Deep Learning Based Automatic Multiclass Wild Pest Monitoring Approach Using Hybrid Global and Local Activated Features

Liu Liu, Chengjun Xie, Rujing Wang, Po Yang, *Senior Member, IEEE*, Sud Sudirman, Jie Zhang, Rui Li, and Fangyuan Wang

***Abstract*—Specialized control of pests and diseases have been a high-priority issue for the agriculture industry in many countries. On account of automation and cost effectiveness, image analytic pest recognition systems are widely utilized in practical crops prevention applications. But due to powerless hand-crafted features, current image analytic approaches achieve low accuracy and poor robustness in practical large-scale multiclass pest detection and recognition. To tackle this problem, this article proposes a novel deep learning based automatic approach using hybrid and local activated features for pest monitoring. In the presented method, we exploit the global information from feature maps to build our global activated feature pyramid network to extract pests' highly discriminative features across various scales over both depth and position levels. It makes changes of depth or spatial sensitive features in pest images more visible during downsampling. Next, an improved pest localization module named local activated region proposal network is proposed to find the precise pest objects positions by augmenting contextualized and attentional information for feature completion and enhancement in local level. The approach is evaluated on our seven-year large-scale pest data-set containing 88.6 K images (16 types of pests) with 582.1 K manually labeled pest objects. The experimental results show that our solution performs over 75.03% mean average precision (mAP) in industrial circumstances, which outweighs two other state-of-the-art methods: Faster R-CNN with mAP up to 70% and feature pyramid network mAP up to 72%.**

***Index Terms*—Convolutional neural network (CNN), global activated feature pyramid network, local activated region proposal network, pest monitoring.**

Nomenclature

$\hat{t}_i$	Ground truth coordinate of bounding box.
$\hat{t}_i$	Ground truth label.
a	Output of convolution operation and activation function.
AP	Average precision.
b_k	Bias of convolution kernel k.
C	Number of channels of feature map.
FN	False negatives.
FP	False positives.
H	Height of feature map.
IoU	Intersection-over-union.
Pr	Precision.
Re	Recall.
t_i	Predicted coordinate of bounding box.
t_i	Predicted label.
TP	True positives.
W	Width of feature map.
W_k	Weight of convolution kernel k.

I. Introduction

Specialized and effective pest control and monitoring in agriculture is becoming an increasingly serious issue all around the world [1]. The urgent demand for efficiently controlling and inspecting the occurrence of agricultural pests in fields has driven the rapid development of industrial pest prevention solutions and intelligent pest monitoring systems, such as chemical pesticides [2], image analytic systems [3], automatic adjustable spraying device [4], status estimation of wheat plants [5], remote sensing [6], etc. On account of automation and cost effectiveness, image analytic based pest recognition and monitoring systems are widely utilized in practical crops prevention applications. Typically, these systems install some stationary pest trap devices or facilities in the wild fields for real-time acquisition and transmission of trap images, and then employ advanced image analytic techniques [7]–[10] into these

Manuscript received August 8, 2019; revised November 27, 2019 and March 18, 2020; accepted May 3, 2020. Date of publication May 18, 2020; date of current version July 26, 2021. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61773360 and Grant 31671586, in part by the Major Special Science and Technology Project of Anhui Province under Grant 201903a06020006, and in part by the Innovate U.K. (U.K.-China: Precision for Enhancing Agriculture Productivity) under Grant 671197. Paper no. TII-19-3668. *(Corresponding authors: Rujing Wang; Chengjun Xie.)*

Liu Liu and Fangyuan Wang are with the University of Science and Technology of China, Hefei 230026, China (e-mail: liuliu66@mail.ustc.edu.cn; wfy710@mail.ustc.edu.cn).

Chengjun Xie, Rujing Wang, Jie Zhang, and Rui Li are with the Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China (e-mail: cjxie@iim.ac.cn; rjwang@iim.ac.cn; 76609080@qq.com; lirui@iim.ac.cn).

Po Yang is with the Department of Computer Science, University of Sheffield S10 2TG Sheffield, U.K. (e-mail: po.yang@sheffield.ac.uk).

Sud Sudirman is with the Department of Computer Science, Liverpool John Moores University, L3 5UG Liverpool, U.K. (e-mail: S.Sudirman@ljmu.ac.uk).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2020.2995208

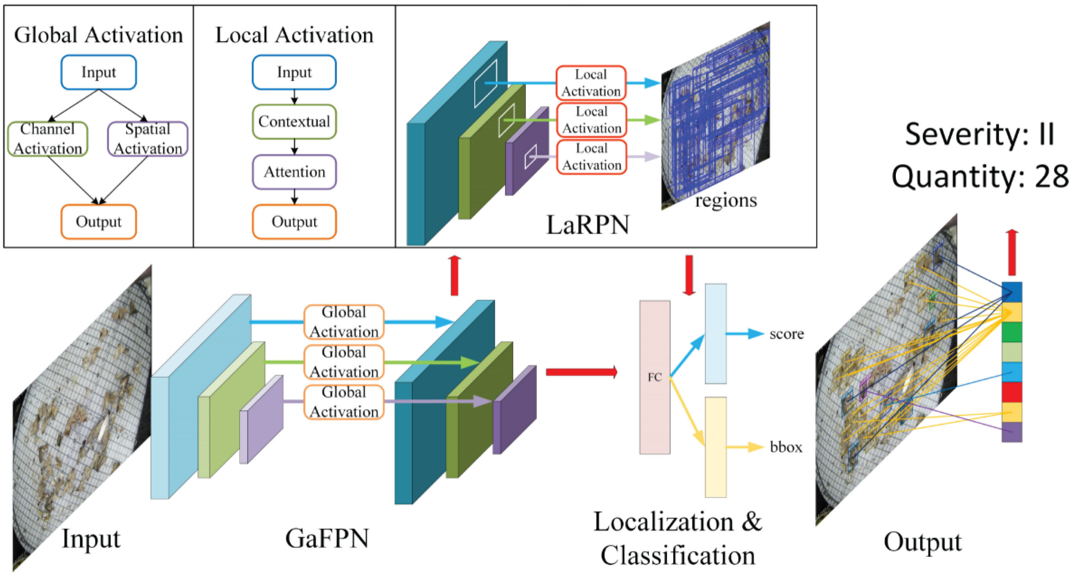

Fig. 1. Pipeline of our pest monitoring system in field environment. The pest detection approach consists of two parts: GaFPN and LaRPN. Our system takes an image captured from in-field environment by our pest monitoring equipment as input, extracts high-quality feature representation by GaFPN and then localize pest regions as well as enhance the features with local activation module for pest recognition. Finally, our system evaluate the pest severity by the high-level detection results.

images for identification and extraction of pest-associated data in support of intelligent prediction and prevention.

Abovementioned advanced image analytic techniques enable abundant success in effectively detecting and recognizing specific pest species. Yet, utilizing these techniques in designing as well as developing practically useful and robust pest monitoring system is still unsatisfied. The first reason for this problem is that extracted features as pest descriptors are short of sufficient details for tiny and blurred pest objects in 2-D static images captured by stationary devices. These pose a fundamental dilemma that it is hard to distinguish small objects from the generic clutter in the background. Also, traditional machine learning approaches have been suffering from many limitations such as powerless hand-crafted features and the lack of expert consensus. Besides, most of current systems focus on whole pest image classification rather than detection, which aims to localize and identify each pest instance in the image that is necessary for high-level pest analysis promoting more efficient pest monitoring systems in the wild. Therefore, toward large-scale multiclass pest monitoring, it is highly necessary to develop a novel automatic approach by mining more valuable information as highly discriminative features for pest detection.

Recently, adv ances in deep learning techniques have led to significantly promising progress in the field of generic object detection, like SSD [11], faster R-CNN [12], feature pyramid network (FPN) [13], and other extended variants of these networks [14], [15]. Among these approaches, two-stage object detection architectures are the most popular in dealing with practical problems due to higher detection accuracy. In faster R-CNN, region-of-interest (RoI) pooling is used to extract features on a single-scale feature map. But targeting at small object detection, FPN is a better state-of-the-art technique over COCO data set [16] with mean average precision (mAP) up to 56.9%. By building up a multiscale image pyramid, FPN enables a model to detect all objects across a large range of scales over both positions and pyramid levels. Besides, feature pyramid structure built on convolutional neural network (CNN) has become a wide selection as it covers low-level object features and high-level semantic features simultaneously. This property is particularly useful to tiny object detection like pest detection.

In this context, this article targets at finding out a practically effective and robust pest monitoring solution by studying the state-of-the-art deep learning methods to solve the problems in current large-scale multiclass pest detection task. As shown in Fig. 1, in our presented method, we first construct a CNN based feature pyramid architecture to ensure the pests across various scales could be found, and then propose a global activated feature pyramid network (GaFPN) for retrieving depth and spatial attention over different levels in the pyramid network. Compared to [12] and [13], this approach, the adjusted network will enable variance or changes of spatial or depth sensitive features in images more visible in the pooling layers. This property will allow some missing features of tiny pests in pooling layers in one level to be redetected by many pyramid levels. Next, an improved pest localization module named local activated region proposal network (LaRPN) is proposed to find the precise pest objects positions by augmenting contextualized and attentional information for feature completion and enhancement in local level. Following this idea, we integrate GaFPN and LaRPN into a two-stage CNN approach. It is evaluated over our newly published large-scale pest detection specific image data set containing 88.6 K raw images with 582.1 K manually labeled pest objects. The image data were collected in the wild field using mobile camera over seven years. The experimental results show that our approach achieves over mAP of 75.03%, which outweighs two other state-of-the-art methods [12] with mAP of 70% and [13] mAP of 72%.

The major contributions of this article are as follows.

1) A novel two-stage CNN based pest monitoring approach using hybrid global and local activated feature is designed for large-scale multiclass pest data set. It is implemented as a practically automatic pest monitoring system, which accurately and effectively detects 16 types pest in fields.
2) The proposed approach introduces two novel global and local activation branches: GaFPN and LaRPN for automatic multiscale feature extraction and efficient region providing and fine tuning, respectively. Our approach could help recognize and extract discriminative features of tiny objects as well as accommodate large variations and changes of distribution of tiny objects over images. It benefits the precise measure and prediction of pest in complex circumstances with multiclass pest insects.
3) A comprehensive and in-depth experimental evaluation on practical industry level large-scale pest data set (88.6 K images) is provided for verifying the usefulness and robustness of the proposed system and approaches. The results show that our approach delivers an mAP of 75.03% over 16 types of pest detection, which outweighs two other state-of-the-art methods: Faster R-CNN [12] with mAP up to 70% and FPN [13] mAP up to 72%.

II. Related Work

In agriculture systems, artificial intelligence and machine learning techniques have been widely used in various monitoring tasks. Ruan *et al.* [17] propose to mine valuable information from agriculture big data to guide the precise management of apple plant for growers. However, environment information might not be enough for building a mature and comprehensive monitoring system while pest is one of the major risks in agriculture applications. In this case, typical image analytic techniques for pest monitoring focus on the study of object identification, including feature extraction and pattern recognition. Early works on insect classification include RGB multispectral analysis [8] and principle component analysis algorithm [18]. Then, more valuable and representative features are considered for precise pest recognition such as size, color [19], shape, and texture [20]. But these types of features were too weak to be insensitive to rotation, scale, and translation. Thus, scale-invariant feature transform in modern computer vision techniques are popular to realize rotational variance for pest classification [21]. On the other hand, classifiers are key to achieve better model training performance, such as support vector machine [22], K-nearest neighbors [23], and artificial neural network [24]. While the aforementioned approaches achieved success to some extent, their results rely too much on the quality of hand-crafted features selection. Toward large-scale multiclass insect data set, one consequence is that within species, extracted descriptors show strong similarity to others. Feature vectors with different species are highly close in feature space to relative variability of their texture, color, shape, and so on. It is hard to utilize these approaches in practical pest monitoring applications, since the process of manually selecting and designing features is laborious and insufficient for multiclass pest species.

Fortunately, the emergence of deep learning techniques has led to significantly promising progress in computer vision techniques that facilitates industrial applications development such as human activity recognition [25], automatic fruit classification [26], plant disease recognition [27], and cloud workload prediction [28]. In smart agricultural applications, under the combination with Internet of Things, various systems are built based on deep learning techniques such as U-Net employed in yellow rust disease monitoring [29] and wild aphid detection system [30]. But the difficulty of remote sensing image capturing limits the real-world applications in this work. In generic image classification and object detection task, CNN has exhibited superior capacities in learning invariance in multiple object categories from large amounts of training data [31]. It enables suggesting object proposal regions in detection process; and extract more discriminative features than hand-engineered features. By detecting locations [12], [14] and fine tuning [32] general representation to a specific object category, CNNs perform well in object detection. Some two-stage approaches [12] utilizes dense sliding window to find out the possible object regions with low-level cues. They localize the better proposals and share the weights of convolutional layers compared with other detectors. They perform even better than one-stage CNN based approaches with higher accuracy of object detection. The abovementioned deep learning methods [11]–[14] have shown great accuracies in many general object detection applications beyond what can be achieved by previous methods [22]–[24], but they are often intractable for pest monitoring applications.

Toward large-scale multiclass pest monitoring, deep learning methods need to integrate with other techniques like feature pyramids [13] for improved performance. The experiment results on the Microsoft COCO data set [16] shows that two-stage object detection framework such as faster R-CNN is an effective region-based object detector toward general object detection with a mAP up to 42.7% because of region proposals are computed at the first stage. But for small object detection, FPN is a better state-of-the-art technique over COCO data set with mAP up to 56.9% due to the fused low-level object features and high-level semantic features. Despite the fact that faster R-CNN has shown great accuracies in generic object detection applications, they are often intractable for use in practical real-world small object detection. Taking our targeted pest detection in the wild as an example, designing an effective deep learning approach is extremely difficult due to many constraints.

1) The intuitive features of pest like texture, shape, or color, are easily confused with background information.
2) Features of tiny pest like rotation, scale, and translation, are too weak and insensitive to be recognized.
3) Many deep learning approaches focus on solving classification of different pests, rather than pest detection (localization and counting).
4) Large variations of density distribution and sizes of tiny pests make the activation of some objects even smaller

and insensitive with each pooling layer through a deep learning architecture.

In order to overcome the abovementioned obstacles, we attempt to propose a new effective deep learning approach toward large-scale multiclass pest monitoring by using hybrid global and local activated features.

III. Approach Overview

Our proposed approach is a two-stage CNN based pest detection and classification pipeline shown in Fig. 1. Two major stages in this approach are GaFPN for automatic multiscale feature extraction and LaRPN for generated boxes classification and regression. Under the powerful global and local feature extracted, the output of our system consists of pest localization, classification, and severity estimation tasks.

In the first stage of feature extraction, our system relies on traditional CNN backbone by introducing a new GaFPN, which is aggregated on each convolutional block for screening and activating depth and spatial information from feature maps outputted by each block. Multiscale image features extracted from GaFPN are used to rebuild the feature maps. This design has two considerations: first, sufficient shallow layers enables mining more valuable semantic features for classification. Second, the bottom layers with high spatial information are fully utilized for avoiding some features vanish in deep CNN block.

In the second stage, based on feature maps extracted from stage one, an improved LaRPN is proposed for providing region proposals and fully connected layers, which are adopted for pest classification and position regression. Different from the standard region proposal network (RPN), we augment local contextualized and attentional information into region proposals for providing more efficient and precise regions.

Finally, we adopt several fully connected layers for the final pest localization and classification results in addition to high-level semantic analysis outputs for pest severity estimation including pest quantity counting and severity prediction. The entire training and inference phase run automatically to achieve effective pest recognition and classification without any human intervention so our method is an end-to-end system.

IV. Materials and Methods

A. Data Set Setup for Large-Scale Multiclass Pest Detection

To the best of our knowledge, while there exist some open insect databases released, no existing large-scale data sets that cover multiclass pests in the wild or natural environments are released for study yet. We establish our large-scale multiclass pest data set by designing an industrial stationary pest monitoring equipment shown in Fig. 2. This equipment uses multispectral light trap for attracting various types of pests, where the wavelengths vary with time according to the habit of pests in the day. Meanwhile, HD camera abovementioned the tray of our equipment is set to take pictures at 2592 × 1944 resolution periodically at 15-s intervals. Pests in the trays were swept away after photographing to avoid images containing 582,170 pests of 16 different types after manual screening to deleting obscure and over-occulted images are used to build our data set.

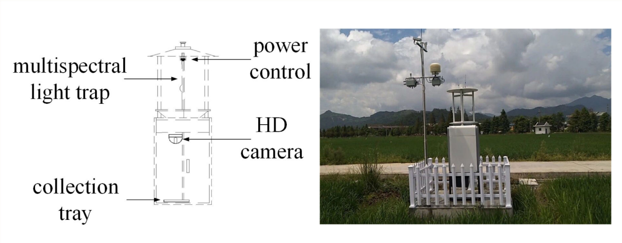

Fig. 2. Stationary pest monitoring equipment in our work. The data set published in this article is collected by this equipment. Besides, we also deploy our system into it for practical pest monitoring.

TABLE I
Statistics on Two Subsets for Our Data Set With Training Subset and Validation Subset. For Each Class, the Number of Images and Objects are Shown in This Table

Pest name	ID	Training Subset		Validation Subset	
		#images	#objects	#images	#objects
CM	1	6,663	11,663	768	1,332
CMw	2	2,956	7,548	367	914
MS	3	11,280	23,055	1,222	2,741
HA	4	22,854	67,426	2,510	7,143
OF	5	17,586	39,126	1,950	4,190
PL	6	21,675	110,309	2,366	12,200
SL	7	7,301	9,857	782	1,079
SE	8	13,212	25,589	1,403	2,544
SI	9	5,136	7,645	583	830
AI	10	8,952	13,844	992	1,553
MB	11	6,389	9,345	719	1,065
HT	12	11,827	21,051	1,287	2,251
HP	13	8,905	30,792	963	3,460
AC	14	13,765	108,112	1,606	12,141
GO	15	9,632	17,432	1,038	2,056
AS	16	4,756	21,728	546	2,219
total		79,800	524,522	8,870	5,7648

Note that because single image may contain objects of several classes, the total shown in the #images columns are not simply the sum of the corresponding columns. (CM: Cnaphalocrocis medinalis, CMw: Cnaphalocrocis medinalis (walker), MS: Mythimna separate, HA: Helicoverpa armigera, OF: Ostrinia furnacalis, PL: Proxenus lepigone, SL: Spodoptera litura, SE: Spodoptera exigua, SI: Sesamia inferens, AI: Agrotis ipsilon, MB: Mamestra brassicae, HT: Hadula trifolii, HP: Holotrichia parallela, AC: Anomala corpulenta, GO: Gryllotalpa orientalis, AS: Agriotes subrittatus).

Hereafter, images are labeled by agricultural experts with pest categories, localization, and severity. we randomly split entire collected images into two subsets for model training and validation respectively at the ratio of 9:1, in which training subset is employed to supervise our model because of labels with expert consensus and validation subset is used to evaluate our system's performance. The statistics of our data set are provided in Table I.

B. Convolutional Neural Network (CNN) Framework

The approach built on a standard CNN framework is composed of three parts: convolutional layer, activation function, and pooling layer. Typically, many combinations of these layers are

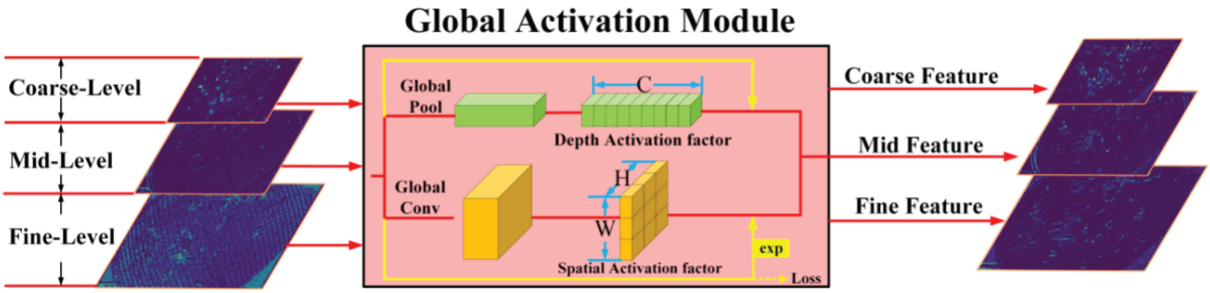

Fig. 3. Structure of GaFPN. There are two parallel branches in GaFPN for activating depth and spatial features, respectively. Note that the GAM is applied in various feature levels (coarse, mid, and fine).

adopted to extract 3-D image features, in which images are input into convolutional layers computed with several convolutional kernels for feature extraction.

Standard convolutional layer takes a set of convolutional kernels to the input and the output feature map in each subsequent layer are regarded as abstract transformations of image. Generally, for each convolutional kernel k, the forward propagation process of convolution in layer l could be represented by

$$a_k^l = \sigma(z_k^l) = \sigma(a^{l-1} * W_k^l + b^l) \tag{1}$$

where the a_k^l and a^{l-1} are output of kernel k from layer l and $l-1$. $\sigma(\cdot)$ represents ReLU function [33] for nonlinear transformation in our approach. $*$ indicates the convolution operation. W_k^l and b_k^l represent the convolution kernel and bias in layer l, respectively. Therefore, the output convolutional layer could be computed as the sum of outputs from the set of kernels

$$a^l = \sigma(z^l) = \sigma\left(\sum_{k-1}^{M} z_k^l\right) = \sigma\left(\sum_{k=1}^{M} (a_k^l * W_k^l) + b^l\right) \tag{2}$$

C. Global Activated Feature Pyramid Network (GaFPN)

Based on standard CNN architecture, we design our feature extraction network named GaFPN whose structure is shown in Fig. 3. The motivation of designing feature pyramid is the observation that recognizing pests at vastly different scales in images is challenging for detectors in a single feature map. Thus, we exploit the inherent multiscale hierarchy of CNN to achieve feature extraction at various scales to ensure that pests with different sizes are recognized with sufficient information and avoid missing features of some tiny pests during downsampling. In GaFPN, the powerful representative information from all convolutional blocks, including high-resolution levels and high-semantic levels, could be featurized to produce a multiscale pest feature descriptor.

Different from the popular object detection framework FPN [13], our GaFPN makes full use of global information between each convolution block to avoid information loss during downsampling operation. As it is well known, feature maps outputted from CNN layers could be a result of convolutional operation with set of kernels. The number of kernels corresponds to be the feature depth and each kernel is learned to extract the specific feature type such as shape and texture. Therefore, we attempt to make our system automatically mine the depth activation vector that weighs the convolutional kernels for highlighting the requirement of various feature types for pest detection in our work. As for spatial activation, we observe that limited receptive field of convolution operations might lead to powerless features in pests positions without appropriate supervision. So, we propose a novel supervised mask to learn the spatial activation vector that could activate the potential positions of pests. Under these motivations, our GaFPN is proposed to achieve depth and spatial activation in global level that aims to improve the feature discriminating power of pest objects from background.

Fig. 3 shows the structure of our GaFPN, in which global activation module (GAM) contains two branches for depth and spatial activation, respectively. In the upper branch of depth activation, the 3-D feature map with shape of $W \times H \times C$ outputted from CNN block is first processed by a global pooling layer that averages all the pixels in each channel (depth) and generates a lower-dimensional (1-D) feature vector ($1 \times 1 \times C$) so the effect of spatial information is ignored. By taking global pooling, the averaged feature vector describes the global feature in depth level. Next, we apply two sets of fully connected layers with nonlinear activation ReLU [33] and sigmoid, respectively, in which the latter aims to map the feature vector into (0, 1). In this way, the output 1-D vector could be learned as depth activation factor in training phase. The final output of depth activation module is the broadcast element-wise product of the input 3-D feature maps ($W \times H \times C$) and 1-D depth activation factor ($1 \times 1 \times C$). In this way, the feature maps are activated in depth.

The second branch of GAM in Fig. 3 is designed for activating spatial position that introduces a novel supervised mask to learn a spatial activation vector. Specifically, the spatial activation branch is a segmentation-like training method, in which the supervised mask is obtained by fulfilling 1 into the ground truth positions and 0 into the background areas. In this part, the input feature map with shape of $W \times H \times C$ is input into a global convolution operation that reduce the number of channels to 1 and the output is a $W \times H \times 1$ feature vector, which could ensure the spatial activation vector is learned in spatial level. Then, we employ two sequential convolution operations with ReLU and Sigmoid function followed which is similar to depth activation branch. For training the spatial activation weights, we adopt pixel-wise sigmoid across entropy as the supervised

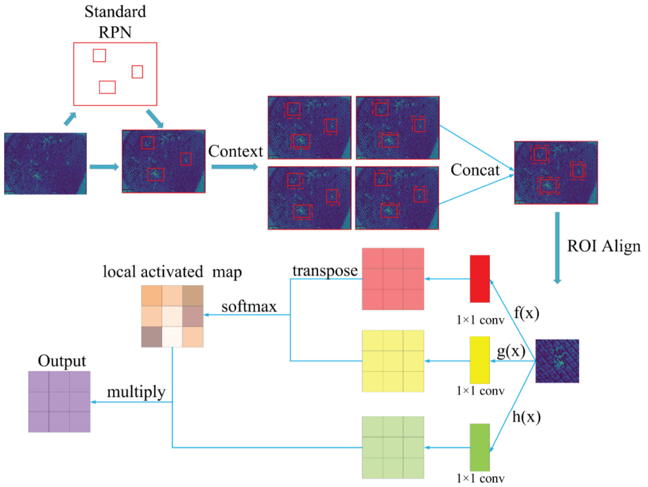

Fig. 4. Structure of LaRPN. This module is improved from standard RPN, in which we introduce contextual and spatial attention mechanism to enhance the feature quality for each pest region.

attention training loss. At last, the learned spatial activation factor is fed into exponential operation to maintain the original global information and then multiplied with the input 3-D feature maps in each position. In this way, our spatial activation branch could enhance the feature maps in pest objects area and diminish the opposition.

Finally, the output of each block in GaFPN is the sum of the activated feature maps from two parallel branches in different CNN levels.

D. Local Activated Region Proposal Network (LaRPN)

Under the GaFPN module, our system could localize the pests' positions in global level of input image. For achieving precise pest recognition in our work, we propose LaRPN to further refine the pest features in local level.

The first motivation of "local activated" is that part of region proposals provided by standard RPN might not cover complete information of target objects. This would result in inaccurate box regression with insufficient features because RoI Align [34] is used to "crop" the regions into local level from global feature maps. To solve this problem, we augment extra contextual information [15] to ensure that enough object features could be considered into box regression. Second, the local spatial positions might also contribute to pest recognition task because the key feature for precise classification might be the fine-grained characteristics such as colors or shapes of pests' wings.

Motivated by these observations, we propose an improvement of standard RPN named LaRPN to take contextual and attentional information into consideration to locally activate the region proposals provided by RPN, whose structure is shown in Fig. 4. There are three steps in our LaRPN. First, apply the standard RPN [12] with our assigned anchors in each output feature maps from GaFPN with various levels of feature pyramid structure. During training phase, the anchors with intersection-over-union to ground truth more than 0.7 are regarded as preliminary pest regions. Next, we expand these positive regions to be 1.5 times larger in four different directions to ensure the regions could cover larger areas as contextual information. And the enriched pest regions are mapped to feature maps and processed by RoI Align [34] to extract local features. Finally, we introduce self-attention mechanism [35] with three parallel branches to obtain the local attention vector in spatial level. Therefore, the relationships among different positions of pests could be learned and the output is multiplication of regions and spatial activated map. Finally, the output feature is used for pest recognition and box fine tuning.

E. Training and Evaluation

We use large-scale pest data set for training and validating our proposed approach. Different loss functions are selected as supervisory indicators for pest localization, classification, and estimation training. Besides, a number of evaluation metrics were built to access the performance of our system on these tasks.

Pest Localization: Pest localization aims to predict bounding boxes of pest regions for input image. To validate the performance of pest localization task, we pay more attention to the positioning accuracy rather than the categories of pest species. Therefore, we consider sigmoid cross entropy loss as the criterion for indicating pest region objectness as well as smooth L1 loss for pest region box regression in this task referred by [12], which is the combination of L1 and L2 norm defined as

$$\mathrm{Loss_L} = \sum_{\mathrm{i}\in\{\mathrm{x,y,w,h}\}} \begin{cases} 0.5(t_i - \hat{t}_i)^2 & |t_i - \hat{t}_i| < 1 \\ |t_i - \hat{t}_i| - 0.5 & |t_i - \hat{t}_i| \geq 1. \end{cases} \quad (3)$$

In this loss function, a region could be characterized by $\{t_x, t_y, t_w, t_h\}$ in which $\{t_x, t_y\}$ are the upper-left coordinates of boxes and $\{t_w, t_h\}$ are the width and height. Thus, t_i and $\hat{t}_i$ represent the ground truth and localized bounding boxes, respectively.

In terms of evaluation metrics, binaryprecision (Pr) and recall (Re) are chosen to validate the pest localization performance in our work. During validation phase, the regions are predicted into two categories: objectness and background, in which objectness (positive) samples are the regions with overlap more than 0.7 with the ground truth bounding boxes while the background regions are negative ones. The binary Pr and Re are calculated by

$$\mathrm{Pr} = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FP}} \quad (4)$$

$$\mathrm{Re} = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FN}} \quad (5)$$

in which #TP, #FP, and #FN represent the number of true positive, false positive, and false negative samples, respectively so the Pr measures the samples that are incorrectly detected while higher Re indicates the lower misdetection rate.

Furthermore, average precision (AP) for binary pest localization is applied as a comprehensive evaluation metric to take the precision and recall into consideration together. In pest localization task, the $\mathrm{AP_L}$ is computed by the integration of precision–recall (PR) curve

$$\mathrm{AP_L} = \int_0^1 \mathrm{Pr}\, d\mathrm{Re} \tag{6}$$

Pest Classification: while localizing pest objects in images, we classify each bounding box of pest into the corresponding category. Different from binary classification in localization task (objectness or background), the bounding boxes are classified into 16 types that are the major pest species in practical agricultural pest monitoring demand in our work. In this task, we use multiclass cross-entropy loss for this pest classification problem

$$\mathrm{Loss_C} = \sum_{i=1}^{N_{cls}} -y_i \log(\hat{y}_i) \tag{7}$$

where N_{cls} represents the number of pest categories (in our task, $N_{cls} = 16$). y_i and $\hat{y}_i$ indicate the truth label and predicted category, respectively. From the perspective of evaluation metrics for pest classification, we combine the localization and classification evaluation methods together to update the AP value [16] for different categories. Thus, in our system, we calculate APs for 16 pest species based on their corresponding PR curves as

$$\mathrm{AP(c)} = \int_0^1 \mathrm{Pr(c)dRe(c)} \tag{8}$$

In addition, the final metric for pest classification task mAP is obtained by taking average of APs from all the pest categories

$$\mathrm{mAP} = \frac{1}{\mathrm{N_{cls}}} \sum \mathrm{AP(c)} \tag{9}$$

Pest severity estimation: the high-level task, pest severity estimation targets at predicting the severity of pest occurrence from the input image. According to agricultural experts' consensus, the severities are divided into five levels from "general" to "serious" that describes the occurrence of pests in the field, so the images are labeled to I–V by agricultural experts after image acquisition. In the process of pest severity prediction, the input features are the combined results from localization and classification tasks abovementioned as well as the visual features extracted from pest image. In terms of encoding method, we adopt a variant of one-hot encoder to transform the pest detection results into N_{cls}-dimensional vector, where each element in this vector indicates the number of detected pests with the corresponding category. In this input vector, we only focus on the quantity of detected pests from each category rather than their positions.

In pest severity estimation task, we build a neural network with consequent two fully connected layers for feature extraction and softmax predictor for severity estimation. As criterion, we employ a weighted multiclass cross-entropy loss defined as

TABLE II
PEST LOCALIZATION RESULTS AP_L

CNN Backbone	Method	AP_L
Inception [36]	Faster R-CNN [12]	74.99%
	FPN [13]	76.65%
	Ours	80.11%
ResNet50 [37]	Faster R-CNN [12]	78.74%
	FPN [13]	80.29%
	Ours	83.61%

$$\mathrm{Loss_E} = \sum_{i=1}^{N_{cls}} -\lambda_i y_i \log(\hat{y}_i) \tag{10}$$

where λ_i is a hyperparameter to weight the loss function which measures the risk of different misclassification samples. We define the risk parameter λ_i as the difference between predicted severity and truth severity. As for evaluation, we consider total accuracy as evaluation metric for pest severity estimation task.

V. Experimental Results and Discussion

A. Implementation Details

We use inception [36] and ResNet50 [37] as CNN backbones to train our pest monitoring model and also build some experiments to evaluate the performance of our system. During the training phase, The proposed method and other state-of-the-art approaches are trained via back propagation and stochastic gradient descent, with momentum 0.9 [38] and initialize learning rate to 0.001 that will be dropped by ten referred by [39]. The mini-batch size is set to four in training phase. In terms of weight initialization, we adopt transfer learning [40] that copy the CNN backbones' weights pretrained on ImageNet data set [41]. In order to avoid over-fitting problem, we utilize early stopping strategy [42] to select the best training epoch. We conduct our experiments using two GTX1080Ti GPUs with 12 GB memory. The performance of our approach is evaluated on our built data set across multiple tasks: pest localization, classification, and severity estimation.

B. Pest Localization Task

For pest localization task, the experimental results are presented in Table II, in which we compare our method with two state-of-the-art baseline approaches faster R-CNN [12] and FPN [13] that are the popular detectors utilized in practical monitoring systems in industrial circumstance. As it can be seen, our proposed method could dramatically surpass the localization performance of faster R-CNN using different CNN backbones for feature extraction, which achieves 5.12% and 4.87% AP_L improvement, respectively. Besides, compared with another feature pyramid method FPN, our system could also obtain a slight improvement in pest localization task. Among these results of

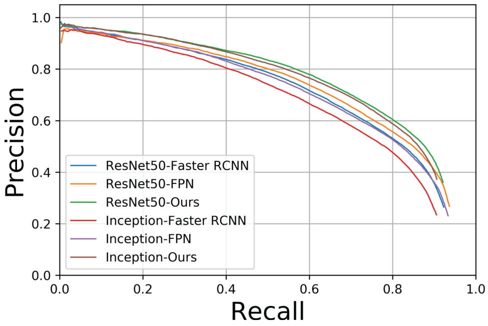

Fig. 5. Precision-recall curve for pest localization task.

TABLE III
PEST CLASSIFICATION TASK RESULTS AP VALUE (%). FR-CNN INDICATES FASTER R-CNN METHOD

Pest ID	Inception [36]			ResNet50 [37]		
	FR-CNN [12]	FPN [13]	Ours	FR-CNN [12]	FPN [13]	Ours
1	51.62	60.24	63.91	57.12	62.13	64.24
2	56.26	61.00	63.11	59.70	62.96	65.98
3	64.27	67.33	68.85	69.75	70.16	73.96
4	80.74	82.10	84.37	83.73	82.82	85.68
5	65.65	69.73	74.39	70.17	71.22	76.48
6	65.36	68.45	69.38	68.60	68.98	70.32
7	63.09	63.30	66.76	68.39	69.46	70.43
8	45.31	49.70	53.43	48.57	53.47	54.19
9	69.93	71.17	76.57	72.56	72.91	77.81
10	75.55	76.27	79.33	79.92	80.58	81.13
11	50.71	51.74	57.04	54.45	57.35	64.02
12	63.17	66.78	68.17	66.26	69.20	71.15
13	77.48	83.31	85.38	84.94	85.18	86.65
14	79.43	86.93	88.51	87.86	88.03	88.76
15	89.81	89.77	90.27	89.93	89.97	90.31
16	69.13	72.51	76.46	73.38	74.37	79.40
mean	66.72	70.02	72.87	70.96	72.42	75.03

our method, the best performance occurs in ResNet50 backbone which achieves localization accuracy with 83.61% AP_L.

It is interesting to note the detailed pest localization performance between our approach and other state-of-the-art methods in Fig. 5, which shows the precision–recall curve of various approaches. Obviously, our proposed global and local activated approach outperforms faster R-CNN by a large margin and improves FPN architecture slightly. This improvement could be contributed to two reasons. First, our method with GaFPN applies a pyramid feature extraction architecture and localize pests' regions on multilevel feature maps that could help precisely find pests positions on various scales, which is also evidence from AP_L values of our method in Table II. Second, holding global activation factors by our presented global activated features for enhancing the depth and spatial information in global level makes it easier to localize pests' positions due to the result that much more highly discriminative features between foreground and background could be extracted.

C. Pest Classification Task

For pest classification task, we show the AP results in Table III for 16 pest categories performed by our method and other state-of-the-art models. Observed from Table III, having pest localization information associated with the predicted bounding boxes to pests, our method could achieve more accurate pest recognition on these pest species. It is obvious that our approach could significantly outperform faster R-CNN in pest classification over almost all the pest categories under inception as CNN backbone. The homologous phenomenon occurs in that using ResNet50 network with 3.28% mAP improvement. In addition, our approach could also largely improve mAP compared to the feature pyramid object detection structure FPN. This gain is largely due to our LARPN's ability to introduce the contextual and local activated information before fully connected layers for pest classification, which is helpful to sufficiently learn the features of pest regions in local level.

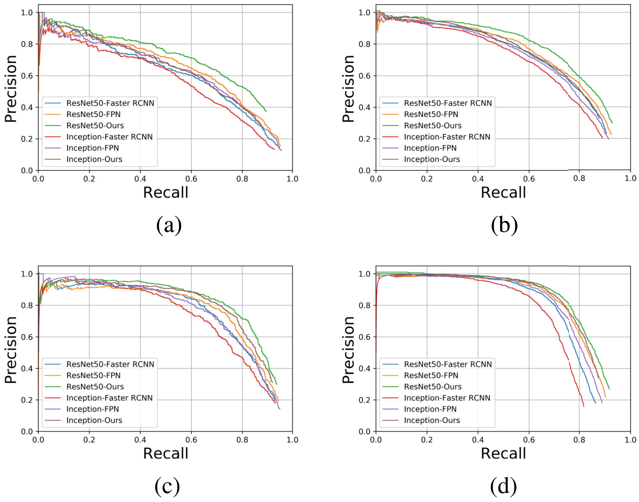

Fig. 6. Illustration of PR curves in our experiments. Note that only four curves are shown here due to the limited space. (a) PR Curve for class 2. (b) PR Curve for class 3. (c) PR Curve for class 9. (d) PR Curve for class 16.

Apart from mAP results, there are obvious differences within classes that can be discussed in Table III. Specifically, pest #8 seems to be the most difficult to be categorized on these precalculated regions with lowest AP value while almost all the models could classify pest #15 well on various CNN backbone. This can be explained by that the pests in the "easy" class hold up a large number of training examples, which help reduce difficulty to classify them comparing Tables III and I. Even though, the amount of data might not be the main factor affecting performance of our approach, where pest #16 still could be categorized with a large AP value (more than 79% AP in our method) even if there are only 4756 training images containing pests of this class. Therefore, our method could overcome the sample limitation and imbalance problem with a great improvement.

Fig. 6 illustrates precision–recall curves for part of pest categories in our experiments. As it is shown, precision could keep a high value with the recall increasing in various models. Especially, our approach using different CNN backbones could obtain a larger precision and recall compared to faster R-CNN, which indicates that it could effectively reduce false positive rate as well as misdetections rate. Concretely speaking, pest #2 is relatively difficult to classify so the PR curve for this class is further away from the point (1,1). In addition, PR curve for

TABLE IV
PEST SEVERITY ESTIMATION TASK RESULTS

CNN Backbone	Method	Accuracy
Inception [36]	Softmax [43]	80.5%
	ours	83.0%
ResNet50 [37]	Softmax [43]	84.9%
	ours	86.9%

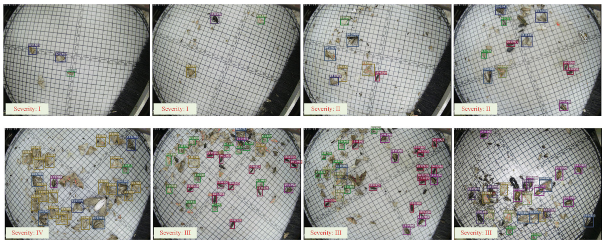

Fig. 7. Examples of pest monitoring results demonstration. The background of the input image from top to bottom is getting complicated.

pest #16 represents that it is hard to obtain a high-recall value but could get satisfied precision value so this curve signifies that our system could make sure that almost all the detected insects of this class are correct but might not detect all of the insects. Furthermore, among these illustrated PR curves, our system performs best on class #3 that maintains high precision in addition to recall simultaneously.

D. Pest Severity Estimation Task

For pest severity estimation, our method regards this task as a classification problem so we achieve severity estimation based on the encoded results outputted from previous pest localization and classification tasks as well as the visual features of input image. In this case, we compare our severity estimation predictor with the state-of-the-art CNN based models that estimate severity by softmax classifier using the whole image as input. Table IV illustrates the comparable results in our experiments. As it is shown, under the high-level pest detection information guidance, our method could beat these CNN approaches with approximately 2% classification accuracy improvement on severity estimation task.

E. Result Visualization

We visualize part of the pest monitoring results in Fig. 7 that fuses localization, recognition, and severity estimation tasks together. These results are outputted by our system based on ResNet50 backbone. The environments of input images from top to bottom are more and more complicated. As it can be seen, our method could achieve multiclass pest localization and recognition under both simple and complicated environments and provide the predicted severity estimation, despite the intractable challenges such as noisy image and tiny objects. Some feature maps outputted from two middle blocks with FPN (left) and our method (right) using ResNet50 are visualized in Fig. 8. It could be found that, the feature maps in our system diminish the highlights of nonobjects and focus more attention on pest regions with lighter activation points with our designed GaFPN architecture. Therefore, our method could perform better on pest detection and progressively learn the pests' features well.

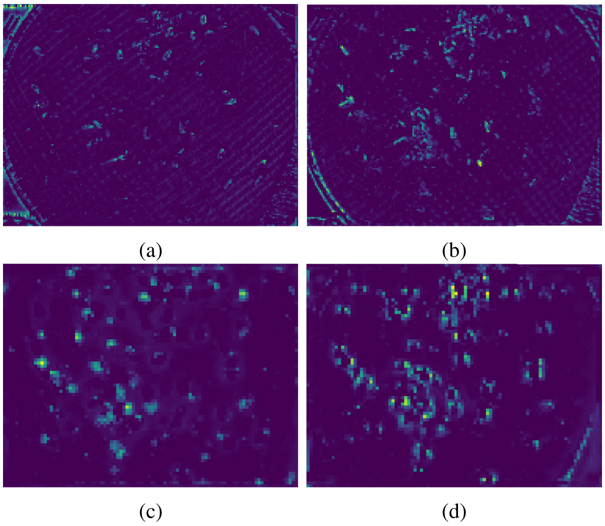

Fig. 8. Part of feature maps generated by FPN (left) and our method (right). (a) From shallow block by FPN. (b) From shallow block by our method. (c) From deep block by FPN. (d) From deep block by FPN.

F. Future Work

Despite that we develop a novel deep learning based system for pest monitoring task in the field and achieve a successful performance in our data set, there are several limitations of our method that could be improved in future smart agriculture innovation. First, the unbalanced data structure could be alleviated in the next work. Specifically, due to the difficulty in capturing pests of some rare categories in our pest monitoring equipment, our system tends to identity an unknown pest into the common species, which might improve the risk of inaccurate pest severity warning. Besides, it is necessary to achieve the real-time pest image recognition and detection performance in our system, in which current inference time might be an important factor that limits the advances in agricultural applications. Therefore, future work would target at solving the problem of unbalanced data set and focus on developing real-time automatic pest monitoring system.

VI. CONCLUSION

This article proposed a novel deep learning approach using hybrid global and local activated features for automatic pest monitoring in industrial equipment to simultaneously perform three key tasks: localization, classification, and severity estimation. Our method successfully realized efficient and automatic feature extraction with global activated feature pyramid GaFPN structure. Furthermore, we presented local activation to enhance position-sensitive features of pest boxes by LaRPN for powerful regions proposal. Under our enriched pest data set captured by our designed stationary pest monitoring equipment, our method outperformed the state-of-the-art methods in pest localization, classification, and severity estimation tasks. Future work will consider developing more efficient deep learning architecture for real-time pest monitoring.

References

[1] G. D. Santangelo, "The impact of FDI in land in agriculture in developing countries on host country food security," *J. World Bus.*, vol. 53, no. 1, pp. 75–84, 2018.

[2] B. L. Bures, K. V. Donohue, R. M. Roe, and M. A. Bourham, "Nonchemical dielectric barrier discharge treatment as a method of insect control," *IEEE Trans. Plasma Sci.*, vol. 34, no. 1, pp. 55–62, Feb. 2006.

[3] H. Liu, S.-H. Lee, and J. S. Chahl, "A multispectral 3-D vision system for invertebrate detection on crops," *IEEE Sensors J.*, vol. 17, no. 22, pp. 7502–7515, Nov. 2017.

[4] R. Berenstein and Y. Edan, "Automatic adjustable spraying device for site-specific agricultural application," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 2, pp. 641–650, Apr. 2017.

[5] S. B. Sulistyo, W. L. Woo, and S. S. Dlay, "Regularized neural networks fusion and genetic algorithm based on-field nitrogen status estimation of wheat plants," *IEEE Trans. Ind. Informat.*, vol. 13, no. 1, pp. 103–114, Feb. 2016.

[6] J. Luo, W. Huang, J. Zhao, J. Zhang, C. Zhao, and R. Ma, "Detecting aphid density of winter wheat leaf using hyperspectral measurements," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 2, pp. 690–698, Apr. 2013.

[7] W. Ding and G. Taylor, "Automatic moth detection from trap images for pest management," *Comput. Electron. Agriculture*, vol. 123, pp. 17–28, 2016.

[8] I. Zayas and P. W. Flinn, "Detection of insects in bulkwheat samples with machine vision," *Trans. ASAE*, vol. 41, no. 3, 1998, Art. no. 883.

[9] J. Cho *et al.*, "Automatic identification of whiteflies, aphids and thrips in greenhouse based on image analysis," *Red*, vol. 346, no. 246, 2007, Art. no. 244.

[10] C. Wen, D. E. Guyer, and W. Li, "Local feature-based identification and classification for orchard insects," *Biosystems Eng.*, vol. 104, no. 3, pp. 299–307, 2009.

[11] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[14] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[15] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 951–959.

[16] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[17] J. Ruan, H. Jiang, X. Li, Y. Shi, F. T. Chan, and W. Rao, "A granular GA-SVM predictor for big data in agricultural cyber-physical systems," *IEEE Trans. Ind. Informat.*, vol. 15, no. 12, pp. 6510–6521, Dec. 2019.

[18] P. J. Weeks, M. A. O'Neill, K. Gaston, and I. Gauld, "Species–identification of wasps using principal component associative memories," *Image Vis. Comput.*, vol. 17, no. 12, pp. 861–866, 1999.

[19] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 761–769.

[20] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *6th Indian Conf. Comput. Vis., Graph. Image Process.*, 2008, pp. 722–729.

[21] L. O. Solis-Sánchez *et al.*, "Scale invariant feature approach for insect monitoring," *Comput. Electron. Agriculture*, vol. 75, no. 1, pp. 92–99, 2011.

[22] N. Larios, B. Soran, L. G. Shapiro, G. Martinez-Munoz, J. Lin, and T. G. Dietterich, "Haar random forest features and SVM spatial matching kernel for stonefly species identification," in *Proc. 20th Int. Conf. Pattern Recognit.*, 2010, pp. 2624–2627.

[23] X.-L. Li, S.-G. Huang, M.-Q. Zhou, and G.-H. Geng, "KNN-spectral regression LDA for insect recognition," in *Proc. 1st Int. Conf. Inf. Sci. Eng.*, 2009, pp. 1315–1318.

[24] Y. Kaya and L. Kayci, "Application of artificial neural network for automatic detection of butterfly species using color and texture features," *Vis. Comput.*, vol. 30, no. 1, pp. 71–79, 2014.

[25] Q. Zhu, Z. Chen, and Y. C. Soh, "A novel semisupervised deep learning method for human activity recognition," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 3821–3830, Jul. 2019.

[26] M. S. Hossain, M. Al-Hammadi, and G. Muhammad, "Automatic fruit classification using deep learning for industrial applications," *IEEE Trans. Ind. Informat.*, vol. 15, no. 2, pp. 1027–1034, Feb. 2019.

[27] F. Wang, R. Wang, C. Xie, P. Yang, and L. Liu, "Fusing multi-scale context-aware information representation for automatic in-field pest detection and recognition," *Comput. Electron. Agriculture*, vol. 169, 2020, Art. no. 105222.

[28] Q. Zhang, L. T. Yang, Z. Yan, Z. Chen, and P. Li, "An efficient deep learning model to predict cloud workload for industry informatics," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3170–3178, Jul. 2018.

[29] J. Su *et al.*, "Aerial visual perception in smart farming: Field study of wheat yellow rust monitoring," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2242–2249, Mar. 2020.

[30] R. Li *et al.*, "A coarse-to-fine network for aphid recognition and detection in the field," *Biosystems Eng.*, vol. 187, pp. 39–52, 2019.

[31] P. Li, Z. Chen, L. T. Yang, Q. Zhang, and M. J. Deen, "Deep convolutional computation model for feature learning on big data in internet of things," *IEEE Trans. Ind. Informat.*, vol. 14, no. 2, pp. 790–798, Feb. 2017.

[32] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. 2nd Int. Conf. Learn. Representations*, 2014.

[33] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*.

[34] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[35] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[36] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[38] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer, 2012, pp. 421–436.

[39] K. Chen *et al.*, "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.

[40] H. K. Suh, J. IJsselmuiden, J. W. Hofstee, and E. J. van Henten, "Transfer learning for the classification of sugar beet and volunteer potato under field conditions," *Biosyst. Eng.*, vol. 174, pp. 50–65, 2018.

[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[42] R. Caruana, S. Lawrence, and C. L. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 402–408.

[43] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.