



# A Method for Residual Network Image Classification with Multi-scale Feature Fusion

Guo Ru<sup>1,2</sup>, Peng Sheng<sup>1,2(✉)</sup>, Anyang Tong<sup>1,2</sup>, and Zhenyuan Li<sup>1,2</sup>

<sup>1</sup> School of Artificial Intelligence and Big Data, Hefei University, Hefei, China  
rugu0809@163.com, cppjava@163.com

<sup>2</sup> Hefei Economic and Technological Development Zone, Jinxiu Avenue No.99. Shushan District, Hefei City, Anhui Province, China

**Abstract.** The traditional convolutional neural network (CNN) only focuses on the features of the last layer of the network, ignoring those of other layers. The detailed information of the shallow network can improve the accuracy of classification to some extent. With the deepening of the network, the problems such as gradient disappearance and explosion are obvious. In order to alleviate this problem, a method for residual network image classification with multi-scale feature fusion is proposed in the paper. First of all, the Random Image Cropping and Patching (RICAP), a data augmentation method, is adopted to cut and splice the new training samples from the training set, based on which feature maps of different sizes are obtained by residual modules. Secondly, after reducing the dimension of high-level feature maps, all the feature maps with the same dimension are processed by the multi-scale fusion. Finally, these feature vectors are input into the Softmax classifier for training and classification, and the method of learning rate decay is used in the training process. The experiment results indicate that this method has a good performance on image classification based on the public datasets of MNIST and CIFAR-10.

**Keywords:** Multi-scale feature fusion · Residual network · Data augmentation · Image classification

## 1 Introduction

Image classification is one of the basic tasks of computer vision. Traditional image classification algorithms extract the color, texture and spatial features of images, which perform well in simple tasks while output unsatisfactory results in complex tasks. As a typical representative of deep learning, Deep Convolutional Neural Network (DCNN) has excellent performance in computer vision tasks. Compared with traditional image classification algorithms that manually extract features, DCNN, extracting features from input images via convolution operation, can effectively learn representations from a large number of samples, and has stronger model generalization ability. However, it remains the following problems: (1) As the network becomes more complex and the parameters increase, the risk of over-fitting in the process of training also grows, and the generalization ability becomes poor. (2) In practice, due to insufficient quantity or poor

quality of samples, the training model has poor effect and generalization ability. (3) Most networks only focus on the features of the last layer, ignoring the feature information in different layers.

In view of the above problems, the method we propose in the paper can be divided into three parts: ResNet-18 Network, RICAP [1], and FPN. Firstly, RICAP is used to randomly crop four images from the training dataset, and patch them to generate new training images which are mixed with the class labels of the four images. Secondly, feature maps of different sizes are obtained by successive and fixed down-sampling in ResNet-18, and then the high-level feature maps are downscaled by  $1 \times 1$  convolution, followed by multi-scale fusion of feature maps with the same dimensionality in the high and low layers using bilinear interpolation, so as to improve the representational capacity of the feature vectors. Finally, the fused feature vectors are input into the Softmax classifier for training and classification, and meanwhile the learning rate is decayed by constant segment. The contributions of this paper are as follows:

- We adopt RICAP to randomly select four images from the training samples for cropping, and the cropped images are patched to produce a new training sample. Under the condition of the original label unchanged, features of samples can be changed according to prior knowledge, so that the new sample approximately conforms to the real distribution of the data. In the era of deep learning, the scale and quality of data determine the upper limit of model learning, and therefore large-scale and high-quality data will greatly improve the generalization ability of the model. Data augmentation technique can increase the number of training samples on a limited dataset, thus prevent over-fitting to a certain extent.
- We use ResNet-18 Network and Shortcut Connect to transmit part of the original input information to the next network layer without matrix multiplication and non-linear transformation. This method effectively reduces the difficulty of deep network training, protects the integrity of information, and solves such problems as gradient disappearance or explosion.
- We use Feature Pyramid Network (FPN) as the network structure and improve the network. It is often used in object detection, which can achieve multi-scale detection without increasing the computational load of the model. In this paper, its characteristic of multi-scale feature fusion is applied for image classification with ResNet-18 as the backbone network, thus establishing a feature pyramid with strong semantic information on all scales. In this way, the multi-scale feature fusion is completed by the fusion of high and low level features through up-sampling.

## 2 Related Research

The purpose of image classification [2] is to determine the category to which the image belongs by using some classification algorithms given a pair of input images, and its main process usually includes three steps: image preprocessing, feature extraction and classifier design. As one of the popular research directions in the field of computer vision, image classification based on vision has a wide range of applications, including unmanned driving [3], object detection[4, 5], attitude estimation [6], facial recognition [7], etc. Therefore, the technique has very high value in research and application.

## 2.1 Data Augmentation

Data augmentation has always been a hot spot in the field of image classification, and rich data is an important guarantee to solve computer vision problems. Traditional data augmentation methods are based on a series of known affine transformation and image processing, which can greatly improve the generalization ability of models, reduce the risk of over-fitting, and improve robustness. Random erasing [8] can correctly classify the object even part of it is covered, forcing the network to use partially uncovered data for recognition, which increases the difficulty of training and improves the generalization ability of the network to a certain extent. Cutout [9] is used to randomly select a square area of a fixed size in the image and set the pixel value in this area to 0 or other uniform values. Similar to Random erasing, the latter method allows the network to better utilize the global information of the image, rather than relying on a small set of specific features. Mixup [10] can mix two images to form a new one and also has the function of soft labels which can mix the class labels of two images and enhances the linear expression between training samples. These new data augmentation techniques have been applied to deep CNNs and made a great breakthrough. Compared with Mixup, the method used in this paper is RICAP has three obvious differences: (1) It selects four images to synthesize; (2) It synthesizes images from space through splicing.; (3) It adds the operation of crop before splicing.

## 2.2 Convolutional Neural Network

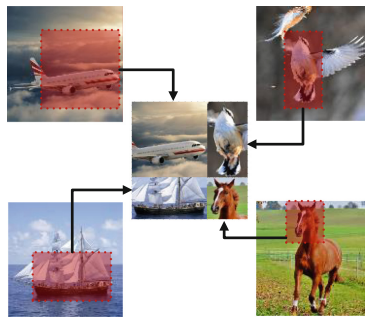
Yann LeCun et al. [11] put forward the LeNet-5 neural network, which is based on the gradient back propagation algorithm for training. By successively alternating connected convolutional and down-sampling layers, the input image can be transformed into a series of feature maps passed to the fully connected neural network, and the final step is to use the Sigmoid function for activation. However, disadvantages of the method include weak generalization ability and small size of training dataset. Krizhevsky et al. [12] again proposed AlexNet neural network, which introduced Dropout technique and ReLU activation function, in order to speed up the gradient descent and alleviate the problem of network over-fitting. Based on the method, Simonyan et al. [13] introduced VGG network, which studied the relationship between performance and depth of CNN and was equipped with excellent generalization ability, but network degradation will be caused with the increase of network depth. To solve the problem, He et al. [14] proposed the ResNet neural network, which can greatly speed up the training of deep network and greatly improve the accuracy of classification. Its greatest contribution is to put forward the residual structure, which can fitting the input and output of each layer through the residual connection, thus retains the characteristic information of the input image to the maximum extent. But the Image classification method based on ResNet neural network only focuses on the features of the last layer of the network, ignoring those of other layers. Although the high-level features are equipped with strong semantic information, the resolution ratio is low and the perception ability of the detailed features is poor. In view of the fact that the detailed information can improve the detection accuracy to a certain extent, this paper proposes a method for residual network image classification

with multi-scale feature fusion, so as to study how to efficiently fuse the high and low-level features, to further improve the representational capacity of the network model, as well as to further improve the accuracy of image classification. Firstly, RICAP is used to randomly crop four images from the training dataset, and patch them to generate new training images which are mixed with the class labels of the four images. Secondly, feature maps of different sizes are obtained by successive and fixed down-sampling, and then the high-level feature maps are downscaled by  $1 \times 1$  convolution, followed by multi-scale fusion of feature maps with the same dimensionality in the high and low layers using bilinear interpolation, so as to improve the representational capacity of the feature vectors. Finally, the fused feature vectors are input into the Softmax classifier for training and classification, and meanwhile the learning rate is decayed by constant segment, which is helpful to the convergence of the algorithm and easier to approach the optimal solution. The experiment results indicate that this method has a good performance on image classification based on the public datasets MNIST and CIFAR-10.

### 3 Methodology

#### 3.1 RICAP

In recent years, with CNN having achieved excellent results in different fields, the main reason is that it contains a large number of parameters capable of fitting a wide variety of data distributions. However, smaller data with too many parameters will suffer from a certain degree of overfitting. Data augmentation by increasing samples of the dataset can effectively alleviate the problem. As shown in Fig. 1, the data augmentation technique RICAP [1] is performed by firstly selecting four images from the training set randomly, and then clipping each image, and finally patching the clipped images into a new training image and inputting into CNN.



**Fig. 1.** The concept of RICAP

As shown in Fig. 2, detailed steps by RICAP are as follows:

- 1) Randomly select four images  $k$  ( $k \in \{1, 2, 3, 4\}$ ) from the training set, and paste them to the top left, top right, bottom left and bottom right respectively.

- 2) With  $I_x$  and  $I_y$  denoting the width and height of the original training images, boundary positions  $(w, h)$  of the four images  $k$  ( $k \in \{1, 2, 3, 4\}$ ) are drawn by obeying the uniform distribution. Therefore,  $(w_k, h_k)$  represents the size of image  $k$ :

$$\begin{aligned}
 w_1 &= w_3 = w, \\
 w_2 &= w_4 = I_x - w, \\
 h_1 &= h_2 = h, \\
 h_3 &= h_4 = I_y - h.
 \end{aligned} \tag{1}$$

- 3) The position of the top left corner of each image  $k$  is determined as  $(x_k, y_k)$ :

$$\begin{aligned}
 x_k &\sim u(0, I_x - w_k), \\
 y_k &\sim u(0, I_y - h_k).
 \end{aligned} \tag{2}$$

- 4) In the process of image classification, the new image combined by the four images  $k$  is blended with the unique thermal encoding class labels  $c_k$ , whose scale is proportional to the area of the new image, thus defining the target label  $c$  as follows:

$$c = \sum_{k \in \{1,2,3,4\}} W_k c_k \text{ for } W_k = \frac{w_k h_k}{I_x I_y} \tag{3}$$

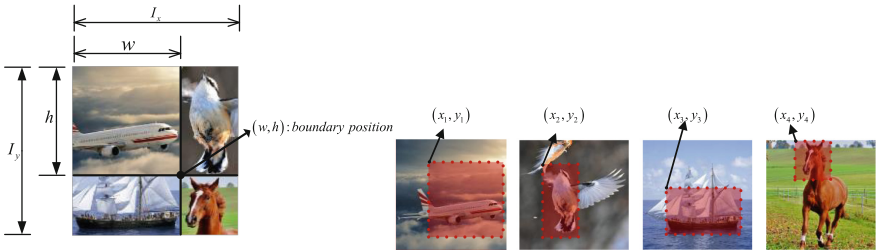


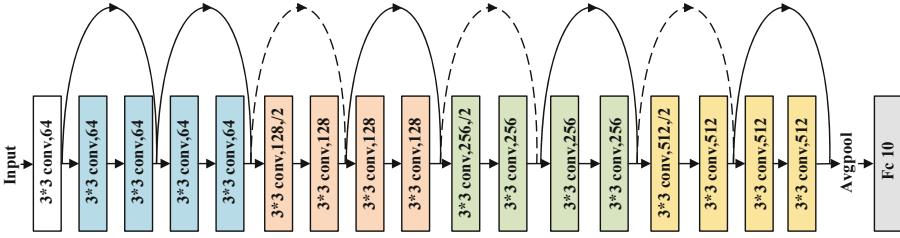
Fig. 2. Graphical Representation of RICAP

### 3.2 ResNet-18

The feature extraction network in the proposed method is combined by ResNet-18 and FPN. The specific parameters of ResNet-18 are shown in Table 1, and its network structure is shown in Fig. 3. Four residual modules denoted as C2 to C5 are included, each of which contains two residual structures. Each structure needs to be convoluted

**Table 1.** Network structure of ResNet-18

Layer name	Conv1	Conv2		Conv3	Conv4	Conv5	
output size	32×32	32×32		16×16	8×8	4×4	1×1
18-layer	7×7,64 Stride 2	7×7 max pool Stride 2	$\begin{bmatrix} 3 \times 3, & 64 \\ 3 \times 3, & 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, & 128 \\ 3 \times 3, & 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, & 256 \\ 3 \times 3, & 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, & 512 \\ 3 \times 3, & 512 \end{bmatrix} \times 2$	Average pool, 10-d fe,softmax

**Fig. 3.** ResNet-18 Structure Diagram

twice, and the size of the convolution kernels is  $3 \times 3$ . For C2 to C5, the number of each convolution kernel is 64, 128, 256, and 512 respectively. The whole network consists of 18 layers, including 17 convolutional layers and 1 pooling layer.

The images processed by RICAP are input to ResNet-18, and the feature maps with half the size of the previous residual module are obtained, thus completing the bottom-up process in the feature pyramid fusion network structure. Subsequently, multi-scale feature fusion will be conducted by FPN.

### 3.3 FPN

FPN changes the number of 512 channels at the highest level of C5 in ResNet-18 to the number of 64 channel with a  $1 \times 1$  convolutional kernel, thus obtaining the new feature map P5. Bilinear interpolation is adopted to up-sample P5 twice, and then connected horizontally to Conv64 of C4 with a  $1 \times 1$  convolutional layer, thus obtaining P4. P3 is obtained by the same way. Figure 4 shows the whole process of multi-scale feature fusion of at high and low levels. The up-sampling of bilinear interpolation is conducted as follows:

Bilinear transformation will happen in the process. Given the value of the four points  $Q_{12} = (x_1, y_2), Q_{21} = (x_2, y_1), Q_{22} = (x_2, y_2)$  at the known function  $f$ , to get the value of interpolation point  $P = (x, y)$  at the unknown function  $f$ , we have:

$$\begin{aligned}
 f(x, y) &\approx \frac{f(Q_{11})}{(x_2 - x_1)(y_2 - y_1)}(x_2 - x)(y_2 - y) + \frac{f(Q_{21})}{(x_2 - x)(y_2 - y)}(x - x_1)(y_2 - y) \\
 &+ \frac{f(Q_{12})}{(x_2 - x_1)(y_2 - y_1)}(x_2 - x)(y - y_1) + \frac{f(Q_{22})}{(x_2 - x)(y_2 - y)}(x - x_1)(y - y_1)
 \end{aligned} \tag{4}$$

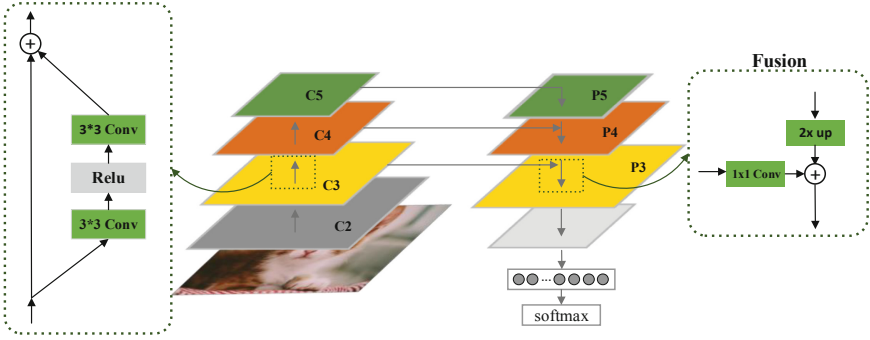


Fig. 4. Feature pyramid network

Finally, the fused feature vectors are input to the Softmax classifier for training and classification, and the learning rate is decayed by constant segments in the process. Afterwards, the Cross Entropy Loss (CEL) function is used to calculate the loss values. The detailed steps are as follows:

- 1) RICAP randomly select four images from the training dataset for clipping and patching, and the new training image, the size of which is specified to  $32 \times 32$ , will be output.
- 2) After inputting the image to ResNet-18, the feature maps of each residual module will be obtained, size of which is half of the previous residual module.
- 3) The convolution kernel of 11 is used to change the number of channels, and the bilinear interpolation is used for up-sampling, thus obtaining the multi-scale fusion features.
- 4) The activation layer is the ReLU function, which has a relatively wide excitation boundary that allows the network to introduce sparsity on its own, and can overcome the problem of gradient disappearance, thus speeding up the training, i.e.:

$$f(x) = \max(0, x) \tag{5}$$

- 5) The pooling layer aims at dimension reduction of features, in order to obtain feature invariance and prevent over-fitting to some extent. In this paper, the average pooling method is adopted.
- 6) The CEL function is used to measure the variability between the true and predicted distributions to calculate the loss values:

$$Loss = - \sum_{i=1}^n p(x_i) \ln(q(x_i)) \tag{6}$$

In this equation,  $p(x_i)$  represents the true probability distribution corresponding to the variable  $p(x_i)$ , while  $q(x_i)$  represents its predicted probability distribution.  $n$  is the number of categories in the sample.

- 7) The learning rate is gradually decayed by constant segments while training, which helps the algorithm to converge and approach the optimal solution more easily.

- 8) After multi-scale feature fusion, the images enter full connection through the averaging pooling layer. The fully connected layer forms the complete image through the weight matrix, which takes the local features extracted by the feature extractor [15], and then finish the process of image classification.

## 4 Experimental Results and Analysis

### 4.1 Experimental Environment

The experimental environment of this paper is: Windows, AMD Ryzen 7 4800H CPU, NVIDIA GeForce GTX 1650 Ti GPU, Pytorch-based deep learning framework, Python programming language. Experiments were carried out on the commonly used datasets of MNIST and CIFAR-10 [16] to prove the effectiveness of the proposed method.

The following recognition accuracy is used as the evaluation criteria:

$$Accuracy = \frac{TP}{Total} \times 100\% \quad (7)$$

In the formula,  $TP$  denotes the number of correctly classified samples and  $Total$  the total number of classified samples.

### 4.2 Introduction to Datasets

**MNIST Dataset.** The MNIST dataset was proposed by the National Institute of Technology and Standards. Its training set consists of handwritten digits from 250 people, 50% of which are census Bureau employees and 50% high school students, and its test set contains the same proportion of handwritten digits. A total of 70,000 grayscale images of 1-channel with the size of  $28 \times 28$  are included, 60,000 of which are used as the training set and 10,000 as the test set.

In the experiment, in order to improve the generalization ability of the model, RICAP is introduced in the training process based on the data augmentation technique of random flipping. In terms of parameter settings, the batch size is set to 64 with the help of SGD optimizer. The learning rate decay strategy is used in the process with the initial learning rate  $\eta = 0.01$ , and  $\eta$  will decrease to 0.1 times of the original learning rate when the accuracy reaches its peak.

**Cifar-10 Dataset.** The cifar-10 dataset is also a more commonly used image classification dataset for recognizing real objects in a small dataset. It contains 60,000 RGB images of size  $32 \times 32$ , of which 50,000 images are the training set and 10,000 images are the test set, divided into 10 categories: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck. Unlike the MNIST dataset, the cifar-10 dataset is more challenging than the MNIST dataset because all the images are in color and the objects to be recognized are more complex.

In the experiment, in order to improve the generalization ability of the model, the RICAP method is introduced in the model training process based on the data enhancement technique using brightness variation, random flip. In terms of parameter settings, the batch size is set to 128, the SGD optimizer is used, and the learning rate decay strategy is used during the training process, with the initial learning rate, which decreases to 0.1 times of the original learning rate when the accuracy reaches the peak.



### 4.3 Comparative Experiment

Based on the MNIST dataset, Table 2 shows the comparison between the experimental results of other algorithms and the method proposed in the paper. It can be seen that the recognition accuracy of ResNet on the test set is 94.0%, CNN-SVM (PSO) 96.0%, VGG-16 97.89%, the method in paper [17] 98.1%, and the method in this paper 99.42%, which shows that the last one has the best result on the MNIST dataset.

**Table 2.** Recognition accuracy on MNIST dataset

Network Models	Accuracy (%)
ResNet	94.0
CNN-SVM(PSO)	96.0
VGG-16	97.89
Paper [17]	98.1
Method of this Paper	99.42

Based on the Cifar-10 dataset, Table 3 demonstrates the comparison between the experimental results of other algorithms as follows and the method of this paper. The recognition accuracy of ResNET-CE on the test set is 90.41%, WideResNet 95.03%, DSENet(depth = 40) 92.52%, and the method in this paper 96.25%, which also shows that the method in this paper has the best effect on the CIFAR-10 dataset.

**Table 3.** Recognition accuracy on CIFAR-10 dataset

Network Models	Accuracy (%)
ResNet-CE	90.41
SqueezeNet	90.26
WideResNet	95.03
DSENet(depth = 40)	92.52
Method of this Paper	96.25

The confusion matrices of the method of this paper on both MNIST and Cifar-10 datasets are shown in Tables 4 and 5 respectively. As the most basic and intuitive means to measure the recognition accuracy of the image classification model, the confusion matrix describes the correspondence between the true category in the horizontal direction and the predicted category in the vertical direction.

**Table 4.** Confusion MATRIX for MNIST dataset based on the method of this paper

MNIST Type	The Method of this Paper									
	0	1	2	3	4	5	6	7	8	9
0	996	0	0	2	0	0	0	0	2	0
1	0	998	0	0	0	0	0	2	0	0
2	0	0	994	0	0	0	0	5	1	0
3	0	0	0	995	0	3	0	0	1	1
4	0	0	0	0	994	0	0	0	1	5
5	0	0	0	6	0	992	2	0	0	0
6	6	0	0	0	1	0	993	0	0	0
7	0	0	5	0	0	0	0	993	0	2
8	3	1	0	0	0	0	0	0	996	0
9	0	0	0	0	7	2	0	0	0	991

**Table 5.** Confusion matrix for Cifar-10 dataset based on the method of this paper

Cifar-10 Type	The Method of this Paper									
	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Airplane	97	1	1	0	0	0	0	0	0	1
Automobile	1	97	0	0	0	0	0	0	0	2
Bird	1	0	98	0	0	0	1	0	0	0
Cat	0	0	1	96	0	2	1	0	0	0
Deer	0	0	2	0	97	0	0	1	0	0
Dog	0	0	0	2	0	96	2	0	0	0
Frog	1	0	2	2	0	2	92	0	1	0
horse	0	0	1	0	2	1	0	96	0	0
Ship	0	1	0	0	0	0	0	0	97	2
Truck	1	3	0	0	0	0	0	0	1	95

The test results of the model in this paper on MNIST dataset are plotted into Table 4, which shows that its accuracy is higher for number 0, 1, 8 and lower for numbers 5, 7, 9. The reason lies in the fact that some numbers are so visually similar that the model is unable to distinguish them accurately. For example, the number 5 may be incorrectly identified as 3 or 6, the number 7 as 2 or 9, and the number 9 as 4 or 5. Similarly, Table 5 shows the result on the Cifar-10 dataset. Due to the high similarity in visual category, the Frog category is easily misidentified as Bird, Cat or Dog, and Truck as

Automobile and Ship. By comparison, the recognition accuracy of the method in this paper is significantly improved on both datasets.

## 5 Conclusion

This paper introduces a method for residual network image classification with multi-scale feature fusion. Based on the residual neural network, the author adopts RICAP and FPN to fuse the features extracted from high and low layers, and meanwhile uses the method of learning rate decay for training. In order to validate the effectiveness of the method, the classification results on MNIST and Cifar-10 datasets serves as the proof. It is suggested that the method of this paper should be applied to more complex and large-scale datasets in the future, such as Cifar-100 and ImageNet. The data augmentation technique RICAP in this method only uses mixed new images in the training process where the original images are not involved, which requires more time to achieve better results. What's more, this method utilizes the fusion of low-level high-resolution and high-level strong semantics, resulting in spending a lot of time training the network. Therefore, further study is needed to improve the method in the near future.

**Acknowledgment.** This work was supported by “the Scientific Research Foundation of HEFEI University” (Grant No. 20RC19) and “Graduate Science Research Project of Anhui Universities”(No.YJS20210564).

## References

1. Takahashi, R., Matsubara, T., Uehara, K.: Data augmentation using random image cropping and patching for deep CNNs. *IEEE Trans. Circuits Syst. Video Technol.* **30**(9), 2917–2931 (2020)
2. Rawat, W., Wang, Z.H.: Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.* **29**(9), 2352–2449 (2017)
3. Radwell, N., Johnson, S.D., Edgar, M.P., Higham, C.F., Murray-Smith, R., Padgett, M.J.: Deep learning optimized single-pixel LiDAR. *Appl. Phys. Lett.* **115**(23), 5 (2019)
4. Qiang, B.H., Chen, R.D., Zhou, M.L., Pang, Y.C., Zhai, Y.J., Yang, M.H.: Convolutional neural networks-based object detection algorithm by jointing semantic segmentation for images. *Sensors* **20**(18), 14 (2020)
5. Hamouda, M., Etabaa, K.S., Bouhleb, M.S.: Smart feature extraction and classification of hyperspectral images based on convolutional neural networks. *IET Image Proc.* **14**(10), 1999–2005 (2020)
6. Wu, C.R., Chen, L., Wu, S.Q.: A novel metric-learning-based method for multi-instance textureless objects' 6D pose estimation. *Applied Sciences-Basel* **11**(22), 12 (2021)
7. Niu, J.-Y., Xie, Z.-H., Li, Y., Cheng, S.-J., Fan, J.-W.: Scale fusion light CNN for hyperspectral face recognition with knowledge distillation and attention mechanism. *Appl. Intell.* **52**(6), 6181–6195 (2021). <https://doi.org/10.1007/s10489-021-02721-8>
8. Zhong, Z., Zheng, L., Kang, G.L., Li, S.Z., Yang, Y.: Assoc Advancement Artificial, I.: Random Erasing Data Augmentation. In: 34th AAAI Conference on Artificial Intelligence. Assoc Advancement Artificial Intelligence. New York (2020)

9. Devries, T., Taylor, G.W.: Improved Regularization of Convolutional Neural Networks with Cutout (2017)
10. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond Empirical Risk Minimization (2017)
11. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE 1998, vol. **86**(11), pp. 2278–2324 (1998)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**(6), 84–90 (2017)
13. Simonyan, K., Zisserman, A.J.C.S.: Very deep convolutional networks for large-scale image recognition. *Computer Science* (2014)
14. He, K., Zhang, X., Ren, S., Sun, J.J.I.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016 (2016)
15. Sainath, T.N., Mohamed, A.R., Kingsbury, B., Ramabhadran, B.: Deep convolutional neural networks for LVCSR. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Ieee. Vancouver, Canada (2013)
16. Krizhevsky, A., Hinton, G.J.H.o.S.A.D.: Learning multiple layers of features from tiny images. *Computer Science* **1**(4) (2009)
17. Zhao, H.H., Liu, H.J.G.C.: Multiple classifiers fusion and CNN feature extraction for handwritten digits recognition. *Granular Computing* (2019)