DOI: 10.19884/j.1672-5220.202303004

Fabric Image Retrieval Based on Fine-Grained Features

LUO Xin*, XIA Dongmei, TAO Ran, SHI Youqun
College of Computer Science and Technology, Donghua University, Shanghai 201620, China

Abstract: Fabric image retrieval is crucial for textile mills to manage their inventory and samples, but it is challenging due to the diverse appearance and fine-grained texture of fabrics. This paper proposes an algorithm based on fine-grained features to deal with this issue. The algorithm uses the coordinate attention (CA) module to extract precise location information of the fabric images and scales the overall network structure of MobileNetV3 to reduce the training time and model parameters. The optimized model is selected based on the scaling factor method, and fabric retrieval experiments are conducted on the fabric image dataset (FID). The results show that the algorithm effectively improves the accuracy of fabric image feature extraction, with a retrieval accuracy (Acc) of 91.82% and floating point operations (FLOPs) of 175.34 MB. The Acc is improved by 13.49 percentage points compared with that of the original MobileNetV3 model, while the training time is reduced, and the inference speed is improved by 25.14%. The algorithm has practical application value.

Key words: fabric image retrieval; MobileNetV3; fine-grained feature; attention mechanism; scaling factor
CLC number: TP3-05 **Document code**: A
Article ID: 1672-5220(2024)02-0115-15

Open Science Identity (OSID)

0 Introduction

Image retrieval technology has been extensively used in the textile industry for applications such as apparel search, online shopping and garment styling. In recent years, many industries have gradually shifted from mass production to multi-variety customized production. The management and reuse of existing products is the main way to shorten the production cycle. When a factory receives a sample to be reproduced, information about the same or similar existing products is retrieved from the database, thus facilitating production.

The traditional method to obtain product parameters was the sample analysis method, i.e., manually searching for identical or similar fabrics in the warehouse. It is a very time-consuming process, and the manual comparison search is highly subjective and inefficient. Subsequently, text-based image retrieval (TBIR)[1] was proposed. The TBIR method is an image retrieval technique most often used by modern textile mills. This method annotates fabric images by manual text annotation based on fabric categories, colors, materials, area densities and other parameter informations, establishes a fabric image annotation database, and retrieves similar fabric products using text keyword combinations. The method improves the efficiency of fabric retrieval to a certain extent, but it relies on the text keywords manually marked on the images, which is highly subjective and difficult to meet the requirements of fabric manufacturers for retrieval accuracy (Acc) and efficiency. With the development of image processing techniques, content-based image retrieval (CBIR)[2] was proposed. The CBRI method uses the visual features of an image to represent the image and determines whether the retrieved image is the best match by calculating the similarity between the query image and the image features in the image database. After decades of research and exploration, various carefully designed hand features have been widely used in fabric image representation. For example, Suciati et al.[3] used the fractal dimension and the gap degree to characterize geometry and spatial distribution of the fabric texture, and hue-saturation-value (HSV) color quantization to extract the primary colors to comprehensively characterize the fabric image information. Jing et al.[4] proposed regional color moments and generalized search tree (GIST) features to characterize printed fabrics and fused them using weight assignment for print fabric retrieval. Li et al.[5] extracted the aspect ratio, stenosis factor, rectangularity and curvature variance of images as shape features, and used a grayscale co-occurrence matrix to extract contrast, entropy, correlation and homogeneity as texture features to jointly characterize lace fabric images. The CBIR method achieves a higher Acc than the TBIR method and also saves a lot of annotation time, thus significantly reducing the workload and improving the efficiency of retrieval.

In recent years, the end-to-end framework based on deep learning has made significant breakthroughs in the

Received date: 2023-03-09
Foundation item: National Key Research and Development Program of China (No. 2020YFB1707700)
* Correspondence should be addressed to LUO Xin, email: xluo@dhu.edu.cn
Citation: LUO X, XIA D M, TAO R, et al. Fabric image retrieval based on fine-grained features [J]. *Journal of Donghua University (English Edition)*, 2024, 41(2): 115-129.

image analysis, and the method based on the convolution neural network (CNN) has been proven to be the most effective.

The CNN GoogLeNet[6] and ResNet[7] were widely used in fabric image classification and fabric image retrieval. Shen et al.[8] established two branches of the CNN to extract features from RGB color model images and local binary pattern (LBP) texture images, respectively, and fused the two features using a nonlinear depth feature fusion structure. Zhou et al.[9] proposed a branch network that aggregated multi-scale and attention features to classify clothing images. Zha et al.[10] proposed a framework of image retrieval with text manipulation by local feature modification which could focus on the related image regions and attributes and perform modification. Li et al.[11] proposed a method of fine-tuning feature extraction network based on masked learning, the encoder of the masked autoencoders used the ViT to extract global features and performed self-supervised fine-tuning by reconstructing masked area pixels.

Nevertheless, the most challenging problem is still to associate pixel-level features with high-level semantic features of human perception, which is called the "semantic gap". Moreover, using the labeling method in the network model, we roughly define the images with completely similar labeling as positive samples, and all other images as negative samples, which is unfair to those with partially similar labeling. In addition, these methods ignore the relationship between labels and lose some information during model training.

The goal of fabric image retrieval is to find fabrics with similar textures, colors and materials. As a special case of general image retrieval, fabric image retrieval has become a research hotspot due to its potential value in many fields such as textile design, e-commerce and inventory management. However, due to the particularity of the fabric itself, the common retrieval methods are difficult to apply directly to fabric retrieval. The main reason is that the fabric images do not contain three-dimensional (3D) shape features. For example, the natural images contain obvious contour and shape features, while the main features in the fabric images are colors and textures[12]. They are global features, and there are huge differences in periodicity and texture complexity. Due to the fundamental differences, the retrieval methods applicable to general images may not perform well on fabric images.

To address the above problems, a fabric image retrieval model based on fine-grained features is proposed to effectively distinguish local detail features in fabric images. The MobileNetV3 network model is pre-trained using the large dataset ImageNet[13]. The squeeze and excitation (SE) attention mechanism[14] in the original MobileNetV3 network model is improved, and the network scaling coefficient method from EfficientNet[15] is referenced to obtain the most suitable network structure for extracting fine-grained features in fabric images.

1 Related Work

1.1 Fine-grained features of fabric images

The appearance of fabrics not only consists of initial features such as stripes, checks and patterns but also possesses detailed features consisting of fabric tissues, densities and weave structures, resulting in complex fabric characteristics. Figure 1 shows three different fabrics with similar composition. Fabrics in Figs. 1(a) – 1(c) all belong to plain fabrics, but the fabric weaving methods are different. For the fabric in Fig. 1(a), fine particles are evenly distributed on the surface of the fabric, and the concave-convex is not obvious. The fabric in Fig. 1(a) belongs to the category of plain crepe. The surface of fabric in Fig. 1(b) has fine and uniform wrinkles, mainly woven horizontally, and vertically woven at intervals. The fabric in Fig. 1(b) belongs to the category of plain change. The surface of the fabric in Fig. 1(c) is uniform and fine, and mainly horizontal weave. The fabric in Fig. 1(c) belongs to the category of plain flat. Though the fabrics in Fig. 1 are divided into plain crepe, plain change and plain flat according to industrial needs, they are very similar visually. Therefore, the image retrieval of the fabric, like the image retrieval of wallpaper and wood, requires a high precision. Due to the characteristics of the fabric, the manually marked features cannot represent the various appearances of the fabric. Although advanced features can represent different fabric components based on different classification methods, they need a large number of training samples in different categories. Fabric image retrieval is still a very challenging task.

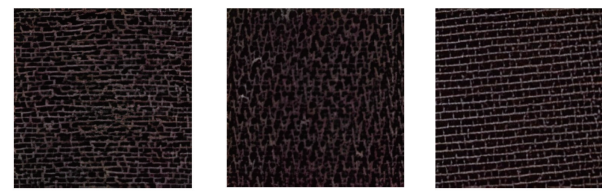

Fig. 1 Sample images of different fabrics with similar composition: (a) plain crepe fabric; (b) plain change fabric; (c) plain flat fabric

After the success of CNN, fine-grained feature extraction of fabric images has also entered the stage of deep learning. Although Zhang et al.[16] emphasized the importance of fine-grained features for wool fabric image retrieval and used hand-labeled features to further distinguish subtle differences between fabric images, the GoogLeNet network that they used had a large number of parameters and needed a long training time. The extraction of fine-grained features relied on manual annotation, which was costly and might be difficult to scale to handle more diverse types of fine-grained classes. Lu et al.[17] proposed an aggregation method that combined low-level features and deep features for image retrieval. Duan et al.[18] proposed a multilevel similarity-aware method

based on deep local descriptors for deep metric learning.

To solve the problems of the low accuracy of the network model and long training time due to the large number of parameters that emerge from the above studies, this paper adopts the lightweight neural network architecture MobileNetV3[19], uses depth-separable convolution to reduce the model size and the number of parameters, and applies the attention mechanism proposed in the transformer to improve the Acc[20].

1.2 MobileNetV3 network structure

MobileNetV3 is a lightweight neural network that combines MobileNetV1[21] and MobileNetV2[22], with higher accuracy and efficiency. Figure 2 shows the network structure of MobileNetV3. The RGB format of the image is used as input. After convolution operation, batch normalization (BN) processing and h-swish activation function, it enters the Bneck structure of the network. The small and the large are two versions of the MobileNetV3 network, which are respectively applicable to different resource requirements. Then, after average pooling, BN is performed again in the fully connected layer and the h-swish activation function is used to finally output a one-dimensional (1D) feature vector.

The Bneck is the core structure of the MobileNetV3 network, which consists of a depthwise separable convolution block and a SE block, as shown in Fig. 3. the MobileNetV3 network continues to use the depthwise separable convolution blocks in MobileNetV1 and MobileNetV2. The difference is that the SE module is introduced in the Bneck structure of the MobileNetV3 network. The MobileNetV3 network uses depthwise separable convolution to change the traditional convolution block and reduces the model capacity. Also during training, the SE module is used to focus more on the relevant features of each channel.

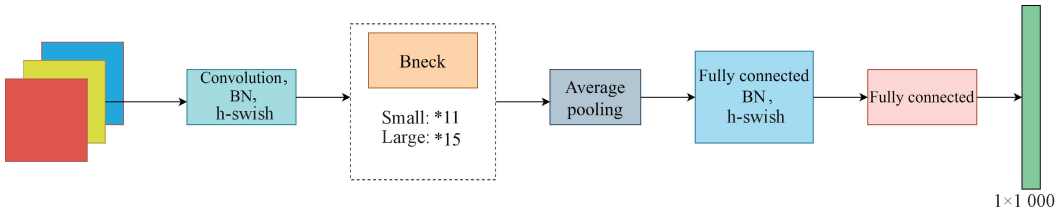

Fig. 2 MobileNetV3 network structure

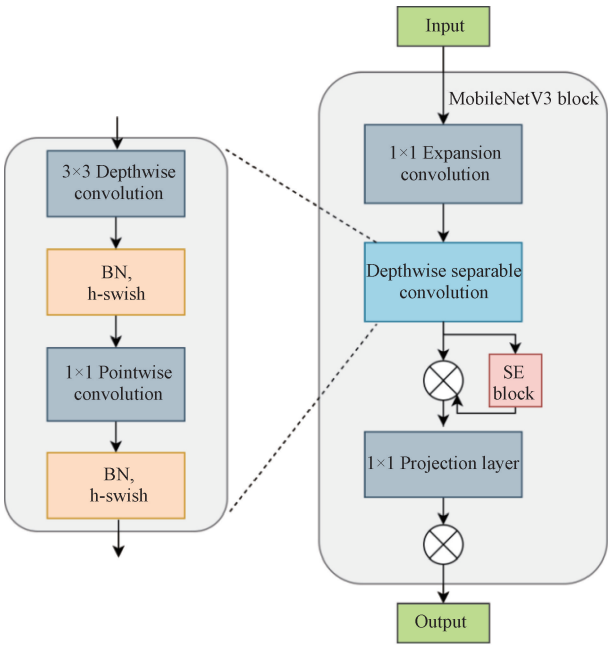

Fig. 3 MobileNetV3 Bneck structure

The SE attention module effectively establishes the interdependencies between channels by simply compressing each two-dimensional (2D) feature map. The specific workflow is shown in Fig. 4. It consists of two main parts: squeeze and excitation. The squeeze part is to perform the global pooling operation on the original feature map, and the excitation part is to perform the fully connected and rectified linear unit (ReLU) function activation on the squeeze feature map before performing the fully connected and sigmoid activation functions. After these two steps, the SE module calculates the weight value of each channel by multiplying the channel weights with the 2D matrix of the corresponding channels of the original feature map using scaling. H, W and C in the image refer to the height, width and number of channels of a feature map, respectively.

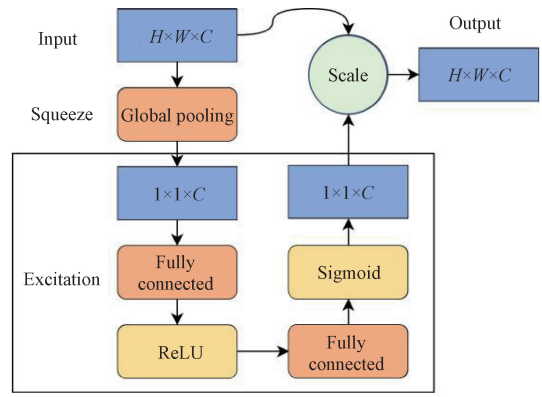

Fig. 4 Workflow of SE module

However, the SE module focuses on the channel dimension of the feature map and ignores the location information of the target. The location information is important for generating spatially selective attention maps. In addition, the SE module increases the total number of parameters and the total computation of the network. This is because although the computation of the fully connected layer used by the SE module is smaller

than that of the convolutional layer, it significantly increases the number of parameters of the model. Therefore, the SE module is replaced with the coordinate attention (CA)[23] module to improve the network. To improve the accuracy of the model for fabric image retrieval, we focus on the key features of the images, i.e., fine-grained features, and suppress unnecessary features by an improved attention mechanism to overcome the drawbacks of its original attention model.

2 Fabric Fine-Grained Feature Extraction

2.1 Network structure

One sign of fine-grained recognition is that it needs rich and expressive appearance descriptors, because traditional descriptors such as scale-invariant feature transform (SIFT)[24] and histogram of oriented gradients (HOGs)[25] may not achieve the correct balance between the discrimination and the invariance of fine-grained classes. The fine-grained feature extraction of the fabric first needs to distinguish the fine details of the fabric's appearance, including the underlying texture details formed by the fabric organizations, densities and weave structures. Therefore, it needs a visual representation to retain the details that are important to distinguish and discard unnecessary information. In addition, it is necessary to find and locate different details, including the general appearance composed of stripes, checks, patterns or dots of the fabric. The central idea of fabric fine-grained feature extraction is to learn fabric features and details to form a unified object description. To this end, the MobileNetV3 neural network model is used to train the feature descriptor end-to-end, so that the descriptor can adapt to the characteristics of each category.

Deep separable convolution is a key feature of the MobileNet series and a major factor in its lightweight role. As shown in Fig. 5, the deep separable volume integration is divided into two processes: a channel-wise separable convolution (depthwise); a normal 1×1 convolution (pointwise). Together, the two processes output the specified number of channels. C' in Fig. 5 represents different types of convolutions.

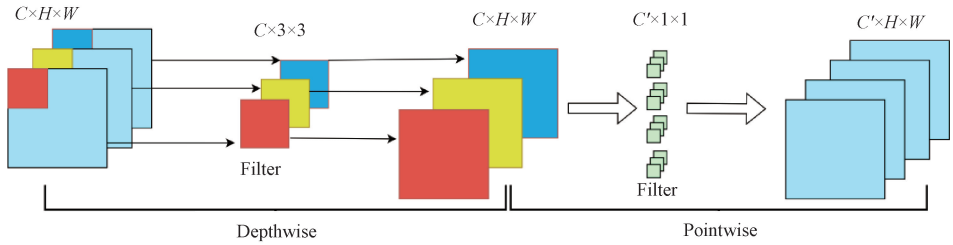

Fig. 5 Deep separable convolution

2.1.1 CA module replacement of SE module

We replace the SE module in the network layers of the original MobileNetV3 network which uses 5×5 convolutional kernels with a CA module that superimposes the mapping formed by the width and the height on the image's precise location information, allowing the network to perform unsupervised detail mining for fine-grained detection by learning appearance descriptors, including detailed features and spatial information. On the one hand, the CA module uses a smaller total number of parameters and a smaller total computation than the SE module; on the other hand, the network model with the addition of the CA module will mine the important features of the image from a low level to a high level along the two main dimensions of the channel and the space. By learning features that are suitable for describing the class of objects of interest, we let the data determine which features are effective for differentiation and which features help avoid losing information useful for classification.

2.1.2 Scale scaling factor

In addition, subtle differences in the appearance of fine-grained features may still be lost in the quantization of the neural network. Consider the probability that a CNN accepts pixels as its input and output classes. For this purpose, a scale scaling factor is used to change the channel coefficients and depth coefficients of the MobileNetV3 network. The improved model is shown in Fig. 6.

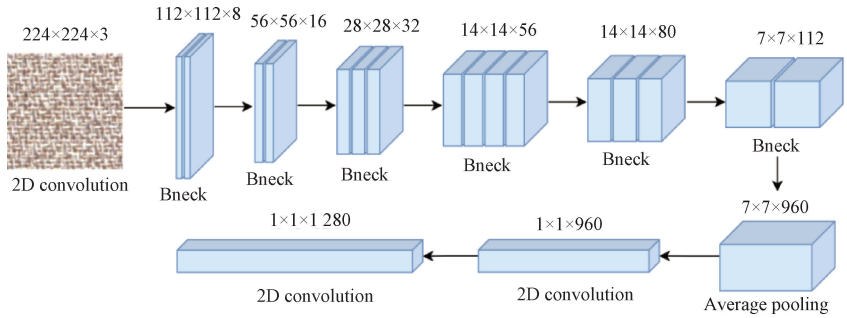

Fig. 6 Improved MobileNetV3 network framework

Figure 6 mainly shows the overall structure of the network and the change of the parameters of each network layer with a parameter scaling of 0.7 times of the original MobileNetV3 network. Firstly, the input fabric images are resized uniformly to $224 \times 224 \times 3$, and then after a series of Bneck structures, average pooling and 2D convolution, a desired 1D fabric fine-grained feature vector is obtained. The main purpose of this design is to prevent overfitting to preserve richer appearance descriptors in the output representation.

2.1.3 Improved Bneck structure

The improved Bneck structure is shown in Fig. 7. The *X* average pooling and the *Y* average pooling refer to 1D horizontal global pooling and 1D vertical global pooling, respectively. Firstly, the extracted fabric fine-grained features are mapped into the network layer where 5×5 convolutional kernels are located through 3×3 convolutional kernels. Secondly, the average pooling operation is performed from both horizontal and vertical directions, followed by stitching together the feature maps in both horizontal and vertical directions that obtain the global perceptual field. Thirdly, the stitched feature maps are fed into the convolutional module with shared convolutional kernels of 1×1 to obtain two feature maps in horizontal and vertical directions with the same number of channels as the original one, respectively. Finally, the fabric fine-grained feature maps with attention weights in both horizontal and vertical directions will be obtained by multiplicative weighting calculation.

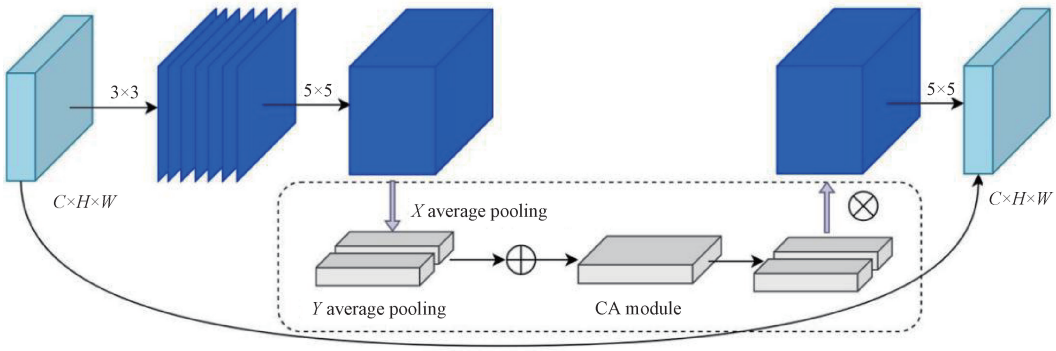

Fig. 7 Diagram of improved Bneck structure

2.2 Adjustment of network depth and width

Previous experiments have shown that for the improvement of CNN, researchers focus on three dimensions: network depths, network widths and resolutions. EfficientNet is designed from these three dimensions by using a series of fixed-scale scaling factors to unify the dimensions of the network. Although increasing the depth of the network can obtain richer features, if the depth of the network is too deep, it will face the problem of gradient disappearance. Similarly, increasing the width of the network can obtain higher fine-grained features, but this will increase the computational overhead and the storage cost. This study uses the idea of a series of fixed-scale scaling factors to adjust the width and the depth of the model. Since the MobileNetV3 network requires the channel value to be set to an integer multiple of eight, the channel value is obtained by multiplying the channel dimension by multiple factors and taking the integer multiple of eight as the final channel value, which is the smallest difference from the channel value. The depth dimension is then multiplied by a multiple and rounded off to obtain the depth value. After several comparison experiments, the channel values of the Bneck structure are finally determined to be eight, 16, 32, 56, 80 and 112, which are 0.7 times that of the original Bneck structure, while the depth remained the same as the original Bneck structure. This can reduce the total number of network parameters and reduce the computational cost on the basis of obtaining higher image fine-grained features. A comparison of the original model Bneck structure and the improved Bneck scaled model structure is shown in Fig. 8.

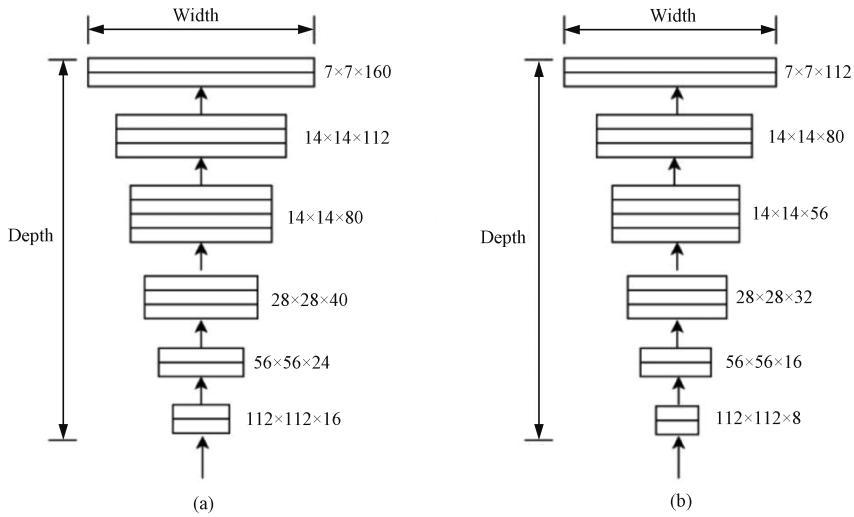

Fig. 8 Bneck structure: (a) original model; (b) improved scaled model

2.3 CA module

The CA module considers a more efficient way to obtain image location information and channel relationships to enhance the feature representation of the MobileNetV3 network and obtain richer texture fine-grained features such as densities and weave structures in fabric images. The CA module is an attention block that combines channel attention, *X*-directional space and *Y*-directional space information. Its specific operation is divided into two steps, namely coordinate information embedding and coordinate information generation. It captures and saves the precise location information of fine textures in fabric images through coordinate information embedding. Through coordinate information generation, the saved fabric texture location information is used to help the network, and can locate the area of fine textures in the image more accurately, thus improving the Acc.

2.3.1 Coordinate information embedding

The coordinate information embedding in the CA module is able to capture remote spatial interactions with precise location information, as shown in Fig. 9. The global pooling is decomposed and converted into a one-to-one 1D feature encoding operation, according to Eq. (1).

$$z_c = \frac{1}{H_c \times W_c} \sum_{i=1}^{H_c} \sum_{j=1}^{W_c} x_c(i, j), \tag{1}$$

where c represents one of the three channels; z_c represents the output of the image fine-grained features encoded horizontally and vertically in channel c; H_c and W_c represent the height and width of the region where the network is concerned with the image fine-grained features, respectively; x_c represents the input of the image texture features into channel c after the convolution layer.

Each channel is first encoded along the horizontal and vertical directions using pooling kernels of size $(H, 1)$ or $(1, W)$, respectively, to obtain the output of channel c at height i:

$$z_c^h(h) = \frac{1}{W_c} \sum_{0 \leqslant i < W_c} x_c(h, i), \tag{2}$$

where i represents the height.

Similarly, the output of channel c at width j is

$$z_c^w(w) = \frac{1}{H_c} \sum_{0 \leqslant j < H_c} x_c(j, w), \tag{3}$$

where j represents the width.

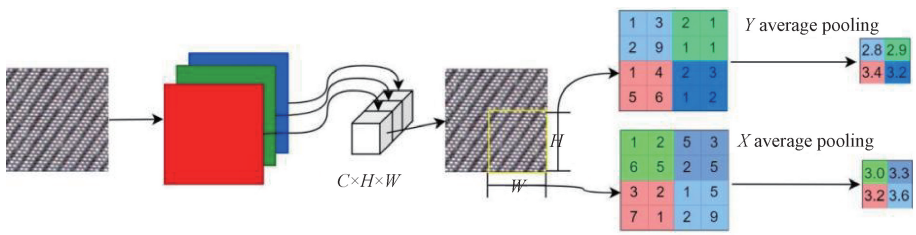

Fig. 9 Coordinate information embedding structure

The above two transformations can extract the fine-grained features of fabric images along two directions separately and obtain a pair of feature maps based on different directional perceptions, which is superior to the SE module that generates a single feature vector. The two transformations also enable the CA module to capture long-term dependencies along one spatial direction and preserve precise location information along the other spatial direction, thus helping the network to locate the region where the fabric fine-grained features are placed

more accurately.

2.3.2 Coordinate information generation

Coordinate information generation can better utilize fine-grained features of images generated by incorporating coordinate information embeddings, as shown in Fig. 10.

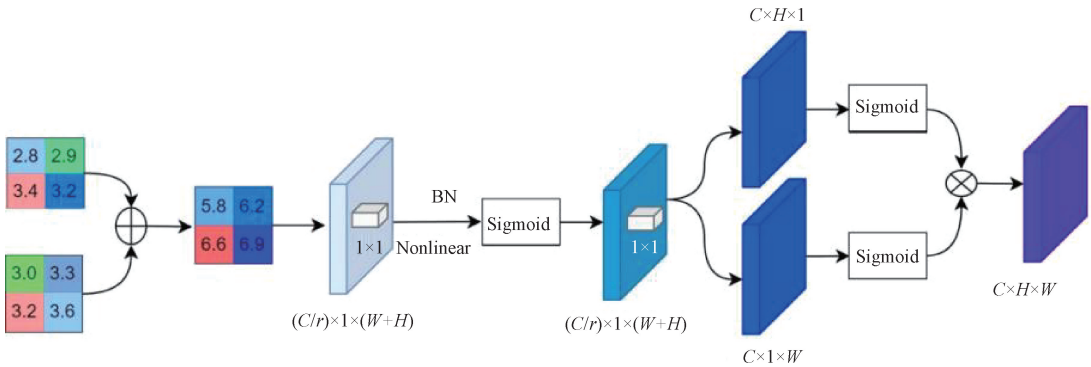

Fig. 10 Coordinate information generation structure

Figure 10 shows that after going through the transformations in the above information embedding, the part performs a concatenate operation on the above transformations and then performs a transform operation on them using the 1×1 convolutional transform function, as shown in

$$f = \delta(F_1([z^h, z^w])), \tag{4}$$

where $[z^h, z^w]$ represents the concatenate operation along the spatial dimension; δ is the nonlinear activation function; f is the intermediate feature mapping that encodes the spatial information in horizontal and vertical directions, and is decomposed into two separate tensors $\boldsymbol{f}^h \in \mathbf{R}^{(C/r)\times H}$ and $\boldsymbol{f}^w \in \mathbf{R}^{(C/r)\times W}$ along the spatial dimension, and r is the reduction factor. The attention weights g^h in the height direction and g^w in the width direction of the fabric feature map are obtained using two additional 1×1 convolutional transformations F_h and F_w, which are transformed into tensor inputs with the same number of channels to x, respectively:

$$g^h = \sigma(F_h(\boldsymbol{f}^h)), \tag{5}$$

$$g^w = \sigma(F_w(\boldsymbol{f}^w)), \tag{6}$$

where σ denotes the sigmoid activation function.

The outputs g^h and g^w are combined into a weight matrix by multiplicative weighting calculation on the original fabric feature map to obtain the final fabric feature map with attention weights in height and width directions:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j), \tag{7}$$

where g_c is the result of the channel number transform and the convolutional transform.

In summary, the workflow diagram of the CA module is shown in Fig. 11.

The feature maps of the height and the width of the global perceptual field are concatenated together and fed into a convolution module with a shared convolution kernel of 1×1 to reduce their dimensionality to the original C/r. The BN and nonlinear feature maps are fed into the sigmoid activation function to obtain a feature map shaped as $(C/r)\times 1\times(W+H)$. Then the feature map is convolved with a convolution kernel of 1×1 according to the original height and width to obtain two feature maps with the same number of channels as the original, respectively. After the sigmoid activation function, the attention weights of the feature maps in terms of height and width are obtained. The fabric feature maps with attention weights in height and width directions will be obtained by multiplicative weighting calculation.

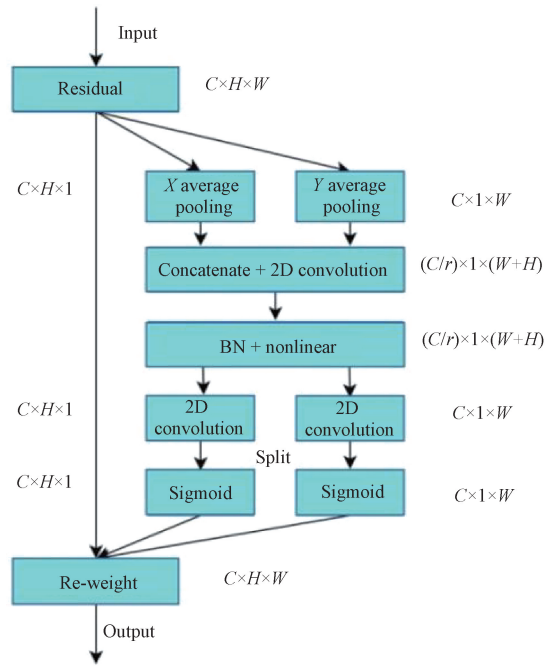

Fig. 11 CA module workflow diagram

2.3.3 Improvement of attention mechanism

The CA module can locate efficiently on the pixel coordinate system, so that the network model can focus on the area where the fine-grained features of the fabric are located and obtain information in a larger range, thus

obtaining better retrieval results. The original SE module in the MobileNetV3 network only considers the information encoded between channels and ignores the spatial information, which will waste the texture information of fabric images obtained from the 5 × 5 convolutional kernels in the Bneck structure in the MobileNetV3 network. Thus the corresponding SE module is replaced with the CA module. At the same time, the SE module used in the original 3 × 3 convolutional kernel is retained, because for the 3 × 3 convolutional kernel, the use of the CA module not only has little effect on improving the Acc, but also increases the total number of network parameters. The improved Bneck structure is shown in Fig. 12. The design pattern of the SE module used in the original 3×3 convolutional kernel is kept unchanged, and only the corresponding SE module is replaced by the CA module between the convolutional layers using the 5×5 convolutional kernel. In Fig. 12, Dwise refers to depthwise separable convolution, and FC stands for fully connected layer.

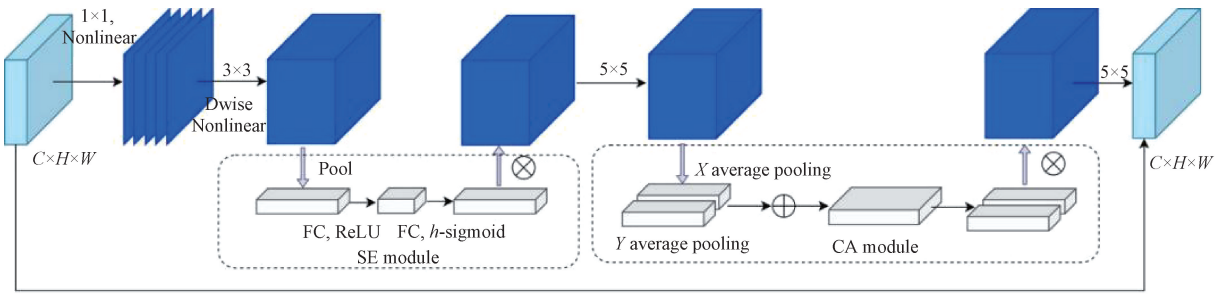

Fig. 12 Structure of Bneck with SE module retained and CA module added

The results of the experiments comparing the two design patterns are shown in Table 1. Table 1 shows that the number of network parameters after replacing all SE modules with CA modules reaches 4.27 MB and the Acc is 87.36%, while the total number of network parameters obtained by the proposed method is 4.54 MB. Although the number of parameters increases by 0.27 MB, the Acc reaches 89.45%, with an improvement of 2.09 percentage points. It is worthwhile to improve the overall Acc of the network at the cost of increasing the number of subtle parameters, and the experimental results verify the superiority of the improved method proposed in this paper.

Table 1 Test results of different Bneck model designs

Method	Parameter/MB	Acc/%
Bneck	5.48	78.33
Bneck(CA)	4.27	87.36
Bneck(SE+CA)	4.54	89.45

Note: Bneck refers to the original modules; Bneck (CA) refers to the replacement of all the original SE modules with CA modules; Bneck (SE+CA) refers to the replacement of the SE modules with CA modules in the original Bneck structure using 5×5 convolutional kernel network layers.

With the above design, the improved network can not only capture remote correlations in one direction but also retain accurate position information in the other direction. In addition, the computational burden from the 5×5 convolutional kernel can be offset due to the low computational effort of the CA module. The parameters of the Bnecks model are shown in Table 2.

Table 2 Parameters of Bneck models in MobileNetV3 large+CA

Bneck ID	Kernel size	Exp size	#out	Attention block
1	3×3	16	16	0
2	3×3	64	24	0
3	3×3	72	24	0
4	5×5	72	40	CA
5	5×5	120	40	CA
6	5×5	120	40	CA
7	3×3	240	80	0
8	3×3	200	80	0
9	3×3	184	80	0
10	3×3	184	80	0
11	3×3	480	112	SE
12	3×3	672	112	SE
13	5×5	672	160	CA
14	5×5	960	160	CA
15	5×5	960	160	CA

Note: Exp size means the dimension of the first ascending convolution output; #out means output dimensions; 0 means that no attention module is used; SE means that the SE module is used; CA means that the CA module is used.

2.4 Compound scaling

To verify the effectiveness of the CA module proposed in this paper, the gradient-weighted class activation mapping (Grad-CAM) is used to visualize the fabric images. Grad-CAM[26] is a CNN feature visualization method proposed in 2017, and Grad-CAM is an upgraded version of class activation mapping (CAM). Compared to CAM[27], Grad-CAM can visualize CNNs with any structure of the model, does not require modifying the network structure or retraining, and is more general than CAM while being able to find the most

interesting regions of the network. Figure 13 shows the Grad-CAM display images. It shows the specific region of the fabric image that the network focuses on more intensely after the attention mechanism in the MobileNetV3 network has completed the improvement, i. e. , the feature information that the attention mechanism focuses on. The redder the color, the more important the area is to the classification task. The improvement found by the comparison shows that the area of interest to the input images increases and is closer to the center. It shows that the improved network has stronger feature learning ability and classification ability, which improves the retrieval performance.

Figure 13 shows the Grad-CAM display images. The heat map activated by generating class activation for the input image can be understood as the contribution distribution to the predicted output. The redder the color, the higher the response and contribution of the corresponding area of the original image to the network, indicating the importance of each position to the class. The larger the red area, the stronger the ability of the method to extract fine-grained features.

The comparison shows that compared with the MobileNetV3 network, the improved attention mechanism of the network can more accurately capture the texture information and the local areas that have a significant impact on the retrieval results, and thus focus on the fine-grained features of the fabric.

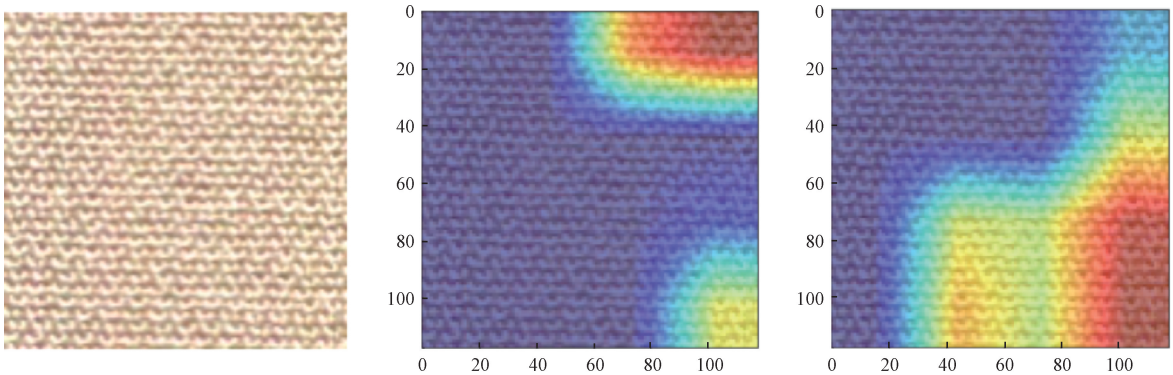

Fig. 13 Grad-CAM display images: (a) input image; (b) Grad-CAM image of the MobileNetV3 network; (c) Grad-CAM image after improving attention mechanism

3 Analysis of Experimental Results

To verify the accuracy of the fabric image retrieval algorithm based on fine-grained features proposed in this paper, fabric image retrieval experiments are designed, as shown in Fig. 14. Firstly, the images in the fabric image dataset are input to the pre-trained network model, and the fine-grained features such as the textures of the fabric images are extracted to build the feature database. Secondly, the fine-grained features of the query images are extracted and their feature vectors are compared with the vectors in the feature database to calculate the similarity. Finally, the top five most similar images are obtained by sorting.

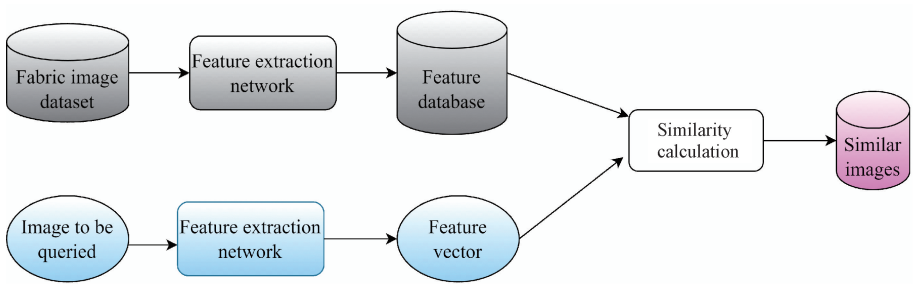

Fig. 14 Fabric image retrieval framework

3.1 Dataset

According to the investigation, there is no public and standard dataset as a benchmark for fabric image retrieval. To verify the effectiveness of the proposed method and provide data support for the experiments, the fabric image dataset from a textile factory, referred to as FID, is therefore used in this paper. A scanner with a resolution of 300 dpi is used to capture 32 860 images, which are uniformly cropped to 354 pixel×354 pixel and divided into 14 categories of images. These 32 860 images are randomly divided into the training set, the validation set and the test set according to the ratio of 7∶2∶1. In this paper, the fine-grained features of fabrics are mainly studied. Thus the macroscopic features of fabrics are restricted to the broad category of plain class, and the plain class is subdivided into 14 subcategories under the plain class, such as plain variable qiao class (PBQ), plain crepe type (PBZ), plain satin fabrics (PDC), plain satin

trees (PDQ), plain satin and crepe fabrics (PDZ), plain plain silk (PPC), plain pingqiao class (PPQ), plain flat crepe fabric (PPZ), plain bilayer class (PSC), plain shunyu class (PSY), plain diagonal silk (PXC), plain oblique coarse class (PXCU), plain oblique arbor class (PXQ) and plain diagonal crepe (PXZ). The abbreviations of the fabric names in brackets come from the factory's habitual descriptions of these fabrics. Figure 15 shows the examples of the fabric images from different subcategories under the same broad category. It can be seen that the appearance of fabrics can be very different even for the same broad category.

Fig. 15 Example of fabric images

3.2 Similarity measurement

The feature vectors are extracted from the query images using the feature extraction network proposed in this paper. The distances between them and all the feature vectors in the fabric feature library are calculated and the results are sorted and output. Since the cosine similarity is not affected by dimensionality and performs well in high-dimensional cases, and the algorithm is relatively simple, this paper uses the cosine distance to perform the similarity matching of fabric images:

$$d(x, y) = 1 - \cos(x, y) = 1 - \frac{xy}{|x||y|}, \quad (8)$$

where x and y denote two n-dimensional eigenvectors, i.e. $x = \{x_1, x_2, \cdots, x_n\}$, $y = \{y_1, y_2, \cdots, y_n\}$; $d(x, y)$ denotes the distance between the two vectors; $\cos(x, y)$ is used to indicate the similarity of these two vectors, and the closer the value to 1, the more similar the candidate image to the query image.

3.3 Evaluation indicators

Since the purpose of the algorithm proposed in this paper is to provide a reference for industrial applications, both precision and recall are important when evaluating the retrieval results of fabric images. Precision is the proportion of the number of similar images among all retrieved images in the total retrieval results. Recall, also known as the full search rate, refers to the proportion of similar images retrieved among all similar images in the dataset during a single search. The formulae for the precision P and recall R are shown as

$$P = \frac{N_r}{R_t} \times 100\%, \quad (9)$$

$$R = \frac{N_r}{N_t} \times 100\%, \quad (10)$$

where N_r is the number of similar images retrieved; R_t is the total number of images retrieved; N_t is the total number of relevant images.

3.4 Compound scaling

The proposed algorithm refers to the idea of EfficientNet scaling coefficients, and the following schemes are designed for a series of scale coefficients of the Bneck structure in the MobileNetV3 network: a width factor of 0.6 combined with a depth factor of 1.0 ($w_m \times 0.6 + d_m \times 1.0 + \mathrm{SE}$), a width factor of 0.7 combined with a depth factor of 0.7 ($w_m \times 0.7 + d_m \times 0.7 + \mathrm{SE}$), and a width factor of 0.7 combined with a depth factor of 1.0 ($w_m \times 0.7 + d_m \times 1.0 + \mathrm{SE}$), where w_m is the width of the model and d_m is the depth of the model. Then the dichotomous idea is used to select coefficients from both the width and the depth, and the precision and the recall of the MobileNetV3 network extracted fabric fine-grained features are compared. The experimental results are shown in Fig. 16. When the scaling coefficients of the Bneck structure is chosen as 0.7 for w_m and 1.0 for d_m, the detection precision improves by 15.50 percentage points and the recall rate improves by 15.40 percentage points for all 14 kinds of fabric images compared with the original MobileNetV3 network model. This is enough to prove the effectiveness of the Bneck scale scaling coefficients proposed in this paper on image fine-grained feature extraction.

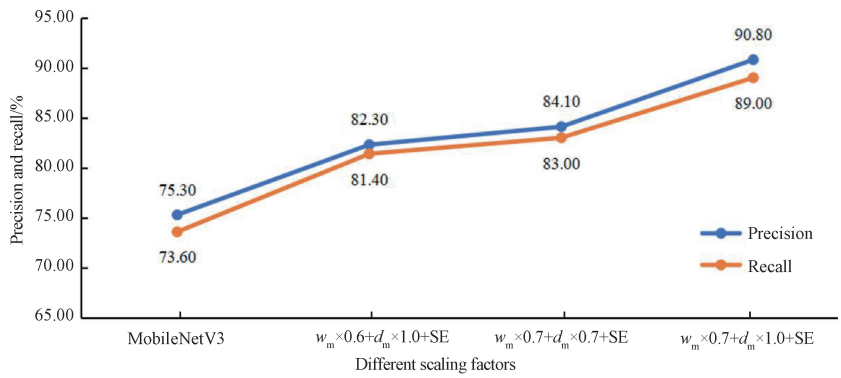

Fig. 16 Comparison analysis with original MobileNetV3 network using different scaling factors

In this study, the lightweight neural network model MobileNetV3 is chosen as the base network, and the scale scaling factor is used to change the channel coefficient and the depth coefficient of the MobileNetV3 network. Through experimental comparison, the empirical value of the Bneck channel scaling factor is finally determined to be 0.7 and the depth scaling factor to be 1.0. This setting further reduces the overall parameter operations of the network on top of enhancing the Acc achieved in the previous study on improving the attention mechanism, thus reducing the network training time and mitigating the computational cost. The results of the comparison experiments are shown in Table 3. Based on the 89.45% Acc achieved by the improved SE module in the MobileNetV3 network, the highest Acc of 91.82% is achieved by setting the empirical value of the Bneck channel scaling factor to 0.7 and the depth scaling factor to 1.0, while the parameter size was reduced by 0.8 MB. The computational cost is reduced.

Table 3 Test results with different scaling factors

Method	Parameter size /MB	Acc/%
MobileNetV3+CA+SE	4.54	89.45
MobileNetV3+w_m×0.6+d_m×1.0+CA+SE	3.55	90.31
MobileNetV3+w_m×0.7+d_m×0.7+CA+SE	3.62	90.96
MobileNetV3+w_m×0.7+d_m×1.0+CA+SE	3.74	91.82

Notes: MobileNetV3+CA+SE refers to the network after improved attention design; MobileNetV3+w_m×0.6+d_m×1.0+CA+SE refers to the combination of width factor 0.6 and depth factor 1.0 on the basis of improved attention design; MobileNetV3+w_m×0.7+d_m×0.7+CA+SE refers to the combination of width factor 0.7 and depth factor 0.7 on the basis of improved attention design; MobileNetV3+w_m×0.7+d_m×1.0+CA+SE refers to combining width factor 0.7 with depth factor 1.0 based on improved attention design.

3.5 Model training

The experiments are implemented using the PyTorch framework, with a server environment of Ubuntu 18.04 and 22 GB of video memory. The training parameters batch_size is set to 16, the epoch is set to 100, and the Adam optimizer is used. The learning rate is initially set to 0.000 1, and the hyperparameter is set to 1.0, which gradually decays with each training round.

The process of loss change in the training and validation sets is shown in Fig. 17.

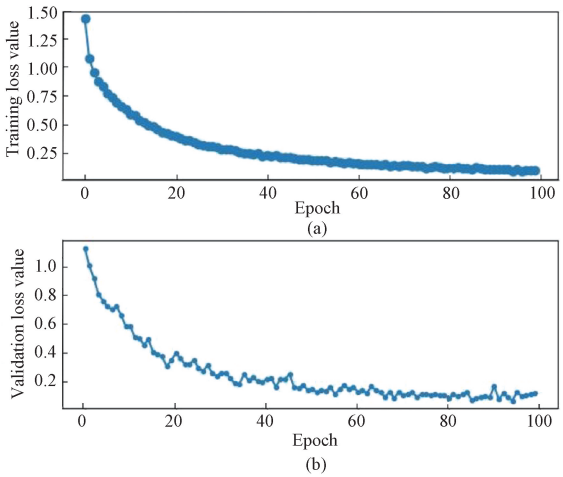

Fig. 17 Loss variation diagram: (a) training set; (b) validation set

As seen in Fig. 17, the loss decreases very quickly within the first 20 epochs and stabilizes around the 60th epoch. After the 80th epoch, the losses in the training and validation sets decrease to about 0.1. It is clear that the model has achieved better results after 80 epochs and can be used to extract fine-grained features such as fabric textures.

3.6 Experimental results

FID has five different position image data for the same fabric and has corresponding numbers. In the test phase, when the search results are returned, the precision and the recall can be determined based on the number of search results. Select 100 pictures randomly from the test set for retrieval, and the retrieval example is shown in Fig. 18. It can be seen that after inputting a fabric image, the five most similar images are displayed and the five images obtained are all in the same category, differing only in colors. Comparison experiments are then conducted to verify the superiority of the feature extraction network proposed in this paper. The effect of the improved attention mechanism on the results of the first five images obtained by retrieval is verified by the precision and the recall as indicators.

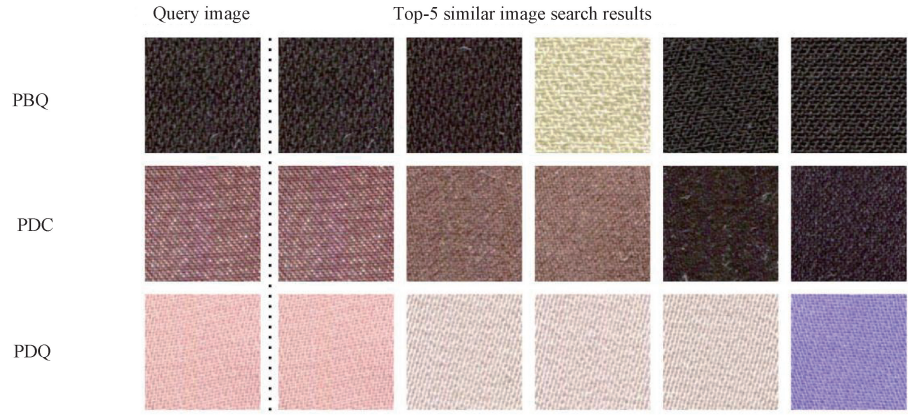

Fig. 18 Example of partial retrieval of query images for different categories

The experimental results are shown in Fig. 19, where MobileNetV3 refers to the retrieval result obtained by using the original MobileNetV3 network, MobileNetV3 + CBAM refers to the retrieval results obtained by replacing the SE attention module of the Bneck structure of MobileNetV3 with the convolutional block attention module (CBAM) using 5×5 convolutional kernels, and MobileNetV3 + CA refers to the retrieval results obtained by replacing the SE attention module with the CA module using 5×5 convolutional kernels.

Figure 19 shows the comparison of the precision and the recall of different attention mechanisms applied to the MobileNetV3 feature extraction network model, from which it can be seen that the precision and recall of CBAM using the hybrid attention mechanism are 88.80% and 86.80%, respectively, both of which are higher than the original MobileNetV3 benchmark values of 75.30% and 73.60%. CA used in this paper performs better, with the precision and the recall improving to 91.00% and 89.60%, respectively, indicating that CA used in this paper is more suitable for fabric image retrieval and has a better ability to extract fine-grained features.

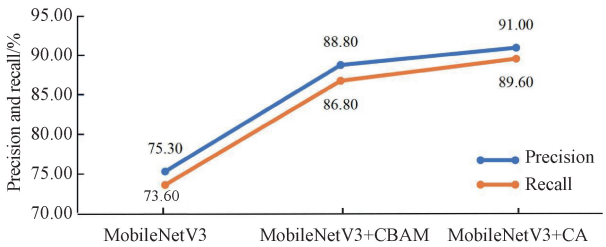

Fig. 19 Comparison of precision and recall with improved attention mechanism

In order to verify the classification ability of the proposed model on 14 kinds of fabric images, the pre-trained model is used to predict the classification of fabric images that are not involved in the training, and the confusion matrix results are shown in Fig. 20. Each column represents the predicted category of fabric images, each row represents the actual category of fabrics, and each cell represents the number of fabric images in which the true category of the current row is predicted as the predicted category in the current column. Take the first row and the first column of PBQ as an example, all the values in the first row sum to 225, and all the values in the first column sum to 222, where 218 represents the number of results that are currently predicted as PBQ and are actually PBQ, i.e. the number of results that are correctly predicted. The number 4 in the first row represents the number of results predicted to be PBQ but actually PBZ, i.e., the number of misclassified results.

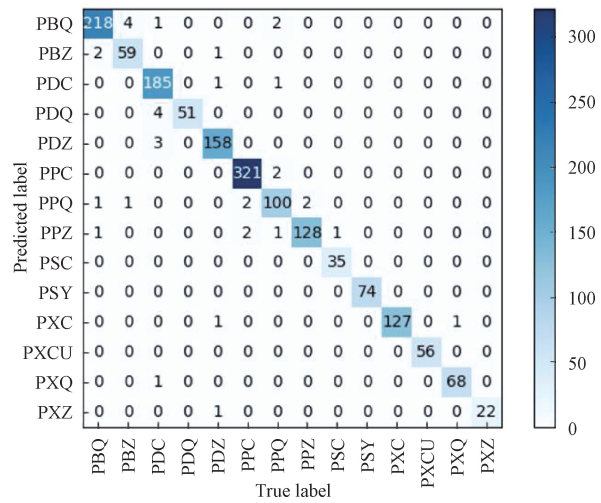

Fig. 20 Confusion matrix results for fabric images not involved in training

From Fig. 21, it can be seen that the lowest recall among the 14 fabric images is PBZ with 92.20%, and the lowest precision is PDQ with 92.70%. The reason for this is that the samples provided by the two fabric enterprises are too small compared with other kinds of fabrics, resulting in poorer prediction results compared with other kinds of fabric images. However, the overall

analysis combined with the graphs shows that the prediction precision and recall of each fabric image are above 90.00%, the average prediction precision and the average prediction recall are 97.50% and 97.90%, respectively, which proves the effectiveness and the practicality of the model in this paper.

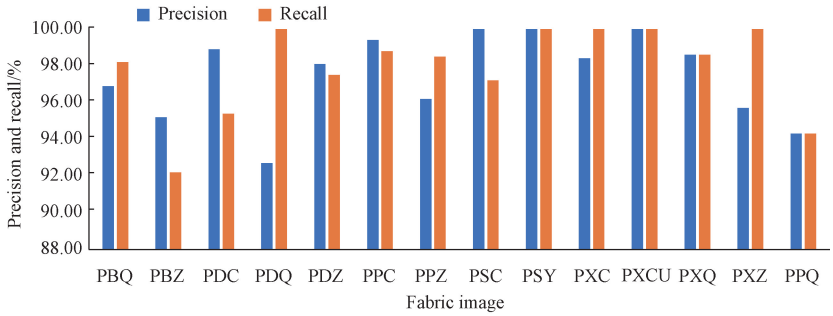

Fig. 21 Prediction results for fabric images not involved in training

To test the performance of the proposed algorithm, we calculate the floating point operations (FLOPs), parameter size, Acc and average retrieval time (Te) for the original MobileNetV3, MobileNetV3 with a scaling factor of 0.7 (MobileNetV3+0.7 factor) and the network designed in this paper (MobileNetV3 + 0.7 factor + CA), multiple part-level feature ensemble (MPEE)[28], and hard decorrelated centralized loss (HDCL)[29]. The results are shown in Table 4. For MobileNetV3 + 0.7factor+CA, the FLOPs is 175.34 MB, indicating that the inference speed improves by 25.14% compared with the FLOPs of MobileNetV3 (234.24 MB); the parameter size is 3.74 MB, indicating that the number of output parameters reduces by 1.74 MB compared with MobileNetV3 (5.48 MB); the Acc is 91.82%, improves by 13.5 percentage points compared with MobileNetV3 (78.33%).

Table 4 Test results of different network models

Method	FLOPs/MB	Parameter size/MB	Acc/%	Te/s
MobileNetV3	234.24	5.48	78.33	5.2
MobileNetV3+0.7 factor	173.41	4.68	90.96	2.0
MobileNetV3+0.7 factor+CA	175.34	3.74	91.82	1.8
MPFE	221.76	5.89	84.37	3.2
HDCL	265.84	6.03	86.13	3.5

4 Conclusions

In this paper, a fabric image retrieval algorithm based on fine-grained features is proposed. Firstly, the algorithm employs the CA to replace part of the SE attention mechanism of the original model to extract fine-grained features, further reducing the error of fabric image retrieval. Secondly, by proposing a Bneck scale scaling factor applicable to fabric fine-grained feature extraction tasks, the feature extraction accuracy is further improved and the parameter computation of the original MobileNetV3 model is reduced, which provides a new idea for future research directions. Meanwhile, the fabric retrieval experiments show that the proposed algorithm has an excellent performance in terms of the parameter size, FLOPs and the Acc compared with the benchmark MobileNetV3 model, with the parameter size reaching 3.74 MB, FLOPs reaching 175.34 MB and Acc reaching 91.82%. The Acc is 13.49 percentage points higher than the original MobileNetV3 model, while the training time of the network is reduced and the inference speed increases by 25.14%. Additional work is currently under way to develop a more robust fabric image retrieval system by employing additional color and texture features to improve the performance of the system.

References

[1] LIU P Y, JIA K B, WANG Z Z. An Effective image retrieval method based on color and texture combined features [C]//Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007). Los Alamitos, California: IEEE, 2007: 169-172.

[2] LATIF A, RASHEED A, SAJID U, et al. Content-based image retrieval and feature extraction: a comprehensive review [J]. *Mathematical Problems in Engineering*, 2019, 2019(2019): 1-21.

[3] SUCIATI N, HERUMURTI D, WIJAYA Y A. Fractal-based texture and hsv color features for fabric image retrieval [C]// 2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE). New

York: IEEE, 2015: 178-182.
[4] JING J F, LI Q, LI P F, et al. A new method of printed fabric image retrieval based on color moments and gist feature description[J]. *Textile Research Journal*, 2016, 86(11): 1137-1150.
[5] LI Y Y, LUO H C, JIANG G M, et al. Content-based lace fabric image retrieval system using texture and shape features[J]. *The Journal of The Textile Institute*, 2019, 110(6): 911-915.
[6] SZEGEDY C, LIU W, JIA Y Q, et al. Going Deeper with Convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2015: 1-9.
[7] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2016: 770-778.
[8] SHEN F, WEI M W, LIU J J, et al. RGB and LBP-texture deep nonlinearly fusion features for fabric retrieval[J]. *High Technology Letters*, 2020, 26(2): 196-203.
[9] ZHOU H L, PENG Z F, TAO R, et al. Feature fusion multi-MNet convolution neural network for clothing image classification[J]. *Journal of Donghua University (English Edition)*, 2021, 38(6): 519-526.
[10] ZHA J H, YAN C R, ZHANG Y T, et al. Image retrieval with text manipulation by local feature modification[J]. *Journal of Donghua University (English Edition)*, 2023, 40(4): 404-409.
[11] LI F, PAN H S, SHENG S X, et al. Image retrieval based on vision transformer and masked learning[J]. *Journal of Donghua University (English Edition)*, 2023, 40(5): 539-547.
[12] XIANG J, PAN R R, GAO W D. Mélange fabric image retrieval based on soft similarity learning[J]. *Journal of Engineered Fibers and Fabrics*, 2022, 17: 155892502210888.
[13] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2009: 248-255.
[14] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 42(8): 2011-2023.
[15] TAN M X, LE Q V. EfficientNet: rethinking model scaling for convolutional neural networks[C]//Proceedings of the 36th International Conference on Machine Learning (PMLR). Los Angeles: ICML, 2019: 6105-6114.
[16] ZHANG N, SHAMEY R, XIANG J, et al. A novel image retrieval strategy based on transfer learning and hand-crafted features for wool fabric[J]. *Expert Systems With Applications*, 2022, 191(191): 1-14.
[17] LU Z, LIU G H, LU F, et al. Image retrieval using dual-weighted deep feature descriptor[J]. *International Journal of Machine Learning and Cybernetics*, 2023, 14(3): 643-653.
[18] DUAN C C, FENG Y, ZHOU M L, et al. Multilevel similarity-aware deep metric learning for fine-grained image retrieval[J]. *IEEE Transactions on Industrial Informatics*, 2023, 19(8): 9173-9182.
[19] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Los Alamitos, California: IEEE, 2019: 1314-1324.
[20] VASWANI A, SHAZEER N M, PARMAR N, et al. Attention is all you need[C]// Conference and Workshop on Neural Information Processing Systems (NIPS). La Jolla, California: NIPS, 2017: 5998-6008.
[21] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2023-01-30]. https://arxiv.org/abs/1704.04861.
[22] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, California: IEEE, 2018: 4510-4520.
[23] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, California: IEEE, 2021: 13708-13717.
[24] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. *International Journal of Computer Vision*, 2004, 60(2): 91-110.
[25] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2005: 886-893.
[26] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. *International Journal of Computer Vision*, 2020, 128(2): 336-359.
[27] WANG H F, WANG Z F, DU M N, et al. Score-CAM: score-weighted visual explanations for convolutional neural networks[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Los Alamitos, California: IEEE, 2020: 111-119.

[28] CAO G, ZHU Y, LU X. Fine-grained image retrieval via multiple part-level feature ensemble [C]//2021 IEEE International Conference on Multimedia and Expo. New York: IEEE, 2021.

[29] ZENG X X, LIU S, WANG X D, et al. Hard decorrelated centralized loss for fine-grained image retrieval [J]. Neurocomputing, 2021, 453: 26-37.

基于细粒度特征的面料图像检索

罗 辛*, 夏冬梅, 陶 然, 史有群

东华大学 计算机科学与技术学院, 上海 201620

摘 要: 面料图像检索对于纺织工厂面料库存和样品管理意义重大, 但面料外观的多样性以及织物纹理的精细性, 使得在面料检索时面料的特征提取较困难。该研究提出一种基于细粒度特征的面料图像检索算法。该算法使用坐标注意 (coordinate attention, CA) 模块来提取图像的精准位置信息, 并将缩放系数法用于在宽度和高度方面整体缩放 MobileNetV3 的网络结构以减少模型参数数量, 达到减少网络训练时间的目的。据此筛选出提取面料图像细粒度特征的最佳模型, 在面料图像数据集 (fabric image dataset, FID) 上进行面料检索实验。结果表明, 该算法有效提高了面料图像细粒度特征提取的准确性, 检索精度达到 91.82%, 浮点运算数达到 175.34 MB。检索精度比 MobileNetV3 原模型提高了 13.49 个百分点, 同时减少了网络训练时间, 速度提高了 25.14%。该算法具有实际应用价值。

关键词: 面料图像检索; MobileNetV3; 细粒度特征; 注意力机制; 缩放系数