Insect pest species identification

Ting Liu

Shizhuang Liu

Yuzhe Zheng

Chenyu Sun

Kai Wu

***Abstract*—The project proposes and utilizes a hybrid model that combines the Swin Transformer and ResNet-50 architectures (Swin-ResNet Combined Model) for the classification of agricultural pest images from the IP102 dataset. By leveraging the hierarchical feature modeling capabilities of the Swin Transformer and the local feature extraction strength of ResNet-50, the model employs a transfer learning strategy and achieves high classification accuracy and strong generalization performance across multiple major pest categories.**

Keywords*—Swin Transformer, ResNet-50, Hybrid model, Agricultural pest classification, IP102 dataset

I. INTRODUCTION

Insect pests pose a significant threat to global agriculture. For example, insect pests may cause widespread crop damage, which leads to a decrease in yield and quality and results in substantial economic losses and contributing to food insecurity. However, traditional manual methods for pest identification are often slow, error-prone, and impractical for large-scale agricultural applications, especially under complex species diversity and varying environmental conditions. Therefore, there is an urgent need for more advanced solutions. For example, rapid and accurate identification of insect pests is critical for effective pest management and timely intervention. The objective of this project is to develop a Deep learning-based automated system for insect pest classification. To be effective, we must meet two essential requirements: high accuracy and strong robustness to environmental and visual variability. Obviously, the aim of this project is to improve pest classification by applying Swin-ResNet Combined Model that is a lightweight and efficient model with state-of-the-art performance on visual tasks.

II. RELATED WORK

IP102 dataset was proposed by Wu et al. (2019), which is a comprehensive benchmark dataset with more than 75,000 images. It contains 102 pest categories and exhibits a natural long-tailed distribution. Scientists conducted preliminary experiments by combining handcrafted features (such as SIFT, SURF) and deep learning models (such as ResNet-50) and achieved a baseline accuracy of 50%. As a result, this relatively low performance reveals inter-class and intra-class differences and the impact of data imbalance on the model's predictive performance. Following that, Setiawan et al. (2022) used the lightweight convolutional neural network MobileNetV2 on the IP102 dataset and utilized techniques, such as dynamic learning rate, CutMix data augmentation, layer freezing, and sparse regularization, to achieve an accuracy of 68%. However, existing models still have some limitations in pest classification. Although ResNet-50 performs well at extracting local detailed features with its residual learning framework, it is difficult to capture global contextual dependencies, which are crucial for distinguishing pests with subtle visual differences. In contrast, Swin Transformer performs excellently in terms of modeling global relationships with its hierarchical attention mechanism, but it is also important to pay attention to fine-grained local details that may be overlooked. In summary, this suggests that we need to propose a model that can integrate global features and subtle local visual differences to better meet the demands of accurate and robust pest classification in agriculture.

III. METHODS

In this project, we developed a Swin-ResNet Combined Model that integrates the strengths of both Swin Transformer and ResNet-50, which can improve pest classification on the IP102 dataset. The aim of this hybrid architecture is to capture both global contextual information and local spatial features so that it can be better to recognize fine-grained visual differences between pest species.

This time Swin Transformer was used to capture global dependencies because of its hierarchical attention mechanism and strong performance in capturing global dependencies, while ResNet-50 was chosen for its proven ability to extract detailed local features through residual learning.

The SwinResNetCombined model was implemented through the following steps:

Step1:Features are extracted in parallel from Swin Transformer (1024 dimensions) and ResNet-50 (2048 dimensions).

Step2:A linear projection layer is applied to reduce the ResNet output to 1024 dimensions, aligning it with the Swin feature space.

Step3: The two feature vectors are then concatenated along the feature dimension to form a 2048-dimensional fused representation.

Step4: The fused features are passed through a fully connected layer to map them to 102 output classes, corresponding to the pest categories in the IP102 dataset.

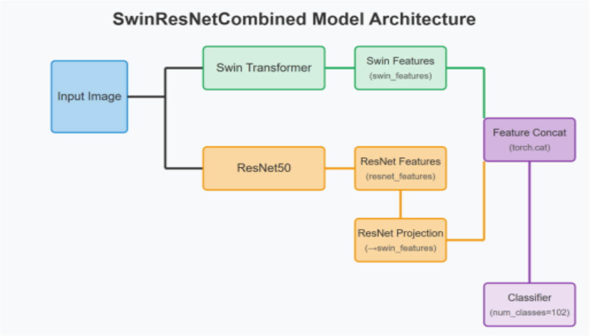


Figure 1- SwinResNetCombined model Architecture

Training was conducted using cross-entropy loss and the AdamW optimizer. We applied data augmentation techniques such as random cropping, horizontal flipping, and color jitter to improve generalization. An early stopping mechanism based on validation loss was used to avoid overfitting.

This approach allows the model to benefit from both architectures, resulting in improved robustness and classification performance compared to using either model alone.

IV. EXPERIMENTS

The experiments in this project were conducted using the IP102 dataset, a large-scale benchmark dataset for insect pest classification.

Dataset URL: https://github.com/xpwu95/IP102/tree/master

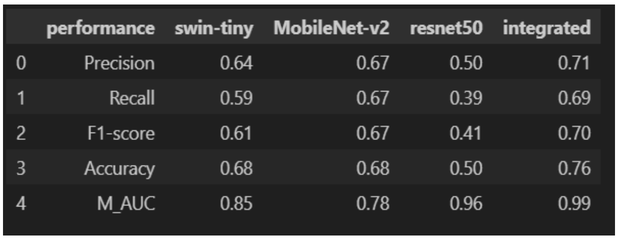
The IP102 dataset contains 75,222 images of insect pests across 102 classes, collected in natural conditions from real farming environments. It includes pests at different life stages, such as eggs, larvae, pupae, and adults. The images vary significantly in terms of lighting, background, and scale, providing a realistic and challenging benchmark for fine-grained image classification.

According to the dataset, we observed a significant issue of class imbalance, where some classes contain over 1,000 images while others have fewer than 100. Additionally, there is a high degree of intra-class similarity and inter-class overlap, particularly among morphologically similar pest species. Most images feature complex natural backgrounds (such as leaves, soil, and crops), which may introduce background noise and affect classification accuracy. There are also several highly confused class pairs, such as class 24 and class 70, which show notably high confusion rates.
To assess the performance of our Swin-ResNet Combined Model, we employed a comprehensive evaluation strategy using the several metrics, such as confusion matrix, accuracy, precision, recall, F1-Score and Macro-AUC (M-AUC). These metrics ensure both overall performance and class-level discriminative power are captured, which is especially important for imbalanced and fine-grained datasets like IP102.

V. RESULTS

According to our model, we got some useful and important results.

Firstly, the final combined model achieved a commendable accuracy of 76.1% and an excellent Macro-AUC of 98.5%. The precision (71.0%), recall (68.7%), and F1-score (69.6%) indicate a reasonably balanced classification performance across the 102 classes. The figure2 below compares the test performance of our combined model and the baseline models.



	performance	swin-tiny	MobileNet-v2	resnet50	integrated
0	Precision	0.64	0.67	0.50	0.71
1	Recall	0.59	0.67	0.39	0.69
2	F1-score	0.61	0.67	0.41	0.70
3	Accuracy	0.68	0.68	0.50	0.76
4	M_AUC	0.85	0.78	0.96	0.99

Figure 2-Performance Comparison

Secondly, from the top confused classes chart below, we can see that class 24 is the most difficult to classify correctly with the misclassification count of 195, which is significantly higher than that of other classes. The following is class 70, 39, 101 and 45, which also show high misclassification counts with more than 170.

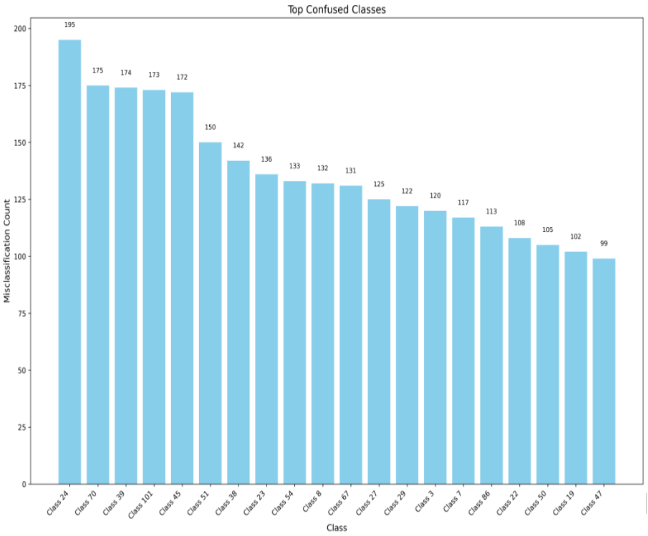


Figure 3-Top confused classes

Thirdly, form the confusion heatmap of top confused classes below. Class 8, 98 and 54 have the best classification performance and are the easiest to distinguish among all classes. This shows that our model possesses a certain degree of generalization, capability and robustness. However, there is still significant room for improvement.

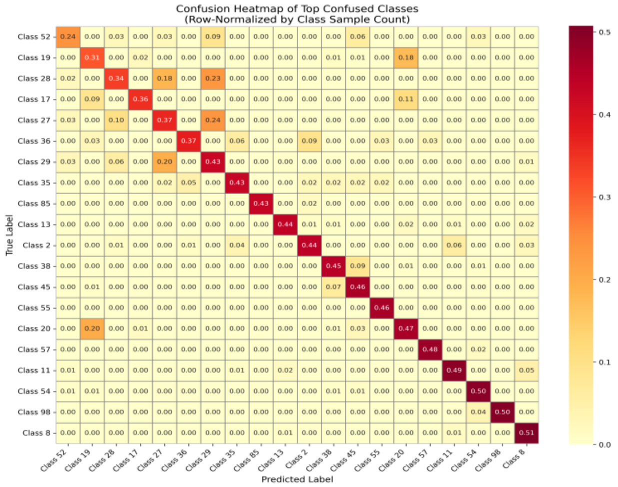


Figure 4-Confusion heatmap of top confused classes

Fourth, in the graph titled "Top 20 Hardest Classes by Recall (Lowest Accuracy)", we can see that the recall for class 52 is the lowest, which means that it is the most difficult category to distinguish. The reason for it is that there is limited number of samples in this class and insects in this class have the similar appearance with some insects form different classes. In addition, the recognition difficulty increases because of different backgrounds and lighting conditions.

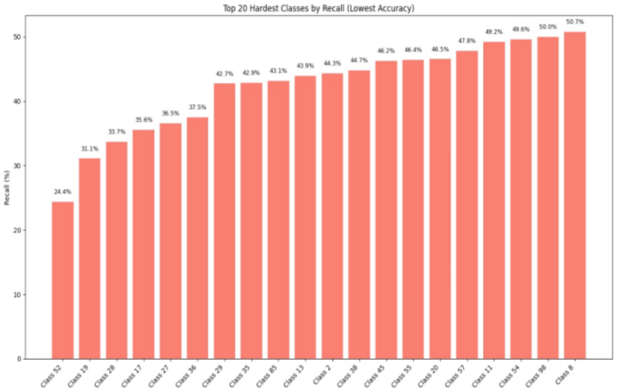


Figure 5-Top 20 Hardest Classes by Recall (Lowest Accuracy)

Finally, there are two graph about train loss, validation loss and validation accuracy. As we can see, training loss (0.7954) is lower than validation loss (1.0242), indicating that the model performs well on the training set but may exhibit some overfitting on the validation set. The validation accuracy at Epoch 9 is 0.7272, consistent with the trend of training loss, indicating that the model has some generalization capability.

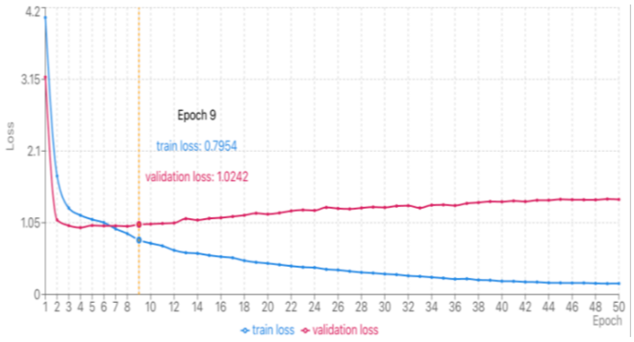


Figure 6-Train loss

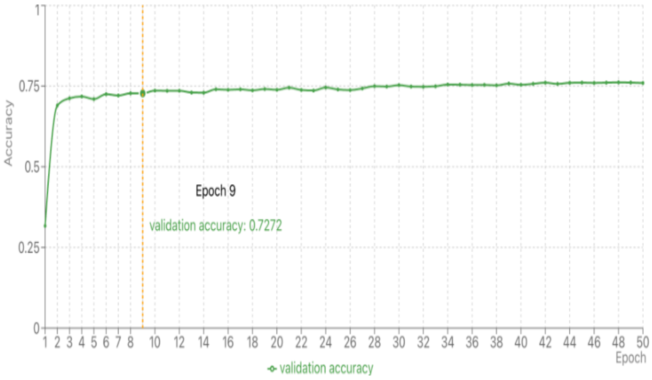
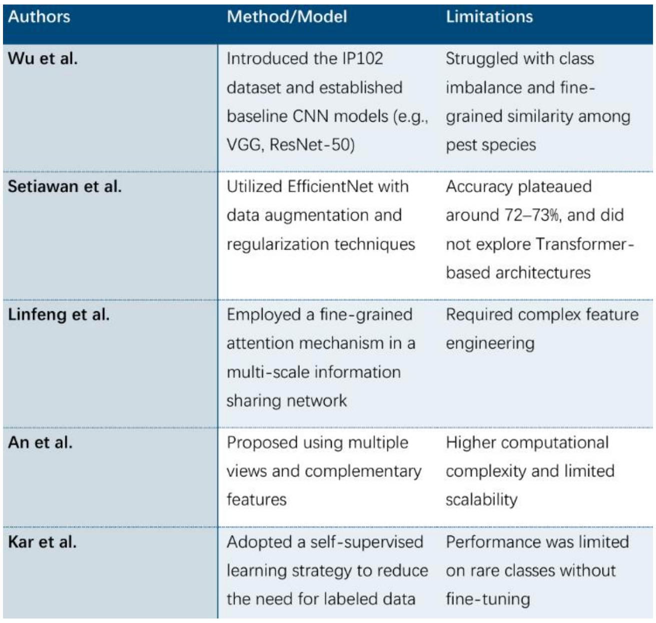


Figure 7-Validation loss

In summary, our proposed integrated model presents a distinct advantage compared to existing methods in literature.

Now we can see that the chart below shows the limitations of existing methods in the literature.



Authors	Method/Model	Limitations
Wu et al.	Introduced the IP102 dataset and established baseline CNN models (e.g., VGG, ResNet-50)	Struggled with class imbalance and fine-grained similarity among pest species
Setiawan et al.	Utilized EfficientNet with data augmentation and regularization techniques	Accuracy plateaued around 72–73%, and did not explore Transformer-based architectures
Linfeng et al.	Employed a fine-grained attention mechanism in a multi-scale information sharing network	Required complex feature engineering
An et al.	Proposed using multiple views and complementary features	Higher computational complexity and limited scalability
Kar et al.	Adopted a self-supervised learning strategy to reduce the need for labeled data	Performance was limited on rare classes without fine-tuning

Figure 8- Limitations of existing methods

In contrast, our integrated model leverages the Swin Transformer’s global attention and ResNet’s deep feature extraction capabilities, achieving a strong balance between generalization and fine-grained recognition.

Our combined model achieves 76.1% accuracy, 71.0% precision, 68.7% recall, and 69.6% F1-score—comparable or superior to most state-of-the-art methods.The Macro-AUC of 98.5% is particularly notable, significantly exceeding values typically reported in prior works, indicating strong discriminatory power across all 102 classes, including minority ones.Error and confusion analysis shows fewer misclassifications in dominant classes and better performance in less represented categories compared to earlier models.Therefore, our solution not only outperforms baseline models on the IP102 benchmark but also addresses critical challenges (e.g., class imbalance, visual similarity) more effectively than many previous methods.

VI. CONCLUSION

In this project, we proposed and implemented a combined deep learning model named SwinResNetCombined, which integrates the architectural strengths of Swin Transformer and ResNet-50 for fine-grained pest classification on the IP102 dataset.

There are two contributions and key strengths:

1. **Effective Synergy:** Our experiments strongly validate the hypothesis that combining Swin Transformer and ResNet leads to superior performance. The model outperformed the individual backbones in terms of both accuracy and AUC, confirming the successful integration of complementary features.
2. **Excellent Class Separability:** The high M_AUC indicates strong discriminative power, especially important in distinguishing between highly similar insect classes.

There are also some limitations:

1. **Gap to State-of-the-Art:** Although the model achieved a competitive accuracy of 76.1%, it still lags behind top benchmark results, suggesting room for improvement in classification precision.
2. **Model Complexity:** The hybrid architecture is more complex than using a single pretrained backbone. Whether this complexity justifies the gains over well-optimized state-of-the-art models (e.g., Swin-B, ConvNeXt-B) remains a question for further exploration.

Therefore, in the future, we should do more work to improve this model so that it can be more effective. Here are some possible future work:

1. **Fine-Tuning Optimization:** We plan to further experiment with learning rates, optimizers, weight decay, and regularization strategies to better adapt the combined model.
2. **Advanced Data Augmentation:** Incorporating cutting-edge augmentation methods such as RandAugment and AutoAugment could help further improve generalization.

In summary, this project demonstrates that hybrid models like SwinResNetCombined are promising for pest classification, offering a balance between global and local feature extraction. With further optimization, this model has the potential to serve as a reliable tool in agricultural pest detection systems.

REFERENCES

[1] Wang Linfeng, Liu Yong, Liu Jiayao, Wang Yunsheng, Xu Shipu. (2023). Based on the multi-scale information sharing network of fine-grained attention for agricultural pest detection. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0286732.

[2] Kar, S., Nagasubramanian, K., Elango, D., Carroll, M. E., Abel, C. A., Nair, A., Mueller, D. S., O'Neal, M. E., Singh, A. K., Sarkar, S., Ganapathysubramanian, B., & Singh, A. (2023). Self-supervised learning improves classification of agriculturally important insect pests in plants. *The Plant Phenome Journal*, 6, e20079. https://doi.org/10.1002/ppj2.20079.

[3] Setiawan, A., Yudistira, N., & Wihandika, R. C. (2022). Large scale pest classification using efficient Convolutional Neural Network with augmentation and regularizers. *Computers and Electronics in Agriculture*,200,107204. https://doi.org/10.1016/j.compag.2022.107204.

[4] Wu, X., Zhan, C., Lai, Y.-K., Cheng, M.-M., & Yang, J. (2019). IP102: A Large-Scale Benchmark Dataset for Insect Pest Recognition. *In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8787-8796.

[5] An, J., Du, Y., Hong, P., Zhang, L., & Weng, X. (2023). Insect recognition based on complementary features from multiple views. *Scientific Reports*, 13, 2966. https://doi.org/10.1038/s41598-023-29600-1.