pest Dataset, which contains genuine images of 21 different pest classes for classification purposes. The dataset comprises images depicting the diverse periods of pest development, including the egg, larval, pupal, and adult periods. Nevertheless, as there is a scarcity of images for certain pests in the egg and pupa periods, only pest categories with images of the larval and adult periods were selected for classification in this experiment. Moreover, since some pest species had an extremely limited number of larvae or adult images, I gathered additional pictures from the internet to augment the dataset. Figure 2 shows examples from the dataset with distinct morphological differences in different growth periods of some pests. To address this class imbalance issue, we have implemented data augmentation. Data augmentation techniques are implemented for pest classes that have a restricted number of samples within the dataset to enhance their representation. The data augmentation strategies we use include randomly rotating the images by 90, 180, and 270 degrees to address the issue of limited images for each pest period. Next, the dataset is divided into training and test sets in a 7:3 ratio. The training set is further augmented by randomly combining techniques such as random cropping, random horizontal flipping, padding, random color jittering, and Gaussian blurring, resulting in a six-fold expansion. In the end, we used 23,200 images for training and 1,584 images for validation. By employing these strategies, we aim to overcome the challenges associated with class imbalance, enhancing the model's proficiency and its ability to generalize across classes with limited sample sizes.

Experimental environment

To guarantee fairness and consistency in the experimental results, all tests are performed under identical conditions. The operating system used for experiments is Ubuntu 20.04.6 LTS, with a CPU of AMD® Epyc 7452 32-core processor; the GPU is an NVIDIA RTX3090 GPU with 24G of memory, and the CUDA version is 12.2. The deep learning toolkit is PyTorch 2.2.0. The input image size is 224×224, the batch size is 32, the learning rate

is 0.001, the optimizer used is SGD, and the weight decay is 5e-4.

Model training

In the experimental section , pre-trained model weights are not utilized. Instead, the models are trained using the dataset provided in this paper. The primary focus is on analyzing the impact of incorporating prior knowledge of text features on classification accuracy. This design choice is intended to evaluate the effectiveness of the proposed method more objectively, particularly the impact of incorporating prior textual knowledge on classification accuracy. When training the GPPK4PCM, the ResNet50 model for pest growth period recognition is trained first.

Evaluation metrics

In the realm of classification tasks, the metric of Accuracy (Acc) stands as a pivotal measure of a model's efficacy. This metric reflects the ratio of samples accurately identified by the model relative to the overall sample count. The formula for calculating accuracy is presented below:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(Eq. 16)}$$

where: *TP* indicates the count of true positive instances correctly identified, *TN* indicates the true negative instances that were accurately recognized, *FP* indicates the false positive cases that were incorrectly identified, and *FN* indicates the false negative cases that were also incorrectly classified. The confusion matrix is indispensable for evaluation purposes, providing a straightforward view of the model's success in classifying data. In the confusion matrix, the predicted labels are displayed along the horizontal axis, while the true labels are shown on the vertical axis. Using this matrix, we can derive key performance metrics for classification tasks:

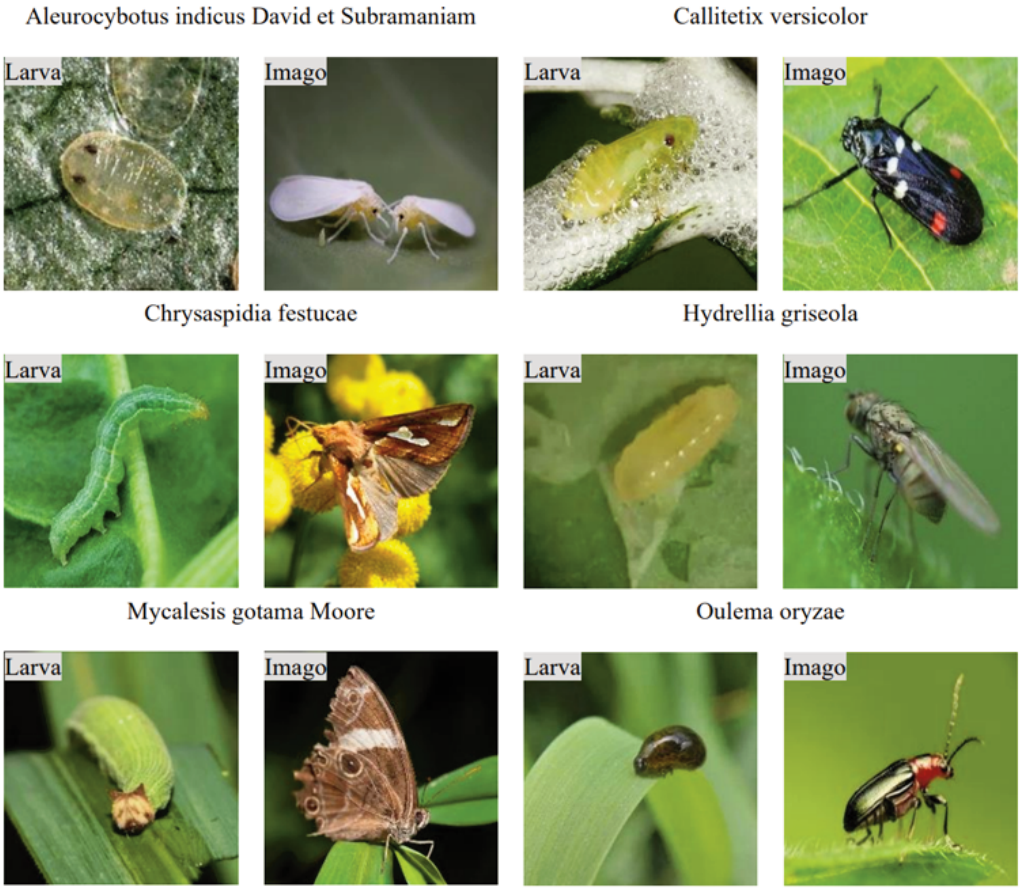

Figure 2. Examples with distinct morphological differences at different growth periods.

Table 1. Ablation experiment results.

Model	Accuracy	Precision	Recall	F1	FLOPs (G)	FPS (ms)
Xception	82.70%	83.12%	82.70%	82.91%	4.6	3.94
Xception+CLIP(XC)	85.16%	85.40%	85.16%	85.28%	8.69	8.72
Xception+CLIP+LMF(XCL)	86.36%	86.42%	86.16%	86.39%	8.70	8.98

Precision (P) and Recall (R). Precision assesses the accuracy of the model in identifying true positive cases as a ratio of all cases predicted to be positive. The formula is as follows:

$$P = \frac{TP}{TP+FP} \quad \text{(Eq. 17)}$$

$$R = \frac{TP}{TP+FN} \quad \text{(Eq. 18)}$$

Additionally, the metrics of Precision and Recall are often trade-off measures, which means that as precision increases, recall may decrease. In some scenarios, it becomes necessary to strike a balance between both precision and recall, and the most commonly used method for achieving this is by utilizing the F1 Score for evaluation. The F1 Score represents the weighted harmonic mean of Precision and Recall.

$$F1 = \frac{2\times(P\times R)}{P+R} \quad \text{(Eq. 19)}$$

Experimental comparison

By assessing the performance on the validation set, we evaluated single-modal versus multi-modal classification to determine if combining image and text information enhances pest classification effectiveness. Furthermore, we validated this on another five excellent neural network models and ultimately compared it to recently prominent classification models in the agricultural field.

Ablation experiment

The GPPK4PCM includes three core modules. To evaluate the individual contribution of each component, an ablation study was conducted, with the results summarized in Table 1. In this context, Xception+CLIP signifies the classification model that integrates visual data from Xception and textual insights from CLIP, utilizing element-wise addition for multi-modal fusion. Xception+CLIP+LMF denotes the combination of Xception-extracted image features and CLIP-extracted text features through the LMF module. For convenience, Xception+CLIP will be referred to as XC, and Xception+CLIP+LMF as XCL.

As shown in Table 1, XC achieved improvements in classification metrics over the single-modal Xception, with increases of 2.46% in accuracy, 2.28% in precision, 2.46% in recall, and 2.37% in F1 score. This indicates that by incorporating prior knowledge of text features from CLIP, the model enhanced its ability to recognize the morphological differences at various periods of pests, thereby improving the classification performance. Despite the moderate increase in computational complexity due to multi-modal fusion (with FLOPs rising from 4.6G to 8.73G), the inference speed remains within a reasonable range for real-time requirements (8.74ms/frame), which fully validates that the gain in classification performance from the textual prior knowledge far outweighs the marginal cost in computational resources. Furthermore, XCL outperforms XC, yielding an additional 1.2% gain in accuracy, 1.02% in precision, 1.0% in recall, and 1.11% in F1 score. These improvements demonstrate the added value of the LMF module in capturing more complex feature interactions. Moreover, as illustrated in Figure 3, with an increase in epochs, the loss and precision of XC and XCL gradually outperformed the single-modal Xception. The curves shifted from oscillation to stability, suggesting that element-wise addition enables access to more sources of information and exhibits better classification performance compared to using a single modality's features. It is worth noting that the FLOPs of XCL (8.74G) are nearly identical to those of XC, and the computational efficiency is further optimized by the parallel tensor decomposition technique, with the inference time increasing only marginally from 8.74 ms to 8.99 ms per frame. This result suggests that the LMF

Table 2. Experimental results on various models.

Model	Accuracy	Precision	Recall	F1
ResNet50	79.04%	80.11%	79.04%	79.57%
ResNet50+CLIP+LMF	83.02%	83.14%	83.02%	83.08%
ViT	69.19%	69.63%	69.19%	69.41%
ViT+CLIP+LMF	71.28%	70.40%	71.28%	70.84%
EfficientNet	75.88%	75.85%	75.88%	75.86%
EfficientNet+CLIP+LMF	76.70%	76.52%	76.70%	76.61%
GhostNet	76.20%	76.85%	76.20%	76.52%
GhostNet+CLIP+LMF	80.56%	80.69%	80.56%	80.62%
InceptionV4	72.47%	73.44%	72.47%	72.95%
InceptionV4+CLIP+LMF	73.97%	73.88%	73.93%	73.90%
Xception	82.32%	83.42%	82.32%	82.87%
Xception+CLIP+LMF	86.36%	86.42%	86.16%	86.39%

Table 3. Experimental results of different agricultural classification models.

Model	Accuracy	Precision	Recall	F1	FLOPs (G)	FPS (ms)
DNVT	79.92%	79.93%	79.92%	79.91%	4.48	23.41
Two-branch-DCNN	75%	75.25%	75%	75.12%	3.46	12.68
ResNet8	75.21%	75.24%	75.21%	75.22%	2.67	15.47
GPPK4PCM (Ours)	86.36%	86.42%	86.16%	86.39%	8.70	8.98

module, through an efficient feature fusion mechanism, maximized the high-order interactions between cross-modal features without significantly sacrificing computational efficiency, ultimately achieving a synergistic optimization of both performance and efficiency. However, it should be noted that merely adding features element-wise signifies a mere linear amalgamation and does not encapsulate the intricate nonlinear interdependencies among the features. In contrast, after features are fused through the LMF module, the model acquires stronger feature representation capabilities. It effectively retrieves higher-order interactions among multi-modal features and greatly enhances the model's performance in challenging classification tasks involving pests at various periods. To further demonstrate the effectiveness of the proposed models in capturing features associated with different pest growth periods, attention heatmaps generated by three models were visualized. These heatmaps highlight the image regions each model focuses on during the classification process.

Figure 4 presents a comparative visualization of the attention

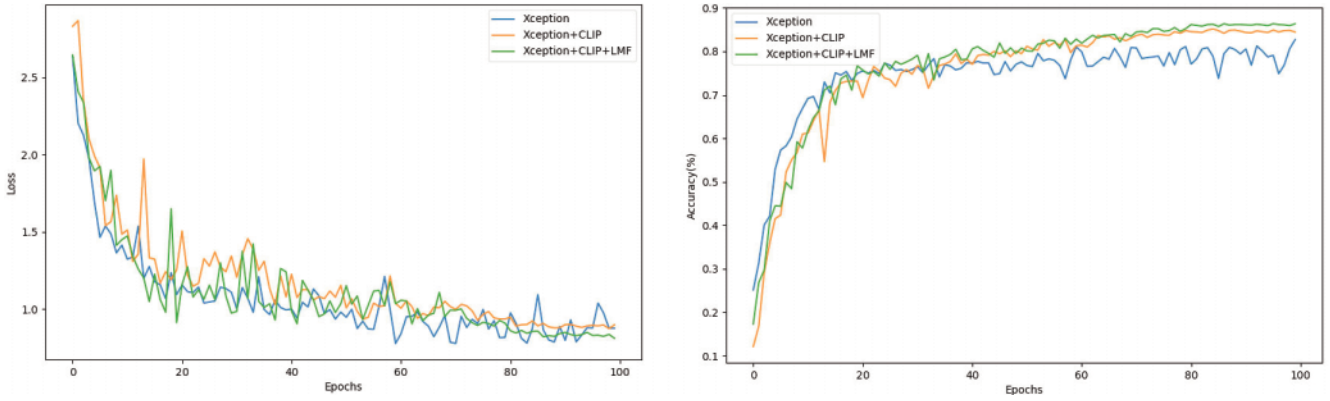

Figure 3. Comparison of loss and accuracy in ablation experiments.

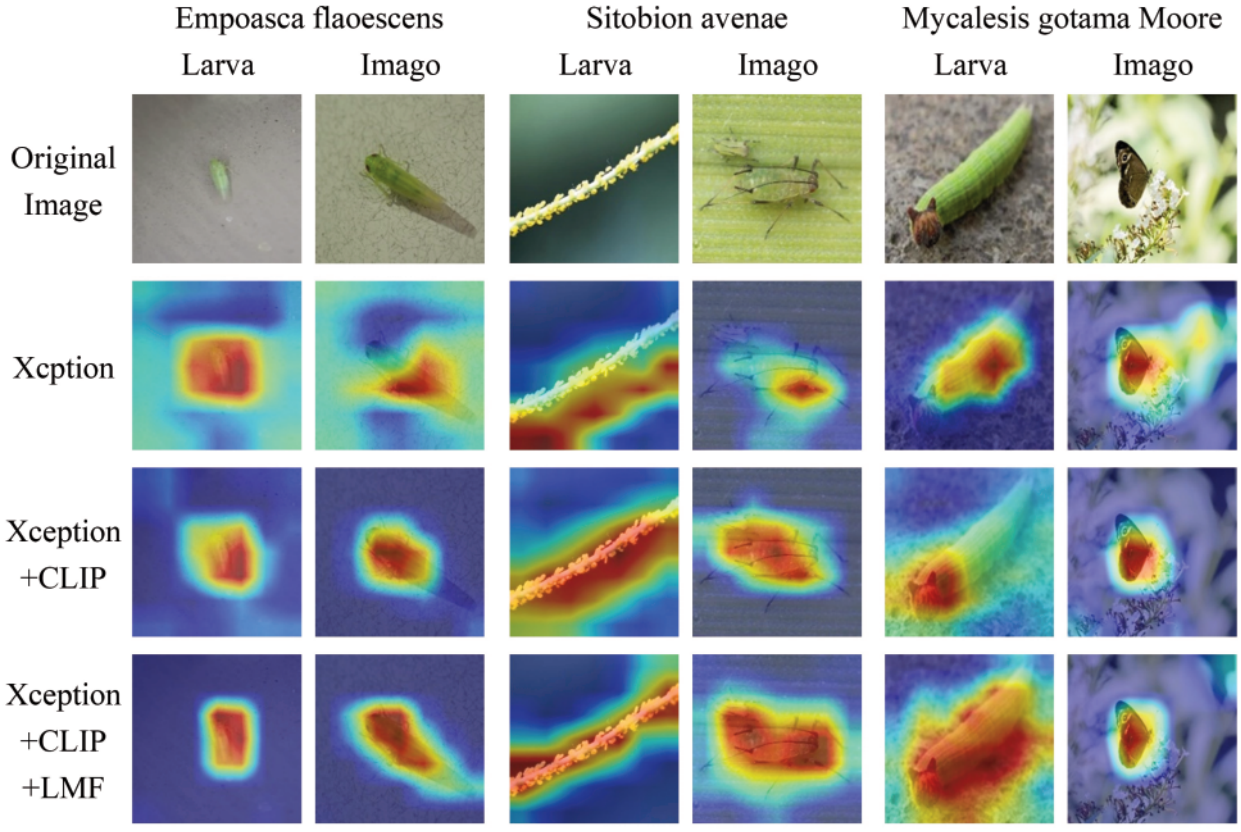

Figure 4. Grad-CAM visualization results for different models.

heatmaps for three representative pest species: *Emposasca flavescens*, *Sitobion avenae*, and *Mycalesis gotama Moore*. Each row corresponds to a different model, while each column represents the larval and adult periods of the pests. Warmer colors (e.g., red and yellow) indicate higher levels of model attention.

The heatmap generated by the baseline model displays a broader, less focused attention distribution, lacking specificity in distinguishing between developmental periods. With the integration of text-based prior knowledge of pest growth periods, the model exhibits improved attention localization, especially in regions relevant to developmental period differentiation. The proposed model, incorporating low-order multimodal fusion (LMF), further enhances this effect. Its heatmaps show more precise, period-specific focus, indicating that the LMF module effectively captures higher-order interactions between visual and textual features.

These visualization results support the conclusion that integrating prior knowledge *via* the CLIP text encoder and applying multimodal fusion through the LMF module significantly enhances the model's ability to discriminate between pest growth periods. This qualitative evidence is consistent with the quantitative improvements reported in the ablation study (Table 1), where the proposed model outperforms both the baseline and intermediate models in classification accuracy.

Comparison of incorporating prior knowledge

Building on earlier discussions, the approach delineated in this study has been designed to enable the utilization of any deep-learning framework for the detection of image characteristics. To demonstrate the general applicability of this method, the experiment tested several image feature extraction networks, including ResNet50, ViT, EfficientNet, GhostNet, Inceptionv4, and Xception. As illustrated in Table 2, the integration of prior textual features derived from the CLIP component has resulted in accuracy improvements for ResNet50, ViT, EfficientNet, GhostNet, Inceptionv4, and Xception by respective increments of 3.98%, 2.09%, 0.82%, 4.36%, 1.50%, and 4.04%. Furthermore, enhancements were also recorded in the Precision, Recall, and F1 Score metrics. These results indicate that the incorporation of the proposed approach into conventional neural networks led to notable improvements in pest classification accuracy, thereby enhancing the model's ability to identify pests across different developmental periods.

Comparison with the latest models

In order to appraise and investigate the performance of the GPPK4PCM in the context of agriculture, this study juxtaposes it against contemporary superior classification models, encompassing DNVT (Xia *et al.,* 2023), two-branch-DCNN (Schuler *et al.,* 2022), and ResNet8 (Guo *et al.,* 2024). A brief introduction to these models is as follows: Schuler *et al.* (2022) introduced a dual-branch deep convolutional neural network (DCNN) aimed at classifying plant diseases, employing three convolutional layers to discern features from the CIE Lab color space and chromatic aberration. Experimental results show that it outperforms traditional single-branch RGB image classification performance. A lightweight, open-world pest image classification model was presented by Guo and associates (2024), featuring a matching network and NT-Xent loss function, all integrated within the ResNet8 framework. The classifier operates by harnessing a ResNet8-based trained matching network to measure the closeness between the prototypes of the support classes and the representations of the query images, exceeding the performance of competing lightweight networks. Xia *et al.* integrated a convolutional neural network (CNN) with an enhanced visual transformer to craft a novel classification model known as the DenseNet Vision Transformer (DNVT). The DNVT framework addresses both long-range dependencies and local feature modeling, significantly enhancing the precision of pest classification. Among the above three models, DNVT and ResNet8 are used for classifying agricultural pests, similar to the application in this paper. Although two-branch-DCNN is used for plant disease classification, it is also used for image classification in the agricultural field and has a similar application scenario. To further evaluate the effectiveness and practicality of GPPK4PCM in agricultural image classification, a comparison was conducted against the three above models. As shown in Table 3, GPPK4PCM achieves an accuracy of 86.36%, a precision of 86.42%, a recall of 86.16%, and an F1 score of 86.39%. Compared with the other three models applied in similar agricultural scenarios, it exhibits clear advantages, with accuracy improvements of 6.44%, 11.36%, and 11%, respectively. In addition to classification accuracy, computational complexity and inference speed were also considered. GPPK4PCM records the highest FLOPs (8.74G) among the models but maintains a reasonable inference time (8.98 ms), making the performance-cost trade-off acceptable. In contrast, models such as ResNet8 and two-branch-DCNN achieve lower FLOPs (2.67G and 3.46G) and faster inference speeds (14.18 ms and 12.68 ms), but at the cost of reduced accuracy. These findings indicate that GPPK4PCM not only delivers superior classification performance but also offers a well-balanced trade-off between model complexity and practical deployment, reinforcing its potential for real-world agricultural applications.

Conclusions

This study presents a classification model that incorporates prior knowledge of growth periods of pests. By incorporating such prior knowledge, the proposed method effectively integrates the developmental period information present in pest images and employs an efficient feature fusion mechanism to enhance the model's classification precision. We developed a dataset that encompasses pests at different developmental periods, drawing from the pest dataset of Sichuan Agricultural University, and conducted experiments to evaluate the efficacy of our proposed method. The findings reveal that GPPK4PCM achieves higher classification accuracy in addressing the significant morphological differences observed throughout the pest life cycle. The proposed method in question serves purpose: enhancing the accuracy of pest identification while simultaneously providing effective technical support for agricultural pest control. However, the current dataset lacks a sufficient number of images for the larval and pupal periods in order to identify pests at different periods. Therefore, this paper's method only focuses on the classification of pests in the larval and adult periods. Future studies could prioritize the advancement of methods for extracting features and classifying pests at different growth periods, while also optimizing the data collection methodology for each period to bolster the model's accuracy and generalization capabilities.

References

Albattah, W., Masood, M., Javed, A. 2023. Custom CornerNet: a drone-based improved deep learning technique for large-scale

multiclass pest localization and classification. Compl. Intell. Syst. 9:1299-1316.
Bai, J., Liu, X., Wang, Y. 2024. Integrating prior knowledge and contrast feature for signal modulation classification. IEEE Internet Things 11:21461-21473.
Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017, pp. 1800-1807.
Dai, G., Fan, J., Dewi, C. 2023. ITF-WPI: Image and text based cross-modal feature fusion model for wolfberry pest recognition. Comput. Electron. Agr. 212:108129.
Deng, X., Feng, S., Lyu, G., 2022. Beyond word embeddings: Heterogeneous prior knowledge driven multi-label image classification. IEEE T. Multimedia 25: 4013-4025.
Dosovitskiy, A., Beyer, L., Kolesnikov, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929.
Guo, Q., Wang, C., Xiao, D. 2024. A lightweight open-world pest image classifier using ResNet8-based matching network and NT-Xent loss function. Expert Syst. Appl. 237:121395.
Han, K., Wang, Y., Tian, Q. 2020. Ghostnet: More features from cheap operations. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Seattle, pp. 1577-1586.
He, K., Zhang, X., Ren, S. 2016. Deep residual learning for image recognition. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, pp. 770-778.
Howard, A.G., Zhu, M., Chen, B. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861.
Kasinathan T, Singaraju D. 2021. Uyyala S R. Insect classification and detection in field crops using modern machine learning techniques. Available from: https://papers.nips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
Khanramaki, M., Asli-Ardeh, E.A., Kozegar, E. 2021. Citrus pests classification using an ensemble of deep learning models. Comput. Electron. Agr. 186:106192.
Krizhevsky A, Sutskever I. 2012. Hinton G E. Imagenet classification with deep convolutional neural networks. Adv. Neural Information Processing Syst. 25.
Li, S., Sun, L., Li, Q. 2023. CLIP-ReID: exploiting vision-language model for image re-identification without concrete text labels. Proc. AAAI Conf. on Artificial Intelligence 37:1405-1413.
Lin, Y., Chen, M., Wang, W. 2023. CLIP is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Vancouver, pp. 15305-15314.
Liu, W., Wu, G., Ren, F. 2020. DFF-ResNet: An insect pest recognition model based on residual networks. Big Data Mining Anal. 3:300-310.
Liu, Z., Shen, Y., Lakshminarasimhan, V.B. 2018. Efficient low-rank multimodal fusion with modality-specific factors. Proc. 56th Annual Meet. Assoc. Computational Linguistics 1:2247-2256.
Lu, W., Wang, X., Jia, W. 2022. Root hair image processing based on deep learning and prior knowledge. Comput. Electron. Agr. 202:107397.
Radford, A., Kim, J.W., Hallacy, C., 2021. Learning transferable visual models from natural language supervision. Proc. 38th Int. Conf. Machine Learning, PMLR 8748-8763.
Schuler, J.P.S, Romani, S., Abdel-Nasser, M., Rashwan, H., Puig, D. 2022. Color-aware two-branch DCNN for efficient plant disease classification. Mendel 28:55-62.
Setiawan, A., Yudistira, N., Wihandika, R.C. 2022. Large scale pest classification using efficient Convolutional Neural Network with augmentation and regularizers. Comput. Electron. Agr. 200:107204.
Sichuan Agriculture University, 2020. Sichuan Agricultural University plant diseases and pests open dataset. Accessed 15 July 2020. Available from: https://github.com/SAUTEG/version_1.0
Simonyan, K., Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556.
Szegedy, C., Liu, W., Jia, Y. 2015. Going deeper with convolutions. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Boston, pp. 1-9.
Szegedy, C., Ioffe, S., Vanhoucke, V. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. Proc. AAAI Conf. on Artificial Intelligence 31:4278-4284.
Tan, M., Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. Proc. 36th Int. Conf. Machine Learning, PMLR 6105-6114.
Tuda, M., Luna-Maldonado, A.I., 2020. Image-based insect species and gender classification by trained supervised machine learning algorithms. Ecol. Inform. 60:101135.
Wang, Q., Wang, C., Lai, Z., 2024. Insectmamba: Insect pest classification with state space model. arXiv:2404.0361.
Wei, D., Chen, J., Luo, T. 2021. Classification of crop pests based on multi-scale feature fusion. Comput. Electron. Agr. 194:106736.
Xia, W., Han, D., Li, D. 2023. An ensemble learning integration of multiple CNN with improved vision transformer models for pest classification. Ann. Appl. Biol. 182:144-158.
Yi, C., Ren, L., Zhan, D.C., 2024. Leveraging cross-modal neighbor representation for improved CLIP classification. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Seattle, pp. 27402-27411.
Zadeh, A., Chen, M., Poria, S. 2017. Tensor fusion network for multimodal sentiment analysis. arXiv:1707.07250.
Zhang, Y., Chen, L., Yuan, Y. 2023. Multimodal fine-grained transformer model for pest recognition. Electronics 12:2620.
Zhou, J., Li, J., Wang, C. 2021. Crop disease identification and interpretation method based on multimodal deep learning. Comput. Electron. Agr. 189:106408.