IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 August 2023, accepted 22 September 2023, date of publication 28 September 2023, date of current version 5 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3320270

RESEARCH ARTICLE

Image Multi-Feature Fusion for Clothing Style Classification

YANRONG ZHANG[1,2], **KEMIN HE**[1], **AND RONG SONG**[1]
[1]School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150080, China
[2]Heilongjiang Key Laboratory of Electronic Commerce and Information Processing, Harbin 150080, China

Corresponding author: Yanrong Zhang (zhangyanrong_5@163.com)

This work was supported in part by the 2021 Harbin University of Commerce Teacher Innovation Project Support Plan Project; and in part by the 2023 Harbin University of Commerce Cultivation Project for 'Young Innovative Research Talents'.

ABSTRACT E-commerce platforms are evolving rapidly, and consumers have adapted to shop online more effectively by utilizing product images on shopping websites. In the case of apparel, consumers pay attention not only to the information about practical attributes but also to the information about style attributes. In this study, to specifically analyze the respective style recognition capabilities of the depth feature extractor for studying different clothing styles, an empirical investigation was conducted on four commonly utilized networks: AlexNet, InceptionV3, ResNet50, and VGG16. To complete style classification, we propose the IMF model, which integrates shallow neural networks subsequent to feature engineering to incorporate the benefits of basic features and depth features. This process reinforces the basic features through secondary extraction and merges them with depth features to generate style features, ultimately accomplishing the objective of clothing style classification. Furthermore, a DeepCluster based unsupervised learning approach is used in this study as a comparison. Its classification outcomes are compared with those of the IMF model to authenticate the efficacy of IMF model in this study.

INDEX TERMS Clothing style classification, shallow neural network, VGG16 extractor, deep feature fusion.

I. INTRODUCTION

With the rise of a new generation of e-commerce users, consumers have a higher and more individualistic pursuit of dressing, with many psychological and aesthetic differences. A person's clothing style reveals his or her individual characteristics and attitude. In the demand for clothing, consumers will not only evaluate the functionality of garments but also the overall aesthetic design.

In order to bridge the gap between visual features and aesthetic language, Jia et al. [1] introduced an intermediate layer to form a novel three-level framework. Gao [2] proposed a clothing subjective style recognition method based on distance metric learning and multi-view learning methods to effectively solve the multi-label classification problem of clothing subjective style. Wang [3] refined clothing features and measured the distance between features to solve the clothing style classification problem. Yi [4] construct an AlexNet convolutional neural network model to solve the clothing style recognition and classification task. Zhang et al. [5] used Inception-V3 structure-based transfer learning techniques to accomplish clothing style recognition in their study of clothing style recognition. To solve the problem of fine-grained clothing style recognition, Li et al. [6] proposed an improved bilinear-CNN model for clothing style recognition.

It has been demonstrated that style information can be quantified and recognized and that it is possible to classify apparel images by style. Fewer studies evaluate the aesthetic style methods of apparel product images. In the study of clothing style recognition, only single underlying features (such as color and fabric) as well as the integration of several basic visual features are considered, or only depth features. All of this contributes to a less comprehensive focus on style feature elements and a lack of fine and rich visual comprehension in aesthetic evaluation. Existing clothing datasets

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

lack public datasets with style tags, and clothing datasets in existing studies are primarily line drawings and clothing images from e-commerce platforms.

The main contributions of this paper can be summarized as follows:

- A clothing style classification model based on image multi-feature fusion (IMF) is constructed.
- In order to take into account the advantages of basic features and depth features, shallow neural networks are added after feature engineering, and the basic features are enhanced by secondary extraction and fused with depth features to obtain style features.
- The best features for the quantified style were found by comparing different feature combinations composed of basic features, enhanced features of shallow features, and deep features.
- Four networks, AlexNet, InceptionV3, ResNet50, and VGG16, are trained and tuned to select the appropriate deep feature extractors for building the style classification model.

II. RELATED WORKS

A. DEFINITION OF CLOTHING STYLE CLASSIFICATION

The essence of clothing style classification is an aesthetic evaluation of images, and aesthetic evaluation is a subfield within the field of computational aesthetics research. Florian Hoenig defines "computational aesthetics" as the research of computational methods that can make applicable aesthetic decisions in a similar fashion as humans can [7]. As computer vision continues to evolve, how computers learn to understand "beauty" and output aesthetically meaningful content for aesthetic evaluation has become an important area of research today [8].

B. CLASSIFICATION STUDY OF CLOTHING STYLE ATTRIBUTES

Vittayakorn et al. [9] proposed a method for studying fashion computationally on a large scale using computer vision on the runway and in more real-world environments. Five distinct characteristics, including color, texture, shape, resolution, and style descriptors, were used to identify three distinct visual trends, namely floral prints, pastels, and neon colors, from the runway to street fashion. Jia et al. [1] construct a three-level framework consisting of visual features, image scale space, and aesthetic vocabulary space to create a bridge between visual features and aesthetic language. Gao [2] proposed a clothing subjective style recognition method based on distance metric learning and multi-view learning methods, which effectively solving the multi-label classification problem. Wang [3] extracted clothing features and measuring the distance between features, so as to achieve clothing style classification. Yu et al. [10] introduced aesthetic information highly relevant to user preferences when designing an apparel recommendation system. The clothing style learning recommendation system proposed by Guan et al. [11] can not only identify the main features of clothing design in depth but also identify the style attributes of clothing and the meaning of clothing related to the body. In the study of clothing style, Wu and Wen [12] focused on the material and proposed a clothing style similarity matching algorithm based on the image gradient analysis method, which increased the diversity of research on the clothing style similarity matching algorithm for the majority of fabric images. Yi [4] constructed the AlexNet convolutional neural network model to solve the problem of clothing style recognition and classification, extracting clothing style features through convolutional alternation and down sampling operations, and then classifying clothing style into eight categories. In the study of clothing style recognition, Zhang et al. [5] utilized Inception-V3 structured migration learning to achieve clothing style recognition, which makes full use of clothing style information.

Chen et al. [13] provided a summary of the research in the field of computational aesthetics that has been applied to the understanding of aesthetic style in the aesthetic evaluation of images. Zhang et al. [14] analyzing clothes category classification and attribute recognition from the perspective of clothing attributes. They explore the usefulness of landmarks and texture features, then they proposed to use two streams to enhance the extraction of shape and texture, respectively. Li et al. [6] proposed an enhanced bilinear-CNN model for clothing style recognition as a solution to the problem of fine-grained clothing style recognition. A single-feature pathway bilinear pooling method is designed in which global average pooling and global maximum pooling are used jointly to mine fine-grained features, and the parameters and computation are decreased. Zhu et al. [15] combined deep learning and Kansei engineering to recognize and analyze the kawaii style by taking female fashion clothing as an example. They built a perceptual annotation database and trained two neural models which performed better than manual recognition for the kawaii style. Naka et al. [16] presents an embedding learning framework that uses novel style description features available on users' posts, allowing image-based and multiple choice-based queries for practical clothing image retrieval. Wang [17] proposed an improved framework HSR-FCN and integrated the regional suggestion network and HyperNet network in R-FCN in the new framework which changed the learning approach of image features in HSR-FCN and achieved the higher accuracy in a shorter training time.

C. CLASSIFICATION STUDY OF PRACTICAL ATTRIBUTES OF CLOTHING

Although there are relatively few studies devoted to the task of clothing style classification, the literature on clothing practical attribute classification is voluminous and useful for our clothing style classification research. In the earliest research on practical attribute classification in clothing, the three fundamental features of color, shape, and texture were mainly extracted. Chao et al. [18] combined two traditional basic features as the input features for the clothing image retrieval

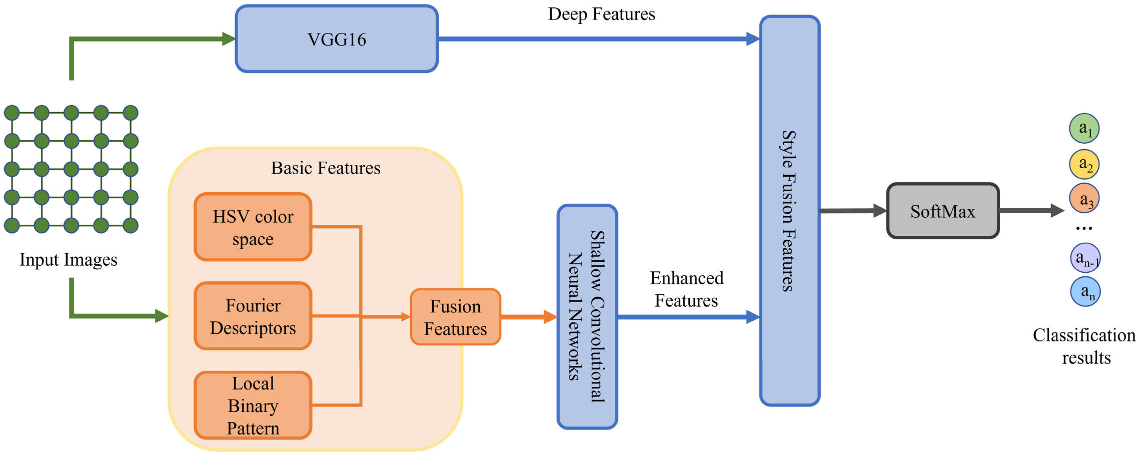

FIGURE 1. **Schematic diagram of IMF model.**

task. Chen and Pan [19] selected garment image color feature such as color histogram and color distance, fused features, and returned retrieval results by calculating the Euclidean distance of image features. Deep learning was introduced to image processing research with the advancement of computer vision in order to extract more abstract features from images. Lao et al. [20] developed a model based on the AlexNet convolutional neural network for recognizing clothing and fashion categories. Elleuch et al. [21] studied the apparel image recognition problem using deep learning and migration learning and validated the model using the ImageNet dataset. To reduce the training time and to improve test accuracy, Xia and Zhang [22] design a novel transfer learning model which is constructed by a pre-trained neural network and novel classification layers with Squeeze-and-Excitation Blocks.

Image feature fusion has a substantial impact on the improvement of the model's classification accuracy. Hou et al. [23] proposed combining the depth features of the residual network extracted based on ResNet50 with color features in order to solve the problem of large color differences when retrieving clothing images. Cheng et al. [24] first extracted global features based on migration learning using enhanced VGG19, then extracted local LBP features and fed them into a shallow convolutional neural network, and finally achieved good classification accuracy. Kayed et al. [25] proposed CNN based LeNet-5 architecture to train parameters of the CNN on Fashion MNIST dataset. Chen et al. [26] proposed a network structure based on the dual attention mechanism to improve the accuracy of final identification by adding different attention mechanisms at different stages of the network to enhance the performance of network features. Liao et al. [27] proposed a multi-scale deep network model which used the modified Inception V3 as the backbone network and expands the perceptual field and embedded the CBAM attention module in the improved backbone network to suppress the interference of noisy information.

III. MODEL

A. MODEL FRAMEWORK

Figure 1 depicts the overall structure of the IMF model. At first, the depth features are extracted using the VGG16 pre-trained model. Then, traditional methods are used to extract color, shape, and texture features for the basic feature extraction. Finally, the shallow convolutional neural network is used to enhance the basic visual features for the second extraction of basic features. The fully connected layer of the convolutional neural network, which serves as the classifier, maps the learned distributed feature representations into the sample label space in order to complete the classification task.

B. STYLE CLASSIFICATION TASK

In response to different occasions and psychological needs of consumers, designers will design various styles of clothing. The designers express the richness of the style through different ways of composing multiple garment attributes. The attribute description of a garment is a unit of a garment [28]. The basic units that make up clothing style usually include color, fabric, decorative elements and so on, Multiple garment attributes together express the style characteristics of clothing, and the formation of a style is precisely the combination of fine-grained attributes [29]. However, some styles of clothing overlap in many attributes, for example, simple style, street style and Korean style are overlapped in some attributes [26]. Therefore, our task is to mine the deep features of clothing and extract the basic attributes to realize the improvement of classification accuracy of style features.

C. DEEP FEATURES EXTRACTION

We compare the networks used in existing research methods for clothing image classification using the data set of this paper for experimental combinations. And specific experimental details are shown in the next section. The results of our classification accuracy experiments after the quantization of styles can be concluded that among the four models, Alex net, InceptionV3, ResNet50, and VGG16, the use of InceptionV3

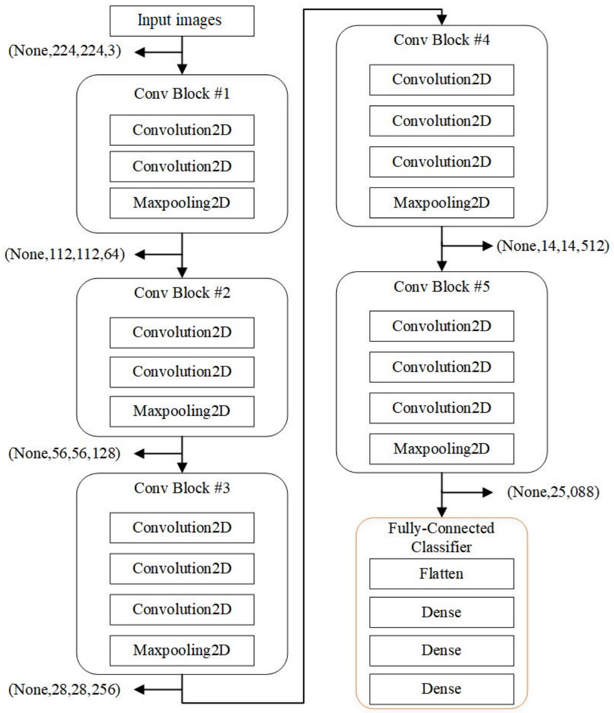

FIGURE 2. **VGG16 structure block diagram.**

and VGG16 as the deep feature extractor for styles gives better final recognition results. Although VGG16 is more time consuming, it has a small advantage in its accuracy. In addition, compared with the other three networks, VGG16 has better generalization ability. Therefore we choose to use VGG16.

The architecture of the convolutional neural network comprises several key components, namely the convolutional base layer, pooling layer, fully connected layer, and output layer. Figure 2 depicts the block diagram of the network structure.

The IMF model uses the optimal effect model parameters obtained from VGG16 training on our dataset. For the clothing style recognition task, the classifier employs the Softmax function and the loss function employs cross-entropy. ReLU is selected as the activation function.

The assessment of image classification involves the utilization of a loss function, which serves to evaluate the efficacy of the model's recognition capabilities and determine the extent of deviation between the predicted and actual values. The present study employs the cross-entropy loss function to ascertain the dissimilarity between the actual output value and the anticipated output value. A smaller cross-entropy value indicates higher accuracy of the experimental model when the actual output value approaches the desired output result. The cross-entropy loss function is defined using (1) for the purpose of solving the multi-classification problem.

$$C = -\sum_{x=1}^{n}[ylog\hat{y} + (1-y)log(1-\hat{y})] \tag{1}$$

where x is the input value, n is the number of batch samples, y represents the label value, and $\hat{y}$ represents the actual output value. In addition, to combat the overfitting problem of high variance, the activation matrix is dropout and L2 regularization is introduced. That is, the sum of squares of all parameters of all layers is added on top of the original loss function.

D. FEATURE FUSION

1) COMPARISON OF CLASSIFICATION ACCURACY OF DIFFERENT FEATURE COMBINATIONS

We compare the classification efficiency of single features, basic fusion features, and shallow features enhanced by secondary extracted features for clothing image classification using the data set of this paper for experimental combinations. The single feature is automatically selected the data with the highest classification accuracy in color, contour, and texture. The comparison of classification accuracy of different feature combinations is shown in Table 1.

The table shows that the style recognition using a single basic feature is low, and it is difficult to get correct classification results from one of the color, shape and texture to determine the style category of the garment. When the three basic features are fused, the style recognition ability has an improvement of 3.92 percentage points in the recognition accuracy value, and the enhanced basic features obtained through the secondary extraction of the shallow neural network have a more effective role in style recognition. Enhanced extraction of basic features by shallow neural network can make a small improvement in style classification effect, and the features extracted by shallow network secondary extraction are more convenient to be combined with deep features. Therefore, we add the step of shallow convolutional neural network feature extraction, so as to achieve better fusion feature extraction ability, laying the foundation for subsequent fusion with deep features.

TABLE 1. **Style classification accuracy for different feature combinations.**

Feature combination	Accuracy
single feature	0.433
fusion feature	0.472
shallow features enhanced by secondary extracted features	0.487

2) COLOR FEATURE QUANTIFICATION

The expression of color features is simply divided into the following two perspectives: (1) choosing a suitable color space; (2) using the corresponding method to express color information as a feature vector. In clothing design, color has a greater impact on its style, and there are more significant differences in the characteristics of color between different styles. The study of the relationship between color and style shows that color can make people feel warm and cold, which brings abstract judgment. RGB color space is not significant in quantifying color information because of the linear combination of three color components, which leads to the connection between the three components and all colors, and the strong correlation between R, G and B components, so the continuous color change in this color space is not significant. And a change in one condition, such as the degree of lightness or darkness, will make all three components change. Therefore, the HSV color space is closer to the human eye's perception of color than the commonly used RGB color space, which is more concerned with the expression of hue, lightness and vividness in color. in the HSV color model, the V component representing lightness information is separated from color information; the H component representing hue and the S component representing saturation are closely related to the way people visually perceive color. Therefore, it is chosen to quantify the color characteristics based on HSV color space.

Figure 3 depicts the process of feature extraction from the input image using the HSV color space. At first, the input is the target image object. Then, the image undergoes a conversion from the initial RGB color space to the HSV color space. The three components, including H, S, and V, are quantized separately, and a color feature histogram is constructed. Finally, the extracted color feature information is represented as a one-dimensional feature vector, thereby accomplishing the objective of quantifying the color feature.

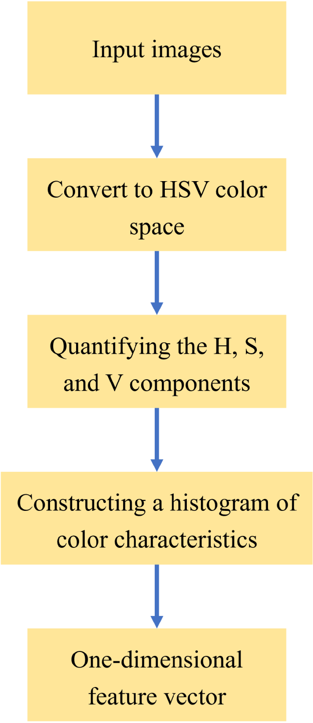
Input images
Convert to HSV color space
Quantifying the H, S, and V components
Constructing a histogram of color characteristics
One-dimensional feature vector

FIGURE 3. Flowchart of HSV color feature extraction.

In this paper, the color features are extracted and visualized for each style, with cute style, simple style, and workplace style images serving as examples for extracting and constructing color feature histograms for each of the three HSV channels. Figure 4 depicts the histogram of color features for the sample of cute fashion; Figure 5 depicts the histogram of color features for the sample of simple fashion; and Figure 6 depicts the histogram of color features for the sample of workplace fashion.

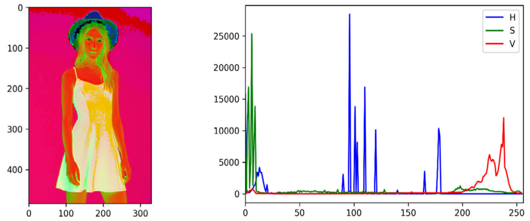

FIGURE 4. Histogram of the color of the cute style example.

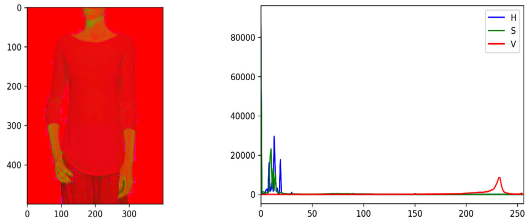

FIGURE 5. Histogram of the color of the Minimalist style example.

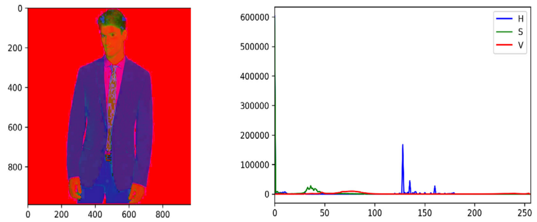

FIGURE 6. Histogram of the color of the workplace style examples.

Each type of image's color feature histograms are compared, analyzed, and summarized by quantifying the three channels of HSV. The following are comparison examples for the cute style, the minimalist style, and the workplace style.

The majority of images of cute-style clothing reveal: The H component is widely dispersed, and there are many values in each range, indicating that the style incorporates the majority of the colors in the spectrum, as cute-style garments typically feature rich and varied color schemes and the hue distribution is not uniform. The large S component in the case of a small H component indicates that the warm portion of the style has greater color saturation than the cool portion, which has less saturation. Because the lovely style is colorful and vibrant, the cooler part has a lower saturation than the warmer part; the V component is larger than the other styles, being greater than the retro style, workplace style, and sports style but lower than the minimalist style.

The majority of minimalist-style images have low H- and S-component values, indicating a single hue and low color saturation due to their simplicity and lightness. The V-component values are significantly greater than the V-component values of the other seven types of style images, indicating that the colors are not only light and elegant but also extremely bright in terms of lightness and darkness.

H component values are concentrated between 240 and 300 degrees, indicating that the style is primarily concentrated in the cooler portion of the spectrum, as the workplace style is primarily formal and sober, and the colors convey a sensible and steady feeling. Greater S component values indicate greater color saturation and less differentiation between styles, such as casual. V component values are dispersed, indicating that the style is concentrated in the part color tone, but there is an even distribution of varying degrees of light and dark.

3) SHAPE FEATURE QUANTIFICATION

Before extracting shape features from a garment image, pre-processing work is required, mainly image pre-processing and edge detection.

a: PRE-PROCESSING WORK

In the pre-processing work, the color image of the garment is converted to a binary image with the main target object in white and the background in black, which is advantageous for the subsequent shape extraction implementation. Figure 7 depicts this pre-processing workflow.

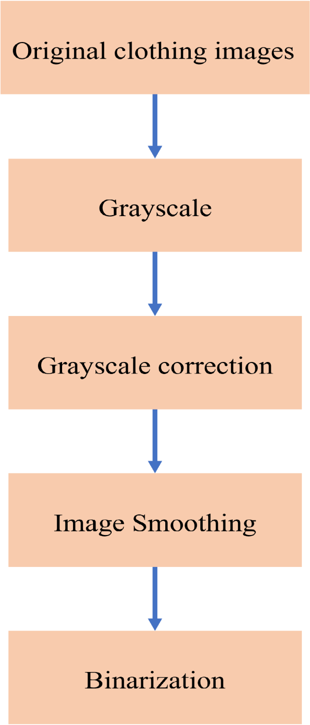

FIGURE 7. Shape extraction pre-processing flow chart.

b: IMAGE CORRECTION AND SMOOTHING

The gray level correction of the grayscale image is performed using the contrast stretching transform function in order to make the distinction between the target subject shape and the background portion of the image more pronounced and thus ensure the quality of subsequent shape feature extraction. The function can amplify the difference between the gray level of the target subject pixel and the gray level of the background pixel [30], and the transformed image after function processing can visually distinguish the subject part from the background region more intuitively and clearly, emphasizing the target region pixel information. Image smoothing is the application of median filtering to remove noise. After median filtering, the image may still contain irrelevant information that must be removed for shape feature extraction; therefore, the garment image must undergo further binarization.

c: BINARIZATION PROCESS

The selection of the threshold value is the primary concern in this process. By choosing an appropriate threshold τ, all pixel values less than τ are set to 0, and all pixel values greater than τ are set to 1. After grayscale correction, the pixel distinction between the subject of focus and the background is clear, making the binarization process straightforward to determine.

After the preceding pre-processing steps, the edge detection algorithm is used to identify and extract the shape of the garment's image's edge contour. The outer contour curve is obtained by extracting the coordinates of the boundary points and tracking the image curve, while the image edge contour is described using the sequentially distributed pixel point coordinates. Multiple areas can be segmented using the coordinate tracking method, which extracts the shape of the image boundary contour through point-by-point tracking based on the connectivity of the boundary.

The Fourier descriptors are the Fourier transform coefficient of the thing contour boundary curve, which can be obtained after analyzing the frequency domain signal of the shape of the thing boundary curve. Among the Fourier series, multiple coefficients directly influence the shape of the closed contour curve, which are defined as Fourier descriptors. When the order of the coefficient terms reaches a sufficient value, the Fourier descriptors can fully extract the shape information and recover the shape of the object. Fourier descriptors are based on a sequence of thing coordinates and have good shape recognition performance.

4) TEXTURE FEATURE QUANTIFICATION

The primary visual presentation of fabric is its color sense, light sense, shape, and texture. For instance, glossy fabric is more exquisite and elegant; soft fabric has better drape and stretch, making people feel natural and comfortable; high-tech coated fabric has a sense of technology; metallic luster and other fabrics have a more avant-garde style. Therefore, the extraction of texture features is required. The circular local binary pattern(LBP), which is widely used in texture analysis and extraction, is employed to describe the local texture features of the image in this paper's research plan.

a: FEATURE VOLUME DESIGN

Based on the LBP histogram, the design feature vectors are added: energy (E), information entropy (H), contrast (C), local smoothness (L), and maximum probability (P_{max}), defined as shown in (2) to (6),

$$E = \sum_{i} P_i^2 \tag{2}$$

$$H = -\sum_{i} P_i log P_i \tag{3}$$

$$C = \sum_{i} i^2 P_i \tag{4}$$

$$L = \sum_{i} P_i/(1 + i^2) \tag{5}$$

$$P_{max} = max P_i \tag{6}$$

where P_i represents the frequency of each pattern within the LBP value. E can reflect the texture's consistency, uniformity, and coarseness. If the current area's texture distribution is uniform and the texture lines are thicker, the energy value is greater; otherwise, it is lower. H is used to measure the image texture information, and the value of entropy is proportional to the amount of image texture information. The value of C describes the depth of the texture and can also indicate the clarity of the texture effect, with the clarity increasing as C increases. L describes the rate at which image texture information changes; as the value of L decreases, the local texture change of the image increases, indicating that the local texture information is not smooth; conversely, as the value of L increases, the local texture change of the image decreases, indicating that the local texture information is smooth. The maximum probability P_{max} is used to identify the unit with the most texture occurrences. The greater the value of P_{max} the larger the texture area occupied by the unit and the more consistent the texture.

E. SHALLOW CONVOLUTIONAL NEURAL NETWORK QUADRATIC QUANTIZATION

Processing an image in the convolutional layer entails sliding a given convolutional kernel over the target image in a specific pattern and performing an inner product calculation at each slide position, with the convolutional kernel serving as the weight of the operation. The convolution process can be mathematically represented by (7), which denotes a linear operation of weighted summation of the input image with the convolution kernel's weight coefficients.

$$h(x, y) = \sum_{s=-m}^{m} \sum_{t=-n}^{n} k(s, t) f(x - s, y - t) \tag{7}$$

where $f(x, y)$ represents the pixel value at the (x, y) point in the image, $k(s, t)$ represents the convolution kernel, and $h(x, y)$ obtained after convolution processing represents the feature map obtained after filtering the original image with the convolution kernel. The process of convolution operation is shown in Figure 8.

Suppose the size of the input processed image is $I_w \times I_h$ and the size of the convolution kernel is set to $k_w \times k_h$, the image needs to be expanded by P_w and the top and bottom borders are expanded by P_h. The value of the expanded part is assigned to 0, and the convolution step is S_c. The size of the feature map after filtering the image with the convolution kernel is $H_w \times H_h$. The size of the feature map is calculated by (8) and (9).

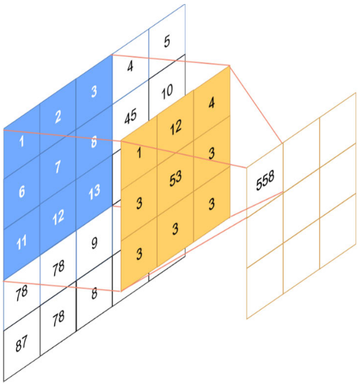

FIGURE 8. Convolution operation process diagram.

$$H_w = \lfloor \frac{I_w + 2 \times P_w - k_w}{S_c} + 1 \rfloor \tag{8}$$

$$H_h = \lfloor \frac{I_h + 2 \times P_h - k_h}{S_c} + 1 \rfloor \tag{9}$$

where S_c represents the distance between two sliding windows when the convolution kernel performs image filtering. When the convolution step size is greater than 1, the convolution layer can achieve the effect of downsampling, which can reduce the network's complexity and computation load. In addition, there is a limit to the setting of the convolution step size. If the step size is set too large, some important information will be lost during the convolution operation. Therefore, when designing the network parameters, discretion is required to ensure the accuracy of the model output information as well as the efficiency of the network operation.

Upon extraction of fundamental characteristics such as color, shape, and texture, these features are subsequently integrated into a shallow convolutional neural network to enhance their basic attributes. The aforementioned network utilizes a shallow architecture to derive a feature vector through vector convolution. This process involves the use of three convolutional layers, each of which comprises ten 3×3 convolutional kernels. Additionally, local connectivity and weight sharing are employed to operate each layer. Equation (10) displays the formula for the shallow convolution calculation.

$$f_{q,h+1} = \mu(\sum_{p} f_{p,h} g_{p,h+1} + c_{q,h+1}) \tag{10}$$

where h represents the number of layers in the convolutional layer; p and q represent the distinct feature maps that are acquired from the convolutional kernels within the current convolutional layer; g and c correspond to the intercepts of the convolutional kernels and functions, respectively; and μ denotes the segmented linear activation function ReLU that is utilized within the convolution layer.

IV. EXPERIMENT

This section commences with the presentation of a dataset that has been labeled with clothing styles, followed by a detailed exposition of the methods employed for data pre-processing. Subsequently, a comparative analysis is conducted to evaluate the efficacy of four distinct depth feature extractors in the context of clothing style recognition and classification. The optimal classification approach is then identified based on the results of this analysis. Finally, the model's style recognition outcomes are assessed and scrutinized in the dataset, followed by a comparative analysis.

a: EXPERIMENTAL ENVIRONMENT

The present study employs PyCharm Community Edition 2021.2.3 within a Windows 10 operating system environment to conduct feature extraction, shallow neural network secondary feature extraction, and deep feature extraction utilizing the VGG16 extractor. Furthermore, the aforementioned environment and tools are utilized to accomplish the IMF model for style identification as well as a comparison model based on an unsupervised learning method. The Tensorflow framework is employed as the primary framework for deep learning. AliCloud servers are utilized in experiments that involve deep convolutional neural networks with the aim of enhancing the speed and efficiency of said experiments. Table 2 presents the pertinent details regarding the experimental setting.

TABLE 2. Experimental environment.

Category	Parameters
Operating System	Windows 10
Hard Disk	1TB
CPU	Intel Core i5-11300H
Memory	DDR4 8GB*2
Operating Environment	64-bit processor
Deep Learning Framework	Tensorflow
Python Environment Software	Anaconda3
Model Implementation Tools	PyCharm Community Edition 2021.2.3

b: EXPERIMENTAL SETUP

This study employs the clothing item images from the publicly available DeepFashion dataset, which was compiled by the Multimedia Laboratory of the University of Hong Kong. During the pre-processing stage, the DeepFashion dataset's clothing images are assigned style labels based on eight distinct styles: workplace style, sports style, elegant style, retro style, simple style, cute and lively, ethnic style, and casual style. After pre-processing, the dataset comprises a total of 6,000 images, with 750 images allocated to each category.

A. DATASET

This study utilizes the clothing product images in the DeepFashion dataset, a public dataset compiled by the Multimedia Laboratory of the University of Hong Kong. The clothing images in the DeepFashion dataset are summarized based on aesthetic aspects to derive style elements, and the data are pre-processed by eight styles: professional business style, casual and comfortable style, elegant and mature style, cute and light style, minimalist style, ethnic style, retro style, and avant-garde style. Sports style and avant-garde style are pre-processed by 8 styles with labeling style tags. The total number of data is 6,000, including 750 data of each category.

1) DATA PRE-PROCESSING

Pre-processing work is needed to ensure the data's quality. First, the uniform size as well as format of the clothing merchandise images helps to guarantee the quality of the experiment, ensure the clear quality of the images to avoid affecting the classification accuracy, and obtain a stronger recognition model; then, the median filtering method is used to smooth the noise, thus improving the image quality and protecting the image edge information; finally, methods such as random data enhancement are used to maximize the use of the limited data set, so that the image data The final method is random data enhancement to maximize the use of the limited data set so that the image data has enough scale to adapt to the clothing style recognition model.

2) EVALUATION METRICS

The following four cases are usually used to measure the effectiveness of the classification task:

True Positive (TP): the number of positive classes predicted to be positive classes;

True Negative (TN): the number of negative classes predicted to be negative classes;

False Positive (FP): the number of negative classes predicted to be positive classes;

False Negative (FN): the number of positive classes predicted to be negative classes.

The accuracy of a multi-classification problem is the ratio of the number of correctly classified samples to the total number of samples, which can be expressed by (11).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

For the multi-category problem in this paper, the confusion matrix needs to treat each category itself as a positive category and all categories other than itself as a negative category, and then update the confusion matrix definition as TP_i, FP_i, FN_i, $TN_i$with i representing the category label, i.e., the ith category. Calculate the $Accuracy_i$ of each category using the newly defined confusion matrix values. The final evaluation needs to aggregate all categories of indicators: macro-average, micro-average, and weighted-average.

TABLE 3. Alex net network implicit layer parameters.

Number of layers	Convolution type	Convolution kernel size	Number of convolution kernels	Step length
1	Conv1	11×11	96	4
2	Max poo	3×3	1	2
3	Conv2	5×5	256	1
4	Max pool	3×3	1	2
5	Conv3	3×3	384	1
6	Conv4	3×3	384	1
7	Conv5	3×3	256	1
8	Max pool	3×3	1	2
9	Fully6	1×1	1	1
10	Fully7	1×1	1	1
11	Fully8	1×1	1	1

The formula for calculating the macro-average accuracy used in the evaluation of the IMF model is defined as follows:

$$\begin{aligned} Accuracy_{macro\_avg} &= \sum_{i=1}^{N} \frac{1}{N} \times Accuracy_i \\ &= \sum_{i=1}^{N} \frac{1}{N} \times \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \end{aligned} \tag{12}$$

B. COMPARISON OF DEPTH FEATURE EXTRACTORS

1) APPAREL CLASSIFICATION BASED ON ALEXNET

The present study employed the AlexNet convolutional neural network architecture, which consists of five convolutional layers, three maximum pooling layers, three fully connected layers, and one Softmax layer. Following three convolutional layers, a maximum pooling layer is incorporated into the architecture to remove extraneous information from the features obtained through convolutional operations. This approach avoids the drawbacks of average pooling and feature blurring while also enhancing computational efficiency and recognition performance. ReLU as an activation function is employed to expedite the convergence of the model. The Local Response Normalization (LRN) layer is utilized to amplify the response for higher values and inhibit the response for lower values, thereby augmenting the network's generalizability. The utilization of dropout has been observed as a potential solution to address the issue of overfitting. The utilization of dual GPU training architectures is a requisite for the network model, with the corresponding network parameters being presented in Table 3.

2) APPAREL CLASSIFICATION BASED ON INCEPTIONV3

The Inception deep learning model is a network structure formed by building the underlying neurons. The network structure is divided into V1, V2, V3 and V4. The network structure is divided into V1, V2, V3 and V4. V1 network structure is shown in Figure 9 [31].

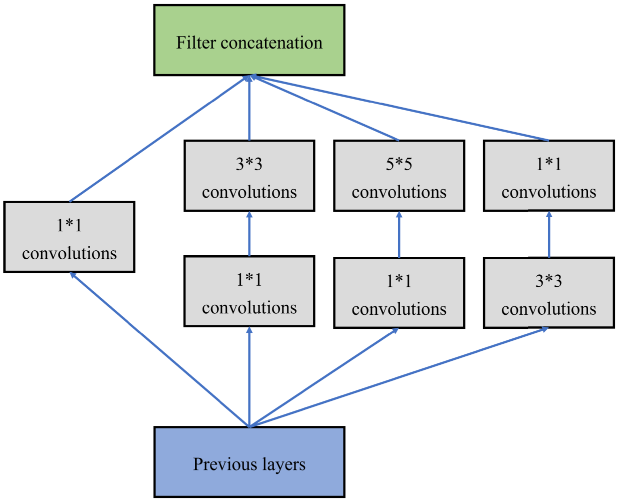

FIGURE 9. Schematic diagram of Inception V1 network structure.

The architecture uses a convolutional overlay pooling operation to enhance the network for size adaptation by adding 1*1 convolution to reduce features and correct ReLU before 3*3, 5*5, and maximum pooling, respectively. Inception V3 is an asymmetric decomposition training method based on V1, which divides n*n into two convolutional layers of 1*n and n*1 to speed up computation. The feature extraction component of the image classification model, which was trained on the ImageNet dataset, is used in this paper to train a new classification layer on it for the feature extraction experiments using the Inception V3 architecture. This part of the training process adjusts the parameters and the final suitable values used are shown in Table 4.

3) APPAREL CLASSIFICATION BASED ON RESNET50

Since the conventional CNN is prone to gradient disappearance with increasing depth in image classification tasks, the literature [32] introduces jumping connection lines to build a deep residual network. The structure of the residual unit is shown in Figure 10.

TABLE 4. Parameter values used in training.

Parameters	Value
Learning Rate	0.01
Step length	8000
Training set batch size	256
Validation set batch size	128

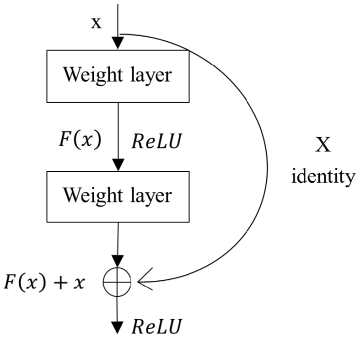

FIGURE 10. Residual cell structure.

After inputting x into the first weight layer in the figure, the mapping function $F(x)$ is obtained. The ReLU activation function is then used to enter the second weight layer, and a jump connection is added to obtain the mapping function $F(x)+x$. This structure enables the shallow features to be mapped directly to the deep layer, thereby facilitating communication between the two layers.

4) APPAREL CLASSIFICATION BASED ON VGG16

The VGG16 convolutional neural network model is a classification model for large-scale data. The network has 13 convolutional layers and 3 fully connected layers. The image classification using VGG16 requires the input image data to be 224*224 pixels, the initial convolutional kernel size is set to 3*3, the stride size and the effective padding size are both 1, and the pooling layer uses a max pooling function of 2*2.

The VGG16's convolutional layer can serve as a means of feature extraction, whereby the initial low-level features are progressively transformed into high-level semantic features through a deeper network architecture. This section utilizes the basic VGG16 network for clothing style recognition, and the classification function uses the Softmax function, whose mathematical expression is shown in (13), which refers to the conversion of an L-dimensional vector to a real vector $\sigma(z)$.

$$\sigma(z)_i = \frac{e^{z}i}{\sum_l^L e^{z}l} \tag{13}$$

5) ANALYSIS OF EXPERIMENTAL RESULTS

Table 5 provides the accuracy of quantitative style identification of these four models. Based on the experimental results of classification accuracy following style quantification, it can be inferred that among the four models,

TABLE 5. Stylistic feature recognition results for different models.

Model	Accuracy
AlexNet	0.677
InceptionV3	0.7
ResNet50	0.689
VGG16	0.706

namely AlexNet, InceptionV3, ResNet50, and VGG16, utilizing InceptionV3 and VGG16 as the depth feature extractors of style yields superior recognition outcomes for the data samples presented in this paper. Despite VGG16's longer processing time, its accuracy is comparatively higher. Furthermore, the VGG16 convolutional neural network architecture exhibits a high parameter count while utilizing a relatively shallow depth, thereby enabling optimal utilization of limited image data and the extraction of more comprehensive and detailed features. In comparison to the other three networks, it exhibits superior generalization capability.

Therefore, we chose to train our dataset using VGG16 and migrate the resulting parameters into a feature fusion model. And the accuracy of VGG16 trained on our dataset is 0.798.

C. EVALUATION AND COMPARISON OF MODEL RESULTS

1) MODEL IDENTIFICATION RESULTS AND EVALUATION

Figure 11 display the final performance evaluation of the multi-category ROC curve. The IMF model performs the classification task of clothing style recognition. 0 7 represent the eight categories of clothing styles in the diagram. Using categories 0, 1, 2, and 3 as examples, category 0 has an AUC of 0.98, category 1 has an AUC of 0.87, category 2 has an AUC of 0.88, and category 3 has an AUC of 0.78. Among these four style classification tasks, the model has the highest accuracy and best results in recognizing and classifying workplace styles; it is weaker in classifying sports styles, elegant and mature styles, and vintage styles than workplace styles; and it is weakest in classifying vintage styles. In general, the ROC curve values for each category surpass those of the

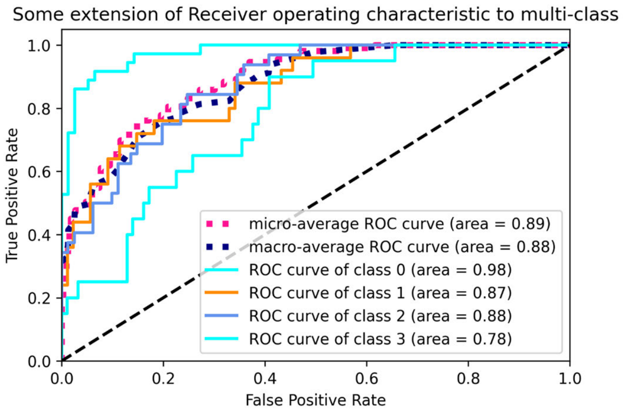

FIGURE 11. ROC curves for category labels 0-4.

invalid ROC curve, indicating the validity of the model in clothing style classification.

2) COMPARISON OF UNSUPERVISED AND SUPERVISED CLASSIFICATION RESULTS

The study compares the effectiveness of two clothing style classification models: the proposed IMF model and the DeepCluster network model that combines k-means and deep learning. The results indicate the reliability of the dataset and highlight the superior performance of the supervised IMF model in recognizing clothing styles.

Deep Clustering for Unsupervised Learing of Visual Features [33] is an unsupervised clustering network learning method based on deep convolutional neural network and combined with k-means clustering to achieve classification using for-label technique.

Following experimental analysis, the unsupervised classification approach needs more data of higher quality, and whether or not the dataset is balanced and the volume of the data will have an impact on the final classification accuracy capability. Table 6 presents the classification accuracy of the unsupervised clustering method utilizing the DeepCluster network on both balanced and unbalanced datasets.

TABLE 6. Classification accuracy based on unsupervised learning methods.

Accuracy	DeepCluster-based unsupervised model	IMF model
Balanced dataset	57.00%	85.58%
Unbalanced dataset	52.37%	85.21%

Visualizing the data as Figure 12, we can visually assess that the unsupervised learning method based on the DeepCluster network has good classification accuracy if the dataset is balanced; if the balanced dataset is unbalanced, the style classification accuracy decreases by almost 5 percentage points, and the overfitting problem is serious. The supervised-style IMF model, in contrast, is unaffected by the unbalanced dataset, demonstrating the model's stability.

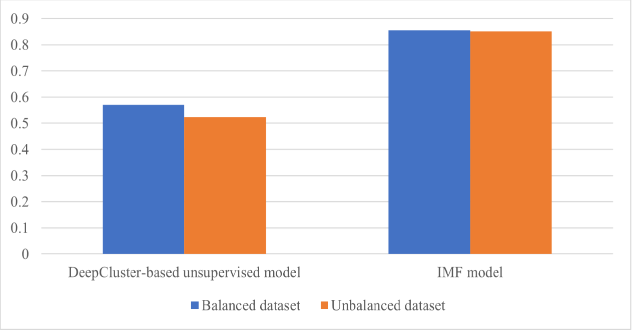

FIGURE 12. Comparison of classification results between supervised and unsupervised methods.

Unsupervised classification methods require both higher data size and quality, and whether the dataset is balanced or not, and the size of the data volume will affect the final classification accuracy ability. And the IMF model for style categorization with supervised methods constructed in our paper is not affected by the imbalance of the dataset. It can be seen that the style classification accuracy of the unsupervised learning network based clustering method is lower than the supervised learning method, and the style recognition accuracy is poor, which proves that the recognition ability of the IMF model is better and reliable in the style recognition classification task.

D. CONCLUSION

There are relatively few studies devoted to the classification of clothing styles, and the challenge is establishing a classification model for style classification that takes into account the benefits of basic and depth features. How to combine subjective and objective defining elements to jointly achieve the evaluation of style characteristics and enhance the precision of style recognition must also be investigated. This paper determines the composition of style features suitable for style classification. And a clothing style classification model utilizing Image Multi-Feature Fusion is proposed and compares it with unsupervised learning methods based on the DeepCluster network to indicate that the model constructed in this paper is valid. Scaling up the dataset to obtain a better performing classification model and extracting fine-grained style features for more accurate recognition are the future directions to further improve the accuracy of clothing style classification. This experiment will hopefully provide useful ideas and methods for scholars and researchers in related fields. In the future, research oriented to clothing style classification tasks will be further investigated and more innovative developments will be made.

REFERENCES

[1] J. Jia, J. Huang, G. Shen, T. He, Z. Liu, H. Luan, and C. Yan, "Learning to appreciate the aesthetic effects of clothing," in *Proc. 30th AAAI Conf. Artif. Intell.*, vol. 30, no. 2 2016, pp. 1216–1222.

[2] S. Gao, "An approach of subjective clothing styles recognition based on distance metric learning and multi-view learning," M.S. thesis, Zhejiang Univ., Hangzhou, China, 2016.

[3] A. Wang, "Research on clothing classification and recommendation method based on image content," M.S. thesis, Kunming Univ. Sci. Technol., Kunming, China, 2017.

[4] G. Yi, "Research on clothing style recognition and recommendation based on convolutional neural network," M.S. thesis, Zhejiang Sci-Tech Univ., Hangzhou, China, 2020.

[5] Y. Zhang, G. Liu, S. Sheng, G. Wang, and L. Wang, "Research on clothing style recognition based on deep learning," *Intell. Comput. Appl.*, vol. 10, no. 5, pp. 14–17, 2020.

[6] Y. Li, R. Huang, and A. Dong, "Fashion style recognition based on an improved Bilinear-CNN," *J. Donghua Univ., Natural Sci.*, vol. 47, no. 3, pp. 90–95, 2021.

[7] F. Hoenig, "Defining computational aesthetics," in *Computational Aesthetics in Graphics, Visualization and Imaging*, L. Neumann, M. Sbert, B. Gooch, and W. Purgathofer, Eds. Eurographics Association, 2005.

[8] H. Duan, "The development process and application of computable aesthetics," *Electron. Technol. Softw. Eng.*, vol. 15, 2019.

[9] S. Vittayakorn, K. Yamaguchi, A. C. Berg, and T. L. Berg, "Runway to realway: Visual analysis of fashion," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 951–958.

[10] W. Yu, H. Zhang, X. He, X. Chen, L. Xiong, and Z. Qin, "Aesthetic-based clothing recommendation," in *Proc. World Wide Web Conf. (WWW)*, 2018, pp. 649–658.

[11] C. Guan, S. Qin, and Y. Long, "Apparel-based deep learning system design for apparel style recommendation," *Int. J. Clothing Sci. Technol.*, vol. 31, no. 3, pp. 376–389, Jun. 2019.

[12] F. Wu and Z. Wen, "A matching algorithm for clothing style similarity based on the gradient of fabric image," *Sci. Technol. Inf.*, vol. 36, 2019.

[13] Y. Chen, Y. Ma, and J. Jia, "Computational aesthetics," *CCCF*, vol. 16, no. 10, pp. 27–34, 2020.

[14] Y. Zhang, P. Zhang, C. Yuan, and Z. Wang, "Texture and shape biased two-stream networks for clothing classification and attribute recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13535–13544.

[15] D. Zhu, X. Lai, and P.-L.-P. Rau, "Recognition and analysis of kawaii style for fashion clothing through deep learning," *Human-Intell. Syst. Integr.*, vol. 4, nos. 1–2, pp. 11–22, Jun. 2022.

[16] R. Naka, M. Katsurai, K. Yanagi, and R. Goto, "Fashion style-aware embeddings for clothing image retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2022, pp. 49–53.

[17] J. Wang, "Classification and identification of garment images based on deep learning," *J. Intell. Fuzzy Syst.*, vol. 44, no. 3, pp. 4223–4232, Mar. 2023.

[18] X. Chao, M. J. Huiskes, T. Gritti, and C. Ciuhu, "A framework for robust feature selection for real-time fashion style recommendation," in *Proc. 1st Int. Workshop Interact. Multimedia Consum. Electron.*, Oct. 2009, pp. 35–42.

[19] Q. Chen and Z. Pan, "Research and improvement of color feature extraction algorithms in the content-based clothing image retrieval system," *Laser J.*, vol. 37, no. 4, p. 7, 2016.

[20] B. Lao and K. Jagadeesh, "Convolutional neural networks for fashion classification and object detection," *CCCV Comput. Vis.*, vol. 546, pp. 120–129, 2015.

[21] M. Elleuch, A. Mezghani, M. Khemakhem, and M. Kherallah, "Clothing classification using deep CNN architecture based on transfer learning," in *Proc. 19th Int. Conf. Hybrid Intell. Syst. Hybrid Intell. Syst. (HIS)*. Bhopal, India: Springer, Dec. 2021, pp. 240–248.

[22] T.-E. Xia and J.-Y. Zhang, "Clothing classification using transfer learning with squeeze and excitation block," *Multimedia Tools Appl.*, vol. 82, no. 2, pp. 2839–2856, Jan. 2023.

[23] Y. Hou, R. He, and J. Liu, "Fusion color feature and depth feature clothing image retrieval algorithm," *Comput. Appl. Softw.*, vol. 37, p. 10, 2020.

[24] W. Cheng, X. Zhang, K. Lin, and A. Li, "Deep convolutional neural network algorithm fusing global and local features," *J. Frontiers Comput. Sci. Technol.*, vol. 16, no. 5, pp. 1146–1154, 2022.

[25] M. Kayed, A. Anter, and H. Mohamed, "Classification of garments from fashion MNIST dataset using CNN LeNet-5 architecture," in *Proc. Int. Conf. Innov. Trends Commun. Comput. Eng. (ITCE)*, Feb. 2020, pp. 238–243.

[26] X. Chen, Y. Deng, C. Di, H. Li, G. Tang, and H. Cai, "High-accuracy clothing and style classification via multi-feature fusion," *Appl. Sci.*, vol. 12, no. 19, p. 10062, Oct. 2022.

[27] L. Liao, S. Zhang, Z. Li, P. Yuan, and Y. Yang, "Clothing classification method based on convolutional network and attention mechanism," *Proc. SPIE*, vol. 12285, pp. 348–358, Jun. 2022.

[28] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool, "Apparel classification with style," in *Proc. 11th Asian Conf. Comput. Vis. (ACCV)*. Daejeon, Korea: Springer, Nov. 2013, pp. 321–335.

[29] Y. Seo and K.-S. Shin, "Image classification of fine-grained fashion image based on style using pre-trained convolutional neural network," in *Proc. IEEE 3rd Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2018, pp. 387–390.

[30] X. Zhao, *Modern Digital Image Processing Technology Improvement and Application Cases Detailed*. Beijing, China: Beihang Univ. Press, 2012.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

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

[33] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2018, pp. 139–156.

YANRONG ZHANG received the Ph.D. degree in engineering from the Northeast Asia Service Outsourcing Mobility Station, in 2016.

She has been engaged in scientific research and teaching and has accumulated richer scientific research experience in artificial intelligence and data mining. She also participated in six projects at provincial and departmental levels. She has published more than 20 papers in important domestic and international journals and international academic conferences as the first author and the corresponding author, including more than ten SCI/EI retrieved papers and two core Chinese journals of Peking University. She has published one textbook as an associate editor. She received the Second Prize of the Heilongjiang Province Science and Technology Progress Award.

KEMIN HE received the B.S. degree in computer science and technology from the Shandong University of Technology, Shandong, China, in 2021. She is currently pursuing the master's degree with the School of Computer and Information Engineering, Harbin University of Commerce, China, under the supervision of Prof. Yanrong Zhang. Her current research interests include natural language processing, classification, and their applications.

RONG SONG received the B.S. degree in computer and information engineering from the Harbin University of Commerce, Harbin, China, in 2019, and the M.S. degree from the School of Computer and Information Engineering, under the supervision of Prof. Yanrong Zhang, in 2023. Her research interests include computer vision and image processing.