Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5577393, 14 pages
https://doi.org/10.1155/2021/5577393

Research Article

Multideep Feature Fusion Algorithm for Clothing Style Recognition

Yuhua Li,[1] **Zhiqiang He**,[1] **Sunan Wang**,[2] **Zicheng Wang**,[1] **and Wanwei Huang**[1]

[1]*Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou 450001, China*
[2]*School of Electronic & Communication Engineering, Shenzhen Polytechnic, Shenzhen 518055, China*

Correspondence should be addressed to Sunan Wang; wsntemp@163.com

Received 7 January 2021; Revised 11 March 2021; Accepted 3 April 2021; Published 17 April 2021

Academic Editor: Amr Tolba

In order to improve recognition accuracy of clothing style and fully exploit the advantages of deep learning in extracting deep semantic features from global to local features of clothing images, this paper utilizes the target detection technology and deep residual network (ResNet) to extract comprehensive clothing features, which aims at focusing on clothing itself in the process of feature extraction procedure. Based on that, we propose a multideep feature fusion algorithm for clothing image style recognition. First, we use the improved target detection model to extract the global area, main part, and part areas of clothing, which constitute the image, so as to weaken the influence of the background and other interference factors. Then, the three parts were inputted, respectively, to improve ResNet for feature extraction, which has been trained beforehand. The ResNet model is improved by optimizing the convolution layer in the residual block and adjusting the order of the batch-normalized layer and the activation layer. Finally, the multicategory fusion features were obtained by combining the overall features of the clothing image from the global area, the main part, to the part areas. The experimental results show that the proposed algorithm eliminates the influence of interference factors, makes the recognition process focus on clothing itself, greatly improves the accuracy of the clothing style recognition, and is better than the traditional deep residual network-based methods.

1. Introduction

Due to the prosperity of economy, people pursue personal spiritual value on the basis of satisfying the material life. People's aesthetics standard of clothing is also gradually improving unconsciously. They are no longer satisfied with the basic functional characteristics of covering up and heating and begin to pay attention to the aesthetics and personalized decorative characteristics of clothing [1]. Nowadays, people prefer to "look different" in their clothes and want to have a unique personal style [2]. Therefore, successful clothes always have distinct style characteristics.

With the introduction of the concept of deep learning, computer vision has been greatly developed [3, 4]. The computer vision technology completes the recognition and classification of images. The computer is also used to analyze and understand the image content, simulate the thinking mode of human, and automatically extract the image features [5–7]. At present, deep learning performs well in visual recognition, speech recognition, image recognition, and other aspects. In this background, based on deep learning and style characteristics of clothing, this paper proposes to take the advantages of object detection and improved deep residual network to automatically extract image features to recognize clothing styles.

He et al. [8] combined the needs of comfort, security, and beauty with clothing fabric, sewing quality, style, size, and other aspects to obtain the design elements of student clothing. Bengio et al. [9] established a connection between the Kansei engineering theory and fashion style design elements, analyzed fashion styles, colors, fabrics, and other elements of clothing, and applied the Kansei engineering theory to fashion style evaluation. Szegedy et al. [10] firstly used the action-movement tracking technology to find the parts that could most influence the style of clothing and ranked them according to their influence weight from high to low. Secondly, they collected the vocabulary describing the style of clothing and obtained the representative factors describing

the style by using the semantic difference method, which were made up of three words. Finally, they utilized Kansei engineering and fuzzy mathematics theory to establish a clothing style model, which is used for quantitative analysis of the relationship between clothing components and style. Bengio et al. [9] applied Kansei engineering to the field of clothing research. The research designed an evaluation scale first, combining with consumers' subjective evaluation of the dress style, and finally analyzed the style characteristics represented by each style and sorted them out.

YOLO [5, 11] proposed by Redmon is an earlier end-to-end detection method. The input image is first divided into $s \times s$ grid cells, and then, the direct input is resized to the convolution neural network structure that consisted of 24 convolutions with two full connection layers. The network output as a tensor includes the dimensions of each unit and is responsible for detecting the target frame of four coordinates, a positioning confidence level, and the probability value belonging to each category. YOLO also sets a multitask loss function that is compatible with border position coordinate prediction, confidence prediction, and target category prediction meanwhile for model training. Statistics in the paper show that YOLO can be as fast as 45 fps but YOLO still has many drawbacks. The most typical defects include the poor performance of YOLO in detecting small objects and nearby features. At the same time, fixed YOLO input leads to slow detection speed and an unstable network structure requires a lot of calculation.

Through in-depth study, this paper proposes a multifeature fusion recognition algorithm for clothing style based on the improved residual network and target detection model based on YOLOv3 [12]. In order to eliminate the interference factors such as the background of the clothing image, at the same time extract the comprehensive and detailed features, the proposed method extracts multicategory areas of the global areas, main parts, and part areas from the clothing image by our model. In order to extract the features of the areas, this paper improves the residual network (ResNet). The improved ResNet is trained by multilabel images beforehand, to enhance the ability of feature extraction for multicategory areas. By combining together the features of the multicategory parts extracted through improved ResNet of different categories, this paper uses an effective multifeature fusion method to realize the recognition of clothing style.

The main contribution of this paper includes the following aspects:

(i) A multicategory feature extraction model (MFEM) is proposed. We designed a multicategory clothing area extraction strategy to extract the three category areas from an image, namely, the global areas, main parts, and part areas, meanwhile eliminating the interference factors in the process of clothing style recognition. In this process, we used the target detection technology

(ii) An improved ResNet model is proposed, by improving the order of the " batch normalization layer with the activation layer with the convolutional layer" in the traditional residual block and adjusting the structure of the network convolutional kernel

(iii) A multifeature fusion method is proposed. The features of global areas, main parts, and part areas extracted by MFEM will play different roles in retrieval due to the different image scales they focus on. Although direct fusing can improve the effect, there will also be mutual influence and weakening. The multifeature fusion technology can effectively fuse different features of multicategory areas of the input image

The rest of this paper is organized as follows. In the next section, we give a brief review of the existing clothing style recognition algorithms. The proposed method is described in Section 3. In Section 4, we report experimental results on two different datasets. Finally, we conclude this paper in Section 5.

2. Related Work

With the number of images growing in the Internet, the clothing style recognition technology has become a hot research area for scientific researchers and internet companies. Schroff et al. [13] used a simple spatial local attribute classification method combined with a naive classifier for image style learning and recognition classification. Zheng et al. [14] proposed a clothing image classification method combining the face and hairstyle. This method first segments the input image into the face, hair, and clothing area, meanwhile applying PCA and GMM to each area, and then uses some known classification results to output a single score for every area, according to the user's face and hairstyle recommend appropriate clothing for the image. Yang et al. [15] believed that a jacket could be defined by its style elements, such as the collar, the printing style, and the existence of sleeves, especially the collar. Therefore, style elements such as the collar shape are important clues to distinguish clothing types. Noh et al. [16] designed a system that could be independent of the model pose, image background, and image resolution, realizing automatic classification from the input image to the jacket. Tola et al. [17] extracted the color texture and other factors of jacket models and then analyzed the extracted features using the random forest, so as to complete the classification of clothing.

He et al. [18] used the Kansei engineering method to decompose various components of men's shirts, extracted influencing factors of styles, studied the relationship between the style and components, and achieved the style quantification of men's shirts. Ketkar [19] analyzed the modeling factors of dresses and introduced the triangle fuzzy number to fuzzy quantify the relationship between the clothing modeling factors and clothing perceptual words, so as to achieve the quantification of clothing style characteristics. Redmon and Farhadi [20] designed a fine-grained deep model and multimedia search program. First, the property vocabulary is constructed using human annotations obtained on the new fine-grained garment dataset. Then, this vocabulary is

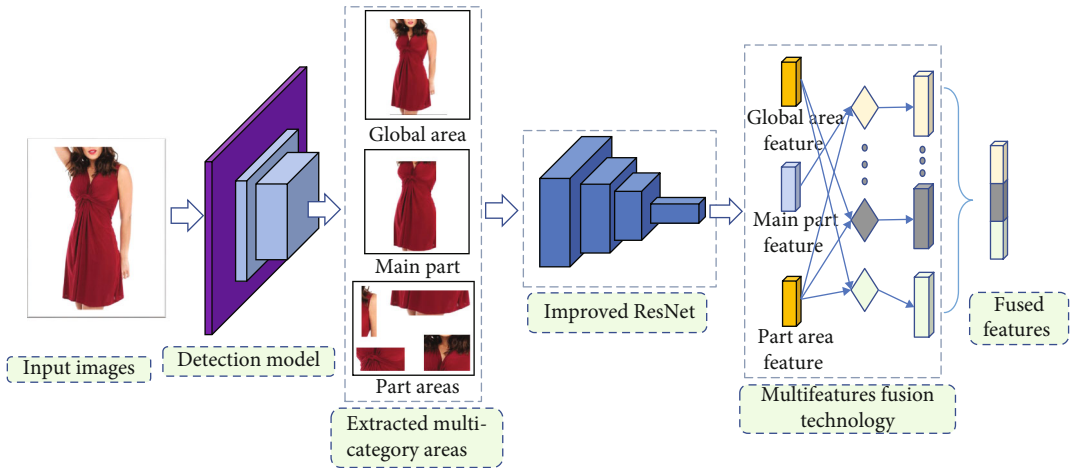


Figure 1: Multifeature extraction and fusion process. Firstly, using the improved detected method to extract the global area, main part, and part areas of the input image. Secondly, according to the results of the method, the multicategory areas is inputted to the improved residual network. And the residual network separately outputs three 128 dimensional features of global area, main part, and parts areas. Finally, the three category features are effectively fused by the multicategory feature fusion technology.

used to train a fine visual recognition system of clothing style to realize the recognition of the clothing style. Mehmood et al. [21] first found similar styles from a large database of tagged fashion images, parsed queries using these examples, and then trained the global model to implement style recognition.

However, the deep convolutional neural network is still inadequate for clothing style recognition. Khan et al. [22, 23] proposed the famous deep residual network ResNet. Compared with the traditional convolutional neural network, the deep residual network introduces a residual module into the network, which effectively alleviates the gradient disappearance of back propagation during network model training, thus solving the problems of difficult training and performance degradation in the deep network. In this paper, a kind of improved deep residual network structure and target detection model are proposed to improve the performance of recognition of clothing style.

Target detection models based on deep learning are generally divided into two categories, one is the target detection model based on candidate regions and the other is the target detection model based on the regression method [24–26].

The target detection model based on the candidate area process is divided into two steps and therefore also known as the two-phase-type (two-stage) target detection model, the first generation contains the ROI (region of interest) [27, 28]; the ROI is used to detect the target location of the candidate region and each candidate region of the generated target category of estimation and the return of border position [29]. This kind of model relies on the design of the convolutional neural network structure, but its real-time performance is poor due to the multistage characteristics.

Compared with the target detection model based on candidate regions, the target detection model based on the regression method does not need to extract the candidate box but directly completes the target border detection through convolution computation, which is called the one-stage method. In literature [30, 31], Redmon proposed the YOLOv2 model [32] and conducted BN normalization operation [33] for the input of each layer of the network. The anchor box was introduced to replace the full-connection layer, and a clustering method was used to screen the anchor box, which improved the detection accuracy of YOLO. Compared with the v2 version, YOLOv3 proposed in literature [34] has made more optimization. For example, binary cross-entropy loss [35] is used by the classification target function branch to replace the original Softmax and the underlying network with darknet-53 [36]. As a result, the detection efficiency is higher and the universality is stronger.

According to the theory of the receptive field, the deeper the convolutional layer is, the more abstract the semantics are and the local detail features of the bottom layer are blurred. Many local fine margin and shape changes become less and less obvious after multilayer convolution processing. Most of the detection models directly extract the features of the last layer of the network for analysis, which directly leads to the loss of the bottom detail features and has little impact on the accuracy of large-scale target objects, but the detection accuracy of small target objects will drop sharply [37]. Multiscale feature fusion is used to solve this problem, that is, instead of choosing the convolution output of the last layer as the feature of the image, it adopts the method of multiscale feature fusion. The above, based on deep learning and traditional feature fusion algorithms, have their own advantages in extracting the overall semantic features and specific local features of the clothing image. It is difficult to use one method alone to make the features of the clothing image more effective and comprehensive.

Therefore, we firstly propose a multicategory feature extraction model (MFEM) to extract the three category areas from an image, namely, the global areas, main parts, and part areas, meanwhile eliminating the interference factors in the

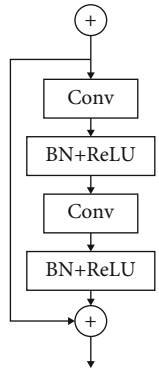


Figure 2: The sequence of the traditional residual block. The main path represents the feature diagram first through the convolution layer to BN and ReLU. The input feature diagram is not feature normalized, so the existence of the BN layer does not play a big role.

process of clothing style recognition. And then, we propose an improved ResNet model, improving the order of the batch normalization layer with the activation layer with the convolutional layer in the traditional residual block and adjusting the structure of network convolutional kernel. Finally, we designed a multifeature fusion technology to solve the problem that the single neural network cannot extract the local feature when extracting the global feature.

3. Methods

At present, the image recognition algorithm based on deep learning to extract features for the original image is widely using a single CNN [12, 17, 22]. But sometimes, the area of clothing identified is a little part of the original image and the areas that are not relevant to the identity of the clothing will have a negative impact on the recognition results. Only using a single CNN to identify the features of the clothing from the global is not comprehensive, leading to the image recognition of the clothing images not being focused on the clothing itself. In this paper, the image recognition algorithm based on improved ResNet and multifeature fusion is proposed, as shown in Figure 1. First, use the improved detected method to extract the global, main, and part areas of the image. Then, according to the results of the method, the multicategory areas are input to the improved residual network. By setting the dimension of the last layer of the residual network to 128, the residual network separately outputs three 128 dimensional features of the global, main, and parts areas of the clothing image and the three category features are effectively fused by the multicategory feature fusion technology.

3.1. Improved Residual Network. At present, most researchers choose AlexNet [38] and VGGNet to extract features for the clothing image. VGGNet has been improved on AlexNet, and the network structure is concise. The literature [39] uses VGGNet on the clothing recognition, but in the face of multiple clothing categories, the network layers of AlexNet and

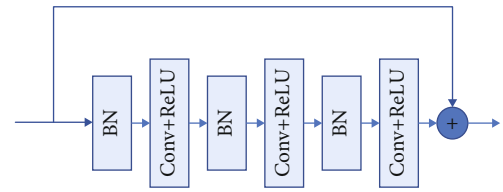


Figure 3: The sequence of the residual block in the improved residual network.

VGGNet are less, which directly affects the feature learning ability of the network.

3.1.1. Traditional Residual Network. ResNet has obtained the first place in the 2015 ImageNet Large-Scale Visual Recognition Competition [40]. The deep residual network is made up of the residual block. Each residual block can be expressed as follows:

$$y_i = h(x_i) + F(x_i, w_i), \tag{1}$$

$$x_{i+1} = f(y_i), \tag{2}$$

where the F is the residual function, f is the *ReLU* function, w_i is the power value matrix, and x_i and y_i are the input and output, respectively, of the I layer. The number h is by

$$h(x_i) = x_i. \tag{3}$$

The residual function F is defined as follows:

$$F(x_i, w_i) = w_i \cdot \sigma\left(B\left(w_i'\right) \cdot \sigma(B(x_i))\right). \tag{4}$$

$B(x_i)$ is batch normalization, "•" is convolution, and $\sigma(x) = \max(x, 0)$.

The residual units in ResNet, like the traditional CNN convolution layer, are not included in the system. Instead, the shortcut connection is introduced from the input end to the output end of each convolution layer. Using identity mapping as a shortcut connection reduces the complexity of the residual network and makes the deep network faster trained. In addition, all these shortcuts do not spread the gradient, which is the reason for the faster optimization and training of the disabled network. As the number of network layers deepens, the accuracy is not falling.

3.1.2. Improved ResNet. The weight of a certain layer of the deep convolution neural network is changed, and the output feature diagram of the layer changes, and the weight of the next layer of network needs to be studied again, and each layer of network weight will be affected. Adding activation functions to ResNet can improve the nonlinear ability of building network models. The deep residual network adopts linear modification unit ReLU [41], function $f(x) = \max(0, x)$ as activation function. The gradient of the ReLU function has been reduced by the gradient at $x0$, and the gradient dispersion is alleviated.

Table 1: Comparison of two kinds of network convolution structures.

	ResNet50	Improved ResNet
Structure	$\begin{bmatrix} C:1\times1,64 \\ C:3\times3,64 \\ C:1\times1,256 \end{bmatrix}\times3$	$\begin{bmatrix} C:1\times1,128 \\ C:3\times3,128 \\ C:1\times1,128 \end{bmatrix}\times3$
	$\begin{bmatrix} C:1\times1,128 \\ C:1\times1,128 \\ C:1\times1,512 \end{bmatrix}\times4$	$\begin{bmatrix} C:1\times1,256 \\ C:1\times1,256 \\ C:1\times1,256 \end{bmatrix}\times4$
	$\begin{bmatrix} C:1\times1,256 \\ C:1\times1,256 \\ C:1\times1,1024 \end{bmatrix}\times6$	$\begin{bmatrix} C:1\times1,512 \\ C:1\times1,512 \\ C:1\times1,512 \end{bmatrix}\times6$
	$\begin{bmatrix} C:1\times1,512 \\ C:1\times1,512 \\ C:1\times1,2048 \end{bmatrix}\times3$	$\begin{bmatrix} C:1\times1,1024 \\ C:1\times1,1024 \\ C:1\times1,1024 \end{bmatrix}\times3$

With the deepening of the convolution neural network, the speed of convergence of the network and the dispersion of the gradient are found in the course of training. This problem can be solved effectively. The specific solution is to normalize the input signal of the same layer, and the formula is as follows:

$$\widehat{x} = \frac{X - E(x)}{\sqrt{\mathrm{Var}(x) + \varepsilon}}, \tag{5}$$

where $\widehat{x}$ is the activation value of the network normalization, X is the activation value of a layer of the network, $E(x)$ is the average, $\mathrm{Var}(x)$ is the variance, and ε is the minimum. The BN algorithm formula is as follows:

$$y^{(k)} = \gamma^{(k)} x \wedge^{(k)} + \beta^{(k)}. \tag{6}$$

Each neuron x^k has a pair of γ, β. When $\gamma^{(k)} = \sqrt{\mathrm{Var}[x^{(k)}]}$, $\beta^{(k)} = E[x^{(k)}]$, the model can maintain the original learning features of a layer and can reconstruct the parameters γ, β and restore the feature distribution of the initial network learning. The BN layer is an activation method of the normalized neural network, and the algorithm of batch normalization is used to process the input signal of each layer, stabilize the distribution of the data, and set up a large learning rate in the training, so that the network converges speed and the training speed is faster. Figure 2 shows the sequence of the convolution layer with the BN layer with the ReLU layer in the traditional residual network.

The sequence of traditional residues is defective in deep convolution ResNet, such as the input of the identical blocks from two paths to the deep network. The main path represents the feature diagram first through the convolution layer to BN and ReLU. The input feature diagram is not processed first, so the existence of the BN layer does not play a big role. The method of arrangement of new residual blocks proposed in this paper is to preserve the identity of the shortcut and also maintain the learning ability of the nonlinear network path on the right, as shown in Figure 3.

Table 1 is the main structure of the convolutional layer of the original ResNet50 network and the main structure after the number of parameters has been changed. There are 3 convolution kernels, 4 convolution kernels, 4 convolution kernels, 6 convolution kernels, and 512 convolution kernels. There are 3 residual blocks containing 1024 convolution kernels and two fully connected layers. The dimensions of the model output are 8 and 10, correspond to the classification categories of the two datasets.

3.2. Extracting Multicategory Areas Based on Target Detection. In order to realize the effective extraction of the global area, main part, and parts areas, the improved target detection model is used to detect the areas. At present, in the field of target detection, there are two popular types of CNN used for feature extraction, namely, VGG and ResNet, both of which are deep network structures. ResNet is more efficient than VGG due to its efficient residual components, and the extracted image feature semantics are more abundant. Therefore, the improved ResNet is used in our model. The improved ResNet trained by simple stochastic gradient descending has fast convergence speed and the ability to use memorized information to avoid repeated computation.

First, the improved ResNet is used to extract the image features, and the region proposal network (RPN) is used to complete the recommendation of candidate boxes on the image features, and a set of candidate boxes is selected. Then, the corresponding feature area is intercepted for the candidate box, and the size of $7 \times 7 \times 512$ is inputted to the full connection layer after pooling. Finally, the classification layer and regression layer are used for target classification and border regression. Our model has been optimized in many aspects.

(1) Through our improved residual block, a total of five feature maps are generated, each of which is at a different level. Therefore, the semantic and resolution information contained vary in strength and weakness

(2) The second module is the RPN recommendation candidate box. Different from faster R-CNN, RPN of this model is a cascading structure and anchors of different scales take the feature map of the corresponding level through a selector. After the first layer RPN selects the candidate box set, the optimized non-maximum suppression (NMS) method is also used to filter the candidate box set, so as to improve the efficiency of candidate box screening

(3) For each candidate region recommended by RPN, the corresponding feature map fragment is intercepted and dimensionally reduced using the ROI Align pooling layer to form the final feature with the size of $7 \times 7 \times 512$ and the full connection layer is inputted. And the ROI Align pooling method uses bilinear

Input: $B = \{b_1, \cdots, b_2\}$, $S = \{S_1, \cdots, S_N\}$, N_t, Where B is the sequence of candidate boxes, S is the score of the candidate box, N_t is the threshold of the IOU.
Output: $D = \{d_1, \cdots, d_2\}$, $S = \{S_1, \cdots, S_k\}$, Where D is the final winning candidate box and S is the score of the output candidate box.

```
Begin:
  D ← {}
  While B! = {} do:
    m ← arg max {S}
    M ← b_m
    L ← M
    B ← B ← M
    For b_i in B
      If IOU{M, b} > N_t
        L ← L ∪ b_i
      End if
    s_i ← s_i f(IOU(M, b_i))
    End for
    M ← f_2(L)
    D ← D ∪ M
End while
```

Algorithm 1: Optimized NMS method.

Table 2: Label categories of the three level areas.

Area category	Label category
Global area	Whole body
Main part	Upper, bottom
Part areas	Collar, sleeve, skirt, trouser

interpolation to avoid precision mismatch caused by quantization

In the prediction stage, this model also makes some optimization operations in order to improve the recommendation efficiency of RPN. Firstly, an RPN module is connected after the RPN model for the refinement of the secondary border of the candidate box. In addition, an optimized NMS algorithm is introduced to suppress and screen the candidate boxes generated by the RPN in the first layer. Because there is no definite proportional relationship between the confidence of the classification result and the confidence of the rectangular box position, traditional NMS will cause many candidate boxes with different targets to be mistakenly deleted. Therefore, the NMS algorithm has been improving. For example, soft-NMS in literature [42, 43] uses a method that does not eliminate high-overlapping candidate boxes but subdivides the candidate boxes. In literature [44], soft-NMS adopts the Gaussian function weighting method to integrate high-overlapping candidate boxes and these methods have certain effects. In this paper, an optimized NMS method is obtained through integration and the pseudocode is shown as Algorithm 1.

In this approach, there are two aspects. Firstly, the soft-NMS scoring inhibition method was used and the scoring formula was shown in equation (7). Secondly, the soft-NMS weighted adjustment method is used to adjust the weight of the candidate box's optimal position coordinates according to the score. As shown in equation (8), each box whose candidate box IOU with the maximum score value is larger than the threshold value is added according to the weight of the score value to get a new box to be added to the final set of candidate boxes.

$$s_i = s_i e^{-\frac{iou(M,b_i)^2}{\sigma}}, \quad \forall b_i \notin D, \tag{7}$$

$$M'_x = \sum_i^m \frac{b_{i,\text{score}}}{\sum_j^m b_{j,\text{score}}} b_{ix}, \tag{8}$$

where b is the sequence of candidate boxes, S is the score of the candidate box, and m is the maximum value of s.

When training our model, we firstly use the annotation tool to label the three level category areas of the clothing image; the specific labeling category is shown in Table 2.

The global area, main part, and part areas are as mentioned before, where the global area is the area for removing the background and retaining the full clothing and human body. The main part is the coat or the lower part of the image, and the dress and the attachment are also part of the coat. The part area is the collar, sleeve, and other local areas. In this paper, the proposed model outputs the coordinate and category information of each area box and then extracts and generates the result according to the area box coordinates, and then, the result map is inputted to the improved residual network for feature extraction.

3.3. Multifeature Fusion. Because CNN has rich features of high-level semantic information, the fusion of features of different scales can not only retain the details of the high-level bottom but also retain the basic features of high-level semantic information. However, different fusion strategies have different effects on the test results. A more complex integration strategy will only increase the

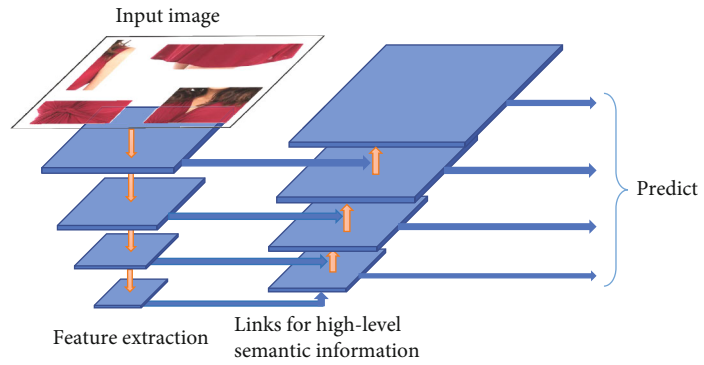


Figure 4: Structure of the feature pyramid network (FPN) model. The top-down link is used for feature extraction of the input image by the improved ResNet; a bottom-up link is used for the downward transmission of high-level semantic information.

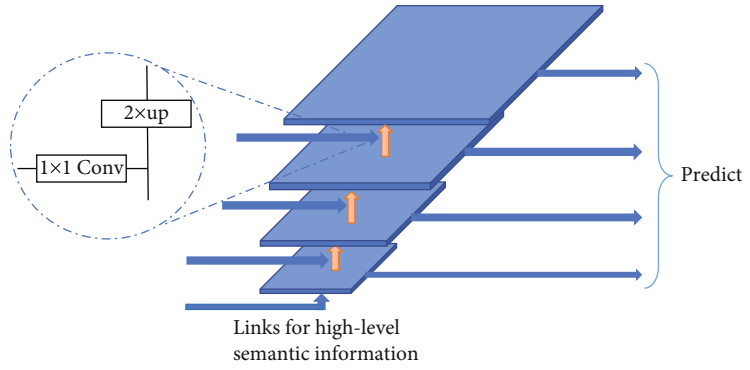


Figure 5: Internal details of the feature pyramid network (FPN) model.

Table 3: The six styles of clothing. We use the multilabel dataset to train the classification.

Attribute category	Specific label category
Sleeve of clothing	Long sleeves, short sleeves, sleeveless
Color of clothing	Pure color, color, pattern
Length of clothing	Long, ordinary, short
Type of clothing	Loose, flat, straight
Material of clothing	Cotton, hemp, cowboys, lace, mix
Collar of clothing	Circle collar, v collar, erect collar

computational complexity of the model but will have a subtle impact on the results. At present, in the target detection model based on candidate regions, the feature pyramid scheme proposed in literature [45] is a multiscale feature fusion strategy with good effect.

As shown in Figure 4, the output of each layer of the pyramid is independent and can be used as the selection of features. Such a feature formation method is also known as the feature pyramid network (FPN).

As shown in Figure 4, FPN has two links and a horizontal connection; a top-down link is used for feature extraction of the input image by the improved ResNet; a bottom-up link is used for the downward transmission of high-level semantic information; a horizontal link is used for the fusion output of features and transmitted semantic information of this layer.

As can be seen in Figure 5, there are actually three fusion links in this model. The first one is the feedforward calculation of improved ResNet, which only needs to use convolution computation to complete the feature extraction of the input image and save the features of each layer. In addition, there are two information transmission links and lateral links on both sides. As mentioned in Figure 4, the left side is the top-down information transmission link and the right lateral link. High-level semantic information is transmitted down through this link, from the third layer all the way to the first layer. The feature fusion method of the two adjacent layers is to carry out up-sampling of the upper layer features. Since the output scale of the two layers of features differs by two times in ResNet, the scale of the upper layer features can be the same as that of the lower layer features only by using deconvolution and sampling twice.

Meanwhile, the lower layer features need to be convolved through 1×1. Then, the two-layer features are added to the element to obtain the features {C1, C2, C3}. Similarly, the other two links are the right bottom-up resolution information transfer path and the left transverse connection link, from the first layer all the way to the third layer. The feature fusion method of the two adjacent layers is to pool the features of the lower layer, and the scale of the features of the upper layer can also be the same as that of the lower layer. At the same time, the upper layer features need to be convolved through 1×1. Then, the two layers of features are

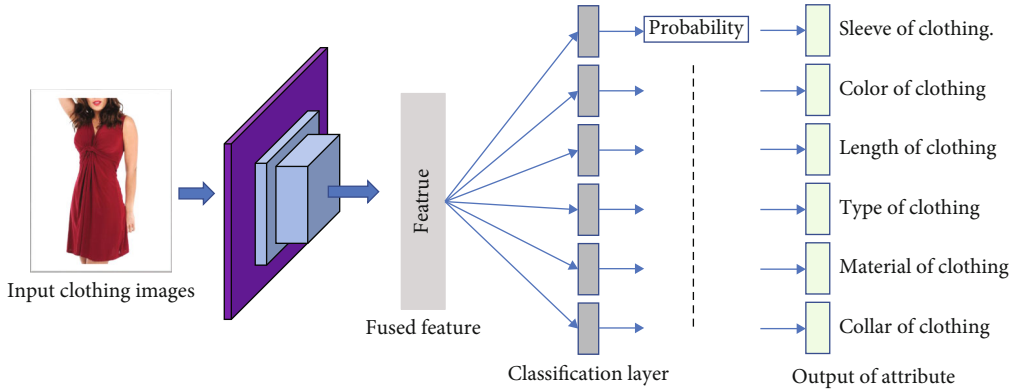


Figure 6: The clothing style classification model. Firstly, the fused feature is obtained by our model. Then, the clothing is classified by using the six Softmax classifiers in the classification layer. The number of classifiers is equal to the number of categories defined for clothing styles.

added by element to obtain features {N1, N2, N3}. Finally, the feature of the corresponding layer is added by element to obtain the final feature vector.

$$L = \left[k_0 + \log_2 \left(\frac{\sqrt{wh}}{224} \right) \right], \tag{9}$$

where 224 is the standard scale of the input of the model and the model takes it as the reference base value of the length and width of the candidate box, which represents the output of layer L that can be used. w and h are the length and width of the candidate box, where $k_0 = 4$.

Multilayer feature fusion is composed of three multiscale features, some of which are biased towards high-level semantic information and some towards low-level resolution information. For the targets with different scales, using the characteristics of different scales is more beneficial to the final result. For example, small-scale targets need rich resolution information, so they can be followed up with features that are biased towards the bottom layer. Large-scale targets are more concerned with the richness of semantic information, so they naturally tend to follow up the calculation with higher-level features.

The output is represented as L_i, 128 dimensional vectors, as the feature of the image. The features of the input multicategory area extraction are represented as $L(\text{global}), L(\text{main}), L(\text{parts})$. The fusion of the output of the system is a weighted set of 128 dimensional vectors, as shown in Figure 1. The output of the encoder contains the multicategory features of the input image, and the three multicategory areas are required to merge into the decoder. The current moment of the input image can be expressed as follows:

$$G = \sum_{i=1}^{n} \alpha_i^{(t)} L_i, \tag{10}$$

where $\alpha_i^{(t)}$ is the poutput weight of t times, $\sum_{i=0}^{n} \alpha_i^{(t)} = 1$, and $\alpha_i^{(t)}$ changes in the change of the t and dynamically adjusting the weights of different locations. And $\alpha_i^{(t)}$ is related to the visual weight of the input of the t moment and the information before the t. $\alpha_i^{(t)}$ update mechanism can be expressed as follows:

$$\begin{aligned} \beta_i^{(t)} &= w^T \varphi\left(W_h h_{t-1} + W_f f_i + b \right), \\ \alpha_i^{(t)} &= \frac{\beta_i^{(t)}}{\sum_{j=1}^{n+1} \beta_j^{(t)}}. \end{aligned} \tag{11}$$

f_i is a subset vector for I, $f_i \in \{G, L_1, L_2, \cdots, L_n\}$ and $\beta_i^{(t)}$ indicates that the corresponding visual vector f_i is weighted under the weight relative to the corresponding score weight that has been produced before. h_{t-1} is the output of a hidden layer; w, W_h, W_f and b are the weighted variables that need to be learned; $\varphi(\cdot)$ is the activation function.

3.4. Clothing Style Recognition. CNN is usually used for single label classification, and the image is the most difficult image category, with a large number of clothing features, except for the rich visual information. In the classification problem of clothing style properties, each image is represented by multiple labels, so single label learning does not apply. In this paper, the paper uses multilabel learning to conduct the classification training of clothing style properties for the improved ResNet and classifies each type of attribute and gets the model of the image classification of the clothing after training. Effectively, to solve the problems of the correlation between the different types of properties and the ability to directly put these properties in the same class set, this article defines multiple general clothing style attributes and several specific category tags, as shown in Table 3.

Based on the above definition, this paper designs the model of the clothing attribute multilabel classification, as shown in Figure 6:

The input image is first obtained by the multicategory deep features by improved detection and the improved ResNet, and then, the properties of several Softmax classifiers in the clothing style classification layer are calculated, and the number of classifiers is equal to the number of category properties of the dress style. The number of neurons in each

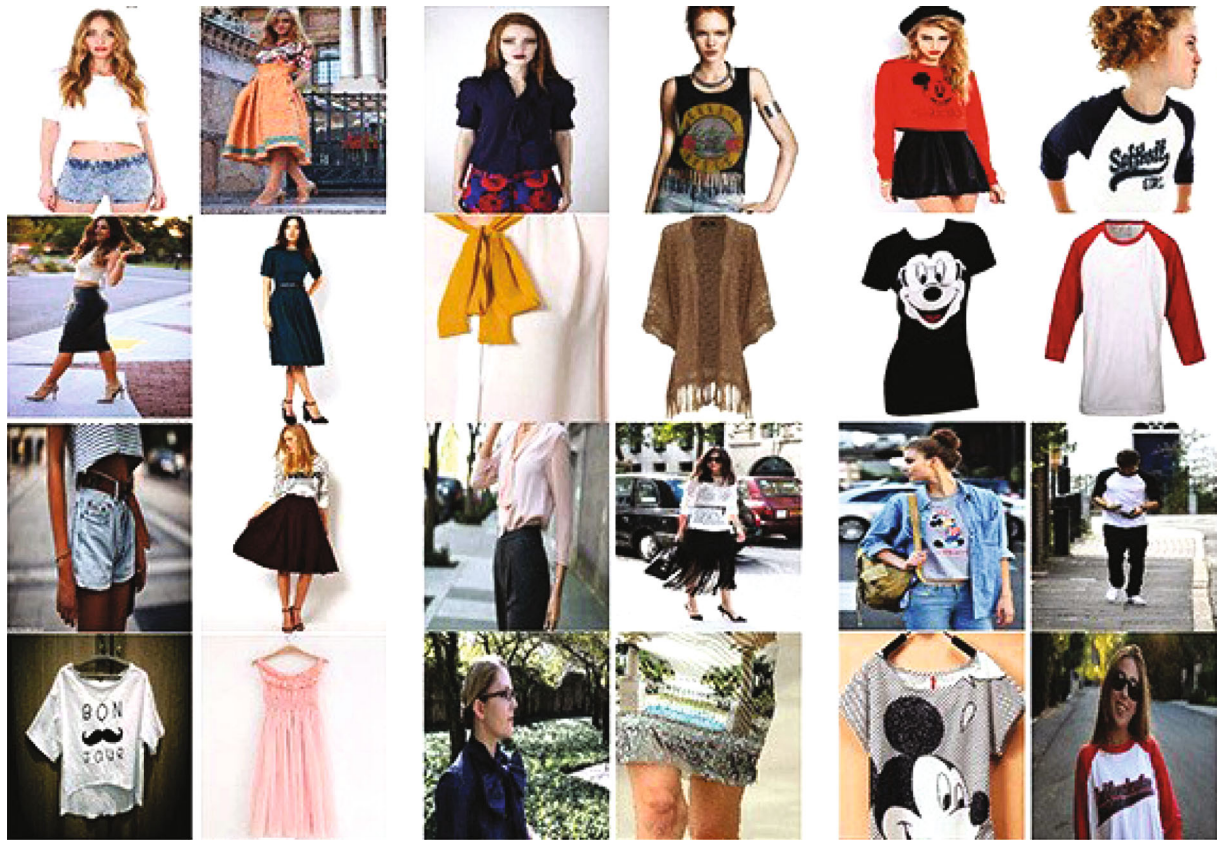


Figure 7: Sample images from the DeepFashion dataset.

classifier is equal to the number of specific labels for the clothing styles corresponding to the classifier.

3.5. Training of Model. Firstly, we randomly selected 100 styles of clothing images from the training dataset, each with a unique ID. Secondly, 3 images (a triplet) are randomly selected for each style of clothing, a total of 300 images. We only choose 3 pictures because the number of images of some clothing styles is relatively small. Then, the model with random initialization parameters is used to extract the features of each image. The retained information of each image has three categories: path, ID, and feature vector. Finally, twice loops are used for each image under each ID to select the matching positive and negative samples from the remaining 299 images according to equation (1) for training our model.

A triple consists of x_i^a (anchor), x_i^p (positive), and x_i^n (negative). x_i^a and x_i^p are the same style, while x_i^a and x_i^n are different styles. In triples, the Euclidian distance between the x_i^a and x_i^p plus the threshold should be greater than the Euclidian distance between the x_i^a and x_i^n. We use the selected triples to train the proposed improved ResNet model and then reselect the triples with the new parametric model.

$$\left\|\mathrm{Net}(x_i^a) - \mathrm{Net}\left(x_i^p\right)\right\|_2^2 + thre > \left\|\mathrm{Net}(x_i^a) - \mathrm{Net}(x_i^n)\right\|_2^2, \quad (12)$$

where i represents the i-th triple. There is the threshold value. By trying different thresholds, the triplet similarity measurement is learned. It is found that the best effect is when the global area, main part, and part area branches are set at 0.2, 0.18, and 0.15, respectively. Net $(\bullet)$ represents the feature vector extracted from the proposed model.

When using triples for training, the feature vectors Net(x_i^a), Net(x_i^p), and Net(x_i^n)) of the three samples are inputted into the triplet loss function. If it is not equal to equation (12), the parameters of the model will not be changed; otherwise, it will be calculated according to equation (13) of the loss function:

$$L = \left\|\mathrm{Net}(x_i^a) - \mathrm{Net}\left(x_i^p\right)\right\|_2^2 + thre - \left\|\mathrm{Net}(x_i^a) - \mathrm{Net}(x_i^n)\right\|_2^2. \quad (13)$$

Obtain the loss L of the model, and then, adjust the parameters of the model. The proposed model trained on the triplet similarity measure can reduce the feature distance of the same clothing image, increase the feature distance of different clothing images, and further improve the recognition ability.

4. Experiments

In this section, we will demonstrate the benefits of our approach. We start with an introduction to the dataset and then present our experimental results with performance comparison to several state-of-the-arts on the public

Short sleeves	Long sleeves	Sleeveless
Pure color	Pure color	Pure color
Ordinary	Ordinary	Long
Flat	Loose	Straight
Cotton	Hemp	Lace
Circle collar	V collar	Circle collar

Long sleeves	Sleeveless	Short sleeves
Pure color	Pattern	Pure color
Ordinary	Long	Ordinary
Loose	Straight	Flat
Mix	Cotton	Cotton
Erect collar	Circle collar	Circle collar

Figure 8: Image classification results on datasets. The black font represents the result of correct recognition, and the bold font represents the result of incorrect recognition. The first row indicates input images. The next table represents the predicted results of every category. The first row to the last row in the table represent the sleeve, color, length, type, material, and collar of the input clothing. It can be seen from the experimental results that our model has a good recognition effect for different types of clothes. For example, the second images are predicted wrongly because the hair covers the collar and the model incorrectly recognizes it as a V collar.

datasets, DeepFashion and FashionMNIST datasets. Finally, the scalability and effectiveness of our method are verified on the datasets. And the experimental device is a GTX 2080 GPU and 32 GB of RAM.

4.1. Datasets

4.1.1. DeepFashion. DeepFashion is a large-scale dataset opened by The Chinese University of Hong Kong. It contains 800000 pictures, including different angles, different scenes, and buyer shows. There are a total of four main tasks, namely, clothing category and attribute prediction, in-shop and C2S clothing retrieval, key points, and external rectangular box detection. Each image also has a wealth of annotation information, including categories, attributes, feature points, and other information. Figure 7 shows some sample images from the DeepFashion dataset.

4.1.2. FashionMNIST. The FashionMNIST dataset contains 10 categories of images, namely, T-shirt, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, and ankle boot. The training data set contains 6000 samples for each category, and the test data set contains 1000 samples for each category. There are altogether 10 categories.

4.2. Evaluation Metrics. In this article, we use mean average precision (mAP) as the measurement standard of the algorithm. mAP is the average on the basis of AP. The formula for calculating mAP is shown in equation (14).

$$\text{mAP} = \frac{1}{|Q_R|}\sum_{q\in Q_R} AP(q) \tag{14}$$

In equation (14), q means a query, which is the image to be retrieved, Q_R means the entire image collection, and $AP(q)$ means the average accuracy rate. In simple terms, AP is to calculate the average accuracy of a query image and mAP is to take the average of the accuracy of all query images. The ordering of target images in the search results is also within the consideration of mAP.

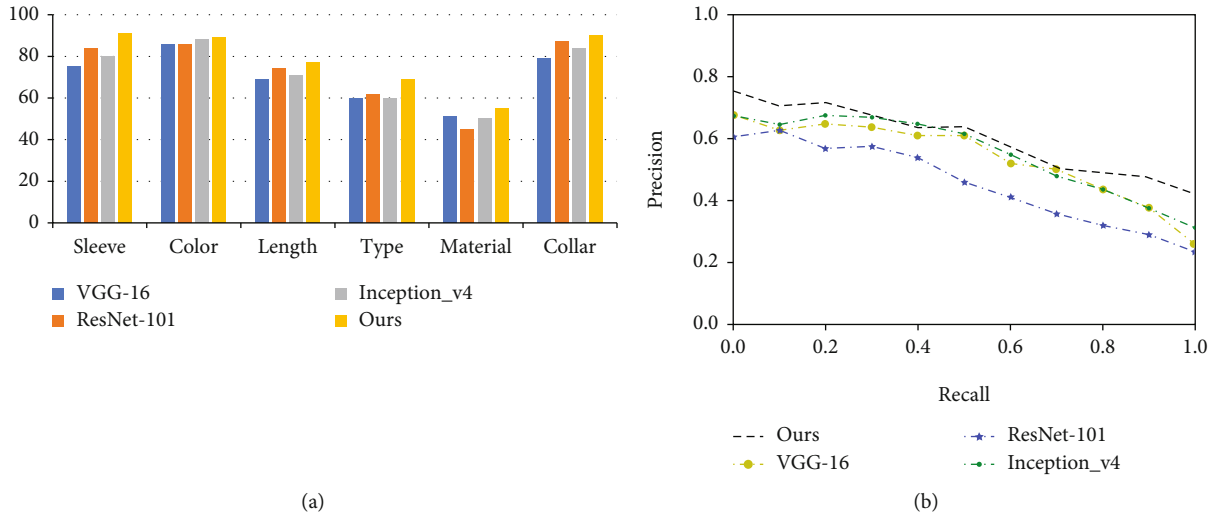
Although mAP is a statistical evaluation of the proportion of correct search results, there is a lack of evaluation of the location information of the search results. This paper uses the PR curve as the evaluation of the location information of the retrieval results. P in the PR curve represents precision, and R represents recall (recall rate).

$$\begin{aligned}\text{precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}},\\ \text{recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}}.\end{aligned} \tag{15}$$

Among them, the positive examples are correctly classified as positive examples, denoted as TP (true positive), and the positive examples are incorrectly classified as negative examples, denoted as FN (false negative). Negative cases are correctly classified as negative examples, denoted as TN (true negative), and negative examples are incorrectly classified as positive examples, denoted as FP (false positive).

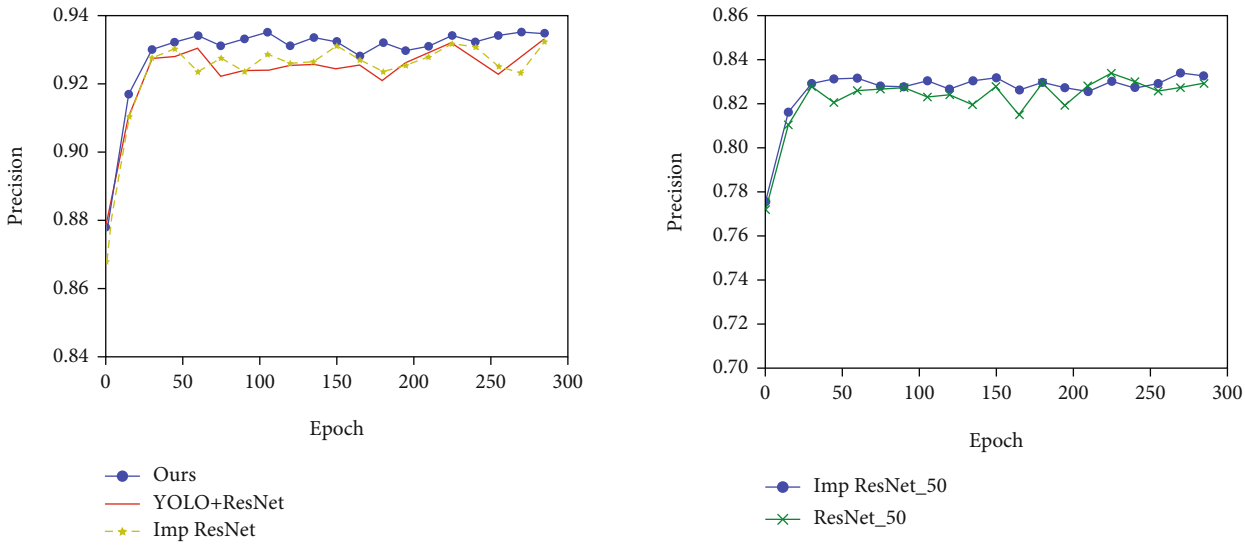
4.3. Experiment of the Proposed Method for Clothing Style Recognition. To demonstrate the scalability and effectiveness of our approach, we tested it on large-scale DeepFashion and FashionMNIST datasets. Both of these datasets are composed of a large number of clothing images, which include people with noiseless background or not. At the same time, the people have different postures. This experiment mainly reflects the classification effect of the clothing jacket. We set the number of neurons in the classification layer as 17 and h in the latent layer as 156. Then, we fine tune our network with the entire dataset. After 10000 training iterations, our proposed method achieved very high accuracy in 17 categories of clothing classification tasks (obtained by the last layer).

As shown in Figure 8, although the background of the clothing image is background free or noisy, the method presented in this paper shows good classification performance with or without people. The recognition effect of six clothes is shown in Figure 8. These 6 pieces of clothing include short sleeves, shirts, dresses, and cardigans, which, respectively, represent different genders and styles of clothing. The table under each photo shows six attributes of the clothing, including the sleeve, color, length, type, material, and collar. The



(a) (b)

Figure 9: (a) is the precision of the classification of the clothing image style and (b) represents the PR curve. As shown in (a), although the different network models are given different rates, however, the classification accuracy of the sleeves, collars, and patterns of the six types of attributes is higher and the classification accuracy of the three kinds of clothing is low. This is due to the easy distinction between the sleeve length, the collar, and the three kinds of species and the length, the plate, and the clothing are more difficult to distinguish. It can also be seen in (b) that our model has better performance than the other three models.



(a) The final model (ours) is compared with two other improved models (YOLO +ResNet: YOLOv3+ResNet_101; Imp ResNet: improved ResNet_101 without YOLOv3)

(b) The improved ResNet_50 compared with ResNet_50

Figure 10: The final model (ours) is compared with other models. As shown in (a), it can be found that no matter whether adding a YOLO model to the traditional residual network or just using improved ResNet, the best experimental results have been obtained with our proposed models. As shown in (b), the improved residual network of this paper is compared with the traditional residual network and the mAP of our model is better than that of the traditional residual network. When the epoch was around 140 and 200, MAP dropped sharply, forming two valleys and peaks.

black font represents the result of correct recognition, and the red font represents the result of incorrect recognition. For example, the second images are predicted wrongly because the hair covers the collar and the model incorrectly recognizes it as a V collar. Please note that some of the images are predicted wrongly because products can be ambiguous between certain categories. For example, as shown in Figure 8, the looseness of white short sleeves is difficult to distinguish.

4.4. Comparison of the Proposed Method with Other Methods on DeepFashion. In the course of the classification of clothing styles, 28500 images were selected from the training center, with 23000 images as a training set, 5500 images as a test

Table 4: The results of different networks.

Network model	Accuracy rate (%)	Training time (h)
VGG16	89.76	44
Inception_v3	91.06	39
ResNet_101	93.26	81
Ours	94.97	83

set, in order to find the suitable network model for the classification of clothing images, selecting the model of VGG-16, ResNet-101, Inception-v4, and ours for the comparison of the classification of clothing styles. Using the training parameters on the ImageNet to initialize each network, the classification layer parameters are randomly initialized by Gaussian distribution and then training the network using the training set.

Finally, Figure 9 shows the precision of the test set in four different networks. As Figure 9 reveals, although the different network models are given different rates, however, the classification accuracy of the sleeves, collars, and patterns of the six types of attributes is higher and the classification accuracy of the three kinds of clothing is low. This is due to the easy distinction between the sleeve length, the collar, and the three kinds of species and the length, the plate, and the clothing are more difficult to distinguish. For the length of the dress, the classification accuracy of the dress is low because of the influence of the fashion style and the height of the model. For the type, due to the angle of shooting and the position of the model, the classification of the type is not high. And clothing is harder to distinguish. The results are common logic, and in the four network models, the overall performance of the network model is the best.

4.5. Comparison of the Proposed Method with Other Methods on FashionMNIST. Standard dataset FashionMNIST has 70000 images from 10 different categories of goods. There are 60000 images as the training set and 10000 images as test set validation. The image size of the dataset is consistent with the MNIST dataset. As shown in Figure 10(a), the recognition algorithm presented in this paper is better than the other two improved networks and the ability to be more powerful in mAP and convergence. It can be found in Figure 10(a) that no matter whether it is adding a YOLO model to the traditional residual network or just using improved ResNet, the best experimental results cannot be achieved. Although our model is slightly worse than the other two models with an epoch of 170, our model is better than the other two models overall. However, the overall performance of YOLO+RseNet and Imp ResNet is basically the same. As shown in Figure 10(b), the improved residual network of this paper is compared with the traditional residual network and the mAP of our model is better than that of the traditional residual network. When the epoch was around 140 and 200, MAP dropped sharply, forming two valleys and peaks. Therefore, in conclusion, it can be found that our model has better and more stable performance in these two groups of comparative experiments. Therefore, in conclusion, our model has better and more stable performance.

As shown in Table 4, the network models are fully trained to identify accuracy and training time. The use of CNN has not resulted in the difficulty of training in the network, the inception_v3 layer is more than VGG16, but inception_v3 is more accurate and effective for image recognition classification, and the accuracy of the improved depth of the network is more accurate than VGG16. In the case of our improved ResNet, the precision is 1.32% better than the traditional residual network precision. The network precision is combined with the module, which improves the accuracy of the network by 2.21%. After the two method junctions, the final model is 0.95% more than the traditional residual network precision. The results show that the proposed model can improve the characteristic learning ability of the convolution neural network. Table 4 shows that using our method to solve the problem of clothing image recognition is very effective. The different experimental results of our method compared with other methods prove that our method of using improve ResNet has significant advantage on image recognition.

5. Conclusions

In this paper, we presented a new method for clothing style recognition, which is based on the target detection and multideep feature fusion. It first introduces and implements the improved target detection model to extract multicategory areas and the improved ResNet to extract deep features. Lastly, by feature pyramid network, the shape, soft-NMS, and multideep features fusion technology, the three multideep features are greatly fused together. In the end, an accurate and fast clothing style recognition of clothing style was achieved. By comparing the experimental results and the evaluation of recognition performance, it can be seen that the proposed algorithm has not only good efficiency but also excellent robustness in the clothing style recognition. Since the method in this paper needs to recognize every detail of clothing, the recognition rate of the proposed method will be greatly reduced if the clothing image is severely occluded. During the experiment, we found that the shirt would cover the pants in most cases, so the recognition rate was not high; so, our method did not support the multicategory recognition of pants for the time being. In the future work, we will further study the solution to this problem.

Data Availability

The DeepFashion data used to support the findings of this study have been deposited in the Google Drive or Baidu Drive repository (http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html). The FashionMNIST data used to support the findings of this study have been deposited in the Github repository (https://github.com/zalandoresearch/fashion-mnist).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is jointly supported by the National Natural Science Foundation of China (62072414 and U1504608), and the Key Scientific and Technological Project of Henan Province (212102210540, 192102210294, and 202102210383), and the Key Scientific Research Projects of Henan Higher School (20B520039).

References

[1] Y. Wang and S. Zhi-Feng, "Clothing image classification and retrieval based on metric learning," *Computer Applications and Software*, vol. 34, pp. 255–259, 2017.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, , San Diego, CA, USA, 2015.

[3] F. Afza, M. A. Khan, M. Sharif et al., "A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection," *Image and Vision Computing*, vol. 106, 2021.

[4] H. Arshad, M. A. Khan, M. I. Sharif et al., "A multilevel paradigm for deep convolutional neural network features selection with an application to human gait recognition," *Expert Systems*, vol. 27, 2020.

[5] M. A. Khan, Y. D. Zhang, S. A. Khan, M. Attique, A. Rehman, and S. Seo, "A resource conscious human action recognition framework using 26-layered deep convolutional neural network," *Multimedia Tools and Applications*, vol. 80, 2020.

[6] N. Naheed, M. Shaheen, S. A. Khan, M. Alawairdhi, and M. A. Khan, "Importance of features selection, attributes selection, challenges and future directions for medical imaging data :a review," *Computer Modeling in Engineering & Sciences*, vol. 125, no. 1, pp. 315–344, 2020.

[7] M. Rashid, M. A. Khan, M. Alhaisoni et al., "A Sustainable Deep Learning Framework for Object Recognition Using Multi-Layers Deep Features Fusion and Selection," *Sustainability*, vol. 12, no. 12, p. 5037, 2020.

[8] Z. He, Y. Li, L. Deng, P. Li, X. Shi, and X. Han, "A new two-stage image retrieval algorithm with convolutional neural network," in *Proceedings of the 2019 8th International Conference on Networks, Communication and Computing*, pp. 98–102, Luoyang, Henan, China, 2019.

[9] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, Las Vegas, NV, USA, 2016.

[11] D. Yarotsky, "Error bounds for approximations with deep ReLu networks," *Neural Networks*, vol. 94, pp. 103–114, 2017.

[12] D. G. Lowe, "Distinctive image features from scale-invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: a unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, Boston, MA, USA, 2015.

[14] Y. Zheng, S. Wu, D. Liu, R. Wei, S. Li, and Z. Tu, "Sleeper defect detection based on improved YOLO V3 algorithm," in *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 955–960, Kristiansand, Norway, 2020.

[15] X. Yang and L. J. Latecki, "Affinity learning on a tensor product graph with applications to shape and image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*, pp. 2369–2376, Colorado Springs, CO, USA, 2011.

[16] H. Noh, A. Araujo, and J. Sim, "Large-scale image retrieval with attentive deep local features," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3476–3485, Venice, Italy, 2017.

[17] E. Tola, V. Lepetit, and P. Fua, "Daisy: an efficient dense descriptor applied to wide-baseline stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 815–830, 2010.

[18] Y.-F. He, L. Zhou, J.-Q. Yu, T. Xu, and T. Guan, "Image retrieval based on locally features aggregating," *Chinese Journal of Computers*, vol. 34, no. 11, pp. 2224–2233, 2011.

[19] N. Ketkar, "Convolutional neural networks," in *Deep Learning with Python*, pp. 63–78, Springer, 2017.

[20] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," 2018, http://arxiv.org/abs/1804.02767.

[21] A. Mehmood, M. A. Khan, M. Sharif et al., "Prosperous human gait recognition: an end-to-end system based on pre-trained CNN features selection," *Multimedia Tools and Applications*, vol. 80, 2020.

[22] N. Hussain, M. A. Khan, M. Sharif et al., "A deep neural network and classical features based scheme for objects recognition: an application for machine inspection," *Multimedia Tools and Applications*, vol. 80, 2020.

[23] M. A. Khan, K. Javed, and T. Saba, "Human action recognition using fusion of multiview and deep features: an application to video surveillance," *Multimedia Tools and Applications*, vol. 80, 2020.

[24] K. Yamaguchi, M. H. Kiapour, and T. L. Berg, "Paper doll parsing: retrieving similar styles to parse clothing items," in *2013 IEEE International Conference on Computer Vision*, pp. 3519–3526, Sydney, NSW, Australia, 2013.

[25] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception - v4, inceptionresnet and the impact of residual connections on learning," 2016, http://arxiv.org/abs/1602.07261.

[26] X. Wang, T. Zhang, D. R. Tretter, and Q. Lin, "Personal clothing retrieval on photo collections by color and attributes," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 2035–2045, 2013.

[27] A. Sarwar, Z. Mehmood, T. Saba, K. A. Qazi, A. Adnan, and H. Jamal, "A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a support vector machine," *Journal of Information Science*, vol. 45, no. 1, pp. 117–135, 2019.

[28] H. J. Kim, E. Dunn, and J.-M. Frahm, "Predicting good features for image geo-localization using per-bundle VLAD," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1170–1178, Santiago, Chile, 2015.

[29] Y. Li, H. Lei, S. Lin, and G. Luo, "A new sketch-based 3D model retrieval method by using composite features," *Multimedia Tools and Applications*, vol. 77, no. 2, pp. 2921–2944, 2018.

[30] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 36–45, Boston, MA, USA, 2015.

[31] X. Han, Y. Li, Q. Zheng et al., "A Multiple Feature Fusion Based Image Retrieval Algorithm," in *Proceedings of 2019 the 8th International Conference on Networks, Communication and Computing*, pp. 104–109, Luoyang, Henan, China, 2019.

[32] L. Wei, S. Zhang, and H. Yao, "GLAD: global-local-alignment descriptor for pedestrian retrieval," in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 420–428, Buenos Aires, Argentina, 2017.

[33] D. Mishkin, F. Radenovic, and J. Matas, "Repeatability is not enough: learning affine regions via discriminability," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 284–300, Munich, Germany, 2018.

[34] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Minneapolis, MN, USA, 2007.

[35] S. Gammeter, "I know what you did last summer: object-level auto-annotation of holiday snaps," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 614–621, Kyoto, Japan, 2009.

[36] S. S. Husain and M. Bober, "REMAP: multi-layer entropy-guided pooling of dense CNN features for image retrieval," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 5201–5213, 2019.

[37] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, 2019.

[38] F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Revisiting Oxford and Paris: large-scale image retrieval benchmarking," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5706–5715, Salt Lake City, UT, USA, 2018.

[39] B. Cao, J. Zhao, P. Yang, P. Yang, X. Liu, and Y. Zhang, "3-D deployment optimization for heterogeneous wireless directional sensor networks on smart city," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1798–1808, 2019.

[40] I. Jung, K. You, H. Noh et al., "Real-time object tracking via meta-learning: Efficient model adaptation and one-shot channel pruning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11205–11212, Venice, Italy, 2020.

[41] B. Cao, J. Zhao, P. Yang et al., "Multiobjective 3-D topology optimization of next-generation wireless data center network," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 3597–3605, 2020.

[42] Y. Hou, H. Zhang, and S. Zhou, "Evaluation of object proposals and ConvNet features for landmark-based visual place recognition," *Journal of Intelligent and Robotic Systems*, vol. 92, no. 3-4, pp. 505–520, 2018.

[43] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 4, pp. 754–766, 2010.

[44] H. Lu, L.-P. Nolte, and M. Reyes, "Interest points localization for brain image using landmark-annotated atlas," *International Journal of Imaging Systems & Technology*, vol. 22, no. 2, pp. 145–152, 2012.

[45] A. Mikulik, M. Perdoch, O. Chum, and J. Matas, "Learning vocabularies over a fine quantization," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 163–175, 2013.