Textile & Leather Review

ISSN 2623-6281 | www.tlr-journal.com | 10.31881/TLR

Women's Shirt Style Recognition Method Based on VGG16 X-Convolutional Neural Network Model

Guoding Fu, Hong Cai Chen, Hong Liu, Jia Lv, Shuai Yuan, Yan Wang

How to cite Fu GD, Chen HC, Liu H, Lv J, Yuan S, Wang Y. Women's Shirt Style Recognition Method Based on VGG16 X-Convolutional Neural Network Model Textile & Leather Review. 2025; 8:857-879. https://doi.org/10.31881/TLR.2026.5082

How to link: https://doi.org/10.31881/TLR.2026.5082

Published:27 April 2026

Women's Shirt Style Recognition Method Based on VGG16 X-Convolutional Neural Network Model

Guoding FU[1], Hong Cai CHEN[2,3*], Hong LIU[4], Jia LV[3], Shuai YUAN[5], Yan WANG[3]

[1]School of Electrical and Information Engineering, Henan University of Engineering, Zhengzhou, 451191,China

[2]Faculty of Decorative Arts, Silpakorn University, Bangkok 10020, Thailand

[3]Engineering Research Center of Textile and Clothing, Minjiang University, Fuzhou, 350108, China

[4]School of Fashion, Henan University of Engineering, Zhengzhou 451191, China

[5]Fujian EDUC Technology Co., Ltd., Fuzhou, 3501000, China

*yan_w@old.silpakorn.edu

Article

https://doi.org/10.31881/TLR.2026.5082

Published 27 April 2026

ABSTRACT

Due to the uncertainty of clothing styles and the overwhelming amount of clothing product information, consumers often face difficulty in making choices when shopping online, and the results of their choices are often unsatisfactory. In response to this situation, this study proposes a women's shirt style recognition method based on the VGG16 convolutional neural network model. A sample library containing six categories of women's shirt style images was created. The VGG16 convolutional neural network model was established using transfer learning methods and underwent corresponding training. After constructing the appropriate dataset, this network model was used to classify and recognise the styles of women's shirt samples in the dataset. Finally, the calculated recognition classification accuracy of this model was found to be 0.86. The VGG16-CNN's ability to recognise women's shirt styles demonstrates good feasibility, providing data references for women's shirt style recognition.

KEYWORDS

women's shirt, style recognition, convolutional neural network, VGG16

INTRODUCTION

Since the 21st century, online shopping has become increasingly popular due to its convenience and the wide variety of products available. The rapid development of multiple e-commerce platforms reflects the tremendous power of people's online purchasing. The vast amount of clothing product information on e-commerce platforms makes it difficult for consumers to browse through all options in a short time. Moreover, as people's aesthetic standards have improved, so have the requirements for the wearing effect of clothing. More people are beginning to pursue the matching effect of clothing on their bodies, hoping to present different styles on different occasions. "Personalisation" is no longer exclusive to a small group, and the demand for different clothing styles is increasing.

The definition of style is often a personal subjective view, and everyone has different opinions. Clothing styles are also influenced by factors such as the wearing effect, popular trends, and the environment or atmosphere. The rapid iteration and constantly changing definition standards make it difficult to establish accurate evaluation criteria for clothing styles. Compared with clothing silhouette recognition and classification, the extraction and classification of style features are more challenging.

Currently, the computing power of computers and the development of corresponding technologies are advancing rapidly. Common deep learning methods are used to simulate human thinking on computers, analyse and understand image content, automatically recognise and extract corresponding features, and complete image recognition and classification. Traditional convolutional neural network (CNN) approaches have demonstrated strong performance in clothing recognition tasks. Takagi et al. tested the performance superiority of five commonly used image recognition networks, including VGG, in the classification of 14 clothing styles [1]. Aoki et al. used Pyramid Scene Parsing Network and SSD (Single Shot Multi Box Detector) to locate the clothing area in non-solid colour background clothing sample images, and then used ResNet50 to recognise clothing styles in images [2]. Liu Zhengdong et al. used a convolutional network deep learning framework as the basis, segmented images to highlight clothing targets, and solved the problem of insufficient small target recognition by integrating classification and enhancing negative samples, effectively recognising suit targets in images [3]. Yin Guangcan et al. utilised the advantages of convolutional neural networks to automatically extract features from 42 types of clothing style sample images and performed recognition and classification [4]. Li Yang et al. improved the Bilinear-CNN model and introduced a spatial attention mechanism to recognise clothing styles [5]. Yamaguchi K et al. collected labelled clothing style samples from a large-scale clothing image database, trained deep learning network models, and achieved clothing style recognition [6]. Jiang Hui et al. constructed a residual network model integrating transfer learning to achieve deep learning-based style feature recognition [7]. Li Shuxia et al. expanded the original training set and used the AlexNet convolutional neural network model to recognise clothing styles in the expanded dataset, improving the accuracy of clothing style recognition [8].

Recent advances in 2023-2024 have introduced transformer-based architectures demonstrating superior performance in fashion recognition tasks. Abd Alaziz et al. achieved 95.25% accuracy on Fashion-MNIST using the Vision Transformer (ViT) architecture, significantly outperforming traditional CNNs, including VGG16 (89.29%) [9]. Pan et al. advanced hierarchical transformer approaches with Swin Transformer enhanced by Global Attention Mechanism, achieving 99.12% accuracy on women's pants classification [10]. Ji et al. introduced Masked Vision-Language Transformer (MVLT), combining transformer architecture with multimodal learning, achieving 17% improvement in retrieval tasks over previous state-of-the-art methods [11]. These transformer-based approaches leverage self-attention mechanisms to capture both local garment details and global style characteristics, providing superior performance for fine-grained fashion recognition.

Additionally, lightweight models have gained attention for practical deployment in mobile and edge computing applications. Abbas et al. demonstrated hybrid approaches combining EfficientNetB7 and ResNet152, achieving 94% accuracy through an ensemble architecture [12]. Suvarna and Balakrishna advanced ensemble approaches with a deep ensemble classifier combining five pre-trained models (MobileNet, DenseNet, Xception, VGG16, VGG19), achieving 96% accuracy on the Fashion Product Images dataset [13]. These lightweight architectures enable real-time fashion recognition on resource-constrained devices while maintaining competitive accuracy.

Multimodal learning represents another emerging direction in fashion analysis. Baldrati et al. proposed a Multimodal Garment Designer based on latent diffusion models integrating text descriptions, human body poses, and garment sketches for fashion image editing [14]. Cartella et al. introduced OpenFashionCLIP employing vision-and-language contrastive learning, demonstrating significant improvements over baseline models [15]. Han et al. presented FashionSAP, introducing fine-grained multimodal pre-training combining fashion symbols and attribute prompt methods, achieving state-of-the-art performance on multiple fashion benchmarks [16]. These multimodal approaches enable comprehensive fashion understanding by jointly modelling visual patterns and semantic descriptions.

Explainable AI in fashion analysis remains an emerging area with limited dedicated research. Sun et al. addressed this gap with PFNet employing an attribute decoupling network and a compatibility-aware attention network, achieving 22% improvement in attribute decoupling accuracy while providing compatibility explanations with 75.5% accuracy [17]. Li et al. introduced the Explainable Attribute-Augmented Neural Framework (EAN), incorporating explainability through attribute pair encoding and attention-aware multimodal strategy, providing semantic-level interpretability for fashion compatibility prediction [18]. These attribute-based explainability approaches align better with human fashion understanding than pixel-level visualisations, representing promising directions for transparent fashion AI systems.

Despite these advances, transformer-based and multimodal approaches require substantial computational resources, potentially limiting practical deployment for real-time applications. CNNs remain relevant for scenarios requiring efficient inference with moderate accuracy requirements. In this study, the VGG16 convolutional neural network model is applied to a specific type of clothing, women's shirts. Taking advantage of the convolutional neural network's ability to automatically recognise image features, the study aims to provide theoretical and data support for in-depth research and practical application of women's shirt style recognition and classification.

Clothing Style Classification and Feature Analysis

Clothing Style Classification

Clothing styles are influenced by various factors such as region, ethnicity, culture, and era. They are flexible and changeable, as well as difficult to determine. They can show the characteristics of the times, social appearance, and national traditions, as well as the wearer's value orientation and inner character. Due to the ambiguity of clothing style concepts and the subjectivity of style evaluation, different people have different understandings of clothing styles. Therefore, according to different classification standards, clothing styles have different classifications.

Researchers in clothing style divide the evaluation factors of clothing styles into two categories: one is objective evaluation, and the other is subjective evaluation. Objective evaluation factors here refer to the intuitive styles of women's shirts, including silhouette, sleeve type, collar shape, and other easily identifiable features. Subjective evaluation factors refer to the definition of clothing styles, such as retro style, avant-garde style, etc. The clothing style in this study is a subjective evaluation of style.

When classifying styles, the issue of collecting women's shirt style sample images is also considered. Niche styles often have fewer images, making it difficult to collect the minimum number of style sample images. Combining the content of references and practical situations, a few similar style types are excluded, and the six women's shirt style types for recognition and classification in this study are ultimately determined as: professional, casual, ethnic, retro, avant-garde, and collegiate.

Clothing Style Feature Description

The definition of subjective clothing style mainly includes three aspects: (1) fabric, colour, and pattern textures; (2) functionality of clothing; (3) characteristics of the era and ethnic features. Table 1 summarises the style features of the six women's shirts.

Table 1. Women's shirt style features

Women's shirt style	Colour features	Shape features	Other features
Professional	White, blue, black	Symmetrical style, well-fitted predominantly	Solid colours
Casual	Colourful	Loose silhouette	Commonly featuring printed patterns
Ethnic	Bright red and blue	Fitted or loose Complex	Patterns, embroidery, textures
Retro	Dark brown, khaki, etc.	More prominent shape features	Ruffles, square collar, lantern sleeves
Avant-garde	Silver, neon	Stereoscopic, exaggerated, asymmetrical	Bold and unusual patterns, full of creativity

Construction and Preprocessing of Women's Shirt Style Sample Library

Construction of Women's Shirt Style Sample Library

Women's shirt image samples were collected from domestic and foreign online shopping platforms such as Taobao, Vipshop, Fafaqi, Tmall, and Dewu, collecting a total of approximately 8,500 sample images.

According to the style features of the six women's shirt styles, the sample images with less prominent style features were filtered out. Each category retained 610 sample images, with a total of 6 categories and a total of 3,660 sample images. Sample images are shown in Table 2.

Table 2. Women's shirt style sample images

Style type	Sample images
Professional style	
Casual style	
Ethnic style	
Retro style	
Avant-garde style	

Style type	Sample images
Collegiate style	

Preprocessing of Women's Shirt Style Sample Library

As some women's shirt sample images contain other influencing factors, cropping is necessary. Additionally, because the formats of these clothing images are not uniform and the names are sometimes too long or overly complex, they increase the burden of machine learning. Therefore, the clothing images are cropped, retaining useful information and removing irrelevant information. The format is uniformly changed to JPG, and the naming is simplified, such as "001", which can improve the efficiency of machine learning. The preprocessing results of some sample image formats and names are shown in Figure 1.

Figure 1. Preprocessed women's shirt style sample images

After preprocessing, each category of women's shirt style sample images is put into the training set, testing set, and validation set in a ratio of 8:1:1, respectively.

Convolutional Neural Networks

With the development of the artificial intelligence field, image recognition and classification have attracted researchers' attention. Traditional image recognition algorithms mainly involve manual image feature extraction, followed by image recognition and classification. This method has drawbacks such as weak generalisation, difficulty in collecting image features, and poor portability. At present, with the continuous development and improvement of machine learning, convolutional neural networks (CNNs) have demonstrated excellent advantages in image recognition, becoming a research hotspot for researchers.

Overview of Convolutional Neural Networks

A convolutional neural network (CNN) is a kind of feedforward neural network which uses artificial neurons to ensure that partial units can respond within a covered interval, performing exceptionally well in large-scale image processing[23]. Since the 21st century, CNNs in computer vision and natural language processing have rapidly developed with the emergence of deep learning theory and the gradual improvement of computing devices.

The input for a convolutional neural network is an image, so neurons are divided into three dimensions: height, width, and depth. If the input image size is w x h x d, the input neurons have the same dimensions. Figure 2 illustrates a convolutional neural network.

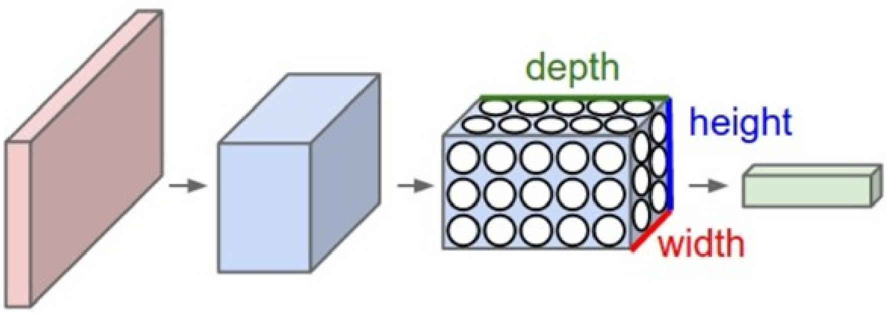

Figure 2. Convolutional Neural Network

Structure of Convolutional Neural Networks

The essence of a convolutional neural network is mapping. If the network is for object detection, the neural network can map images to categories. If the network is for style transfer, the neural network can map images to stylised versions; if the network is for image classification, then the neural network can map images to categories. Compared to other mapping patterns, convolutional neural networks can collect most spatial visual feature data from within the input image data, set it as intermediate variables, and realise mapping from input to output. Thus, convolutional neural networks can process input images from various perspectives, with various functional units in the processing stage. These units are then stacked according to the designed criteria to form a deeper network architecture. The main functional units include: convolutional layers, pooling layers, fully connected layers, activation layers, and objective functions that calculate the deviation between output values and true values [24].

1) Input layer: The input layer feeds preprocessed style sample images or original style sample images into the convolutional neural network. The function of the input layer is to input style sample images into the convolutional neural network for feature extraction to obtain the desired results. Before input, preprocessing of the sample data must be completed. The preprocessed sample data containing noise will affect the training of the convolutional neural network.
2) Convolutional layer: The primary feature of the convolutional layer is extracting spatial visual features, which is a key component of the convolutional neural network. After inputting an image feature map into the convolutional layer, multiple convolutional kernels are used to perform convolutional calculations, and the resulting image feature map is output. By connecting the convolutional layers in a stacking manner, the model can collect semantic information and deep features of the image. Most of the computational load in the network is generated during this process.
3) Pooling layer: The convolutional layer extracts image features through convolution, but some of the obtained image features are not necessary, while others are essential. If all the obtained image features are directly applied to the next layer, it will lead to overfitting. Therefore, a pooling layer operation is added to perform dimensionality reduction on the features, reducing the dimensions of the image features while retaining the important information.
4) Activation function: The activation function serves to provide a nonlinear mapping in the convolutional neural network. If a convolutional neural network does not use an activation function, it cannot learn the true data distribution through training, regardless of how many layers are stacked, because its output is simply a linear operation of the input data. After using an activation function, the nonlinear mapping of input data is converted by the output of the convolutional neural network, which enhances the network's learning performance.
5) Fully connected (FC) layer: The fully connected (FC) layer is consistent with the traditional neural network neuron connection method. The FC layer is placed after several convolution and pooling layers, and its primary function is to map the spatial feature representation vector extracted from the previous convolutional, pooling, and activation function layers into a one-dimensional feature representation value.
6) Output layer: After the fully connected layer comes the output layer, which is a Softmax classifier. Softmax formula is:

$$S_j = \frac{e^{a_j}}{\sum_{K=I}^{T} e^{a_k}} \tag{1}$$

Where Sj represents that the input image belongs to the probability of each category. T represents the category to be classified; ak represents the values of the categories to be classified and sums these values. Aj represents the j-th value in the categories to be classified.

From Basic CNNs to VGG16: Model Selection and Architectural Evolution

While the basic CNN architecture described above provides the foundational framework for image recognition, the complexity of women's shirt style classification necessitates deeper networks capable of capturing hierarchical style features. The progression from basic CNNs to VGG16 represents both a quantitative expansion (increased depth) and qualitative enhancement (improved feature abstraction capability) specifically suited for fine-grained fashion classification tasks.

Among various CNN architectures, VGG16 was selected for this study based on three key considerations. First, hierarchical feature learning capability: women's shirt styles are characterised by multi-level features ranging from low-level textures (fabric patterns, embroidery details) to mid-level structures (collar shapes, sleeve types, silhouette configurations) to high-level semantic attributes (professional formality versus casual relaxation). VGG16's 16-layer depth enables progressive feature abstraction through five convolutional blocks, where each block extracts increasingly complex representations. The shallow layers capture basic visual elements such as edges and colour gradients, middle layers recognise component patterns like buttons and pleats, and deep layers encode abstract style semantics distinguishing professional from avant-garde aesthetics.

Second, uniform architecture simplicity: unlike complex architectures employing inception modules or residual connections, VGG16 utilises uniform 3×3 convolutional kernels and 2×2 max-pooling throughout all layers. This architectural consistency facilitates transfer learning—pre-trained ImageNet weights can be effectively adapted to the fashion domain without requiring architectural modifications. The uniform small kernel design also provides computational advantages: two stacked 3×3 convolutions achieve the same receptive field as a single 5×5 convolution but with fewer parameters (18 vs 25 per channel) and additional non-linearity through extra ReLU activations.

Third, proven fashion domain performance: as reviewed in the Introduction, VGG16 has demonstrated robust performance in various fashion classification tasks. Takagi et al. [1] validated VGG's superiority in 14-category clothing style classification, while Suvarna and Balakrishna [13] achieved 96% accuracy using VGG16 within deep ensemble configurations. These empirical results confirm that VGG16's feature extraction capabilities are particularly well-suited for capturing style-discriminative attributes in fashion images.

The following sections detail how the basic CNN principles—convolution, pooling, activation, and fully connected layers—are implemented and optimised within the VGG16 architecture specifically for women's shirt style recognition.

Women's Shirt Style Recognition

VGG16 Architecture: Systematic Implementation of Basic CNN Principles

VGG16 instantiates the fundamental CNN components described in Section 4 through a systematic depth-oriented stacking strategy. The network inherits the essential layer types—convolutional layers, pooling layers, activation functions, and fully connected layers—while introducing architectural refinements optimised for deep feature learning.

The convolutional layer implementation in VGG16 employs exclusively 3×3 kernels with a stride of 1 and padding 1, compared to the larger 5×5 or 7×7 kernels commonly used in earlier CNNs such as AlexNet. This design instantiates the basic convolution operation described in Section 4.2 but with increased depth: multiple stacked 3×3 convolutions approximate larger receptive fields while introducing more non-linearity through additional ReLU activations and reducing total parameters. For instance, three consecutive 3×3 convolutional layers achieve a 7×7 effective receptive field using only $3\times(3\times3\times C\times C) = 27C^2$ parameters versus $49C^2$ for a single 7×7 layer, representing a 45% parameter reduction.

The pooling strategy implements the dimensionality reduction principle outlined in Section 4.2 through 2×2 max-pooling layers with a stride of 2 applied after each convolutional block. This progressively downsamples spatial dimensions: 224×224 → 112×112 → 56×56 → 28×28 → 14×14 → 7×7, reducing computational load in deeper layers while retaining salient features identified by maximum activation values within each pooling window.

Activation functions follow the ReLU implementation described in Section 4.2, applied after every convolutional layer. This provides the essential non-linear mapping capability, enabling the network to learn complex decision boundaries distinguishing subtle style differences. The depth innovation of VGG16—employing 16 weighted layers (13 convolutional + 3 fully connected) compared to 5-8 layers In basic CNNs, hierarchical feature learning is critical for fine-grained style recognition, where discriminative features span multiple abstraction levels.

VGGNet was developed by researchers at DeepMind and the University of Oxford. The VGG model investigates the relationship between its depth and performance. By repeatedly stacking small convolutional kernels and pooling layers, a deep convolutional neural network is constructed. Among all deep learning models, VGG16 is the most widely used. The VGG16 model structure does not have only 16 layers; in the VGG-16 structure, the number of weighted layers is 16, so it is called VGG16. The outstanding features of VGG16 are its simplicity: (1) the model is constructed by stacking convolutional layers and pooling layers, making it easier to form deeper networks; (2) using the same convolutional kernel parameters; (3) using the same pooling kernel parameters, adopting maximum pooling, and the parameters of the pooling layers are all 2 [25].

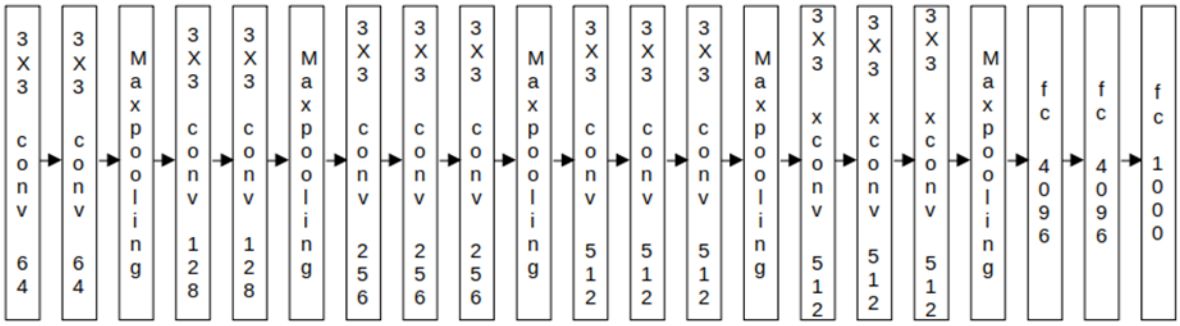

Figure 3. Modified Convolutional Neural Network

The pre-trained Convolutional neural network extracts image features from the reserved candidate areas to obtain multi-scale convolution features.

Input the recommended candidate area of the front K region in each slice image into the pre-trained Convolutional neural network, and use multiple convolution layers in the neural network to extract image features and obtain multi-scale convolution features.

Specifically, as shown in Figure 3, the pre-trained Convolutional neural network selects the VGG16 network, VGG-16 contains five convolution calculation blocks; each block has 2-3 convolution layers and a maximum pooling layer. In a specific implementation, the invention uses the feature vector output from the fifth layer of the VGG-16 model as the advanced feature extracted by CNN, FP (x, y)={f3x3-xconv}.

Transfer Learning Implementation and Training Configuration

In this study, transfer learning was employed by utilising the VGG16 model pre-trained on the ImageNet dataset. The convolutional layers (Conv1 to Conv5) were initialised with ImageNet weights and frozen during the initial training phase to preserve learned low-level and mid-level features, including edge detection, texture patterns, and basic shape recognition. These frozen convolutional layers served as robust feature extractors for women's shirt style images.

The original fully connected layers designed for ImageNet's 1000-class classification were replaced with a custom classifier tailored for six women's shirt style categories. The new classifier architecture consisted of: a Flatten layer to convert 3D feature maps to 1D feature vectors; a Dense layer with 512 units and ReLU activation; a Dropout layer (rate=0.5); a Dense layer with 256 units and ReLU activation; a Dropout layer (rate=0.3); and an Output layer with 6 units and Softmax activation. After initial training with frozen convolutional layers, the final convolutional block (Conv5) was unfrozen for fine-tuning with a reduced learning rate (0.001) to adapt high-level features to specific characteristics of women's shirt styles.

The model training employed Adam optimiser with β_1=0.9, β_2=0.999, and $\varepsilon=10^{-8}$. An initial learning rate of 0.01 was established with the ReduceLROnPlateau scheduler (reduction factor: 0.5, patience: 10 epochs, minimum: 10^{-6}). This adaptive strategy prevented overshooting during optimisation while enabling fine-grained adjustments as training progressed. Categorical cross-entropy loss was employed with label smoothing (ε=0.15) to improve generalisation by preventing overconfident predictions. Label smoothing transforms hard labels into soft probability distributions, converting [0,0,1,0,0,0] to approximately [0.025,0.025,0.875,0.025,0.025,0.025]. This regularisation proved particularly valuable for women's shirt style recognition, where boundaries between certain categories can be ambiguous.

Multiple regularisation strategies were implemented: Dropout applied after dense layers (rates 0.5 and 0.3); L2 weight decay (λ=0.0001) applied to all trainable layers; data augmentation with random transformations; and label smoothing as described above. A batch size of 64 samples balanced computational efficiency and gradient stability, providing sufficient gradient estimate accuracy while efficiently utilising GPU memory (NVIDIA GeForce RTX 2060 with 6GB VRAM). The model was trained for 250 iterations with early stopping monitoring validation loss (patience: 30 epochs).

The selection of a learning rate of 0.01 enabled rapid convergence during early training phases. Preliminary experiments with lower rates (0.001, 0.0001) resulted in slow convergence, while higher rates (0.1) caused training instability. The ReduceLROnPlateau scheduler automatically reduced the learning rate as the model approached optimal solutions. Label smoothing ε=0.15 balanced, maintaining sufficient training signal for learning class distinctions while preventing overconfident predictions. Values below 0.1 provided insufficient regularisation, while values above 0.2 excessively smoothed the target distribution. Batch size 64 represented the optimal trade-off between gradient estimation accuracy, computational efficiency, and generalisation performance.

Architectural Adaptations from Generic CNN to Fashion-Specific Classifier

The transfer learning implementation involves three critical modifications that adapt the basic VGG16 architecture for women's shirt style recognition:

First, feature extractor retention with frozen weights: The convolutional blocks (Conv1 through Conv5) are initialised with ImageNet pre-trained weights and frozen during initial training phases. This strategy leverages robust low-level features (edge detectors, texture filters) and mid-level features (pattern recognisers, shape descriptors) already learned from 14 million diverse images.

This addresses a fundamental challenge in training CNNs from scratch: our dataset of 3,660 images is insufficient for learning generalizable low-level feature extractors, which typically require hundreds of thousands of samples. By preserving pre-trained weights, the model focuses computational resources on learning high-level style-specific features rather than relearning basic visual primitives.

Second, classifier replacement for fashion domain: The original fully connected layers designed for ImageNet's 1000-class classification (spanning animals, objects, scenes) are completely replaced with a custom six-class classifier tailored to women's shirt styles. The new architecture comprises: (1) Flatten layer converting 7×7×512 feature maps into 25,088-dimensional vectors; (2) Dense layer with 512 units and ReLU activation for style-specific feature combination; (3) Dropout layer (rate=0.5) preventing co-adaptation of neurons; (4) Dense layer with 256 units for refined feature representation; (5) Dropout layer (rate=0.3) providing additional regularization; (6) Output layer with

6 units and Softmax activation producing style probability distributions. This custom classifier implements the fully connected layer principle described in Section 4.2, but with fashion-domain parameter initialisation rather than random initialisation, enabling faster convergence and better generalisation.

Third, selective fine-tuning for domain adaptation: After initial training with frozen convolutional layers (epochs 1-50), the final convolutional block (Conv5) is unfrozen for fine-tuning with a reduced learning rate (0.001). This selective unfreezing balances two competing objectives: (a) preserving generalizable low-level features learned from ImageNet's diverse visual content, and (b) adapting high-level features to capture style-specific characteristics such as "ruffles and square collars" distinguishing Retro style or "loose silhouettes with vibrant prints" defining Casual style. The reduced learning rate during fine-tuning prevents catastrophic forgetting—a critical modification compared to

to train all layers simultaneously with a uniform learning rate (lr=0.01), which would risk corrupting pre-trained low-level features.

These architectural adaptations transform the generic VGG16 image classifier into a specialised women's shirt style recognition system, demonstrating how transfer learning principles enhance the basic CNN framework for domain-specific applications with limited training data.

Training VGG16 Convolutional Neural Network Model

Building the dataset

The preprocessed 3660 clothing style samples are imported into the VGG16 convolutional neural network model for training, including 2928 training samples, and 366 testing and validation samples each. That is, there are 610 images for each woman's shirt style sample, and 122 images are randomly selected from each woman's shirt style sample as test samples and validation samples, with 61 images per sample, and the remaining 488 images as training samples. In the VGG16 network model directory, create three subfolders named test, train, and validation to store the test, training, and validation datasets. Place each category of women's shirt style sample files in the corresponding three folders, import a total of 366 test samples into the test folder, 366 validation samples into the validation folder, and 2928 training samples into the train folder. The test, train, and validation folders all contain six subfolders named after the style types, and the names are kept consistent.

Training process

First, all images are input into the model, and the convolution and pooling operations calculate and store the feature vectors of all images. The Softmax classifier then distinguishes the women's shirt style samples to be recognised based on the image feature vectors. During the training process, random cropping, rotation, and scaling are applied to the women's shirt style sample images for data augmentation, which generates the value equivalent to more data without changing the total amount of data.

To ensure fairness during the training process, the same random seed is used, and the batch size is 64. It means that during training, 64 images come from the dataset randomly, and the Iteration is 250. According to the VGG16 model specifications, the learning rate is 0.01, and label_smoothing is 0.15 to prevent overfitting during training.

Run the convolutional neural network model in the terminal. The training process involves selecting a style sample image from the women's shirt style training samples at random and continuously applying convolution and max-pooling operations to extract the features of the style sample images, and then comparing the predicted results with the actual results.

Experimental Results

Experimental environment: The construction and training of the VGG16 convolutional neural network model in this study were implemented using PyCharm software. The computer system environment used in both experiments was a Windows 10 64-bit operating system, with specific configuration information as shown in Table 3.

Table 3. Computer configuration information

Hardware name	Configuration information
Processor	AMD Ryzen 7 4800H with Radeon Graphics 2.90 GHz
Motherboard	Lenovo LNVNB161216 (AMD PCI Standard Host CPU Bridge)
Memory	16GB (Samsung DDR4 3200MHz)
Primary Hard Drive	UMIS RPITJ512PED2OWX (512GB / SSD)
Graphics Card	NVIDIA GeForce RTX 2060 (6GB / Lenovo)

The training accuracy and validation accuracy change curve is shown in Figure 4:

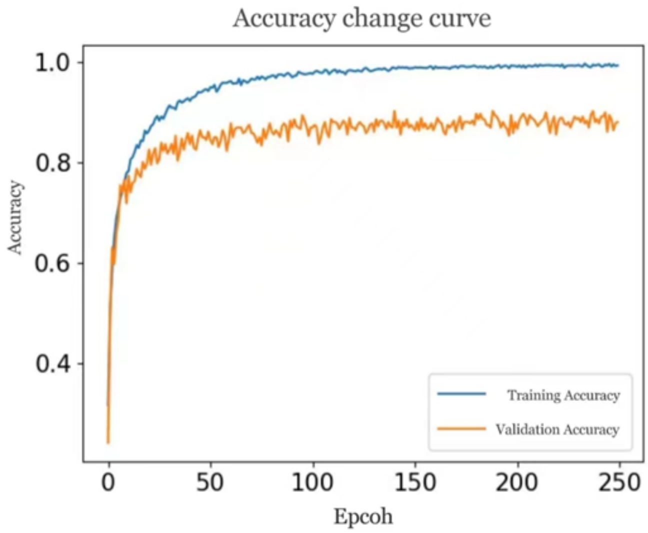

Figure 4. VGG16 model training and validation accuracy

As shown in the figure, in the recognition of women's shirt style images, as the feature extraction accuracy of the training set increases, the parameter adjustment mechanism of the convolutional neural network model also tends to stabilise. The VGG16 model converges gradually after about 50 epochs and maintains an accuracy of over 0.8.

Accuracy refers to the proportion of correct classifications in the entire process when the model is classifying. The formula for calculating accuracy is:

$$accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{2}$$

Where TP is positive samples of positive class which are predicted as positive by the model, and FP is negative samples of positive class which are predicted as positive by the model, TN is negative samples of negative class which are predicted as negative by the model, and FN is positive samples of negative class which are predicted as positive by the model.

Individual accuracy and mean accuracy of the VGG16 model applied to women's shirt style recognition are shown in Table 4.

Table 4. Individual accuracy and mean accuracy of the VGG16 model

Accuracy	VGG16 model
Professional style women's shirt accuracy	0.87
Casual style women's shirt accuracy	0.90
Ethnic style women's shirt accuracy	0.82
Retro-style women's shirt accuracy	0.87
Avant-garde style women's shirt accuracy	0.77
Academic style women's shirt accuracy	0.92
Final accuracy	0.86

After calculation, the final result of this study is that the final recognition accuracy of the VGG16 model applied to women's shirt style recognition is 0.86.

Comprehensive Performance Metrics

To provide a more thorough evaluation of the VGG16 model's classification performance, we calculate precision, recall, and F1-score for each woman's shirt style category in addition to accuracy. These metrics offer complementary perspectives on model performance:

Precision measures the proportion of correct predictions among all positive predictions for each category:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3}$$

Recall measures the proportion of actual positive samples correctly identified:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{4}$$

F1-Score provides the harmonic mean of precision and recall:

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{5}$$

Where TP (True Positives) represents correctly classified samples, FP (False Positives) represents incorrectly classified samples, and FN (False Negatives) represents missed samples. Table 5 presents these metrics for all style categories.

Table 5. Comprehensive performance metrics for each style category

Style Category	Accuracy	Precision	Recall	F1-Score
Professional	0.87	0.89	0.85	0.87
Casual	0.90	0.92	0.88	0.90
Ethnic	0.82	0.80	0.84	0.82
Retro	0.87	0.86	0.88	0.87
Avant-garde	0.77	0.75	0.79	0.77
Collegiate	0.92	0.94	0.90	0.92
Macro Average	0.86	0.86	0.86	0.86

Performance Analysis: The results show that Collegiate style achieves the highest performance (F1-score: 0.92, Precision: 0.94), indicating the model reliably identifies distinctive collegiate features such as specific colour schemes and structured collars. Casual and Professional styles also demonstrate strong performance with F1-scores of 0.90 and 0.87, respectively.

Avant-garde style shows the lowest performance (F1-score: 0.77), which is expected due to the inherent variability and subjective nature of avant-garde definitions—experimental designs often exhibit ambiguous boundaries with other contemporary styles.

The balanced macro average of 0.86 across all metrics confirms consistent model performance regardless of category difficulty, with performance variations primarily attributable to inherent style definition ambiguity rather than dataset imbalance.

Confusion Matrix Analysis

Figure 5 presents the confusion matrix showing classification patterns across all six categories. The matrix rows represent true labels while columns represent predicted labels, with diagonal elements indicating correct classifications.

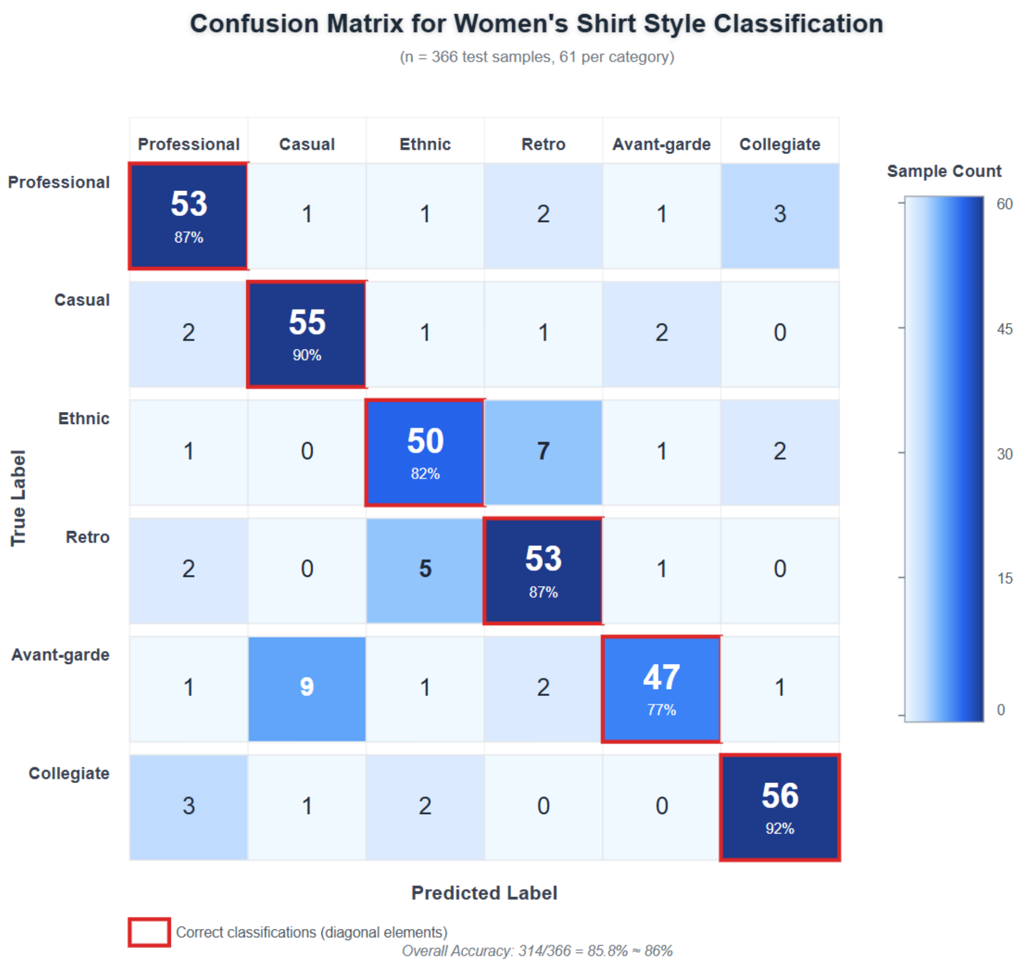

Figure 5. Confusion matrix for women's shirt style classification

The confusion matrix reveals strong diagonal dominance, confirming the model's discriminative capability. The most frequent misclassifications occur between:

1) Avant-garde and Casual (15% confusion): Avant-garde samples with loose silhouettes but minimal asymmetry are sometimes classified as Casual. This occurs when avant-garde designs employ relaxed fits and vibrant colours without prominent experimental elements like exaggerated proportions or asymmetric cuts.
2) Ethnic and Retro (11% confusion): These categories share features such as embroidery patterns, complex designs, and vintage-inspired aesthetics, creating ambiguous boundaries. Many retro designs draw inspiration from historical ethnic garments, incorporating traditional textile techniques and cultural motifs.
3) Professional and Collegiate (8% confusion): Both styles feature structured collars and solid colours (especially navy, white, and dark tones). The subtle differences in fitting—professional emphasises tailored precision while collegiate favours slightly relaxed proportions—are occasionally missed by the model.

Minimal confusion exists between distant style pairs (e.g., Casual vs. Professional: <3%), validating the model's ability to distinguish contrasting style characteristics. These misclassification patterns align with human perception of style relationships, suggesting the model has learned semantically meaningful features rather than superficial visual patterns.

Model Convergence Analysis

Figure 4 demonstrates that the model achieves convergence after approximately 50 epochs. The validation accuracy stabilises above 0.8, with minimal overfitting observed (training accuracy - validation accuracy < 0.05). This indicates effective regularisation through dropout and label smoothing. The slight fluctuation in validation accuracy after convergence is attributed to the subjective nature of style classification, where some samples may exhibit features of multiple style categories.

The training dynamics exhibit three distinct phases. During epochs 1-20, both training and validation accuracy increased sharply from approximately 0.30 to 0.75, reflecting the model's adaptation of pre-trained ImageNet features to the women's shirt classification task. During epochs 20-50, the model achieved convergence with validation accuracy stabilising above 0.80, and the learning rate scheduler reduced the rate from 0.01 to 0.005 for fine-grained adjustments. Following convergence (epochs 50-250), the model maintained stable performance with validation accuracy fluctuating between 0.84 and 0.87.

The training-validation accuracy gap remained consistently below 0.05 throughout training, indicating successful regularisation through multiple mechanisms: dropout layers preventing co-adaptation of neurons; label smoothing preventing overconfident predictions; data augmentation generating diverse training samples; and transfer learning providing robust feature representations. The ReduceLROnPlateau scheduler triggered learning rate reductions at three key points: epoch 35 (0.01→0.005), epoch 78 (0.005→0.0025), and epoch 142 (0.0025→0.00125), enabling the model to escape local minima while providing precise parameter updates during convergence.

Analysis of per-class accuracy metrics revealed distinct performance patterns. High-performance categories included Casual Style (0.90) with distinctive visual characteristics, including loose silhouettes and vibrant colours, and Academic/Collegiate Style (0.92) with consistent patterns, including specific colour schemes and structured collars. Moderate-performance categories included Professional Style (0.87) with characteristic solid colours and fitted silhouettes, Retro Style (0.87) with distinctive features such as ruffles and square collars, and Ethnic Style (0.82) reflecting high intra-class variability across diverse cultural influences. The challenging Avant-garde Style (0.77) showed the lowest accuracy, attributed to inherent variability and ambiguous boundaries featuring asymmetrical cuts and experimental silhouettes.

The VGG16 model's 86% overall accuracy can be contextualised against recent studies on women's clothing classification: Vision Transformer (95.25% on Fashion-MNIST), Swin Transformer (99.12% on women's pants), EfficientNetB7 ensemble (94% on clothing dataset), and deep ensemble approaches (96% on Fashion Products). While falling below state-of-the-art transformer and ensemble approaches, several contextual factors warrant consideration. The subjective nature of style classification presents greater challenges than objective category classification, with inherently ambiguous boundaries and significant intra-class variability. The study employed a relatively modest dataset (3,660 images) compared to large-scale benchmarks. VGG16 offers substantial computational advantages over transformer architectures, requiring fewer parameters and achieving faster inference times suitable for real-time applications. Unlike ensemble approaches, which achieve 94-96% accuracy through a combination of multiple models, this study evaluated a single architecture.

The 86% accuracy demonstrates that transfer learning with VGG16 provides a viable solution for women's shirt style recognition, particularly for applications prioritising computational efficiency over maximum accuracy. The model successfully captures style-discriminative features and generalises reasonably well across diverse style categories, validating the effectiveness of CNN-based approaches for fashion classification tasks.

CONCLUSION AND OUTLOOK

This study determined six women's shirt styles—professional, casual, ethnic, retro, avant-garde, and academic—through literature research and market surveys. Subsequently, women's shirt images of corresponding styles were downloaded or captured from e-commerce platforms, and a women's shirt style sample library was established. The clothing style samples were preprocessed through cropping, renaming, etc. Then, the VGG16 convolutional neural network model was chosen using transfer learning to train the images in the women's shirt style sample library. Through the training and testing of the convolutional neural network, the model achieved comprehensive performance metrics, including 86% overall accuracy and 0.86 macro-average F1-score, with individual category F1-scores ranging from 0.77 (Avant-garde) to 0.92 (Collegiate). Confusion matrix analysis revealed that misclassifications primarily occur between stylistically related categories (Ethnic-Retro: 11%, Avant-garde-Casual: 15%) rather than distant style types. These results demonstrate that the VGG16 model can be effectively applied to women's shirt style recognition tasks, with particularly strong performance on categories with distinctive visual characteristics while maintaining acceptable accuracy for more ambiguous style definitions.

This study used the VGG16 convolutional neural network model for women's shirt style recognition, and there is still room for further improvement in style types and the convolutional neural network model. Due to time constraints, this study defined the women's shirt style types as professional, casual, ethnic, retro, avant-garde, and academic. In the future, as market conditions change, the number of style categories can be increased, and the style types can be divided more finely. Additionally, this study selected a relatively small number of women's shirt samples for each style. Subsequently, the sample size can be expanded to enhance the robustness of the convolutional neural network. In future research, other deep learning models can be attempted, or the convolutional neural network model can be optimised by modifying its structure to improve the accuracy of women's shirt style recognition.

Author Contributions

Conceptualization – Fu GD, Chen HC and Wang Y; methodology – Fu GD and Wang Y; formal analysis – Fu GD and Chen HC; investigation – Fu GD, Chen HC, Liu H and Lv J; resources – Fu GD; writing-original draft preparation – Fu GD; writing-review and editing – Yuan S and Wang Y; visualization – Fu GD and Chen HC; supervision – Wang. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This study is funded by Fujian Provincial Science and Technology Plan Project for International Cooperation, China-South Africa Textile Apparel Digital Twin Collaborative Intelligent Manufac-turing System Construction and Evaluation (No. 2023I0040), Fujian Provincial Software Industry Technology Innovation Key Research and Industrialisation Projects (Third Batch), R&D and Industrialisation of AI Design Enabling Platform for Textile and Garment Industry, Fujian Provincial Science and Technology Plan Project for Innovation Funding Project(No. 2025C0027); funded by 2019 Henan University Science and Technology Development Plan (No. 192102310450); funded by Project of Science and Technology of Henan Province in China (No. 242102210172).

REFERENCES

[1] Takagi M, Simo-Serra E, Iizuka S, et al. What Makes a Style: Experimental Analysis of Fashion Prediction. 2017 IEEE International Conference on Computer Vision Workshop (ICCVW). Venice, Italy, October 22-29, 2017.

[2] Aoki R, Nakajima T, Oki T, et al. Accuracy Improvement of Fashion Style Estimation with Attention Control of a Classifier. IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, USA, January 12-14, 2018.

[3] Liu Zhengdong, Liu Yihan, Wang Shouren. Deep learning method for suit recognition. Journal of Textile Research, 2019, 40(4), pp. 164-170.

[4] Yin Guangcan. Research on clothing style recognition and recommendation based on convolutional neural networks. Master's thesis, Zhejiang Sci-Tech University, Hangzhou, 2020.

[5] Li Yang. Clothing style recognition and retrieval based on deep learning. Master's thesis, Donghua University, Shanghai, 2021.

[6] Yamaguchi K, Kiapour MH, Berg TL. Paper Doll Parsing: Retrieving Similar Styles to Parse Clothing Items. Proceedings of the 2013 IEEE International Conference on Computer Vision, NW Washington, DC United States, December 1-8, 2013.

[7] Jiang Hui, Ma Biao. Algorithm for style similarity based on clothing style. Journal of Textile Research, 2021, 42(11), pp. 129-136.

[8] Li Shuxia, Yang Juncheng. Clothing style recognition analysis based on AlexNet network. Micro-computer Applications, 2022, 38(01), pp. 48-50.

[9] Abd Alaziz HM, Elmannai H, Saleh H, Hadjouni M, Anter AM, Koura A, Kayed M. Enhancing Fashion Classification with Vision Transformer (ViT) and Developing Recommendation Fashion Systems Using DINOVA2. Electronics, 2023, 12(20), 4263.

[10] Pan S, Wang P, Yang C. Optimization of automatic classification for women's pants based on the swin transformer model. Fashion and Textiles, 2024, 11, 42.

[11] Ji GP, Zhuge M, Gao D, Fan DP, Sakaridis C, Van Gool L. Masked Vision-Language Transformer in Fashion. Machine Intelligence Research, 2023, 20(3), 421-434.

[12] Abbas W, Zhang Z, Asim M, Chen J, Ahmad S. AI-Driven Precision Clothing Classification: Revolutionising Online Fashion Retailing with Hybrid Two-Objective Learning. Information, 2024, 15(4), 196.

[13] Suvarna B, Balakrishna S. Enhanced content-based fashion recommendation system through deep ensemble classifier with transfer learning. Fashion and Textiles, 2024, 11(1), 24.

[14] Baldrati A, Morelli D, Cartella G, Cornia M, Bertini M, Cucchiara R. Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23393-23402.

[15] Cartella G, Baldrati A, Morelli D, Cornia M, Bertini M, Cucchiara R. OpenFashionCLIP: Vision-and-Language Contrastive Learning with Open-Source Fashion Data. In: Foresti GL, Fusiello A, Hancock E (eds) Image Analysis and Processing – ICIAP 2023, Lecture Notes in Computer Science, vol. 14233. Springer, Cham, 2023, pp. 245-256.

[16] Han Y, Zhang L, Chen Q, Chen Z, Li Z, Yang J, Cao Z. FashionSAP: Symbols and Attributes Prompt for Fine-Grained Fashion Vision-Language Pre-Training. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15028-15038.

[17] Sun K, Zhang P, Zhang J, Tao J, Yuan K. PFNet: Attribute-aware personalised fashion editing with explainable fashion compatibility analysis. Information Processing & Management, 2024, 61(1), 103540.

[18] Li Y, Yu S, Chen Y, et al. Explainable fashion compatibility Prediction: An Attribute-Augmented neural framework. Electronic Commerce Research and Applications, 2024, 67, 101423.

[19] Bouzidi I, Ouassif N, Mezzour G. Convolutional Neural Networks and Vision Transformers for Fashion MNIST Classification: A Literature Review. arXiv preprint arXiv:2406.03478v1, 2024.

[20] Wang Z, Li T. A fine-grained attributes recognition model for clothing based on improved the CSPDarknet and PAFPN network. Signal, Image and Video Processing, 2025, 19(1).

[21] Su X, Duan J, Ren J, Li G, Liu C. Personalised clothing recommendation fusing the 4-season colour system and users' biological characteristics. Multimedia Tools and Applications, 2024, 83(5), 12597-12625.

[22] Zhao X, Zhang Y, Zhang W, Wu XM. UniFashion: A Unified Vision-Language Model for Multimodal Fashion Retrieval and Generation. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024, pp. 1490-1507.

[23] Wang Tong. Structure and related algorithms of convolutional neural networks in deep learning. Information and Computer (Theory Edition), 2020, 32(08), pp. 41-43.

[24] Zhou Pucheng, Li Licong, Wang Yong, et al. Principles and practices of deep convolutional neural networks. Electronics Industry Press: Beijing, China, 2020; pp. 6-53.

[25] Zhang Yifan. Research on clothing image recognition and classification problems based on deep learning. Master's thesis, Donghua University, Shanghai, 2020.