IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 May 2024, accepted 31 July 2024, date of publication 9 August 2024, date of current version 11 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3441321

RESEARCH ARTICLE

Visual Attention Focusing on Fine-Grained Foreground and Eliminating Background Bias for Pest Image Identification

XINYUAN XU[1], HENG LI[1], (Member, IEEE), QI GAO[1], MEIXUAN ZHOU[1], TIANYUE MENG[1], LIPING YIN[2], AND XINYU CHAI[1,3]

[1]School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Technical Center for Animal Plant and Food Inspection and Quarantine of Shanghai Customs, Shanghai 200002, China
[3]Vision Science and Rehabilitation Engineering Laboratory, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding author: Xinyu Chai (xychai@sjtu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFD1400100 and Grant 2021YFD1400102, in part by the National Natural Science Foundation of China under Grant 62103269 and Grant 62073221, and in part by the Med-X Research Fund of Shanghai Jiao Tong University under Grant YG2022QN077.

ABSTRACT Plant diseases and pests caused by harmful insects has always been a significant threat to agricultural and forestry production. In addition, the threat of invasive insects causes damage to local ecosystems with a decrease in biodiversity and even the extinction of some species, seriously harming the local economy. Governments around the world have invested a significant number of efforts in insect detection and control. With the development of AI, automated identification is an irreversible trend to improve efficiency and reduce government input. Recent researches attempt to apply deep learning tools into the detection and identification of insects, but meeting a series of difficulties. Insect identification abstracted to a fine-grained vision classification task provides unique challenges including the small difference between classes and the large difference within a class. In this study, we propose a pest identification model guided by visual attention, designed to address the above challenges. We establish an attention mechanism from these two perspectives, enhancing attention to foreground features by amplifying fine-grained features and eliminating attention to background biases through counterfactual inference. Our approach ultimately achieves a classification accuracy of 74.5% for 102 insect categories on the IP102 dataset, and similarly, achieves an exceptional 99.8% accuracy for 40 insect categories on the D0 dataset. The approach proposed in this study will contribute to the automatic insect detection and identification system in the future as the core technique.

INDEX TERMS Counterfactual inference, deep learning, insect identification, visual attention.

I. INTRODUCTION

Plant diseases and pests caused by harmful insects have become a significant challenge faced by agriculture and forestry globally, severely impacting the production and quality of agricultural and forestry products such as grains, vegetables, fruits, and timber [1]. Pests not only directly result in reduced crop yields and quality decline but may also lead to issues such as pesticide residues and environmental degradation, posing a serious threat to the sustainable development of agriculture and forestry. According to statistical data, global annual crop losses due to pests and diseases amount to billions of dollars, significantly affecting agricultural production and food security. Particularly in developing countries, where agricultural technological levels are relatively low and monitoring and warning systems are inadequate, pest issues are more pronounced, causing significant losses to farmers and national economies. According to [2], with the acceleration of global climate change and the process of globalization, the issue of pest invasions exacerbates the

The associate editor coordinating the review of this manuscript and approving it for publication was Huiyu Zhou.

consequences and challenges of pest and disease control. Governments worldwide have consistently prioritized this issue, investing significant research efforts in establishing pest detection systems. Pest identification, as a crucial component, represents a highly challenging task in this regard.

The realm of artificial intelligence (AI) has witnessed rapid advancements, showing its robust vitality in various fields as a pivotal catalyst for the Fourth Industrial Revolution. Under this technological backdrop, our research goal could be transformed into fine-grained insect classification. It shows challenging work considering the nuanced similarity in insect appearances, the intricacy of color, texture, shape, and the varied complexity found within insect image backgrounds. Through the lens of deep learning, the task of insect classification can be inherently categorized as a fine-grained classification challenge [3]. In contrast to coarse-grained tasks, fine-grained classification deals with subtler distinctions and narrower inter-class variations. Current researches have predominantly been confined to the agricultural sphere including plant pests and diseases. These studies typically leverage common benchmark models to conduct relatively simplistic insect classification, characterized by a limited scope of categories and sample sizes. However, our research delves into a more complex multi-class classification task. It has been observed that benchmark models fail to cover fine-grained tasks [12]. Recent researches [13], [14], [15], [16], [17], [18] attempt to use hybrid attention mechanisms, ensemble models or large language models to achieve their objectives.

In this study, we analyze the unique characteristics of insect classification issues and propose innovative methods to handle the hurdles of fine-grained insect classification. Specifically, we focused on two aspects including fine-grained features enhancement and background biases elimination, which culminated in an efficacious classification of insects without inflating computational demands and simultaneously elevating the model's interpretability. The main contributions of our paper are as follows:

- We have developed a fine-grained feature enhancement approach utilizing a mixed attention mechanism and attention-based resampling.
- We have devised an interpretable counterfactual attention learning methodology to effectively eliminate background biases.

II. RELATED WORKS

In this section, we review existing research work on insect classification and provides an in-depth analysis of the strengths and limitations of existing research.

A. DATASETS

It is found that there are very few large public insect data sets. Most of the published studies collect insect images themselves, and the types and numbers of established data sets are small. The shortage of these studies is the limited application range that they only suit to insect recognition tasks in specific environments or under specific classes or orders. For example, the study by Alves et al. published in 2020 focused on cotton-plant pests [4]. Moreover, environment interference makes it impossible to catch pictures of insects with clean background in the real world. Some research captures insects and collects images under ideal conditions, such as the research on field insect recognition published by Yuan et al. in 2020 [5]. Although clean datasets can obtain better results under specific research objectives, it leads to the poor generalization. In addition, half of the datasets used in related studies also contains multimodal information like sound, dynamic motion, DNA sequences or 3D images, which is very difficult to collect and not conducive to practical applications. The largest public insect dataset available is TensorFlow's open source i_naturalist2017 dataset [6], but this dataset covers 5089 insect species, which is too many categories and too large a volume of data to be suitable for our study. We will conduct this study on the insect dataset IP102 published on 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019) [7]. Dataset IP102 contains tens of thousands of images on 102 categories. The rather dataset size allows the utilization of more advanced computer vision models in this study.

B. DEEP LEARNING MODELS

Deep learning models have been extensively applied to insect classification, with a predominant focus on the agricultural domain in published research. The investigations primarily address pest and disease occurrences within specific botanical gardens, resulting in a relatively constrained scope with a limited number of insect species. Consequently, the classification tasks undertaken in these studies are generally uncomplicated, allowing common models to achieve optimal performance. Noteworthy among the employed models are CNN architectures such as AlexNet, ResNet-50, Inception-v3, which have demonstrated efficacy in achieving commendable classification results on smaller datasets [8], [9], [10]. However, these models exhibit limitations in terms of generalization across diverse datasets. A notable research trend in this domain revolves around optimizing computational efficiency, particularly the realization of lightweight classification models. Although several studies have made advancements in this direction, the overarching objective remains enhancing the models' adaptability to diverse datasets. Notably, only a limited number of studies have extended their investigation to encompass more than 100 species classification tasks. Notable among these is the work by Sagar et al. in 2020 [11], where they successfully employed deep convolutional neural networks to identify over 100 species of Indian butterflies and moths, marking a significant achievement in the field.

C. CURRENT RESEARCHES ON DATASET IP102

The classification of the IP102 dataset poses a notable challenge, prompting extensive research endeavors [12],

[13], [14], [15], [16], [17], [18]. Key contributions in this domain are highlighted in the following studies. In 2019, the creators of the IP102 dataset conducted comprehensive evaluations using various state-of-the-art architectures, including AlexNet, GoogleNet, VGGNet, and ResNet. Notably, ResNet emerged as the top-performing model, achieving a classification accuracy of 49.5%. Addressing the issue of imbalanced data distribution within the IP102 dataset, Reza et al. proposed a solution in 2019 [12] by employing data augmentation techniques such as rotation, translation, and flip. They utilized three deep neural networks—VGG16, ResNet, and Inception-v3—individually for classification, with Inception-v3 demonstrating the highest accuracy at 57.08%. In 2020, Liu et al. [13] introduced a novel approach by constructing a Deep Multi-branch Fusion Residual Network (DMF-ResNet). This architecture facilitated the extraction of image features from three distinct branches, enabling the learning of multi-scale features. The application of DMF-ResNet resulted in a notable improvement, achieving a classification accuracy of 59.22% on the IP102 dataset. These studies collectively contribute to advancing the understanding and methodologies for effective classification tasks on the challenging IP102 dataset.

In 2021, Yang et al. [14] introduced the Convolutional Rebalancing Network (CRN), addressing the challenge of unbalanced datasets by incorporating two sampling methods—balanced sampling and reverse sampling. The model further enhanced classification performance through an image enhancement module based on region cropping and region coverage. The feature fusion module optimized the training process, resulting in a notable classification accuracy of 70.42% on the IP102 dataset. Another significant contribution in 2021 comes from Yang et al. [15], who proposed a model integrating channel attention and spatial attention into the CNN architecture. By combining the Spatial Transformer Network (STN) with the ResNet50 backbone network, the interference of image background on classification tasks was minimized. This approach yielded an impressive classification accuracy of 73.29% on the IP102 dataset. Additionally, Luo et al. [16] presented the Salience-Guided Discriminant Learning Network (SGDL-Net) in 2021, featuring original and fine-grained branches. Coarse-grained features were extracted through primitive branches, and image clipping was guided by the Salient Object Location Module (SOLM). Fine-grained features were subsequently mined by the Fine-Grained Feature Mining Module (FFMM). When utilizing DenseNet121 as the backbone network, the SGDL-Net achieved a classification accuracy of 72.65% on the IP102 dataset. In the same year, Ung et al. [17] proposed a model integrating three modules including Attention Network, Feature Pyramid Network, and Multi-Branch and Multi-Scale Attention Network (MMAL-Net). This model demonstrated remarkable effectiveness, achieving a classification accuracy of 74.13% on the IP102 dataset. There are also some related works [18] published in recent years.

In recent years, thanks to advances in artificial intelligence, Transformers have shown remarkable progress in various fields. While some researches [19] on Transformer-based models excel in representation learning and cross-domain generalization, their massive parameter sizes and extensive pre-training data limit the application in specific domains like agriculture. On the one hand, agricultural images require visual attention avoiding the disturb of background information rather than self-attention with global information. On the other hand, agriculture demands high practical effectiveness, imposing strict constraints on model efficiency and resource consumption. Despite the theoretical potential of large models, their high computational costs hinder widespread adoption in agricultural practices. To address these challenges, our research focuses on exploring model methods suitable for practical implementation in agriculture. We have chosen Convolutional Neural Networks (CNNs) for their efficiency and suitability for agricultural applications, offering advantages in parameter size, computational efficiency, and resource consumption over large models.

However, above studies observed that fine-tuned CNN models exhibited unsatisfactory classification performance on the IP102 dataset. The findings underscore the limitations of relying solely on a single, fine-tuned CNN model for fine-grained classification tasks, particularly in scenarios with small inter-specific differences but substantial intra-specific variations and uneven data distribution. The research indicates that while improved CNN models or transfer learning may be effective for small-scale or distinctly different datasets, they prove insufficient for fine-grained tasks such as insect classification on datasets like IP102. Integrating advanced models, although enhancing classification performance, entails high computational requirements and training costs, presenting challenges that need careful consideration. In summary, the deficiencies of existing researches inspire us to think about solutions from the perspectives of fine-grained issues and data balance.

III. MATERIALS AND METHODS

A. DATA PREPROCESSING

Data Source: In this study, we utilize two main datasets as experimental materials, i.e., IP102 and D0. The dataset IP102 was first proposed in 2019 as a benchmark dataset like ImageNet and COCO, containing 75,222 images from 102 different species of insects. Correctly classifying insects on the dataset IP102 is a challenging task because of the poor quality of images with noisy background especially under serious data imbalance. To make matters worse, this dataset also meets significant obstacle of fine-grained classification, namely, the same class shows difference while different classes show similarity as illustrated in Fig. 1. During implementation, we follow the standard partition ratio to divide dataset IP102 into a training set, a validation set and a test set by 6:1:3.

The dataset D0 was first proposed by Xie et al. in 2018 [20], containing 4508 images from 40 different species of insects. The typical characteristics of this dataset include clean background and large target proportion as shown in Fig. 2. To compare with existing research, we divide dataset D0 into a training set, a validation set and a test set with the ratio of 7:1:2.

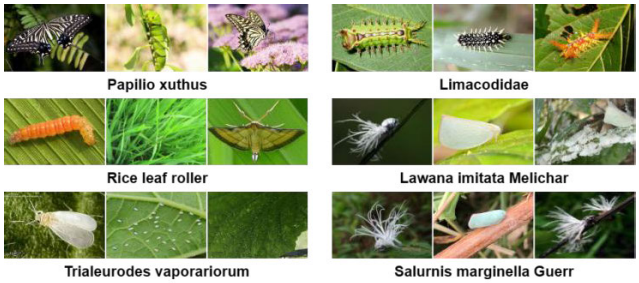

FIGURE 1. Illustrating examples from dataset IP102.

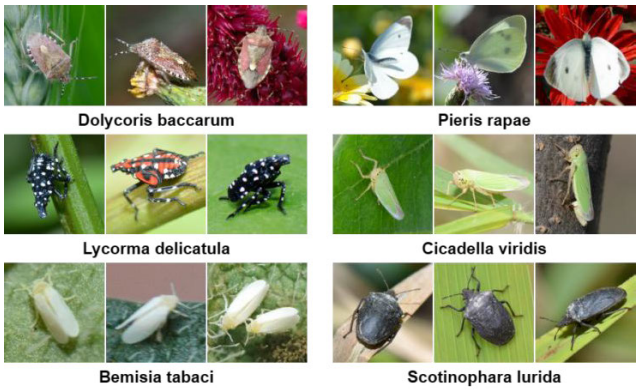

FIGURE 2. Illustrating examples from dataset D0.

Preprocessing: Data preprocessing is a crucial step before we start training our model. Image normalization is an essential step to feed the neural network which often requires input images being resized to 224 × 224 or 448 × 448. Then a set of transformations including random flipping, cropping and color variation are utilized to augment input images. The operation of data augmentation could improve the robustness and generalization of model. Under the conditions of dataset IP102, data augmentation can help getting rid of data noise. Under the conditions of dataset D0, data augmentation can effectively reduce the risk of model overfitting resulted from the limited training sample size.

B. ENHANCING FINE-GRAINED FEATURES

The framework of the proposed end-to-end system is illustrated in Fig. 3 where the two core designs are highlighted in colors.

Network With Attention: In tasks of image processing, especially in fine-grained classification, attention mechanism is indispensable. Attention helps to locate the target objects and extract fine-grained features. Many researches coming out suggesting a variety of different approaches to build attention including spatial attention, channel attention and mixed attention. Inspired by CBAM (Convolutional Block Attention Module) proposed by Woo et al. [21], we add the mixed spatial and channel attention into our identification model. Given the features with the dimension of $\boldsymbol{b} \times \boldsymbol{c} \times \boldsymbol{h} \times \boldsymbol{w}$, we realize the feature weighting on spatial domain and channel domain sequentially. Channel feature weighting first squeezes feature maps by global pooling to obtain the flattened features with the dimension of $\boldsymbol{b} \times \boldsymbol{c} \times \mathbf{1} \times \mathbf{1}$. The flattened features are then fed into MLP layers to generate weight matrix $\boldsymbol{W_c}$. Weighting the feature maps with $\boldsymbol{W_c}$ to obtain feature vectors with the dimension of $\boldsymbol{b} \times \mathbf{1} \times \boldsymbol{h} \times \boldsymbol{w}$, we then realize spatial attention. Similarly, the channel weighted features are convolved and activated to obtain the final weight matrix $\boldsymbol{W}$. The mixed attention can be added into the network by weighting features with $\boldsymbol{W}$.

Attention Resample: Fine-grained objects and discriminative features usually make a small part of the image resulting in the introduction of redundant information after convolution of the input images. Attention mechanism aims to figure out the significant pixels where fine-grained objects and discriminative features lies. Every attention mechanism represents the attention as a weight matrix to indicate the importance of each element's position. In order to make better usage of the learned attention, different methods were used to guide the update of model parameters. A straight-forward approach is to crop out the fine-grained objects from the original images. But the simple cutting operation fails to fully utilize the attention information by exchanging the attention weight matrix into a binary crop mask. Therefore, we adopt an attention-based sampling method to replace the simple cropping, which introduces attention gradient characteristics when cropping images.

Given the input data $\boldsymbol{I}$ ***(with the dimension of*** $\boldsymbol{b} \times \boldsymbol{c} \times \boldsymbol{h} \times \boldsymbol{w}$***)*** and the average attention matrix $\bar{A}$ ***(with the dimension of*** $\boldsymbol{b} \times \mathbf{1} \times \boldsymbol{h} \times \boldsymbol{w}$***)***, we generate a resample network to resample the input data. A first step is the normalization of the attention maps by matching shape with bilinear interpolation. The most crucial step is to generate sampling net according to the attention maps following the principle that the sampling points are more densely distributed higher up in attention weight value. Considering that the spatial distribution of sampling points is two-dimensional, this nonlinear sampling task can be decomposed into two dimensions $\boldsymbol{h}$ and $\boldsymbol{w}$. Taking the $\boldsymbol{h}$ dimension as an example, sampling rate is in direct proportion to the attention weight value on $\boldsymbol{h}$ dimension. First, we calculate the maximum of attention weight on $\boldsymbol{h_i}$ according to attention matrix $\boldsymbol{A_{h \times w}}$:

$$A_{h_i \cdot} = \max_{1 \le w_j \le w} A_{h_i w_j}, \tag{1}$$

the value of $\boldsymbol{A_{h_i}}$. represents the degree of attention. Then we calculate the integral of $\boldsymbol{A_{h_i}}$. on the axis of $\boldsymbol{h}$:

$$S(h)_w = \int_1^h A_{h_i} dh_i = \sum_1^h A_{h_i}, \tag{2}$$

where the gradient of $S(h_i, w)$ represents the sampling density. At this occasion, target nonlinear sampling distribution can be calculated by finding the corresponding coordinate h_{sample} when uniformly sampling $S(h)_w$:

$$h_{sample} = S^{-1}(h)_w. \tag{3}$$

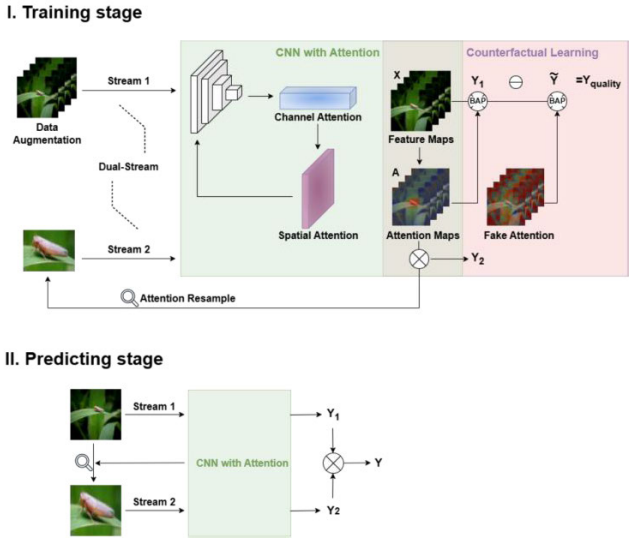

FIGURE 3. Framework of proposed approach (the two core designs are colored). I. In training stage, input batch images are first fed to the feature extractor with a mixed attention module to generate corresponding feature maps and attention maps for further process. The attention guides the image resampling for data augmentation. Then the features are optimized through counterfactual learning module. II. Predicting stage follows the dual-stream strategy, the test image is first fed to the main model to produce a prediction and the attention to guide resampling. The final prediction is calculated by the first prediction and that of the resampled image.

Similar operations on w dimension could lead to target nonlinear sampling distribution w_{sample}. So far, we are able to combine h_{sample} and w_{sample} to build the two-dimensional sampling net. Applying it to the input data I, we use interpolation to generate the resampled image $I_{resample}$ with the same shape as original I.

Dual-Stream Training and Predicting: To implement the above in an end-to-end system, we adopt a dual-stream training and predicting strategy where the first stream achieves the workflow from features learning to attention weighting while the second stream achieves attention-based resampling. As shown in Fig. 3 I, in the model backbone, the feature extractor maps the input image to feature vector. The convolution features of input images are then weighted through bilinear attention pooling (BAP) in the first stream. Given the features $F(\textit{with the dimension of } b \times c \times h \times w)$, we obtain the attention maps A to weight F:

$$\bar{F} = G(F, A) = \frac{1}{h \times w}(F * A), \tag{4}$$

where $*$ means element-wise matrix multiplication. The weighted features $\bar{F}$ are then fed to the fully connected layer to output the first prediction Y_1. In the second stream, attention-based resampling is performed following the main backbone. The resampled images are fed to the beginning of the model as augmented data to output the second prediction Y_2. In the gradient backward stage, Y_1 and Y_2 are respectively calculated by loss function to obtain L_1 and L_2. During prediction stage (shown in Fig. 3 II), the inference result is determined by a weighted summation of Y_1 and Y_2. The weight matrix is selected from a group of suitable weights according to the experience of model training. This strategy will be further analyzed in the Discussion part since it is not the core technique of the study.

C. COUNTERFACTUAL ATTENTION

Enhancing the fine-grained features of foreground objects by attention is efficient to train our model to focus on more discriminative features. But this attention mechanism not always leads to an ideal result. Because it depends on the assumption that we have already found the correct attention which is uncertain. The uncertain attention augments discriminative features as well as redundant features under the former enhancement mechanism leading to the degradation of model performance. We intend to analyze this dilemma through causal inference [22]. The crucial issue is the uncertainty of attention, which stems from the failure on the part of model training to catch sight of the causal relations between attention learning and ground truth. This neglect is resulted from current strategy of attention learning, depending on the loss function based on final prediction to indirectly supervise the quality of the learned attention. Bias brought from datasets produce inaccurate attention and influence model performance under this traditional training strategy. For example, Xylotrechus usually appears on tree trunk while wheat sawfly usually appears on wheat leaves, therefore, it is likely to infer what on the tree trunk is Xylotrechus while what on wheat leaves is wheat sawfly. However, Xylotrechus may also appear on the background of wheat leaves and wheat sawfly may also appear on the background of tree trunk. This example shows a typical bias current attention learning cannot avoid, especially when it comes to long-tailed distribution. Considering the task of insect identification, bias introduced by datasets is represented as background bias. We adopt causal inference to optimize the former attention learning mechanism.

Causal inference makes an explainable connection between the learned attention and class prediction. A traditional workflow of attention model includes the following steps as shown in Fig. 3: (1) a CNN backbone extracts feature maps X from input images, (2) the attention module generates attention maps A from input features, (3) (X, A) jointly predict the final outcome $Y = G(X, A)$, (4) the model trains to update parameters by supervising the consistency between class prediction and ground truth. Traditional attention model overlooked the quality of the learned attention could impact the model prediction. We adopt counterfactual causal inference to guide the attention learning by quantifying the contributions of attention to proper predictions.

Counterfactual causal inference jumps out of the positive feedback loop and shift the focus from the established facts to the contrary possibilities. In the assessment of attention quality, counterfactual causal inference establishes a fake attention called counterfactual attention $\tilde{\boldsymbol{A}}$ to intervene the learning procedure. $\tilde{\boldsymbol{A}}$ is generated through uniform probability distribution and then throwed into the same workflow as original attention maps $\boldsymbol{A}$. $(\boldsymbol{X},\tilde{A})$ jointly predict the counterfactual outcome $\tilde{Y}$:

$$\tilde{Y} = G(X, \tilde{A}), \tag{5}$$

which represents the wrong predictions produced by wrong attentions. Therefore, the quality of the learned attention can be represented as follows:

$$Y_{quality} = Y - \tilde{Y}. \tag{6}$$

During training, $\boldsymbol{Y_{quality}}$ is added into the objective function to adapt loss function:

$$L_3 = L_{ce}(Y_{quality}, y). \tag{7}$$

In addition, we utilize center loss $\boldsymbol{L_4}$ to assess feature center. Together with the loss $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$ of dual-stream procedure, the final loss function is defined as:

$$L_{final} = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3 + L_4, \tag{8}$$

where $\boldsymbol{\lambda_1}$, $\boldsymbol{\lambda_2}$ and $\boldsymbol{\lambda_3}$ are hyper-parameters optimizing the training process.

D. EVALUATING INDEX

In order to evaluate the performance of our identification model, we use the metrics of accuracy on test set, i.e., Top-1 accuracy. Accuracy is a popular evaluation index can be expressed as the following equation:

$$Acc = \frac{\sum_i^N y_i}{N}, \tag{9}$$

where $\boldsymbol{N}$ represents the total numbers of test samples, $\boldsymbol{y_i}$ represents the correctness of the $\boldsymbol{i}$th prediction:

$$y_i = \begin{cases} 1, \textit{ground truth} = 1st \textit{ prediction} \\ 0, \textit{ground truth} \neq 1st \textit{ prediction} \end{cases} \tag{10}$$

IV. RESULTS

We assess the performance of proposed method on dataset IP102 and D0. This section details the implementation of experiments and lists the results of ablation and comparison research.

A. IMPLEMENTATION

All the experiments are carried out on a single RTX 3090 GPU. We utilize ResNet50 and ResNet101 as benchmark models to design the experiments. Considering the matching of computing resources and different size of dataset IP102 and D0, images are resized to $\mathbf{224 \times 224}$ and $\mathbf{448 \times 448}$ respectively. We apply random cropping and flipping as data augmentation to improve model performances. During training phase, we use stochastic gradient descent (SGD) optimizer with the initial learning rate of $\mathbf{10^{-3}}$, momentum of 0.9 and weight decay of $\mathbf{10^{-5}}$. The batch size is set to 32 and the maximum of epochs is set to 100. Hyper-parameters in the loss function are set to $\boldsymbol{\lambda_1 = 1/3, \lambda_2 = 2/3, \lambda_3 = 1}$.

B. ABLATION EXPERIMENTS

Previous section detailed the two crucial techniques constituting our proposed method. We design ablation experiments to compare our method with the baseline model whose results are shown in Table 1. Baseline model is set as a classical ResNet with BAP optimization. The first ablation experiment aims to verify the effectiveness of fine-grained enhancement designed to teach the model to focus on fine-grained foreground. The proposed dual-stream strategy of attention weighing and resampling successfully improve the baseline performance on dataset IP102 and D0 by 1.3% and 0.4% respectively. The second ablation experiment aims to verify the effectiveness of counterfactual causal inference designed to eliminate the background bias. The proposed counterfactual attention learning successfully improve the model performance on dataset IP102 and D0 by 1.1% and 0.2% respectively. All the ablation experiments executed with backbone ResNet50 are repeated with backbone ResNet101 indicating the same conclusion as that on ResNet50. Furthermore, a stronger backbone like ResNet101 can lift the identification performance leading to the best results of 74.5% in this study.

TABLE 1. Results of ablation experiments indicating the effectiveness of each module proposed.

Dataset	Methods	ACC (Resnet50 backbone)	ACC (Resnet101 backbone)
IP102	baseline	71.7%	72.9%
	baseline+att	73.0%	74.1%
	baseline+att+CAL	**74.1%**	**74.5%**
D0	baseline	96.6%	97.5%
	baseline+att	97.0%	99.6%
	baseline+att+CAL	**97.2%**	**99.8%**

C. COMPARISON WITH PREVIOUS RESEARCHES

We compare our method with the up-to-date and best results approaches in Table 2. The best results of our experiments reach state-of-the-art stage among homogeneous studies. In this study, identification accuracy on dataset IP102 and D0 reaches the best 74.5% and 99.8% respectively, outperforming other researches.

Table 3 and 4 shows the top 10 classes with the highest identifying accuracy and the lowest accuracy implementing on dataset IP102. Analyzing the tables that identification accuracy differs in different species, we find that it has a lot to do with the dataset itself. Three characteristics of the classes with low accuracy are figured out comparing with that high accuracy: (1) multiple objects in a image, (2) multiple forms

TABLE 2. Comparisons of accuracy with the SOTA insect classification researches on IP102 and D0.

Dataset	Models based on CNN	ACC
IP102	ResNet [7]	49.5%
	Inception-v3 [12]	57.1%
	DMF-ResNet [13]	59.2%
	CRN [14]	70.4%
	STN [15]	73.3%
	SGDL-Net [16]	72.7%
	EM [17]	74.1%
	Dise-Efficient [18]	64.4%
	Ours	**74.5%**
D0	D0 (original paper) [20]	89.3%
	EM [17]	99.8%
	Ours	**99.8%**

in the same class, (3) few samples in these classes caused by long-tail distribution.

TABLE 3. Top 10 classes with the highest identifying accuracy in dataset IP102.

Number	Class name	Accuracy
1	Papilio xuthus	99.3%
2	mole cricket	98.2%
3	oides decempunctata	96.5%
4	Locustoidea	95.5%
5	Pieris canidia	95.0%
6	wheat blossom midge	93.8%
7	Aleurocanthus spiniferus	92.8%
8	grub	92.2%
9	Limacodidae	92.2%
10	flea beetle	91.6%

TABLE 4. Top 10 classes with the lowest identifying accuracy in dataset IP102.

Number	Class name	Accuracy
1	therioaphis maculata Buckton	15.4%
2	green bug	25.5%
3	large cutworm	26.4%
4	Bactrocera tsuneonis	31.4%
5	white margined moth	33.3%
6	Polyphagotars onemus latus	34.6%
7	english grain aphid	35.0%
8	bird cherry-oataphid	37.1%
9	Mango flat beak leafhopper	39.3%
10	beet fly	40.6%

V. DISCUSSION

A. CONTRIBUTIONS OF FINE-GRAINED ENHANCEMENT

The attention mechanism and resampling operations figure out where the discriminative features locate and then extract them to generate a new image contains the fine-grained objects as Fig. 4(b) shows. During the procedure of CNN, an input image with the dimension of $\mathbf{3} \times \boldsymbol{h} \times \boldsymbol{w}$ will be convolved into an output feature vector with the dimension of $\mathbf{1024} \times \boldsymbol{h} \times \boldsymbol{w}$ (taking Inception Net as an example). Therefore, the model can find the optimal solution easily with a larger occupancy of effective information in the image.

In previous experiments, we set a fixed ratio to combine the predictions of original images and resampled images to obtain the final prediction. To analyze the essence of fine-grained features advancing identification performance, we design an extra experiment. Requiring a minimal variant, we alternatively set a learnable ratio $\boldsymbol{\alpha}$ to sum up the predictions of raw features and fine-grained features:

$$Y = (1/2 + \alpha)Y_1 + (1/2 - \alpha)Y_2. \tag{11}$$

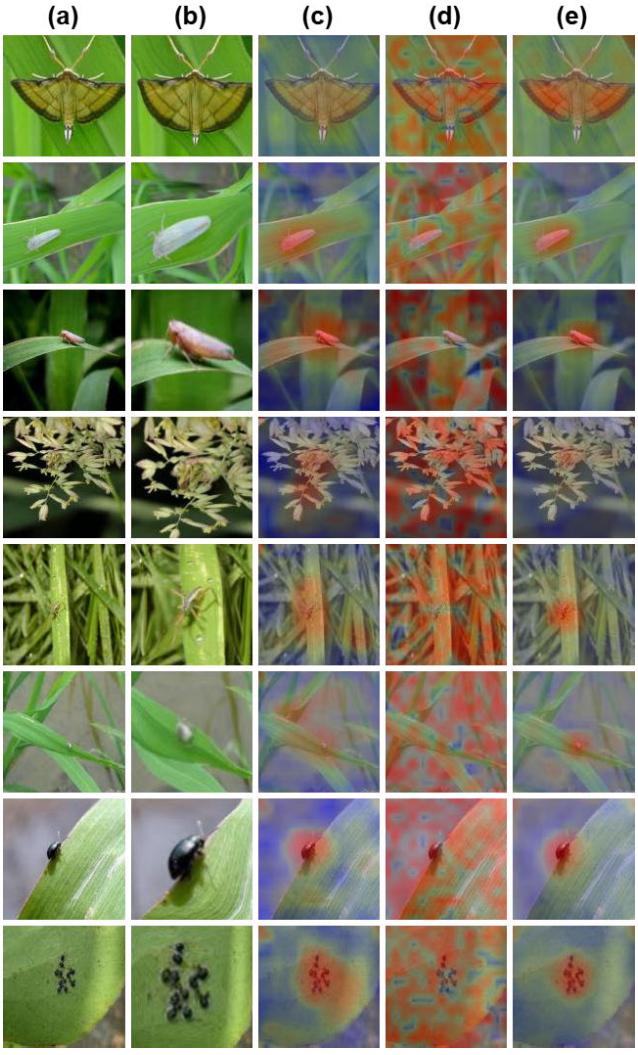

FIGURE 4. Visualization of attention learned by proposed approach. Illustration of fine-grained attention enhancement process and counterfactual attention learning process (image samples from dataset IP102): (a) original image, (b) attention based resample, (c) attention heatmap before counterfactual optimization, (d) fake attention (counterfactual attention) heatmap, (e) attention heatmap after counterfactual optimization.

Under this setting, we implement the same end-to-end training to analyze the change brought by the variable ratio. The final model finds that the optimal ratio is 0.5 indicating

that the original image plays a major role in prediction. Results above indicate that all the enhancing approaches of fine-grained features including the attention mechanism and attention-based resample enhance the capability of feature extraction rather than directly increase the quality of prediction decision. Testing the convergence speed in training phase, we find that the model with enhancing approaches tends to the optimal solution more quickly.

B. CONTRIBUTIONS OF COUNTERFACTUAL CAUSAL INFERENCE

We implemented causal inference through counterfactual attention optimization in this study aiming to eliminate the background bias. Fig. 4(c)(d)(e) shows the heatmaps of attention produced in different phases: (c) attention before optimization has large scope of interests and some mistakenly focuses on environment background, (d) fake attention generated from uniform distribution simulating counterfactual conditions, (e) attention after optimization obviously narrow the region of interest to focuses more on the targeting object. Comparing the heatmaps of columns (c) and (e), it is found that the counterfactual optimization process successfully eliminates the wrong attention focusing on the background and makes the area of interest more concentrated. The wrong attention roots in the preference that exists in dataset sources leading to an unexplained result. It is considered a kind of overfitting that an incorrect attention obtains a proper prediction defying the causal relationship. Therefore, we introduce the counterfactual inference to establish the relationship between the learned attention and the final prediction. Counterfactual inference [23], also known as causal inference, is to answer the counterfactual question. In the forward learning process, the model could learn either true attention or incorrect attention, which can be expressed as follows:

$$Y = \begin{cases} 1, U = 1(\textit{True attention}) \\ Z, U = 0(\textit{Fake attention}) \end{cases}, \tag{12}$$

$$P(Z = k) = p^k(1-p)^{1-k}, k = 0, 1, \tag{13}$$

where $\boldsymbol{Y} = \mathbf{1}$ represents a correct prediction while $\boldsymbol{Y} = \mathbf{0}$ represents an incorrect prediction, $\boldsymbol{U}$ represents the factor of concern, namely attention in this study. The above expressions mean that the true attention leads to a correct prediction and the fake attention leads to a correct prediction with a probability $\boldsymbol{p}$, respectively. There is no evidence to tell $\boldsymbol{U}$ when we observe that the model outputs a correct prediction. But we can get:

$$P(Y = 1 \,|U = 1)\,, \tag{14}$$

meaning we first obtain a fact that the learned attention leads to a correct prediction. The corresponding counterfactual question is what the fake attention would lead to. Applying counterfactual intervene to get $\boldsymbol{Y_x}$, the question can be expressed mathematically as:

$$P(Y_x = 0 \,|Y = 1, U = 0) \tag{15}$$

Factual output minus counterfactual output $\boldsymbol{Y(1) - Y(0)}$ represents individualized treatment effect, namely the quality of the learned attention in this study.

We consider the dataset bias as background bias in this study. In fact, bias is a very common problem also treated by many other different algorithms like re-weighting methods, stratification methods, matching methods, tree-based methods, representation-based methods, multi-task methods and meta-learning methods etc.

C. LIMITATIONS AND PROSPECTS

Classification accuracy performed on dataset IP102 is significantly lower than that on other popular fine-grained datasets owing to the characteristics of dataset itself. Dataset IP102 contains many limitations including noisy images, noisy labels and unbalanced data distribution etc. We listed the three challenges in Section IV leading to the unsatisfying performance. We need extra supervision to instruct multiple objects in one image. Multiple forms in one species and long-tail distribution are two challenges of dataset IP102 itself, the first of which requires a specific classification strategy. Although the counterfactual inference alleviates the problems caused by unbalanced distribution to a certain extent, more straightforward approaches should work to handle the long-tail bias. Moreover, a possible utilization of data cleaning and uncertain label deleting could greatly improve the model performance.

Our approach captures the most useful fine-grained features without significant decreasing performance across different field backgrounds and imaging devices. Therefore, this study provides the algorithmic basis for future pest control system including both a small-scale usage like farm pest control and a large-scale usage like wild insect detection. In practical application, our model can be integrated into the pest control and management systems as a computing module. The integration of the module requires an extra lightweight procedure like knowledge distillation. Another way is incorporating cloud computing to reduce the reliance on local computing resources. These practical procedures make our approach available in future agricultural engineering projects.

VI. CONCLUSION

In this paper, we have designed an end-to-end deep learning system of insect identification. We analyzed the challenges existed in current researches to propose approaches for fine-grained characteristics. First, for insect objects are small, we designed fine-grained enhancement approach to extract target foreground. Second, for image background is interference, we adopted counterfactual inference to eliminate the bias. Experiments carried on dataset IP102 and D0 showed identification accuracy of 74.5% and 99.8% respectively with the above two techniques. Our proposed deep learning approach can be trained end-to-end and requires limited computing resources which will contribute to the automatic insect detection and identification system in the future.

REFERENCES

[1] X. Wang, L. Ma, S. Yan, X. Chen, and A. Growe, "Trade for food security: The stability of global agricultural trade networks," *Foods*, vol. 12, no. 2, p. 271, Jan. 2023. [Online]. Available: https://www.mdpi.com/2304-8158/12/2/271

[2] T. D. Ramsfield, B. J. Bentz, M. Faccoli, H. Jactel, and E. G. Brockerhoff, "Forest health in a changing world: Effects of globalization and climate change on forest insect and pathogen impacts," *Forestry*, vol. 89, no. 3, pp. 245–252, Jul. 2016, doi: 10.1093/forestry/cpw018.

[3] W. Li, T. Zheng, Z. Yang, M. Li, C. Sun, and X. Yang, "Classification and detection of insects from field images using deep learning for smart pest management: A systematic review," *Ecol. Informat.*, vol. 66, Dec. 2021, Art. no. 101460. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S157495412100251X

[4] A. N. Alves, W. S. R. Souza, and D. L. Borges, "Cotton pests classification in field-based images using deep residual networks," *Comput. Electron. Agricult.*, vol. 174, Jul. 2020, Art. no. 105488. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169919311342

[5] Z. Yuan, H. Yuan, and Y. Yan, "Lightweight field insect recognition and classification model based on deep learning," *J. Jilin Univ.*, vol. 51, pp. 1131–1139, Jun. 2021.

[6] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The iNaturalist species classification and detection dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8769–8778.

[7] X. Wu, C. Zhan, Y.-K. Lai, M.-M. Cheng, and J. Yang, "IP102: A large-scale benchmark dataset for insect pest recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8779–8788.

[8] H.-W. Pang, P. Yang, X. Chen, Y. Wang, and C.-L. Liu, "Insect recognition under natural scenes using R-FCN with anchor boxes estimation," in *Proc. 10th Int. Conf. Image Graph.* Beijing, China: Springer-Verlag, Aug. 2019, pp. 689–701, doi: 10.1007/978-3-030-34120-6_56.

[9] S. Lim, S. Kim, S. Park, and D. Kim, "Development of application for forest insect classification using CNN," in *Proc. 15th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, Singapore, Nov. 2018, pp. 1128–1131.

[10] H. X. Huynh, D. B. Lam, T. V. Ho, D. T. Le, and L. M. Le, "CDNN model for insect classification based on deep neural network approach," in *Proc. ICCASA/ICTCC*, 2019, pp. 127–142. [Online]. Available: https://api.semanticscholar.org/CorpusID:209060600

[11] V. Sagar, R. Sachin, K. Chandrashekara, and K. N. Ganeshaiah, "Identification of Indian butterflies and moths with deep convolutional neural networks," *Current Sci.*, vol. 118, no. 9, pp. 1456–1462, May 2020.

[12] M. T. Reza, N. Mehedi, N. A. Tasneem, and M. A. Alam, "Identification of crop consuming insect pest from visual imagery using transfer learning and data augmentation on deep neural network," in *Proc. 22nd Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2019, pp. 1–6.

[13] W. Liu, G. Wu, and F. Ren, "Deep multibranch fusion residual network for insect pest recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 3, pp. 705–716, Sep. 2021.

[14] G. Yang, G. Chen, C. Li, J. Fu, Y. Guo, and H. Liang, "Convolutional rebalancing network for the classification of large imbalanced rice pest and disease datasets in the field," *Frontiers Plant Sci.*, vol. 12, 2021, Art. no. 671134. [Online]. Available: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2021.671134

[15] X. Yang, Y. Luo, M. Li, Z. Yang, C. Sun, and W. Li, "Recognizing pests in field-based images by combining spatial and channel attention mechanism," *IEEE Access*, vol. 9, pp. 162448–162458, 2021.

[16] Q. Luo, L. Wan, L. Tian, and Z. Li, "Saliency guided discriminative learning for insect pest recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.

[17] H. T. Ung, Q. H. Ung, T. T. Nguyen, and B. T. Nguyen, "An efficient insect pest classification using multiple convolutional neural network based models," in *Proc. 21st Int. Conf. New Trends Intell. Softw. Methodol., Tools Techn. (SoMeT)*, vol. 355, H. Fujita, Y. Watanobe, and T. Azumi, Eds. Kitakyushu, Japan: IOS Press, Sep. 2022, pp. 584–595, doi: 10.3233/FAIA220287.

[18] H. Guan, C. Fu, G. Zhang, K. Li, P. Wang, and Z. Zhu, "A lightweight model for efficient identification of plant diseases and pests based on deep learning," *Frontiers Plant Sci.*, vol. 14, 2023, Art. no. 1227011. [Online]. Available: https://api.semanticscholar.org/CorpusID:259927881

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2017, pp. 6000–6010.

[20] C. Xie, R. Wang, J. Zhang, P. Chen, W. Dong, R. Li, T. Chen, and H. Chen, "Multi-level learning features for automatic classification of field crop pests," *Comput. Electron. Agricult.*, vol. 152, pp. 233–241, Sep. 2018, doi: 10.1016/j.compag.2018.07.014.

[21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, vol. 11211, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Munich, Germany, Sep. 2018, pp. 3–19.

[22] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1005–1014.

[23] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. New York, NY, USA: Basic Books, 2018.

XINYUAN XU was born in 2000. She received the bachelor's degree from the School of Biomedical Engineering, Shanghai Jiao Tong University (SJTU), China, in 2022, where she is currently pursuing the master's degree in biomedical engineering. Her research interests include computer vision and machine learning.

HENG LI (Member, IEEE) received the Bachelor of Engineering degree from PLA Information Engineering University, China, in 2010, the Master of Engineering degree from Zhengzhou University, China, in 2013, and the Ph.D. degree from Shanghai Jiao Tong University (SJTU), China, in 2018. From 2018 to 2022, he was a Postdoctoral Researcher with the School of Biomedical Engineering, SJTU, where he is currently a Senior Assistant Researcher. His research interests include artificial intelligence in computer vision and natural language processing, pattern recognition and intelligent systems, artificial intelligence, and neural engineering.

QI GAO received the Bachelor's degree from the School of Biomedical Engineering, Beihang University, China, in 2021. She is currently pursuing the Ph.D. degree in biomedical engineering with Shanghai Jiao Tong University (SJTU), China. Her research interests include computer vision and multimodal machine learning.

MEIXUAN ZHOU received the bachelor's degree from the School of Life Science and Technology, Xi'an Jiaotong University, China, in 2018. He is currently pursuing the Ph.D. degree in biomedical engineering with Shanghai Jiao Tong University (SJTU), China. His research interests include neural engineering and electrical stimulation of neural modulation.

TIANYUE MENG received the bachelor's degree from the School of Biomedical Engineering, Chongqing University, China, in 2022. She is currently pursuing the Ph.D. degree in biomedical engineering with Shanghai Jiao Tong University (SJTU), China. Her research interests include medical image processing and computer vision.

LIPING YIN received the master's degree from Nanjing Agricultural University, China, in 1988. From 1989 to 1998, she was an Associate Researcher with Shanghai Animal Plant and Food Inspection Bureau, China. Since 1998, she has been a Researcher with the Technical Center for Animal Plant and Food Inspection and Quarantine of Shanghai Customs, China. Her research interests include intelligent detection systems and signal processing.

XINYU CHAI received the M.S. and Ph.D. degrees in biomedical engineering from Xi'an Jiaotong University, China, in 1991 and 1998, respectively. From 2001 to 2011, he was an Associate Professor and a Professor with the School of Life Science and Biotechnology, Shanghai Jiao Tong University (SJTU), China. Since 2011, he has been a Professor with the School of Biomedical Engineering, SJTU. His research interests include intelligent biomedical instrument, computer vision, and visual prosthesis.