Retraction Notice

The Editor-in-Chief and the publisher have retracted this article, which was submitted as part of a guest-edited special section. An investigation uncovered evidence of systematic manipulation of the publication process, including compromised peer review. The Editor and publisher no longer have confidence in the results and conclusions of the article.

ZL and JH either did not respond directly or could not be reached.

Design of street art image retrieval system based on virtual simulation technology for the public health environment

Zhiyong Li* **and Jing Hua**
Hebei Academy of Fine Arts, Shijiazhuang, China

Abstract. Creating a good public health environment can improve the public's environmental health literacy level. Therefore, we propose a street art image retrieval system to address the problems of artists' single creation method and the exhaustion of creative inspiration in the process of creating street art images in a public health environment. The system can retrieve the relevant image categories and cultural backgrounds, according to the street art images to be drawn, which reduces the artists' creation burden and helps the public to better appreciate and understand art. In the image classification module, a ResNet34 street art image classification network with a nonlocal attention mechanism is proposed by combining the characteristics of street art images. The experimental results show that the method can achieve accurate classification of art images and can accurately retrieve relevant images based on the input art images, helping artists to better create and improve the public health environment. © *2023 SPIE and IS&T* [DOI: 10.1117/1.JEI.32.6.062506]

Keywords: deep learning; ResNet34; street art portraits; image retrieval; image classification.

Paper 221277SS received Nov. 11, 2022; accepted for publication Dec. 14, 2022; published online Jan. 4, 2023.

1 Introduction

As an emerging technology, virtual simulation technology represented by computer vision is playing an irreplaceable role in industrial manufacturing, entertainment, real estate, education, and other industries.[1] The application of computer vision technology to the field of art portrait creation can break through the dimensional limitations that exist in traditional art design and has great potential for development.[2] As a cultural carrier in the public health environment, street art portraits occupy a very important position in the urban humanistic landscape, but at present, when artists create street art portraits, they often encounter difficulties, such as the single creation method and exhaustion of creative inspiration, and are unable to create street art portraits with more humanistic value.[3] Therefore, using computer vision technology to classify and retrieve street art portraits can bring more inspiration for artists' creations, enhance the expressiveness of art portraits, improve the level of public environmental health literacy, and play an irreplaceable role in the construction and development of the public health environment.[4]

Extracting features from images is a key step for traditional image classification methods, and researchers generally need to design feature extraction methods tailored to the features of the images in the classification task. For example, scale invariant feature transform (SIFT)[5] extracts object features for object recognition by taking advantage of the presence of extreme points in the image, whether far or near and blurred or clear, and the LBP operator[6] extracts the edge information of a face for face recognition by taking advantage of the relationship between the gradient of a pixel and the surrounding pixels. The HOG feature[7] exploits the change in directional gradient at the junction of the target and background in an image to extract the edge gradient features of the human body in the image for the pedestrian detection task.

Nowadays, convolutional neural networks have been able to bring amazing accuracy improvements to various image classification tasks due to the rapid development of deep learning techniques in computer vision that bring new life to increasingly complex image

*Address all correspondence to Zhiyong Li, zhiyong1903@163.com

1017-9909/2023/$28.00 © 2023 SPIE and IS&T

classification tasks.[8] For example, VGG[9] deepens the network with small convolutional kernels and improves the performance, and ResNet and DenseNet enable the network to perform tens of thousands of iterations to obtain even better image recognition than humans using residual connections and dense connections, respectively.[10] Although the above methods improve the accuracy of the network for image classification to some extent, the number of parameters in the network is too large, making it unable to have better robustness or maintain a high retrieval speed.

We aim to solve the problems of poor network robustness and slow retrieval speed of convolutional neural networks in classifying and retrieving art images while further improving the accuracy of classifying street art portraits in a public health environment, providing artists with a more creative and imaginative creative platform, free from material constraints, such as space and materials, and creating a better public health environment for the public. In this paper, we propose a classification and retrieval system for art portraits based on improved ResNet34. Experimenting on the collected street art portrait image dataset, the improved ResNet network in this paper can better identify the categories of art portraits and related cultural elements, achieving over 95% classification accuracy and 98% retrieval rate to help people develop good environmental health literacy.

2 Related Research Works

The classification of street art portraits in public health settings is an important research component. The introduction of convolutional neural networks in deep learning techniques in art portraits can help artists filter cultural elements in images, reduce the stress that artists encounter when making art portraits, and improve public health literacy. An increasing number of researchers are working on classifying street art images in public health settings through computer vision techniques.

Ever since the LeNet-5 model was proposed,[11] which used stacked convolutional and pooling layers to extract features from input handwritten character images, allowing a significant improvement in the classification accuracy of handwritten characters, the convolutional neural network approach has brought a new idea to the bottlenecked image classification task. AlexNet[12] added three convolutional layers to LeNet-5, turning it into a deeper and larger model and achieving classification performance that was not possible with traditional methods. Later, VGG increased the depth of the network to 19 layers, further improving the classification ability of the network. However, the brutal deepening of the network layers made it difficult to train the network, and the image classification task was again bottlenecked. DenseNet incorporates the idea of feature reuse from ResNet and proposes a dense connection that connects each layer, further enhancing the network's ability to dig deeper into image information. From LeNet-5 to DenseNet, the deepening of the number of layers of the network brings rapid growth in classification performance, and there is a positive correlation between the number of layers of the network and its generalization ability. However, when the number of model layers increases, the corresponding training difficulty also increases significantly, which is mainly related to the gradient dispersion phenomenon. The gradient information is passed from the last layer of the network layer by layer. In the case of a large depth, the gradient may be close to 0, and it does not meet the application requirements on various hardware devices, so the number of layers should be controlled appropriately to avoid the serious gradient dispersion phenomenon.

To solve the problems of training difficulties and gradient dispersion due to the deep layers of the network, ResNet34[13] introduced a skip connection mechanism to make it possible to fall back to a shallow neural network, that is, the network could fall back to a shallow substate under certain conditions.[14]

However, for street art portraits in public environments, sometimes the target size of the portrait that represents a specific art element is small, which increases the difficulty of network recognition. Some recent studies have exploited the human-specific visual mechanism so that the network can gradually deepen its focus on these small-sized traditional cultural elements during the training process. Nonlocal networks[15] focus on the information that is important in the whole feature map by calculating the similarity between pixels, and this approach effectively improves

the network's ability to capture features. Some researchers have since started to consider this approach to capture features. DANet[16] enhances the features to be focused on by considering the dependencies between the correlation of each location on the feature map and the channel mapping using nonlocal learning. ANLNet[17] samples are obtained from high-level features using a pyramidal pooling of the null space with four different expansion rates, respectively, and the low-level features are convolved after being used, effectively reducing the cost of collecting global contextual information by the network. Similarly, to reduce the cost of collecting contextual information by the network, CCNet[18] focuses only on the relationship between pixels on the feature map and pixels in the same row and column during one computation of attention and obtains the relationship of the full image pixels by two computations, obtaining better results while effectively reducing the computational cost.

3 Improved Street Art Portrait Retrieval System

The processing flow of the proposed street art portrait retrieval system is shown in Fig. 1. First, the ResNet34 network with improvements is properly trained so that the parameters of the network are learned, and the Softmax classifier that meets the requirements is obtained. Then the feature extraction operation is performed according to the requirements; the corresponding image features are obtained; and based on these, the feature library is further determined. After processing the features and categories of the input image, the similarity measure is performed with the features of the image library, and the corresponding retrieval results are determined after sorting. The corresponding process is as follows.

3.1 *ResNet34 Street Art Portrait Classification Network*

The CNN network structure used in this paper is a 34-layer deep residual network ResNet34 with an input size of 32×32. The connection line between every two convolutional layers is a skip connection mechanism that gives the neural network a fallback capability. The last layer is a fully connected layer with a Softmax classifier added. The specific network structure design is shown in Fig. 2. Because the classification task in this paper has four classifications, the network

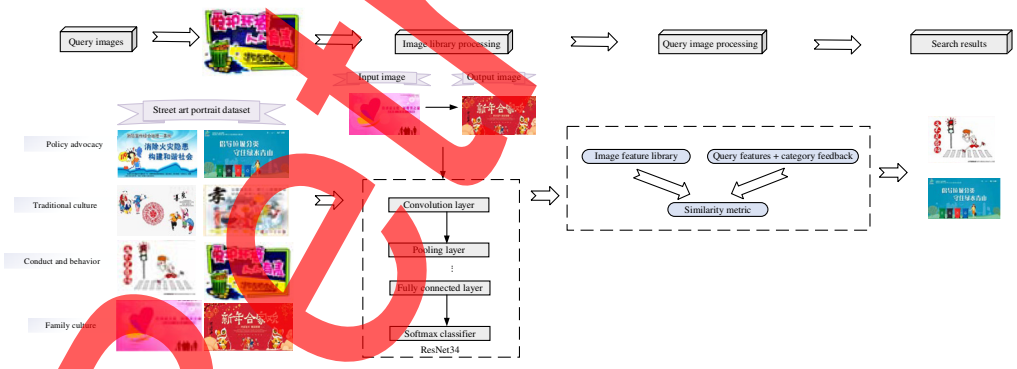

Fig. 1 Flowchart of search processing.

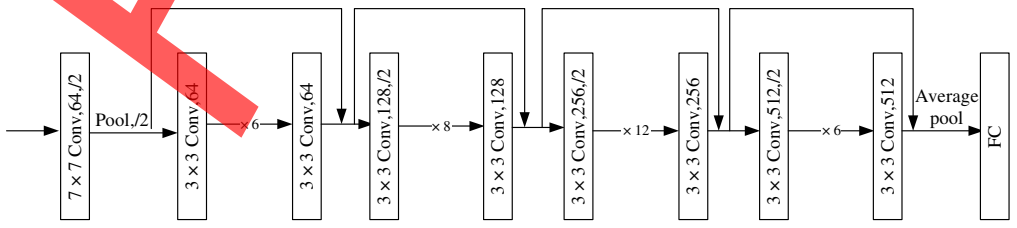

Fig. 2 Structure of the ResNet34 network.

structure is fine-tuned to change the final fully connected layer output to 4. Considering that the residual module in the ResNet34 network has only a single convolutional kernel, it cannot obtain feature information of the data from multiple scales and thus cannot obtain richer input features. The nonlocal attention mechanism and dropout layer are not shown separately but are integrated into the convolutional layer. The pooling layer operates by sampling the input feature maps, thus reducing the number of connected units in adjacent convolutional layers to support the subsequent computational analysis. In this process, a high proportion of pooling types, such as mean pooling, is used. In this paper, the maximum pooling method is chosen to consider the requirements of related problems, to significantly reduce the impact of the convolutional layer parameter errors, and to preserve the texture information adequately. The output of the fully connected layer is appropriately preserved in the specific implementation for building the image feature library.

The ResNet34 network selected in this paper is mainly carried out through the residual block when performing layer rollback, and the specific structure of the residual block is shown in Fig. 3.

The input X of the residual block is processed by the convolutional layer to obtain A. The C obtained by summing A with the original input X is called the residual block.

The most direct way to improve the performance of the network is to increase the depth and width of the network, i.e., the number of convolutional units per layer, so this paper proposes a multiscale residual block based on the residual block, which combines the bottleneck[19] residual block and the convolution residual block[20] to further increase the "width" to improve the model performance. The specific design of the residual module is shown in Fig. 4.

The proposed nonlocal attention mechanism draws its lessons from the idea of nonlocal mean filtering. Nonlocal mean filtering first constructs a region in the image and then compares the proximity (i.e., Euclidean distance) of neighboring regions to this region. Those with higher proximity are given a larger weight, whereas those with lower proximity are given a smaller weight, which highlights their similarity and eliminates their differences. Nonlocal attention mechanisms can ignore irrelevant information in the feature map and focus on point information, extract more semantic information, are computationally efficient, and are easily embedded in various network structures. Therefore, in this paper, the nonlocal attention mechanism is effectively integrated with the improved ResNet34 network structure to effectively improve the classification accuracy of the network. The specific structure of the nonlocal attention mechanism is shown in Fig. 5.

The nonlocal attention mechanism takes the feature maps X and Y as input, first adds a global average pooling layer with 1×1 convolution before X to obtain global feature information, and then performs linear mapping of X and Y to obtain the Q, K, V features. Through the reshape operation, the dimensions of the above three features except the number of channels are combined, and then the matrix dot product is performed on Q and K. The self-correlation in the features is calculated by the reshape operation, and the relationship of each pixel in each frame

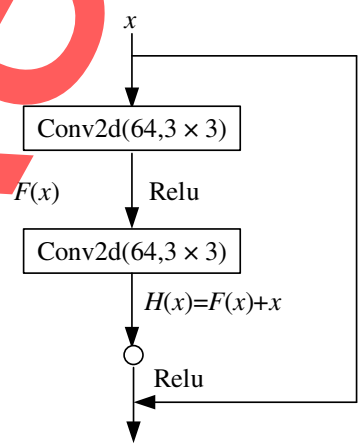

Fig. 3 Residual module.

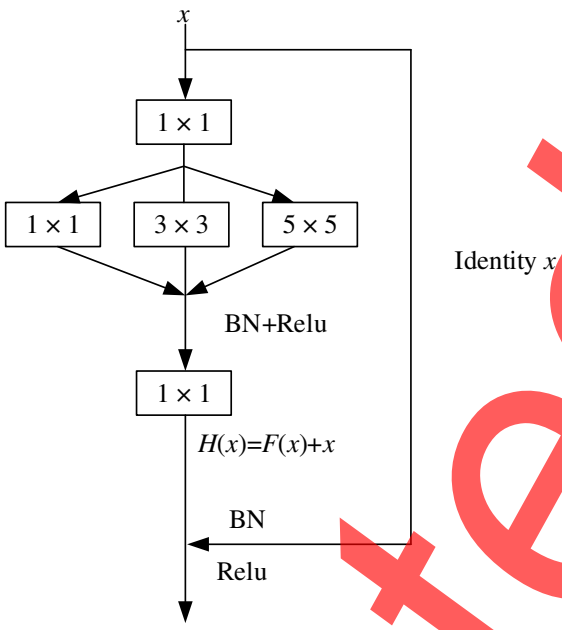

Fig. 4 Improved multiscale residual module.

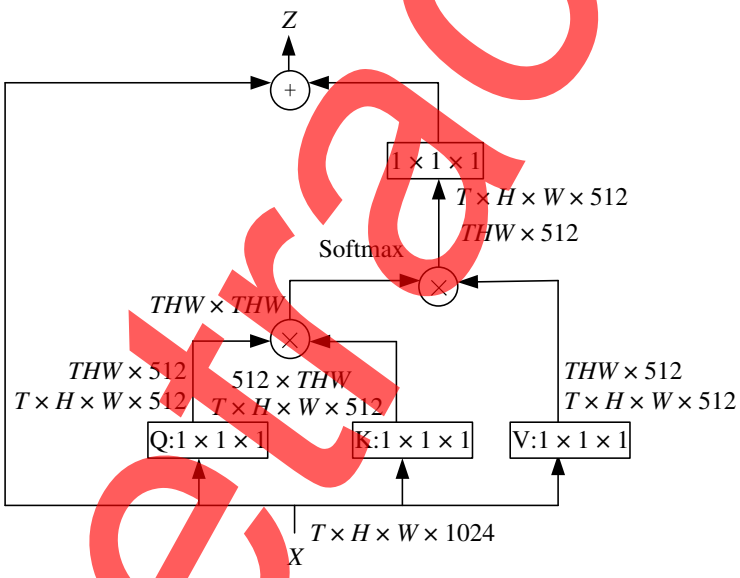

Fig. 5 Nonlocal attention mechanism.

to all pixels in all other frames is obtained. The Softmax operation is performed on the self-correlation features to obtain the weights with the value range of [0 to 1], i.e., the required self-attention coefficients. Finally, the self-attention coefficients are multiplied back into the feature matrix V, and then the 1×1 convolution is performed with the original. Finally, the 1×1 convolution is done with the original input feature map A for residual operation to obtain the output Z.

The dropout layer is introduced in the network design implementation to randomly disconnect the neural network, significantly reducing the number of parameters in the model during training and supporting the speed of operation. The dropout layer restores all connections during the testing process, thus ensuring the best performance of the network during testing. The principle is that, when training is performed, each input sample is updated in the weights, and a part of the nodes in the hidden layer appears through the threshold set, so the weights do not depend

on the interaction between the nodes and the application performance of the system is improved, which can be based on the prevention of overfitting.

3.2 *Similarity Metric*

In this analysis process, this paper measures the similarity between two street art portraits based on Euclidean distance.[21] x and y correspond to the feature vectors of both, and the corresponding Euclidean distance expressions are as follows:

$$D(x,y) = \left(\sum_{i=1}^{n} \|x_i - y_i\|^2\right)^{1/2}. \tag{1}$$

For the queried street art portraits, the relevant image features are determined by processing based on the trained convolutional neural network, the Softmax classifier is used for classification, and the category of portraits is determined. Then the feature maps of portraits of the same category are retrieved, and the Euclidean distance between the two is calculated to achieve portrait retrieval. After sorting the results according to the size of the feature distance, the retrieval results are returned in order. After the search is completed, the system displays the category of the street art portrait and the related cultural knowledge on its own.

4 Experiment

The experiments are conducted in Python on Windows 10, using the GPU version of the deep learning framework TensorFlow 1.4.0, and the GPU provided by CUDA for acceleration when using the CPU for training. The processor is Intel i7-10875H, the memory is 15 GB, and the graphics card is NVIDIA GeForce RTX 2070s.

To test the superiority of the improved ResNet34 network, 16,000 images from the publicly available PASCAL VOC2007 and VOC2012 datasets were selected to train the network for 150 iterations, and 4953 images from the VOC2007 test set were used for testing.

ResNet18,[22] ResNet34,[23] ResNet152,[24] and the improved ResNet34 network were compared in terms of accuracy, recall, training time (s) per step, and evaluation test set consumption time (s), respectively. The comparison results are shown in Table 1, and the accuracy and recall curves of the improved ResNet34 network during training are shown in Fig. 6.

As can be seen from Table 1, compared with the original ResNet34 network, the improved ResNet34 network achieves the best results in terms of accuracy and recall because the improved residual module extracts more multiscale information, and the introduction of the attention mechanism also helps the network to extract better quality features. Although the training time per step and the evaluation test set consumption time are increased compared with the ResNet18 network, the accuracy and recall of the ResNet18 network are too low to meet the demand of image classification.

As can be seen from Fig. 6, the improved ResNet34 network converges quickly and has good robustness.

Table 1 Comparison of the performance of several networks.

Network name	ResNet18	ResNet34	ResNet152	Proposed method
Accuracy	0.683	0.768	0.802	0.854
Recall	0.796	0.874	0.911	0.962
Training time per step (s)	0.280	0.310	0.498	0.300
Evaluation test set consumption time (s)	199.70	235.20	330.00	223.40

Fig. 6 Accuracy and recall graphs of the improved ResNet34 network.

4.1 *Comparison of Search Results*

To verify the effectiveness of the proposed system for the classification and retrieval of street art portraits, this paper crawled a dataset containing 1105 street art portraits from several domestic news websites and microblogs and expanded it to 4420 images using data enhancement. The category division and the number of types of portraits are shown in Table 2.

There are many different metrics to evaluate the performance of image retrieval approaches, and in this paper, capability of precision and recall precision are used to evaluate the processing retrieval results during the study.

The correlation network of Fig. 2 was used for training on the street art portrait dataset, and after 150 iterations of operation, the network was found to be more than 95% accurate in classifying street art portraits. The corresponding classification accuracy curves are shown in Fig. 7.

From Fig. 7, the improved ResNet34 network is in a steadily increasing stage of accuracy overall from 0 to 40 iterations. At 40 to 150 iterations, the accuracy curve remains stable. Therefore, the proposed model has good convergence speed and generalization ability.

To verify the improvement effect of the proposed multiscale residual module and nonlocal attention mechanism on the improved ResNet34 network, 20 images were randomly selected from the dataset for examination, and the corresponding returned images were set to 20, 40, 60, 80, and 100 and their corresponding retrieval accuracies were compared with that of the original ResNet34 network. The results are shown in Fig. 8.

From the above figure, the improved ResNet34 network has a 100% image retrieval rate when the returned images are 20 to 60. The retrieval rates are 4% and 3.6% higher than the original ResNet34 network when the returned images are 80 and 100, respectively. Therefore, the improved ResNet34 network has a good retrieval performance.

Table 2 Classification of street art portrait categories and number of images in each category.

Street art portrait categories	Training set	Test set	Total
Policy advocacy	400	390	1290
Traditional culture	1130	400	1530
Conduct and behavior	660	240	900
Family culture	510	190	700

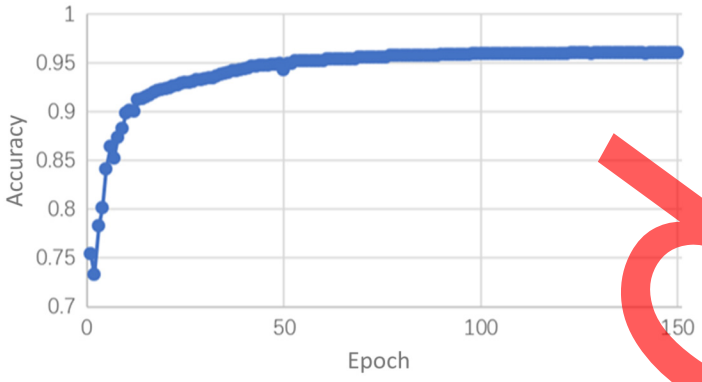

Fig. 7 Classification accuracy curve.

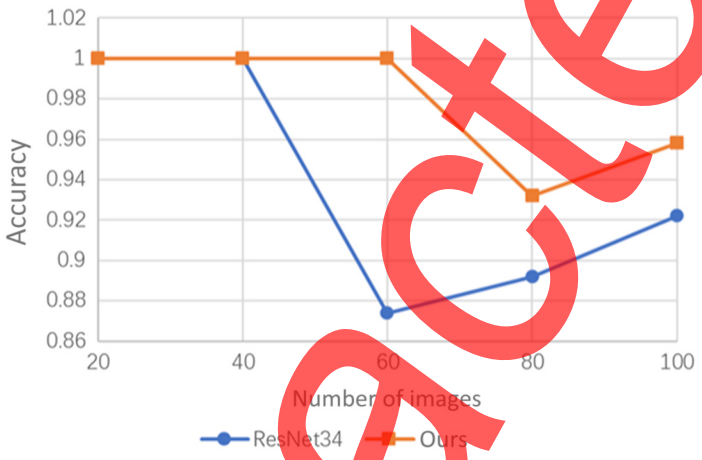

Fig. 8 Street art portrait dataset retrieval accuracy results.

4.2 *Model Analysis*

The model architecture of the convolutional neural network directly determines the accuracy of the classification results and thus the accuracy of the retrieval results. Therefore, the classification study was conducted by three other network architectures in the present study, and a comparative analysis was done to process the obtained results, The correlation between the network structure models and the classification accuracy under the corresponding conditions is shown in Table 3.

As can be seen from Table 3, the proposed network outperforms AlexNet and ResNet in terms of training time and classification accuracy.[18] This is because the skip connection

Table 3 Network knot model classification accuracy comparison chart.

Network name	AlexNet	Vgg16	ResNet18	Proposed in the present study
Policy advocacy	0.742	0.842	0.928	0.949
Traditional culture	0.753	0.874	0.937	0.961
Conduct and behavior	0.740	0.869	0.931	0.955
Family culture	0.769	0.903	0.940	0.967
Accuracy	0.751	0.872	0.934	0.958
MB	42	49	60	19

mechanism introduced in this paper greatly reduces the complexity and computational loss of the network and improves the reuse rate and transfer efficiency of the art portrait features. In terms of the classification accuracy compared with these three algorithms, it improves 21.3%, 9.2%, and 3%, respectively. This is because the proposed multiscale residual module effectively widens the "depth" of the network, effectively integrates the feature layers at different scales, and extracts more feature information about the targets in the street art portraits. Therefore, the improved ResNet34 network can achieve real-time and accurate art portrait classification and retrieval tasks.

5 Conclusion

In this paper, we proposed a street art portrait retrieval technique for the ResNet34 network with a nonlocal attention mechanism. First, the model parameters of the convolutional neural network were properly trained, and then the trained network was used to extract features from the art images and build a library of image features based on certain recognition requirements. Finally, a SoftMax classifier was used to classify the query image and perform retrieval analysis within the class, and the corresponding retrieved results were obtained after the similarity metric. The experiments showed that the proposed ResNet34 network-based street art painting method performed well on the self-constructed street art painting dataset, and the classification accuracy of the network model was over 95%. When the returned image was 100, the average retrieval rate of the selected four categories was 98%. Compared with other network structures, the improved ResNet34 network retrieval performance is good and can meet the requirements related to street art portrait retrieval, and the corresponding generalization ability also reaches a high level, broadening the creative thinking of artists, improving the public understanding and cognition of art portraits, and further highlighting the street art. The humanistic value of art portraits is further highlighted. In future work, we will investigate further improvements of the classification performance and stability while reducing the number of network parameters to better empower the field of art image classification.

Acknowledgments

This work received no funding. The authors declare that there are no conflicts of interest.

Code, Data, and Materials Availability

The dataset can be accessed upon request.

References

1. D. Cazzato et al., "A survey of computer vision methods for 2D object detection from unmanned aerial vehicles," *J. Imaging* **6**(8), 78 (2020).
2. R. Matan, B. Ofer, and O. Gal, "An end-to-end computer vision methodology for quantitative metallography," *Sci. Rep.* **12**(1), 4776 (2022).
3. L. Weisheng and C. Junjie, "Computer vision for solid waste sorting: a critical review of academic research," *Waste Manag.* **142**, 29–43 (2022).
4. Z. Pei, "Research on the application of virtual reality technology in public art design and display," *Ind. Des.* **2019**(10), 136–137 (2019).
5. D. G. Low, "Distinctive image features from scale-invariant key-points," *Int. J. Comput. Vis.* **60**(2), 91–110 (2004).
6. C. Cheng et al., "Pavement crack detection and classification based on fusion feature of LBP and PCA with SVM," *Int. J. Pavement Eng.* **23**(9), 3274–3283 (2022).
7. L. Yang and Z. Wei, "A novel SVM network using HOG feature for prohibition traffic sign recognition," *Wireless Commun. Mob. Comput.* **2022**, 6942940 (2022).
8. S. D. Thepade et al., "Face presentation attack identification optimization with adjusting convolution blocks in VGG networks," *Intell. Syst. Appl.* **16**, 200107 (2022).

9. Z. Wei et al., "Hybrid ResNet based on joint basic and attention modules for long-tailed classification," *Int. J. Approx. Reason.* **150**, 83–97 (2022).
10. A. Saleh, A. Nasir, and S. Muhammad, "Coronavirus disease (COVID-19) detection using X-ray images and enhanced DenseNet," *Appl. Soft Comput.* **110**, 107645 (2021).
11. S. Yongyi et al., "A new hydrogen sensor fault diagnosis method based on transfer learning with LeNet-5," *Front. Neurorob.* **15**, 664135 (2021).
12. G. Heng et al., "A novel fault diagnosis method of wind turbine bearings based on compressed sensing and AlexNet," *Meas. Sci. Technol.* **33**(11), 115011 (2022).
13. K. He et al., "Deep residual leaning for image recognition," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 770–778 (2016).
14. R. C. Gerum et al., "Sparsity through evolutionary pruning prevents neuronal networks from overfitting," *Neural Netw.* **128**, 305–312 (2020).
15. T. Meral and T. Patrizia, "Topology optimization of scale-dependent non-local plates," *Struct. Multidiscip. Optim.* **65**(9), 248 (2022).
16. J. Fu et al., "Dual attention network for scene segmentation," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit. (CVPR)*, IEEE (2020).
17. Z. Zhu et al., "Asymmetric non-local neural networks for semantic segmentation," in *IEEE/CVF Int. Conf. Comput. Vision (ICCV)* (2019).
18. Z. Huang et al., "CCNet: criss-cross attention for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* (2020).
19. T. Y. Lin et al., "Focal loss for dense object detection," in *IEEE Int. Conf. Comput. Vision (ICCV)*, pp. 2999–3007 (2017).
20. X. Li et al., "Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection," *Adv. Neural Inf. Process. Syst.* **33**, 21002–21012 (2020).
21. N. Bharatha Devi, "Satellite image retrieval of random forest (rf-PNN) based probablistic neural network," *Earth Sci. Inf.* **15**, 941–949 (2022).
22. L. Yan, S. GuoRong, and C. ShuXaing, "Magnetic resonance image diagnosis of femoral head necrosis based on ResNet18 network," *Comput. Methods Programs Biomed.* **208**, 106254 (2021).
23. Z. Qingbin, G. Senzhong, and Z. Liangyu, "Human–computer interaction based health diagnostics using ResNet34 for tongue image classification," *Comput. Methods Programs Biomed.* **226**, 107096 (2022).
24. M. Raouia and H. Mariem, "CADNet157 model: fine-tuned ResNet152 model for breast cancer diagnosis from mammography images," *Neural Comput. Appl.* **34**(24), 22023–22046 (2022).

Zhiyong Li obtained his master's degree in fine arts from the PLA Academy of Arts in 2015. He is now an associate professor of the Plastic Arts School of Hebei Academy of Fine Arts. His main research interests are painting art, traditional culture, artistic creation, etc.

Jing Hua received her bachelor's degree in Chinese from Hebei University in 2005 and is now an associate professor of the Basic Department of Hebei Academy of Fine Arts. Her main research interests are aesthetics, traditional culture, art theory, etc.