IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 28, 2020, accepted August 8, 2020, date of publication August 12, 2020, date of current version September 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3016116

A Novel Multi-Branch Channel Expansion Network for Garbage Image Classification

CUIPING SHI[1,2], (Member, IEEE), RUIYANG XIA[1], AND LIGUO WANG[2], (Member, IEEE)
[1]College of Communication and Electronic Engineering, Qiqihar University, Qiqihar 161000, China
[2]College of Information and Communications Engineering, Harbin Engineering University, Harbin 150001, China

Corresponding author: Cuiping Shi (scp1980@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 41701479 and Grant 61675051; in part by the Project Plan of Science Foundation of Heilongjiang Province of China under Grant QC2018045; in part by the China Postdoctoral Science Foundation under Grant 2017M621246; in part by the Postdoctoral Science Foundation of Heilongjiang Province of China under Grant LBH-Z17052; in part by the Fundamental Research Funds in Heilongjiang Provincial Universities of China under Grant 135309342 and Grant 135309456; and in part by the Key Projects of Natural Science Foundation of Heilongjiang Province under Grant ZD2019F004.

ABSTRACT Due to the lack of data available for training, deep learning hardly performed well in the field of garbage image classification. We choose the TrashNet data set which is widely used in the field of garbage image classification, and try to overcome data deficiencies in this field by optimizing the network structure. In this article, it is found that the deeper network and short-circuit connection, which are generally accepted in the field of deep learning, will not work well on the TrashNet data set. By analyzing and modifying the network structure, we propose an effective method to improve the network performance on TrashNet data set. This method widens the network by expanding branches, and then uses add layers to realize the fusion of feature information. It can make full use of feature information at slight additional computational cost. Using this method to replace the core structure of the Xception network, the performance of the improved network has been improved greatly. Finally, the M-b Xception network proposed by us achieves 94.34% classification accuracy on the TrashNet data set, and has certain advantages over some state-of-the-art methods on multiple indicators. The python code can be download from https://github.com/scp19801980/Trash-classify-M_b-Xception.

INDEX TERMS Garbage image classification, deep learning, feature information fusion, multi-branch, small data sets.

I. INTRODUCTION

With the development of human society, the problem of environmental pollution is becoming more and more serious [1], and environmental pollution has great harm to the earth and all its organisms [2]. Among them, most of the pollution is caused by domestic garbage. The decomposition of some domestic garbage may lead to the high concentration of chemical substances in the environment [3], damaging the ecological environment. There are also some domestic wastes that rarely biodegrade. For example, plastics are ubiquitous pollutants in all marine environments around the world [4]. Therefore, the first step to solve the waste pollution is to classify the waste according to its nature. Many countries in the world require to separate dumping of wastes [5]. Nevertheless, for those who lack professional knowledge, it is very difficult to identify all kinds of domestic garbage accurately. The intelligent garbage classification system can solve this problem. Applying it to the intelligent garbage bins or smart phones can guide people to dump the domestic garbage correctly. At present, the difficulty is that the intelligent garbage classification system cannot classify garbage images accurately.

In the past decade, due to the improvement of computing power and theoretical system, deep learning accessed a period of rapid development [6]. Now deep learning has penetrated into all aspects of computer vision, and has achieved exciting results in image classification, target detection and image semantic segmentation tasks [7]. The main advantages of deep learning method over other machine learning methods is its powerful modeling ability [8], as well as its end-to-end learning method, which free people from heavy manual work

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

[9]. Deep learning is highly dependent on data [10]. As a relatively new field, garbage classification has no standardized data sets for neural networks to learn, so in 2016, Mindy Yang and Gary Thung established the TrashNet Database for garbage image classification [11]. After that, the work about garbage image classification based on deep learning is gradually increased. Although the TrashNet data set has been widely used, due to the small number of images and the lack of feature information, these work based on it have not achieved good results.

For the garbage image classification based on deep learning, people tend to use deeper neural network gradually. In fact, image processing based on deep learning is mostly developing towards deeper network [12]. Generally speaking, deepening the network will make the work of each layer simpler, which helps to obtain a better nonlinear representation ability [13], and better fitting effect for complex feature information. However, the result of deepening the network is not certain. It is usually appropriate to use a deeper neural network in target detection, image semantic segmentation, or complex scene classification [14]. For the TrashNet, a small data set with a single background, the difficulty comes from the data set itself, i.e., the small amount of feature information, the small number of data samples, and the large similarity among classes. In this case, classification performance is hardly to be improved by increasing network depth. Therefore, we should consider the particularity of garbage classification and explore a more effective method.

In summary, the main contributions of this article are listed as follows.

- We not only analyze the characteristics of the TrashNet data set and the reason why the deeper neural network is not suitable for the TrashNet data set but also prove that much deep neural network will reduce the accuracy. We shortened the Xception network and achieved higher accuracy on the TrashNet data set.
- A new method that can expand the branch of a specific network layer is proposed. This method can widen network structure and extract feature information more effectively, which is beneficial for garbage image classification.
- The proposed network broadening method was used to improve the shortened Xception network and the network structure with better performance obtained. We called it as M-b Xception. Moreover, comparing with some state-of-the-art methods, the M-b Xception can provide higher classification accuracy, and more balanced single category classification ability, which allows it to be used in practical applications.

The rest of this article is organized as follows. Section 2 discusses some related work on garbage image classification. In Section 3, the TrashNet data set was tested and a method based on multi-branch feature information fusion is proposed. We used this method to establish the M-b Xception that performed better on the TrashNet data set. In Section 4, First, we evaluated the M-b Xception and two contrast networks in this article in several different aspects, and proved the advantage of the M-b Xception. After that, we optimized the number of channels for M-b Xception to further improve the accuracy. Then we compared it to some new works. Finally, conclusions are provided in Section 5.

II. RELATED WORK

In the past few years, researchers have done a lot of work for garbage image classification, which can be divided into two categories, based on traditional machine learning methods and based on end-to-end learning system.

A. TRADITIONAL MACHINE LEARNING METHODS

1) SUPPORT VECTOR MACHINE (SVM)

SVM is a powerful classification algorithm developed from Vapnik's statistical learning theory [15], which deals with machine learning tasks on the basis of optimization theory [16]. Its main operation is to find an optimal hyperplane to separate the two categories. For multi-classification problems, we can combine multiple SVM classifiers, or use the one-time solution method, which is to optimize the parameters of all categories by an optimized formula. SVM provides different processing methods for linear separable problems and nonlinear separable problems [17]. In 2016, Mindy Yang *et al.* utilized the SVM algorithm to work on the TrashNet data set, with the final accuracy of 63%.

2) K-NEAREST NEIGHBOR (KNN)

KNN classification algorithm is one of the simplest methods in data classification [18]. KNN algorithm has no training stage. In KNN algorithm, the category of the samples to be divided is determined according to the category of the nearest one or several samples [19]. KNN is widely used in image classification and prediction because of its efficiency and simple implementation [20]. In 2018, Bernardo S.Costa *et al.* used KNN algorithm to classify six kinds of garbage images in the TrashNet data set, with the final accuracy of 88% [21].

3) RANDOM FOREST (RF)

RF algorithm is to generate a number of different data sets by sampling, and then train a classification tree on each data set. Each tree will participate in the final decision of prediction results [22]. The advantage of RF algorithm is that it has good robustness when dealing with missing data and strong reliability when dealing with tasks with more variables [23]. In the meanwhile, it has fast training speed [24]. In 2018, Mandar Satvilkar used RF algorithm to classify garbage images on the TrashNet data set, achieving 62.61% accuracy [25].

4) EXTREME GRADIENT BOOSTING (XGBoost)

XGBoost algorithm is improved based on GBDT (Gradient Boosting Decision Tree) algorithm [26], which is a kind of supervision algorithm [27]. The idea is to establish a certain number of classification regression trees, so that the predicted number value of the tree group is as close to the real number value as possible and has the greatest generalization ability [28]. The advantage of XGBoost algorithm is that it is hard to over-fitting, it can specify the default direction of branch for missing number value or specified number value. In 2018, Mandar Satvilkar used XGBoost algorithm to classify garbage images on the TrashNet data set, achieving 70.1% accuracy.

B. END-TO-END LEARNING SYSTEM

Generally speaking, the traditional machine learning method has achieved good results in the field of image processing due to its long development time and effective theoretical system [29]. Nevertheless, these methods are usually composed of several independent steps, so they need a lot of storage space to store the intermediate results [30]. The implementation process is cumbersome and not intelligent enough. The emergence of end-to-end learning system solves this problem. As long as the training data and test data are given in advance, the end-to-end learning system will automatically calculate the error results between the prediction and the real data. In addition, it will update the weight and solve the gradient by using the back propagation method, and find the minimum value of the loss function with the gradient descent method [31]. The whole process of convergence is accomplished independently and coherently. In recent years, there are many garbage image classification work based on end-to-end learning, and they all use the TrashNet data set.

In early 2018, Kennedy Tom proposed the OscarNet network (fine-tuned by vgg19), which achieved 88.42% classification accuracy [32]. In October 2018, Bernardo S. Costa *et al.* proposed a fine-tuned AlexNet network with 91% accuracy and a fine-tuned VGG16 network with 93% accuracy [21]. Stephen L. Rabano *et al.* proposed a fine-tuned MobileNet network, which achieved 87.2% accuracy [33]. In December 2018, Rahmi Arda Aral *et al.* tested multiple classic networks on the TrashNet data set. They achieved 89% accuracy using Inception-Resnet V2 and 89% accuracy using DenseNet121. They also attempted to fine-tune the weight of the pre-trained model on the ImageNet data set, where fine-tuned DenseNet121 achieved 95% accuracy and fine-tuned Inception-ResNet V2 achieved 94% accuracy [34]. And in June 2019, Victoria Ruiz *et al.* achieved an accuracy of 87.71% by using the Inception network, 88.34% by using the Inception-ResNet network and 88.66% by using ResNet network [35]. These methods are based on the classic networks with outstanding performance in large-scale target detection, image semantic segmentation and multi-category image classification competitions. Moreover, these classic networks are improved with common methods in the field of deep learning, without considering the particularity of the TrashNet data set. Our work just fills in the gaps in the above works.

III. MATERIALS AND METHODS

A. DATA SET AND ITS SPECIFICITY

1) TrashNet DATA SET

The data set used in this article is the TrashNet, which was produced by Mindy Yang and Gary Thung in 2016. There are 2527 RGB images, including 501 glass images, 594 paper images, 403 cardboard images, 482 plastic images, 410 metal images and 137 trash images. The background of all images is white, and they are all taken under sufficient illumination. All images are 512 × 384 pixels in size. Fig. 1 shows some garbage images of the TrashNet data set.

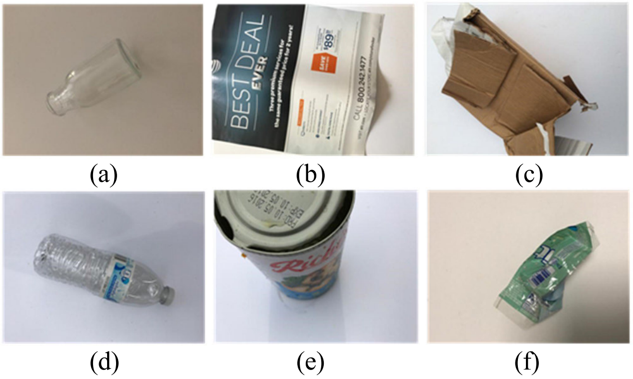


FIGURE 1. Some garbage images of the TrashNet data set. (a) glass, (b) paper, (c) cardboard, (d) plastic, (e) metal, and (f) trash.

Different from other classification data sets, each image in TrashNet data set only contains a single object, which may make the task easier for human eyes, but not for computers. Convolutional neural network has the ability of feature extraction far beyond human eyes. For computers, it's not difficult to find the details of all positions in the image by using the trained model [36]. However, for the images of the TrashNet data set which merely contain a single object, the number of features that can be extracted is few, so the fault tolerance is poor. There are no other objects in the image can provide extra feature information. When there are some differences between the sample object and its class, it will finally show great differences. It is one of the difficulties of garbage classification task. Another difficulty is the limited amount of data in the TrashNet data set. Deep learning relies on large-scale data for massive parameters training due to the very cumbersome gradient back-propagation optimization [37]. When the data is insufficient, the training process will be very difficult, and even significantly overfitting [38]. Owing to the TrashNet dataset's two challenges, we next tested it quantitatively.

2) DEEP CNN IS NOT SUITABLE FOR THE TrashNet DATA SET

With the improvement of computing power and the solutions to the gradient disappearance problem, people tend to use deeper neural network to realize the complex scene image classification. The advantage of this is obvious. A deeper

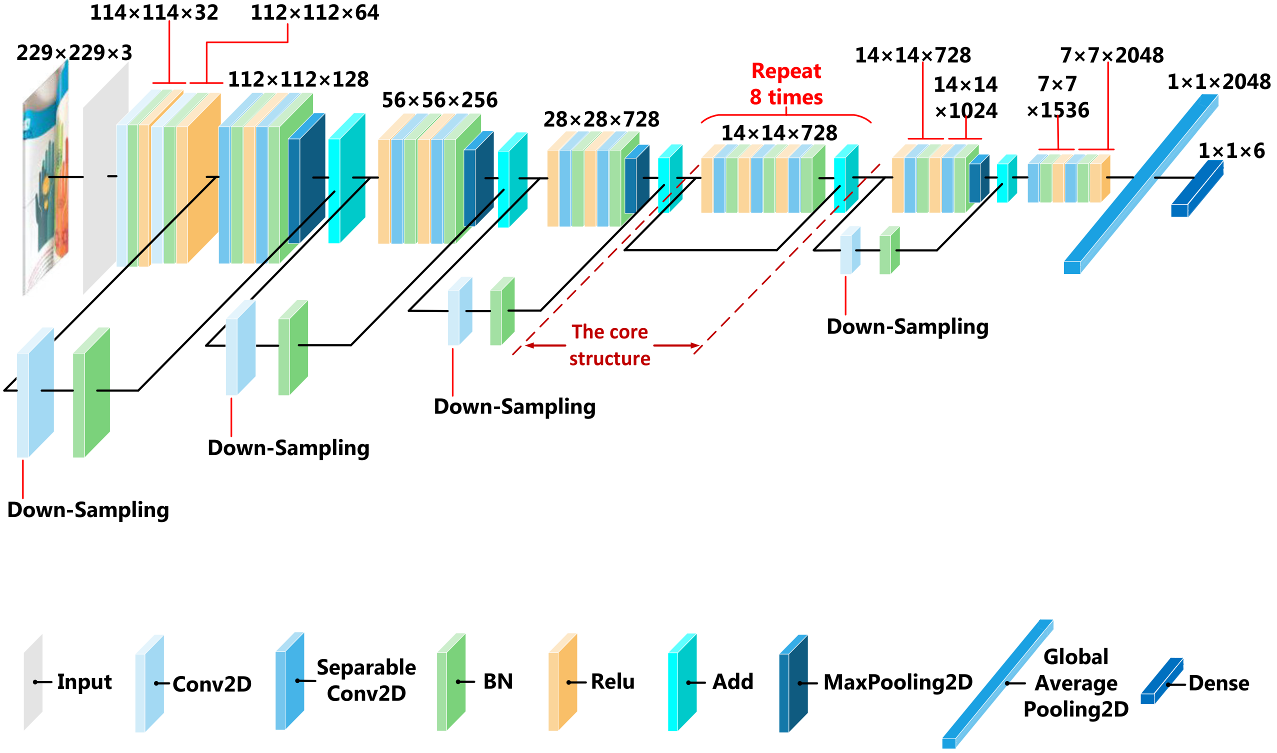


FIGURE 2. The Xception network structure.

network means better nonlinear representation ability, learning more complex transformations, and fitting more complex feature inputs [39]. However, experiments show that deeper network is not beneficial for image classification on the TrashNet data set.

Firstly, the Xception network is used to test the data set. The Xception network structure is shown in Fig. 2. The part of the network receiving 14 × 14 × 728 images is called the core structure part. Owing to the limited space, it is not expanded. Actually, the core structure is composed of a 9-layer structure repeated 8 times of linear stacking. The 9-layer structure consists of 3 Relu layers, 3 separable Conv2D layers and 3 batch normalization (BN) layers. Here, the depth separable convolution (DSC) uses separable Conv2D under the Keras framework instead of depthwise Conv2D. separable Conv2D must be used with the activation layer [40].

After that, some network layers of Xception are removed to explore the relationship between network depth and network performance on TrashNet data set. Here, the non-core structure of Xception is preserved. The reason is that there are few convolution layers in non-core structure. If the number of convolution layers is further reduced, the performance of the network will be greatly affected [41], [42]. Therefore, we chose to remove some network layers from the core structure of Xception to achieve network shortening.

For more intuitive, the complete core structure of Xception network is shown in Fig. 3. By removing part of the network layer, 7 new network structures are constructed and named L-w Xception (Lightweight Xception) 1-7. The core structures of L-w Xception 1-7 are shown in Fig. 3, and the non-core structure is exactly the same as that of the Xception network.

TABLE 1. Comparison of accuracy and number of parameters of 7 lightweight Xception networks and Xception networks.

Method	Accuracy	parameters
Xception	0.9150	20,873,774
L-w Xception 7	0.9150	19,255,430
L-w Xception 6	0.9172	17,637,086
L-w Xception 5	0.9172	16,018,742
L-w Xception 4	0.9172	14,400,398
L-w Xception 3	0.9172	12,782,054
L-w Xception 2	0.9194	11,163,710
L-w Xception 1	0.9194	9,545,366

The 7 newly constructed networks L-w Xception 1-7 and the Xception are trained under the same conditions. The accuracy and the number of parameters of them are listed in Table 1. It can be seen that in Table 1, with the shortening of the network, the number of network parameters decreases continuously, and the accuracy even increases slightly. This shows that the deeper network cannot play a positive role for the classification on TrashNet data set.

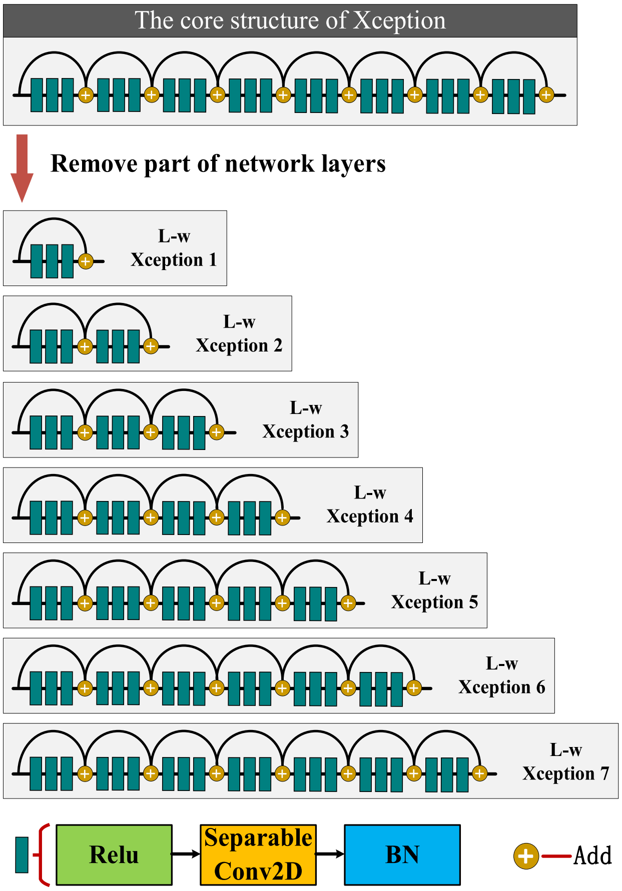


FIGURE 3. Build 7 lightweight Xception networks.

In this experiment, the L-w Xception 1 with the shortest core structure is proved to have the best performance. In the following experiment, the L-w Xception 1 is adopted and simply referred as "L-w Xception".

B. THE PROPOSED METHOD

1) THE PROPOSED NETWORK BROADENING METHOD

Depth and width are two important attributes of the convolutional neural network. Enough depth can make the network have good nonlinear representation ability, and enough width can make the network learn abundant features [43]. In general, we prefer to deepen a network rather than widen it when improving the network structure. This is because when the depth and width of the network are small, deepening the network usually yields higher performance gains [44]. However, when the network is deep enough, further deepening the network will not improve the network performance but make the network more difficult to train and even make it performance decline [45]. Some related works have also shown that shallow and wide networks may work better than deep and narrow ones [46]. Sergey Zagoruyko *et al.* have proved that a wide ResNet can achieve at least as much accuracy as a deep ResNet [47]. In 2016, Junting Pan *et al.* proposed a shallow network and a deep network, which proved to have similar prediction errors in saliency prediction [48]. Work by Zifeng Wu *et al.* has shown that well-designed shallow networks can perform better than many deep networks [49]. In the previous experiments, the shallow L-w Xception network is proved more suitable for the Trash net data set than the deep Xception network. Therefore, we propose a network broadening method to further improve the network performance.

The network broadening method we proposed is to build branches for the target network layer and to realize the feature information fusion among branches with the add layer. It is important to note that the use of the add layer here is different from the traditional residual connection method. The traditional residual connection usually adds a short-circuit mechanism to the linear network, as shown in Fig. 4.

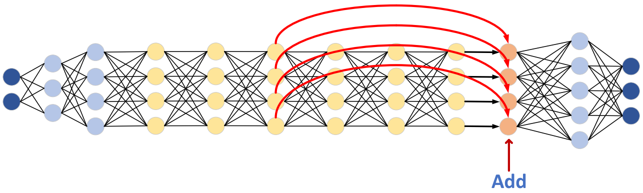


FIGURE 4. The traditional residual connection method.

Its idea is to map the information identity in the lower layer of the network to the higher layer of the network. In calculation, it can be understood as put the input x into the output as the initial result, so that the output $H(x) = f(x) + x$ [50]. Even when $f(x) = 0$ is output in the deep network, $H(x)$ can also be made equal to x. The advantage of it is that the transmission gradient can be lossless. Therefore, the negative effect of back propagation, i.e., the problem that the learning rate of the network layer near the input is smaller than that near the output is solved.

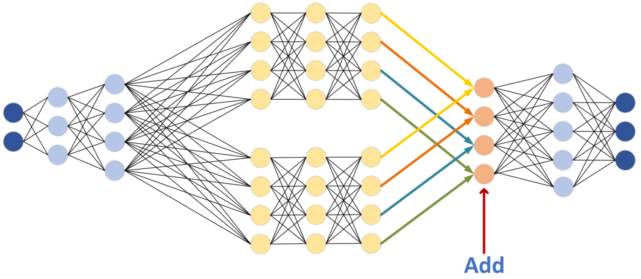


FIGURE 5. Our proposed method of network broadening.

The add layer used in this article can realize horizontal corresponding channel information fusion, that is, stacks the output of multiple branches. As shown in Fig. 5. There are three ways to increase the width of network. The common method is to expand the number of channels or use concatenate layer for channel merging. As a matter of fact, the use of add layer will also have the same effect. The difference is that enlarging the number of channels can directly increase the number of features extracted from each convolution layer; the concatenate layer can stack the channels horizontally, which will increase the number of features, but the information of Each channel will not increase; the add layer will increase the amount of information in each channel, but the number of feature maps will not increase.

Now, three network widening methods are compared from the perspective of the number of network parameters. K represents the size of convolution kernel, N represents the number of channels of the layer, and M represents the number of feature maps of the previous layer. The number of parameters of Conv2D can be calculated as

$$Param = K^2 \cdot M \cdot N \tag{1}$$

Taking the structure of Fig. 5 as an example, the size of input image is 256 × 256. Here, it is considered that the convolution type is Conv2D. Both add layer and concatenate layer will not increase the additional parameter amount, but concatenate layer will increase the number of channels and the parameter amount of the later network layer. The total number of parameters of convolution layer is 1962 by doubling the number of convolution layer channels directly, 1386 by concatenate layer and 1206 by add layer.

Next, in terms of time complexity, the three network widening methods are compared. The time complexity of convolutional neural network is evaluated as

$$Time \sim O\left(\sum_{l}^{D} S_l^2 \cdot K_l^2 \cdot M_l \cdot N_l\right) \tag{2}$$

The size of the input feature maps S is determined by the size P of the input matrix, the size K of the convolution kernel, the *Padding*, and the *Stride* together. The corresponding relationship can be represented as

$$S = \frac{(P - K + 2 \cdot Padding)}{Stride} + 1 \tag{3}$$

Obviously, it is more advantages in time complexity by using add layer to increase network width. Lower time complexity means less training time, faster prediction, and lower demand for computing power.

Next, some experiments are performed to compare the three network broadening methods quantitatively. In the experiments, the three methods are used to double the width of the core structure of L-w Xception network respectively, as shown in Fig. 6. Fig. 6 does not show the full network structure. The parts that are omitted are the same as that of the Xception.

In Fig. 6, the method 1 is to double the number of convolution layer channels in the core structure. Since the add layer can only connect to the network layer with the same number of channels, we add a 1 × 1 convolution layer, followed by a BN layer. The method 2 is to expand a branch of the core structure, and to connect the outputs of the two branches with a concatenate layer. Since the number of channels after connection becomes 1456, we add a 1 × 1 convolution layer and a BN layer. In method 3, the proposed network broadening method is utilized. A branch of the core structure is expanded, and then the outputs of the two branches are connected with the add layer. Since the number of channels before and after the connection remains unchanged, there is no need to add 1 × 1 convolution layer in the residual connection.

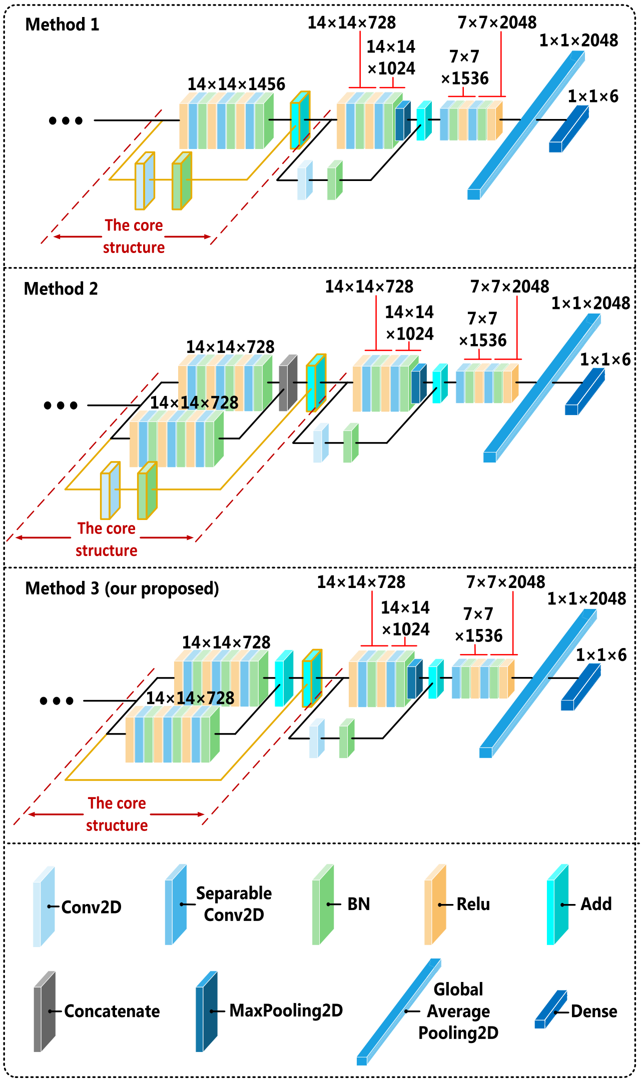


FIGURE 6. Double the width of the core structure of L-w Xception network by three network broadening methods.

	Residual connection	Accurancy	Parameters
Method1	yes	0.9216	15,624,894
	no	0.9237	14,559,102
Method2	yes	0.9216	13,511,510
	no	0.9216	12,445,718
Method3	yes	0.9237	11,163,710
	no	0.9259	11,163,710

FIGURE 7. Comparison of the three network broadening method with and without residual connection.

The above three methods are trained under the same settings, and the experimental results are shown in Fig. 7. Moreover, in order to analyze the influence of residual connection on network performance, we also remove the residual connection in the above three methods, i.e., to remove the yellow line and the network layer with yellow border in Fig. 6. The experimental results are also shown in Fig.7. It can be seen that, under the same conditions, the accuracy of the proposed network broadening method is higher than that of method 1 and method 2, and the number of parameters is also less, whether there is residual connection or not. This proves the network broadening method with add layer is more suitable. Moreover, for the proposed method, after removing residual connections, the network performance can be further improved. Therefore, we choose the network broadening method without residual connection as the final method.

2) THE PROPOSED LEARNING NETWORK

According to the characteristics of TrashNet data set, the Xception network is improved by using the proposed channel expansion method. The network structure is shown in Fig. 8. After a large number of tests on the Xception network, it is found that the network layer with four Downsampling times has the best effect on the feature extraction of TrashNet data set. Therefore, we broaden this part of the network structure, the specific method is to change the original 8 times repeated linear structure to 8 branches parallel structure. The reason why the number of branches is set to 8 is that it can make full use of the advantages of this part of structure feature extraction. In addition, it ensures that the proposed network has the same number of network layers as the Xception network. This method is convenient for quantitative analysis as well as comparison.

We call this new network as M-b Xception (Multi-branch Xception). The M-b Xception finally achieve an accuracy of 0.9325 on the TrashNet data set, which was 1.75% higher than the Xception. We also try to add a residual connection to the core structure of M-b Xception, but we don't get satisfactory results. The accurancy obtained by adding the residual connection was 0.88% lower than before, so we ultimately chose not to add the residual connection to the core structure of M-b Xception.

In the process of forward propagation of the proposed learning network, the input of the first convolution layer of the l branch is:

$$Z^{(l)} = \sum_{n=1}^{728} \left[\sum_{m=1}^{728} \left(W_{mn}^{(l)} \cdot a_n^{(l)} + b_n^{(l)} \right) \right] \tag{4}$$

W is the weight of the m-th neuron in the input assigned to the n-th neuron in the target convolution layer. $a_n^{(l)}$ is the output of the n-th neuron in the previous layer. $b_n^{(l)}$ is the corresponding bias. Although the information source of the eight branches is the same one, the information assigned to the first convolution layer among them is different due to different weights as well as bias, which is controlled by back propagation. The first convolution layer of the 8 branches receives the information of each branch and then carries out the cross-correlation operation of each branch. The output is given by

$$a^{(l)} = g\left(Z^{(l)}\right) \tag{5}$$

Here, $a^{(l)}$ and $Z^{(l)}$ are the output and input of the first separable Conv2D layer of the l-branch, respectively. $g(\cdot)$ represents the cross-correlation operation. The outputs of the 8 convolution layers are respectively passed to the next network layer of the current branch until the last layer is completed. This process is independent and will not interfere with each other. At last, the output of each branch will be fused by 7 add layers.

There are four short-circuit connections in the non-core structure of the network. Consider the network structure between each short-circuit connection as a short-circuit block. According to the chain rule, the back propagation gradient from deep short-circuit block D to shallow short-circuit block S can be defined as

$$\frac{\partial loss}{\partial x_S} = \frac{\partial loss}{\partial x_D} \cdot \left[1 + \frac{\partial}{\partial x_D} \sum_{i=S}^{D-1} g\left(x_i, W_i\right) \right] \tag{6}$$

In formula (6), $loss$ is the error distance between the actual output value and the label value. Gradient $\frac{\partial loss}{\partial x_S}$ consists of two parts. One part is $\frac{\partial loss}{\partial x_D} \cdot \left[\frac{\partial}{\partial x_D} \sum_{i=S}^{D-1} g\left(x_i, W_i\right) \right]$ that propagated through the weight layer, and another part is $\frac{\partial loss}{\partial x_D}$ that propagated directly [51]. Gradients in the deep layers of the network can spread to the shallow layers of the network, which ensures that the non-core structure of M-b Xception does not cause the gradient to disappear.

In the core structure of M-b Xception, the back propagation process is shown in Fig. 9. The outputs of the 8 branches after passing the 7 add layers are

$$z\left(x_i\right) = \sum_{i=1}^{8} x_i \tag{7}$$

It can be seen that formula $\frac{\partial z}{\partial x_i} = 1$ holds no matter when integer i takes from 1-8. In Fig. 9, a represents the derivative of the back propagation to the core structure of the network. Suppose $a = \frac{\partial L}{\partial z}$, according to the chain rule, $b = \frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial z} \cdot 1$ will be obtained. Therefore, in the core structure of M-b Xception, gradients can be transmitted to each branch losslessly. In addition, the network structure of each branch is relatively short, so the core structure of M-b Xception network will not cause the gradients to disappear.

IV. EXPERIMENTS AND DISCUSSIONS

A. THE PARAMETER SETTIN

In this section, the data enhancement settings and optimization settings are listed. In order to ensure that the experimental results only reflect the performance of the network structures,

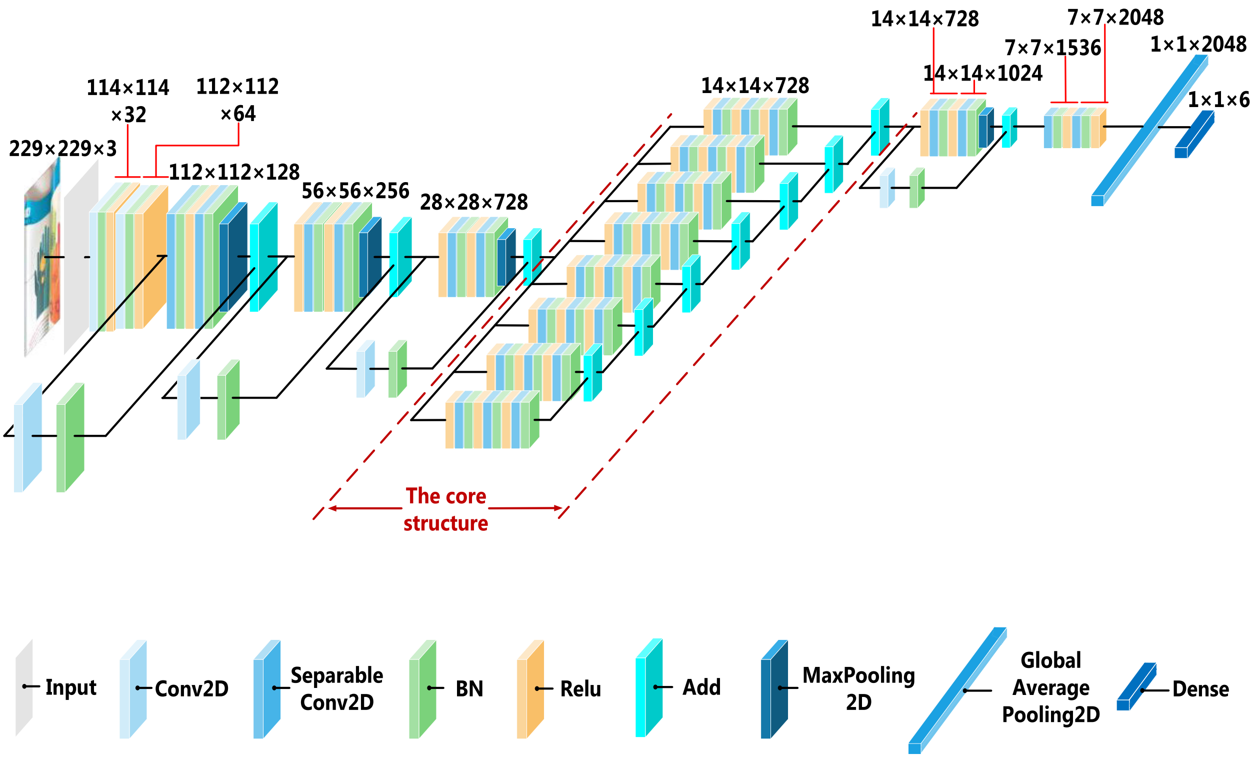


FIGURE 8. The M-b Xception network structure.

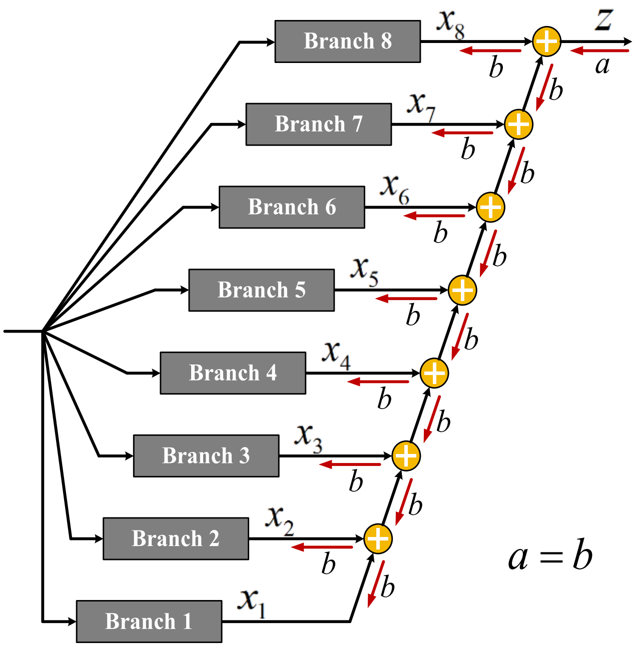


FIGURE 9. The back propagation in the core structure of M-b Xception.

and then verify the effectiveness of the proposed method, data enhancement is used in the training process of all networks, and the relevant settings are exactly the same.

1) DATA ENHANCEMENT

Due to the lack of images in the TrashNet data set, data enhancement is utilized to suppress the phenomenon of over fitting during training [52]. It is a process of expanding data samples in the training data set [53], including cutting, scaling, rotation and so on. The data enhancement settings are listed in Table 2.

TABLE 2. Data enhancement settings.

Project	Settings
shear_range	0.2
zoom_range	0.2
horizontal_flip	True
rotation_range	40
width_shift_range	0.2
height_shift_range	0.2
fill_mode	nearest

2) THE OPTIMIZATION AND REGULARIZATION SETTINGS

The deep learning method usually requires a lot of training times to achieve an excellent performance of a model. If the parameters are not initialized correctly, the training process would require a long time and fall into a local minimum [54]. After a large number of experimental tests, the appropriate optimal parameter settings of our model are listed in Table 3.

B. THE EXPERIMENTAL COMPARISON OF THE THREE NETWORKS

In this section, we will further examine the performance of the three representative networks in the methodology. These networks are the Xception, the L-w Xception obtained by

TABLE 3. Optimization and regularization settings.

Project	Settings
Optimizer	SGD
Momentum	0.9
Initial learning rate	1e-3
Learning rate decay	ReduceLROnPlateau
L2 regularizers	1e-4

TABLE 4. Accuracy, number of parameters and time of single image training of three networks.

Method	Accurancy	parameters	S-I training time
Xception	0.9150	20,873,774	44.69 ms
L-w Xception	0.9194	9,545,366	30.52 ms
M-b Xception	0.9325	20,873,774	44.69 ms

shortening the Xception core structure and the M-b Xception obtained by network broadening.

1) COMPARISON OF ACCURACY, NUMBER OF PARAMETERS, AND TIME COMPLEXITY

Table 4 shows accuracy, the number of parameters and time of single image training of three networks. Our operating system is Windows 10 with an Intel Core i5-8300K, 8 GB of RAM and GeForce GTX 1050 Ti. Since the L-w Xception is obtained by shortening the Xception, it has a much smaller number of parameters than the Xception. The M-b Xception is obtained from the L-w Xception extension branch. Since the number of network layers and the number of channels in each layer of the M-b Xception is exactly equal to the Xception, they also have the same number of parameters. It is satisfactory that the S-I training time (Single image training time) of the M-b Xception is only a little higher than that of the L-w Xception, which indicate that the channel capacity expansion method proposed in this article could widen the network at the cost of relatively small increase in time complexity.

In general, the M-b Xception network can trade an acceptable increase in time complexity for the accuracy higher than the L-w Xception 1.31%. Compared with the Xception, the accuracy of the M-b Xception network is 1.75% higher than that of Xception without increasing the number of network parameters and time complexity.

In addition, the three networks are also trained without data enhancement, and other settings remain unchanged. The results show that the accuracy of Xception is 0.8431, that of L-w Xception is 0.8497, and that of M-b Xception is 0.8540. This shows that the performance improvement comes from the improvement of network structure, and further proves the effectiveness of the proposed method.

2) COMPARISON OF Grad-CAM VISUALIZATION

Grad-CAM (Gradient-weighted Class Activation Mapping) is a visualization technique proposed by Selvaraju R R in 2020, which can produce a localization map to highlighting the important regions in the image [55]. In this article, the Grad-CAM is adopted as the visualization method. The visualization results of some test images with the Xception, the L-w Xception, and the M-b Xception are shown in Fig. 10. It can be seen that in Fig.10, the extracted regions of M-b Xception are more complete and accurate, which are most consistent with the important region of the garbage images. This indicates that the M-b Xception has the strongest ability to extract garbage image features, and then come L-w Xception, and finally Xception. This result is consistent with the accuracy comparison of the three networks.

3) COMPARISON OF TRAINING PROCESS

Fig. 11 shows the comparison results of training process of the M-b Xception network and the Xception network.

From the perspective of network convergence speed, the Xception converges after 105 iterations, while the M-b Xception converges after 82 iterations. The fast convergence speed can save computing resources during training. From the fitness of training loss and verification loss curve, the M-b Xception is better than that of the Xception, which proves that the M-b Xception has a good fitting ability for the TrashNet data set. Therefore, the final verification loss is lower than that of the Xception.

The visualization of the training process can reflect the learning effect of the network through the convergence speed and fitting effect, but it cannot accurately reflect the gap between the network performance. In this experiment, the visualization results of the training process of L-w Xception and the M-b Xception are very similar, so they are not shown separately. In the following experiments, we will quantitatively analyze the performance of these networks through a large number of experimental results. For the L-w Xceotion network and the M-b Xception network, the comparison results we obtained will reflect the key differences between them.

4) COMPARISONS OF ROBUSTNESS

Some studies have shown that neural networks are vulnerable to external interference [56]–[59]. Network models are prone to make false predictions when the ambient light around the object changes, or the object is obscured [60], [61]. For garbage classification, the most common situation in real life is that garbage samples are incomplete or obscured. Therefore, some experiments are carried out to test the robustness of the Xception, the L-w Xception, and the M-b Xception.

A gray square with RGB = (192,192,192) is used to occlude the images of the original test set. 9 new test sets are established according to the different occlusion locations, as shown in Fig.12. The occlusion location of the new test set 1 is the location of number 1, and the occlusion location of test set 2 is the location of number 2, and so on.

Next, the 9 new test sets are used to test the three network models that we have trained. Fig. 13 shows the overall accuracy of the three models on the 9 test sets. It can be seen that

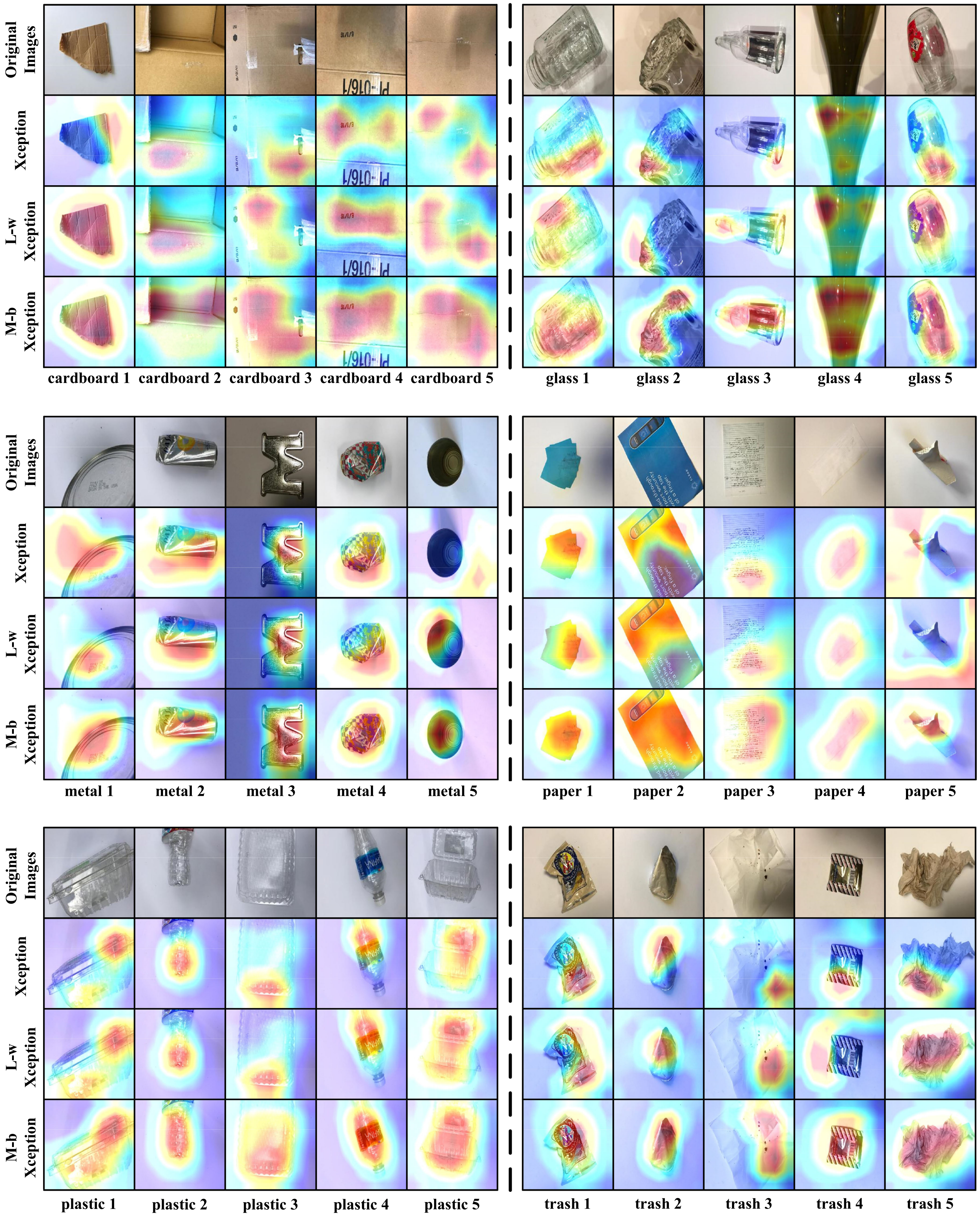


FIGURE 10. The Grad-CAM visualization results of the Xception, the L-w Xception and the M-b Xception.

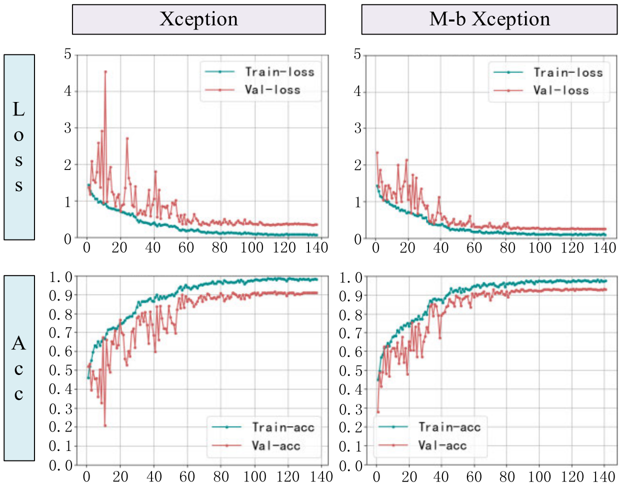


FIGURE 11. Accuracy and loss curves.

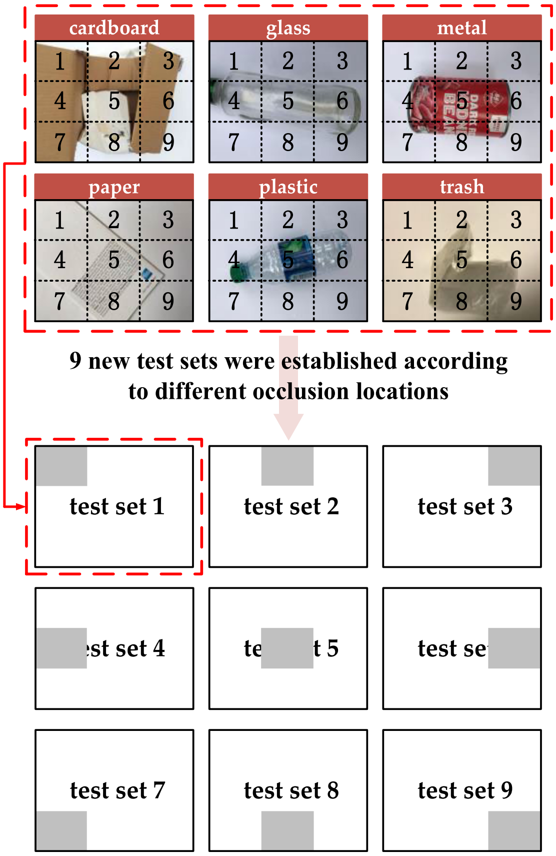
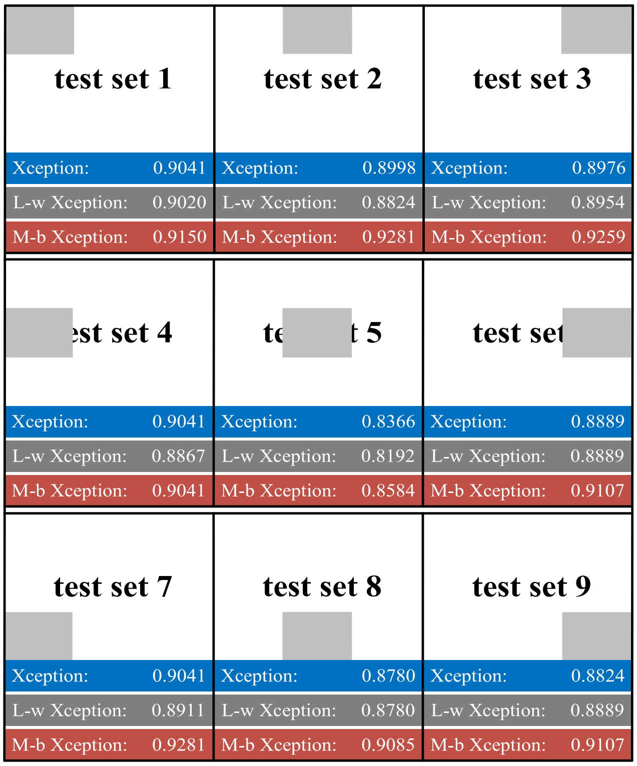


FIGURE 12. Occlude the images in the original test set, and establish 9 new test sets according to the different occlusion locations.

the performance of L-w Xception is the worst, which shows its poor robustness. Although the accuracy of L-w Xception is higher on the original test set than that of the Xception, it does not mean that L-w Xception also has good performance under other conditions, such as occlusion. In practical application, the performance of garbage image classification may be more affected by the robustness of the model. In this experiment, the M-b Xception model shows the best performance as it does on the original image set. This shows that the M-b Xception model has good robustness and further proves its effectiveness.



test set 1	test set 2	test set 3
Xception: 0.9041	Xception: 0.8998	Xception: 0.8976
L-w Xception: 0.9020	L-w Xception: 0.8824	L-w Xception: 0.8954
M-b Xception: 0.9150	M-b Xception: 0.9281	M-b Xception: 0.9259
est set 4	te t 5	test set
Xception: 0.9041	Xception: 0.8366	Xception: 0.8889
L-w Xception: 0.8867	L-w Xception: 0.8192	L-w Xception: 0.8889
M-b Xception: 0.9041	M-b Xception: 0.8584	M-b Xception: 0.9107
test set 7	test set 8	test set 9
Xception: 0.9041	Xception: 0.8780	Xception: 0.8824
L-w Xception: 0.8911	L-w Xception: 0.8780	L-w Xception: 0.8889
M-b Xception: 0.9281	M-b Xception: 0.9085	M-b Xception: 0.9107

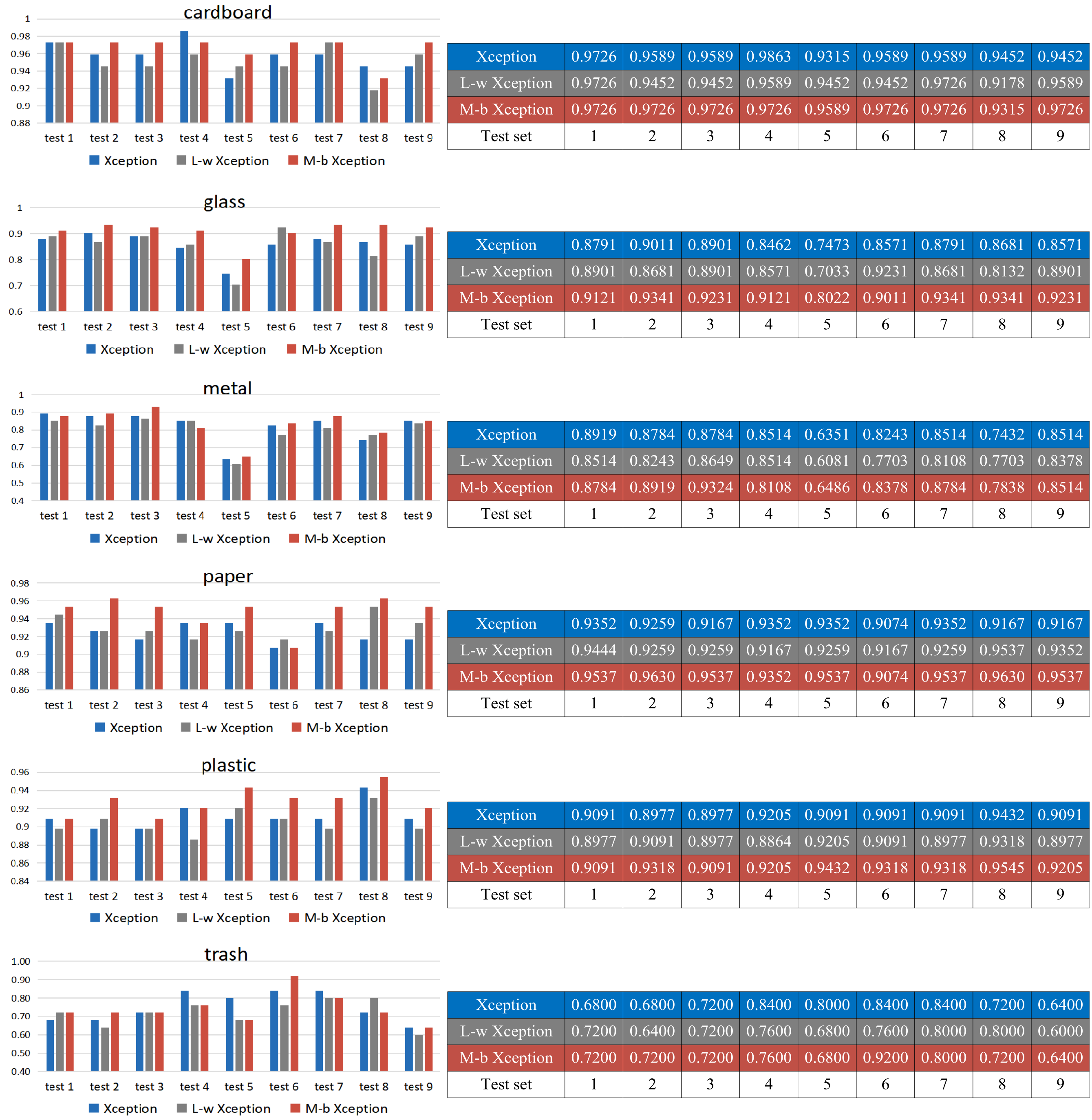
FIGURE 13. Test accuracy of the three network models on the 9 new test image sets.

Fig. 14 shows the single category accuracy of the three models when garbage images are occluded from different positions. The results show that in most cases, the M-b Xception model performs best, followed by the Xception, and finally the L-w Xception. The accuracy of the M-b Xception model in the glass, paper, and plastic categories is significantly higher than that of the other two models. It also has some slight advantages in the cardboard, metal, and trash categories. It should be noted, especially trash, the samples of these categories are messy and the images within the class are quite different, which makes it difficult to achieve significant improvements in accuracy.

The above experiments prove that the proposed M-b Xception model has good robustness. Therefore, the M-b Xception model can provide reliable classification results, even if the test sample is incomplete or occluded.

C. OPTIMIZATION OF CONVOLUTION CHANNEL NUMBER OF THE M-B Xception CORE STRUCTURE

Both width multiplier α in MobileNet and scale factor s in ShuffleNet play the same role, i.e., quickly adjust the number of channels in the convolution layer of the network [62], [63]. This design intends to make it easier for optimizing the



Xception	0.9726	0.9589	0.9589	0.9863	0.9315	0.9589	0.9589	0.9452	0.9452
L-w Xception	0.9726	0.9452	0.9452	0.9589	0.9452	0.9452	0.9726	0.9178	0.9589
M-b Xception	0.9726	0.9726	0.9726	0.9726	0.9589	0.9726	0.9726	0.9315	0.9726
Test set	1	2	3	4	5	6	7	8	9

Xception	0.8791	0.9011	0.8901	0.8462	0.7473	0.8571	0.8791	0.8681	0.8571
L-w Xception	0.8901	0.8681	0.8901	0.8571	0.7033	0.9231	0.8681	0.8132	0.8901
M-b Xception	0.9121	0.9341	0.9231	0.9121	0.8022	0.9011	0.9341	0.9341	0.9231
Test set	1	2	3	4	5	6	7	8	9

Xception	0.8919	0.8784	0.8784	0.8514	0.6351	0.8243	0.8514	0.7432	0.8514
L-w Xception	0.8514	0.8243	0.8649	0.8514	0.6081	0.7703	0.8108	0.7703	0.8378
M-b Xception	0.8784	0.8919	0.9324	0.8108	0.6486	0.8378	0.8784	0.7838	0.8514
Test set	1	2	3	4	5	6	7	8	9

Xception	0.9352	0.9259	0.9167	0.9352	0.9352	0.9074	0.9352	0.9167	0.9167
L-w Xception	0.9444	0.9259	0.9259	0.9167	0.9259	0.9167	0.9259	0.9537	0.9352
M-b Xception	0.9537	0.9630	0.9537	0.9352	0.9537	0.9074	0.9537	0.9630	0.9537
Test set	1	2	3	4	5	6	7	8	9

Xception	0.9091	0.8977	0.8977	0.9205	0.9091	0.9091	0.9091	0.9432	0.9091
L-w Xception	0.8977	0.9091	0.8977	0.8864	0.9205	0.9091	0.8977	0.9318	0.8977
M-b Xception	0.9091	0.9318	0.9091	0.9205	0.9432	0.9318	0.9318	0.9545	0.9205
Test set	1	2	3	4	5	6	7	8	9

Xception	0.6800	0.6800	0.7200	0.8400	0.8000	0.8400	0.8400	0.7200	0.6400
L-w Xception	0.7200	0.6400	0.7200	0.7600	0.6800	0.7600	0.8000	0.8000	0.6000
M-b Xception	0.7200	0.7200	0.7200	0.7600	0.6800	0.9200	0.8000	0.7200	0.6400
Test set	1	2	3	4	5	6	7	8	9

FIGURE 14. **Single category accuracy of the three models when garbage image is occluded from different positions.**

network structure for some data sets. It also reflects that finding the number of convolution channels suitable for the data set used is an important problem in network optimization.

In this article, we change the number of convolution channels of the core structure of M-b Xception network to realize the better performance on TrashNet data set. Fig. 15 shows some typical experimental results. It can be seen that in Fig. 15, when the number of convolution channels in the core structure increased to 896, the network performance is the best, and the accuracy is increased by 1.09%.

However, when the number of convolution channels is further increased to 1024, the accuracy is reduced. The reason is that for a narrow network, increasing the number of convolution channels can improve the network performance [64]. However, when the number of convolution channels can meet the current task, the further increasing will bring negative

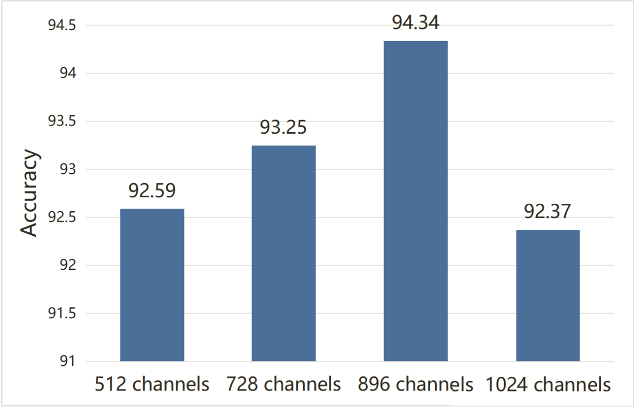


FIGURE 15. Performance comparison of different channel numbers in core structure.

effects. For example, the increase of parameters may make the network relatively difficult to converge, and then affect the network performance [65].

Fig. 16 shows the comparison results of confusion matrix with or without the channel number optimization. The vertical axis shows the real category, and the horizontal axis shows the predicted category. The main diagonal element is the percentage of the number of images identified correctly. It shows that when the channel number of the core structure is set to 896, the three hard-to-predict categories, i.e., glass, metal and trash, accuracy increased by 4%, 2% and 8%, respectively. This has greatly improved the classification balance of the M-b Xception, and reduces the occurrence of unreliable detection for a certain category in practical application.

D. COMPARISON BETWEEN THE M-B Xception AND SOME STATE-OF-THE-ART WORKS

In this section, we compare the M-b Xception to some state-of-the-art works. All of these new works are being done on the TrashNet data set.

1) COMPARISON OF OVERALL ACCURACY

The comparison results of M-b Xception and other related new works are shown in Fig. 17. The methods used for comparison include four machine learning methods, i.e., SVM, XGB, RF, KNN, and 10 deep learning methods. Among them, the fine-tuned DenseNet121 proposed by R.A. Aral *et al.* has a slightly higher accuracy than the method in this article, but this method belongs to transfer learning. It uses pre-trained weights on the ImageNet data set. However, the model in this article is trained in the TrashNet data set with the weight of random initialization, so the comparison of accuracy is not fair enough. R. A. Aral *et al.* also tried to train DenseNet121 on the TrashNet data set with the weight of random initialization, and obtained an accuracy of 0.89, which is 5.34% lower than our accuracy. Therefore, the performance of M-b Xception network proposed by us on TrashNet data set is better than other existing networks. It should be noted that the accuracy of M-b Xception in Fig. 17 is different from that in Table 4. The reason is that the number of convolution channels of M-b Xception is optimized in Section IV.C.

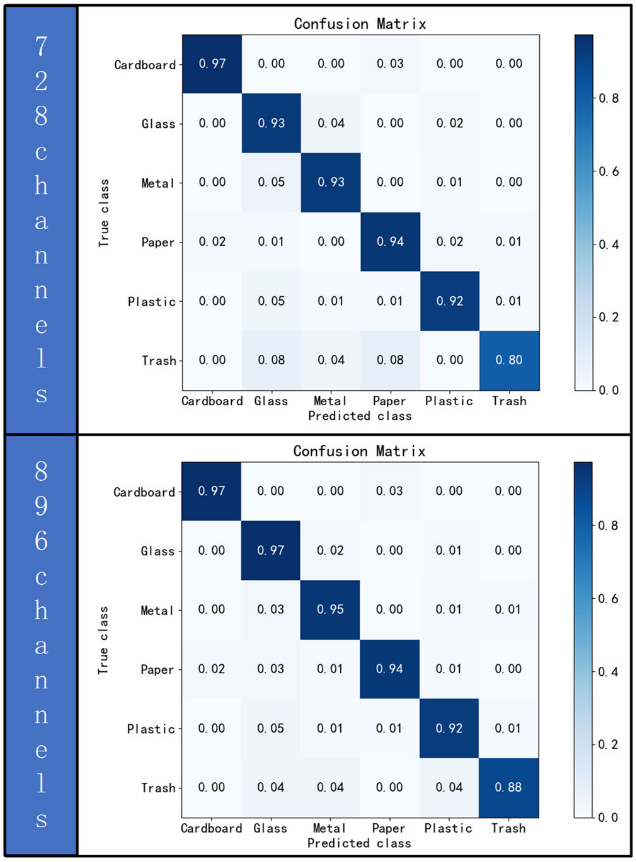


FIGURE 16. Confounding matrix before and after optimization of the number of convolution channels in core structure.

2) COMPARISON OF SINGLE CATEGORY ACCURACY

Following, the confusion matrices of several other methods are converted into some single category accuracy. In practical application, we can't predict which type of garbage needs to be predicted at any given moment. Here we assume that each kind of garbage that needs to be classified appears at the same odds. Therefore, we add the accuracy of each kind of garbage and then divide them by 6 to get the average value, so as to evaluate the accuracy in practical application. The comparison results are listed in Table 5. To compare the overall performance of different methods, the accuracy comparison results of single category are shown in Fig. 18.

In Fig. 18, the other four methods are described as bar and the M-b Xception as broken line. It shows that the average recognition accuracies of the six categories of the M-b Xception is generally higher than that of other methods. The M-b Xception has no significant difference in recognition accuracies of different categories. Compared with other methods, the M-b Xception has obvious advantages in average accuracy of six classes.

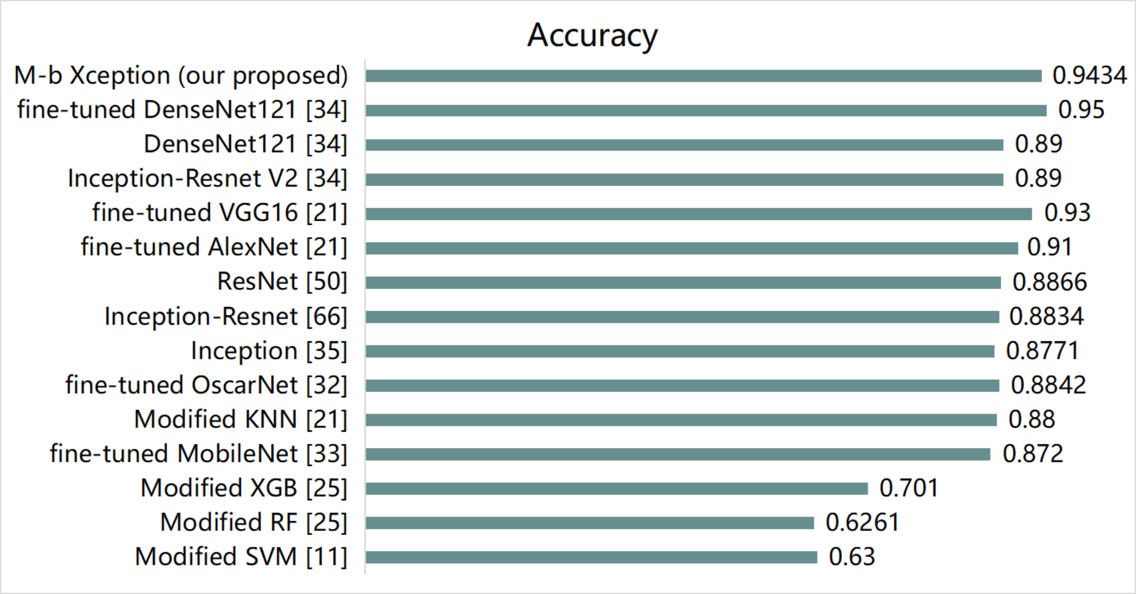


FIGURE 17. The accuracy comparison between the M-b Xception method and some state-of-the-art works.

TABLE 5. The accuracy comparison of single category.

	Cardboard	Glass	Metal	Paper	Plastic	Trash	Average
Inception [35]	0.93	0.86	0.78	0.95	0.83	0.92	0.8783
Inception-ResNet [66]	0.88	0.88	0.98	0.9	0.77	0.69	0.85
ResNet [50]	0.93	0.82	0.9	0.97	0.88	0.83	0.8883
Fine-tuned OscarNet [32]	0.9497	0.7522	0.9007	0.7522	0.4014	0.4059	0.6936
M-b Xception (our proposed)	**0.9726**	**0.967**	**0.9459**	**0.9351**	**0.9204**	**0.88**	**0.9368**

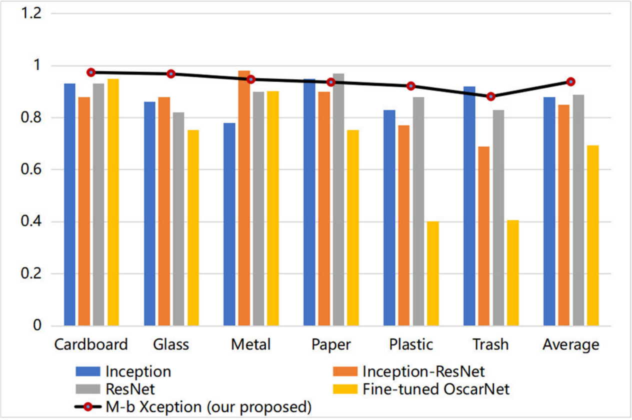


FIGURE 18. The accuracy comparison results of single category.

3) COMPARISON OF F1 SCORE

F1 score is a comprehensive evaluation index to balance accurancy and Recall, which could reflect the comprehensive performance of a model. The F1 score of single category of the above methods is calculated, and the result is shown in Fig. 19. We can see that the M-b Xception has obvious advantages in cardboard, glass, metal, paper and plastic categories, which further proves the effectiveness of this method.

V. CONCLUSION

In this article, a novel network improvement method based on channel expansion is proposed for garbage image classification. It can make full use of feature information at slight additional computational cost. When some common network improvement methods are hardly to work, the proposed method can greatly improve the network performance. Compared with the Xception network, the M-b Xception network has higher accuracy on the TrashNet data set. It also shows good robustness in occlusion experiments. comparing with some related new methods, the M-b Xception can provide higher accuracy and F1-score. It also has more balanced predicting ability, which allows it to be used in practical applications.

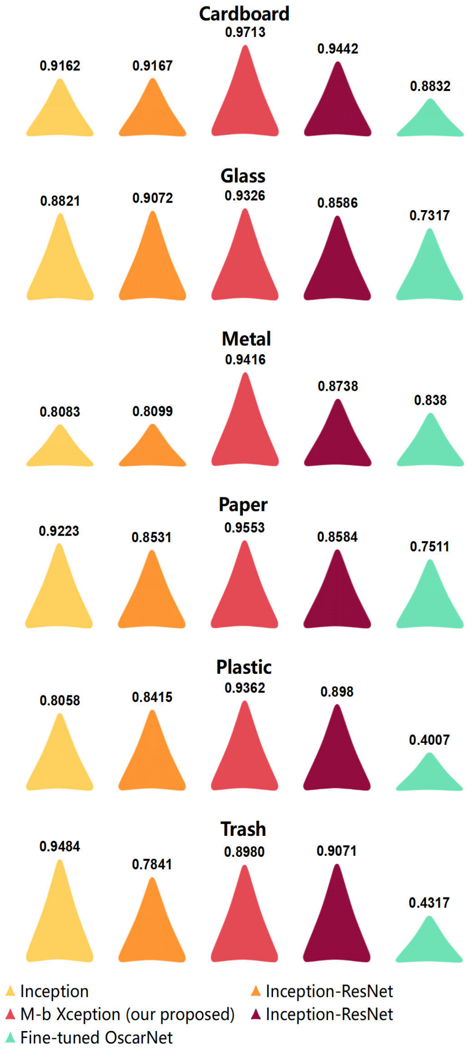


FIGURE 19. The comparison results of F1-score.

In the near future, we will continue to study end-to-end learning systems, and analyze quantitatively the impact of small changes of network structure on classification performance. In addition, it is more pressing to minimize the network volume while maintaining high accuracy. At last but not least, our work will be transplanted into the mobile phone.

REFERENCES

[1] R. Rajamanickam and S. Nagan, "Assessment of comprehensive environmental pollution index of Kurichi industrial cluster, Coimbatore district, Tamil Nadu, India—A case study," *J. Ecol. Eng.*, vol. 19, no. 1, pp. 191–199, 2018.

[2] Z. Deng *et al.*, "AirVis: Visual analytics of air pollution propagation," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 800–810, Jan. 2020.

[3] H. AnvariFar, A. K. Amirkolaie, A. M. Jalali, H. K. Miandare, A. H. Sayed, S. İ. Üçüncü, H. Ouraji, M. Ceci, and N. Romano, "Environmental pollution and toxic substances: Cellular apoptosis as a key parameter in a sensible model like fish," *Aquatic Toxicol.*, vol. 204, pp. 144–159, Nov. 2018.

[4] C. G. Alimba and C. Faggio, "Microplastics in the marine environment: Current trends in environmental pollution and mechanisms of toxicological profile," *Environ. Toxicol. Pharmacol.*, vol. 68, pp. 61–74, May 2019.

[5] J. Huber, W. K. Viscusi, and J. Bell, "Dynamic relationships between social norms and pro-environmental behavior: Evidence from household recycling," *Behav. Public Policy*, vol. 4, no. 1, pp. 1–25, Mar. 2020.

[6] A. Khellal, H. Ma, and Q. Fei, "Convolutional neural network based on extreme learning machine for maritime ships recognition in infrared images," *Sensors*, vol. 18, no. 5, p. 1490, May 2018.

[7] D. de Oliveira and M. Wehrmeister, "Using deep learning and low-cost RGB and thermal cameras to detect pedestrians in aerial images captured by multirotor UAV," *Sensors*, vol. 18, no. 7, p. 2244, Jul. 2018.

[8] J. Enguehard, P. O'Halloran, and A. Gholipour, "Semi-supervised learning with deep embedded clustering for image classification and segmentation," *IEEE Access*, vol. 7, pp. 11093–11104, 2019.

[9] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[10] M. Yan, "Adaptive learning knowledge networks for few-shot learning," *IEEE Access*, vol. 7, pp. 119041–119051, 2019.

[11] M. Yang and G. Thung, "Classification of trash for recyclability status," Mach. Learn., Stanford, CA, USA, Project Rep. CS229, 2016.

[12] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[13] J. Peng, S. Kang, Z. Ning, H. Deng, J. Shen, Y. Xu, J. Zhang, W. Zhao, X. Li, W. Gong, J. Huang, and L. Liu, "Residual convolutional neural network for predicting response of transarterial chemoembolization in hepatocellular carcinoma from CT imaging," *Eur. Radiol.*, vol. 30, no. 1, pp. 413–424, Jul. 2019.

[14] H. Wang, L. Dai, Y. Cai, X. Sun, and L. Chen, "Salient object detection based on multi-scale contrast," *Neural Netw.*, vol. 101, pp. 47–56, May 2018.

[15] S. Yu, X. Li, X. Zhang, and H. Wang, "The OCS-SVM: An objective-cost-sensitive SVM with sample-based misclassification cost invariance," *IEEE Access*, vol. 7, pp. 118931–118942, 2019.

[16] X. Wu, W. Zuo, L. Lin, W. Jia, and D. Zhang, "F-SVM: Combination of feature transformation and SVM learning via convex relaxation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5185–5199, Nov. 2018.

[17] L. Lan, Z. Wang, S. Zhe, W. Cheng, J. Wang, and K. Zhang, "Scaling up kernel SVM on limited resources: A low-rank linearization approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 369–378, Feb. 2019.

[18] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, May 2018.

[19] W. Xing and Y. Bei, "Medical health big data classification based on KNN classification algorithm," *IEEE Access*, vol. 8, pp. 28808–28819, 2020.

[20] B. Tu, J. Wang, X. Kang, G. Zhang, X. Ou, and L. Guo, "KNN-based representation of superpixels for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4032–4047, Nov. 2018.

[21] B. S. Costa, A. C. S. Bernardes, J. V. A. Pereira, V. H. Zampa, V. A. Pereira, G. F. Matos, E. A. Soares, C. L. Soares, and A. F. Silva, "Artificial intelligence in automated sorting in trash recycling," in *Proc. Anais do XV Encontro Nacional de Inteligência Artificial e Computacional. SBC*, Oct. 2018, pp. 198–205.

[22] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection," *IEEE Access*, vol. 7, pp. 180235–180243, 2019.

[23] A. Dapogny, K. Bailly, and S. Dubuisson, "Dynamic pose-robust facial expression recognition by multi-view pairwise conditional random forests," *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 167–181, Apr. 2019.

[24] S. Kim, S. Kwak, and B. C. Ko, "Fast pedestrian detection in surveillance video based on soft target training of shallow random forest," *IEEE Access*, vol. 7, pp. 12415–12426, 2019.

[25] M. Satvilkar, "Image based trash classification using machine learning algorithms for recyclability status," *Image*, vol. 13, pp. 8, Sep. 2018.

[26] Y. Jiang, G. Tong, H. Yin, and N. Xiong, "A pedestrian detection method based on genetic algorithm for optimize XGBoost training parameters," *IEEE Access*, vol. 7, pp. 118310–118321, 2019.

[27] X. Gu, Y. Han, and J. Yu, "A novel lane-changing decision model for autonomous vehicles based on deep autoencoder network and XGBoost," *IEEE Access*, vol. 8, pp. 9846–9863, 2020.

[28] W. Zhang, X. Zhao, and Z. Li, "A comprehensive study of smartphone-based indoor activity recognition via xgboost," *IEEE Access*, vol. 7, pp. 80027–80042, 2019.

[29] R. Tao, S. Zhang, X. Huang, M. Tao, J. Ma, S. Ma, C. Zhang, T. Zhang, F. Tang, J. Lu, C. Shen, and X. Xie, "Magnetocardiography-based ischemic heart disease detection and localization using machine learning methods," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 6, pp. 1658–1667, Jun. 2019.

[30] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-connected covariance network for remote sensing scene classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1461–1474, May 2020.

[31] S. Hussein, P. Kandel, C. W. Bolan, M. B. Wallace, and U. Bagci, "Lung and pancreatic tumor characterization in the deep learning era: Novel supervised and unsupervised learning approaches," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1777–1787, Aug. 2019.

[32] T. Kennedy, "OscarNet: Using transfer learning to classify disposable waste," Deep Learn., Stanford, CA, USA, Project Rep. CS230, 2018.

[33] S. L. Rabano, M. K. Cabatuan, E. Sybingco, E. P. Dadios, and E. J. Calilung, "Common garbage classification using MobileNet," in *Proc. IEEE 10th Int. Conf. Humanoid, Nanotechnol., Inf. Technol., Commun. Control, Environ. Manage. (HNICEM)*, Nov. 2018, pp. 1–4.

[34] R. A. Aral, S. R. Keskin, M. Kaya, and M. Haciömeroglu, "Classification of TrashNet dataset based on deep learning models," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 2058–2062.

[35] V. Ruiz, Á. Sánchez, J. F. Vélez, and B. Raducanu, "Automatic image-based waste classification," in *Proc. Int. Work-Conf. Interplay Between Natural Artif. Comput.*, 2019, pp. 422–431.

[36] F. Ye, H. Xiao, X. Zhao, M. Dong, W. Luo, and W. Min, "Remote sensing image retrieval using convolutional neural network features and weighted distance," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 10, pp. 1535–1539, Oct. 2018.

[37] L. Zhang, J. Liu, B. Zhang, D. Zhang, and C. Zhu, "Deep cascade model-based face recognition: When deep-layered learning meets small data," *IEEE Trans. Image Process.*, vol. 29, pp. 1016–1029, Sep. 2020.

[38] G. Hu, X. Peng, Y. Yang, T. M. Hospedales, and J. Verbeek, "Frankenstein: Learning deep face representations using small data," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 293–303, Jan. 2018.

[39] J. Shi, Z. Li, S. Ying, C. Wang, Q. Liu, Q. Zhang, and P. Yan, "MR image super-resolution via wide residual networks with fixed skip connection," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 3, pp. 1129–1140, May 2019.

[40] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[41] Y. Ding, F. Chen, Y. Zhao, Z. Wu, C. Zhang, and D. Wu, "A stacked multi-connection simple reducing net for brain tumor segmentation," *IEEE Access*, vol. 7, pp. 104011–104024, 2019.

[42] Z. Liu, Y. Chen, B. Chen, L. Zhu, D. Wu, and G. Shen, "Crowd counting method based on convolutional neural network with global density feature," *IEEE Access*, vol. 7, pp. 88789–88798, 2019.

[43] G. F. Montúfar, R. Pascanu, K. Cho, and Y. Bengio, "On the number of linear regions of deep neural networks," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 2924–2932.

[44] Y. Bengio and Y. Lecun, "Scaling learning algorithms towards AI," *Large-Scale Kernel Mach.*, vol. 34, no. 5, pp. 1–41, 2007.

[45] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[46] L. Chen, H. Wang, J. Zhao, D. Papailiopoulos, and P. Koutris, "The effect of network width on the performance of large-batch training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9302–9309.

[47] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*. [Online]. Available: https://arxiv.org/abs/1605.07146

[48] J. Pan, E. Sayrol, X. Giro-i-Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 598–606.

[49] Z. Wu, C. Shen, and A. van den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," *Pattern Recognit.*, vol. 90, pp. 119–133, Jun. 2019.

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

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.

[52] R. A. Al-Falluji, A. A. A. Youssif, and S. K. Guirguis, "Single image super resolution model using learnable weight factor in residual skip connection," *IEEE Access*, vol. 7, pp. 58676–58684, 2019.

[53] T. D. Pham, "Geostatistical simulation of medical images for data augmentation in deep learning," *IEEE Access*, vol. 7, pp. 68752–68763, 2019.

[54] G. Zhong, K. Zhang, H. Wei, Y. Zheng, and J. Dong, "Marginal deep architecture: Stacking feature learning modules to build deep learning models," *IEEE Access*, vol. 7, pp. 30220–30233, 2019.

[55] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[56] I. Evtimov *et al.*, "Robust physical-world attacks on deep learning models," 2017, *arXiv:1707.08945*. [Online]. Available: https://arxiv.org/abs/1707.08945

[57] N. Carlini and D. Wagner "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy*, May 2016, pp. 39–57.

[58] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 427–436.

[59] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, Mar. 2016, pp. 372–387.

[60] F. Cen and G. Wang, "Dictionary representation of deep features for occlusion-robust face recognition," *IEEE Access*, vol. 7, pp. 26595–26605, 2019.

[61] D. K. Shin, M. U. Ahmed, and P. K. Rhee, "Incremental deep learning for robust object detection in unknown cluttered environments," *IEEE Access*, vol. 6, pp. 61748–61760, 2018.

[62] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861

[63] 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.

[64] K. Kawaguchi, J. Huang, and L. P. Kaelbling, "Effect of depth and width on local minima in deep learning," *Neural Comput.*, vol. 31, no. 7, pp. 1462–1498, Jul. 2019.

[65] H. Zhu, Z. An, C. Yang, X. Hu, K. Xu, and Y. Xu, "Rethinking the number of channels for the convolutional neural network," 2019, *arXiv:1909.01861*. [Online]. Available: https://arxiv.org/abs/1909.01861

[66] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4 inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 4278–4284.

CUIPING SHI (Member, IEEE) received the M.S. degree from Yangzhou University, Yangzhou, China, in 2007, and the Ph.D. degree from the Harbin Institute of Technology (HIT), Harbin, China, in 2016. From 2017 to 2019, she held a Postdoctoral Research position at the College of Information and Communications Engineering, Harbin Engineering University, Harbin. She is currently an Associate Professor with the Department of Communication Engineering, Qiqihar University. She has published two academic books about image processing and more than 50 papers in journals and conference proceedings. Her main research interests include image processing, pattern recognition, and machine learning. Her doctoral dissertation won the Nomination Award of Excellent Doctoral Dissertation of HIT, in 2016.

RUIYANG XIA is currently pursuing the bachelor's degree with Qiqihar University, Qiqihar, China. He has applied for two patents of invention. His research interests include digital image processing and machine learning. His research project won the provincial Students Awards.

LIGUO WANG (Member, IEEE) received the M.S. and Ph.D. degrees in signal and information processing from the Harbin Institute of Technology, Harbin, China, in 2002 and 2005, respectively. From 2006 to 2008, he held a Postdoctoral Research position at the College of Information and Communications Engineering, Harbin Engineering University, Harbin, where he is currently a Professor. He has published two books about image processing and more than 130 papers in journals and conference proceedings. His main research interests include image processing and machine learning.