Heliyon 10 (2024) e27178

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Style classification of media painting images by integrating ResNet and attention mechanism

Xinyun Zhang [a,*], Tao Ding [b]

[a] *The University of York, School of Arts and Creative Technologies(2013-2014), York, YO10 5DD, United Kingdom*
[b] *1-19 Torrington Place. University College London, Gower Street, London, WC1E 6BT, United Kingdom*

ARTICLE INFO

Keywords:
Media painting images
Classification model
Attention mechanism
Convolutional neural network

ABSTRACT

The progress of deep learning technology has made image classification an important application field. Image style classification is a complex task involving the recognition of the whole picture, including the recognition of salient features and detailed features. This study is based on the ResNet algorithm and has improved its Resnet 50 version with excellent performance. In the model architecture, we introduce blur pool operation and replace the traditional Relu function with Celu activation function. In addition, the triplet attention mechanism was integrated to further enhance the model performance. Through a series of experiments, it is found that the improved ResNet50 model has the highest classification accuracy of 80.6% on large-scale image data sets, which is 11.7% higher than the traditional ResNet50 model. In terms of recognition of similar style images, the model incorporating triplet attention demonstrated higher average accuracy (74%) and recall (82%). This improvement has achieved certain results and has certain technical reference value for various styles of image classification fields.

1. Introduction

With the advent of the era of big data, the explosive growth of image data has greatly promoted the development of image classification methods based on deep learning (DL) algorithms. These methods have gradually replaced the traditional manual recognition methods, especially in the processing of large-scale and diverse image data show their advantages [1]. Media painting images, including historical photos, magazine pictures, news photos and painting pictures, almost cover all the image styles in our lives. The style of these images is not only reflected in the salient features, but often hidden in the nuances, such as the distinction between Baroque and Rococo styles in oil paintings is a challenge [2]. This requires image classification not only to identify the overall features, but also to focus on the extraction of obvious and detailed features. In the field of deep learning, image classification research is mainly divided into two categories, one is based on traditional manual feature extraction algorithm, and the other is based on convolutional neural network (CNN) method [3]. Traditional manual feature extraction is effective in processing images of specific types and simple structures, but in the face of complex picture elements and style diversity, this method is often limited by feature extraction within its capability [4]. In contrast, convolutional neural networks show obvious advantages in processing large and complex image data because of their strong feature learning ability. Therefore, this study aims to explore how to further improve the performance of image classification models based on convolutional neural networks. The ResNet is selected as the basic framework, and the Resnet 50 version with the best performance is improved. Specifically, the blur pool operation is introduced into the model structure, and the traditional

[*] Corresponding author.
E-mail address: chichi0707@163.com (X. Zhang).

https://doi.org/10.1016/j.heliyon.2024.e27178
Received 18 September 2023; Received in revised form 24 February 2024; Accepted 26 February 2024
Available online 28 February 2024

Relu function is replaced by Celu activation function in order to improve the model's ability to recognize subtle stylistic differences. In addition, considering the potential of the attention mechanism (AM) in image processing, the triplet attention mechanism is incorporated to improve the targeting performance of similar image classification. It is hoped that these improvements will not only improve the accuracy of image classification, but also provide an effective solution for similar image classification.

This study consists of four parts. Firstly, the current research achievements in image classification models and attention concentration are reviewed. Secondly, the methodology of this study is introduced. Next, based on the methods in the second part, experiments are conducted and the results are analyzed. Finally, the conclusion of this study is summarized.

2. Related works

Computer developing has enabled machine learning (ML) to be well applied in image style recognition. Currently, single class label methods are commonly used for image classification. However, Kolisnik B et al. [5]. Believe that images as a hierarchical structure of classes represent more specific classes as the level of layers increases. Therefore, they propose a new hierarchical image classification model. The new model can reduce network training time by using higher-level actual class labels to train lower level branch CNN. In addition, the model also learns the relationship between layers to obtain Conditional probability, which can prevent error propagation. Compared with the basic CNN, this model achieved higher classification accuracy. In medical image recognition and classification, Blaivas M et al. [6] compared and trained CNNs with six different complexities and ages, and tracked and recorded the training time and accuracy separately. In addition, through real-world testing of the algorithm, important evaluation parameters such as descriptive statistics and Pearson R values were calculated for each model. The training time of the visual geometry group network was 232 min, with an accuracy of 96%. The development of swarm intelligence methods and ML in image recognition is very rapid. The applicability of DL based models in different fields of image recognition has also been enhanced. Revathi M et al. [7] proposed an improved DL image classifier. By fusing the two algorithms, the model can autonomously learn and explore the optimal solution. Compared with other image classification methods, this model has a certain improvement in the effectiveness of image classification. Xu Y et al. [8] proposed a new aerial image classification research model to overcome the problem of inaccurate detection of discriminant clues by traditional models. By performing morphological filtering on multiple components obtained from deconstruction of aerial images, it facilitates the feature extraction operation of the quality model generation discrimination mechanism. This model can effectively recognize the types of aerial images and has strong robustness.

For media image detection applications, Tang Z et al. [9] designed a visual attention model based invariant moment hash image algorithm based on commonly used hash algorithm image processing techniques. The introduction of visual AM enables the model to effectively and accurately capture important image information features in the focus area of attention. The weighted discrete wavelet transform extracts invariant moments for hash construction, which makes the model show good classification performance on open image data sets. This model has better image classification performance than other hash algorithms. However, Fang C et al. [10] believe that although the introduction of AM can make image classification models more targeted and accurate, there are still shortcomings for low resolution image classification. So they introduced triple attention AM into the super-resolution image classification model to obtain good representation ability. Compared to existing image classification models with AM, this model has improved classification accuracy by 1–3 percentage points. Andriyanov et al. [11] discussed the importance of object detection and recognition in images and image sequences, and introduced various methods and techniques proposed in this regard. In particular, the research focuses on methods based on image mathematical models, random field models and likelihood ratios, as well as the development of convolutional neural networks in solving object recognition problems. It also introduces some efficient pre-training architectures that do not use mathematical models but are trained using a library of real images. It provides important reference and guidance for the research of object detection and recognition methods. Poongodi et al. [12] proposed a method to adapt media titles and add specific sounds to images. To achieve this effect, a computer vision model of image scene recommendation and natural language processing is combined. In the experiment, the accuracy of Top5 index and Top1 index is 67% and 53% respectively. Although the accuracy is not high, it provides a reference for the application expansion of media image recognition. Li et al. [13] proposed an image-based two-stage data-driven framework for satellite image detection and segmentation. In the first stage of this method, the object detection algorithm Faster-RCNN is used to detect satellite images and present them in the form of bounding boxes. In the second stage, the satellite image is cropped into a small image according to the position information of the boundary box, and the boundary detection algorithm is used to identify the boundary information. The effectiveness, efficiency and universality of the framework are verified by experiments.

In summary, for different styles of media image classification, experts have improved DL to improve the accuracy of image recognition. Some experts have also attempted to integrate AM in DL to improve the performance of special image classification. However, the improvements made by experts on the basis of traditional CNN have overlooked that there are still better choices for the network layer structure of the model. And the fusion of AM has not taken into account the issue of similar and easily confusing image styles. So it is worth studying how to design a model that can recognize a large number of images and accurately distinguish similar images.

3. Design of image classification model integrating ResNet and AM

Compared to traditional image classification, the style of an image often involves the entire image, requiring attention not only to the details of the image, but also to the overall features of the image [14]. The color, stroke information, texture information, structural layout, etc. In painting works can have an impact on the classification of image styles. And image recognition with high similarity requires the model to be able to extract detailed feature information to improve the accuracy of classification [15]. So this study improved the traditional classification model based on ResNet50 network, further improving the accuracy of image style classification recognition. And AM was introduced into the improved ResNet50 model in the experiment to design a new image style classification model that integrates ResNet50 and AM.

3.1. Improved design of traditional image classification model based on ResNet

Traditional image classification models are usually based on CNN, and their recognition performance is related to the amount of feature information data captured by the model. However, when convolved into deep network layers, traditional classification models may experience performance degradation due to the accompanying gradient explosion or disappearance [16]. In response to this issue and to improve the model's image feature recognition performance, this study made improvements on traditional classification models and designed an image classification model based on Residual Network (ResNet). As a ResNet composed of multiple residual blocks stacked together, it will be easier to improve the network model structure. This study is based on ResNet50 for improvement. Fig. 1 shows the network structure diagram of ResNet50 before improvement. The core principle that residual structure in ResNet50 can improve the performance of network models is that the characteristics of residual structure enrich the feature data by combining shallow and deep network feature information [17]. Bo_A and Bo_B of Fig. 1 are the residual blocks in the structure. The input information of the next layer residual group is obtained through the weight layer and activation function.

ResNet50, as a CNN, essentially seeks the optimal weight value for feature extraction, which can be described by equation (1).

$$\theta^{*} = \arg \max_{\theta} L(f(x,\theta),y) \tag{1}$$

In equation (1), L is a loss function of ResNet50 [18]. θ contains all parameters in the parameter space of the text CNN. θ^{*} is the group with the best weight value among all parameters. $f(x,\theta)$ represents the model. y represents the entered data information. y is the true label for each sample. After calculating ownership weight gradient, the weight value will be updated. The basic update weight equation (2) in the algorithm is:

$$\theta_{iter} = \theta_{iter-1} - \eta g_{iter} \tag{2}$$

In equation (2), the number of iterations during the training process is represented by *iter*. BB is all parameters iterated to the *iter*-th parameter space. DL rate is expressed in η. g_{iter} is a gradient. This is the most basic update algorithm that can be used for model training, but there are problems like slow Rate of convergence and long training time. So the design of ResNet50 uses Adam optimization algorithm. It combines RMSprop method and Momentum method to improve training efficiency, adjust the Adaptive learning rate, and retain the last gradient direction information after each iteration update [19]. The process of updating the weight values for each iterating in Adam optimization algorithm is represented by equations (3) and (4):

$$v_{iter} = \beta_1 v_{iter-1} + (1-\beta_1) g_{iter} \tag{3}$$

In equation (3), the update direction at the *iter*-th iteration is represented by v_{iter}, and β_1 is attenuation coefficient.

$$\omega_{iter} = \beta_2 \omega_{iter-1} + (1-\beta_2) g_{iter} \cdot g_{iter} \tag{4}$$

In equation (4), the momentum of the *iter*-th iteration is represented by ω_{iter}, and β_2 is attenuation coefficient. The control of learning rate attenuation is expressed by equations (5)–(7):

$$\hat{v}_{iter} = \frac{v_{iter}}{1-\beta_1^{iter}} \tag{5}$$

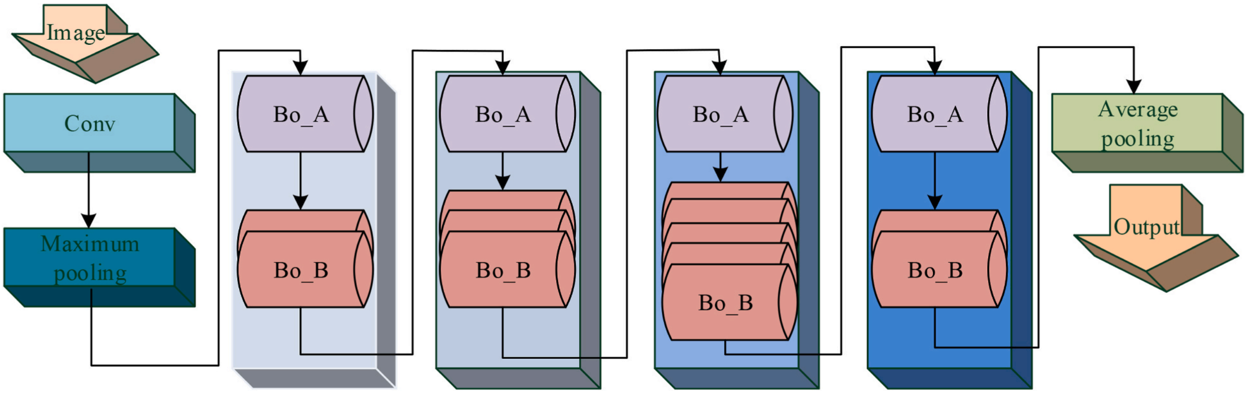


Fig. 1. Network structure diagram of ResNet50 before improvement.

In equation (5), the update direction at the *iter*-th iteration is represented by υ_{iter}, and β_1^{iter} is attenuation coefficient.

$$\overset{\wedge}{\omega}_{iter} = \frac{\omega_{iter}}{1-\beta_2^{iter}} \tag{6}$$

In equation (6), the momentum of the *iter*-th iteration is represented by ω_{iter}, and β_2^{iter} is attenuation coefficient.

$$\theta_{iter} = \theta_{iter-1} - \eta \frac{1}{\sqrt{\overset{\wedge}{\omega}_{iter}} + \varepsilon} \cdot \overset{\wedge}{\upsilon}_{iter} \tag{7}$$

In equation (7), ε is a constant. To avoid the denominator value being equal to 0, ε is usually taken as 10^{-8}.

In this study, a series of improvements are made to the network structure of ResNet50 to improve the image classification effect and reduce the distortion of the model. Firstly, the step size of the maximum pooling layer is shortened, the original step size of the pooling layer is changed to the step size of 1, and the blur pool structure is introduced. This change is designed to increase the network's ability to capture image details by reducing step size, while combining downsampling and low-pass filtering to reduce high-frequency noise and distortion and preserve important image features by introducing a blur pool structure. Then, we also modify the step size of the second convolution layer in the Bo_A structure of the third and fourth layers by setting the step size to 1, and also introduce the blur pool structure after normalization. This adjustment aims to improve the feature extraction capability of the network in these layers and reduce the possible information loss during the downsampling process through the blur pool. In addition, in order to further optimize the performance of the model, the blur pool structure is inserted in front of the subsampling layer of the Bo_A structure of the third and fourth layers, and the step of the subsampling layer is set to 1. Such adjustments not only allow the model to capture more detail while maintaining resolution, but also reduce distortion through low-pass filtering, thus improving the accuracy of image classification. Through these structural improvements, the model can more effectively retain key information when processing complex images, while reducing unnecessary information loss and distortion, thus significantly improving the accuracy and reliability of image classification.

In the pre improvement ResNet50 structure, Relu function was used in the output data activation operation step, described by formula (8).

$$Relu(x) = \max(0, x) \tag{8}$$

In equation (8), x is the input value. When $x \geq 0$, Relu function outputs it directly. On the contrary, the output value is 0. But if the input values into Relu function are not greater than or equal to 0, that is, the output values are all 0, it will reduce the learning performance of ResNet50. Therefore, Celu activation function is used in this study, which is described by formula (9).

$$Celu(x) = \max(0, x) + \min(0, \alpha * (\exp(x / \alpha) - 1) \tag{9}$$

In equation (9), when $x \geq 0$, Celu function outputs it normally, and when x is less than 0, $\alpha * (\exp(x/\alpha) - 1$ is output, and $\alpha = 1$.

Fig. 2 shows the structure diagram of the improved ResNet50 model. Bo_C structure in the residual block group in Fig. 2 is a residual block that introduces the blur pool structure. It also performs weight operations on input information, combines down-sampling and low-pass filtering, and outputs it through Celu function.

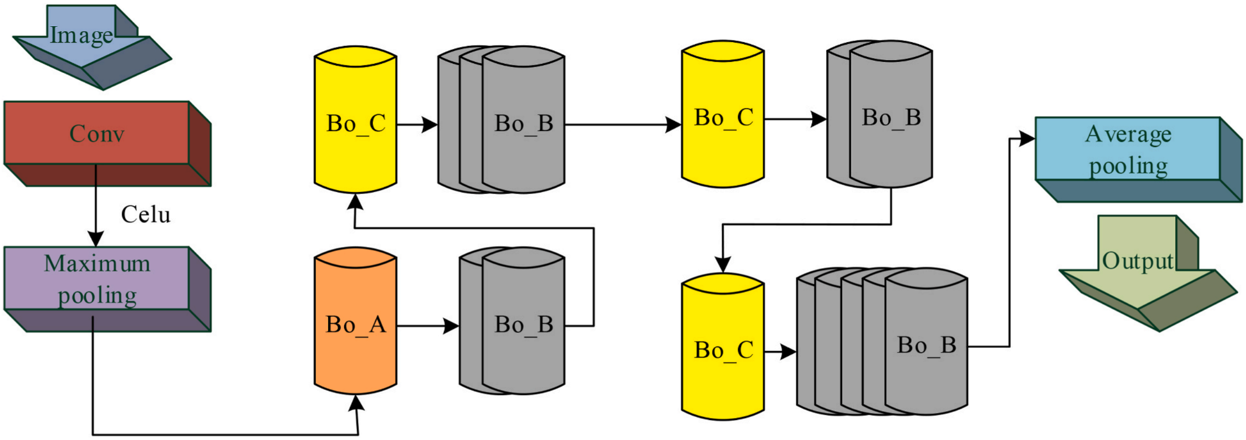


Fig. 2. Structure diagram of the improved ResNet50 model.

3.2. Design of ResNet50 image classification model based on AM fusion

However, for a group of images with high similarity, it is required that the model can extract detailed feature information to improve the accuracy of classification. To enable ResNet50 to extract more detailed feature information in the image, it is also necessary to fuse AM in ResNet50 and emphasize the detailed features in the image by weighting the weight coefficients obtained through ResNet50 DL iteration.

Traditional classification models typically use channel AM or combine it with spatial AM for AM application [20]. The way channel AM captures effective feature information in an image is by learning the relevant information between channels in the network structure. The channel AM combined with spatial AM can capture specific target objects in the image. In traditional structures, the most important feature information of the corresponding feature layer can be captured through maximum pooling operation [21]. By pooling the global maximum value, the two-dimensional data of the feature layer is transformed into a one-dimensional vector. The pooled vector's length is equal to channels number in feature layer. This can achieve that each channel only retains the maximum value of data. This process is described by formula (10):

$$\alpha_i^{(0)} = \max_v \left(\underset{i}{\hat{s}}\, l(v) \right) \tag{10}$$

In equation (10), $\alpha_i^{(0)}$ is the vector, and $\underset{i}{\hat{s}}\, l(v)$ is the i-th channel's v-th feature data in the l-th layer. After obtaining the vector, it is input into the fully connected network, and Relu function activates it, which is expressed by formula (11):

$$\alpha_j^{(1)} = relu\left(\sum_i \alpha_i^{(0)} w_{ij}^{(a_1)} + b_j^{a_1} \right) \tag{11}$$

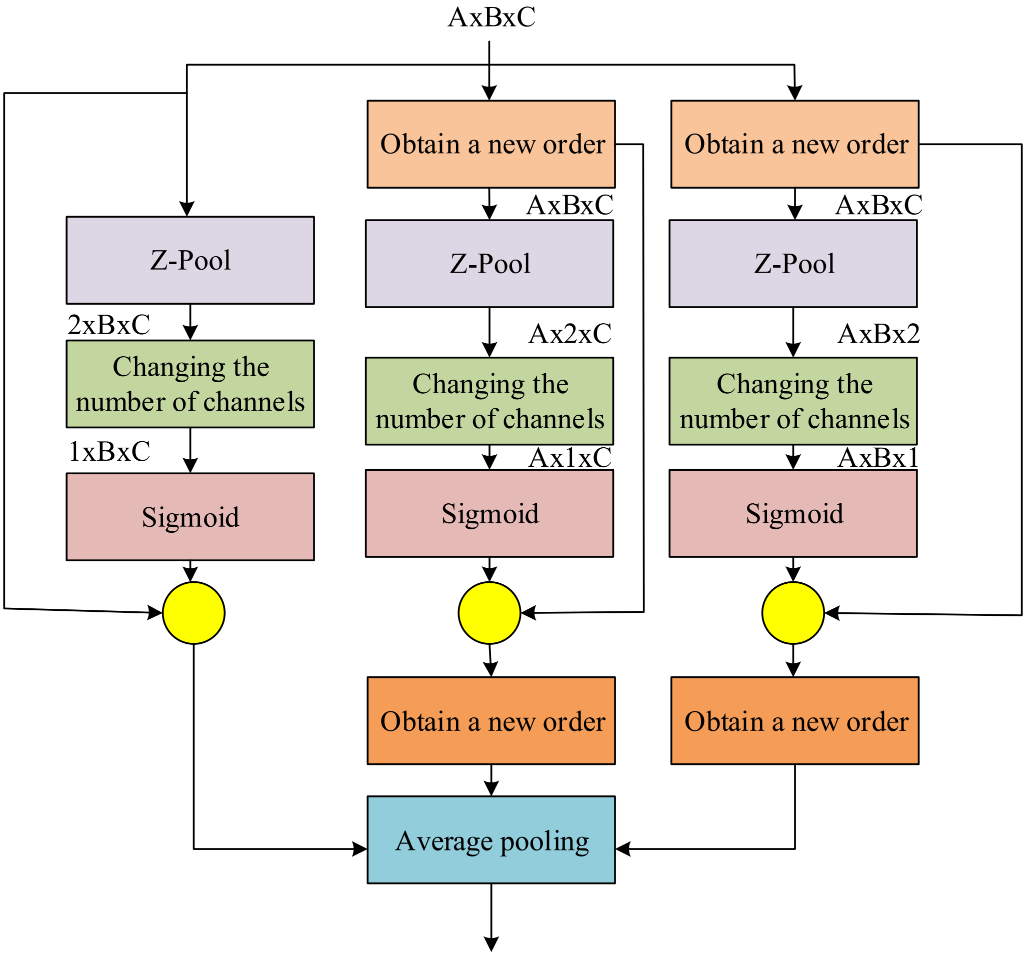


Fig. 3. Triplet attention structure diagram.

In equation (11), $\alpha_j^{(1)}$ is a hidden layer vector. $w_{ij}^{(a_1)}$ and $b_j^{a_1}$ are weight parameters for updating gate operations. After obtaining the hidden layer, the results are transmitted to the output layer. In neural networks, both output and input are vectors with channel quantities equal to length. Activation function in the output layer is sigmoid, and the output result will be a probability value in the interval (0, 1), which is described by formula (12):

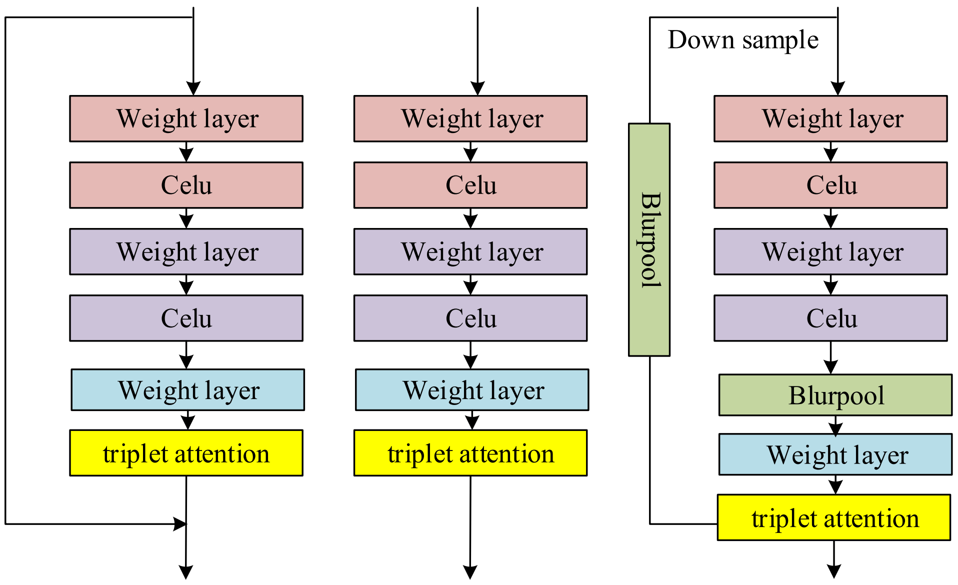
$$\alpha_i^{(2)} = sigmoid\left(\sum_j \alpha_j^{(1)} w_{ji}^{(a_2)} + b_i^{a_2}\right) \tag{12}$$

In equation (12), $w_{ij}^{(a_2)}$ and $b_j^{a_2}$ are the weight parameters for the update gate operation. Finally, to suppress the worthless text data of some channels and retain important channel text data, formula (13) was used for the obtained $\alpha_i^{(2)}$:

$$s_i^l = \alpha_i^{(2)} \underset{i}{\hat{s}} l \tag{13}$$

In equation (13), $\underset{i}{\hat{s}} l$ is the i-th channel's characteristic data in the l-th layer.

However, traditional AM can cause heavy computational burden when processing larger images due to the excessive number of nodes in the image. This study adopts a triple attention structure with low computational complexity to address this issue. Different from the traditional attention mechanism, the triplet attention structure effectively reduces the computational complexity through its three-branch structure, while ensuring the comprehensive capture of information [22]. Triplet attention contains three parallel branches, each focused on a different task. The first branch is responsible for calculating the weight of the feature information, which improves the focusing efficiency of the model by evaluating the importance of each pixel in the image. The second branch focuses on the acquisition of interactive information in the spatial dimension, which enhances the model's understanding of the image structure by analyzing the spatial relationship between the various parts of the image. The third branch focuses on the interaction information in the channel dimension, analyzing the interaction between different channels (or different feature types), so as to improve the richness of feature representation. The combined effect of these three branches allows triplet attention to fully understand image content while remaining computationally efficient. It can significantly reduce the computational burden of large-size image processing without sacrificing performance. In addition, the structural design of triplet attention enables it to perform well in image processing tasks of varying sizes and complexity, especially in application scenarios that require detailed and comprehensive feature understanding. By embedding a triple attention structure after the third convolutional layer of ResNet50 structure and adding the low feature information of the shallow network to triple attention's output, the emphasized detailed feature output can be obtained. Specifically, three components in this structure serve as output tensors, which are obtained in a new order through the triple attention structure and then input into the Z-pool. The entire process only requires minimal computational effort to construct the dimensional weight relationship between space and channels. Fig. 3 shows the structure of a triple attention. After inputting the tensor x with the shape set to AxBxC



(a) Structure of Bo_A (b) Structure of Bo_B (c) Structure of Bo_C

Fig. 4. Structure diagram of residual block after introducing triple attention.

into Z-pool, Z-pool preserves as much information as possible in x by increasing channels number in the first dimension of x to 2. Based on the large amount of information in the tensor, the computational complexity of AM can be reduced, as described by formula (14).

$$Z-pool(x)=[MaxPool_{0d}(x),AvgPool_{0d}(x)] \tag{14}$$

In equation (14), $MaxPool_{0d}(x)$ is a description of the maximum pooling operation. $AvgPool_{0d}(x)$ is a description of the average pooling operation. After changing channels number in Z-pool, tensors with a shape of 2xBxCx are transformed into 1xBxCx through convolutional and batch normalization layers. Finally, the sigmoid function activates the results to obtain the corresponding importance weights for each tensor.In Fig. 4 and (a) shows the structure diagram of the Bo_A; Fig. 4 (b) shows the structure diagram of the Bo_B; Fig. 4 (c) shows the structure duagram of the Bo_C. A tensor with the input shape AxBxC was input into the residual block, which was activated by Celu function after passing through weight layer and AM for output.

This study introduces the evaluation index F-Score (F) to comprehensively consider the designed model, represented by formula (15) as [23]:

$$F=(1+\beta^2)\frac{precision*recall*(1+\beta^2)}{(precision+recall)} \tag{15}$$

In equation (15), β represents the importance description of precision (P) and recall (R). If β is less than 1, it indicates that P is more important than R. If β is greater than 1, it indicates that R is more important than P. β equals 1, indicating that both are equally important. In the experiment, β is 1.

4. Experimental results and analysis

The operating system version number of this experimental platform is Ubuntu 20.04.2, and GPU is GeForce GTX 1080Ti. It is designed and implemented using Python framework. The initial learning rate is set to 1e-3 for 100 iteration cycles. Every 30 iteration cycles, learning rate becomes one tenth of the original. Loss function is a cross entropy loss function, and the optimizer uses Adam optimizer. The evaluation indicators are P, R, and F.

4.1. Experimental results and analysis of the improved ResNet50 classification model

This experiment used three sets of 5242 images as samples, with images of different categories as a group in Table 1.

This experiment consists of two stages. In the first stage, four experiments were conducted to compare different CNN types' classification performance on three images sets. These models include depth CNNAlexNet, visual geometry group VGG19 (VGG), ResNet34, the pre-improved ResNet50 model, and the improved ResNet50 model.

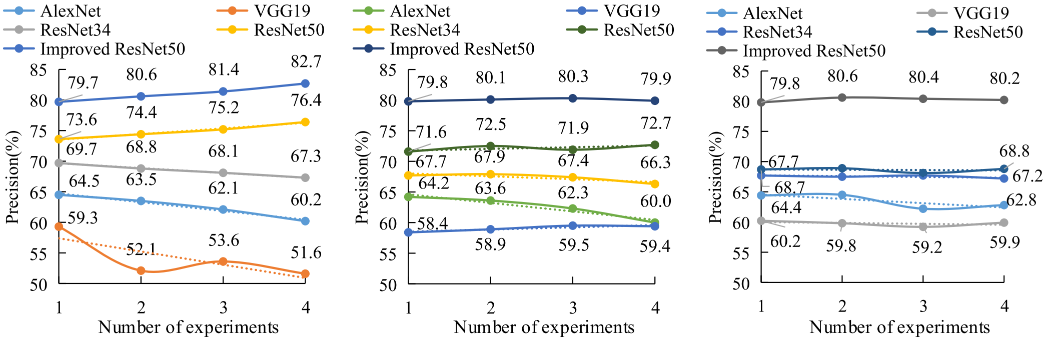
Fig. 5 shows the trend of recognition P changes for these three sets of image sets using different models. In the first set of image set tests, all five classification models showed an unstable state during repeated testing. Among them, ResNet50 model and the improved ResNet50 model showed a slight upward trend in overall P, both increasing by about 3 percentage points, with the highest P being 76.4% and 82.7%. Overall, the P values of AlexNet, ResNet34, and VGG19 all showed a decrease of 2–8 percentage points, with VGG19 exhibiting the most significant instability during the recognition process, with a sharp drop in P to 52.1% in the second test. However, in the second and third set of image set tests, the recognition P of the five models showed a stable trend overall. Except for AlexNet model, all other models fluctuated by about one percentage point, while AlexNet overall showed a downward trend, with the lowest P reaching 60.0%. In the third set of image set tests, the overall P of the improved ResNet50 was much higher than other models, reaching a maximum of 80.6%, which was 11.7% higher than the second highest ResNet50.

Fig. 6 shows the average P comparison of five models tested on each set of images. Regardless of the image set, the average recognition P of ResNet34, ResNet50, and improved ResNet50 based on residual linking is 13–27 percentage points higher than AlexNet and VGG19 models. In the classification model based on residual connection structure, the improved ResNet50 has the highest overall average P, and the highest P was tested in the first image set, which is 81.1%.

In the improvement of ResNet50 model, Celu function was used to replace Relu function used in traditional ResNet50. Next, based on the improved ResNet50 model, the experiment compares the classification P of three groups of pictures under Relu, PRelu, LeakyRelu, Tanh and Celu activation functions. Similarly, four detections were repeated on each set of images. Whether the improvement can improve model recognition P is tested. Fig. 7 shows the recognition P' change trend of the improved ResNet50 model for three groups of image sets when using different activation functions. Fig. 7 (a) shows the comparision of precision between different models

Table 1
Various image sets.

First set of images	Finance and economics	Physical culture	Entertainment	Society	Historic photos
Number of images	350	482	497	431	289
Second set of images	Iconic image	Documentary photos	magazine cover	Special images	Book Illustrations
Number of images	336	384	415	190	365
Third set of images	Oil painting images	Thangka image	Wash painting	Sketch	Comic
Number of images	412	178	210	359	344



(a) Comparison of precision between different models in the first set of images

(b) Comparison of precision between different models in the second set of images

(c) Comparison of precision between different models in the third set of images

Fig. 5. Comparison of accuracy of different models in three sets of images.

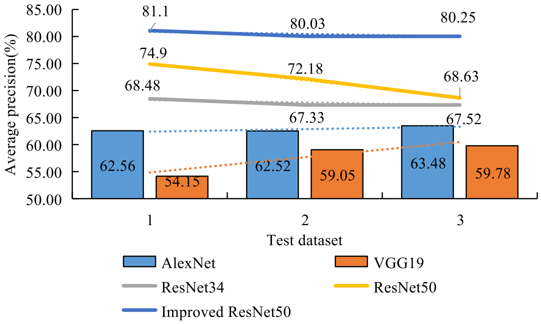
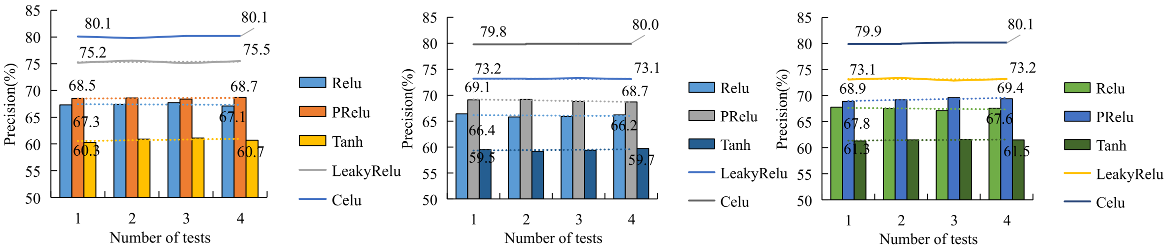


Fig. 6. Comparison of average accuracy of different models in three sets of images.

in the first set of images; Fig. 7 (b) shows the comparision of precision between different models in the second set of images; Fig. 7 (c) shows the comparision of precision between different models in the third set of images. No matter which group of images are recognized and classified, the recognition P under the environment of Celu function and LeakyRelu function is 15–20 percentage points higher than the other three Activation function. Among them, Celu function can outperform LeakyRelu function in P detection, with the former reaching a maximum of 80.2%.

Fig. 8 shows the average P comparison of the improved ResNet50 model in three sets of image set detection experiments under different Activation function environments. Among them, Fig. 8 (a) compares the average P values of five experimental subjects, while Fig. 8 (b) shows the multiple relationships corresponding to the average P of each experimental subject. When Celu Activation function is used in the improved ResNet50 model, the average P of the model is the highest overall. Its highest value is 80.1%, which is about 14 percentage points higher than the traditional ResNet50 model using Relu function. And from the sub-graph (b) of Fig. 9, although the

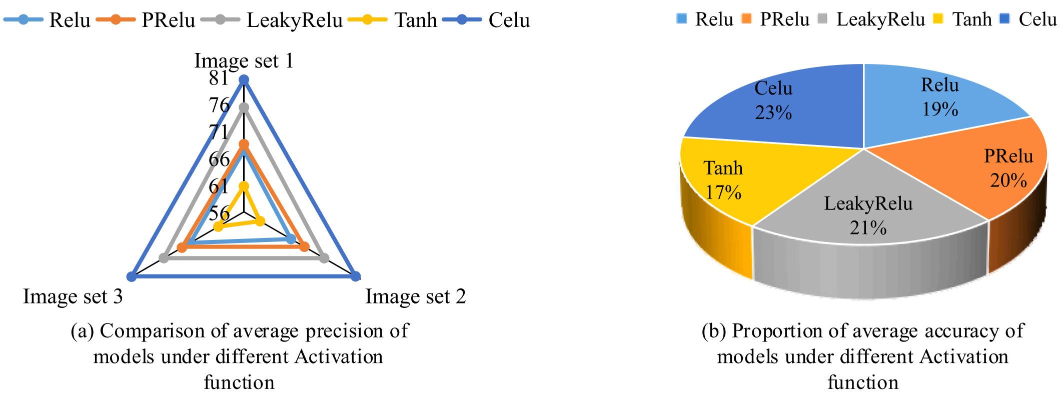


(a) Comparison of precision between different models in the first set of images

(b) Comparison of precision between different models in the second set of images

(c) Comparison of precision between different models in the third set of images

Fig. 7. Classification accuracy.

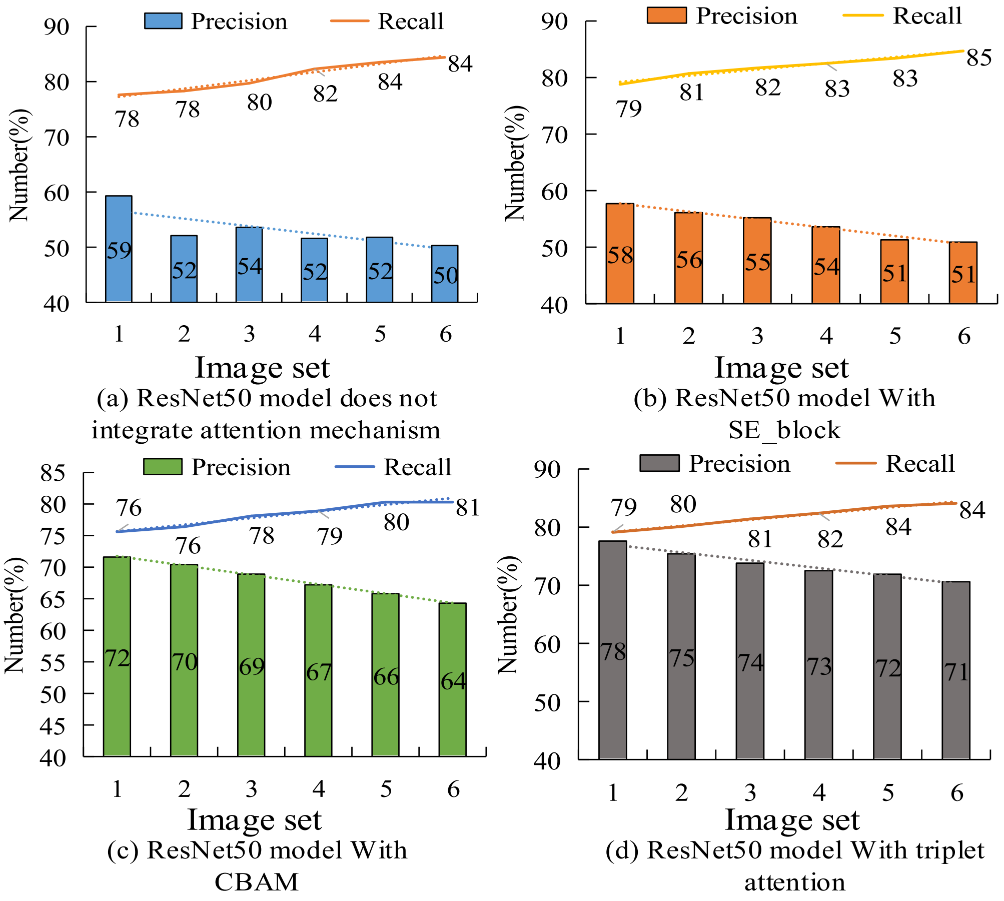


(a) Comparison of average precision of models under different Activation function

(b) Proportion of average accuracy of models under different Activation function

Fig. 8. Comparison of average accuracy of models under different Activation function.

improved functions PRelu and LeakyRelu based on Relu function can improve the P of ResNet50 model, the increase is not significant, ranging from 2 to 6 percentage points. This also shows that activation function replacement of the traditional ResNet50 model in this study has a significant effect on the model recognition P improvement.



(a) ResNet50 model does not integrate attention mechanism

(b) ResNet50 model With SE_block

(c) ResNet50 model With CBAM

(d) ResNet50 model With triplet attention

Fig. 9. Comparison of classification performance of models with different attention mechanism structures for similar images.

4.2. Experimental results and analysis of ResNet50 classification model after AM fusion

On the basis of the improved ResNet50 classification model, triple attention AM is introduced. In this regard, a comparative experiment was conducted on ResNet50 classification models paired with different AMs in a painting image set containing six sets of similar styles for similar image recognition and *P*-detection. Table 2 presents the painting image set data.

Fig. 9 shows the comparison of the recognition detection P and R of the improved ResNet50 classification model for painting image sets. Among them, Fig. 9 (a) shows ResNet50 model does not integrate attention mechanism; Fig. 9 (b) shows ResNet50 model with SE_block (Squeeze and Excitation); Fig. 9 (c) shows ResNet50 model with CBAM (Convolutional block attention module); Fig. 9 (d) shows ResNet50 model with triple attention AM, and AM. The overall recognition P of all experimental subjects decreased due to the increasing difficulty of distinguishing similar styles in the atlas group. But ResNet50 classification model, which is paired with SE_block and triple attention AM, has the smallest decrease of 7 percentage points. But under SE_block AM, the overall P is low, ranging from 51% to 58%. Overall, the P of the model under the triple attention structure is about 20 percentage points lower. The R changes of all experimental subjects ranged from 76% to 85%. But when the model is paired with CBAMAM, both P and R are low at the same time, both being 76%. This is because the use of AM is not stable.

Table 3 shows the average accuracy rate, average recall rate and F-Score data when the model is combined with different attention mechanism structures in the recognition process of similar style images, and compares them with some advanced algorithms. The advanced algorithms involved in the comparison include recurrent neural network - Long short-term memory network (RNN-LSTM) and generative adversarial network (A-GAN) integrating attention mechanism. The higher the F-Score, the higher the model recognition performance. It can be seen that although the average recall rate of each subject varies from 78% to 82%, the accuracy of the model with and without the SE_block attention mechanism is significantly lower, 54% and 53% respectively, which is lower than the advanced methods of RNN-LSTM and A-GAN. However, the method combining CBAM and Triplet attention mechanism is superior to the advanced methods participating in the comparison, and the average accuracy and average recall rate are higher than other experimental methods. When the model is paired with triple attention AM, F is also the highest, at 0.76, which is 0.4 to 0.12 higher than other experimental subjects. Especially compared with ResNet50 classification model without AM, ResNet50 classification model with triple attention has a 0.12 higher F, which significantly improves the recognition performance of similar style images.

In the image style classification task, the computational complexity of the algorithm greatly affects the work, so the comprehensive performance evaluation of the algorithm needs to include the comparison of the computational complexity of different algorithms. In this study, the computational time complexity and computational space complexity are used as indexes to compare different algorithms. As shown in Fig. 10. Fig. 10 (a) is A comparison of computing time. It can be seen from the figure that the computing time of the research method is lower than that of other attention mechanisms, and slightly higher than that of RNN-LSTM and A-GAN. Fig. 10 (b) shows the comparison of computing space. From the perspective of the memory proportion in the calculation process, the calculated memory proportion of the research method is slightly higher than that of the improved ResNet50 without attention mechanism, but lower than that of other methods participating in the comparison. It can be seen that the research method significantly reduces the computational burden compared with other attention mechanisms. Although compared with some advanced algorithms, the computational complexity is not very different, but considering the high performance in precision and recall rate, the study believes that this is completely acceptable.

5. Discussion

The research focuses on the problem of image style classification in media painting, which has significant significance in the automatic analysis of art works and content management system. Media painting images, such as historical photos, magazine pictures, etc., are not only large in quantity, but also diverse in style, and their automatic classification is extremely important for digital cultural heritage protection, art education, copyright management and digital library construction. By combining ResNet and attention mechanism, this study not only improves the accuracy of image classification, but also enhances the model's ability to recognize subtle stylistic differences, providing an effective way to deal with complex and detailed image styles. The improved model combining ResNet50 and triplet attention has wide application potential. For example, in medical image analysis, this model can be used to improve the accuracy of lesion identification. In the field of security monitoring, this model can help realize the rapid identification of subtle changes in the scene. In addition, in the social media content analysis and personalized recommendation system, the model can also provide more accurate image content recognition and classification services. However, this study also has some limitations, such as a high reliance on the quality and diversity of the dataset, and possible differences in effects across different types of images. Future research can focus on improving the diversity and quality of datasets, optimizing the scalability and applicability of models, further improving computational efficiency, and further studying the effects of different attention mechanisms. These efforts will help to further enhance the practicality and effectiveness of the model, and bring more innovations and breakthroughs in the field of image classification.

6. Conclusion

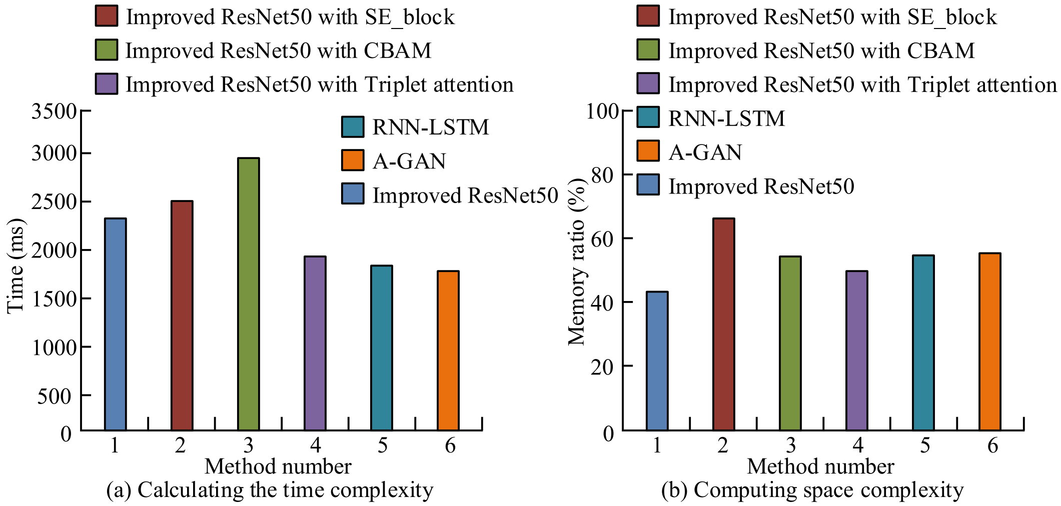
The recognition accuracy of traditional CNN based image style classification models is insufficient, making it difficult to distinguish similar images. In this regard, this study is based on the improvement of the higher quality ResNet50 algorithm to design a new model that can improve image classification P. At the same time, the improved ResNet50 classification model is integrated with triple attention AM to improve the performance of similar style image classification in the model. The improved ResNet50 classification

Table 2
Painting image set data.

Image group Painting style	Baroque	Rococo	Gothic	line drawing	Realistic writing	Cartoon
1	102	98	10	10	10	10
2	20	20	138	20	20	22
3	21	21	20	140	19	19
4	10	10	10	89	96	25
5	20	20	78	45	20	57
6	55	55	10	55	55	10

Table 3
Average accuracy, average recall, and F-score data.

Method		Average precision rate (%)	Average recall rate (%)	F-Score value
Improved ResNet50	No attention mechanism	53	81	0.64
	SE_block	54	82	0.66
	CBAM	68	78	0.72
	Triplet attention	74	82	0.76
RNN-LSTM		59	79	0.66
A-GAN		63	80	0.69



(a) Calculating the time complexity

(b) Computing space complexity

Fig. 10. Comparison of computational complexity of different methods.

model has a significantly higher *P*-value compared to other convolutional network algorithm models in recognizing different style image sets, with a maximum of 80.6%, and 11.7% higher than the traditional ResNet50 model. When ResNet50 classification model uses different Activation function, and the improved ResNet50 model uses Celu Activation function, the average P of the model is the highest overall. Its highest value is 80.1%, which is about 14 percentage points higher than the traditional ResNet50 model using Relu function. In the performance testing of ResNet50 model fused with triple attention, the average P and average R of the model on six similar images were higher than those of the model paired with other AMs, with 74% and 82%, respectively. The F of this model is also the highest, at 0.76, which is 0.4 to 0.12 higher than other experimental subjects. Compared with traditional ResNet50, ResNet50 classification model with triple attention has a 0.12 higher F than the former. The image classification model designed in this study, which combines the attention mechanism and ResNet50 algorithm, has significantly improved the performance of image recognition. However, the research also has some shortcomings. First, the model is highly dependent on data sets, which means that further tuning and optimization may be required on different types or qualities of data sets. Second, although the model is excellent at handling painted image styles, its ability to handle differences in image quality has not been fully validated. In addition, the computational efficiency of the model can be improved, especially in resource-constrained environments. Future work will focus on several key areas. First, testing and optimizing the model on a broader and diverse dataset will be explored to improve its ability to generalize. Secondly, we will study how to improve the model to better handle images of different qualities, which is especially important for practical application scenarios. In addition, the study plans to further explore computing efficiency improvement strategies for effective deployment in environments with limited computing resources. Finally, the research will also consider exploring different types of attention mechanisms to further improve the performance and accuracy of the model. Through these works, the research hopes to bring more innovations and breakthroughs in the field of image classification and provide more powerful tools for automatic image analysis.

Data availability

All data generated or analyzed during this study are included in this published article.

CRediT authorship contribution statement

Xinyun Zhang: Investigation. **Tao Ding:** Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] W. Liang, Y. Wu, M. Li, Y. Cao, Adaptive multiple kernel fusion model using spatial-statistical information for high resolution SAR image classification, Neurocomputing 492 (Jul.1) (2022) 382–395.

[2] Y. Guo, Z. Mustafaoglu, D. Koundal, Spam detection using bidirectional transformers and machine learning classifier algorithms, Journal of Computational and Cognitive Engineering 2 (1) (2022) 5–9.

[3] Z. Lu, Y. Chen, Pyramid frequency network with spatial attention residual refinement module for monocular depth estimation, J. Electron. Imag. 31 (2) (2022), 23005.1-23005.18.

[4] A. Radman, A. Sallam, S.A. Suandi, Deep residual network for face sketch synthesis, Expert Syst. Appl. 190 (Mar) (2022) 115980.1–115980.10.

[5] B. Kolisnik, I. Hogan, F. Zulkernine, Condition-CNN: a hierarchical multi-label fashion image classification model, Expert Syst. Appl. 182 (Nov) (2021) 115195.1–115195.14.

[6] M. Blaivas, L. Blaivas, Are all deep learning architectures alike for point-of-care ultrasound?: evidence from a cardiac image classification model suggests otherwise, J. Ultrasound Med. 39 (6) (2020) 1187–1194.

[7] M. Revathi, I.J.S. Jeya, S.N. Deepa, Deep learning-based soft computing model for image classification application, Soft Comput. 24 (24) (2020) 18411–18430.

[8] Y. Xu, M. Wei, M.M. Kamruzzaman, Inter/intra-category discriminative features for aerial image classification: a quality-aware selection model, Future Generat. Comput. Syst. 119 (4) (2021) 77–83.

[9] Z. Tang, H. Zhang, C.M. Pun, M. Yu, Robust image hashing with visual attention model and invariant moments, IET Image Process. 14 (5) (2020) 901–908.

[10] C. Fang, Y. Zhu, L. Liao, X. Ling, TSRGAN: real-world text image super-resolution based on adversarial learning and triplet attention, Neurocomputing 455 (2) (2021) 88–96.

[11] N.A. Andriyanov, V.E. Dementiev, A.G. Tashlinskii, Detection of objects in the images: from likelihood relationships towards scalable and efficient neural networks, Comput. Opt 46 (1) (2022) 139–159.

[12] M. Poongodi, M. Hamdi, H. Wang, Image and audio caps: automated captioning of background sounds and images using deep learning, Multimed. Syst. 29 (5) (2023) 2951–2959.

[13] H. Li, Y. He, Q. Xu, J. Deng, W. Li, Y. Wei, Detection and segmentation of loess landslides via satellite images: a two-phase framework, Landslides 19 (3) (2022) 673–686.

[14] Z. Ji, Y. Yang, F. Wang, L. Xu, X. Hu, Feature encoding with hybrid heterogeneous structure model for image classification, IET Image Process. 14 (10) (2020) 2166–2174.

[15] C.J. Lin, C.H. Lin, S.H. Wang, Integrated image sensor and light convolutional neural network for image classification, Math. Probl Eng. 2021 (Pt.12) (2021) 5573031.1–5573031.7.

[16] B. Wei, K. Hao, L. Gao, X. Tang, Y. Zhao, A biologically inspired visual integrated model for image classification - ScienceDirect, Neurocomputing 405 (Sep.10) (2020) 103–113.

[17] Y. Xie, H. Chen, Y. Ma, Y. Xu, Automated design of CNN architecture based on efficient evolutionary search, Neurocomputing 491 (Jun.28) (2022) 160–171.

[18] H. Meng, C. Zhang, P. Chen, T. Lei, H. Meng, Triplet interactive attention network for cross-modality person re-identification, Pattern Recogn. Lett. 152 (Dec) (2021) 202–209.

[19] Y. Li, X. Jiang, J.N. Hwang, Effective person re-identification by self-attention model guided feature learning, Knowl. Base Syst. 187 (Jan) (2020) 104832.1–104832.11.

[20] W. Gu, R. Nishikubo, A. Saeki, Coordination of NH 2 - or COOH-appended Pt-porphyrins with CsPbBr 3 perovskite quantum dots to improve a cascade process of two-photon absorption and triplet–triplet annihilation, J. Phys. Chem. C 124 (27) (2020) 11439–11445.

[21] C. Qin, Y. Zhang, Y. Liu, S. Coleman, H. Du, D. Kerr, A visual place recognition approach using learnable feature map filtering and graph attention networks, Neurocomputing 457 (Oct.10) (2021) 277–292.

[22] L. Stojanovic, R. Crespo-Otero, Aggregation-induced emission in the tetraphenylthiophene crystal: the role of triplet states, J. Phys. Chem. C 124 (32) (2020) 17752–17761.

[23] Z.X. Huang, J. Li, Z. Hua, Underwater image enhancement via LBP-based attention residual network, IET Image Process. 16 (1) (2022) 158–175.