IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 August 2023, accepted 8 September 2023, date of publication 12 September 2023,
date of current version 21 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3314676

RESEARCH ARTICLE

Improvements Based on ShuffleNetV2 Model for Bird Identification

LIU-LEI ZHANG, YING JIANG, YOU-PENG SUN, YUAN ZHANG, AND ZHENG WANG
College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China

Corresponding author: Ying Jiang (jiangying0510@sina.com)

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Bird identification and classification are of great significance in bird conservation. Through proper bird classification, the changes in bird populations in a given area can be effectively predicted, thereby ensuring their effective protection. Nowadays, deep learning has achieved high accuracy in classifying bird images. However, most existing models suffer from poor generalization ability and high complexity. To enhance the model's generalization ability, this paper constructed a large dataset consisting of 275 bird species and the dataset contains bird pictures in different situations, such as rainy days, foggy days. To reduce model complexity, a lightweight neural network based on ShuffleNetV2 was constructed. In ShuffleNetV2 network, there is no feature fusion module and efficient attention mechanism to assist the feature learning of the model. Therefore, this paper adds a feature fusion module and two attention mechanisms to make up for this shortcoming. A multi-channel feature fusion structure (MCF) was adopted to improve the network's adaptability to extract information from multiple channel scales. By introducing Squeeze-and-Excitation (SE) Module and Coordinate attention (CA) in the Block module, the model's ability to refine the global features was enhanced. The experimental results show that the accuracy of the model in identifying 275 bird species on the self-built dataset is 92.3%, which is 6.1% higher than the accuracy of ShuffleNetV1 (86.2%) and 1.8% higher than the accuracy of ShuffleNetV2 (90.5%). At the same time, with smaller parameters and floating point operations (FLOPs), its accuracy is 1.2% higher than ResNet50's accuracy (ResNet50's accuracy is 91.1%), which can save the cost better.

INDEX TERMS Deep learning, image classification, lightweight, attention mechanism, multi-channel feature fusion.

I. INTRODUCTION

From remote and inaccessible forests to bustling cities, there are over 10,000 bird species in almost every environment [1], [2]. However, bird species and populations around the world are decreasing to varying degrees, even facing extinction [29]. For example, Hawaii, known as the "extinction capital," has lost 68% of its bird species, which may have a huge impact on the local food chain and lead to an ecological imbalance. By monitoring populations and numbers, researchers can timely detect changes in bird species and populations and formulate corresponding conservation strategies. The first step in this work is to quickly and conveniently detect birds [3], [4].

Currently, many professionals are starting to observe birds for long periods of time to protect their species. In most cases, many monitors choose bird sounds as the main monitoring method, but this method often works well in quieter mountain forest environments, while in environments with high noise levels, such as cities, this method is not as effective as image detection. Furthermore, birds are one of the most photographed animals, and there are many bird photos online, containing various types of birds with different postures, colors, and shapes [5]. The success of bird classification also contributes to research in species identification [6], [7], [8].

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

Researchers in related fields tend to use Internet of Things devices to monitor bird populations remotely online. However, since most bird protection habitats are in the wild, even under good network conditions, it is difficult for the online monitoring system to transmit the captured bird pictures back to the server for data processing, identification and feedback [9]. If off-line monitoring is performed in bird sanctuaries, low cost embedded devices cannot support the high complexity of image recognition algorithms. To solve this problem, this paper not only envisions designing a lightweight bird recognition algorithm that can achieve high accuracy using simple and single features, but also make the model small enough to run on low-cost embedded devices [10], [11].

A. PRIOR WORK

There has been a lot of research on the problem of bird recognition [28], and there are a variety of methods including images, sound, video and so on. For example, Chao Huang proposed a graphics model (GMS) with salience [6]. By over-segmenting the image into several regions, GMS is used to extract objects and classify the image according to the local upper and lower parts of each region and the global upper and lower parts and salience. Meanwhile, in order to improve the accuracy, SVM was used to classify the images according to the features of the annotated birds [34]. Finally, the posterior probability distribution obtained by GMS and SVM was used to classify the images. Xue Han uses a fusion classification method based on error correction output coding (ECOC) and support vector machine (SVM) to classify 11 species of birds, and extracts the Merle cepstrum coefficient (MFCC) of bird sounds as acoustic features [12]. ECOC-SVM was compared with Random Forest (RF), Gaussian mixture model (GMM) and multi-layer perceptron neural networks (MLP-NN) and convolutional neural networks (CNN) for bird sound classification. Passalis et al. proposes a new hypersphere-based Weight Imprinting(WI) method that is able to train neural networks in a regularized, imprint-aware manner, effectively overcoming the problem that WI cannot handle new classes with multimodal distributions [28]. Li further improved SqueezeNet by embedding two different attention mechanism modules in SqueezeNet in different ways, then fusing the attention feature graph with bilinear fusion, and then fusing the new attention feature graph with the last layer of the network to obtain a new tensor. Finally, it is sent to the linear layer for classification, and good results are obtained [30]. Liao proposed a class of Attention transfer CNNS (CAT-CNNs). Transferring part of the attention knowledge from a very large Fine-Grained Visual Categorization (FGVC) network to a small and efficient network significantly improved its expressiveness [21].

Although previous studies have been extensive and solved some problems, some problems still exist [14], [15], [16]. For example, most models use a small number of dataset in terms of categories and pictures, resulting in poor generalization ability of well-trained models. In addition, the number of model parameters used in most studies is very large and the computational complexity is relatively high, which cannot be deployed in low-cost embedded devices, thus making the application in remote areas more difficult. Although there are some models that can meet the requirements of low power consumption and low calculation, they have not conducted professional training and improvement for bird classification, so the accuracy of these models in bird classification is often a little unsatisfactory. Therefore, it is necessary to design a model with strong generalization ability, low computation and high recognition accuracy [24].

B. CONTRIBUTION

In order to solve the above problems, this paper firstly collected a large amount of bird picture data, constructed a dataset of 275 species of birds, and then designed a lightweight recognition model to classify these pictures and get the classification results. The contributions of this paper are summarized as follows:

1) This paper constructs a dataset containing 275 species of birds, which can be used to improve the generalization ability of the model.
2) This paper proposes a convolutional neural network that can be integrated into embedded devices. By adding multi-channel feature fusion Module and Squeeze-and-Excitation Module and Coordinate attention, the accuracy of model recognition is ensured while the number of model parameters remains basically unchanged.
3) A new feature fusion module is designed in this paper. By capturing the information of multiple channels, the features of the image can be extracted well to strengthen the capturing ability of the model on details and improve the accuracy of the model.

The organization of the remainder of this study is as follows:

In Section II, this paper introduces the establishment of the dataset and some processing of the original data. In Section III, this article describes how to build the model. In Section IV, the results of the ablation experimental, the comparison of the results between different models, and the comparison between the proposed scheme and those of previous schemes are given. Finally, the research is summarized in Section V.

II. DATASET AND METHODS

A. DATASET

This dataset is partly based on data provided by Kaggle in various competitions, and partly on pictures we took in Zijin Mountain, Hongshan Zoo in Nanjing, and Rugao area in Nantong [33]. Since the model in this paper is ultimately applied to embedded devices, it is inevitable that various extreme weather conditions will occur in the actual application process, such as rainy days, foggy days, etc. In this case, the image of the bird will be very different from

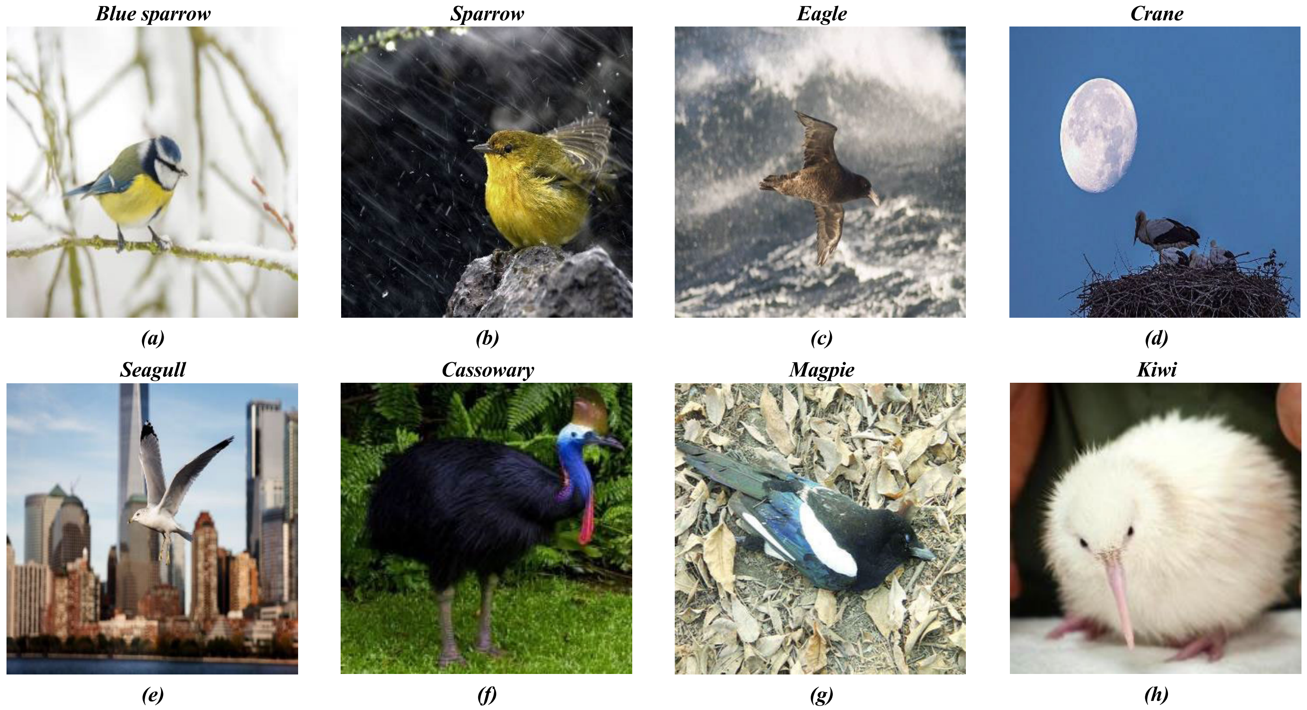


FIGURE 1. Image information on 8 species of birds.

TABLE 1. A subset of data information in a dataset.

Number	The situation	Number of data
0	*Snow day*	*2159*
1	*Rainy day*	*3105*
2	*blurry*	*1909*
3	*Night*	*2166*
4	*city*	*1124*
5	*macrobird*	*1215*
6	*dead*	*1330*
7	*microbird*	*1138*

the normal shooting conditions. However, such images are not reflected in public dataset such as Caltech-UCSD Birds 200 dataset [31] and Birds 525 Species dataset [32]. In order to solve this problem, many such images are added to the constructed dataset in this paper. Compared with Caltech-UCSD Birds 200 dataset and Birds 525 Species dataset, it has a picture of birds in a variety of situations, and about 70 Chinese native birds are added to enrich the regional characteristics of the dataset. This is also the reason why this paper chooses to use self-built dataset. The images in common public bird dataset are usually high in definition and have good shooting conditions, which cannot solve the application problems faced by this paper. Therefore, this paper chooses to construct a new data set. The dataset contains 275 bird species, The training set contains 31,598 images and the test set contains 7,766 images.

In order to better show the difference of dataset, I listed some bird pictures in different situations and made a table with the number of pictures. As shown in Figure 1, a stands for Blue sparrow on a snowy day, b for Sparrow on a rainy day, c for the fuzzy Eagle, d for the Crane at night, e for Seagull in the city, f for the large bird Cassowary, g for the dead Magpie, and so on. h is for Kiwi, a small bird. These are a variety of situations and forms of birds that are missing from some public dataset. Table 1 shows the number of bird pictures in these cases.

B. METHODS

In order to solve the problem of data enhancement and limited size of dataset, this paper manipulates the original image to make about 31000 pictures contained in the dataset rotate, move horizontally or vertically, cut, scale and flip horizontally, and further convert them into training data to enhance the number of training images [6]. These operations will be done before the training of the model and since only some geometric processing has been done to the image, the features of the image will not be changed. Likewise, these actions will not produce a new data set.

III. MODEL STRUCTURE

The model in this article is an improvement over the ShuffleNetV2 model. The main innovation of ShuffleNetV2 is its internal shuffle structure, which makes ShuffleNetV2 work well for problems such as graphic classification. This part is also retained in this paper, but in the structure of

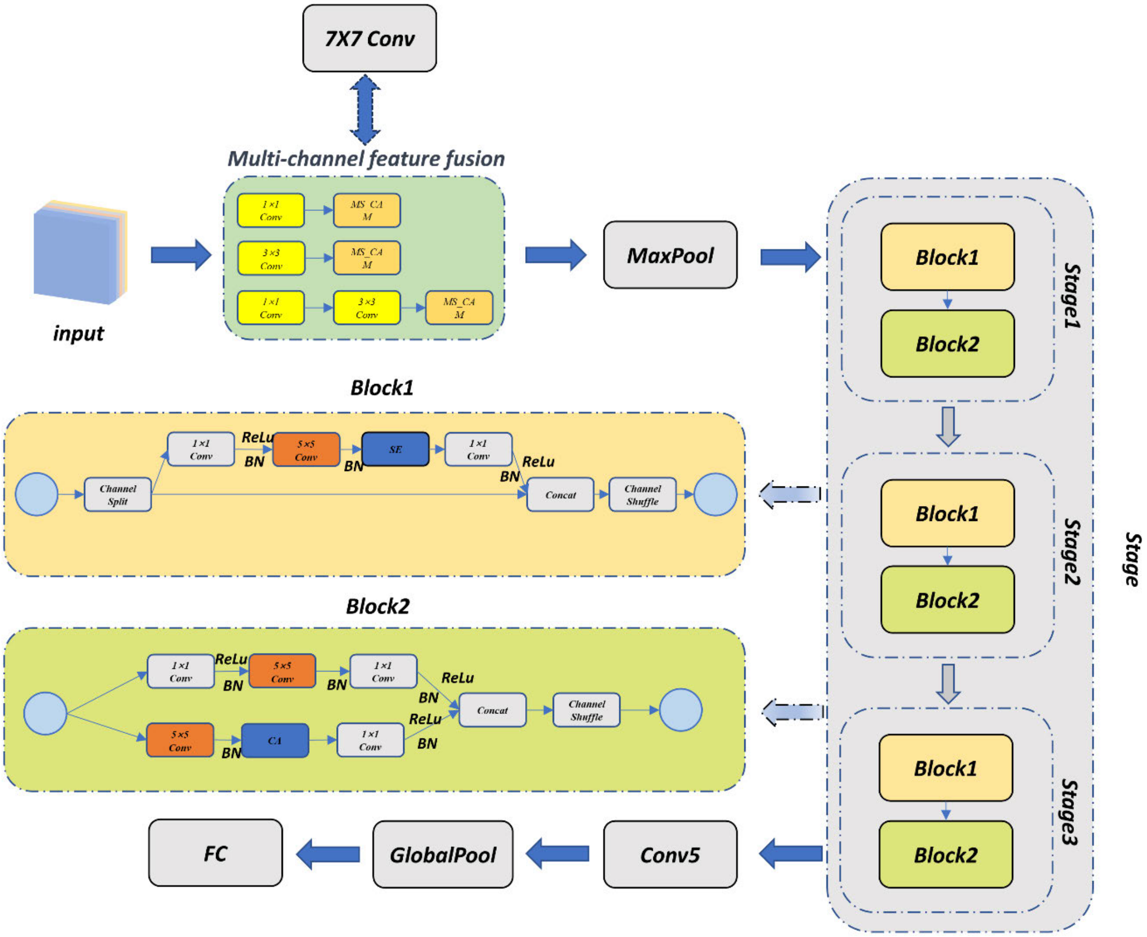


FIGURE 2. Overall structure of network model(The gray part is the original network structure of ShuffleNetV2, the dotted two-way arrows are the replaced parts, the orange is the modified part, and the dark blue is the added part.)

ShuffleNetV2, there is no effective attention mechanism to assist the model to learn features. Based on this starting point, this article adds SE and CA attention mechanisms to the model. At the same time, in order to avoid the increased attention mechanism greatly increasing the complexity of the model, the SE and CA attention mechanisms are placed in two blocks respectively. So these two attention mechanisms are not parallel. In addition, it is noted that ShuffleNetV2 does not have a good feature fusion structure. In order to make up for this regret, the convolutional layer in the model is replaced by the multi-channel feature fusion module proposed in this paper. After these improvements, the model effect has been improved. The structure of the network in this paper is as follows.

A. OVERALL STRUCTURE DESIGN OF THE MODEL

The network backbone model designed in this paper refers to ShuffleNetV2 and introduces a multi-channel feature fusion module in view of the relatively rich features in the initial stage [17], [18], [19], [20]. and since ShuffleNetV2 does not have the ability to capture key information, the Squeeze-and-Excitation Module and Coordinate attention are introduced in this paper. At the same time, the convolution kernel in depth separable convolution is expanded to obtain a larger receptive field under the condition that the number of parameters is basically unchanged. Its network model structure is shown in Figure 2.

As shown in Figure 2, in the initial stage of the model, the original ShuffleNetV2 network used 7 × 7 convolution. However, using this method not only has a large number of parameters, but also cannot extract features efficiently when facing the initial stage with rich image input. For this reason, this paper replaces the original 7 × 7 convolution with the multi-channel feature fusion module created in this paper. The addition of this module can extract the features of the input image in multi-size and multi-channel, which can improve the accuracy of the model and accelerate the convergence speed of the model. Then, the 3 × 3 convolutions in the block module were replaced with 5 × 5 convolutions respectively. The reason is that after replacing the 5 × 5 convolutions, the increase in the number of parameters is small, but the receptive field can be larger. Finally, in the

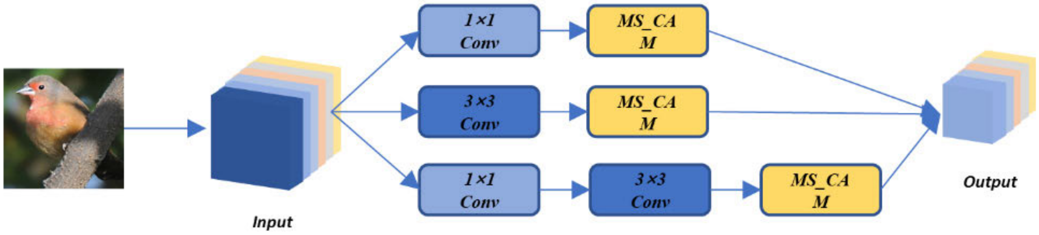


FIGURE 3. Multi-channel feature fusion module.

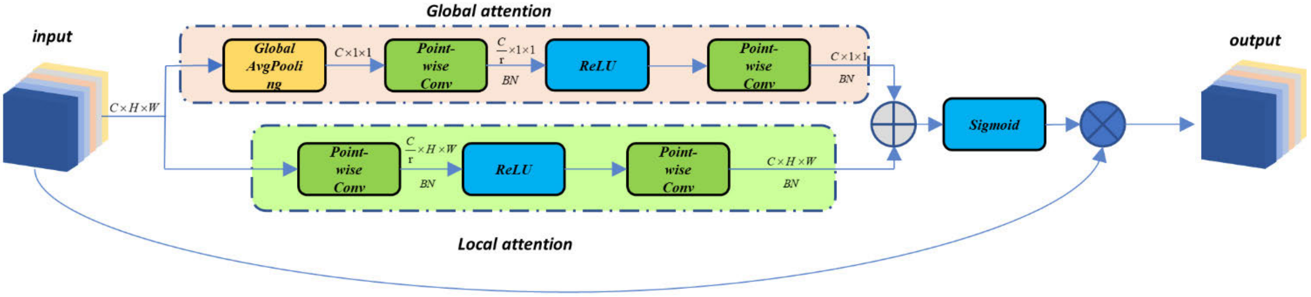


FIGURE 4. Multi-channel feature fusion module.

original block module, the model only performs simple convolution operations when learning features, and does not make the information in each channel get attention. Therefore, in order to enable the model to focus on the information between channels, Squeeze-and-Excitation Module and Coordinate attention are added. The addition of these two attention modules improves the model's ability to perceive channel information, and further makes the model better extract spatial features. Through these operations, the irrelevant information in the global information can be effectively removed, the effective information can be retained to the greatest extent, so that the model can refine the global information to the greatest extent, and the ability of the model to refine the global information can be improved. The model structure is shown in Table 2.

B. MULTI-CHANNEL FEATURE FUSION

In the initial stage of experiment, the input data features are very rich, so it is very necessary to design and use a feature fusion module. A multi-channel feature fusion module is designed and used in this paper, and its structure is shown in Figure 3. In this Module, inception structure and Multi-scale Channel Attention Module (MS_CAM) structure are combined [21], and good results are achieved in practical experiments. In the training process, the convolution of 1 × 1 in inception structure is used to raise and reduce dimensions respectively, which can reduce the complexity of the model while extracting more features. Moreover, due to the convolution at different sizes, features of different scales can also be extracted. Richer features also mean that the final classification judgment will be more accurate.

TABLE 2. Network model structure.

Layer	Size	KSize	Stride	Repeat	Channels
Image	224×224	–	–	–	3
Feature fusion	112×112	3×3	2	1	24
MaxPool	56×56	3×3	2	1	24
Stage1	28×28	–	2	1	116
	28×28		1	3	
Stage2	14×14	–	2	1	232
	14×14		1	7	
Stage3	7×7	–	2	1	464
	7×7		1	3	
Conv5	7×7	1×1	1	1	1024
GlobalPool	1×1	7×7	–	–	–
FC	–	–	–	–	275

In order to further improve the model's ability to capture features, the MS_CAM structure is added to each segment of the inception structure in this paper, as shown in Figure 4. In Figure 4, since scale is not the exclusive problem of spatial attention, channel attention can also have dimensions other than global by changing the dimensions of spatial pooling. By aggregating multi-scale context information along channel dimensions, MS_CAM emphasizes both large objects with more global distribution and small objects with more local distribution, so that the network can recognize and detect objects under extreme scale changes. The feature fusion module is used to improve the ability of the model

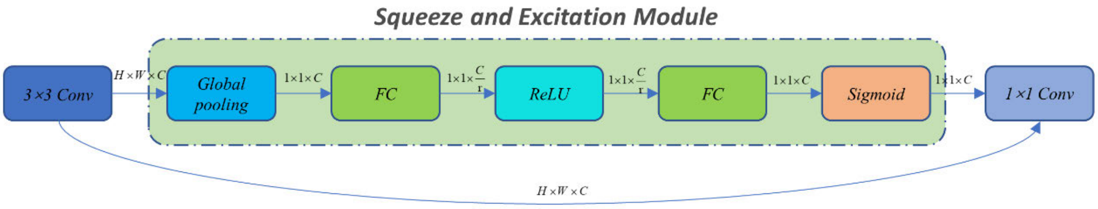


FIGURE 5. Squeeze-and-Excitation Module.

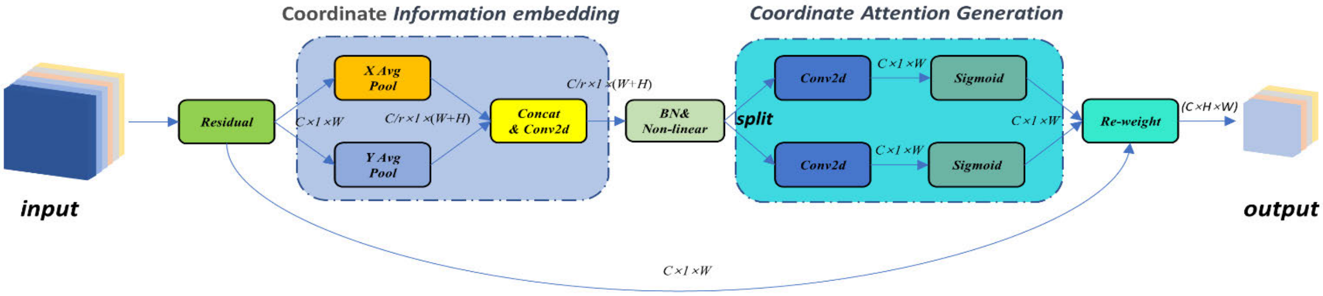


FIGURE 6. Coordinate Attention Module.

to capture important features, thus greatly increasing the recognition accuracy of the model.

C. SQUEEZE-AND-EXCITATION MODULE

For an image, the weight of different channels is generally different. If we can capture this information, then our network can get more information and naturally have higher accuracy. The Squeeze-and-Excitation Module can help us get information inside these channels [22]. The Squeeze-and-relationship between each channel. Finally, the normalized weight obtained before is weighted to the features of each channel, that is, to complete the required features of attention transfer to this paper. Although space is not taken into account in the Squeeze-and-Excitation Module, its structure is simpler. In block1, the model precision can be improved while the number of model parameters remains roughly the same.

D. COORDINATE ATTENTION

Studies in mobile networks show that channel attention mechanisms have a significant effect on model performance [22], but they often ignore location information, which is very important for generating spatially selective attention map. Unlike channel attention, which converts input into a single feature vector by 2D global pooling, Coordinate attention splits channel attention into two 1D feature coding processes that aggregate features along different directions. The advantage is that long range dependencies can be captured along one spatial direction and precise position information can be retained along the other [23]. The generated feature maps are then encoded separately to form a pair of direction-aware and position-sensitive feature maps, which can be applied complementary to the input feature maps to enhance the target features of interest. The two different channel attention directions are coordinate information embedding and coordinate attention generation respectively, and its specific structure is shown in Figure 6.

In the coordinate information embedding part, the channel attention mechanism uses global pooling to encode the spatial information, which results in the global spatial information being compressed into channel descriptors and difficult to store the location information. Therefore, in this Excitation Module is to do attention or gating on the channel dimension. This attention mechanism allows the model to focus on the most excitation channel features while suppressing the less important ones. The implementation of the Squeeze-and-Excitation Module is shown in Figure 5.

As shown in Figure 5, firstly, the two-dimensional features of each channel are compressed into one real number through global averaging pooling, and the global features of the channel level are obtained. After that, a weight value is generated for each feature channel, and the correlation between channels is constructed through two full connection layers. The number of output weight values is the same as the number of channels in the input feature map. The weight of different channels is also obtained by learning the paper will adopt Coordinate Attention global pool is decomposed into a pair of one dimensional feature coding to operate, in the process of actual operation, we will be high and wide use of the image size (H, 1) and (1 W) the pooling of nuclear along the horizontal coordinate direction to encode each channel, can get high and wide feature mapping in two directions.

TABLE 3. Comparison of ablation experiment results.

EXP	MCF	SE	CA	Acc/%	Para(M)	FLOPS(M)
Our module	√	√	√	*92.3*	*2.49*	*166.83*
1	√	×	×	*90.7*	*2.31*	*165.93*
2	×	√	×	*91*	*2.48*	*159.84*
3	×	×	√	*90.5*	*2.32*	*159.58*
4	×	×	×	*90.5*	*2.28*	*152.71*

In this way, long distance correlations can be captured in one spatial direction while accurate positional information can be retained in the other. Then the feature map is encoded as a pair of directional and position-sensitive attention maps, which are complementary to the input feature map. This approach enhances the representation of objects of interest in the model.

IV. RESULT

A. EXPERIMENTAL ENVIRONMENT

This experiment was completed in python3.9 based environment. Model identification and classification were completed in python3.9 and pytoch1.8 based environment. Hardware configuration was 5GHz Intel i7 12700K processor and 32GB 3200Mhz DDR4 memory. Nvidia GeForce RTX3070 and Nvidia GeForce RTX3070Ti graphics cards. The total number of bird picture samples in this experiment is 39,364. 31,598 samples are selected as the training set and 7766 samples as the test set. In the experiment, the initial learning rate was set at 0.1, the dynamic adjustment of the learning rate was used in the training process, the batch size was set at 16, the number of training rounds was set at 300, the model optimizer was used as stochastic gradient descent (SGD), the loss function was used as cross entropy loss function, and the learning rate decline strategy was used as cosine annealing.

B. ATTENTION EXPERIMENT

In order to verify that each improvement point of the proposed model contributes to the improvement of the model performance, a series of ablation experiments are conducted in this paper. In the ablation experiment, the recognition accuracy of the model on the test set was used as the benchmark to judge the effect of model improvement. The ablation experiments included whether a multi-channel feature fusion module was included, whether Squeeze-and-Excitation Module was used, and whether Coordinate attention was used. The ablation experiment is shown in Table 3

As shown in Table 3, when adding different modules individually, the improvement of model accuracy is not obvious, but if they are combined, the effect of the model will be greatly improved, and the number of parameters will remain basically unchanged. In the case of a small change in the number of parameters, it means that the training speed of the model and the test speed after the training will not change greatly, which will increase the benefit of the change.

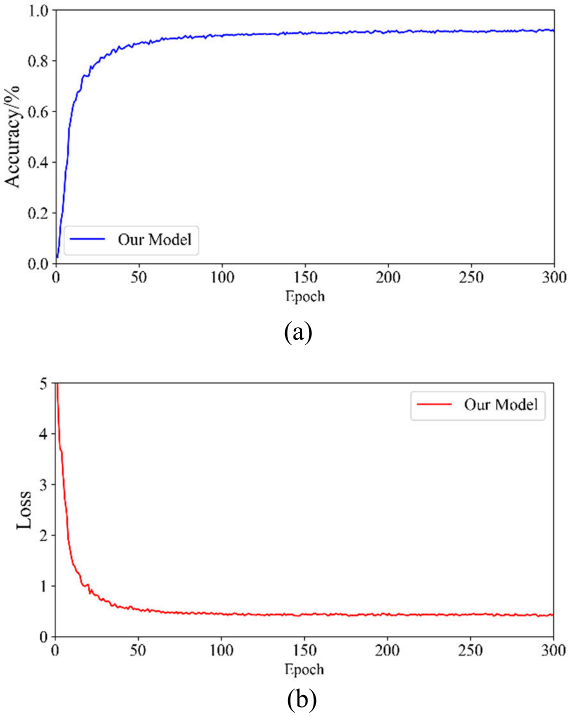


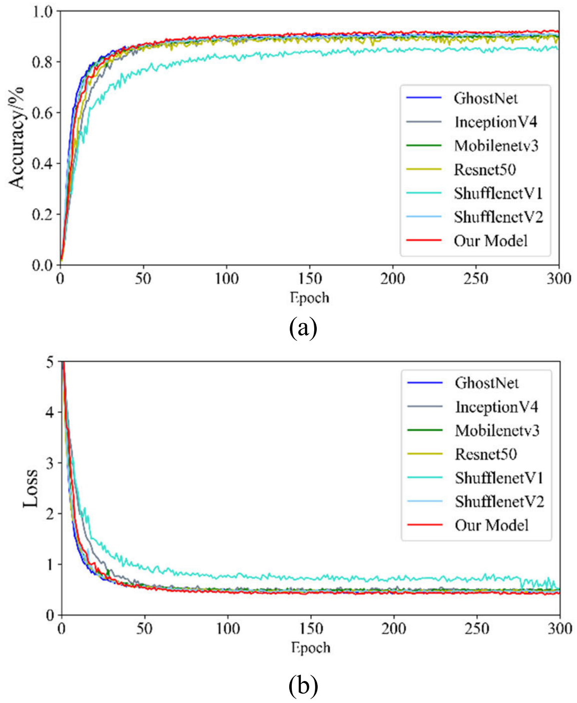
FIGURE 7. A schematic diagram of the training results of the model presented in this paper. (a) Accuracy curve of the test set; (b) Loss curve of the test set.

C. ALGORITHM COMPARISON AND ANALYSIS

Figure 7 shows the performance diagram of the model test set in this article. As can be seen from Figure (a), the convergence rate of the model is fast. The model converges when the model approaches 100 Epoch, and the accuracy rate of the model is as high as 92.3%. As can be seen from Figure (b), the loss rate of the model on the test set is about 0.42. Such an effect can be achieved under the condition of multiple types and fewer pictures, which proves that the model proposed in this paper has a good effect.

In order to compare with the model in this paper, the current classical deep learning models Resnet, InceptionV4 and some lightweight deep learning models MobileNet [35], ShuffleNet [17] and GhostNet [36] are selected respectively to train the dataset constructed in this paper. Test set accuracy and training loss of different models were recorded. In addition, in order to make the comparison model achieve better results, all the comparison models selected in this paper are pre-trained models to achieve better results. as shown in Figure 8. Where epoch is the iteration period of training, Acc is the accuracy of test set, and Loss is the training loss.

As can be seen from Figure 8(a), the model proposed in this paper also has good classification accuracy. At the same time, it can be seen from Figure 8(b) that the training loss ratio of the model proposed in this paper has a fast convergence rate, which is close to convergence at about 100 rounds, indicating



(a)

(b)

FIGURE 8. Diagram of training results of some models including this paper. (a) Accuracy curve of training of some models including this paper; (b) Loss decline curve in the training process of some models including this paper.

TABLE 4. The result representation of each model.

Module	Acc/%	Para(M)	FLOPS(M)
Our module	*92.3*	*2.49*	*166.83*
MobileNetV3	*90.4*	*5.48*	*234.24*
ShuffleNetV1	*86.2*	*2.43*	*168.67*
ShuffleNetV2	*90.5*	*2.28*	*152.71*
GhostNet	*91*	*5.48*	*197.89*
Resnet50	*91.1*	*25.3*	*4288.18*
InceptionV4	*90.9*	*4.81*	*519.99*

that the model has a fast learning ability and can quickly learn some features of birds. The training effect of the lightweight model proposed in this paper is higher than that of ResNet50 and other large networks, and its accuracy is better than that of the lightweight model MobileNet and ShuffleNet.

The statistical results of different models are summarized in Table 4.

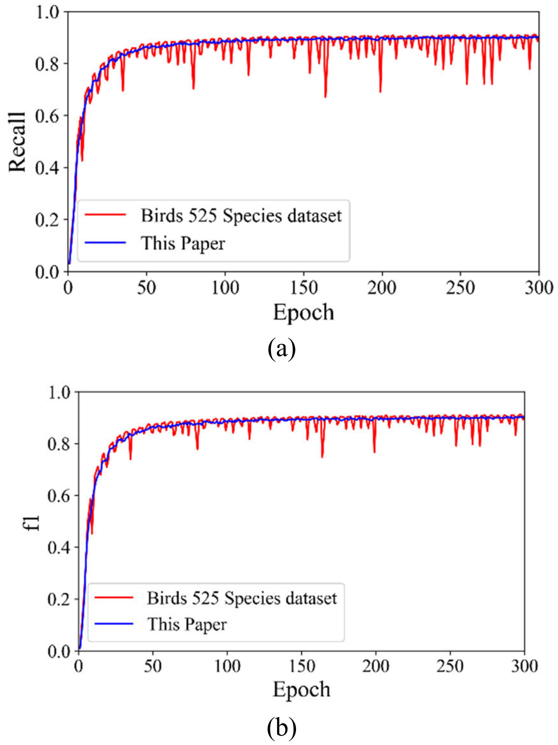
As can be seen from Table 4, the model designed in this paper can get good results with small parameters and calculation amount. Whether compared with the lightweight model of MobileNetV3 or large network Resnet50, has its advantages. Therefore, the model designed in this paper meets the original intention of design.

TABLE 5. Comparison of two hardware platform models.

Device	Acc%	Time(ms)	Price
Jetson Nano	*91.9*	*91*	*$100*
Jetson TX2	*92.2*	*35*	*$800*

TABLE 6. Experimental results of different dataset.

dataset	Acc%	Species
Our	*92.3*	*275*
Caltech-UCSD Birds 200 [31]	*89.1*	*200*
Birds 525 Species dataset [32]	*91.1*	*525*



(a)

(b)

FIGURE 9. Recall and F-1 scores in two different datasets.

D. EMBEDDED APPLICATIONS AND COMPARISON WITH OTHERS' METHODS

In order to verify whether the model designed in this paper has application value on embedded devices, this paper establishes the model on Jetson TX2 and Jetson Nano platform. The results are shown in Table 5.

As shown in Table 5, although there are great differences in reasoning time, classification accuracy is almost the same (the difference in accuracy is due to the compression of the model when deployed on Jetson Nano in this paper) [37]. Moreover, the increase in the number of parameters of the model is very small, so when it is applied to embedded devices, the increase in running time due to the increase in modules is almost negligible. It shows that the application of this model on the hardware platform is feasible, however, when applied to low-cost embedded devices, there are still problems such as the need to compress files (which will lead to the reduction of the accuracy of the final model) and long inference time, which will be our future research direction. Finally, the model is applied to different bird dataset, and the results are shown in Table 6.

As shown in Table 6, when the Caltech-UCSD Birds 200 dataset was used, the result of model training was 89.1%, which was due to the unsatisfactory training results caused by the small number of images in this dataset and poor image quality. When the Birds 525 Species dataset is used, the training result is 91.1%. In order to better evaluate the ability of this model to process very large data and the differences between this dataset and the Birds 525 Species dataset. In this paper, the recall rate and F-1 score curves of the models under the two data sets were plotted, as shown in Figure 9.

As shown in Figure 9, (a) represents the change of Recall value under two different data sets. It is not difficult to see from the curve that the model has a high Recall accuracy in the final result, but with the increase of types, the Recall value will fluctuate greatly. Similarly, from (b), we can also see that the model still has a good F-1 score in training results, but with the increase of types, the curve still produces large fluctuations. Therefore, choosing the data set in this paper can not only solve various situations in practical applications, but also make the model more stable in the training process.

V. CONCLUSION

The model proposed in this paper is improved on the basis of ShuffleNetV2. As can be seen from Table 4, the accuracy of the model proposed in this paper is about 2% higher than that of ShuffleNetV2, and the number of network parameters is not significantly improved. The main reasons are as follows:

1) ShuffleNetV2, as a popular network model, has strong recognition ability. The shuffle structure itself enhances the communication between information, which is of great help to the training and feature extraction of the model.
2) The ShuffleNetV2 network model is referred to in this paper. Although some parameters are introduced in the added structure, the cycles of modules with a high degree of complexity in the ShuffleNetV2 network model are greatly reduced because no structure is added to its branches in the trunk network and the performance is not affected. Therefore, the complexity of this paper is not significantly improved.
3) The multi-channel feature fusion structure is introduced in this model to achieve multi-channel feature capture. In the fusion process, different multi-channel features are extracted through different receptive fields, and then they are splicing, so that the model can more easily capture important features and suppress unimportant features, so as to improve the recognition accuracy of the model.

Finally, as mentioned above, although the model studied in this paper has certain application value on embedded devices, its response speed still needs to be improved when it is actually applied to low-cost devices, which will also be the direction of our follow-up research [26].

REFERENCES

[1] M. D. Dettling, K. E. Dybala, D. L. Humple, and T. Gardali, "Protected areas safeguard landbird populations in central coastal california: Evidence from long-term population trends," *Ornithol. Appl.*, vol. 123, no. 4, p. duab035, 2021.

[2] G. D. Duckworth and R. Altwegg, "Why a landscape view is important: Nearby urban and agricultural land affects bird abundances in protected areas," *PeerJ*, vol. 9, Jul. 2021, Art. no. e10719.

[3] H. S. Oliveira and L. dos Anjos, "Silent changes in functionally stable bird communities of a large protected tropical forest monitored over 10 years," *Biol. Conservation*, vol. 265, Jan. 2022, Art. no. 109407.

[4] H. S. Oliveira, S. F. Gouveia, J. Ruiz-Esparza, and S. F. Ferrari, "Fragment size and the disassembling of local bird communities in the Atlantic forest: A taxonomic and functional approach," *Perspect. Ecol. Conservation*, vol. 18, no. 4, pp. 304–312, Oct. 2020.

[5] X. Yu, W. Zhu, J. Wei, S. Jia, A. Wang, Y. Huang, and Y. Zhao, "Estimation of ecological water supplement for typical bird protection in the yellow river delta wetland," *Ecol. Indicators*, vol. 127, Aug. 2021, Art. no. 107783.

[6] C. Huang, F. Meng, W. Luo, and S. Zhu, "Bird breed classification and annotation using saliency based graphical model," *J. Vis. Commun. Image Represent.*, vol. 25, no. 6, pp. 1299–1307, Aug. 2014.

[7] R. W. Doughty, *Feather Fashions and Bird Preservation: A Study in Nature Protection*. Berkeley, CA, USA: Univ. of California Press, 1975.

[8] R. Brouwer, P. van Beukering, and E. Sultanian, "The impact of the bird flu on public willingness to pay for the protection of migratory birds," *Ecol. Econ.*, vol. 64, no. 3, pp. 575–585, Jan. 2008.

[9] M. Ogueta-Gutiérrez, S. Franchini, and G. Alonso, "Effects of bird protection barriers on the aerodynamic and aeroelastic behaviour of high speed train bridges," *Eng. Struct.*, vol. 81, pp. 22–34, Dec. 2014.

[10] Q. Yang, Z. Zhang, L. Yan, W. Wang, Y. Zhang, and C. Zhang, "Lightweight bird's nest location recognition method based on YOLOv4-tiny," in *Proc. IEEE Int. Conf. Electr. Eng. Mechatronics Technol. (ICEEMT)*, Jul. 2021, pp. 402–405.

[11] Y. Sun, B. Shen, Z. Jin, and Z. Liu, "Fine-grained birds recognition based on lightweight bilinear CNN with additive margin softmax," in *Proc. 14th Int. Conf. Digit. Image Process. (ICDIP)*, Oct. 2022, pp. 242–249.

[12] X. Han and J. Peng, "Bird sound classification based on ECOC-SVM," *Appl. Acoust.*, vol. 204, Mar. 2023, Art. no. 109245.

[13] C. Zhang, Y. Chen, Z. Hao, and X. Gao, "An efficient time-domain end-to-end single-channel bird sound separation network," *Animals*, vol. 12, no. 22, p. 3117, Nov. 2022.

[14] U. Dagan and I. Izhaki, "The effect of pine forest structure on bird-mobbing behavior: From individual response to community composition," *Forests*, vol. 10, no. 9, p. 762, Sep. 2019.

[15] W. Xu, J. Yu, P. Huang, D. Zheng, Y. Lin, Z. Huang, Y. Zhao, J. Dong, Z. Zhu, and W. Fu, "Relationship between vegetation habitats and bird communities in urban mountain parks," *Animals*, vol. 12, no. 18, p. 2470, Sep. 2022.

[16] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[17] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 116–131.

[18] M. G. Hluchyj and M. J. Karol, "Shuffle net: An application of generalized perfect shuffles to multihop lightwave networks," *J. Lightw. Technol.*, vol. 9, no. 10, pp. 1386–1397, Oct. 1991.

[19] S. Türkmen and J. Heikkilä, "An efficient solution for semantic segmentation: ShuffleNet V2 with atrous separable convolutions," in *Proc. 21st Scand. Conf. Image Anal. (SCIA)*, Norrköping, Sweden. Springer, Jun. 2019, pp. 41–53.

[20] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3559–3568.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[22] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.

[23] A. Saad, J. Ahmed, and A. Elaraby, "Classification of bird sound using high-and low-complexity convolutional neural networks," *Traitement du Signal*, vol. 39, no. 1, pp. 187–193, Feb. 2022.

[24] K. Wang, F. Yang, Z. Chen, Y. Chen, and Y. Zhang, "A fine-grained bird classification method based on attention and decoupled knowledge distillation," *Animals*, vol. 13, no. 2, p. 264, Jan. 2023.

[25] Y. Xu, Q. He, Y.-Q. Guo, X.-H. Huang, Y.-R. Dong, Z.-W. Hu, and J. Kim, "Experimental and theoretical investigation of viscoelastic damper by applying fractional derivative method and internal variable theory," *Buildings*, vol. 13, no. 1, p. 239, Jan. 2023.

[26] Z. Zhao and S. Ge, "Aesthetic wideband dielectric resonator antenna based on fractal slot with two independently controllable resonant frequencies," *IEICE Electron. Exp.*, vol. 19, no. 4, 2022, Art. no. 20210508.

[27] Y.-Q. Guo, G. Chen, Y.-N. Wang, X.-M. Zha, and Z.-D. Xu, "Wildfire identification based on an improved two-channel convolutional neural network," *Forests*, vol. 13, no. 8, p. 1302, Aug. 2022.

[28] N. Passalis, A. Iosifidis, M. Gabbouj, and A. Tefas, "Hypersphere-based weight imprinting for few-shot learning on embedded devices," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 925–930, Feb. 2021.

[29] M. Li, L. He, C. Lei, and Y. Gong, "Fine-grained image classification model based on improved SqueezeNet," in *Proc. IEEE 5th Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Mar. 2021, pp. 393–399.

[30] Q. Liao, D. Wang, and M. Xu, "Category attention transfer for efficient fine-grained visual categorization," *Pattern Recognit. Lett.*, vol. 153, pp. 10–15, Jan. 2022.

[31] *Caltech-UCSD Birds-200-2011*. [Online]. Available: https://www.vision.caltech.edu/datasets/cub_200_2011

[32] *Birds 525 Species-Image Classification*. [Online]. Available: https://www.kaggle.com/datasets/gpiosenka/100-bird-species

[33] *200 Bird Species With 11,788 Images*. [Online]. Available: https://www.kaggle.com/datasets/veeralakrishna/200-bird-species-with-11788-images

[34] Z. Zheng, Y. Zhao, A. Li, and Q. Yu, "Wild terrestrial animal re-identification based on an improved locally aware transformer with a cross-attention mechanism," *Animals*, vol. 12, no. 24, p. 3503, Dec. 2022.

[35] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[36] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1577–1586.

[37] S. Xue, Z. Li, R. Wu, T. Zhu, Y. Yuan, and C. Ni, "Few-shot learning for small impurities in tobacco stems with improved YOLOv7," *IEEE Access*, vol. 11, pp. 48136–48144, 2023.

LIU-LEI ZHANG was born in Jiangsu, China, in 2000. He is currently pursuing the master's degree with Nanjing Forestry University, Nanjing, China. His current research interests include artificial intelligence, machine vision, and animal recognition.

YING JIANG received the bachelor's degree from the Nanjing University of Science and Technology, in 2001, the master's degree from Hohai University, in 2005, and the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, in 2014.

She is currently the Master Supervisor with the School of Mechanical and Electronic Engineering, Nanjing Forestry University. Her current research interests include artificial intelligence, deep learning, embedded devices, and sensors.

YOU-PENG SUN was born in Shandong, China, in 1999. He is currently pursuing the master's degree with Nanjing Forestry University, Nanjing, China. His current research interests include artificial intelligence, deep learning, sound signal processing, and machine vision.

YUAN ZHANG was born in Jiangsu, China, in 1998. He is currently pursuing the master's degree with Nanjing Forestry University, Nanjing, China. His current research interests include artificial intelligence, machine vision, vehicle recognition, and embedded devices.

ZHENG WANG was born in Jiangsu, China, in 1997. He is currently pursuing the master's degree with Nanjing Forestry University, Nanjing, China. His current research interests include artificial intelligence, defect recognition, and machine vision.

• • •