IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Recognizing pests in field-based images by Combining spatial and channel attention mechanism

XINTING YANG[1,2,3], YONGCHEN LUO[1,2,3,4], MING LI[1,2,3], ZHANKUI YANG[1,2,3,5], CHUANHENG SUN[1,2,3], AND WENYONG LI[1,2,3]

[1]National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China
[2]Beijing Research Center for Information Technology in Agriculture, Beijing 100097, China
[3]National Engineering Laboratory for Agri-Product Quality Traceability, Beijing 100097, China
[4]School of Information Technology, Jinlin Agricultural University, Changchun 130118, China
[5]School of Computer Science and Technology, Beijing University of Technology, Beijing 100124, China

Corresponding author: Wenyong Li (e-mail: master0808@126.com).

This work was supported in part by the National Key Technology Research and Development Program of China under Grant 2017YFE0122503, in part by the Promotion and Innovation of Beijing Academy of Agriculture and Forestry Sciences, and in part by National Natural Science Foundation of China under Grant 31871525.

ABSTRACT Large scale pest recognition is one of crucial components in pest management in outdoor conditions, which is much more difficult than common object recognition because of the variational image acquisition direction, location, pest size and complex image background. To overcome the challenges, this study proposes a CNN model by combining spatial attention mechanism and channel attention mechanism to realize accurate pest location and recognition in field images. The proposed model consists of two major parts. Firstly, the module Spatial Transformer Networks (STN) is incorporated into a Convolutional Neural Network (CNN) architecture to provide image cropping out and scale-normalization of the appropriate region, which can simplify the subsequent classification task. The second one is called Improved Split-Attention Networks that is used to enable feature-map attention across feature-map groups. The proposed model is evaluated on three different datasets: Li's dataset (10 species), proposed dataset (58 species) and IP102 dataset (102 species), achieving the classification accuracies of 96.78%, 96.50% and 73.29%, respectively. Comparisons with five traditional CNN models and three attention-related state-of-the-art deep learning models show that the current method outperforms these previous models. Besides, to verify the robustness of this proposed model on different image resolutions, six datasets with different image resolutions are constructed and all accuracies exceed 92% with the image resolution of 400×267 pixels reaching the optimal performance. All results show that the proposed method provides a reliable solution to recognize insect pest in field and support precision plant protection in agriculture production.

INDEX TERMS Insect recognition, Attention mechanism, Deep learning, Image processing

I. INTRODUCTION

Agricultural insect pests are responsible for causing significant damage to crops and reducing their quantity and quality. Therefore, it is particularly important to strengthen the ability of pest monitoring and early warning, so as to carry out effective strategies for pest prevention and control [1]. However, one of the prerequisites for these tasks is to identify pests accurately and timely.

Traditionally, insect pest recognition mainly relies on few plant protection experts and technicians to complete according to the typical appearance characteristics of pests in the field, which is a time-consuming and labor-intensive task [2]. With the development of computer vision

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3132486, IEEE Access

techniques, they have been widely applied in object recognition in many fields, including insect pest recognition and detection. In general, the image-based insect pest recognition methods can be summarized into two categories: traditional machine learning methods and deep learning methods. The insect recognition methods based on traditional machine learning mainly include three sequential stages: image preprocessing, feature extraction and feature classification. Yao *et al*. [3] proposed a pest detection method by integrating Adaboost and SVM classifier and achieved a false detection rate of 9.6%. Inspired by the human cognitive neuroscience, Deng *et al*. [4] firstly used saliency model to detect region of interest (RoI) and then the invariant features representing the pest appearance were extracted and trained using SVM classifier, achieving a recognition rate of 85.5%. Unlike the aforementioned methods, Xie *et al*. [5] constructed a dictionary matrix and sparse decomposition to realize species classification, performing well on the classification of 24 common insect species. However, these traditional methods based on handcrafted features cannot adequately extract the characteristics of insect pest images from complex outdoor environment [6]. In addition, it is also difficult to determine the optimal solution for feature design and selection in these methods[7], which limit the improvement of pest recognition accuracy and their plications in field..

Compared with the traditional methods, the emerging deep learning-based models in recent years, such as convolutional neural networks (CNNs) [8], implement self-learning of features and their relations using data itself, which is considered as an end-to-end machine learning method [9]. To further extract the high-level image features and avoid complex modeling procedures from feature extraction to feature classification, some deep learning methods had been proposed to improve the accuracy and efficiency of pest recognition in the field images. Ding and Taylor [10] detected moths by applying a CNN into image patches at different locations, and they achieved a precision-recall rate of 93%. Paddy field pests were located and classified by computing a saliency map and applying a deep convolutional neural network (DCNN), achieving a mean accuracy precision (mAP) of 0.951 [11], which is a significant improvement on previous methods. It is well known that training a complex CNN from scratch to excellent performance level requires a huge set of labeled images and consumes a significant amount of computational resources, which means that it is not realistic to train a dedicated CNN for most image classification tasks [12].

In this study, motivated by the many successful applications of Spatial Transformer Networks (STN) in image classification, co-localization, spatial attention[13-15] and the simple and modular structure of ResNet variants in image classification, object detection and semantic segmentation[16-18]. A cascaded architecture based on STN and ResNest network is developed for large-scale pest recognition, in which region of interest are located and multi-channel features are learned from original images automatically without any preprocessing rather than hand-crafted.

Our purpose is to improve representation performance of insect pest images by using attention mechanism: focusing on important features and suppressing unnecessary ones. Furthermore, it was found that the insect targets in images have many poses and even occupy small area in the whole image, which makes it difficult to focus on important insect features during model learning. To achieve this, we sequentially apply STN network to locate the region of pest target and a novel Split-Attention block to improve the learned feature representations to boost performance across image classification.

The main contributions of this study include:

- An insect pest dataset containing 58 pest species from garden and forest was constructed and could be access by the public.
- We proposed a cascaded yet effective attention architecture that can be applied to improve image representation power.
- The effectiveness of our attention architecture was validated through extensive comparisons on different scale datasets.

II. RELATED WORK

An important phenomenon of the human visual system is that one does not attempt to process a whole scene at once. Instead, human selectively focus on the salient parts in order to capture visual information better, which is the attention mechanism in the human visual system. Recently, the mechanism is also incorporated into CNNs in large-scale classification tasks. From the perspective of attention domain, the implementation of attention mechanism can be divided into three types: spatial domain, channel domain and mixed domain.

A. ATTENTION MECHANISM IN SPATIAL DOMAIN

In object classification using digital images, especially, for the similar species, discriminative information is always reflected in certain regions while the other regions contain much redundancy, which makes object recognition an extremely difficult computer vision task. For solving this problem, many recent studies develop models on the attentional regions, rather than the whole scenes [14, 19].

The attention mechanism in spatial domain is based on the spatial position of feature map without distinguishing the influence brought by channels. Although attentional regions can be learned using deep neural networks, it is hard to train with only class information because they have to simultaneously complete two difficult tasks (i.e., region localization and recognition). To overcome this difficulty,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3132486, IEEE Access

Jaderberg et al. [20] proposed a Spatial Transformer Network (STN) which can be included into a standard neural network architecture to provide spatial transformation capabilities. This model allows networks to not only select regions of an image that are most relevant (attention), but also to transform those regions to a canonical, expected pose to simplify recognition in the following layers. Inspired by the classical non-local means method, Wang et al. [21] presented non-local operations which compute the response at a position as a weighted sum of the features at all positions. Besides categorical labels, the study proposed by Chen et al. [22] requires another ground truth, the facial landmarks, which is quite unique for face detection. In this study, we only use class labels as ground truth.

B. ATTENTION MECHANISM IN CHANNEL DOMAIN

The input image will become a tensor after the convolution transformation with the number of output channels, which is equivalent to the decomposition of the original image, and each channel is the component of the original image on different convolution kernels. In contrast to the spatial domain, the channel domain focuses on the weighting of different channels, regardless of the location difference of each pixel in the channel.

Instead of seeking to strengthen the representational power of a CNN by enhancing the quality of spatial encodings throughout its feature hierarchy, Hu et al. [17] proposed a compact module, SE block, to exploit the inter-channel relationship, which introduces attention mechanism from channel dimension. The SE block obtains the weight of importance of each feature channel and assigns the weight to each feature channel respectively. This design makes the neural network focus on some feature channels, that is, promoting the feature channels that are important to the current task and suppressing others that are of little use to the current task. Xie et al. [23] adopted group convolution in the ResNet bottle block, which results in a homogeneous, multi-branch architecture. Li et al. [24] proposed a dynamic selection mechanism in CNNs that allows each neuron to adaptively adjust its receptive field size based on multiple scales of input information. More close to our work, Zhang et al. [25] generalized the channel-wise attention into feature-map group representation, which can be modularized and accelerated using unified CNN operators.

C. ATTENTION MECHANISM IN MIXED DOMAINS

The works based on channel domain are short of the mechanism of spatial attention which plays an important role in deciding ‘where’ to focus in an image recognition task. Therefore, some researchers provide interesting studies about the combined use of spatial and channel attention. Woo et al. [26] exploited both spatial and channel-wise attention based on an efficient architecture and empirically verify that exploiting both is superior to using only the channel-wise attention. Building long-range dependencies is helpful in most classification tasks using computer vision techniques. Like CBAM [26], NLNet [21], SENet [17] build interdependencies among the channel dimensions introducing spatial attention mechanisms or designing advanced attention blocks. Another way to model long-range dependency is to exploit convolutional operators with large kernel windows. Liu et al. [27] presented SCNet, which is able to heterogeneously exploit the convolutional filters nested in a convolutional layer and adaptively builds long-range spatial and inter-channel dependencies around each spatial location. Inspired by the previous methods, in this study, we exploit a new attention approach by combining spatial transformer module and improved ResNest block to construct a create a simplified network for recognition of insect pest.

III. PRINCIPLE OF THE PROPOSED METHOD

A. INSECT IMAGE DATASET

In the current study, it involved three insect image datasets. In the first dataset, images were captured using NIKON D200 cameras at different locations in suburb of Beijing, China. A total of 58 types of insects, including 7344 images, were collected and this dataset is named as the proposed dataset. Figure 1 shows 10 geographical locations and their insect sample images of the proposed dataset. All images in the proposed dataset and code are available from the corresponding author on reasonable request. The other two datasets are public image sets. One was reported by Li et al. [28], which includes 10 categories and a total of 5629 images. Some samples in the Li’s dataset are shown in Fig. 2. The other one (IP102) was from the literature [29], including 102 types of field crops pests and a total of 75222 images. Some insect images in the IP102 dataset are shown in Fig.3. As a result, three datasets representing different sizes were constructed to evaluate the performance of the proposed model. The detail of the three datasets is summarized in Table 1.

TABLE 1. Composition and comparison of the three datasets.

Dataset names	Number of species	Average image numbers per species	Total images
Li’s dataset	10	563	5629
Proposed dataset	58	126	7344
IP102 dataset	102	737	75222

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3132486, IEEE Access

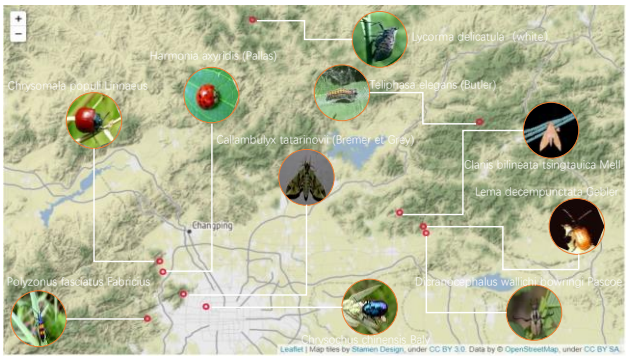


FIGURE 1. **Diagram of image collection locations and samples of pest in theproposed dataset.**

FIGURE 2. **Ten image samples in the Li's dataset [30].**

FIGURE 3. **Ten image samples in the IP102 insect dataset [29].**

B. DATA PREPROCESSING

To expand the image quantities to adequately train the models, a set of online transformations was used to produce extra images from the original datasets in this study. Unlike offline augmentation methods which were implemented by processing the whole dataset directly before training a model, online data augmentation divides the training data into multiple batches and input to the model batch by batch, co-trained with the target learning task This online method is both more efficient, in the sense that it does not require expensive offline training when entering a new domain, and more adaptive as it adapts to the learner state [31]. Generally, it is often applied to augmentation of large-scale datasets, which has been supported in many deep learning frameworks and can be optimized by using GPU calculations.

Furthermore, to fairly compare the results between different methods, the strategies of data preprocessing in this study were kept consistent with the previous methods. For Li's and the proposed datasets, they were preprocessed using the online augmentation to improve the generalization ability of the model, And the large-scale dataset was not processed for the data expansion [32].

C. THE PROPOSED MODEL

The residual network, ResNet, is widely used since its skip connections between different layers. This superior design can transmit the input signal to the higher layer from any lower layer, which solves model degradation problem caused by the increase in the number of convolutional layers. However, the ResNet network lacks cross-channel interaction, so that there are many improvements on it. SE-Net [33] uses a cross-channel attention mechanism on the residual block, which makes the focus of traditional CNN from the global information to the local feature. ResNext [34] uses the idea of grouping convolution to put different channels into different groups, so that each group focuses on different features, and then the results from the groups were merged. Compared with ResNet, the ResNext achieves a trade-off between global and local features. The latest ResNest [35] network is also one of ResNet's variants, which combines the characteristics of SENet and ResNext, and achieves state-of-the-art performance in classification, detection, and segmentation tasks simultaneously.

The channel attention mechanism in ResNest network helps it to assign corresponding weights to the feature maps obtained by different convolution kernels, so as to it can focus on the features of interest. However, the complex background of field pests may still mislead the model to pay attention to features that are not related to pest itself. Therefore, the spatial transformer network(STN)[36] which allows the model to learn the importance of different spaces was introduced into the proposed method to locate the object in the image. As shown in Fig. 4, a pest recognition model was proposed in this study by combining channel and spatial attention mechanism.

In the proposed model, the input image firstly was processed through an STN network based on affine transformation described in Equation (1) to locate the region of interest (RoI), which realized a spatial transformation capabilities and attention mechanism. As shown in Fig. 4, three convolutional layers and two fully connected layers are used as the localization network in STN to obtain the affine transformation matrix A_θ. The generator calculates the coordinate value of each position in the output map,$T_\theta(G)$, and the sampler perform sampling in the original image according to the coordinate information in $T_\theta(G)$. A RoI image is obtained after copying the pixels of the original image to the output image. The number of channels between input and output image is the same in the spatial attention network, but the RoI image will focus on key areas. The image transformed by the spatial attention mechanism is then input the improved ResNest network which integrates the cross-channel attention mechanism through multiple Split-Attention Block modules. Finally,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3132486, IEEE Access

the fully connected layer achieves accurate recognition of multiple types of pests.

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = A_\theta \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & t_x \\ a_3 & a_4 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (1)$$

Here, A_θ is an affine transformation matrix, t_x, t_y represent the amount of image translation, and the parameter a_i (i=1,2,3,4) reflects the changes in image rotation, scaling, etc.

Then, the output of spatial attention operation is processed by the improved ResNest50 network. In this study, this channel attention network is composed of 16 split-attention modules, as shown in the orange box in Fig. 4. The insect pest image firstly is processed through a 7×7 convolutional layer, and then the channel attention module performs feature extraction on insect images. Unlike ResNest, it does not need to pool the information of all channels in the feature map at once. Instead, the feature map is divided into multiple Cardinals by channel, then concatenating multiple groups in Cardinal at the channel level, and performing global average pooling on the feature map after concatenation. In fact, through grouping convolution, the association between different feature maps is reduced, the differences between feature maps are more apparent, and ultimately the complementary feature maps are obtained.

As shown in Fig.5, each Split-Attention Block module consists of a set of group convolution. The input feature maps are divided into K cardinals, and each cardinal is divided into R groups, so there is a total of $G = k \cdot R$ feature map groups. In the improved ResNest network, the self-calibrated convolution [27] was introduced to replace the second convolution layer in each group. Unlike the common convolutions that extract spatial and channel-wise information using small kernels (e.g., 3 × 3), the self-calibrated convolution adaptively builds long-range spatial and inter-channel dependencies around each spatial location through a novel self-calibration operation, which can help CNNs generate more discriminative representations by explicitly incorporating richer information.

After applying the corresponding transformation into each group $\{f_1, f_2, \dots, f_G\}$, The intermediate result is: $U_i = f_i(x), i \in \{1,2 \dots, G\}$. After fusing the intermediate result U_i, global information is obtained through global average pooling, and different groups are given with different weights through the Dense layer (Fig.6). The importance of each channel is automatically obtained by model learning. According to this importance, the useful features are enhanced and the other ones are suppressed.

By combining the spatial attention mechanism network STN with an improved ResNest50 backbone network, the impact of complex background on classification is reduced, and the ability of feature representation for large-scale insect dataset images is strengthened, thereby improving the recognition performance on the insect datasets.

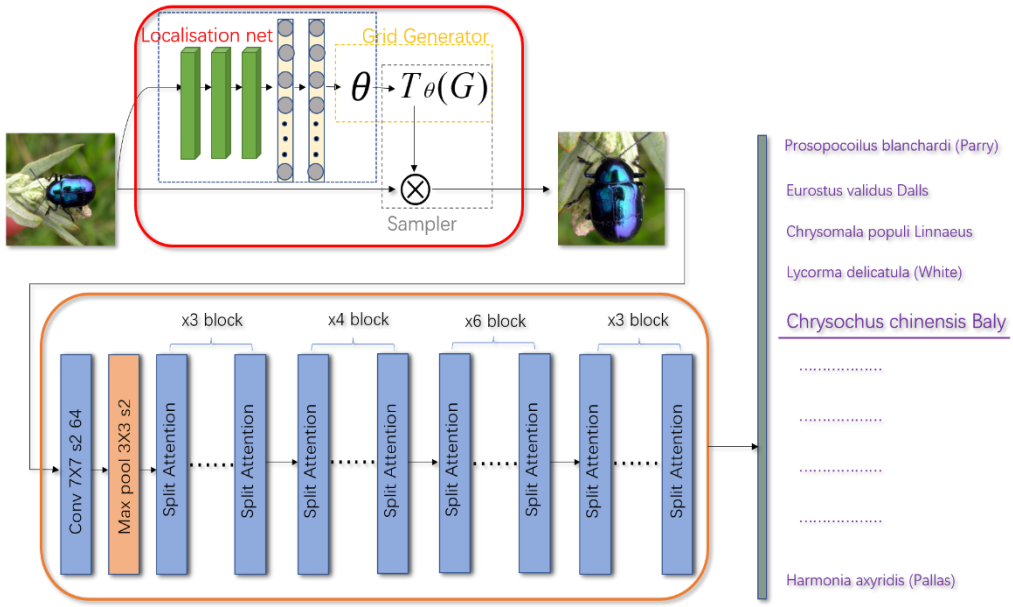


FIGURE 4. **The framework structure of the proposed model.**

IV. RESULTS

In this section, the performance of different models on the three datasets was evaluated and compared. The experiment was implemented on the PyTorch deep learning framework (https://pytorch.org/) and windows10 operating system with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3132486, IEEE Access

RTX2080Ti 11GB GPU hardware platform (https://www.nvidia.com/en-us/geforce/graphics-cards/rtx-2080-ti/). The cross-entropy loss function (*loss*) and the average accuracy (*acc*) was used to train and evaluate the models, respectively. They are calculated as followed:

$$loss = -\sum_i y_i \, log(P_i) \tag{2}$$

$$acc = \frac{Number\ of\ insects\ predicted\ correctly}{Total\ number\ of\ insect\ samples} \tag{3}$$

where y_i is the category label, P_i is the probability that the predicted category of the network output is i.

A. MODEL FINE-TUNING

The model hyperparameters are closely related to the model performance. In current study, the model hyperparameters were tuned by setting different gradient in multiple experiments. In terms of model learning rate, three gradients of 0.01, 0.001, and 0.0001 were constructed, and the training optimizer was chosen between Stochastic Gradient Descent(SGD) plus momentum and Adam [37]. In the recognition experiment of the proposed dataset, the feature extraction layer of the pre-trained model was fine-tuned, and the fully connected layer of the original model was replaced with 58 neurons. To avoid overfitting, dropout [38] was employed into models and set to 0.3. The input image size was fixed to 224×224 pixels, and the data set was randomly divided into training dataset and test dataset at a ratio of 7:3. The fine-tuned results of the five models: AlexNet, VGG19, GoogleNet, ResNet50, and ResNest50 are shown in Fig. 7.

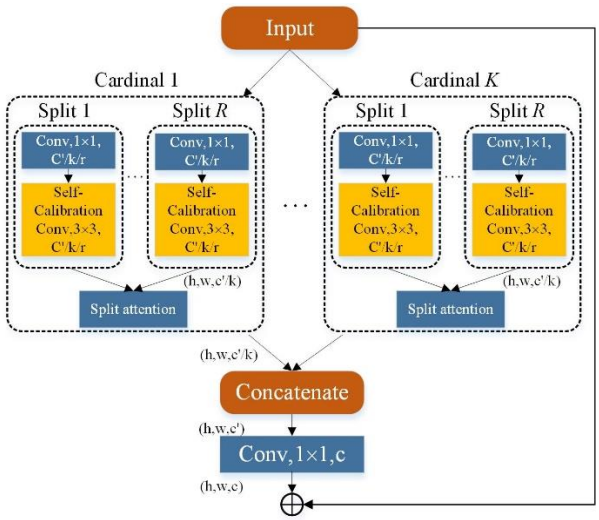


FIGURE 5. The composition of the Split-Attention Block module.

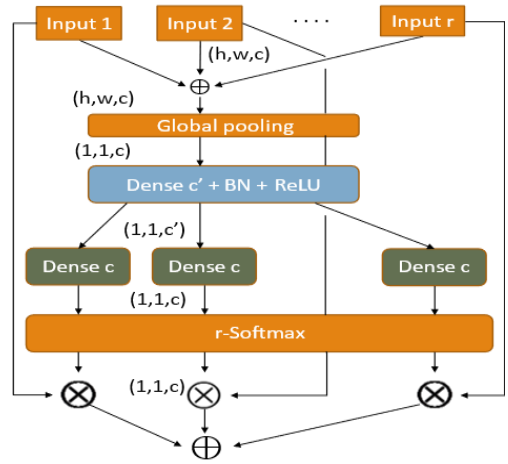


FIGURE 6. The detailed structure of each Cardinal group [25].

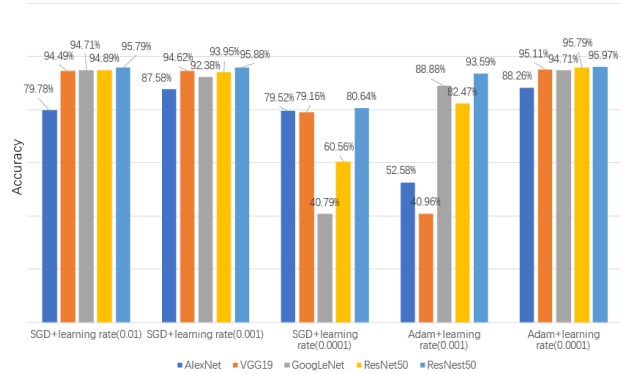


FIGURE 7. Performance comparison of models on different parameter combinations.

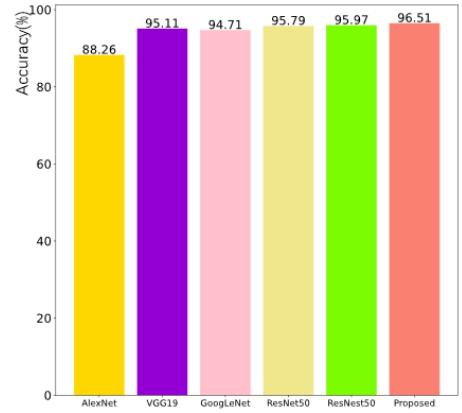


FIGURE 8. Comparison of the recognition performance of multiple models on self-built data sets.

By comparing the fine-tuned results with different optimizers and learning rate, it was found that all models achieved the best recognition accuracy when the Adam was used and initial learning rate was set to 0.0001. Among them, ResNest50 reached the highest accuracy rate of 96.86%.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3132486, IEEE Access

Therefore, we applied these optimal parameters into the proposed model, and then it was trained on the three datasets to obtain the best recognition model, respectively.

B. MODEL PERFORMANCE

Firstly, the recognition results of the proposed model are compared with five CNN models on the proposed dataset. As shown in Fig.8, the proposed method achieves the highest accuracy rate of 96.51%, which is an improvement of 0.64% compared to the original ResNest50 model. The lowest recognition accuracy rate is obtained by AlexNet, which only reached 88.26%. The accuracies of other three models, VGG19,

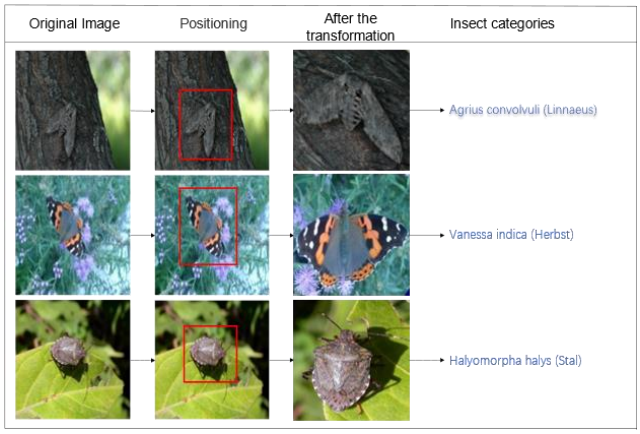


FIGURE 9. **The insect target is focused and corrected after STN.**

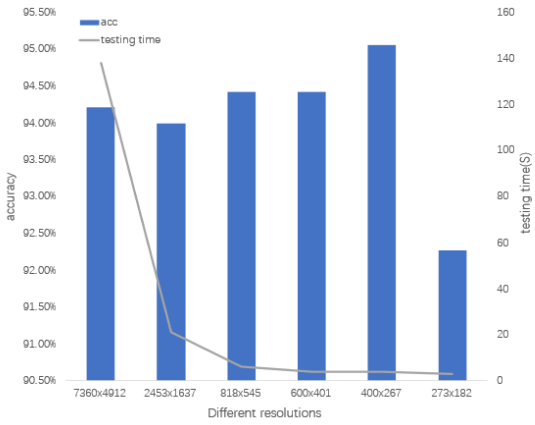


FIGURE 10. **The results of the proposed model for different resolution image data recognition experiments.**

GoogleNet and ResNet50 are similar around 95%.

Insects in field images are often accompanied with complex background, so the recognition network is easy to be misled by the background when it extracted image features. In addition, the insect postures and sizes in the images are different because of the various distances and angles when they are photographed in the field, which increases the recognition difficulty. In this study, a STN structure was trained to locate and correct the insect target. Fig.9 shows the results of the STN procedure. As illustrated in this figure, there are complex background in the original input images, and the insects have different posture and size (the first column in Fig.9). After processed by the STN module, the insect targets are highlighted and adjusted while the image size keeping unchangeable (the third column in Fig.9). In this way, a spatial localization and attention mechanism is integrated into the proposed model, which reduces the probability of being misled by the background, thus improving the recognition accuracy.

C. MODEL ROBUSTNESS

1) THE INFLUENCE OF IMAGE RESOLUTION ON MODEL PERFORMANCE

Field images from different sources often have different image resolutions, so analyzing the model performance on images of different resolutions helps to select the best image resolution in practical tasks. In the current study, 466 images from 10 insect categories in the proposed dataset were randomly selected, and the highest resolution 7360×4912 pixels in this dataset is used as a benchmark to design six different gradients: 7360×4912, 2453×1637, 818×545, 600×401, 400×267, 273×182 pixels. Subsequently, six image datasets with different image resolution are constructed to evaluate the model performance.

The results of recognition accuracy and time are illustrated in Fig.10. It is found that that the recognition accuracies of all datasets with different image resolutions were over 92%, which indicated the proposed method had a good robustness on image resolution. In addition, the recognition time witness a downward trend as the image resolution decreased. However, the model recognition accuracy has not the similar results with the recognition time. In particular, the recognition accuracy reached the highest value of 95.06% when the resolution was 400×267 pixels, and the lowest accuracy of 92.27% is achieved on the image dataset of 273×182 pixels in resolution.

2) MODEL PERFORMANCE ON DIFFERENT DATASETS

To evaluate the model performance on different-scale pest datasets, the IP102 and Li's datasets described in section 2 were also tested by the proposed method.

In the test of the two datasets, all preprocessing methods are consistent with the ones described in original publications. For instance, the parameters of all pre-trained models were fine-tuned without data augmentation for IP102 dataset. However, the parameters of all pre-trained models were frozen on Li's dataset, and online data enhancement including randomly flipping and cropping the image was implemented to train the model. The recognition results of six models on the IP102 and Li's datasets are shown in Fig. 11 and Fig.12, respectively. It can be found that the proposed model achieved the highest recognition

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3132486, IEEE Access

accuracy of 73.29% on the IP102 dataset, which is 1.32% higher than the second highest model, ResNest50. However, the results of VGG16, GoogleNet and ResNet50 model are pretty close and approximately 69%, and the lowest recognition accuracy of 58.81% is reached by the AlexNet model. Likewise, as shown in Fig.12, the highest recognition accuracy of 96.78% is obtained by the proposed model in the test of Li's dataset, while the recognition results of VGG16, ResNet50 and ResNest50 are similar (about 95%). The lowest accuracy of 91.02%, however, is obtained by the GoogleNet model, which is approximate to the result of AlexNet model.

The comparison between Fig.11 and Fig.12 shows

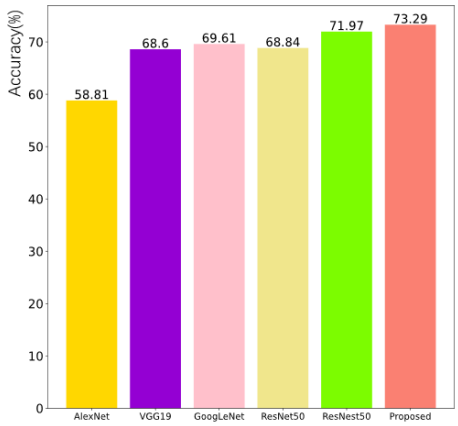


FIGURE 11. Classification performance of different models on the IP102 dataset.

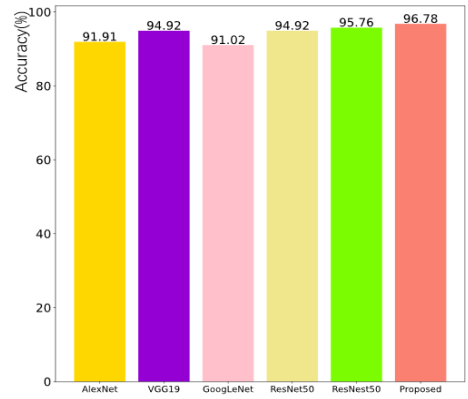
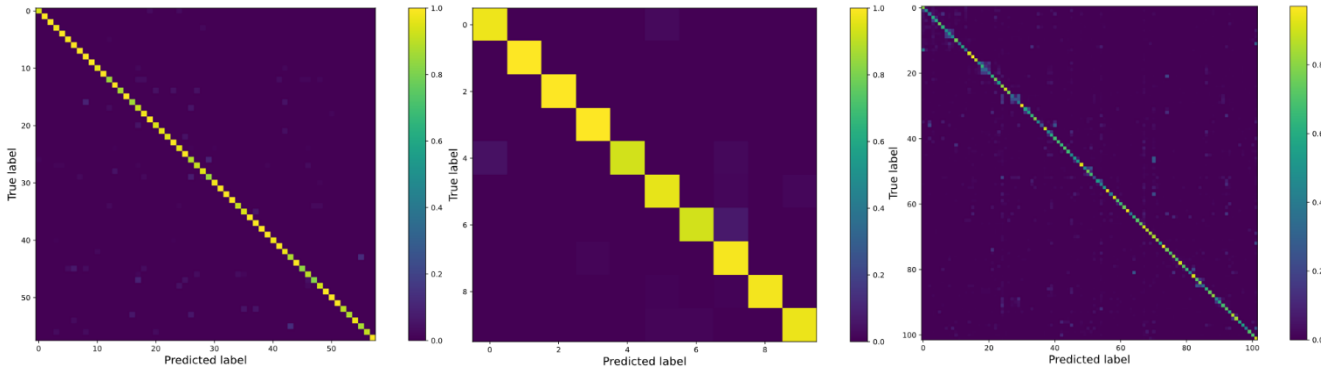


FIGURE 12. Classification performance of different models on the Li's dataset.

that the recognition accuracy of all models on the Li's dataset exceeded 90%, however, for the IP102 dataset, the recognition accuracy of these models is less than 75%. To further illustrated the recognition performance of the proposed method visually, the classification confusion matrixes on the three datasets are plotted in Fig.13. The diagonal elements of the confusion matrix represent the true positives, and the rest of elements in rows mean the false positives of classification. Overall diagonal elements in this figure showed the maximum values as expected. However, compared with the proposed and Li's dataset, more small numbers were found at non-diagonal elements on the IP102 dataset, which also showed that the recognition performances of the proposed method on Li's and proposed datasets are better than that on IP102 dataset. Figure 14 shows some similar insect images of different species from the IP102 dataset. The two images in the same column represents two different insect species, but there are similar appearance features, which is one of main reasons to cause the relative low recognition accuracy on IP102 dataset.



(a) Li's dataset **(b) Proposed dataset** **(c) IP102 dataset**

FIGURE 13. The normalized classification confusion matrix of classification on the three datasets: (a) Li's dataset, (b) Proposed dataset, (c) IP102 dataset.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3132486, IEEE Access

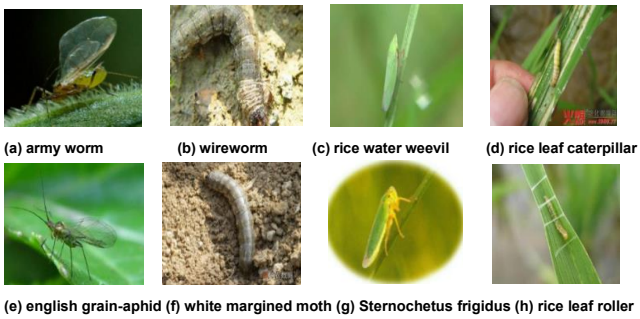


FIGURE 14. Samples of different insect species with similar features.

V. DISCUSSION

A. MODEL ARCHITECTURE

By comparing the results with other models on IP102, proposed and Li's datasets, the excellent performance of the proposed method was proved. Spatial attention mechanism helps the model more accurately locate the insect target in the image with complex backgrounds, and the channel attention mechanism contributes the extraction of discriminant features in the proposed model. Fig.15 shows the visualization results of the key areas that the model focuses on during species recognition and prediction. it is found that most insect areas in most images can be accurately located, which helps to improve the recognition performance of the proposed model. However, some images in the IP102 dataset are not well positioned, such as the image in row 2, column 4. The target in the image is a piece of insect egg, which was not located accurately by the model. The image quality of most images in the IP102 dataset is relatively worse, including advertising words, blurry target and so on. All these factors bring challenges to the recognition model on the IP102 dataset. However, the proposed method still achieved a new benchmark of 73.29% on this dataset.

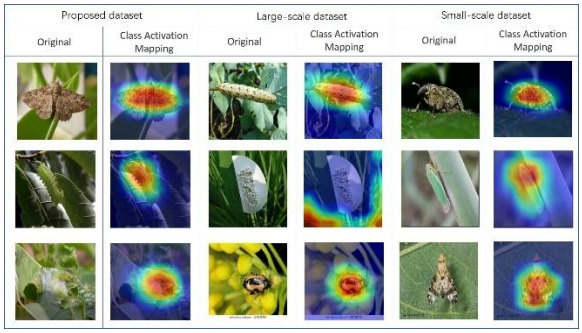


FIGURE 15. The key areas that the model focuses on during image processing.

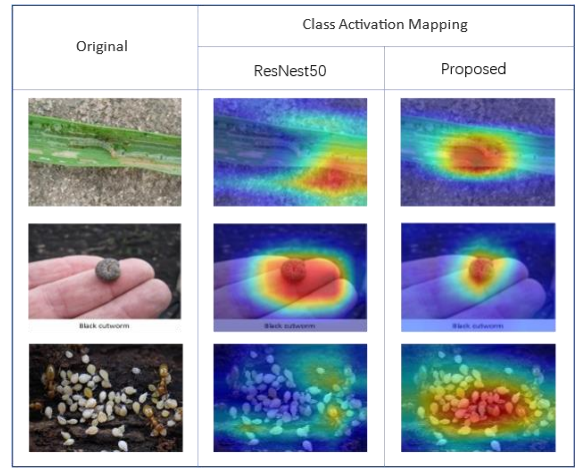


FIGURE 16. Location results of the models with and without STN network on the IP102 dataset.

The proposed method improves the recognition performance by incorporating a spatial attention mechanism into the improved ResNest50 network. The visualization results of the attention regions (Fig.16) in the recognition procedure on the IP102 dataset shows that the model can focus on the target more effectively under the spatial attention mechanism, and the classification feature is more dependent on the insect itself rather than the surrounding background.

B. THE EFFECT OF IMAGE RESOLUTION ON MODEL PERFORMANCE

The experimental results in Fig.10 shows that the classification performance of the proposed model reaches the highest value when the image resolution is 400×267 pixels. This result is not consistent with our intuition—the higher the image resolution is, the easier the insect in the field image is recognized. Actually, the result in current study shows the higher image resolution does not mean the better classification performance. Instead, it spends more time in the image processing. The possible reason is that the image needs to be scaled to 224×224 pixels before being input to the model. The higher the image resolution, more serious the image is distorted and more information will be lost when it is compressed to a fixed resolution, which causes inaccurate features extraction. On the other hand, the insect target in an image with pretty low resolution is blurry, which causes fine-grained features are hard to be extracted by the model, so the classification performance on the low-resolution (273 × 182) image dataset is also not good.

C. COMPARISON WITH PREVIOUS STUDIES

Fig.11 and Fig.12 show the performance of different models on two public datasets (the IP102 and Li's datasets). For the IP102 dataset, Wu *et al.* [32] extracted manual features and deep features and fine-tuned ResNet50 to reach the highest

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3132486, IEEE Access

accuracy rate of 49.7%, while the proposed method achieved a 4.35% improvement compared to the previous result. Li *et al.* [30] fine-tuned GoogleNet model to classify 10 types of pests, achieved 93% accuracy in the Li's dataset, which is lower than the result (96.78%) of the proposed method. Three pre-trained CNN models were used to integrate into a model, and achieved 67.13% classification accuracy on the IP102 dataset [39], while the proposed model achieved 73.29%. Nanni et al. [40] proposed a classifier by the fusion between saliency methods and convolutional neural networks and the classification accuracy on the IP102 dataset is 61.93%, which is 11.36% lower than that of the proposed model. Therefore, all comparisons show the proposed method reaches the state-of-the-art accuracy on the three different-scale datasets and provides a novel approach for the recognition of insect images under complex background in field.

On the other hand, although the proposed model achieves a new benchmark of 73.29% on the IP102 dataset, it is much lower than that on Li's dataset. The reasons may include two aspects: one is the fine-grained differences between many similar insect species and low image quality in the IP102 dataset, as shown in Fig.14, and the other one is the long-tail effect of the dataset, which means insect categories with huge sample sizes have received more attention from the model in the IP102 dataset, while rare categories are often under-focused. So, the two aspects are the difficulties needed to be conquered to further improve the model performance in the future.

D. COMPARISON WITH ATTENTION-RELATED MODELS AND ABLATION EXPERIMENTS

In order to further demonstrate the performance of the proposed model in this study, it was also compared with the other attention mechanism model on the IP102 dataset. Furthermore, the ablation experiment was conducted to prove the effectiveness of the spatial attention mechanism. Table 2 lists the recognition performance of different models on IP102 insect data sets, as well as the calculation consumption and the inference time on the test dataset. It is found that the proposed method has better performance than SE-Net with channel attention mechanism and SA-Net and CBAM with both spatial and channel attention mechanisms. However, the proposed method needs the highest floating-point operations (FLOPs) in image recognition, which means it has the highest model complexity. The testing time of the four attention-related models on the test dataset is approximate.

TABLE 2. Comparison different advanced methods' performance on the IP102 dataset, including some models with attention mechanism and ablation experiments on spatial attention mechanism.

MODELS	ACCURACY	FLOPS	TEST TIME
ALEXNET [41]	58.81%	657.85M	2M 6S
GOOGLENET [42]	69.59%	2.85G	2M 33S
RESNET50 [43]	68.84%	4.12G	2M 31S
VGG19 [44]	68.80%	19.58G	2M 56S
SE-NET [17]	69.72%	4.12G	2M 17S
DENSENET121[45]	67.39%	2.88G	3M 31S
RESNEXT [46]	71.59%	4.26G	2M 39S
RESNEST50 [25]	71.97%	5.41G	2M 42S
SA-NET [47]	70.78%	4.12G	2M 46S
CBAM [26]	69.75%	4.12G	2M 48S
STN-ALEXNET	59.07%	697.04M	2M 6S
STN-GOOGLENET	69.61%	2.89G	2M 27S
STN-RESNET50	69.23%	4.16G	2M 33S
STN-VGG19	68.62%	19.62G	2M 56S
STN-SE-RESNET50	69.84%	4.16G	2M 17S
STN-RESNEXT	72.12%	4.3G	2M 11S
STN-DENSENET121	70.25%	2.92G	3M 48S
THE PROPOSED MODEL	**73.29%**	5.45G	2M 45S

In the ablation experiments, it is found that the STN module can improve the performance of these models listed in table 2 by different extents. For example, the best result is STN-DenseNet121, which has an improvement of 2.86% compared to the original model. The experiments show the STN module can provide a spatial attention mechanism to reduce the impact of image complex backgrounds. However, the metric, FLOPS, indicates that the models with STN will increase the computational load.

V. Conclusion

This study proposed a pest recognition framework by integrating spatial and channel attention mechanism based on STN and ResNest50 networks. A medium-scale dataset collected manually and other two public datasets (Li's and IP102) were constructed to evaluate the proposed model. The experimental results showed the proposed method outperformed other five classic models and three attention-related state-of-the-art methods, and reached a new benchmark of 73.29% on the IP102 dataset. Moreover, an optimal image resolution of 400×267 pixels was determined after multiple experiments using six datasets of different image resolutions. The results show the proposed model has great potential in pest recognition in agricultural field. In the future, more attention should be paid to solve the fine-grained insect identification and the effects from long tail distribution of insect datasets.

REFERENCES

[1] J. G. A. Barbedo, "Detecting and Classifying Pests in Crops Using Proximal Images and Machine Learning: A Review," *AI,* vol. 1, no. 2, pp. 312-328, 2020.

[2] J. Liu and X. Wang, "Plant diseases and pests detection based on deep learning: a review," *Plant Methods,* pp. 17-22, 2021.

[3] Q. Yao, D. Xian, Q. Liu, B. Yang, G. Diao, and J. Tang, "Automated Counting of Rice Planthoppers in Paddy Fields Based on Image

Processing," *Journal of Integrative Agriculture,* vol. 13, no. 8, pp. 1736-1745, 2014.

[4] L. Deng, Y. Wang, Z. Han, and R. Yu, "Research on insect pest image detection and recognition based on bio-inspired methods," *Biosystems Engineering,* vol. 169, pp. 139-148, May 2018.

[5] C. Xie *et al.*, "Multi-level learning features for automatiic classification of field crop pests," *Computers and Electronics in Agriculture,* vol. 152, pp. 233-241, 2018.

[6] V. Partel, L. Nunes, P. Stansly, and Y. Arnpatzidis, "Automated vision-based system for monitoring Asian citrus psyllid in orchards utilizing artificial intelligence," *Computers and Electronics in Agriculture,* vol. 162, pp. 328-336, Jul 2019.

[7] D. I. Patrício and R. Rieder, "Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review," *Computers and Electronics in Agriculture,* vol. 153, pp. 69-81, 2018/10/01/ 2018.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, vol. 86, no. 11, pp. 2278-2324.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, pp. 436-444, 2015.

[10] W. Ding and G. Taylor, "Automatic moth detection from trap images for pest management," *Computers and Electronics in Agriculture,* vol. 123, pp. 17-28, Apr 2016.

[11] Z. Liu, J. Gao, G. Yang, H. Zhang, and Y. He, "Localization and Classification of Paddy Field Pests using a Saliency Map and Deep Convolutional Neural Network," *Scientific Reports,* vol. 6, no. 1, 2016.

[12] M. Valan, K. Makonyi, A. Maki, D. Vondracek, and F. Ronquist, "Automated Taxonomic Identification of Insects with Expert-Level Accuracy Using Effective Feature Transfer from Convolutional Networks," *Syst Biol,* vol. 68, no. 6, pp. 876-895, Nov 1 2019.

[13] C. Luna-Jimenez, J. Cristobal-Martin, R. Kleinlein, M. Gil-Martin, J. M. Moya, and F. Fernandez-Martinez, "Guided Spatial Transformers for Facial Expression Recognition," *APPLIED SCIENCES-BASEL,* vol. 11, no. 16, 2021.

[14] D. Liu, Y. Wang, and J. Kato, "Attention-Guided Spatial Transformer Networks for Fine-Grained Visual Recognition," *IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS,* vol. E102D, no. 12, pp. 2577-2586, 2019.

[15] X. Zhang, T. Gao, and D. Gao, "A new deep spatial transformer convolutional neural network for image saliency detection," *DESIGN AUTOMATION FOR EMBEDDED SYSTEMS,* vol. 22, no. 3, pp. 243-256, 2018.

[16] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classication with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 558-567.

[17] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 42, no. 8, pp. 2011-2023, 2019.

[18] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," in *30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5987-5995: IEEE.

[19] Y. Zhu, C. Zhao, H. Guo, and J. Wang, "Attention CoupleNet: Fully Convolutional Attention Coupling Network for Object Detection," *IEEE Transactions on Image Processing,* vol. 28, no. 1, pp. 113-126, 2019.

[20] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial Transformer Networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems* Montreal Canada, 2015, vol. 2, pp. 2017-2025: MIT Press.

[21] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-

local Neural Networks," in *Computer Vision and Pattern Recognition*, 2018: IEEE.

[22] D. Chen, G. Hua, F. Wen, and J. Sun, "Supervised Transformer Network for Efficient Face Detection," in *European Conference on Computer Vision*, 2016, vol. 9909, pp. 122-138: Springer.

[23] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, vol. 1, pp. 5987-5995.

[24] X. Li, W. Wang, X. Hu, and J. Yang, "Selective Kernel Networks," in *Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 510-519: IEEE.

[25] H. Zhang *et al.*, "ResNeSt: Split-Attention Networks," p. arXiv:2004.08955Accessed on: April 01, 2020Available: https://ui.adsabs.harvard.edu/abs/2020arXiv200408955Z

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module " *Computer Vision – ECCV 2018,* vol. 11211, pp. 3-19, 2018.

[27] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving Convolutional Networks with Self-Calibrated Convolutions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020: IEEE.

[28] Y. Li, H. Wang, L. M. Dang, A. Sadeghi-Niaraki, and H. Moon, "Crop pest recognition in natural scenes using convolutional neural networks," *Computers and Electronics in Agriculture,* vol. 169, Feb 2020, Art. no. 105174.

[29] X. Wu, C. Zhan, Y.-K. Lai, M.-M. Cheng, and J. Yang, "IP102: A Large-Scale Benchmark Dataset for Insect Pest Recognition," in *IEEE International Conference on Computer Vision*, 2019, pp. 8779-8788.

[30] Y. Li, H. Wang, L. M. Dang, A. Sadeghi-Niaraki, and H. Moon, "Crop pest recognition in natural scenes using convolutional neural networks," *Computers and Electronics in Agriculture,* vol. 169, p. 105174, 2020.

[31] Z. Tang, Y. Gao, L. Karlinsky, P. Sattigeri, R. Feris, and D. Metaxas, "OnlineAugment: Online Data Augmentation with Less Domain Knowledge," in *European Conference on Computer Vision*, 2020, vol. 12352, pp. 313-329: Springer.

[32] 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)*, Long Beach, CA, USA, 15-20 June 2019, pp. 8787-8796.

[33] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," in *Proc. IEEE conf. Comput. Vis. Pattern. Recognit. (CVPR)*, Salt Lake City, USA, 18-22 June 2018, pp. 7132-7141.

[34] S. Xie, R. Girshick2, P. Doll´ar, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," in *Proc. IEEE conf. Comput. Vis. Pattern. Recognit. (CVPR)*, Honolulu, HI, USA, 21-26 July 2017, pp. 1492-1500.

[35] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, and Z. Zhang, "ResNeSt: Split-Attention Networks," *2020, arXiv:2004.08955. [online]. Available:https://arxiv.org/abs/2004.08955*.

[36] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial Transformer Networks," *2016, arXiv:1506.02025. [Online]. Available: https://arxiv.org/abs/1506.02025*.

[37] D. P. Kingma and J. L. Ba, "ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION," *2015, arXiv:1412.6980. [online]. Available: https://arxiv.org/abs/1412.6980*.

[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Machine Learning Research,* vol. 15, pp. 1929-1958, 2014.

[39] E. Ayan, H. Erbay, and F. Varçın, "Crop pest classification with a genetic algorithm-based weighted ensemble of deep convolutional neural networks," *Computers and Electronics in Agriculture,* vol. 179, p. 105809, 2020.

[40] L. Nanni, G. Maguolo, and F. Pancino, "Insect pest image detection and recognition based on bio-inspired methods," *Ecological Informatics,* vol. 57, 2020.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *International Conference on Neural Information Processing System,* vol. 25, no. 2, 2012.

[42] C. Szegedy *et al.*, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 1-9.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[44] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations*, San Diego, CA, 2014, p. 1556.

[45] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.

[46] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 5987 - 5995: IEEE.

[47] Q.-L. Z. Y.-B. J. a. p. a. Yang, "SA-Net: Shuffle attention for deep convolutional neural networks," 2021.