Second-hand clothing classification algorithm based on feature fusion and attention

DOI: 10.35530/IT.075.06.2023139

TING CHENG
YUN ZHANG
WEIBO LI
JUNJIE ZHANG

ABSTRACT – REZUMAT

Second-hand clothing classification algorithm based on feature fusion and attention

This study suggests a second-hand clothing categorization method based on enhanced residual networks to enhance the effect of second-hand clothing retrieval and encourage clothing transactions on second-hand platforms. This study first gathers image data on used garments. Web crawlers are utilized to gather internet photos of second-hand clothes to train the network model, while camera equipment is used to take pictures of second-hand clothing. The resulting images are then used to assess the network model's categorization accuracy. The next step is to construct a classification model based on ResNet50, add an attention mechanism, and carry out feature extraction in stages. Finally, the developed classification model's performance is assessed and contrasted with other approaches. The experimental findings demonstrate that this strategy outperforms previous methods in terms of classification accuracy on the self-built dataset and DeepFashion dataset, reaching 79.69% and 82.22%, respectively. Additionally, the sorting and recycling of used clothing is greatly assisted by this method.

Keywords: *clothing image classification, second-hand clothing, residual network, attention mechanism, feature extraction*

Algoritm de clasificare a îmbrăcămintei second-hand, bazat pe fuziune de caracteristici și mecanism de atenție

Acest studiu sugerează o metodă de clasificare a îmbrăcămintei second-hand, bazată pe rețele reziduale îmbunătățite, pentru a spori efectul recuperării îmbrăcămintei second-hand și pentru a încuraja tranzacțiile de îmbrăcăminte pe platformele second-hand. Acest studiu adună mai întâi date imagistice despre articolele de îmbrăcăminte uzate. Programele de tip „web crawler" sunt utilizate pentru a colecta fotografii de pe internet ale îmbrăcămintei second-hand pentru a antrena modelul de rețea, în timp ce echipamentele cu camera web sunt folosite pentru a fotografia îmbrăcămintea second-hand. Imaginile rezultate sunt apoi utilizate pentru a evalua acuratețea de clasificare a modelului de rețea. Următorul pas este construirea unui model de clasificare bazat pe ResNet50, adăugarea unui mecanism de atenție și efectuarea extragerii caracteristicilor în etape. În cele din urmă, performanța modelului de clasificare dezvoltat este evaluată și comparată cu alte abordări. Descoperirile experimentale demonstrează că această strategie depășește metodele anterioare în ceea ce privește acuratețea clasificării pe setul de date autoconstruit și setul de date DeepFashion, atingând valori de 79,69% și, respectiv, 82,22%. În plus, sortarea și reciclarea îmbrăcămintei uzate este îmbunătățită semnificativ prin această metodă.

Cuvinte-cheie: *clasificarea imaginii îmbrăcămintei, îmbrăcăminte second-hand, rețea reziduală, mecanism de atenție, extragerea caracteristicilor*

INTRODUCTION

Although using second-hand clothes reduces carbon emissions significantly, this practice is not widely used worldwide due to the influence of important technology and consumer behaviours. Only 10% to 20% of waste textiles in Europe are sold on the used market to be recycled. The foundation for recycling used garments is an effective sorting method [1]. The two primary methods currently used for sorting textiles are manual sorting and automated sorting. To detect and categorize waste clothing components, automatic sorting mostly uses Near Infrared Spectrometry (NIRS) and Raman Spectrometry (RS) [2]. Since manual sorting typically relies on experienced people to conduct the corresponding sorting procedures on the production line, it is vulnerable to visual fatigue and is more easily influenced by subjective factors. As a result, there are many missed and incorrect detections. Contrarily, even though NIRS and RS identification techniques offer great accuracy for waste textiles, their use in industrial settings is constrained by their high working environment requirements. The most effective option to implement the classification of used clothing is the recognition system based on machine vision since it can more adequately compensate for their limitations [3]. The machine vision recognition system can also be used in the commercial sector in addition to the industrial sector. For instance, improved clothes categorization on the second-hand clothing trading platform makes it easier for buyers to locate listed second-hand clothing through suggested information, encouraging both buyers and sellers to complete deals faster.

Classification methods based on machine vision can be divided into two categories: clothing classification methods based on traditional image content and clothing classification methods based on deep learning. Clothing classification methods based on traditional image content mainly use image processing technology to extract image features to describe clothing. Commonly used image features include Local binary pattern (LBP) [4], Scale-invariant feature transform (SIFT) [5], Histogram of oriented gradient (HOG) [6], etc. Classification is performed by inputting the extracted features or combination of features into classifiers such as Support Vector Machine (SVM) [7] and Random Forest (RM) [8]. Such methods are limited by their reliance on the manual design of features and feature combinations, which are highly subjective and task-specific. At the same time, these models can only extract shallow features, and the models are prone to overfitting and inefficiency due to the large amount of data, which ultimately affects the classification accuracy. However, the clothing classification method based on deep learning is not limited by the above conditions. Among them, Convolutional Neural Network (CNN), as a deep learning technology for analysing visual images, can directly convolve image pixels and extract image features from image pixels. In addition, the weight-sharing property and pooling layer of the CNN greatly reduce the parameters that the network needs to train, simplify the network model, and improve the efficiency of training [9, 10]. Yu et al. [11] used VGG16 as the feature extraction network and then added a convolutional block attention module in the second convolutional block to enhance attention to the target area. Therefore, feature extraction is very suitable as a visual recognition technology applied to the field of second-hand clothing classification.

Images of used clothing have the following issues in comparison to brand-new clothing: first, there are frequently issues with cluttered shooting situations, complicated shooting backgrounds, and inadequate lighting when it comes to second-hand clothes photographs. Second, there are unknown elements in used clothing such as deformation, folds, and occlusion. The current second-hand clothing image classification impact is frequently unsatisfactory due to these ambiguous elements, and it is unable to adequately address the needs of second-hand clothing classification in the industrial and commercial domains [12]. This research suggests a second-hand clothing classification method based on an enhanced residual network to more accurately categorize used apparel. In figure 1, the experimental procedure is depicted.

The paper is structured as follows: we described the procedures used to get the data, as well as the training and test sets in 2nd section. In 3rd section, the classification models are designed and implemented, and the model is trained with the created training set.

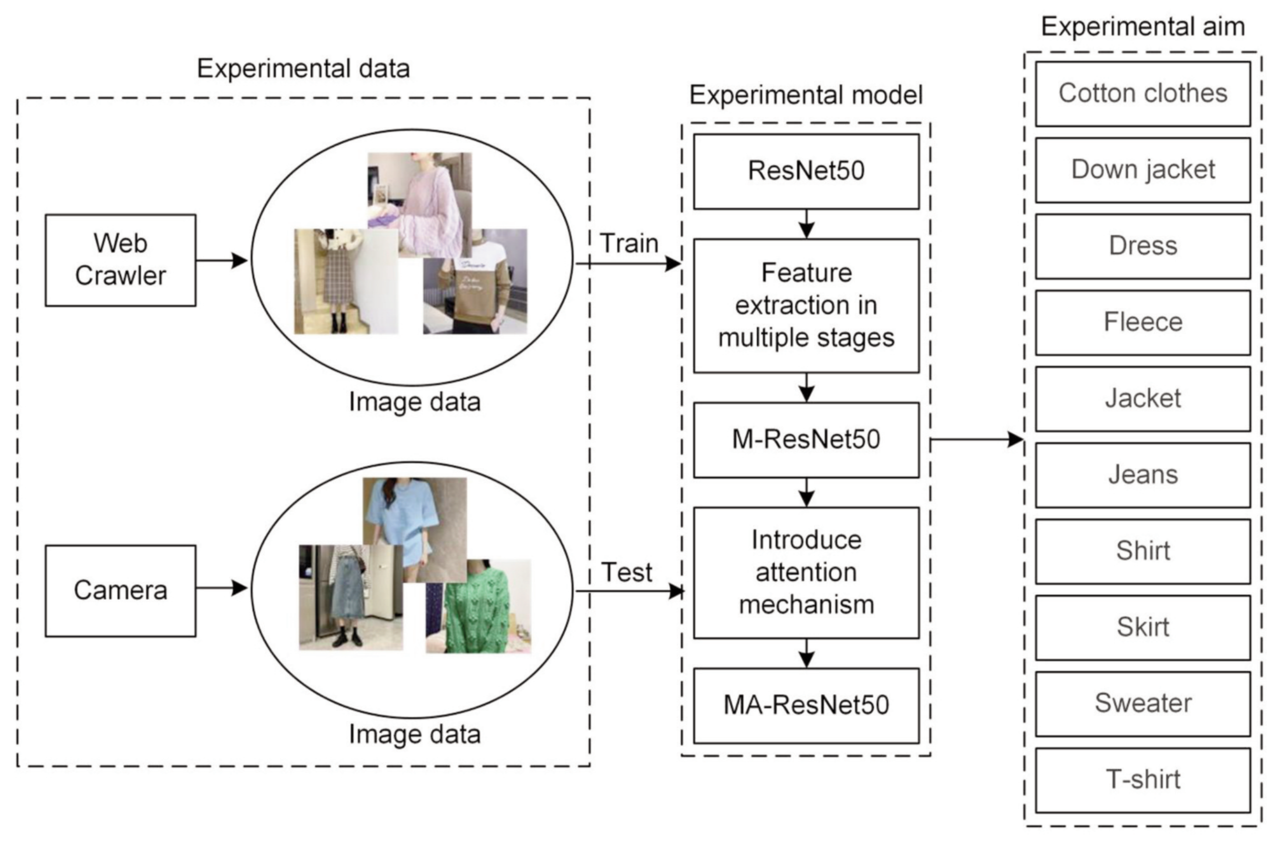

Fig. 1. Experimental process of second-hand clothing classification

In the 4th section, we utilized the trained classification model to identify the second-hand clothing image and compared the accuracy of different models on the second-hand clothing dataset. The advantages and characteristics of this categorization model are summarized in the concluding section.

DATA COLLECTION AND PROCESSING

Users who trade used clothing are likely to need to take pictures of the used clothing as part of the transaction, and they are also likely to utilize online images of the same models as inspiration for their product descriptions. Therefore, when it comes to data collecting, we employ two different methods: the first is to shoot used clothes using photography equipment, and the second is to utilize web crawlers to gather the already-existing photographs of used clothing on online purchasing platforms. We utilized the images captured by the camera equipment as the garment image dataset for the training set and test set because the clothes that are taken with the camera equipment are frequently the ones that require more sorting.

Due to the complicated context in which the garment is positioned, there are several surroundings present as well as a significant amount of noise, distortion, occlusion, and light fluctuations in the image (figure 2). Therefore, it is necessary to manually screen and classify the raw picture data first. The guidelines for screening and categorizing clothes are as follows: if the clothing category is not clearly defined, group conversation is used to decide on the category, and if the group discussion cannot come to a clear decision, the image data is discarded to prevent the problem of misclassification. All manual classification results are further verified thereafter to make sure they are accurate.

Following the collection of all the data, a 500×500 size pixel image of the garment is created by uniformly storing the various pixel sizes of the garment images. The data was then sorted into datasets, with the images being divided into a training set and a training set. The final dataset of second-hand clothing images is shown in table 1.

Table 1

SECOND-HAND CLOTHING DATASET

Class	Train	Test	Sum
Cotton clothes	1412	525	1937
Down jacket	1305	475	1780
Dress	1482	637	2119
Fleece	1487	642	2129
Jacket	1492	666	2158
Jeans	1117	493	1610
Shirt	1472	584	2056
Skirt	821	238	1059
Sweater	1502	619	2121
T-shirt	1276	596	1872
Total	13366	5475	18841

MODEL DESIGN AND IMPLEMENTATION

Overall model architecture

In general, it is believed that the lower-level features of an image can be recovered using CNN's first few layers, while the higher-level characteristics can be extracted using the network's deeper layers. Different features have distinct traits from one another; that is, lower-layer features have a higher resolution and more position and specific details are contained, but since they have undergone less convolution, they are less semantic and noisier. Despite having relatively low resolution and a reduced ability to perceive detail, upper-layer characteristics carry more robust semantic information [13]. It is clear that the properties of CNN retrieved at different levels complement one another.

MA-ResNet50 is built based on ResNet50 [14]. The feature extractor of ResNet50 has five stages, and the spatial size of the feature map is halved after each stage. Considering that deep layers have more semantic information, this paper plugs the attention module at the end of stage 3, stage 4 and stage 5. Figure 3 shows the architecture diagram of MA-ResNet50. The application of the attention mechanism can effectively capture extensive context

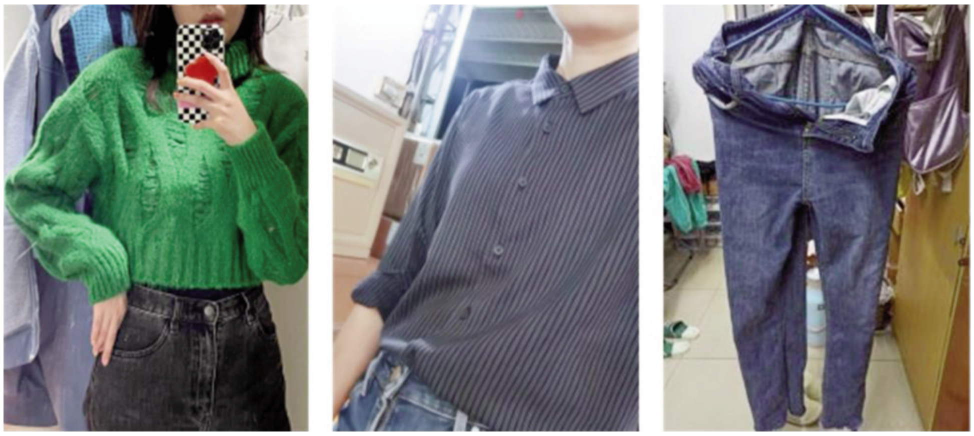

Fig. 2. Sample graph of second-hand clothing

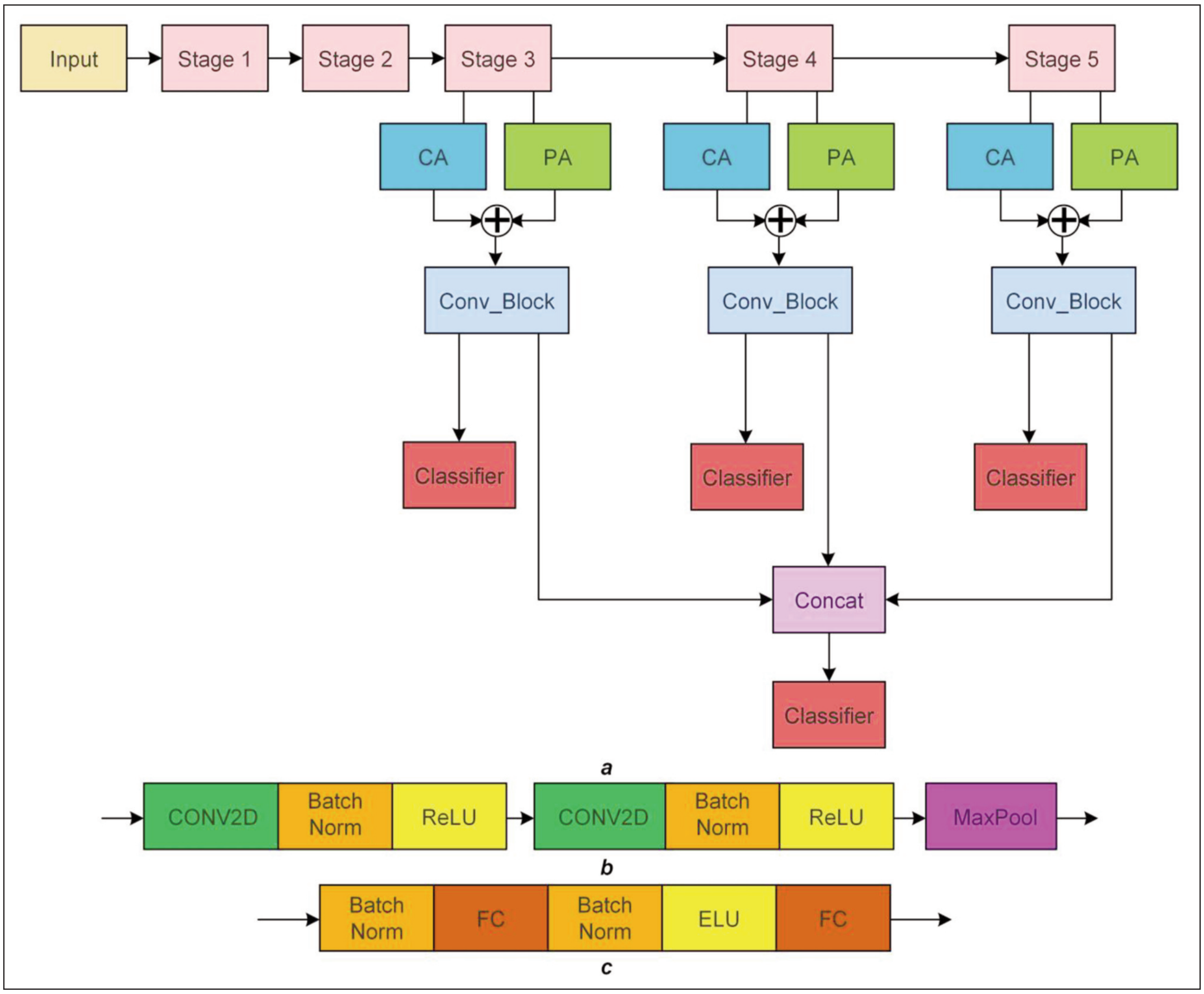

Fig. 3. MA-ResNet50 overall model diagram: *a* – architectural diagram of MA-ResNet50; *b* – Conv_Block architecture in MA-ResNet50; *c* – Classifier architecture in MA-ResNet50

information, and feature extraction in multiple stages can effectively utilize more complementary advantages of CNN features at different levels. Conv_Block represents a combination of two convolutional layers and a max pooling layer. The classifier represents two fully connected layers with a softmax layer at the end.

During training, the output from the corresponding classifier will be used for loss calculation and parameter update. For the training of the output of each stage and the output of the cascaded features, the loss is calculated using the Cross-Entropy (CE) between the ground truth label and the predicted probability distribution:

$$L = -\sum_{i=0}^{C-1} y_i \log(p_i) \qquad (1)$$

where $p = [p_0, \ldots, p_{C-1}]$ is a probability distribution, p_i denotes the probability that the sample belongs to the category i, $y = [y_0, \ldots, y_{C-1}]$ is the one-hot representation of the sample label, as the sample belongs to the category i, $y_i = 1$, otherwise $y_i = 0$.

Channel attention module

The channel attention technique [15] is utilized increasingly frequently in the field of deep learning. Different channel-extracted features contain links with different degrees of tightness, and by collecting feature data from several channels, it is possible to highlight the interrelated image features. The channel attention module is shown in figure 4, *a*. We directly compute the channel attention map $X \in R^{C\times C}$ from the original features $M \in R^{C\times H\times W}$. Specifically, we reshape M into $R^{C\times N}$ and then perform matrix multiplication between M and the transposition of M. Finally, we apply the softmax layer to obtain the channel attention map $X \in R^{C\times C}$:

$$x_{ji} = \frac{exp\,(M_i \cdot M_j)}{\sum_{i=0}^{C} exp\,(M_i \cdot M_j)} \qquad (2)$$

where x_{ij} denotes the influence factor of the i^{th} channel on the j^{th} channel. In addition, transpose X and do matrix multiplication with M and reshape the result as $R^{C\times H\times W}$, then multiply the result by the parameter α and then perform element summation operation with M to obtain the final result $F \in R^{C\times H\times W}$:

$$Fj = \alpha i = 1CxjiMi + Mj \qquad (3)$$

where α gradually learns the weights from 0. From equation 2, the final feature F of each channel is a

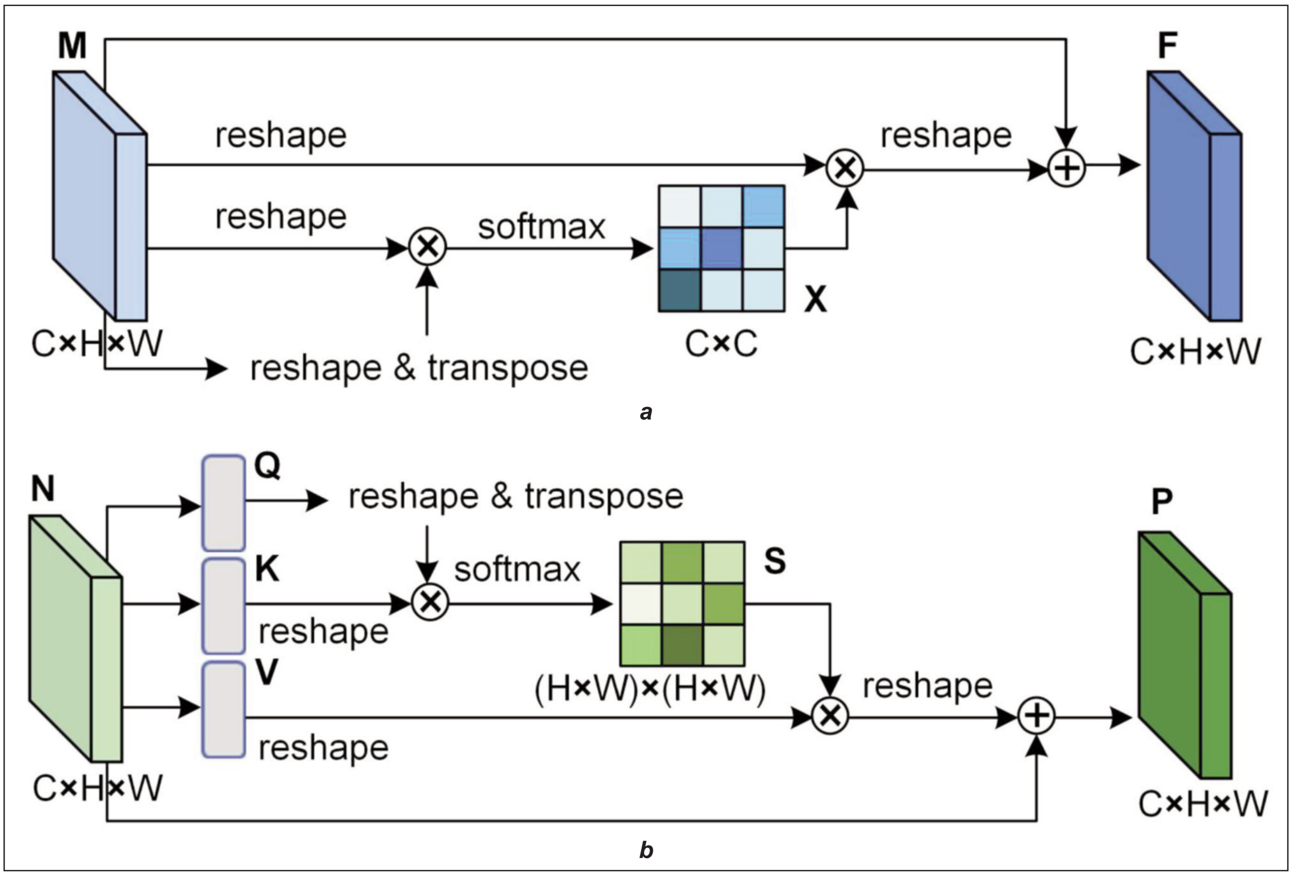

Fig. 4. Attention module: *a* channel attention module; *b* position attention module

weighted sum of the features of all channels and the original features, which models the long-term semantic dependency between feature mappings. It helps to enhance feature discriminability.

Position attention module

We developed the position attention module to model partial features, enriching contextual relations in the process. The position attention module encodes broader contextual information as local features, thus enhancing its representational capability [16]. The position attention module is shown in figure 4, *b*. Given a local feature $N \in R^{C \times H \times W}$, we first input it into the convolution layer to generate two new feature maps Q and K, respectively, where $\{Q,K\} \in R^{C \times H \times W}$. Then we reshape them into $R^{C \times N}$, where $N = H \times W$ is the number of pixels. After that, we perform matrix multiplication between the transpose of K and Q, and apply the softmax layer to compute the spatial attention map $S \in R^{N \times N}$:

$$s_{ji} = \frac{exp\,(Q_i \cdot K_j)}{\sum_{i=1}^{N} exp\,(Q_i \cdot K_j)} \qquad (4)$$

where s_{ij} denotes the influence factor of the i^{th} position on the j^{th} position. In addition, we input the features N into the convolution layer to generate a new feature mapping $V \in R^{C \times H \times W}$ and reshape it into $R^{C \times N}$. Then, we perform matrix multiplication between the transpose of V and S and reshape the result to $R^{C \times H \times W}$. Finally, the result is multiplied by the parameter β and then we perform the element summation operation with N to obtain the final result $P \in R^{C \times H \times W}$:

$$P_j = \beta \sum_{i=1}^{N} (s_{ij} V_i) + N_j \qquad (5)$$

where β is initially 0 and gradually learns to assign more weights [17]. From equation 4, it can be inferred that the resulting feature at each location is a weighted sum of the features at all locations and the original feature. Therefore, it has a global context view that enables superior contextual information aggregation based on the spatial attention graph.

Meanwhile, similar features with higher weights can play a mutually reinforcing role, thus improving intra-class compactness and semantic consistency.

EXPERIMENTS RESULTS AND ANALYSIS

Experiment settings

PyTorch with a version higher than 1.8 was utilized for the testing environment [18], and a single Tesla V100 GPU was employed for each experiment. The model was trained with 13366 images from 10 categories and tested with 5475 images. The ratio of training to test garment image data was 7:3. The only annotations utilized for training are the category labels on the images. During training, input images were resized to a size of 550×550 and randomly cropped to 448×448, and random horizontal flips were applied for data augmentation. During testing, the input images were resized to a size of 550×550 and cropped to 448×448 from the centre. For the optimization of model parameters, Stochastic Gradient Descent (SGD) was used as the model parameter optimizer. The momentum is 0.9 and the weight decay is 0.0005. Additionally, the newly added convolutional and fully connected layers' learning rates were initially set at 0.002 and decreased throughout training using the cosine annealing approach [19]. The learning rate of the pre-trained convolutional layer is kept at 1/10 of the newly added layer. For all models, they are trained for 300 epochs, and the batch size is 16.

Results and analysis

This research compares the CNN classification models VGG16 [20] and ResNet50, which are often used for garment image classification, for analysis to confirm the efficacy of the MA-ResNet50 model. The second-hand clothing dataset is used to train each model until it converges, with all of the parameters being created from scratch. This process makes sure that all of the models are developed under the same circumstances. To visually track changes in the model's classification performance throughout the iterations, the training set is tested once for each iteration that is finished on the test set, and the classification accuracy is output and recorded.

The loss variation of several models on the dataset for used garments is illustrated in figure 5, *a*. It should be noted that the M-ResNet50 network model does not include an attention mechanism and just employs several stages for feature extraction. The M-ResNet50 and MA-ResNet50 models demonstrate a faster drop and convergence in loss values than the other three types of models. The loss curves of the two identical models, M-ResNet50 and MA-ResNet50, are comparable, and the change law and fluctuation range are almost the same. This experimental finding demonstrates that although the structure of the MA-ResNet50 model is more complex than the M-ResNet50 model and there are more parameters and calculations involved, it will not affect the rate at which the model loss value declines and converges when compared to commonly used models.

Figure 5, *b* represents the variance in model accuracy on the dataset for used garments. The classification accuracy is significantly improved by the M-ResNet50 model, which only adds multiple-stage feature extraction branches to the ResNet50 model's structure. This indicates that feature extraction in multiple stages significantly enhances the model's ability to extract features and that the richness of the features contributes to the improvement in classification performance. Figure 5, *b* also shows that the M-ResNet50 model is more stable than the ResNet50 model, with its accuracy curve fluctuating significantly less after roughly 25 iteration cycles. The attention process also strengthened the attractive features while weakening the detrimental ones, which might increase the accuracy of the model while maintaining its stability. According to the experimental results, the MA-ResNet50 model exceeds other deep convolutional networks in terms of classification accuracy.

For comparative trials, this study employed the DeepFashion clothes dataset [21], which has a reduced scene complexity, to further examine the scene applicability and application area of the MA-ResNet50 model. We utilized the DeepFashion dataset, which is a big publicly available apparel dataset established by the Chinese University of Hong Kong's Multimedia Laboratory. DeepFashion contains 800,000 photos, with 50 categories, 1,000 attributes, 4-8 key points, and paired characteristics of the same clothes, making it the largest visual fash-

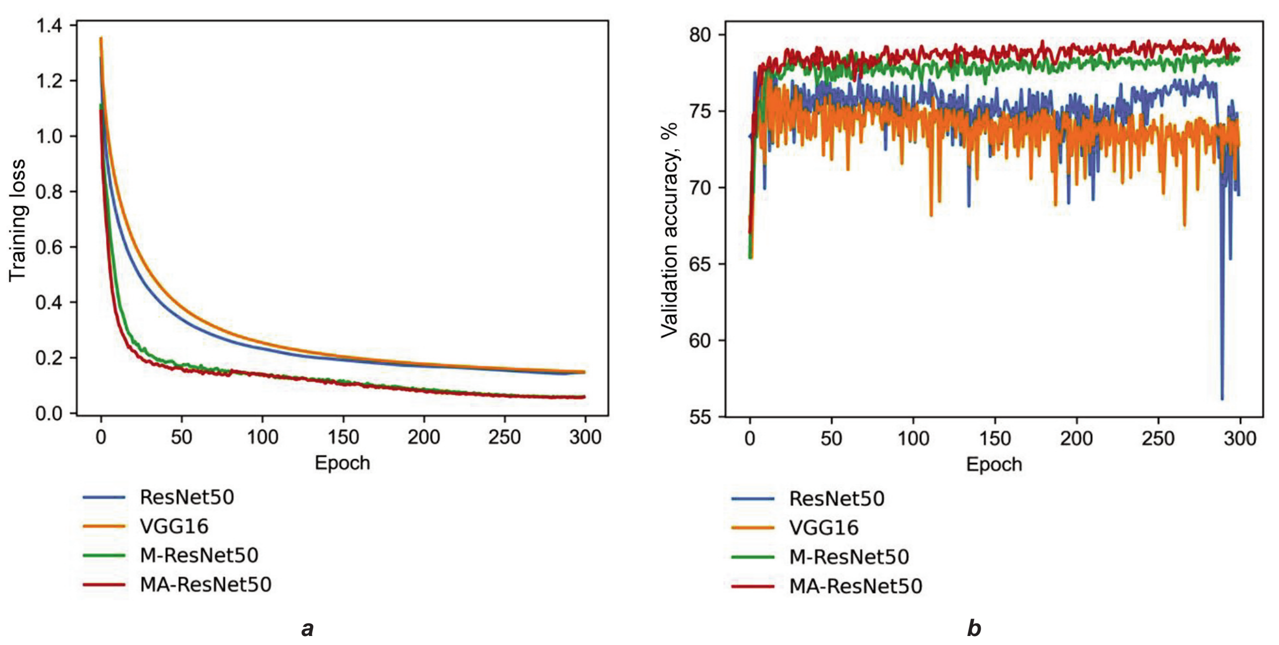

Fig. 5. Comparison of different models on second-hand clothing dataset:
a loss change curves; *b* accuracy change curves

Fig. 6. Sample graph of the DeepFashion dataset

ion analysis database that can be used to compare the classification performance of different deep learning algorithms.

The models were trained on 13112 photos from 10 categories and tested on 4477 images. On this dataset, four network models were evaluated. Table 2 presents the experimental results. In the table, represents the accuracy rate decreased when transitioning from simple to complicated scenarios. The experimental results demonstrate that: (1) The classification accuracy of the MA-ResNet50 model was 1.76% and 1.95% higher than that of the ResNet50 model on the DeepFashion and second-hand clothing datasets, respectively, and (2) When the complexity of the clothing scenes increased from low to high, both the ResNet50 model and the MA-ResNet50 model's accuracy slightly reduced, by 2.72% and 2.53% respectively. The experimental results prove that this model can improve the classification accuracy of clothing images in both simple and complex scenarios, and the classification accuracy of this model is more stable and can maintain high accuracy in complex scenes.

Table 2

COMPARISON OF CLASSIFICATION ACCURACY IN DIFFERENT DATASETS

Network model	Second-hand clothing (%)	DeepFashion (%)	D_A (%)
VGG16	77.46	77.71	0.25
ResNet50	77.74	80.46	2.72
M-ResNet50	78.79	81.19	2.40
MA-ResNet50	79.69	82.22	2.53

CONCLUSIONS

This research proposes a method based on an improved residual network to improve classification accuracy to optimize the effectiveness of second-hand clothes search engines and facilitate clothing transactions on second-hand platforms. Specifically, we introduce a location-attention module and a channel-attention module to capture global dependencies in the spatial and channel dimensions, respectively, and employ several stage branches to acquire CNN features at multiple levels. The attention module can efficiently gather a broad variety of contextual information for feature extraction in stages while also effectively exploiting the complementary capabilities of CNN features at different levels to provide more accurate classification results. The experimental results suggest that the modified residual network classification method outperforms the existing methods. The classification accuracy on the self-built dataset was 79.69% and 82.22% on the DeepFashion dataset, respectively. This result demonstrates this method's excellence and applicability. Furthermore, the solution proposed in this study offers a wide range of potential applications in the classification and recycling of the second-hand clothing market and thus contributes to textile sustainability.

REFERENCES

[1] Yang, X., Li, Q.Z., Wu M., Zhou Y.K., *Circular economy in European Union textile industry chain and key issues of waste textiles treatment*, In: Journal of Textile Research, 2022, 43, 1, 106–112

[2] Zhao, G.L., *Present Situation and Prospect of Comprehensive Recycling Technologies of Waste Textiles in China*, In: Journal of Beijing Institute of Clothing Technology (Natural Science Edition), 2019, 39, 1, 94–100

[3] Zhang, X.Q., Guo, S.C., Liu, D.Z., Wang, C., Liu, J., Gong, Y., *Research progress of automatic sorting method for waste and scrap textiles*, In: Shanghai Textile Science & Technology, 2022, 50, 6, 61–64

[4] Ojala, T., Pietikainen, M., Harwood, D., *Performance evaluation of texture measures with classification based on Kullback discrimination of distributions*, In: Proceedings of 12th International Conference on Pattern Recognition, 1994, 1, 582–585

[5] Lowe, D.G., *Distinctive image features from scale-invariant keypoints*, In: International Journal of Computer Vision, 2004, 60, 2, 91–110

[6] Dalal, N., Triggs, B., *Histograms of oriented gradients for human detection*, In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, 1, 886–893

[7] Joachims, T., *Text categorization with support vector machines: Learning with many relevant features*, In: European Conference on Machine Learning, Springer, Berlin, Heidelberg, 1998, 137–142

[8] Breiman, L., *Random forests*, In: Machine Learning, 2001, 45, 1, 5–32

[9] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., *Gradient-based learning applied to document recognition*, In: Proceedings of the IEEE, 1998, 86, 11, 2278–2324

[10] Akhtar, S.W., Rehman, S., Akhtar, M., Khan, M.A., Riaz, F., Chaudry, Q., Young, R., *Improving the robustness of neural networks using k-support norm based adversarial training,* In: IEEE Access, 2016, 4, 9501–9511

[11] Yu, S., Jin, S., Peng, J., et al., *Application of a new deep learning method with CBAM in clothing image classification*, In: IEEE International Conference on Emergency Science and Information Technology, 2021, 364–368

[12] Bhatnagar, A., Aggarwal, S., *Fine-grained Apparel Classification and Retrieval without rich annotations*, 2018, arXiv preprint arXiv:1811.02385

[13] Yu, W., Yang, K., Yao, H., Sun, X., Xu, P., *Exploiting the complementary strengths of multi-layer CNN features for image retrieval*, In: Neurocomputing, 2017, 237, 235–241

[14] He, K., Zhang, X., Ren, S., Sun, J., *Deep residual learning for image recognition*, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 770–778

[15] Hu, J., Shen, L., Sun, G., *Squeeze-and-excitation networks*, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 7132–7141

[16] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., *Dual attention network for scene segmentation*, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 3146–3154

[17] Zhang, H., Goodfellow, I., Metaxas, D., Odena, A., *Self-attention generative adversarial networks*, In: International Conference on Machine Learning, 2019, 7354–7363

[18] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lerer, A., *Automatic differentiation in pytorch*, 2017

[19] Loshchilov, I., Hutter, F., *Sgdr: Stochastic gradient descent with warm restarts*, 2016, arXiv preprint arXiv:1608.03983

[20] Simonyan, K., Zisserman, A., *Very deep convolutional networks for large-scale image recognition*, 2014, arXiv preprint arXiv:1409.1556

[21] Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X., *Deepfashion: Powering robust clothes recognition and retrieval with rich annotations*, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, 1096–1104

Authors:

TING CHENG[1], YUN ZHANG[2], WEIBO LI[3], JUNJIE ZHANG[4]

[1]Wuhan Sports University, School of Journalism and Communication, Wuhan 430073, China

[2]Wuhan Business University, School of Economics, Wuhan 430056, China

[3]Wuhan Textile University, Wuhan 430073, China

[4]Hubei Key Laboratory of Digital Textile Equipment, Hubei Provincial Engineering Research Center
for Intelligent Textile and Fashion, Wuhan Textile University, Wuhan 430073, China

Corresponding authors:

WEIBO LI
e-mail: leewb@wtu.edu.cn
JUNJIE ZHANG
e-mail: 2007086@wtu.edu.cn