

TANG, Y., ZHAO, J., HUANG, H., ZHUANG, J., TAN, Z., HOU, C., CHEN, W. and REN, J. 2022. Multiscale voting mechanism for rice leaf disease recognition under natural field conditions. *International journal of intelligent systems* [online], 37(12), pages 12169-12191. Available from: <https://doi.org/10.1002/int.23081>


Multiscale voting mechanism for rice leaf disease recognition under natural field conditions.

TANG, Y., ZHAO, J., HUANG, H., ZHUANG, J., TAN, Z., HOU, C., CHEN, W.
and REN, J.

2022

This is the accepted manuscript of the above article, which has been published in final form at <https://doi.org/10.1002/int.23081>. This file is distributed under the publisher's terms and conditions of use for self-archived versions: <https://authorservices.wiley.com/author-resources/Journal-Authors/licensing/self-archiving.html#3>

Multiscale voting mechanism for rice leaf disease recognition under natural field conditions

Yu Tang¹ | Jinfei Zhao¹ | Huasheng Huang¹  | Jiajun Zhuang² | Zhiping Tan¹ | Chaojun Hou² | Weizhao Chen¹ | Jinchang Ren^{1,3}

¹Academy of Interdisciplinary Studies, Guangdong Polytechnic Normal University, Guangzhou, China

²Academy of Contemporary Agriculture Engineering Innovations, Zhongkai University of Agriculture and Engineering, Guangzhou, China

³College of Computing, Robert Gordon University, Aberdeen, UK

Correspondence

Huasheng Huang, Academy of Interdisciplinary Studies, Guangdong Polytechnic Normal University, No. 293 Zhongshan Avenue, Guangzhou 510665, China.
Email: huanghsheng@gpnu.edu.cn

Jinchang Ren, Academy of Interdisciplinary Studies, Guangdong Polytechnic Normal University, No. 293 Zhongshan Avenue, Guangzhou 510665, China.
Email: jinchang.ren@ieee.org

Funding information

National Natural Science Foundation of China, Grant/Award Number: 32071895; Planned Science and Technology Project of Guangdong Province, China, Grant/Award Numbers: 2019B020216001, 2019A050510045, and

Abstract

Rice leaf disease (RLD) is one of the major factors that cause the decline in production, and the automatic recognition of such diseases under natural field conditions is of great significance for timely targeted rice management. Although many machine learning approaches have been proposed for RLD recognition, scale variation is still a challenging problem that affects prediction accuracy, especially in uncontrolled environments, such as natural fields. Also, the existing RLD data sets are collected in laboratory environments or with a constant scale, which cannot be used to develop the RLD classification algorithms under natural field conditions. To tackle these particular challenges, we propose a multiscale voting mechanism for RLD recognition under natural field conditions. First, data from 26 rice fields were collected to build a data set containing 6046 images of RLD. Afterwards, a feature pyramid was embedded into a mainstream classification architecture (EfficientNet) with a bottom-up and top-down pathway for feature fusion at different scales. To further reduce the inconsistency among multiscaled features, a multiscale voting strategy with regard to

2021A0505030075; Natural Science Foundation of Guangdong Province, China, Grant/Award Numbers: 2020B1515120070, and 2021A1515010824; Key Project of Universities in Guangdong Province, China, Grant/Award Number: 2020ZDZX1061; Innovation Team Project of Universities in Guangdong Province, China, Grant/Award Number: 2021KCXTD010; Planned Science and Technology Project of Guangzhou, China, Grant/Award Numbers: 202002020063, 202007040007, and 202103000028; Basic and Applied Basic Research Fund in Guangdong Province, China, Grant/Award Number: 2021A1515110756; Project of Educational Commission of Guangdong Province, China, Grant/Award Number: 2021KQNCX044; Rural Revitalization Strategy Project of Guangdong Province, China, Grant/Award Number: 2019KJ138

probability distribution was proposed to integrate the decisions from various scales. Each proposed module was carefully validated through an ablation study to demonstrate its effectiveness, and the proposed method was compared with a few state-of-the-art algorithms, including the Single Shot MultiBox Detector, Feature Pyramid Networks, Path Aggregation Network, and Bidirectional Feature Pyramid Network. Experimental results have shown that the classification accuracy of our model can reach 90.24%, which is 4.48% higher than that of the original EfficientNet-b0 model and 1.08% higher than that of existing multiscale networks. Finally, we exploit and demonstrate a visualized explanation for the boosted performance from the proposed model. As an extra outcome, our data set and codes are available at <http://github.com/huanghsheng/multiscale-voting-mechanism> to benefit the whole research community.

KEYWORDS

EfficientNet, feature pyramid network, multiscale voting mechanism, rice leaf disease recognition

1 | INTRODUCTION

Rice is a cereal crop widely planted around the world, and it is the staple food of more than half of the world's population.¹ The yield and quality of rice have a significant influence on human life and food security. However, during rice planting, rice leaf disease (RLD) is one of the most serious threats to rice agroecosystems worldwide, and the RLD-caused yield reduction can be up to 75%.² Therefore, rapid and accurate identification of RLD categories is of great significance for timely targeted management and to ensure the food security of most countries in the world.³ Currently, the recognition of RLDs is mainly conducted by manual investigation in real applications,⁴⁻⁶ which suffers from several limitations for accurate identification. On the one hand, the manual investigation is time-consuming, laborious, and unreliable.^{7,8} On the other hand, due to the limited resources of professionals, it is difficult and costly for most farmers to invite experts to identify such diseases in time, due to lack of agronomic knowledge, leading to failure in timely identification and mitigation of such diseases.⁹ Therefore, it is of great significance to develop an automatic RLD identification system for rice cultivation.¹⁰

Recently, with the development of computer vision and machine learning technologies, many researchers have used machine vision and deep learning techniques to identify crop diseases. Xiao et al.¹¹ proposed integrating handcrafted features and the Principal Component Analysis method for feature extraction and applied a backpropagation network for the classification of four rice blast

spots. Sethy et al.¹² introduced 5932 on-field images of four types of RLDs and evaluated the performance of 11 convolutional neural network (CNN) models with the transfer learning and deep feature plus the support vector machine (SVM) methods. Experimental results showed that the deep feature plus SVM method has better classification performance than other CNN counterparts. However, during the construction of the RLD database, the diseased portion was manually extracted from the original large images, which is different from the natural field conditions. Jiang et al.¹³ applied a CNN and SVM model to build a two-step approach to classify and predict RLDs. Experimental results showed that their method surpassed traditional algorithms with a recognition rate of 96.8%. However, the images collected in this paper were almost the same scale and free from the disturbance of the backgrounds, which is also different from the in-field conditions. Lu et al.¹⁴ proposed an RLD identification approach by combining sparse automatic coding and stochastic pooling, which achieved a recognition accuracy of 95.48% on a natural data set containing 10 categories of RLD, surpassing several conventional machine learning models. Also, the data set used for evaluation only employed 500 images and is not collected in the natural field conditions, which is not suitable for our research. Liang et al.¹⁵ proposed a CNN-based method for rice blast identification and established a data set of 2906 positive samples and 2902 negative samples. The experimental results showed that the high-level features extracted by the CNN were more discriminative and effective than conventionally manual-crafted features. The data set used in this paper only involved the rice blast disease, and the recognition was the binary classification problem, which is not suitable for our research. Picon et al.¹⁶ proposed to concatenate the contextual information with deep representation for reducing the misclassification of mainstream models in the recognition of crop diseases. Chen et al.¹⁷ studied the transfer learning of a deep cellular neural network for enhanced learning of small lesions symptoms and proposed a new deep learning structure derived from the Visual Geometry Group Network (VGGNet), Inc-VGGN, for the recognition of plant disease images. The results achieved had an accuracy of over 92.00% for the RLD classification. However, the collected data set only contained 500 rice images, which is inadequate to test the generalization of the recognition algorithms. Jiang et al.¹⁸ designed a multitask model for the recognition of three kinds of RLDs and two kinds of wheat leaf diseases. The model took VGG as the backbone and used the ImageNet pretrained weights for transfer learning and alternate learning. Though the RLD classification had reached an accuracy of 98.75%, the rice images were collected in the laboratory environments, which is different from the natural field conditions.

Although the aforementioned studies exploited deep learning in RLD recognition, few of them considered the scale variation in natural fields. Currently, most of the RLD data sets are collected with approximate scale, thus, the corresponding classification researches do not take the scale variation into consideration. However, in the application of intelligent agricultural machinery, the distance between the camera and the surrounding rice leaves varies inevitably, resulting in scale variation for the rice leaves and their spots in the captured images.¹⁹ Under different scales, features extracted by most CNN models demonstrate different salience at the same layer, which will affect the accuracy of disease classification. At present, there are few studies dealing with this problem in the field of classification. However, many studies have been conducted to solve the multiscale problem in the domain of object detection, in a feature fusion framework even with a feature pyramid as detailed below. The Single Shot MultiBox Detector (SSD)²⁰ is the first attempt to embed a feature pyramid into a convolutional network with a bottom-up workflow and combines the feature maps at different layers for region proposal. Leng and Liu²¹ proposed enhancing the conventional SSD by fusing feature maps of different output layers and proposed a visual reasoning method for small object detection. Zhang et al.²² integrated multiscale and feedback features from different layers to better represent objects of various sizes and provide high-level semantic information.

Sehwag et al.²³ proposed an SSD framework based on self-supervised outlier detection, which used self-supervised representation learning followed by Mahalanobis distance-based detection in the feature space. The results showed that the performance of the detector was much better than most existing detectors based on unlabeled data by a large margin. On the basis of SSD,²⁰ Lin et al.²⁴ further proposed to combine low-resolution, semantically strong features with high-resolution, semantically weak features via a top-down pathway and lateral connections. This study is known as a Feature Pyramid Network (FPN) and builds a baseline for multiscale object detection. Hu et al.²⁵ proposed multiscale feature learning, where the multilevel global context and the adjacent levels were fused with content-aware sampling and channelwise reweighting. The proposed model outperformed an FPN with an increase of 2.1% Map. Xu et al.²⁶ extended multiscale detections with an architecture search framework, and the proposed model achieved a 5% improvement over an FPN in terms of the mean Average Precision while reducing network parameters by 50%. Gong et al.²⁷ proposed a fusion factor to be configured in an FPN, which was demonstrated to improve the performance of the baseline in tiny object detection. Liu et al.²⁸ proposed a bottom-up information flow on the basis of an FPN, which was demonstrated to improve the detection capability for multiscale objects. Liang et al.²⁹ proposed the one-shot path aggregation network architecture search, which was effective and efficient in finding the optimal architecture for mainstream detectors. The method showed superiority in localizing significantly smaller objects with reduced searching costs than the neural architecture search-FPN and Auto-FPN methods.

Rice leaves have irregular shapes with different angles under natural field conditions and cannot be easily detected by mainstream object detection approaches. Therefore, the problem has been converted to a classification task, which is in accordance with most related studies.^{17,18} However, scale variation is still a key problem that affects the classification accuracy.¹⁹ Although an extensive study was conducted to address the multiscale issues in deep learning, there are two main limitations in our research scenario. First, most research has focused on object detection, and few have covered the scale issue in the classification domain. Second, the mainstream solutions fused the multiscale representation for decision making, and few of them considered the integration of decision making from the different scales. Also, the existing RLD data sets are collected in the laboratory environments or with a constant scale, which cannot be used to develop the RLD classification algorithms under natural field conditions. To solve these issues, we propose a multiscale voting mechanism (MVM) for RLD recognition, which focuses on the solution to address the negative influence of multiscale factors on the classification accuracy. The major contributions of this paper are highlighted as follows. (1) We introduce a data set that is designed towards the RLD recognition under natural field conditions; (2) we embed a feature pyramid model in a classification network that integrates the multiscale representation to boost the classification accuracy; and (3) we propose an MVM with regard to the probability distribution to achieve more consistent prediction. The objective of this study is to solve the negative influence of different scales on classification accuracy, which is expected to lay a foundation for the automatic recognition of RLD and to provide decision-making information for management machines, such as agricultural robots.

2 | DATA COLLECTION

The data we used was collected in Zhu Village, Zengcheng District, Guangzhou, China. The experimental sites consisted of 26 different rice fields under natural conditions, as illustrated in Figure 1. For ease of operations for future deployment, data collection was conducted using a

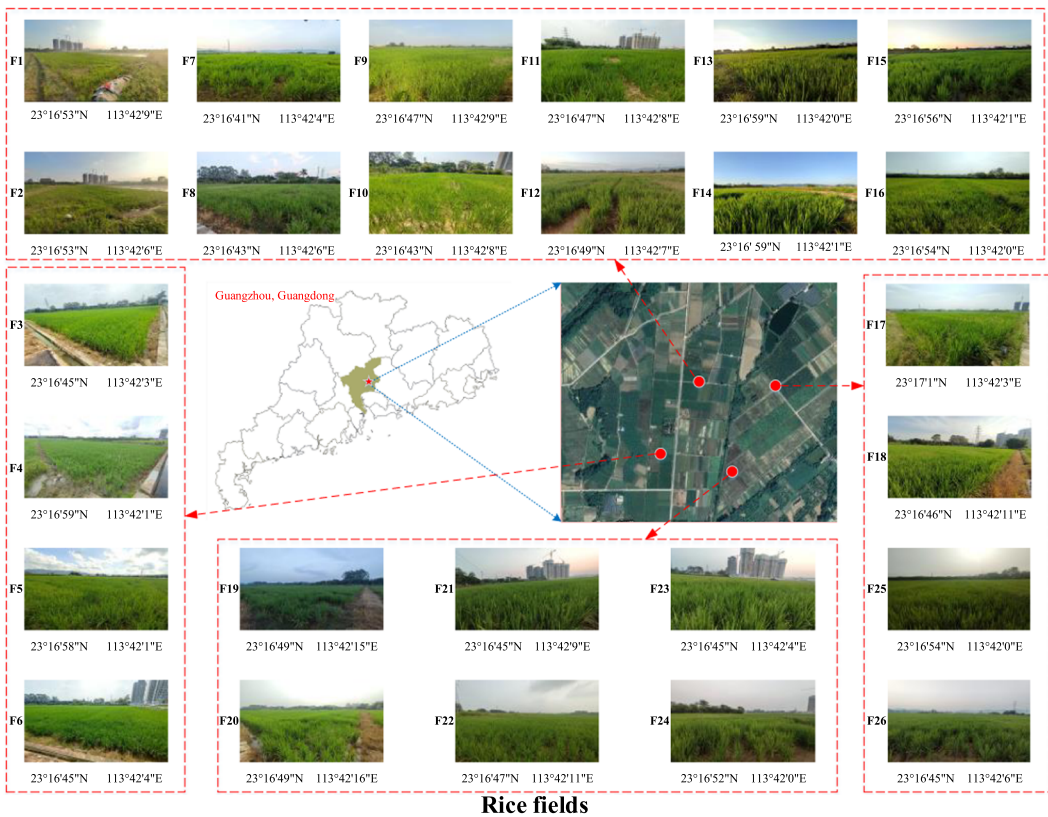


FIGURE 1 General location of the experimental sites. [Color figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

mobile phone (Mi 10s, Xiaomi Corp.). It was equipped with a rear four-camera module, including a 100-megapixel ultra-HD main camera, a 13-megapixel ultrawide angle lens, a 2-megapixel macro lens, and a 2-megapixel depth of field lens. The camera's mode was set to autoexposure and autofocus during data collection, and the lens was 10–50 cm away from the canopy. The resolution of each image taken by the mobile phone was 5792×4344 .

The image was collected from August 21, 2021, to September 30, 2021, and the daily acquisition time periods were 7:00–11:30 a.m. and 16:00–18:30 p.m. The images were collected under different weather conditions, such as sunny, cloudy, and rainy. A total of 6046 rice leaf images were collected, including 1046 *rice bacterial leaf blight* images, 1053 *rice blast* images, 1542 *rice brown spot* images, 823 *rice sheath blight* images, and 1582 *healthy leaf* images. Some examples of RLDs images are shown in Figure 2.

The details of the experimental data are shown in Table 1, namely, RiceDisease5, where (A)–(E) represent the five categories of leaves. The experimental data are divided into a training set and a testing set, as illustrated in Table 2. As seen in Table 2, the samples in the training and testing sets were collected on different dates and fields, this is to ensure the generalization capability of the developed models. Table 3 gives the statistical information on each RLD category, where the samples are shown sufficient for this purpose.

The data set used in our experiment was collected in natural field environments, which is challenging for classification models with various influential factors. One of the main influential factors is spatial scales. Due to the spatial structures of the rice, the distance from

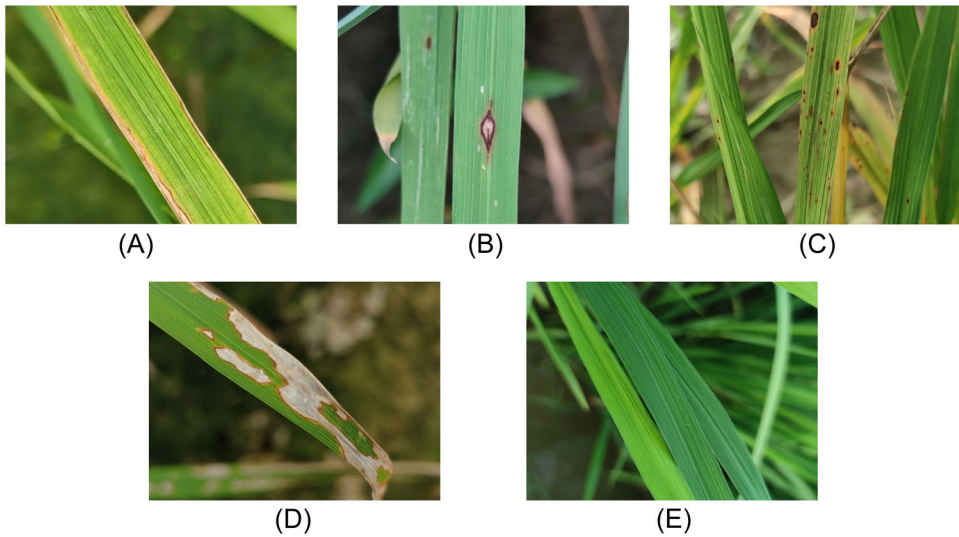


FIGURE 2 In-field photos for each RLD. (A) Rice bacterial leaf blight, (B) rice blast, (C) rice brown spot, (D) rice sheath blight, and (E) healthy rice. RLD, rice leaf disease. [Color figure can be viewed at wileyonlinelibrary.com]

the camera to the rice leaves is inevitably different, leading to noticeable scale variation in the collected images. From the first row and the last two rows in Figure 3, it is obvious that these rice leaves are on different scales. Since most classification models cannot automatically focus on leaf areas of interest, feature extraction on an entire image may involve irrelevant information, which creates significant challenges for recognition. From the third row in Figure 3, it can be seen that the RLD classification under natural field conditions also suffers from other factors, such as illumination variation. However, it was proven that illumination variation produces fewer negative effects on RLD recognition than scale variation. One of the possible reasons is that the deep representation has strengthened the classifier's robustness against the illumination variation. Therefore, this study only focused on the multiscale problem, where the other factors will be left as our future work.

3 | METHODOLOGY

In this study, an MVM was proposed for the automatic identification of RLDs under natural field conditions. The overall framework of our proposed model is shown in Figure 4. Through the comparison of the performance of different network architectures, EfficientNet-b0³⁰ was employed as the backbone, and a feature pyramid with a bottom-up and top-down flow is embedded into the classification model, aiming to reduce the loss of details while preserving the semantic information. Moreover, this feature pyramid can help pay attention to small lesion spots during the feature extraction stage, which is a general drawback for the recognition of early infection of rice diseases. After that, a multiscale voting strategy is adopted to further refine the classification accuracy by reducing the variance from different scales. The details of the backbone network, multiscale fusion embedding, and MVM are introduced in Sections 3.1, 3.2, and 3.3, respectively.

TABLE 1 Details of the experimental data collection

Collection date	Field code	Number of RLD images					Data set name
		A	B	C	D	E	
2021-8-21	F1	1	76	29	0	10	D1
2021-8-22	F2	2	94	18	3	83	D2
2021-8-28	F3	153	16	4	2	41	D3
2021-8-30	F4	0	98	64	46	133	D4
2021-9-5	F5	0	9	331	30	0	D5
2021-9-11	F6	105	12	109	104	1	D6
2021-9-11	F7	5	17	16	4	186	D7
2021-9-11	F8	0	2	0	0	159	D8
2021-9-12	F9	1	5	35	21	222	D9
2021-9-12	F10	150	6	4	0	0	D10
2021-9-13	F11	46	112	54	0	0	D11
2021-9-20	F12	91	59	108	54	61	D12
2021-9-21	F13	16	71	66	32	101	D13
2021-9-25	F14	25	36	48	7	54	D14
2021-9-25	F15	61	22	5	0	38	D15
2021-9-26	F16	0	17	116	0	4	D16
2021-9-26	F17	80	19	31	2	106	D17
2021-9-26	F18	2	4	0	132	1	D18
2021-9-26	F19	6	4	0	55	45	D19
2021-9-27	F20	15	4	20	238	73	D20
2021-9-27	F21	116	84	10	1	46	D21
2021-9-28	F22	8	44	69	78	105	D22
2021-9-28	F23	139	24	14	4	26	D23
2021-9-29	F24	2	28	293	2	0	D24
2021-9-30	F25	2	96	95	0	15	D25
2021-9-30	F26	20	94	3	8	72	D26

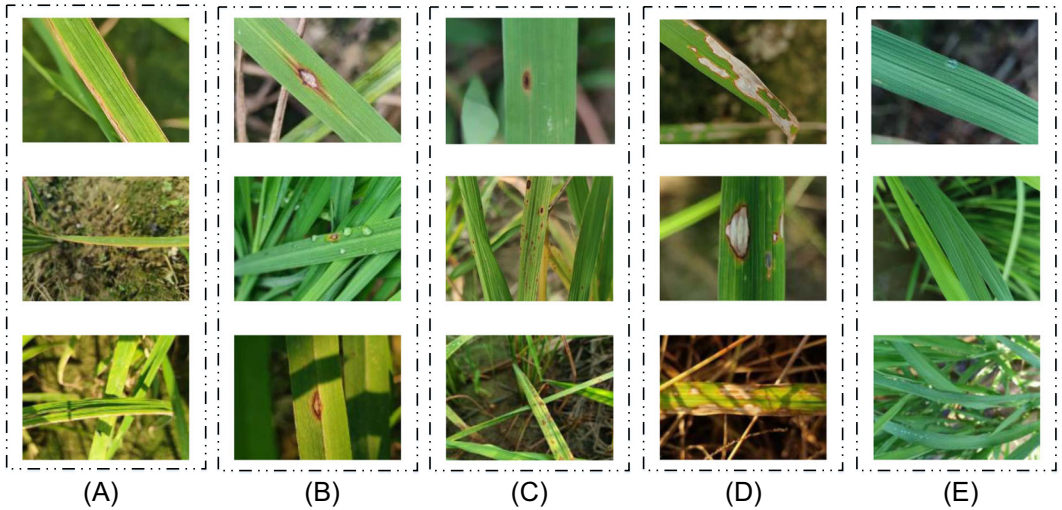
Abbreviation: RLD, rice leaf disease.

TABLE 2 Data split for the training and testing sets

Data set	Data set allocation
Training data set	D1, D2, D3, D4, D5, D6, D7, D10, D12, D14, D15, D17, D20, D22, D25, D26
Test data set	D8, D9, D11, D13, D16, D18, D19, D21, D23, D24

TABLE 3 Detailed information on each category of the data set

Disease	Total	Number of training images	Number of test images
Rice bacterial leaf blight	1046	718	328
Rice blast	1053	702	351
Rice brown spot	1542	954	588
Rice sheath blight	823	576	247
Healthy sample	1582	978	604
Total	6046	3928	2118

**FIGURE 3** Illustration of the scale factor in natural rice fields. (A) Rice bacterial leaf blight, (B) rice blast, (C) rice brown spot, (D) rice sheath blight, and (E) healthy samples. [Color figure can be viewed at wileyonlinelibrary.com]

3.1 | Adaptation to the backbone network

In this study, the EfficientNet-b0 is utilized as the basic model, and its network structure is shown in Figure 5. Different from other network models, EfficientNet uniformly scales all dimensions of depth, width, and resolution using a simple yet highly effective compound coefficient for improved performance. With the difference in input sizes, we modified the final classifier to fit its input vector and output categories. Table 4 lists the detailed parameters of EfficientNet-b0, in which the network was divided into nine stages. The general block structure was as follows. Before the depthwise separable convolution of 3×3 or 5×5 , a 1×1 convolution was used to increase the dimension. After the depthwise separable convolution of 3×3 or 5×5 , an attention mechanism about the channel was added. Finally, a large residual edge was added after dimension reduction by a 1×1 convolution. Among them, compared with conventional convolution operations, deeply separable convolution greatly reduces the number of parameters and operation cost of the network model. A squeeze-and-excitation (SE) attention mechanism allows the model to pay more attention to the channel features with the largest amount of information, which makes it possible to concentrate on the lesion regions and reduces the disturbance from the background.

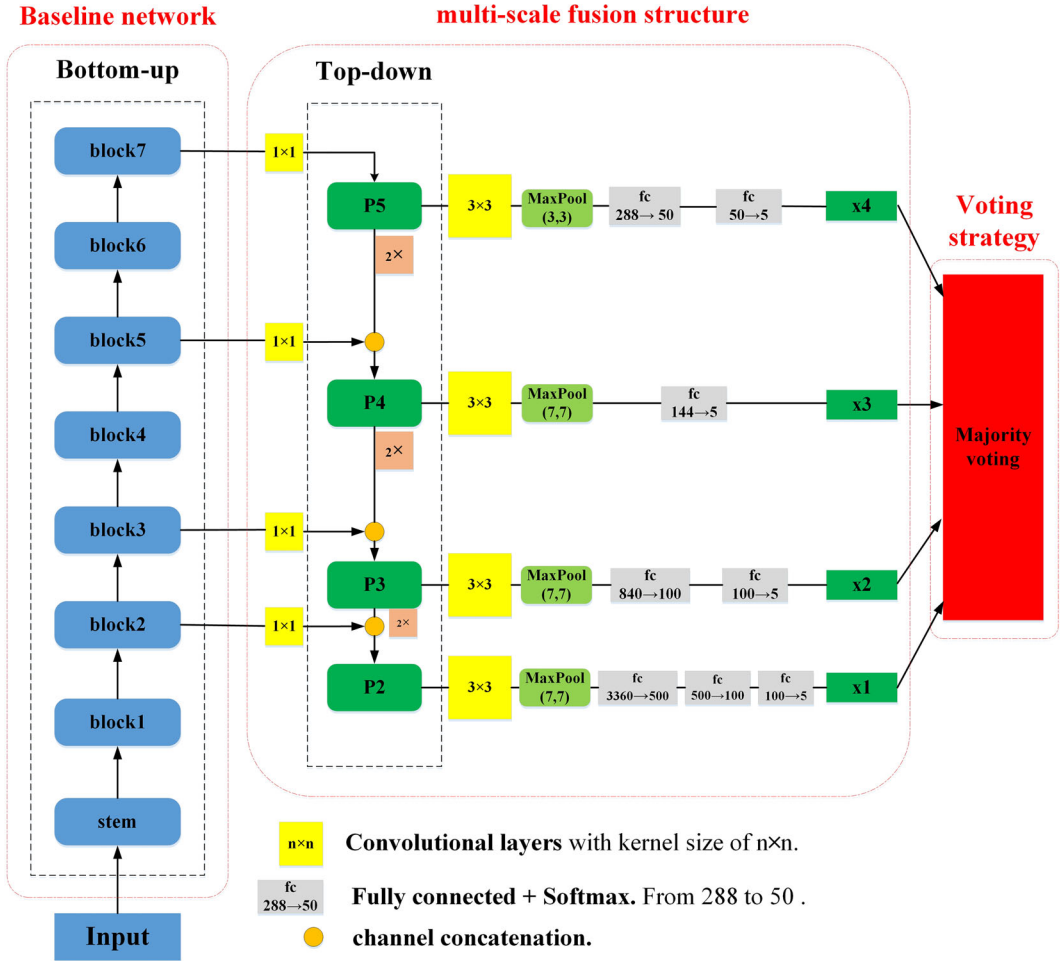


FIGURE 4 General framework of our proposed model. [Color figure can be viewed at wileyonlinelibrary.com]

3.2 | Multiscale embedding for effective classification

The feature pyramid structure has been widely used in the field of object detection, which integrates the semantics of all layers for final decisions of region proposals. Several commonly used feature pyramid models are shown in Figure 6. The SSD²⁰ is one of the first attempts to embed a feature pyramid into a convolutional network with a bottom-up workflow and combines the feature maps at different layers for prediction, as shown in Figure 6A. However, this information flow fails to reuse the high-resolution information at the prediction branch of deep layers. To solve this problem, Faster R-CNN²⁴ proposed another top-down workflow, which seamlessly combines the deep representation with high-resolution feature maps through lateral connection, as shown in Figure 6B.

The rice leaves and their lesions were presented in varying sizes in the collected images, which revealed that the effective representation may be identified in different layers. In this case, we embedded the feature pyramid architecture in the classification models, as shown in Figure 7. From Figure 7, it can be seen that only the first few representations (P_5 , P_4 , P_3 , and P_2)

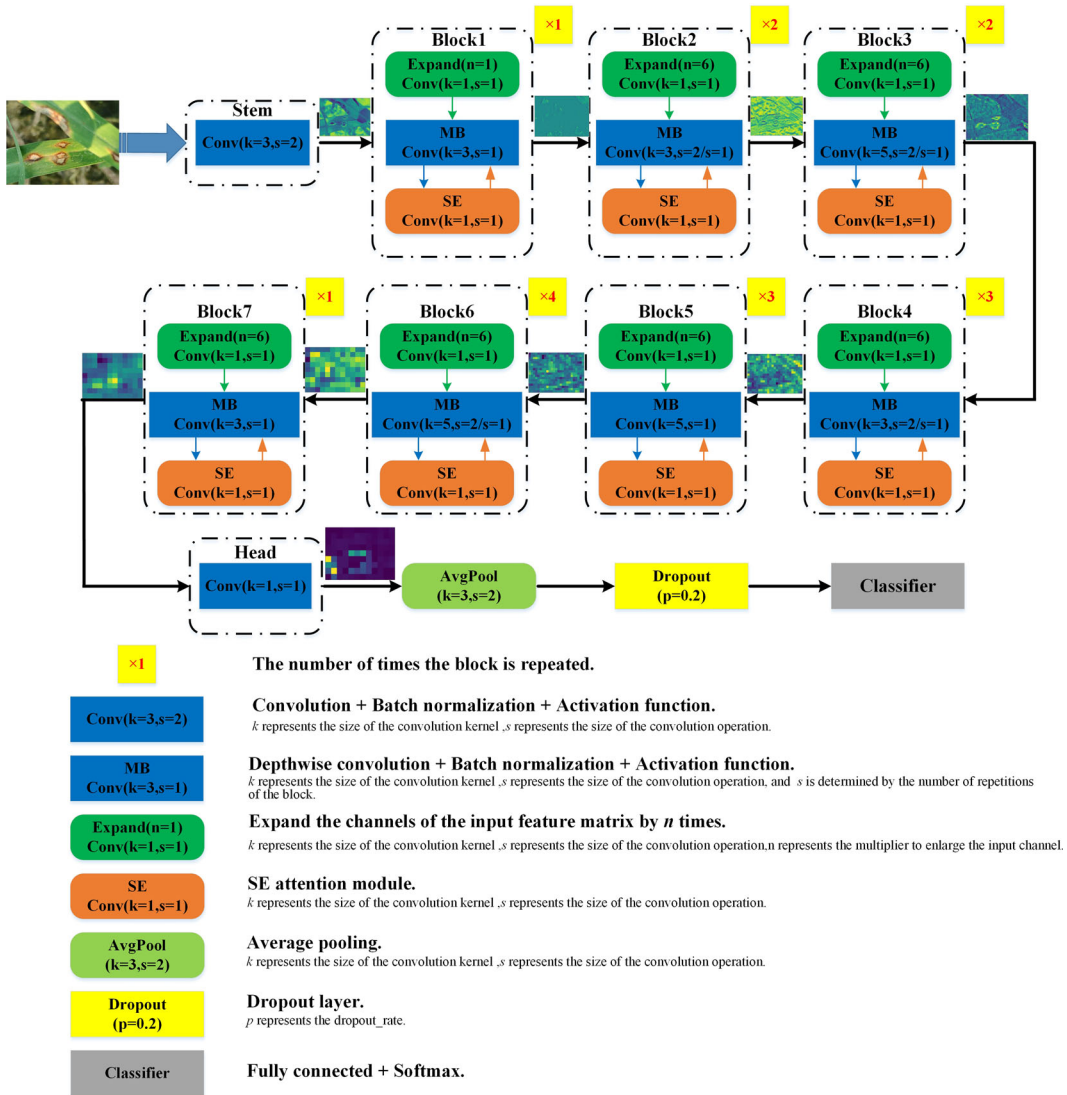
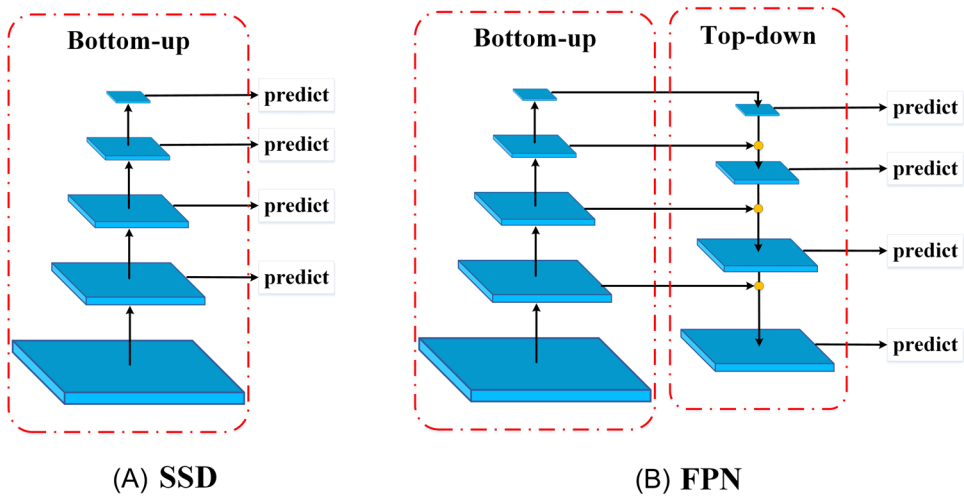


FIGURE 5 Structure of the EfficientNet-b0 model. [Color figure can be viewed at wileyonlinelibrary.com]

in the top-down pathway are adopted, and the last representation (P_1) is ignored. The reason for this is due to the large spatial size in the last representation, which will consume too much memory and computation in the multilayer perceptron during the forward and backward processes. Different from object detection, we did not employ the feature pyramid for the region proposal. Instead, we obtained the probability distribution for all categories in each scale. In addition, we used concatenation for feature fusion instead of elementwise addition, which is adopted by most studies.^{20,24,31} Compared with feature addition, feature concatenation combines the position information of a high-resolution feature map with deep features, which reduces the loss of RLD features caused by the network in the downsampling and avoids the bottleneck of feature representation. Under the backbone of EfficientNet, the feature concatenation does not significantly increase the amount of calculation because of the use of

TABLE 4 Detailed specifications for each layer of the EfficientNet-b0 model

Name	Number of repetitions	Kernel size/stride	Output size
Stem	1	$3 \times 3/2$	$400 \times 300 \times 3$
Block 1	1	$3 \times 3/1$	$200 \times 150 \times 16$
Block 2	2	$3 \times 3/2$ or 1	$100 \times 75 \times 24$
Block 3	2	$5 \times 5/2$ or 1	$50 \times 37 \times 40$
Block 4	3	$3 \times 3/2$ or 1	$25 \times 18 \times 80$
Block 5	3	$5 \times 5/1$	$25 \times 18 \times 112$
Block 6	4	$5 \times 5/2$ or 1	$12 \times 9 \times 192$
Block 7	1	$3 \times 3/1$	$12 \times 9 \times 320$
Head	1	$1 \times 1/1$	$12 \times 9 \times 320$
Fc	—	—	5



● **Element-wise summation**

FIGURE 6 Commonly used feature pyramid models. (A) Single Shot MultiBox Detector (SSD) and (B) Feature Pyramid Network (FPN). [Color figure can be viewed at wileyonlinelibrary.com]

downsampling and depthwise separable convolution. After the softmax function, we sum up the probability distribution for all categories and choose the maximum probability as the predicted value, which can be expressed below:

$$prediction = \underset{j=1,2,\dots,c}{argmax} \left(\sum_{i=1}^s p_{ij} \right), \quad (1)$$

where c denotes the number of categories, s represents the number of scales in the top-down pathway, and p_{ij} refers to the probability of category j predicted by the i th scale.

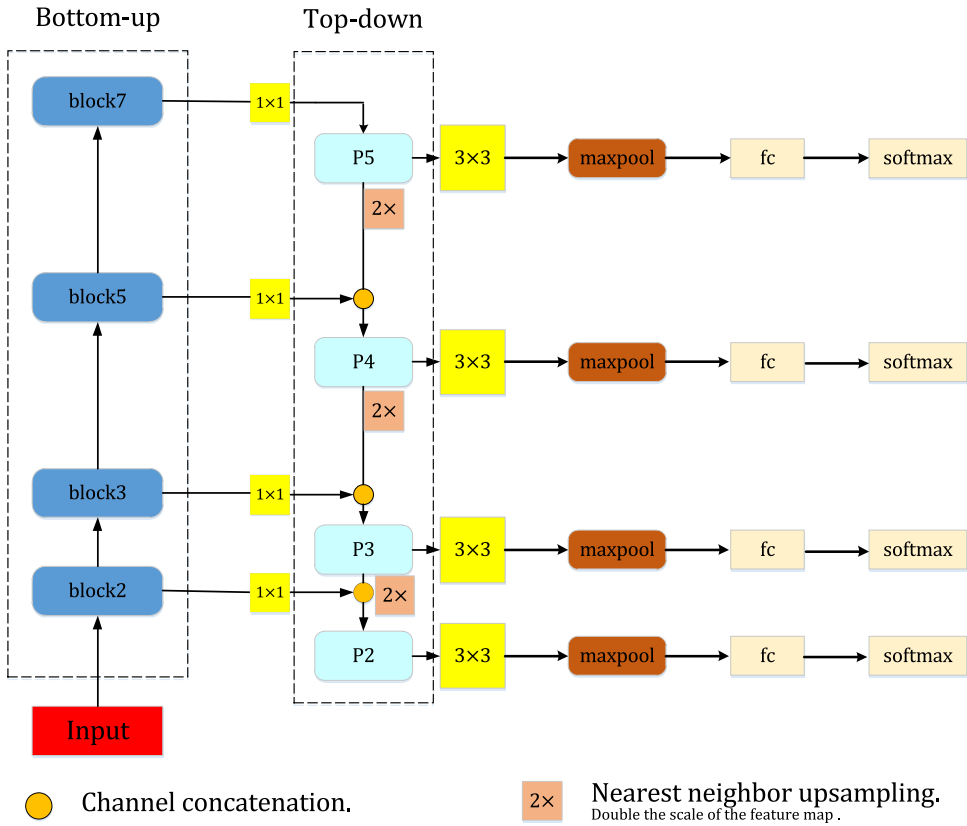


FIGURE 7 Feature pyramid embedded in the classification model [Color figure can be viewed at wileyonlinelibrary.com]

3.3 | Multiscale voting mechanism

With the hypothesis that the RLD can be detected at each scale with a reasonable accuracy, a majority voting mechanism is adopted to reduce the bias of each scale for more consistent classification. Therefore, we propose a voting strategy that is built on a multiscale architecture with regard to the probability distribution, as shown in Figure 8. From Figure 8, each feature level represents one decision branch, and its prediction is adopted for voting. In the voting process, the relative majority principle was used, and the candidate with the most votes is adopted as the final label. However, the counting process can be infeasible when there is more than one winning candidate with the same votes. To solve this problem, we propose three different strategies with regard to the probability distribution output by all scales: (1) *Voting with probability summation*: The probability for each category was added across all scales, and the class with the maximum probability was selected as the final result. (2) *Voting by finest level*: Inspired by the fact that the highest-resolution, strongly semantic feature maps of P_2 in the top-down pathway are most effective for prediction, its output was adopted as the final label when there is a disagreement between different scales. (3) *Voting with the largest probability*: The probabilities for all categories from all scales are compared, and the class with the largest probability value is used as the final label.

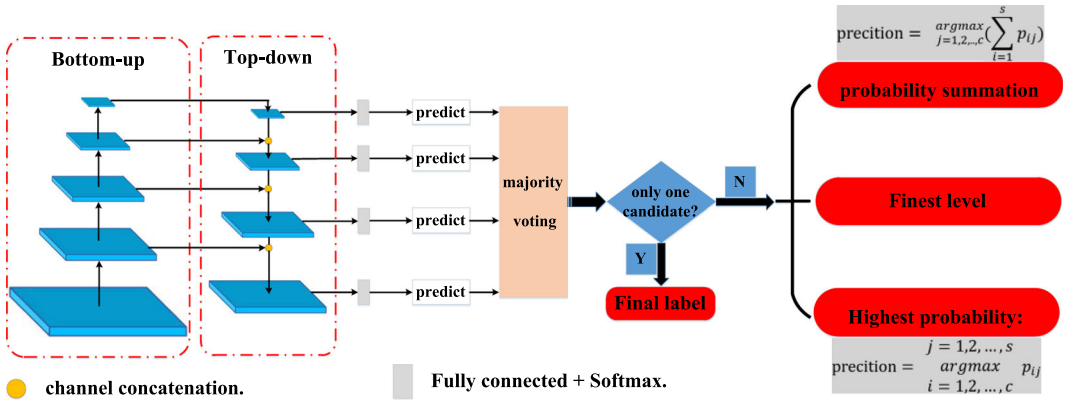


FIGURE 8 Illustration of the multiscale voting mechanism [Color figure can be viewed at wileyonlinelibrary.com]

4 | RESULTS AND DISCUSSIONS

The modeling and experimental implementation were conducted on an Intel Xeon Silver 4210R processor $\times 2$, an NVIDIA RTX3090 24 GB graphics card $\times 4$, and 32 GB memory $\times 8$. In this study, precision, recall, F1 score, and accuracy were adopted as the metrics for quantitative evaluation.

4.1 | Implementation details

The original resolution of the collected images was 5792×4344 (4:3), which will easily cause graphics processing unit exhaustion during the forward and backward processes. Therefore, the image size is normalized to $400 \times 300 \times 3$, which was validated to preserve the details of the rice leaves and their lesions. The EfficientNet-b0 network was divided into a total of nine stages. Each stem block had a convolutional layer with two subsequent steps (containing batch normalization [BN] and the Swish activation function). Block 1–Block 7 are repeatedly stacked mobile inverted bottleneck convolution (MBConv) structures, as shown in Figure 9. As presented in Figure 9, the MBConv structure consists of a 1×1 convolution (with BN and Swish), a 3×3 or 5×5 depthwise convolution (with BN and Swish), an SE module, a 1×1 convolution (with BN) and a dropout layer. The general design idea was an inverted residual structure and residual structure. First, a 1×1 convolution was used to raise the dimension before the 3×3 or 5×5 network structure, then an attention mechanism for the channels was added after the 3×3 or 5×5 network structure, and finally, a large residual edge was added after dimension reduction by a 1×1 convolution. BN is used after each convolution and before activation. The initial learning rate was 0.01, and we transferred the ImageNet pretrained weights and fine-tuned our data set. The batch size was 180, and the number of epochs was 500. The stochastic gradient descent optimizer was used to optimize the model, and a weight decay of 0.001 and the cross entropy were used as the loss function.

4.2 | Comparison between the backbone networks

In this section, four representative backbone networks were selected to compare their performance on our RiceDisease5 data set. The selected backbone networks included

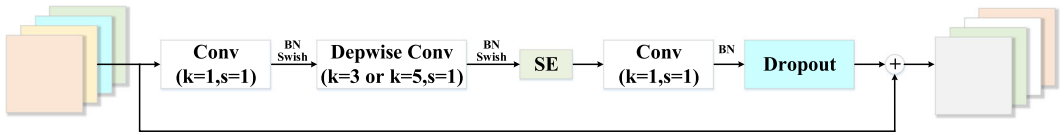


FIGURE 9 Structure of the MBConv module [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 5 Recognition results of various CNN models

Backbone	EfficientNet-b0	GoogLeNet-V3	DenseNet-121	ResNet-50
Accuracy (%)	85.79	84.84	86.07	85.55
Test time (ms)	3.91	4.46	4.33	4.70
Params (M)	15.60	23.83	7.98	25.56
FLOPs (B)	0.47	7.75	6.74	10.10

Note: Bold values indicate the best results.

Abbreviation: CNN, convolutional neural network.

EfficientNet-b0,³⁰ GoogLeNet-V3,³² ResNet-50,³³ and DenseNet-121,³⁴ and the experimental results are shown in Table 5. As seen in Table 5, DenseNet-121 achieves the highest accuracy rate of 86.07%, which is slightly better than that of EfficientNet-b0. However, the computational cost of DenseNet-121 is too heavy because of the dense connection, and the floating-point operations per second (FLOPs) of the model are up to 6.74 billion. In comparison, the FLOPs of the EfficientNet-b0 is only 0.47 billion, which may reduce much computational cost. From the computational complexity, EfficientNet generally uses an order of magnitude fewer FLOPs than other baseline models. We conjecture that the proposed compound scaling method leads to better architectures, better scaling, and better training settings. As a result, the EfficientNet-b0 model strikes the best trade-off between accuracy and speed. Therefore, we applied the EfficientNet-b0 as our backbone structure for the following experiments.

4.3 | Ablation study

4.3.1 | Effect of multiscale embedding

As shown in Section 3.2, we embed a feature pyramid into the EfficientNet-b0 with a top-down pathway and a lateral connection to different layers of the feature extractors. Table 6 shows that the embedding of multiscale architecture can greatly improve the prediction accuracy in all metrics, and the performance boosts are obvious in each category. Specifically, the overall accuracy for all categories was boosted by a margin of 3.45%. For the efficiency, the embedding of multiscale architecture consumes more than 0.14 ms per image, where the extra computation comes from the feature fusion. Generally, the classification accuracy for the *Blast* disease category is relatively low with and without multiscale architecture. One possible reason is that the symptoms of *Blast* disease are similar to the background, which makes it difficult to distinguish it from other categories. However, the embedding of the multiscale architecture still raised the precision by 8.7%.

TABLE 6 Recognition results of various multiscale models

Model	Rice diseases	Precision	Recall	F1 score	Accuracy (%)	Test time (ms)
Backbone	Bacterial leaf blight	0.895	0.912	0.903	85.79	3.91
	Blast	0.727	0.621	0.670		
	Brown spot	0.822	0.830	0.826		
	Sheath blight	0.819	0.879	0.848		
	Health	0.950	0.983	0.966		
With multiscale embedding	Bacterial leaf blight	0.929	0.954	0.941	89.24	4.05
	Blast	0.814	0.672	0.736		
	Brown spot	0.858	0.884	0.871		
	Sheath blight	0.851	0.903	0.876		
	Health	0.960	0.990	0.975		

Note: Bold values indicate the best results.

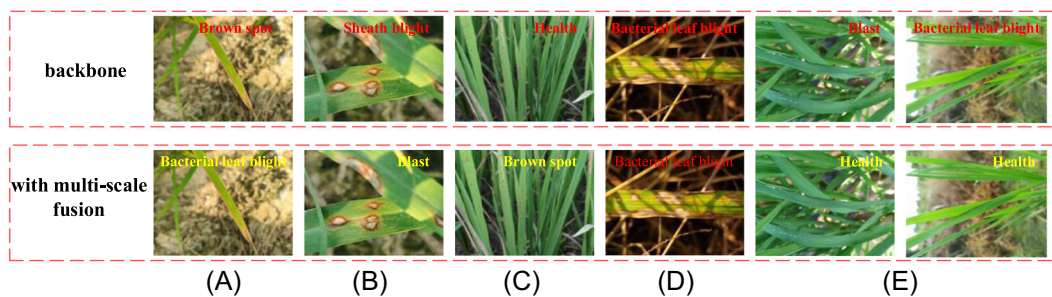


FIGURE 10 Some classification results with and without the multiscale architecture (among them, the red words represent misclassified samples, and the yellow words indicate correctly identified samples). (A) Bacterial leaf blight, (B) blast, (C) brown spot, (D) sheath blight, and (E) health. [Color figure can be viewed at wileyonlinelibrary.com]

Figure 10 presents some cases at different scales, including the lesions that are not obvious at a large distance between the camera and the rice, leading to the most incorrect results under the backbone network. In contrast, by introducing the embedding of multiscale architecture to well address the scale problem, our pyramid representation has greatly improved the robustness of classification to the variation of object scales. Figure 10D also gives one failed example of our approach, where the network misclassified the *Sheath blight* as *Bacterial leaf blight*. The possible reason for this failure is that the rice leaf is under severe uneven illumination conditions, which makes it difficult to extract the effective representation for *Sheath blight*. However, it was proven that illumination variation produces less negative effects on RLD recognition than scale variation. We argue that the robustness brought by the deep representation has strengthened the classifier's capability against the illumination problem. Therefore, this study only focused on the multiscale problem and did not consider the illumination variation in the natural conditions, which will be part of our future work.

TABLE 7 Performance of different design choices in the multiscale architecture

Channel reduction	Maxpooling	Avg-pooling	Dropout	Accuracy (%)	Test time (ms)
×	✓	×	×	87.20	13.59
✓	✓	×	×	89.24	4.05
✓	×	✓	×	88.67	4.14
✓	✓	×	✓	88.62	4.10

Note: Bold values indicate the best results.

Different choices of network architectures with the embedded feature pyramid may affect the final recognition performance, such as channel reduction, pool layer type, and dropout layer. Therefore, we conducted further experiments on different design choices of multiscale architecture, and the modules were tested as the feedforward process of the network. Table 7 shows performance comparisons for different architectural combinations. As shown in the first two rows of Table 7, appending channel reduction after feature concatenation increases the accuracy by 2.04% over the version without channel reduction. We conjecture that feature concatenation without channel reduction has too much redundant information, which increases the difficulty in automatic feature extraction and decreases the performance. Additionally, the maxpooling operation slightly outperforms the average pooling, perhaps because maxpooling can better retain texture features. Finally, the dropout strategy used in the fully connected layers also decreases the prediction accuracy. One of the possible reasons for this result is that the dropout layer loses some information in the representation, which causes more prediction errors. Overall, the design choices with channel reduction, maxpooling, and dropout strategy consistently outperform others in terms of accuracy and efficiency.

4.3.2 | Ablation study on the MVM

In this section, we propose the majority voting with regard to probability distribution to integrate decisions from different scales. To address the problem with more than one winning candidate, three voting strategies were proposed concerning the probability distribution: voting with the probability summation, voting with the finest level (P_2), and voting with the maximum probability. Table 8 shows that voting with probability summation and the finest level can both increase the accuracy, while voting with the maximum probability slightly decreases the performance. We argue that voting with the maximum probability ignores most of the decisions from different scales, leading to increased variance in the classifier. Overall, voting with the finest level achieves the best performance in accuracy boost thanks to the highest-resolution, strongly semantic feature maps of P_2 . For the network efficiency, Table 8 shows that the voting strategies have slightly improved the inference speed. We conjecture that most of the decisions can be made through majority voting, which can actually save the probability summation of different scales for decision making. Generally, the MVM with different probability strategies can effectively improve the performance of the baseline in terms of accuracy and efficiency.

Figure 11 shows example images with and without the multiscale mechanism. As seen, different voting strategies can correct the misclassification of the baseline network to some extent. Overall, the strategy of voting with the finest level can obtain the correct result in most cases. However, our proposed voting strategy still cannot effectively distinguish between the

TABLE 8 Ablation study on the multiscale voting mechanism

Model	Rice diseases	Precision	Recall	F1 score	Accuracy (%)	Test time (ms)
Without multiscale voting mechanism	Bacterial leaf blight	0.929	0.954	0.941	89.24	4.05
	Blast	0.814	0.672	0.736		
	Brown spot	0.858	0.884	0.871		
	Sheath blight	0.851	0.903	0.876		
	Health	0.960	0.990	0.975		
Voting with probability summation	Bacterial leaf blight	0.955	0.963	0.959	89.66	3.74
	Blast	0.793	0.687	0.736		
	Brown spot	0.861	0.883	0.872		
	Sheath blight	0.857	0.899	0.877		
	Health	0.968	0.995	0.981		
Voting with the finest level	Bacterial leaf blight	0.952	0.960	0.956	90.27	3.69
	Blast	0.817	0.698	0.753		
	Brown spot	0.859	0.891	0.875		
	Sheath blight	0.891	0.923	0.907		
	Health	0.966	0.993	0.979		
Voting with the maximum probability	Bacterial leaf blight	0.946	0.960	0.953	88.67	3.60
	Blast	0.804	0.632	0.708		
	Brown spot	0.842	0.879	0.860		
	Sheath blight	0.849	0.911	0.879		
	Health	0.951	0.992	0.971		

Note: Bold values indicate the best results.

Bacterial leaf blight and the *Healthy* category. We conjecture that healthy rice leaves at different angles present different illumination reflections, which is similar to the symptoms of *Bacterial leaf blight* disease, making it difficult to classify in between.

Finally, we compare the confusion matrix of the backbone network and the improved version. As shown in Figure 12, our proposed model improves the accuracy in all categories. Thanks to the proposed multiscale embedding and the MVM, the overall accuracy has achieved 90.27%, significantly outperforming the EfficientNet-b0 baseline by 4.8%. We conjecture that the appended top-down pathway exploits the potential for different scales, and the MVM integrates the decision information from different scales for further accuracy boost.

4.4 | Comparison with state of the art

In this section, we compare the performance of our proposed approach with several state-of-the-art multiscale frameworks, including SSD,²⁰ FPN,²⁴ Path Aggregation Network (PANet),²⁸

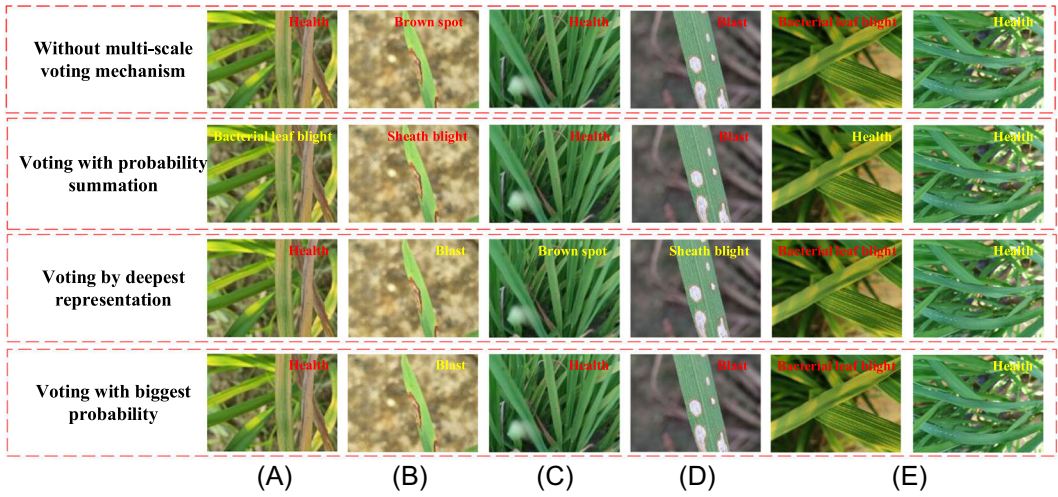


FIGURE 11 Multiscale voting with different probability strategies (among them, the red words represent misclassified samples, and the yellow words indicate correctly identified samples). (A) Bacterial leaf blight, (B) blast, (C) brown spot, (D) sheath blight, and (E) health. [Color figure can be viewed at wileyonlinelibrary.com]

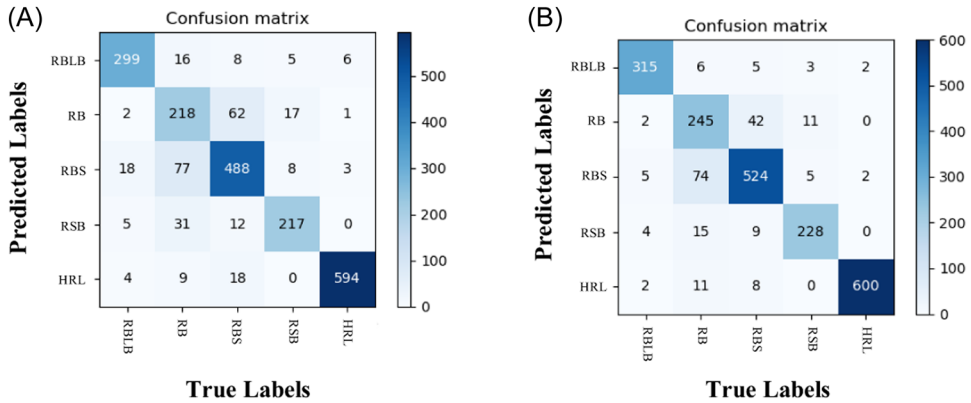


FIGURE 12 Confusion matrix of the backbone network and the improved version. (A) EfficientNet-b0 and (B) our proposed model. [Color figure can be viewed at wileyonlinelibrary.com]

and Bidirectional Feature Pyramid Network (BiFPN),³¹ which are not architecture specific and can be used for both target detection and classification. In this section, these multiscale strategies were embedded in the classification architecture for comparison. For a fair evaluation, all these counterparts used the same backbone network (EfficientNet-b0) and the same hyperparameters as ours. Table 9 shows that our proposed method achieves the best accuracy for most of the categories. On the test set of this study, our method increases over the existing best results by 1.08% of the overall accuracy (90.27% vs. 89.19%). It can be seen that our method outperforms the classical FPN with a 1.41% increase in accuracy. We argue that the FPN only focused on feature-level fusion, while our approach considered the integration of decision information from different scales, which led to a performance boost. Additionally, it is

TABLE 9 Comparison of our proposed method with state-of-the-art methods

Model	Rice diseases	Precision	Recall	F1 score	Accuracy (%)	Test time (ms)
SSD ²⁰	Bacterial leaf blight	0.878	0.918	0.898	86.31	4.27
	Blast	0.745	0.667	0.704		
	Brown spot	0.858	0.820	0.839		
	Sheath blight	0.813	0.899	0.854		
	Health	0.941	0.975	0.958		
FPN ²⁴	Bacterial leaf blight	0.932	0.954	0.943	88.86	4.02
	Blast	0.779	0.692	0.733		
	Brown spot	0.880	0.861	0.870		
	Sheath blight	0.824	0.907	0.864		
	Health	0.957	0.987	0.972		
PANet ²⁸	Bacterial leaf blight	0.939	0.933	0.936	89.05	4.03
	Blast	0.785	0.729	0.756		
	Brown spot	0.857	0.874	0.865		
	Sheath blight	0.862	0.862	0.862		
	Health	0.964	0.988	0.976		
BiFPN ³¹	Bacterial leaf blight	0.939	0.942	0.940	89.19	4.06
	Blast	0.807	0.704	0.752		
	Brown spot	0.869	0.872	0.870		
	Sheath blight	0.824	0.907	0.864		
	Health	0.960	0.987	0.973		
Ours	Bacterial leaf blight	0.952	0.960	0.956	90.27	3.69
	Blast	0.817	0.698	0.753		
	Brown spot	0.859	0.891	0.875		
	Sheath blight	0.891	0.923	0.907		
	Health	0.966	0.993	0.979		

Note: Bold values indicate the best results.

Abbreviations: BiFPN, Bidirectional Feature Pyramid Network; FPN, Feature Pyramid Network; PANet, Path Aggregation Network; SSD, Single Shot MultiBox Detector.

worth noting that the accuracy of our proposed model is 1.22% higher than that of PANet and 1.08% higher than that of BiFPN. We conjecture that PANet and BiFPN add an extra bottom-up pyramid on the basis of FPN to convey location information, but this structure has little influence on classification. Moreover, further information transmission may introduce redundant information, which adversely affects recognition accuracy. For efficiency, our approach outperformed others with a notable margin. Due to the voting strategy built on the multiscale architecture and probability distribution, most samples can be directly identified through majority voting, without the need to sum up the probability from different scales for the final decision.

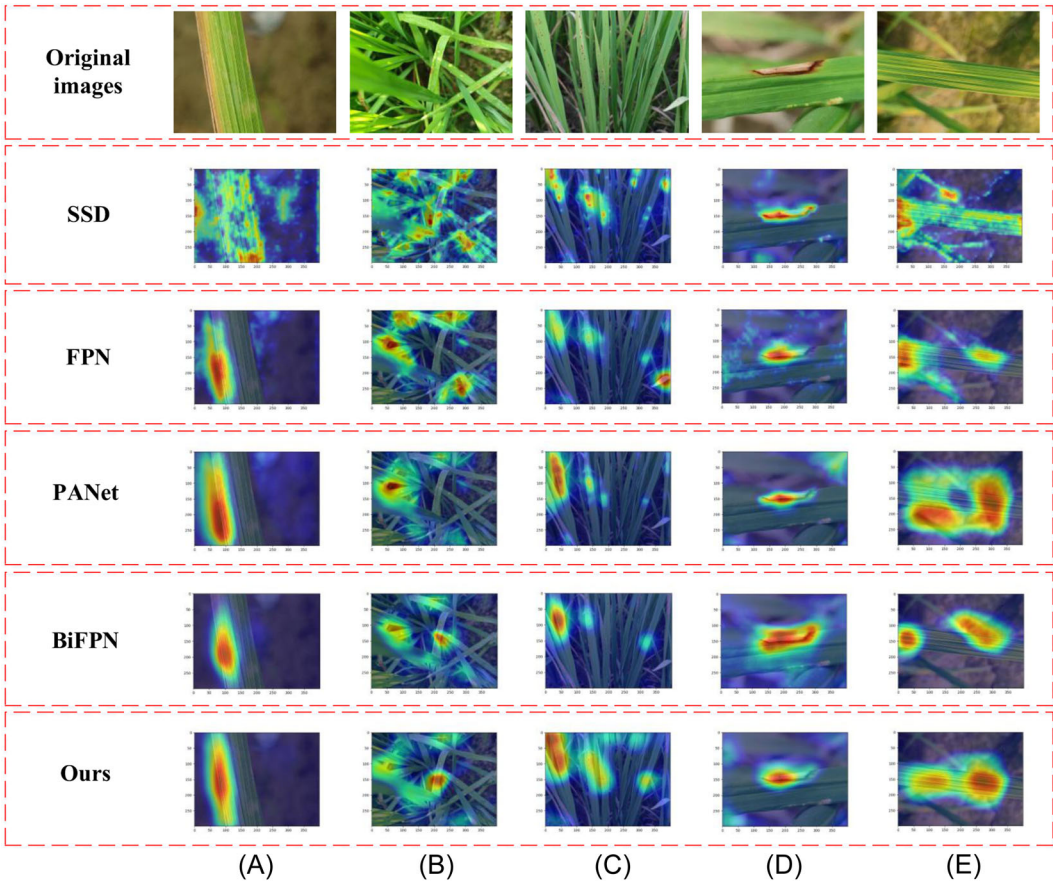


FIGURE 13 Gradient-weighted class activation maps generated using different models. (A) Bacterial leaf blight, (B) blast, (C) brown spot, (D) sheath blight, and (E) health. BiFPN, Bidirectional Feature Pyramid Network; FPN, Feature Pyramid Network; PANet, Path Aggregation Network; SSD, Single Shot MultiBox Detector. [Color figure can be viewed at wileyonlinelibrary.com]

We applied gradient-weighted class activation mapping to produce a visual explanation for decisions from different models. The localization map highlights the important regions in the image, reflecting each model's capability to capture and concentrate on the discriminative regions for classification. Figure 13A,D shows that most models can effectively localize the discriminative regions when the rice leaves are at close range and the background is simple. However, when the rice leaves are distant and the lesions are small, most mainstream models fail to localize the effective regions, as shown in Figure 13B,C. In contrast, our proposed multiscale architecture accurately localizes the discriminative regions despite the disturbance from the complex backgrounds. Additionally, Figure 13E shows that our method can better localize the rice leaves for the health class compared with other counterparts. Overall, our approach can better localize the discriminative areas when the regions of interest are small and the background is complex, which provides a visualization explanation for the accuracy boost brought by our proposed modules.

5 | CONCLUSIONS

In this paper, we proposed an MVM for RLD recognition under natural field conditions. First, we collected data from over 20 rice fields and built a data set containing 6046 images of different RLDs under natural field conditions. After that, we embedded a feature pyramid into a mainstream classification architecture with a bottom-up and top-down pathway to fuse the representations from different scales. Later, we proposed multiscale voting with regard to probability distribution to integrate the decision information from different scales. During the experimental process, each proposed module was carefully validated through an ablation study to demonstrate its effectiveness, and the proposed method was compared with state-of-the-art methods, including SSD, FPN, PANet, and BiFPN. Qualitative and quantitative results showed that the proposed modules can effectively increase the model's robustness for scale variation, which improved the accuracy of the baseline by 4.48%. Additionally, the comparison results showed that our proposed method outperformed other state-of-the-art algorithms in terms of accuracy and efficiency. However, according to the experimental results, the network model proposed in this paper has difficulty solving other influential factors under natural field conditions, such as varying illumination and water droplets, which is the direction of our future work.

AUTHOR CONTRIBUTIONS

Yu Tang: Conceptualization, project administration, methodology, and writing—original draft preparation. **Jinfei Zhao:** Investigation, methodology, and software. **Huasheng Huang:** Investigation, software, and writing—reviewing and editing. **Jiajun Zhuang:** Methodology and writing—reviewing and editing. **Zhiping Tan:** Methodology and software. **Chaojun Hou:** Methodology and writing—reviewing and editing. **Weizhao Chen:** Methodology and software. **Jinchang Ren:** Conceptualization, methodology, and writing—reviewing and editing.

ACKNOWLEDGMENTS

The authors acknowledge support from the National Natural Science Foundation of China (Grant No. 32071895), the Planned Science and Technology Project of Guangdong Province, China (Grant Nos. 2019B020216001, 2019A050510045, and 2021A0505030075), the Natural Science Foundation of Guangdong Province, China (Grant Nos. 2020B1515120070 and 2021A1515010824), the Key Project of Universities in Guangdong Province, China (Grant No. 2020ZDZX1061), the Innovation Team Project of Universities in Guangdong Province, China (Grant No. 2021KCXTD010), the Planned Science and Technology Project of Guangzhou, China (Grant Nos. 202002020063, 202007040007, and 202103000028), the Basic and Applied Basic Research Fund in Guangdong Province, China (Grant No. 2021A1515110756), the Project of Educational Commission of Guangdong Province, China (Grant No. 2021KQNCX044), and the Rural Revitalization Strategy Project of Guangdong Province, China (Grant No. 2019KJ138).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available.

ORCID

Huasheng Huang  <http://orcid.org/0000-0002-6546-6501>

REFERENCES

1. Chen J, Zhang D, Nanehkaran YA, Li D. Detection of rice plant diseases based on deep transfer learning. *J Sci Food Agric*. 2020;100(7):3246-3256.
2. Shen W, Guan Y, Wang Y, Jing D. Research on rice leaf disease recognition based on BP neural network. *J Northeast Agric Univ (English Ed)*. 2019;26(3):75-86.
3. Bera T, Das A, Sil J, Das AK. A survey on rice plant disease identification using image processing and data mining techniques. In: *Emerging Technologies in Data Mining and Information Security*. Springer; 2019: 365-376.
4. Bai X, Cao Z, Zhao L, et al. Rice heading stage automatic observation by multi-classifier cascade based rice spike detection method. *Agric Forest Meteorol*. 2018;259:260-270.
5. Islam A, Islam R, Rafizul Haque SM, Mohidul Islam SM. Rice leaf disease recognition using local threshold based segmentation and deep CNN. *Int J Intell Syst Appl*. 2021;13(5):35-45.
6. Zhou G, Zhang W, Chen A, He M, Ma X. Rapid detection of rice disease based on FCM-KM and Faster R-CNN fusion. *IEEE Access*. 2019;7:143190-143206.
7. Maeda-Gutiérrez V, Galvan-Tejada CE, Zanella-Calzada LA, et al. Comparison of convolutional neural network architectures for classification of tomato plant diseases. *Appl Sci*. 2020;10(4):1245.
8. Guo Y, Zhang J, Yin C, et al. Plant disease identification based on deep learning algorithm in smart farming. *Discrete Dyn Nat Soc*. 2020;2020:2479172.
9. Howlader MR, Habiba U, Faisal RH, Rahman MM. Automatic recognition of guava leaf diseases using deep convolution neural network. In: *International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE; 2019.
10. Barbedo JGA. A review on the main challenges in automatic plant disease identification based on visible range images. *Biosyst Eng*. 2016;144:52-60.
11. Xiao M, Ma Y, Feng Z, et al. Rice blast recognition based on principal component analysis and neural network. *Comput Electron Agric*. 2018;154:482-490.
12. Sethy PK, Barpanda NK, Rath AK, Behera SK. Deep feature based rice leaf disease identification using support vector machine. *Comput Electron Agric*. 2020;175:105527.
13. Jiang F, Lu Y, Chen Y, Cai D, Li G. Image recognition of four rice leaf diseases based on deep learning and support vector machine. *Comput Electron Agric*. 2020;179:105824.
14. Lu Y, Yi S, Zeng N, Liu Y, Zhang Y. Identification of rice diseases using deep convolutional neural networks. *Neurocomputing*. 2017;267:378-384.
15. Liang W, Zhang H, Zhang G, Cao H. Rice blast disease recognition using a deep convolutional neural network. *Sci Rep*. 2019;9(1):1-10.
16. Picon A, Seitz M, Alvarez-Gila A, Mohnke P, Ortiz-Barredo A, Echazarra J. Crop conditional convolutional neural networks for massive multi-crop plant disease classification over cell phone acquired images taken on real field conditions. *Comput Electron Agric*. 2019;167:105093.
17. Chen J, Chen J, Zhang D, Sun Y, Nanehkaran YA. Using deep transfer learning for image-based plant disease identification. *Comput Electron Agric*. 2020;173:105393.
18. Jiang Z, Dong Z, Jiang W, Yang Y. Recognition of rice leaf diseases and wheat leaf diseases based on multi-task deep transfer learning. *Comput Electron Agric*. 2021;186:106184.
19. Fan X, Luo P, Mu Y, Zhou R, Tjahjadi T, Ren Y. Leaf image based plant disease identification using transfer learning and feature fusion. *Comput Electron Agric*. 2022;196:106892.
20. Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: *European Conference on Computer Vision*. Springer; 2016.
21. Leng J, Liu Y. An enhanced SSD with feature fusion and visual reasoning for object detection. *Neural Comput Appl*. 2019;31(10):6549-6558.
22. Zhang H, Tian Y, Wang K, Zhang W, Wang F. Mask SSD: an effective single-stage approach to object instance segmentation. *IEEE Trans Image Process*. 2019;29:2078-2093.

23. Sehwal V, Chiang M, Mittal P. SSD: a unified framework for self-supervised outlier detection. arXiv preprint, arXiv:2103.12051. 2021.
24. Lin T, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2017.
25. Hu M, Li Y, Fang L, Wang S. A2-FPN: attention aggregation based feature pyramid network for instance segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2021.
26. Xu H, Yao L, Zhang W, Liang X, Li Z. Auto-FPN: automatic network architecture adaptation for object detection beyond classification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*; 2019.
27. Gong Y, Yu X, Ding Y, Peng X, Zhao J, Han Z. Effective fusion factor in FPN for tiny object detection. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*; 2021.
28. Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2018.
29. Liang T, Wang Y, Tang Z, Hu G, Ling H. OPANAS: one-shot path aggregation network architecture search for object detection. arXiv e-prints, arXiv:2103.04507. 2021.
30. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning*. PMLR; 2019.
31. Tan M, Pang R, Le QV. EfficientDet: scalable and efficient object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2020.
32. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016.
33. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016.
34. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2017.