electronics

MDPI

Article

Multimodal Fine-Grained Transformer Model for Pest Recognition

Yinshuo Zhang [1,2], Lei Chen [1,*] and Yuan Yuan [1]

[1] Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China
[2] Science Island Branch, Graduate School of University of Science and Technology of China, Hefei 230026, China
* Correspondence: chenlei@iim.ac.cn

Abstract: Deep learning has shown great potential in smart agriculture, especially in the field of pest recognition. However, existing methods require large datasets and do not exploit the semantic associations between multimodal data. To address these problems, this paper proposes a multimodal fine-grained transformer (MMFGT) model, a novel pest recognition method that improves three aspects of transformer architecture to meet the needs of few-shot pest recognition. On the one hand, the MMFGT uses self-supervised learning to extend the transformer structure to extract target features using contrastive learning to reduce the reliance on data volume. On the other hand, fine-grained recognition is integrated into the MMFGT to focus attention on finely differentiated areas of pest images to improve recognition accuracy. In addition, the MMFGT further improves the performance in pest recognition by using the joint multimodal information from the pest's image and natural language description. Extensive experimental results demonstrate that the MMFGT obtains more competitive results compared to other excellent models, such as ResNet, ViT, SwinT, DINO, and EsViT, in pest recognition tasks, with recognition accuracy up to 98.12% and achieving 5.92% higher accuracy compared to the state-of-the-art DINO method for the baseline.

Keywords: pest recognition; multimodal representation; fine-grained image recognition; vision transformer; few-shot learning

Citation: Zhang, Y.; Chen, L.; Yuan, Y. Multimodal Fine-Grained Transformer Model for Pest Recognition. *Electronics* **2023**, *12*, 2620. https://doi.org/10.3390/electronics12122620

Academic Editor: Enzo Pasquale Scilingo

Received: 20 April 2023
Revised: 4 June 2023
Accepted: 7 June 2023
Published: 10 June 2023

1. Introduction

Agricultural pests seriously affect agricultural production and crop storage. Prevention of agricultural pests requires proper recognition of the pest species and targeted control measures. The mainstream methods for image classification are based on deep convolutional neural networks (CNNs) [1]. However, such methods require a large number of high-quality pest datasets manually labeled by experts, which is costly and impractical. Therefore, the pest recognition techniques that can accommodate few-shot and low-quality pest datasets have become a hot topic in current research. Transfer learning-based image recognition methods [2] have achieved remarkable results with few-shot datasets. However, such methods require the data in the source and target domains to be as similar as possible, which is often difficult to satisfy for fine-grained recognition.

In recent years, the vision transformer (ViT) [3] model has achieved remarkable success as a new model for applying transformers to the field of computer vision. ViT uses a self-attentive mechanism to extract global information and is capable of parallel training. However, such a method still requires large sample datasets for training. In contrast, self-supervised learning is based on using positive and negative sample pairs for feature extraction, which can achieve good performance with few-shot datasets. Therefore, fusion of self-supervised learning and ViT has become one of the potential methods to solve the few-shot pest recognition problem.

To address the problem that existing methods require large datasets and do not exploit the semantic associations between multimodal data, this paper proposes a multimodal

fine-grained transformer (MMFGT) model for pest recognition. The MMFGT extends the transformer structure with self-supervised learning and fine-grained recognition methods. In particular, it extracts target features using self-supervised learning to improve recognition accuracy and reduce the reliance on data volume, while focusing attention on subdivision regions of pest images, overcoming the challenge represented by pest images with low proportions of pest targets, which are difficult to identify accurately. As shown in Figure 1, the MMFGT method can improve the accuracy of fine-grained pest recognition by focusing attention on the subdivided regions of the head, thorax, and tail in the pest image. In addition, the MMFGT further improves the performance of fine-grained pest recognition by exploiting the joint multimodal information from images and natural language descriptions of pests. Experimental results showed that the MMFGT achieved more competitive results compared to several advanced image recognition methods in the pest recognition task, with a recognition accuracy of up to 98.12%.

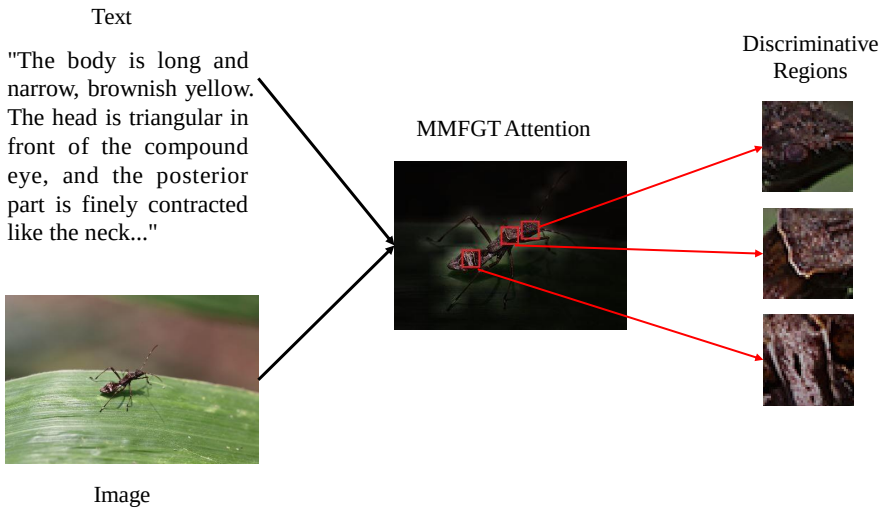


Figure 1. Visualization results from the MMFGT for fine-grained attention, with the selected top three positions shown in red.

The contributions of this paper can be summarized in three points:

- The multimodal fine-grained transformer model for pest recognition is proposed. The MMFGT extends the ViT model through self-supervised learning and fine-grained recognition to address the problem of pest recognition; i.e., the target is small in the image and a large sample dataset is required for training;
- Compared to existing methods, the proposed MMFGT model can provide a better-performing classification strategy for pest recognition by using joint multimodal information from images and natural language descriptions;
- Extensive experimental results demonstrate that the MMFGT can obtain more competitive results compared to existing image recognition models in the pest recognition task. The accuracy levels of the MMFGT with the IDADP pest dataset and tomato pest dataset were 98.12% and 95.83%, respectively, 5.92% and 4.16% higher than the state-of-the-art DINO method for the baseline.

The rest of the work is organized as follows: Section 2 describes the related work, and the materials and the proposed method are discussed in Section 3. Section 4 presents the experiments and discussion. Section 5 summarizes this work.

2. Related Work

2.1. Image Classification

The current mainstream image classification methods include AlexNet, VGG, GoogLeNet, ResNet, InceptionNet, DenseNet, and other CNN methods [1]. Related studies on using these methods for pest recognition have been conducted [4–7]. However, training of a CNN requires

large datasets containing thousands of images, and there is no such large and diverse dataset in the field of pest recognition. Transfer learning, which can effectively improve recognition accuracy by pretraining CNN models with large datasets and re-training them with smaller datasets [8–10], is an effective way to solve the above problem. Dawei et al. [2] proposed a transfer learning model based on AlexNet for a pest detection and recognition diagnosis system that can be trained for and detect 10 pest species. In addition, a deep CNN-based transfer learning framework [11] has been proposed for the classification of tomato pests. Chen et al. [12] studied migration learning with deep convolutional neural networks for plant leaf disease identification. However, transfer learning requires the data in the source and target domains to be as similar as possible, which is often difficult to achieve for fine-grained recognition. In contrast, our method does not require a large and diverse dataset, nor does it impose additional requirements on the dataset. Therefore, it is more suitable for pest recognition where datasets are difficult to obtain.

2.2. Transformer

In recent years, the transformer model [13] has achieved great success in the field of natural language processing as a pure attention mechanism approach that can learn high-quality feature representations by considering the whole context. However, there are significant challenges in using the transformer model for image classification due to the significant differences in many scales between visual signals and textual symbols, as well as the high dimensionality of pixel-level information, which may introduce additional computational complexity. The ViT model [3] was proposed as a transformer model that could be applied directly to image patch sequences to solve the image recognition problem. Liu et al. [14] proposed a hierarchical ViT model called SwinT, which is more suitable for computer vision tasks as it shifts the window computational representation and restricts the self-attentive computation to non-overlapping local windows while also allowing cross-window connectivity. However, the transformer method also relies on a large dataset for training and is not effective for few-shot pest recognition. In this work, the proposed MMFGT model uses self-supervised learning to extend the transformer structure to extract target features through contrastive learning to reduce the reliance on data volume.

2.3. Self-Supervised Learning

Self-supervised learning can effectively solve the few-shot problem. It involves learning by constructing positive and negative samples and comparing the distance difference between them without labeling data. Two of its most famous implementations in the field of computer vision are MoCo and SimCLR. MoCo [15] uses momentum contrast for unsupervised visual representation learning. SimCLR [16] is a simple visual-representation contrastive learning framework. MoCov2 [17] establishes a stronger baseline with better performance than SimCLR by using MLP projection heads and more data augmentation in the MoCo framework, and it does not require large-scale batch training. SimCLRv2 [18] is a simple semi-supervised ImageNet classification framework that includes unsupervised pretraining, supervised fine-tuning, and extraction of unlabeled data. BYOL [19] does not require negative samples and learns its representation by predicting the output of previous versions. Sim Siam [20] is a simple Siamese network that uses neither negative sample pairs nor a momentum encoder. DINO [21] is a recently proposed method combining self-supervision with the ViT to solve the recognition problem for few-shot images. EsViT [22] is a more recently proposed method combining self-supervision with SwinT that is also suitable for solving the few-shot image recognition problem. However, the small percentage of targets in pest images makes identification difficult. To solve this problem, fine-grained recognition was integrated into the MMFGT to focus attention on finely differentiated regions of pest images to improve recognition accuracy.

2.4. Multimodal Learning

Multimodal learning approaches address machine learning problems that contain data from different modalities, and they can by used for tasks such as data classification [23,24], sentiment analysis, semantic computing, cross-modal retrieval [25,26], and visual question answering [27,28]. In image classification, multimodal data information can describe images more comprehensively than single-modality data information. For instance, natural language can complement the description of subtle differences between images to facilitate image classification. He et al. [29] used natural language descriptions to identify the discriminative parts of relevant images, thus enabling multimodal representation for fine-grained image classification. Nawaz et al. [30] proposed a strategy for learning natural language descriptions and joint image representations using a multilayer two-branch network to improve fine-grained classification tasks. Gallo et al. [31] built a multimodal classifier with two different models and adapted it to a stacking technique. In agricultural disease identification, Zhou et al. [32] studied the semantic embedding methods for disease images and disease description texts, as well as knowledge representation and knowledge-embedding mechanisms in the disease recognition domain, and constructed a disease identification model based on "image-text" multimodal collaborative representation and knowledge assistance (ITK-Net). However, little work has been devoted to the use of multimodal information for few-shot image classification because multimodal information is not easy to search and difficult to fuse. In this work, we propose a few-shot image classification method based on multimodal information.

3. Materials and Methods

In this section, we first introduce the two datasets used for the experiments in Section 3.1 and then discuss the detailed structure of the proposed MMFGT method in Section 3.2.

3.1. Datasets

Two datasets were used in the experimental part of this study. One was the IDADP pest dataset [33] containing 1293 images of 29 pest categories; the maximum number of images per category is 124 and the minimum number of images is 6. The amount of data in each category is small and unevenly distributed. The number of images in each category in the IDADP dataset is shown in Table 1. Typical example images and the corresponding description text are shown in Table 2. In each image, the background of the image is complex and the pests are small and inconspicuous in the image. The samples are manually divided into training and validation sets in a ratio of 7:3. The text dataset is based on the corresponding pest description text written for each image in Wiki Encyclopedia.

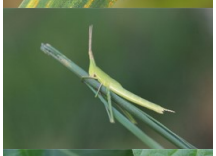
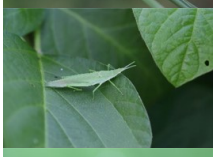
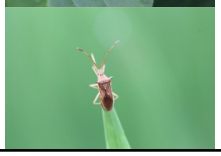
Table 1. The number of images per category in the IDADP dataset.

Pest Name	Number
Colposcelis signata	41
Piezodorus rubrofasciatus	19
Riptortus pedestris	54
Eysacoris guttiger	6
Erthesina fullo	40
Membracidae	41
Acrida cinerea	13
Tingidae	18
Oxya	10
Scurelleridae	9
Spoladea recurvalis	38
Cletus schmidti Kiritshenko	40

Table 1. *Cont.*

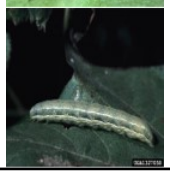
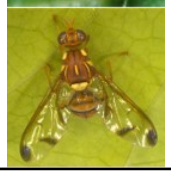
Pest Name	Number
Ascotis selenaria Schiffermuller et Denis	36
Helicoverpa armigera	39
Berytidae	79
Taiwania	25
Aphidoidea	124
Eurygaster testudinarius	19
Spodoptera frugiperda	95
Trigonotylus ruficornis Geoffroy	28
Riptortus linearis Fabricius	65
Rhopalosiphum maidis	19
Pygmy sand cricket	17
Atractomorpha sinensis Bolivar	90
Tropidothorax elegans Distant	30
Cletus punctiger Dallas	93
Dolycoris baccarum	120
Nysius ericae	62
Longhorned grasshoppers	23

Table 2. Typical example images and corresponding description text.

Pest Name	Image	Description Text
Aphidoidea		Body length 2 mm, green, ovoid. Eyes large, small ocular surface. Ventral tube tubular, apical margin protruding, surface smooth or tiled. Tail plate end round
Acrida cinerea		Body medium to large, green in color. Head conical. Face extremely inclined, face bulge extremely narrow. Head protruding with rounded apex. Antennae sword-shaped. Compound eyes long-oval
Atractomorpha sinensis Bolivar		The body is green. Head sharpened, protruding forward, with small yellow tuberculate projections on lateral margins. Forewings green, exceeding the abdomen; hindwings red at the base and light green at the tip
Membracidae		The body is yellowish brown, narrow and long, with dense carving points. The top of the head and the anterior margin of the dorsal plate of the prothorax are dotted with small black grains. Compound eyes maroon, single eye red. The lateral horns are elongated and black at the end

The second database was a database of images of eight common tomato pests [34], including *Tetranychus urticae*, *Bemisia argentifolii*, *Zeugodacus cucuurbitae*, *Thrips palmi*, *Myzus persicae*, *Spodoptera litura*, *Spodoptera exigua*, and *Helicoverpa armigera*. The images were collected from the IPMImages database and the National Bureau of Agricultural Insect Resources (NBAIR), totaling 609 images of pests in eight categories. The number of images and example images for each category are shown in Table 3.

Table 3. Number of images and example images for each category in a database of eight common tomato pests.

Pest Name	Image	Number	Pest Name	Image	Number
Bemisia argentifolii		61	*Spodoptera litura*		97
Helicoverpa armigera		109	*Thrips palmi*		25
Myzus persicae		131	*Tetranychus urticae*		75
Spodoptera exigua		76	*Zeugodacus cucurbitae*		43

3.2. The Structure of the Proposed Model

In this paper, a fine-grained pest recognition model (MMFGT) is proposed to solve the few-shot pest recognition problem. The architecture is shown in Figure 2. At the beginning, the image is segmented into small pieces and projected into the embedding space. The input to the transformer encoder includes the patch embedding, as well as the learnable position embedding. Before the last transformer layer, a part selection module (PSM) is applied to select the tokens corresponding to the discriminative image patches, and only these selected tokens are used as input for the last transformer layer to finally obtain the features of the image. The description text corresponding to the image is fed to the text encoder (ALBERT [35]), and the input is transformed into feature vectors as text features after a multi-layer transformer. Finally, the image features and text features are linearly stitched together for linear classification to obtain the predicted class of pests.

The MMFGT includes improvements of three aspects of the transformer architecture to make it well-suited for few-shot pest recognition: (1) Self-supervision. To address the problem that it is difficult to train few-shot pest datasets on a large scale, the MMFGT uses self-supervised learning to extend the transformer architecture to extract target features by means of contrastive learning, reducing the dependence on data volume; (2) Fine-grained recognition. In order to overcome the challenge represented by the small percentage of pest targets in pest images, the MMFGT integrates fine-grained recognition to focus attention on subdivided areas of pest images and improve recognition accuracy; (3) Multimodality. The MMFGT can utilize the joint multimodal information from image and natural language descriptions encoded using the image encoder (fine-grained transformer model for pest recognition (FGT)) and text encoder (ALBERT [35]), respectively. These extracted image features and text features are combined and fed into a linear classifier for classification, further improving the performance of fine-grained pest recognition. Compared to previous work, although DINO [21] is implemented in a self-supervised manner to train the ViT, which is suitable for small-sample recognition, this model does not perform well in the pest recognition task due to the limited variation in pest features. In contrast, inspired by [21,36], we innovatively improved the accuracy of fine-grained pest recognition by incorporating a fine-grained module, PSM, into the ViT to focus attention on pest segmentation in this

work. In addition, our model further improves the accuracy of pest recognition by fusing image features and text features extracted with our proposed image encoder (fine-grained transformer model for pest recognition (FGT)) and text encoder (ALBERT [35]) for pest classification. To the best of our knowledge, this is the first method that combines fine-grained, self-supervised ViT and multimodal models for pest recognition.

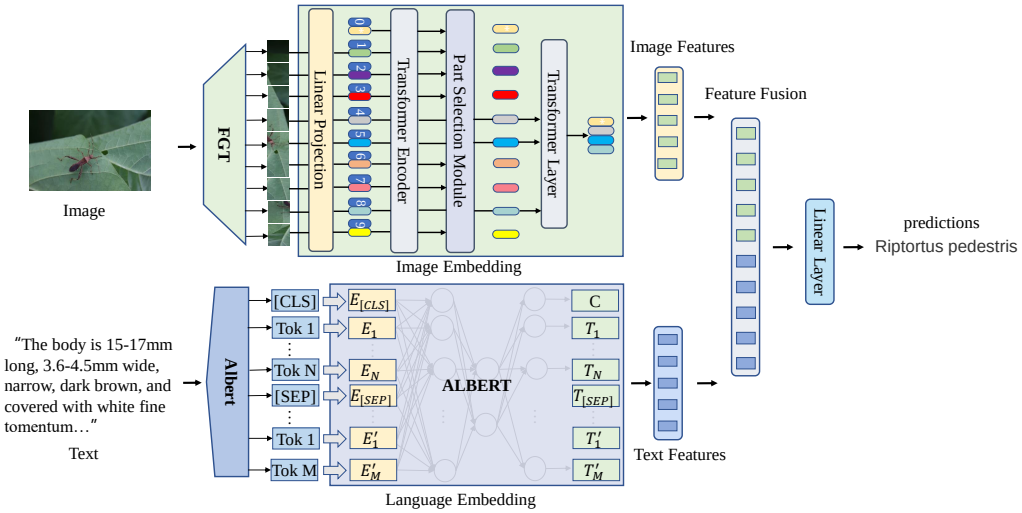


Figure 2. Framework diagram of the MMFGT. The image and the corresponding text description are fed into the image encoder (FGT) and the text encoder (ALBERT), respectively. Then, the obtained image features and text features are stitched together for linear classification.

3.2.1. Image Encoder

The image encoder uses FGT, a fine-grained pest recognition method with a self-supervised transformer architecture. As shown in Figure 3, FGT uses a combination of self-supervised learning and knowledge distillation [21], with the teacher network being dynamically constructed during the training process and having the same architecture as the student network but with different parameters. Two different image transforms of the input image are fed into the student network g_{θ_s} and the teacher network g_{θ_t}, respectively. One N-dimensional feature is output from each of the two networks, and the similarity between the two features is calculated using cross-entropy loss after normalizing the features with softmax. The model propagates the gradient through the student network only, and the parameters of the teacher network are updated using an exponential moving average (EMA) of the student parameters. Both the student and teacher networks in FGT use the ViT model modified with a part selection module (PSM) to better extract nuanced regional features. The implementation details are as follows.

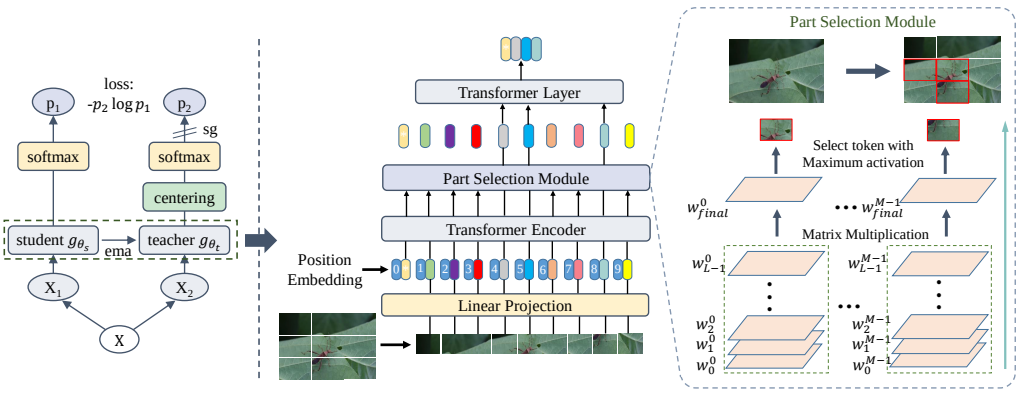


Figure 3. FGT image encoder. Training the ViT model improved with a PSM with self-supervised learning.

Self-Supervised Learning Architecture

The architecture performs two different random transformations on the input image x and generates a set of view sets containing two global views x_1^g and x_2^g and several local views. All views are passed through the student network, while only the global views are passed through the teacher network. The student network g_{θ_s} is trained based on the output of the teacher network g_{θ_t}, and the two networks have the same structure but different parameters, denoted by θ_s and θ_t, respectively. The probability distributions for the N dimensions of the outputs of the two networks are denoted as P_s and P_t. The probability P is obtained by normalizing the output of the network g using the softmax function, denoted as:

$$P_s(x)^{(i)} = \frac{exp(g_{\theta_s}(x)^{(i)}/T_s)}{\sum_{n=1}^{N} exp(g_{\theta_s}(x)^{(n)}/T_s)} \tag{1}$$

where $T_s > 0$ is the temperature parameter that controls the sharpness of the output distribution. Given a fixed teacher network g_{θ_t}, the model learns to match these probability distributions P by minimizing the cross-entropy loss with respect to the parameter of the student network θ_s:

$$\min_{\theta_s} \sum_{x \in \{x_1^g, x_2^g\}} \sum_{x' \in V and x' \neq x} H(P_t(x), P_s(x')) \tag{2}$$

where $H(a, b) = -a\log b$. The parameter θ_s is learned by minimizing Equation (2) using stochastic gradient descent.

Improvement of the ViT Architecture with a PSM

To better extract nuanced regional features, the student and teacher networks use the ViT architecture improved with a PSM. To make full use of the attentional information, the last transformer layer is used as input for the feature layer. The improved ViT has M self-attentive heads, and the hidden features of the last layer input are noted as $a_{L-1} = [a_{L-1}^0; a_{L-1}^1, a_{L-1}^2, \ldots, a_{L-1}^N]$. The attention weights of the previous layers are:

$$w_l = [w_l^0, w_l^1, w_l^2, \ldots, w_l^M] \qquad l \in 1, 2, \ldots, L-1 \tag{3}$$

$$w_l^i = [w_l^{i_0}; w_l^{i_1}, w_l^{i_2}, \ldots, w_l^{i_N}] \qquad i \in 0, 1, \ldots, M-1 \tag{4}$$

The matrix multiplication is applied recursively to the original attention weights of all layers:

$$w_{final} = \prod_{l=0}^{L-1} w_l \tag{5}$$

Compared to the single-layer raw attention weight w_{L-1}, w_{final} is a better choice for selecting discriminative regions because it captures how information is passed from the input layer to higher-level embeddings. Then, the indexes of the maximum values $W_1, W_2, \ldots, W_M$ are chosen for the M different attention heads in w_{final}. These positions are used as indexes of our model to extract the corresponding tokens in a_{L-1}. Finally, the selected tokens with the classification tokens are connected as input to the last transformer layer, denoted as:

$$a_{local} = [a_{L-1}^0; a_{L-1}^{W_1}, a_{L-1}^{W_2}, \ldots, a_{L-1}^{W_M}] \tag{6}$$

By connecting the categorical tokens corresponding to information regions to the last transformer layer instead of the original, entire input sequence as input, the architecture not only preserves the global information but also forces the last transformer layer to focus

on the subtle differences between different subcategories and ignore less discriminative regions, such as background or common features. Figure 4 shows the difference images of the pests captured with the FGT method.

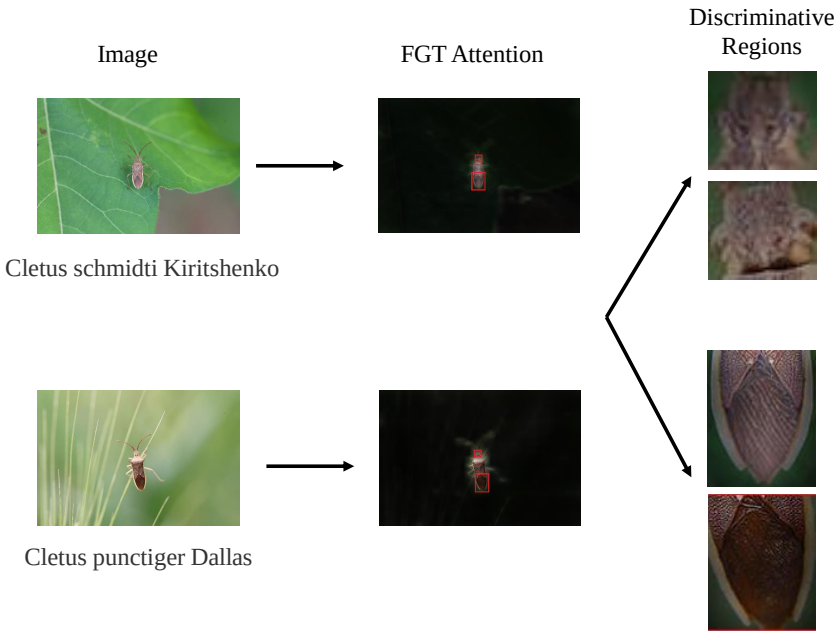


Figure 4. A pair of similar instances from the IDADP pest dataset for which FGT applied the ViT model improved with a PSM to capture subtle differences.

3.2.2. Text Encoder

Single-modality representation using only images often struggles to achieve fine-grained recognition. Therefore, multimodal representation including image and text was used for more robust predictions since the text information can complement the image information. Natural language description provides useful information that can be used for fine-grained image classification. The joint multimodal information from the image and natural language description can yield richer representation information than image description. As shown in Figure 2, when one modality of data is missing, the multimodal system can still operate based on the other modality of information.

The mainstream text encoders are BERT [37], ALBERT [35], etc. As shown in Figure 5, BERT is a deep bi-directional language representation model capable of incorporating contextual information. The input of BERT is the representation corresponding to each token, denoted as E. The representation consists of token embeddings, segment embeddings, and position embeddings. The word dictionary is constructed using the WordPiece algorithm. To accomplish the specific classification task, in addition to the word token, each sequence of the input has a specific classification token $[CLS]$ at the beginning, and the output of the last transformer layer corresponding to this classification token plays the role of aggregating the information of the whole sequence representation. Each sentence is followed by a split token $[SEP]$ to separate the different sentence tokens. After the multilayer transformer, C is the output of the classification token $[CLS]$ corresponding to the last transformer, and T_i represents the output of the ith input token corresponding to the last transformer. The input text goes through the BERT model and is transformed into feature vectors.

Compared to BERT, ALBERT has simplified model parameters, reduced memory consumption, and improved training speed, solving the problem of limited GPU memory. Therefore, we used ALBERT to encode the text information. The backbone of the ALBERT model is similar to BERT in that both use transformer encoders with the GELU nonlinearity activation function. The vocabulary size is denoted as V, the vocabulary embedding size is denoted as F, and the hidden layer size is denoted as H. ALBERT has improved factorized embedding parameterization and cross-layer parameter sharing. ALBERT uses

the factorization of embedding parameters; instead of directly projecting one-hot vectors into the hidden space of size H, it first projects them into the low-dimensional embedding space of size F and then projects them into the hidden space. With this decomposition, the embedding parameters are reduced from $O(V * H)$ to $O(V * F + F * H)$, where $F \ll H$, thus reducing the number of parameters. Moreover, ALBERT shares all parameters across layers, further reducing the number of parameters. When processing pest text, ALBERT's training corpus is large and fully trained, which can solve the problem related to the many technical terms in pest text. Furthermore, ALBERT's network structure adopts a bidirectional transformer, which better solves the problem of the context dependence of pest text. At the same time, the ALBERT model is smaller and has faster convergence, shorter prediction time, and better recognition timeliness. Finally, the obtained text features are linearly stitched with the image features and then fed into a linear classifier for classification prediction.

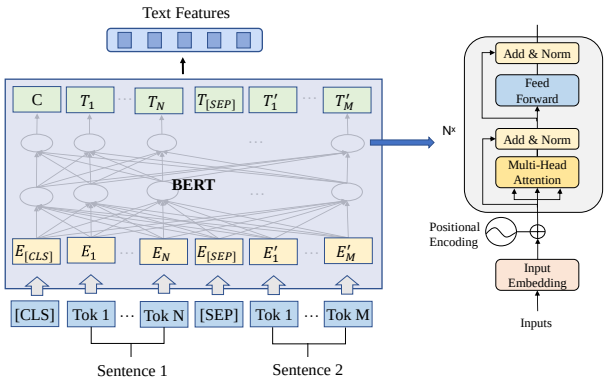


Figure 5. BERT architecture.

4. Experiments and Discussion

4.1. Experimental Settings

Experimental procedure. The MMFGT model was first validated with a 29-class IDADP pest dataset. To verify the effectiveness of the MMFGT model in identifying similar datasets, it was further validated with the 15-class IDADP stinkbug dataset with higher similarity and the remaining 14-class IDADP pest datasets. Furthermore, its effectiveness was evaluated with a database of eight common tomato pest images. The performance of the MMFGT model in recognizing real pest images was evaluated by analyzing the recognition accuracy.

Pretraining. The ViT architecture improved with a PSM was trained in a self-supervised manner on the CUB_200_2011 dataset, and the initial weights of the ViT were loaded from the ViT-B_16 model pretrained on ImageNet21k. The text encoder used the Hugging Face pretrained ALBERT language model.

Parameter Settings. CrossEntropyLoss was used as the loss function and SGD as the optimizer. The batch size was set to 16, the initial learning rate was 0.07, the momentum was 0.9, the drop_rate was 0.1, and the epoch was 500.

Experimental environment. All models used in this work were implemented on a desktop running Ubuntu 18.04 with Nvidia Quadro RTX 6000Ti GPU.

4.2. Evaluation Metrics

In this paper, the performance of the model was evaluated in terms of accuracy (ACC), precision (P), recall (R), and F1 score. The computational equations are shown below:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100\% \tag{7}$$

$$P = \frac{TP}{TP + FP} \times 100\% \quad (8)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (9)$$

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\% \quad (10)$$

where TP refers to the number of samples that are actually positive and predicted to be positive, FP refers to the number of samples that are actually negative but predicted to be positive, FN refers to the number of samples that are actually positive but predicted to be negative, TN refers to the number of samples that are actually negative and predicted to be negative, and TN refers to the number of samples that are actually negative but predicted to be positive.

4.3. Baseline

Several advanced models (ResNet101 [38], ViT [3], SwinT [14], DINO [21], and EsViT [22]) were chosen as the baselines. ResNet101 is a classical CNN method and the most widely used image recognition method. ViT is a recently proposed method applying the transformer architecture to the field of computer vision that has achieved good results in the field of image recognition. The more recently proposed SwinT exploits the prior knowledge of the CNN and is more suitable for dealing with image problems. DINO is a recently proposed method combining self-supervision with ViT that is suitable for few-shot image recognition. EsViT is a more recently proposed method combining self-supervision with SwinT that is also suitable for few-shot image recognition. In this study, a fine-grained transformer model for pest identification (FGT) was developed and used as a baseline to compare the effectiveness of the multimodal MMFGT.

4.4. Experimental Results

The performances of the FGT model, MMFGT model, and the five baseline methods were compared by comparing their accuracy with the validation set. Table 4 shows the highest accuracies achieved by the seven methods with the validation set during the training period, and Figure 6 shows the accuracy curves and loss curves for these seven methods with the IDADP pest validation set during the training period of 500 epochs. It can be observed that: (1) Compared to the existing methods, the MMFGT method proposed in this paper achieved the highest recognition accuracy, precision, recall, and F1 score; (2) The FGT method proposed in this paper outperformed current image recognition methods. This was because the combination of self-supervision and a fine-grained transformer is more suitable for pest recognition under few-shot conditions; (3) The MMFGT model proposed in this paper was more suitable for fine-grained recognition as it includes multimodal information, and the recognition accuracy was improved by 0.8% compared to the FGT method; (4) The overall recognition loss for the ViT, SwinT, DINO, EsViT, FGT, and MMFGT methods with the validation set tended to decrease as the number of epochs increased, but ResNet101 showed overfitting because the dataset was too small and the training set was not similar to the validation set; (5) The image encoder was pretrained with the CUB_200_2011 dataset. The CUB_200_2011 bird dataset is similar to the IDADP pest dataset we used, as both are few-shot and fine-grained datasets, and the text encoder was pretrained with Hugging Face, which gave our proposed MMFGT model highly accurate recognition and improved accuracy speed.

Table 4. Performance of different models with the validation set of the IDADP pest dataset.

Model	ACC (%)	P (%)	R (%)	F1 (%)
ResNet101	68.90	63.79	56.37	52.43
ViT	86.33	86.45	83.50	84.33
SwinT	87.67	87.98	85.05	84.68
DINO	92.20	93.29	91.16	92.07
EsViT	91.20	91.85	91.05	91.23
FGT	97.32	98.67	96.99	97.51
MMFGT	**98.12**	**99.07**	**98.56**	**98.50**

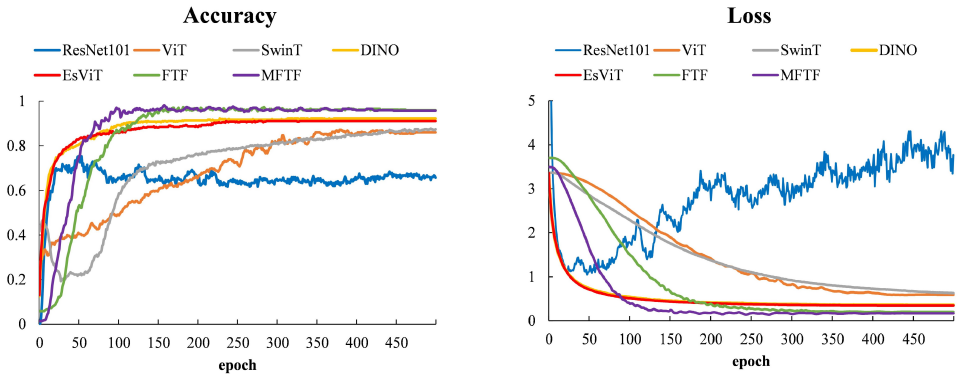


Figure 6. Accuracy curves and loss curves for different pest recognition models with the validation set of the IDADP pest dataset.

Since the stinkbug dataset has more similar data that are more difficult to identify, experiments were conducted with the 15-class stinkbug dataset containing 682 images from the IDADP to further validate the effectiveness of our method. The performances of the different models with the validation set of the IDADP bedbug dataset are shown in Table 5. The accuracy curves and loss curves for the different pest recognition models with the IDADP stinkbug validation set are shown in Figure 7. It can be seen that the MMFGT model achieved the highest recognition accuracy, precision, recall, and F1 score, outperforming the existing methods. Specifically, the accuracy of the MMFGT model was 1.27% better than the FGT method, 5.73% better than EsViT, 5.21% better than DINO, 12.5% better than SwinT, 7.82% better than the ViT method, and 22.92% better than ResNet101.

Table 5. Performance of different models with the validation set of the IDADP stinkbug dataset.

Model	ACC (%)	P (%)	R (%)	F1 (%)
ResNet101	74.48	67.83	62.90	63.59
ViT	89.58	79.39	74.42	75.36
SwinT	84.90	73.84	73.08	73.06
DINO	92.19	92.89	91.66	92.07
EsViT	91.67	92.32	91.97	91.84
FGT	96.13	98.42	96.11	97.16
MMFGT	**97.40**	**98.96**	**96.85**	**97.20**

In addition, validation was performed with the remaining 14-class pest datasets (containing 611 images) in the IDADP dataset, without the stinkbug dataset. The performances of the different models with the validation set for the rest of the IDADP datasets are shown in Table 6. The accuracy curves and loss curves for the different pest recognition models with the remaining class-validation IDADP sets are shown in Figure 8. It can be seen that, due to the low similarity between the remaining class datasets (excluding the stinkbug dataset), the large differences in pest characteristics, and the small number of datasets, the recognition accuracy of the MMFGT model was high, achieving 100% for accuracy,

precision, recall and F1 score. Specifically, the accuracy of the MMFGT model was 3.87% better than the FGT method, 7.18% better than EsViT, 7.73% better than DINO, 15.53% better than SwinT, 14.36% better than the ViT method, and 27.07% better than ResNet101. From the experimental results for the stinkbug dataset and the rest of the datasets, it can be seen that the MMFGT model still showed the best performance with smaller and more similar datasets in comparison to the advanced baseline.

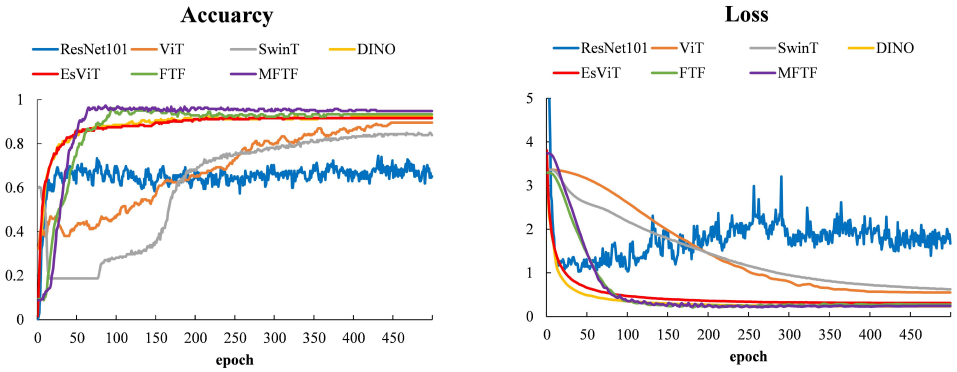


Figure 7. Accuracy curves and loss curves for the different pest recognition models with the validation set of the IDADP stinkbug dataset.

Table 6. Performance of different models with the validation set for the rest of the IDADP datasets.

Model	ACC (%)	P (%)	R (%)	F1 (%)
ResNet101	72.93	66.37	65.65	64.53
ViT	85.64	85.43	85.25	84.72
SwinT	84.53	85.15	84.78	84.37
DINO	92.27	92.45	91.97	91.48
EsViT	92.82	93.01	91.95	91.82
FGT	96.13	98.31	96.37	97.49
MMFGT	**100.00**	**100.00**	**100.00**	**100.00**

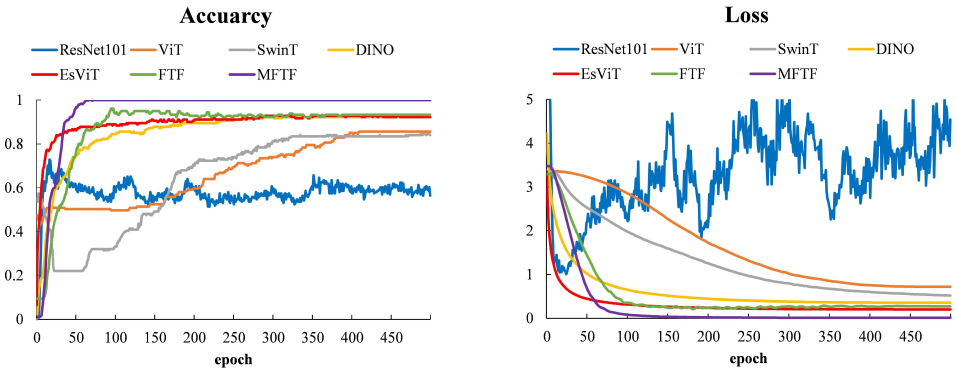


Figure 8. Accuracy curves and loss curves for the different pest recognition models with the validation set for the rest of the IDADP datasets.

4.5. Generalization Performance of the Model

To validate the applicability of our proposed method in different scenarios, experiments were conducted with a publicly available dataset containing eight species of tomato pests. The performances of the different models with the validation set of the public tomato pest dataset are shown in Table 7. The accuracy curves and loss curves for the different pest recognition models with the public tomato validation set are shown in Figure 9. It can be seen that the precision of the FGT method was slightly higher than that of the MMFGT model, and both the recall and the F1 score were the highest for the MMFGT

model. Specifically, the accuracy of the MMFGT model was 0.59% better than the FGT method, 4.76% better than EsViT, 4.16% better than DINO, 10.71% better than SwinT, 11.9% better than the ViT method, and 19.64% better than ResNet101. These experimental results show that the MMFGT method still achieved superior performance with the public dataset.

Table 7. Performances of different models with the validation set of the public tomato pest dataset.

Model	ACC (%)	P (%)	R (%)	F1 (%)
ResNet101	76.19	63.50	67.59	64.57
ViT	83.93	90.78	77.00	76.58
SwinT	85.12	86.29	85.42	84.03
DINO	91.67	91.88	91.38	90.29
EsViT	91.07	91.67	90.82	90.04
FGT	95.24	**96.23**	93.05	94.28
MMFGT	**95.83**	96.12	**96.24**	**96.16**

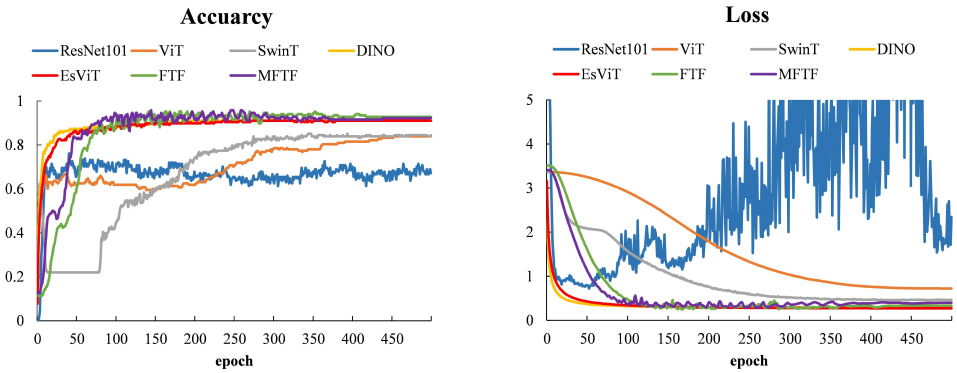


Figure 9. Accuracy curves and loss curves for the different pest recognition models with the validation set of the public tomato validation set.

4.6. Visual Analysis

To further explain the role of each module of the FGT model, several IDADP pest images were visualized, and the visualization results of the different methods are shown in Figure 10. The importance of each region is represented by the color from low to high with blue, green, yellow, and red. As can be seen from Figure 10a–e, the ResNet101 model could not accurately focus on the pest images and the attention images were scattered. The ViT model could focus roughly on the pest image due to the attention mechanism, while DINO could focus more precisely on the pest image compared to the ViT model due to the self-supervision mechanism, and the FGT model could better focus on the head, thorax, and tail of the pests compared to DINO due to the fine-grained mechanism. However, for cases where the proportion of pests in the image was too small and the pests were very close to the background, such as in Figure 10f, the FGT method and the existing Resnet, ViT, and DINO methods were disturbed by the background and could not accurately identify the image, thus reducing the recognition accuracy. In summary, the FGT method could generally improve recognition accuracy by accurately focusing attention on pest-image subdivision regions through the fusion of the attention mechanism, self-supervision mechanism, and fine-grained mechanism, but for cases where the pest target is not obvious, recognition accuracy still needs to be improved.

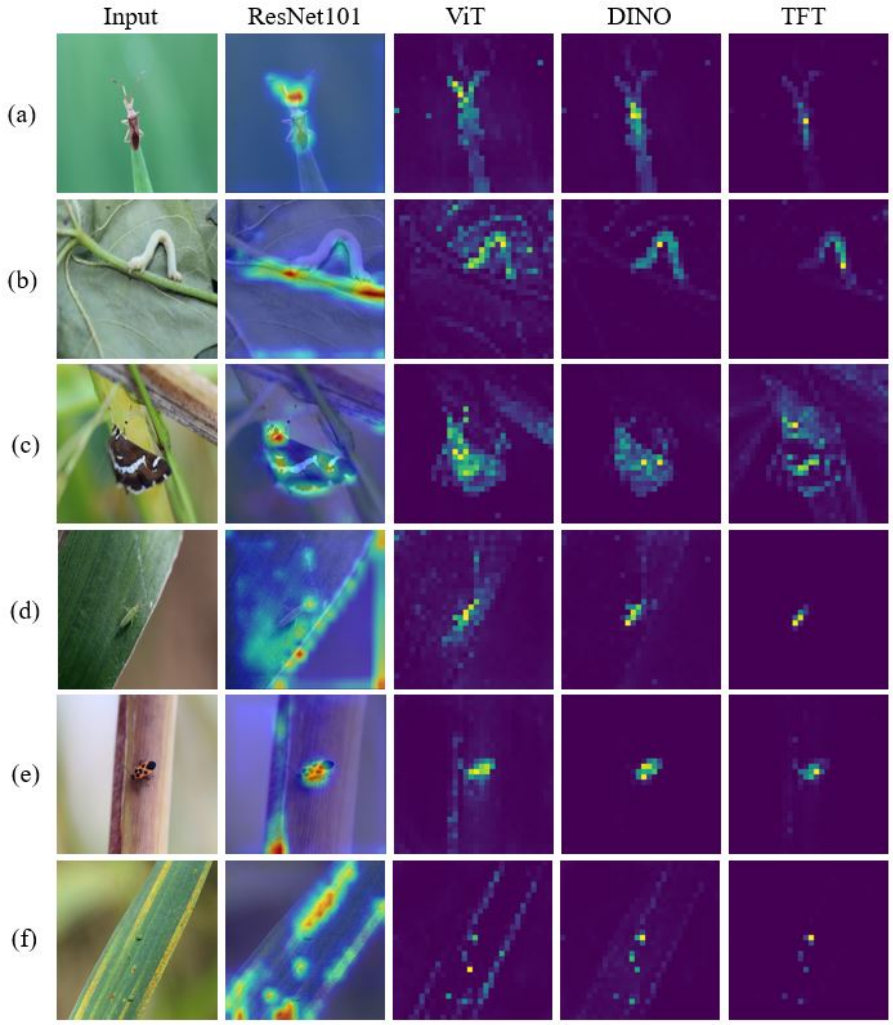


Figure 10. Visualization results of the different methods. (**a**) *Cletus punctiger* Dallas, (**b**) *Ascotis selenaria* Schiffermuller et Denis, (**c**) *Spoladea recurvalis*, (**d**) *Trigonotylus ruficornis* Geoffroy, (**e**) *Tropidothorax elegans* Distant, (**f**) *Aphidoidea*.

4.7. Ablation Experiments

Ablation experiments were performed to demonstrate the role of each module in the model. The modules used by each method and their recognition accuracy are shown in Table 8. It can be seen that the MMFGT method proposed in this paper significantly outperformed the baseline method for the pest recognition task with the few-shot dataset. First, it can be observed that the ViT method achieved 17.43% improved accuracy compared to ResNet101 when classifying the 29-class IDADP pest dataset and 7.74% improved accuracy for the dataset containing eight tomato pests. This was because the attention mechanism focused attention on the pest images without overfitting problems, unlike the CNN model. Secondly, the transformer architecture was extended by self-supervised learning, which can reduce dependence on data volume and is suitable for few-shot image recognition problems. The experimental results show that, compared with the ViT method, DINO achieved 5.87% improved classification accuracy for the 29-class IDADP pest dataset and 7.74% improved accuracy for the dataset containing eight species of tomato pests, which verified the above conclusions. The improved fine-grained transformer architecture made it possible to focus attention on the subdivision regions of the image, thus increasing the accuracy of pest recognition.

To better compare the performance of the models, as shown in Table 9, we compared the training time, inference time, number of parameters, and accuracy of the ViT, SwinT,

DINO, EsViT, FGT, and MMFGT methods with the IDADP dataset. It can be seen that our proposed FGT method and MMFGT method obtained higher accuracy at the cost of more training and longer inference time compared to the baseline method.

Table 8. The modules for each method and their accuracy rates.

Method \ Module	Attention Mechanism	Self-Supervised Learning	Fine-Grained Mechanism	Text	IDADP ACC (%)	Tomato ACC (%)
ResNet101					68.9	76.19
ViT	✓				86.33	83.93
SwinT	✓				87.67	85.12
DINO	✓	✓			92.20	91.67
EsViT	✓	✓			91.20	91.07
FGT	✓	✓	✓		97.32	95.24
MMFGT	✓	✓	✓	✓	**98.12**	**95.83**

Table 9. Training time, inference time, number of parameters, and accuracy of the different models with the IDADP dataset.

Model	Training Time (s)	Inference Time (ms)	Parameters (M)	ACC (%)
ViT	2067	5.78	86	84.33
SwinT	2432	6.95	88	84.68
DINO	2217	6.75	85	92.07
EsViT	2864	7.64	87	91.23
FGT	3164	8.81	85	97.51
MMFGT	3595	11.44	97	98.50

4.8. Discussion

It can be seen that, compared to DINO, our proposed FGT achieved 5.12% improved accuracy in classifying the 29-class IDADP pest dataset and 3.57% improved accuracy in classifying the dataset containing eight species of tomato pests. The multimodal features of the joint image and text representation of the pests enriched the input information and thus enhanced the pest recognition. It can be observed from the experimental results that, compared to the FGT method, the accuracy of the MMFGT method was improved by 0.8% when classifying the 29-class IDADP pest dataset and by 0.59% when classifying the dataset containing eight species of tomato pests. In summary, as it integrates the self-supervised transformer architecture, fine-grained recognition, and multimodal information, the MMFGT model is more suitable for solving the few-shot pest recognition problem and has good prospects compared to the currently existing image classification methods. However, for aphid recognition, as shown in Figure 10f, the MMFGT model could not accurately focus on the pest. This was because the pest was too small in proportion to the image and very close to the background, with high background interference. Ongoing research will address this issue by incorporating multiple fine-grained attention mechanisms or by adding more precise textual description information.

5. Conclusions

In this work, a new multimodal fine-grained transformer architecture (MMFGT) was proposed for pest recognition. Specially, the MMFGT model improved on the transformer architecture in three aspects, making it well-suited for few-shot pest recognition. Firstly, the MMFGT model extracted target features using self-supervised learning to improve recognition accuracy and reduce the reliance on data volume. Secondly, focusing attention on subdivision regions of pest images, the MMFGT model overcame the challenge represented by pest images with small proportions of pest targets, which are difficult to identify accurately. Moreover, the performance of the fine-grained pest recognition was further improved by exploiting joint multimodal information from images and natural language

descriptions of pests. The experimental results demonstrated the superior performance of our method compared to the existing baselines; i.e., the MMFGT model achieved more competitive results compared to several advanced image recognition methods in the pest recognition task, with recognition accuracy up to 98.12% with the IDADP dataset and a 5.92% improvement compared to the state-of-the-art DINO method for the baseline. However, when the proportions of pests in the images were too low and the pests were very close to the background, it was difficult for the MMFGT model to perform accurate pest recognition. In the future, we will address this issue by incorporating multiple fine-grained attention mechanisms or by adding more precise textual description information.

Author Contributions: Methodology, Y.Z.; validation, Y.Z.; investigation, L.C. and Y.Y.; resources, L.C. and Y.Y.; data curation, L.C. and Y.Y.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z. and L.C.; visualization, Y.Z.; supervision, L.C. and Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (grant nos 32071901 and 32271981) and the database of the National Basic Science Data Center (no. NBSDC-DB-20).

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this paper:

MMFGT	multimodal fine-grained transformer model
CNN	convolutional neural network
ViT	vision transformer
IDADP	Image Database for Agricultural Diseases and Pests Research
FGT	fine-grained transformer model
BERT	bidirectional encoder representations from transformer
ALBERT	a lite BERT
PSM	part selection module
SwinT	hierarchical vision transformer using shifted windows
DINO	a form of self-distillation with no labels
EsViT	efficient self-supervised vision transformer

References

1. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.Q.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [CrossRef] [PubMed]
2. Dawei, W.; Limiao, D.; Jiangong, N.; Jiyue, G.; Hongfei, Z.; Zhongzhi, H. Recognition pest by image-based transfer learning. *J. Sci. Food Agric.* **2019**, *99*, 4524–4531. [CrossRef] [PubMed]
3. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. In Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, 3–7 May 2021.
4. Cheng, X.; Zhang, Y.; Chen, Y.; Wu, Y.; Yue, Y. Pest identification via deep residual learning in complex background. *Comput. Electron. Agric.* **2017**, *141*, 351–356. [CrossRef]
5. Ren, F.; Liu, W.; Wu, G. Feature Reuse Residual Networks for Insect Pest Recognition. *IEEE Access* **2019**, *7*, 122758–122768. [CrossRef]
6. Xia, D.; Chen, P.; Wang, B.; Zhang, J.; Xie, C. Insect Detection and Classification Based on an Improved Convolutional Neural Network. *Sensors* **2018**, *18*, 4169. [CrossRef] [PubMed]
7. Huo, M.; Tan, J. Overview: Research Progress on Pest and Disease Identification. In *Lecture Notes in Computer Science, Proceedings of the Pattern Recognition and Artificial Intelligence—International Conference, ICPRAI 2020, Zhongshan, China, 19–23 October 2020*; Lu, Y., Vincent, N., Yuen, P.C., Zheng, W., Cheriet, F., Suen, C.Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12068, pp. 404–415.

8. Brahimi, M.; Boukhalfa, K.; Moussaoui, A. Deep Learning for Tomato Diseases: Classification and Symptoms Visualization. *Appl. Artif. Intell.* **2017**, *31*, 299–315. [CrossRef]
9. Mohanty, S.P.; Hughes, D.P.; Salathé, M. Using Deep Learning for Image-Based Plant Disease Detection. *Front. Plant Sci.* **2016**, *22*, 1419. [CrossRef] [PubMed]
10. Tetila, E.C.; Machado, B.B.; Menezes, G.K.; da Silva Oliveira Junior, A.; Alvarez, M.A.; Amorim, W.P.; de Souza Belete, N.A.; da Silva, G.G.; Pistori, H. Automatic Recognition of Soybean Leaf Diseases Using UAV Images and Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 903–907. [CrossRef]
11. Pattnaik, G.; Shrivastava, V.K.; Parvathi, K. Transfer Learning-Based Framework for Classification of Pest in Tomato Plants. *Appl. Artif. Intell.* **2020**, *34*, 981–993. [CrossRef]
12. Chen, J.; Chen, J.; Zhang, D.; Sun, Y.; Nanehkaran, Y.A. Using deep transfer learning for image-based plant disease identification. *Comput. Electron. Agric.* **2020**, *173*, 105393. [CrossRef]
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
14. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical Vision transformer using shifted windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 9992–10002.
15. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R.B. Momentum contrast for unsupervised visual representation learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; Computer Vision Foundation/IEEE: Piscataway, NJ, USA, 2020; pp. 9726–9735.
16. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G.E. A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, Virtual Event, 13–18 July 2020; Volume 119, pp. 1597–1607.
17. Chen, X.; Fan, H.; Girshick, R.B.; He, K. Improved Baselines with Momentum Contrastive Learning. *arXiv* **2020**, arXiv:2003.04297.
18. Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; Hinton, G.E. Big self-supervised models are strong semi-supervised learners. In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual Conference, 6–12 December 2020.
19. Grill, J.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.H.; Buchatskaya, E.; Doersch, C.; Pires, B.Á.; Guo, Z.; Azar, M.G.; et al. Bootstrap your own latent—A new approach to self-supervised learning. In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual Conference, 6–12 December 2020.
20. Chen, X.; He, K. Exploring simple siamese representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual Conference, 19–25 June 2021; Computer Vision Foundation/IEEE: Piscataway, NJ, USA, 2021; pp. 15750–15758.
21. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 9630–9640.
22. Li, C.; Yang, J.; Zhang, P.; Gao, M.; Xiao, B.; Dai, X.; Yuan, L.; Gao, J. Efficient self-supervised vision transformers for representation learning. In Proceedings of the Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, 25–29 April 2022.
23. Ovalle, J.E.A.; Solorio, T.; Montes-y-Gómez, M.; González, F.A. Gated multimodal units for information fusion. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
24. Kiela, D.; Grave, E.; Joulin, A.; Mikolov, T. Efficient large-scale multi-modal classification. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, LA, USA, 2–7 February 2018; McIlraith, S.A., Weinberger, K.Q., Eds.; AAAI Press: Washington, DC, USA, 2018; pp. 5198–5204.
25. Wang, L.; Li, Y.; Lazebnik, S. Learning deep structure-preserving image-text embeddings. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Piscataway, NJ, USA, 2016; pp. 5005–5013.
26. Faghri, F.; Fleet, D.J.; Kiros, J.R.; Fidler, S. VSE++: Improving visual-semantic embeddings with hard negatives. In Proceedings of the British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, 3–6 September 2018; BMVA Press: Durham, UK, 2018; p. 12.
27. Fukui, A.; Park, D.H.; Yang, D.; Rohrbach, A.; Darrell, T.; Rohrbach, M. Multimodal compact bilinear pooling for visual question answering and visual grounding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, TX, USA, 1–4 November 2016; Su, J., Carreras, X., Duh, K., Eds.; The Association for Computational Linguistics: Toronto, ON, Canada, 2016; pp. 457–468.

28. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; Computer Vision Foundation/IEEE Computer Society: Piscataway, NJ, USA, 2018; pp. 6077–6086.
29. He, X.; Peng, Y. Fine-grained image classification via combining vision and language. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: Piscataway, NJ, USA, 2017; pp. 7332–7340.
30. Nawaz, S.; Calefati, A.; Caraffini, M.; Landro, N.; Gallo, I. Are these birds similar: Learning Branched networks for fine-grained representations. In Proceedings of the 2019 International Conference on Image and Vision Computing New Zealand, IVCNZ 2019, Dunedin, New Zealand, 2–4 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
31. Gallo, I.; Ria, G.; Landro, N.; Grassa, R.L. Image and TEXT FUSION FOR UPMC food-101 using BERT and CNNs. In Proceedings of the 35th International Conference on Image and Vision Computing New Zealand, IVCNZ 2020, Wellington, New Zealand, 25–27 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
32. Zhou, J.; Li, J.; Wang, C.; Wu, H.; Zhao, C.; Teng, G. Crop disease identification and interpretation method based on multimodal deep learning. *Comput. Electron. Agric.* **2021**, *189*, 106408. [CrossRef]
33. Yuan, Y.; Chen, L.; Ren, Y.; Wang, S.; Li, Y. Impact of dataset on the study of crop disease image recognition. *Int. J. Agric. Biol. Eng.* **2022**, *15*, 181–186. [CrossRef]
34. Huang, M.L.; Chuang, T.C. A Database of Eight Common Tomato Pest Images. 2020. Available online: https://data.mendeley.com/datasets/s62zm6djd2/1 (accessed on 12 February 2023).
35. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised learning of language representations. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020.
36. He, J.; Chen, J.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; Wang, C. TransFG: A transformer architecture for fine-grained recognition. In Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022, Virtual Event, 22 February–1 March 2022; AAAI Press: Washington, DC, USA, 2022; pp. 852–860.
37. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Toronto, On, Canada, 2019; Volume 1 (Long and Short Papers), pp. 4171–4186.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Piscataway, NJ, USA, 2016; pp. 770–778.