TSFFNet: texture-shape feature fusion network for clothing style classification

Yaxin Zhao
School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, China

Feng Yu*
School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, China
Engineering Research Center of Hubei Province for Clothing Information, Wuhan, China
yufeng@wtu.edu.cn

Minghua Jiang
School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, China
Engineering Research Center of Hubei Province for Clothing Information, Wuhan, China

Xiaoxiao Liu
School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, China

Hua Wang
School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, China

Xinrong Hu
School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, China
Engineering Research Center of Hubei Province for Clothing Information, Wuhan, China

Abstract

With the explosion of data, clothing classification, as the most basic technology, plays a very essential role in the fashion field. However, garment styles are diverse, which makes garment features are vulnerable to interference. In addition, clothing classification that rely solely on texture characteristics may focus more on garment detailed information, resulting in low robustness and low accuracy. In this paper, we propose a two-stream fashion classification network (texture-shape feature fusion network, TSFFNet) for clothing style images, aiming to achieve an accurate and intelligent clothing classification. In our work, 1) we find that the shape characteristics of clothing help to classify clothing styles, 2) we adopt clothing image acquisition module to remove the interference of complex backgrounds in order to facilitate the network to focus more on the clothing shapes, 3) we fuse the texture features with the shape features of the clothing to extract valid features, and 4) based on the characteristics of texture and shape features in clothing, we improve the network structure to further increase network accuracy. Comprehensive and rich experiments demonstrate our discoveries and the effectiveness of our model. Our network can achieve 74.6% accuracy on the fashion dataset, which gains 12.4% improvement over using the best mainstream classification network alone.

Keywords: clothing style classification, feature extraction, convolutional neural network, two-stream network

*Corresponding author: Feng Yu

1. Introduction

The rapid development of online sales of fashion companies has aroused scholars' attention to sustainable fashion in the field of marketing[31]. Hence, many studies are committed to clothes recognition ([24, 40, 22, 46, 5]), retrieval ([23, 21, 15, 10]), recommendation ([11, 7]) and fashion trend prediction ([1, 27]). However, clothing classification, as the basic technical support in these fields, has a profound impact on the subsequent functions (e.g., clothing retrieval, clothing recommendation).

In this paper, our prime interest is the classification problem in fashion domain, which is highly demanded in all stages of e-commerce. Recently, various deep learning-based algorithms([24, 48, 40, 22, 4, 46, 30, 20, 38]) have been proposed to address the problem of clothes recognition. These algorithms have demonstrated that the performance of clothing classification can be improved with the

adoption of the neural network. These methods are mainly divided into two categories, one is to improve classification effect through landmark detection, and the other is to use hierarchical classification to obtain fine-grained classification results. In the large-scale DeepFashion dataset[24], the landmark information is defined as a set of keypoints on clothing images and each image has 4 to 8 landmark points (e.g., left & right sleeves, left & right collars, left & right waistlines, and left & right hems). Liu et al.[24] proposes a deep model FashionNet to learn the clothing features by jointly predicting clothes category and landmark localization on the DeepFashion dataset. Wang et al. [40] introduces a deep grammar model, Bidirectional Convolutional Recurrent Neural Networks (BCRNNs), with two attention mechanisms for enhancing fashion landmark detection and clothing category classification. Shajini et al.[30] proposes an attention-driven deep learning technique for tackling visual fashion clothes analysis in images, aiming to achieve clothing category classification and attribute prediction by producing regularised landmark layouts. These models in fashion classification show promising results but lack the consideration of the hierarchical nature of fashion annotations. Zhu et al.[48] introduces Branch-CNN (B-CNN), the most well-known branching CNN for hierarchical image classification. Cho et al.[4] proposes a fashion classification model that works in a hierarchical manner can help improve performance of fashion classification. Kolisnik et al.[20] proposes a hierarchical fashion image classification model, Condition-CNN, which addresses some of the shortcomings of the branching convolutional neural network in terms of training time and fine-grained accuracy.

Previous works are aware of the importance of landmark detection and hierarchical classification to the clothes recognition and have achieved particular success. However, there are still some problems in clothing classification: 1) in complex backgrounds, clothing features are easily disturbed, resulting in low classification accuracy, 2) relying only on texture features, it is difficult to improve the accuracy of clothing classification, and 3) clothing styles are diverse and changeable, some clothing parts have similar characteristics, which makes fine-grained classification difficult. Existing methods ignore the influence of clothing shape on the effect of clothing classification. Zhang et al.[46] prove that clothing classification is more dependent on the shape features of clothing. In our experiments, we also try to combine the two features by adding a branch to an existing model, which jointly learns the texture and shape features, and fine-tune it on the dataset. Though the idea of using both texture and shape features is straightforward, it is challenging to design proper network architecture to incorporate them into fashion classification tasks.

To make full use of these two features, we use measurement studies to find the factors that affect the accuracies when integrating the features into different models. Based on our measurement insights, we propose a novel combined classification network called TSFFNet to improve the accuracy of fashion classification. The contributions of this paper are summarized as follows:

- To solve the problem of interference caused by complex backgrounds in fashion images for clothing classification, a clothing acquisition module is proposed to extract clothing shapes from fashion images, allowing the network to focus more on clean clothing shape features.
- To make full use of the features of fashion images, we propose a two-stream network structure to effectively fuse the detail features of fashion images with the shape features of clothing images based on the differences between clothing styles, substantially improving the accuracy of fashion image classification.
- To address the challenge of fine-grained classification of clothing, we improve the structure of the network based on the characteristics of clothing features, allowing the network to extract more accurate features and achieve higher accuracy in the diverse and variable styles.

2. Related work

To achieve fast and accurate clothing image classification, we analyze the characteristics of clothing images and conduct experimental research from different angles to improve the accuracy.

2.1. Clothes Datasets

Several clothes datasets have been proposed such as [44, 12, 17, 43, 24, 47, 8]. They vary in size as well as amount and type of annotations. For example, Clothing1M[43], WTBI[12] and DARN[17] have 1M, 425K and 182K images respectively. They scraped category labels from metadata of the collected images from online shopping websites, making their labels noisy. In contrast, CCP[44], DeepFashion1[24], and ModaNet[47] obtain category labels from human annotators. Moreover, different kinds of annotations are also provided in these datastes. For example, DeepFashion1 labels 4 to 8 landmarks per image that are defined on the functional regions of clothes. The definitions of these sparse landmarks are shared across all categories, making them difficult to capture rich variations of clothing images. Furthermore, DeepFashion does not have mask annotations. By comparison, ModaNet[47] has street images with masks of single person but without landmarks. Unlike existing datasets, DeepFashion2[8] contains 491K images and 801K instances of landmarks, masks, and

bounding boxes, as well as 873K pairs. It is the most comprehensive benchmark of its kinds to date. Moreover, the semantic map provided by Deepfashion2 can help us get the outline of the clothing in the image.

2.2. Clothing Image Understanding

Deep learning based models have achieved great success in fashion field, such as clothes recognition ([24, 40, 22, 46]), retrieval ([23, 21, 15, 10]), recommendation ([11, 7]) and fashion trend prediction ([1, 27]).Earlier works use traditional image analysis methods (e.g. Candy[2], SIFT[26], HOG[6],) to extract fashion image features for the follow-up work, which are hard to grasp the most useful features of fashion images.

Different outer contours and characteristic elements constitute a variety of different clothing category, which reflects the practicality and specificity of clothing. Extracting characteristic elements in clothing images, such as textures, contours, or parts of clothing, can effectively help clothing classification. With the development of deep learning methods, fashion models have achieved prodigious success. As the fashion market grows and more clothing styles flood in the market, different types of clothing may have similar textures, similar parts and similar patterns, and these characteristics create the problem of difficult classification of clothing. However, the previous works([24, 40, 5]) is rarely carried out from the perspective of the clothing shape itself. In this paper, we analyze the characteristics of clothing outline and give corresponding optimization methods.

2.3. Clothing classification

Clothing classification are mainly divided into two categories, one is to improve classification effect through landmark detection, and the other is to use hierarchical classification to obtain fine-grained classification results.

In order to further accurately detect the clothes people are wearing for classification under complex backgrounds, solutions typically attempt to identify the important regions for each piece of clothing to help classifications[40, 22, 46]. Localization is performed by selecting regions in bounding boxes which contain an article of clothing, or by masking unimportant parts of the images using attention mechanisms. The accurate locations and rich amounts of landmarks in the fashion images can be a good assistance to fashion tasks such as clothes attribute recognition and clothing classification. Wang et al.[40] leverages the high-level human knowledge of landmarks and propose two important fashion grammars, dependency grammar capturing kinematics-like relation and symmetry grammar accounting for the bilateral symmetry of clothes. Li et al.[22] proposes a refined implementation of attention localization. The model presents a two-stream network that creates a landmark heatmap which is multiplied with the original image to effectively create an attention mechanism. The altered image is then classified using a straight forward convolutional multi-class classification network.

A hierarchical classification(HC) is a part of the classification task that maps input data into a defined hierarchical relationship, which can be represented in the form of a tree or a graph. To properly catch the relationship between layers, the HC algorithm must be able to label inputs to one or multiple paths in the class hierarchy. Cho et al.[4] proposes a new hierarchical classification model to improve the classification performance by using the hierarchical nature of fashion image annotations, such as ”trousers” and ”skirt” both having ”bottom” as the upper level. The model is designed to improve classification performance by exploiting the hierarchical information.

These techniques often augment the data with other information about the image to aid classification. The branch of research that focuses on object localization and classification techniques frequently use images of clothing worn by models provided in the DeepFashion dataset.

2.4. Attention Mechanism

In recent years, attention mechanisms have been widely used in various domains of deep learning, such as image classification and object detection. It is generated by borrowing the human visual attention mechanism to quickly screen out high-value information from a large amount of information. Specifically, it quickly scans the global image to obtain the target area that needs to be focused and then devotes more attention resources to this area to obtain more texture information related to the target that needs attention, while suppressing other useless information. In the direction of computer vision, the attention mechanism has many different forms. In terms of the domain of attention, it can be divided into spatial domain, channel domain, layer domain, and mixed domain. And the attention modules we commonly used are spatial transformer network [19], SE-Net [16], Convolutional Block Attention Module[42], Non-local Neural Networks[41], and Multi-Head Attention [37]. The attention mechanism technology has also been researched and applied in the field of fashion. Wang et al.[40] proposes two kinds of attentions, namely category-directed and landmark-aware attentions to enhance clothing category classification. Zhang et al.[45] introduces the attention mechanism to incorporate the impact of positions for clothing attribute prediction with only image-level annotations. Shajini et al.[30] incorporates two attention pipelines: landmark-driven attention and spatial–channel attention for improving the accuracy of clothing classification. Inspired by the success of the attention mechanism, we study deeply about the use of the best combination of the attention mechanism module for fine-grained image classification tasks.

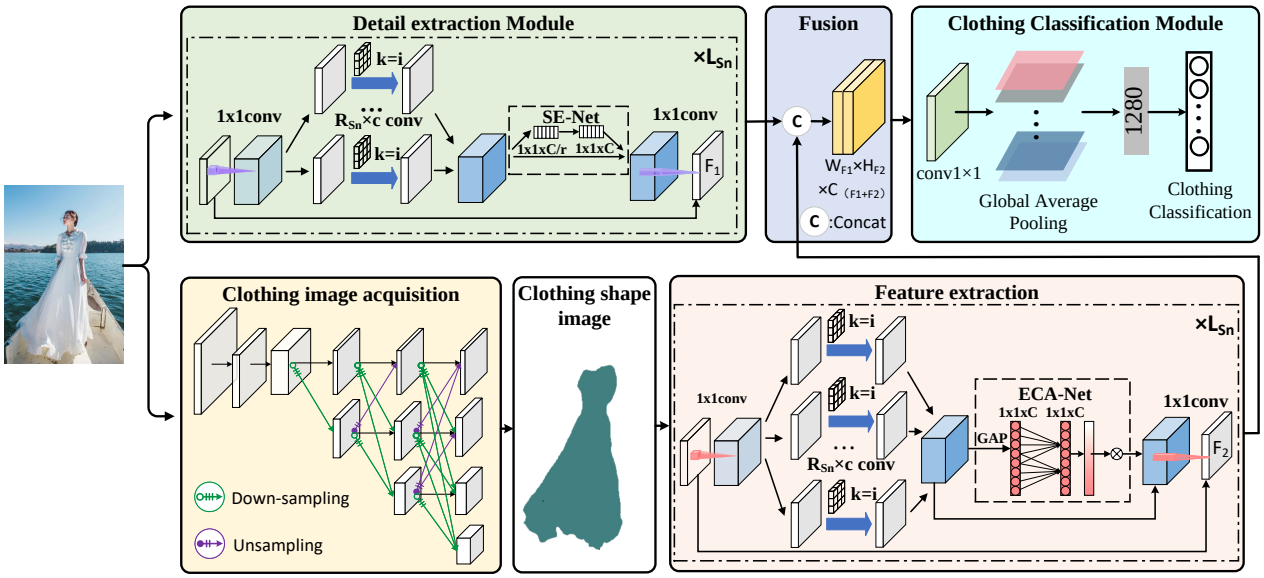

Figure 1. The network structure of TSFFNet. The network obtains texture features F_1 and clothing shape features F_2 of the fashion image through two data streams, and then fuses the obtained F_1 and F_2 features and uses a classifier to obtain classification results. In the figure, S_n means the stage layer of the extraction module, L_{Sn} means the number of stage repetitions, and R_{Sn} means the multiple of the current stage channel expansion.

3. Methodology

The main task of the algorithm in this paper is to quickly and accurately identify the category of the given clothing images. The category of clothing greatly depends on the sleeve length, the fabric, the pattern and so on. We summarize these features and find that clothing classification mainly relies on the shape and the clothing textures. If these corresponding features can be extracted and learned in a better way, they will bring a massive promotion to clothing classification.

In our work, we propose an effective method to extract the textures and shape of the clothing as much as possible. The network we proposed named Two-Stream Clothing Classification Network(texture-shape feature fusion network, TSFFNet),as shown in Figure 1, in which one stream is texture extraction stream, and the other is shape extraction stream. For the shape extraction stream, we use image segmentation technology to help extract the shapes of the fashion images, and then a classification network improved from the fashion dataset is used to further extract features of clothing shapes; for the texture extraction stream, we use a model normalised for width, height and resolution settings, which can help us to maximise extract effective texture features from garments. Then we concatenate the features extracted by the two streams together to classify the clothes categories.

3.1. Shape Extraction Module

For the task of clothing image classification, we guess that the category of clothing largely depends on the shape of the clothing by analyzing the characteristics of the dataset. And we conduct a small experiment to verify our conjecture. Based on the mask(polygon) information about the clothing provided in Deepfashion2[8], we extract the shape images of the clothing as a brand new dataset, and some of the images are shown in Figure 2. We use the same classification network(i.e. EfficientNet-b0) for image classification, and the accuracy of the original image dataset and the shape dataset obtained after processing increased from 53.82% to 84%. By analyzing the experimental results, we find that the shape of the clothing may be very helpful to improve the accuracy of clothing image classification.

Then, we adopt the technique of image segmentation to obtain the clothing shape information in the images. Most existing methods[13, 3] pass the input through a network, typically consisting of high-to-low resolution blocks that are connected in series. The HR-Net[33] present a novel architecture, namely High-Resolution Net (HRNet), which is able to maintain high-resolution representations through the whole process. It consists of parallel high-to-low resolution subnetworks with repeated information exchange across multi-resolution subnetworks(multi-scale fusion). The use of parallel high-to-low resolution subnetworks is able to maintain the high resolution instead of recovering the resolution through a low-to-high process. And the use of re-

Figure 2. The clothing shape image extracted according to the polygon information provided in Deepfashion2.

peated multiscale fusions to boost the high-resolution representations with the help of the low-resolution representations of the same depth and similar level, and vice versa, resulting in that high-resolution representations are also rich. After the segmentation map predicted by the HRNet network, we can get clean clothing images in complex images.

Stage	0perator	Input	Channels	Layers	Stride
1	Conv3 ×3	224 × 224	32	1	2
2	IMBConv1, k3 × 3	112 × 112	16	1	1
3	IMBConv6, k3 × 3	112 × 112	24	2	2
4	IMBConv6, k7 × 7	56 × 56	40	2	2
5	IMBConv6, k3 × 3	28 × 28	80	3	2
6	IMBConv6, k7 × 7	14 × 14	112	3	1
7	IMBConv6, k7 × 7	14 × 14	192	4	2

Table 1. The structure of feature extraction module.

We then perform feature extraction on the obtained pure clothing image to obtain the shape features of the image. The structure of the feature extraction module is shown in Table 1. We use the improved MBConv [35] module, which is modified according to the characteristics of the clothing shape dataset as a stacked module for feature extraction. The network structure of the improved MBConv block is shown in Figure 3(a).

In order to extract richer shape features, we use a 7×7 convolution kernel to increase the receptive field of the model. Compared with the SE-attention mechanism used in EfficientNet, we adopt the ECA-attention mechanism[39]. Unlike SE attention, which first performs channel compression on the input feature map, and such compressed dimensionality reduction has a detrimental effect on learning the dependencies between channels. Based on this concept, the ECA-attention mechanism avoids dimensionality reduction, and uses 1-dimensional(1D) convolution to efficiently realize local cross-channel interaction, which can better extract the dependencies between channels. Specific steps are as follows: 1) a global average pooling operation is performed on the input feature map to obtain a 1×1×C feature map, 2) the obtained Feature map is subjected to a 1D convolution operation with a convolution kernel size of 3, and the weights w of each channel are obtained using the Sigmoid activation function. The formula is as follows:

$$\omega = \sigma\left(\mathbf{C1D}_k(\mathbf{y})\right) \tag{1}$$

where C1D indicates 1D convolution, k represents the size of the convolution kernel, which we set to 3 here, y means the given aggregated feature without dimensionality reduction, and 3) multiply the weights with the corresponding elements of the original input feature map to get the final output feature map. The network structure of ECA-attention is shown in Figure 3(b).

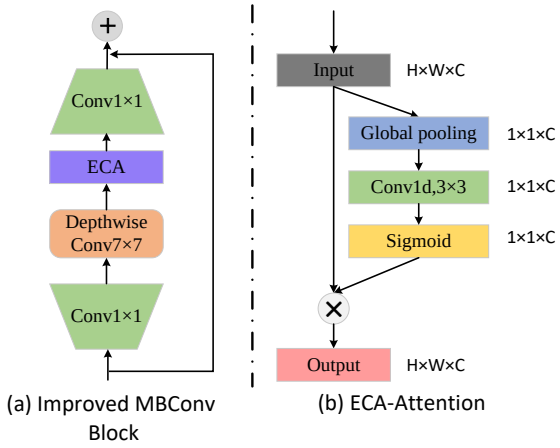

Figure 3. The network structure of (a) improved MBConv block and (b) ECA-attention mechanism.

By removing redundant convolutional layers, using large convolutional kernels and ECA-attention, we can extract the shape information of clothing quickly and accurately.

3.2. Texture Extraction module

Clothing types are diverse and varied, but the similarities that garments maintain in certain textural characteristics make them easily grouped together. The classification accuracy of some attributes (such as stripe, print, graphic) of the clothing is also highly dependent on the understanding of the texture features[46]. Robert Geirhos et al.[9] find that convolutional neural networks prefer to use color and texture for prediction rather than shape. Thus, We use a network with standardised settings for width, height and resolution to extract texture features directly from the input image. This structure facilitates our overall joint scaling network to obtain the optimal extraction structure. Also, the direct extraction of the original image can avoid the

errors generated by our clothing acquisition module when performing image segmentation.

We compare the existing mainstream classification networks to find the best network for fashion image classification as the texture extraction network through experiments. According to the experimental results, it can be found that the network that achieves the best results on other datasets does not work well for clothing feature extraction. Finally, we identify EfficientNet-b0 as the extraction network in the texture extraction module.

3.3. Network Structure

We have tried two methods to integrate the shape extraction stream and texture extraction stream. The first approach is a class score fusion[32]. This is a weighted fusion of the scores obtained by the softmax layers in the two extraction networks. The second approach is to perform the corresponding feature extraction for different input datasets (shape dataset and original dataset) separately. It is to concatenate the feature maps of the two streams together before entering the classification module. We have tried these two methods separately, and find that the second method is better. This is also in line with our assumption: the ability to specifically enhance the neural network's extraction and learning of texture features and shape features, respectively, can help to improve the accuracy of clothing recognition.

To validate our conception, we propose TSFFNet, whose network structure is shown in Figure 1. We adopt the idea of two-stream networks commonly used in action recognition to extract the texture and shape features of clothing separately. For the clothing texture extraction module, we compare various networks, and finally select EfficientNet as the feature extraction network. Then for shape feature extraction, we first use HRNet to segment the image with complex background to get clean clothing shape images. Then we improve the existing network according to the characteristics of the clothing dataset to achieve lighter and more accurate extraction of clothing characteristics, the results of the experiment are shown in Table 4. Finally we fuse the clothing features with the texture features of the fashion images and then use the classifier to classify them accurately.

4. Experiments and Discussion

The experimental platform system is Linux, the graphics processing unit (GPU) is Tesla V100, which relies on python3.7, numpy1.18.5, torchvision0.4.0, and pytorch1.7.1. The size of the image input to the network is 224×224, the batch-size is 32, the model epoch is 100.

4.1. Cloth Category Dataset

We evaluate the performance of the proposed model using three large-scale benchmark fashion clothes datasets: DeepFashion2 dataset[8], Deepfashion1[24] and Clothing1M[43] dateset.

4.1.1 DeepFashion2 Dateset

DeepFashion2[8], is a large-scale benchmark with comprehensive tasks and annotations of fashion image understanding. DeepFashion2 contains 491K images of 13 popular clothing categories. A full spectrum of tasks are defined on them including clothes detection and recognition, landmark and pose estimation, segmentation, as well as verification and retrieval. All these tasks are supported by rich annotations. For instance, DeepFashion2 totally has 801K clothing items, where each item in an image is labeled with scale,occlusion, zooming, viewpoint, bounding box, dense landmarkss (e.g. 39 for 'long sleeve outwear' and 15 for 'vest'), and per-pixel mask.

Our aim is to achieve accurate classification of flat clothing images of seller shows in online stores. Since there are multiple categories in a picture in Deepfashion2 dataset, it is not suitable for us to do a single clothing classification. Thus, we use code to select images that are slightly obscured and contain only one category based on the category number and clothing occlusion information provided by DeepFashion2 dataset. As some of the categories in the dataset have few images, such as, "sling", "skirt" and "short sleeve outwear". In order to keep the numbers even under each category, we further randomly filter from the obtained 25132 images to obtain a small dataset with 10 categories and a total of 5500 images,of which 5000 are used as the training set and 500 as the validation set. Also, we filter out the categories with less than 500 images.

4.1.2 Clothing1M Dateset

Clothing1M[43] contains 1M clothing images in 14 classes. It is a dataset with noisy labels, since the data is collected from several online shopping websites and include many mislabelled samples. This dataset also contains 47570, 14313, and 10526 images with clean labels for training, validation, and testing, respectively. In this article, we will mainly use the clean version for experimental verification.

4.1.3 DeepFashion1 Dateset

DeepFashion1[24] dataset is used to evaluate the proposed category classification. It contains 289,222 annotated fashion clothes images. Each image is labelled with 46 clothing categories.

4.2. Performance Evaluation

For clothing category classification, we apply top-1 classification accuracy, precision rate and F1-score

to evaluate our model. In order to ensure the fairness of the results, all trainings are re-trained on the same experimental platform and the resolution of input image is the same. By comparing classical networks such as GoogleNet[34], ResNet[14], DenseNet[18], and mainstream image classification networks, such as MobileNetv2[29] EfficientNet[35], EfficientNetv2[36], ESPNetv2[28], ConvNext[25]. All the experimental results are summarised in Table 2.

Model	Accuracy	Precision	F1-score
GoogleNet[34]	0.522	0.538	0.526
DenseNet201[18]	0.420	0.436	0.428
ESPNetv2[28]	0.542	0.543	0.542
MobileNetv2[29]	0.518	0.522	0.519
Resnet34[14]	0.576	0.575	0.575
ResNet50[14]	0.524	0.564	0.544
EfficientNet-b0[35]	0.622	0.621	0.621
EfficientNet-b2[35]	0.610	0.611	0.610
EfficientNetv2[36]	0.602	0.604	0.603
ConvNext-tiny[25]	0.504	0.512	0.508
ConvNext-base[25]	0.496	0.502	0.499
Our Method (TSFFNet)	**0.746**	**0.755**	**0.750**

Table 2. Experimental results for clothing category classification on validation set using top-1 accuracy.

As shown in Table 2, the shape and texture features of the garment images are extracted separately, which can improve the classification accuracy very substantially than using a single network. Our TSFFNet model achieves 74.6%, 75.5%, and 75.0% for top-1 accuracy, precision and F1-score, respectively. The TSFFNet model outperforms state-of-the-art methods in clothing category classification.

We show the results of the classification of some types of clothing in Figure 4. Accurate segmentation can help us remove invalid background information and obtain the shape information of the clothing. From the shape of the clothing in the figure, the classification accuracy of the categories with distinct characteristics will be higher. However, relying only on the shape information of the clothing can also produce errors on the classification results, such as similarity of some categories on some parts caused by different poses and occlusions. For example, there will be some similarities between long sleeve tops and short sleeve tops in the hands-around-the-clasp position. At such times, the original image may provide more effective feature information for enhancing the accuracy of clothing classification.

4.3. Clothing Segmentation

As our network is designed according to the shape and texture characteristics of the clothing, the method of clothing shape acquisition plays an important role. We conjecture whether the accuracy of segmentation would further affect the accuracy of the network. So we select several classical segmentation networks for experimental validation. The MIoU and MPA of the segmentation network, the accuracy of the segmented clothing image dataset and the accuracy of the fusion of the extracted original image features and the segmented image features are shown in Table 3.

Model	MIoU	MPA	Shape-ACC	TSFFNet
Deeplab V3+[49]	47.76	56.64	0.668	0.704
HRNet-w18[37]	24.58	32.66	0.632	0.638
HRNet-w32[37]	46.49	54.39	**0.724**	**0.732**
Unet-resnet50[50]	42.97	52.20	0.720	0.726
Unet-vgg[50]	32.09	42.39	0.658	0.672
PspNet[51]	45.58	55.34	0.658	0.690

Table 3. The MIoU and MPA of the segmentation network, the accuracy of the segmented clothing image dataset and the accuracy of the fusion of the original and the segmented images.

Table 3 shows the relationship between the performance of segmentation and the classification accuracy of the segmented image, and it can be seen that it is not the higher the MIoU or MPA of segmentation, the better the classification. By comparing the garment images obtained by Deeplab V3+ and HRNet-w32 segmentation, Deeplab V3+ has a good segmentation effect, but the classification effect on the segmented map is not as good as that of HRNet-w32. But the higher the classification accuracy of the shape features obtained after segmentation, the higher the classification accuracy of the fusion of shape and texture features.

We select HRNet-w32 as the network for the clothing acquisition module based on the final classification results in the obtained images, and obtain an image of pure garment shape on the right in each cell in Figure 4. Then based on this, we fuse the texture features of the original image and the shape features obtained from clothing shape image. And we enhance the classification accuracy of our network by enhancing the extraction ability of these two features separately. We find the extraction network that best suits these two features through comparative experiments.

4.4. Extraction of Clothing Shape Feature

As shown in Table 3, the higher the classification accuracy of the shape features obtained after segmentation, the higher the classification accuracy of the fusion of shape and texture features. We use several different classical networks to re-run experiments on the segmented obtained garment shape image dataset. The experimental results are shown in Table 4.

According to the experimental results, we observe that the ConvNext, which achieve the best effect in the ImageNet dataset, achieve poor results in the clothing classification problem. We speculate that the optimal solution of this model is difficult to obtain on our hardware because there

Figure 4. The results of TSFFNet. The image column shows the original image of the input and the obtained image of the clothing shape. The probability displays the accuracy of our network for classifying these images.

Model	Accuracy
GoogleNet[34]	0.700
DenseNet201[18]	0.670
ESPNetv2[28]	0.712
MobileNetv2[29]	0.692
Resnet34[14]	0.720
ResNet50[14]	0.702
EfficientNet-b0[35]	0.724
EfficientNet-b2[35]	0.722
EfficientNetv2[36]	0.716
ConvNext-tiny[25]	0.692
ConvNext-base[25]	0.694
Our Method (TSFFNet)	**0.746**

Table 4. Quantitative results for clothing category classification on segmented clothing shapes.

is no transfer learning. Another possible reason is that the accuracy may be reduced due to overfitting caused by the complexity of the network. At the same time, comparing the deeper structure of the same network, it can be seen that ResNet50 is not as good as ResNet34, and EfficientNet-b2 is not as good as EfficientNet-b0. We conjecture that perhaps a deeper network does not help to improve clothing classification accuracy. Therefore, we choose lightweight and accurate EfficientNet-b0 as the texture and shape feature extraction network, and we also adapt and improve the EfficientNet-b0 according to the characteristics of the apparel dataset.

Combining Table 2 and Table 4 show that our method (i.e., using both the texture features of the image and the shape features of the clothing) works better than extracting one type of feature alone.

To further illustrate that the fusion of the shape of the clothing and the texture features of the image will play a better role, we also conducted a comparison experiment. We segment the clothing images, keeping the clothing images with texture features(i.e. removing the redundant body parts as well as the background from the images). In this way, the resulting clothing dataset has both texture and shape information of the clothing without the interference of redundant information. Then, we select the classical networks, such as GoogleNet[34], MobileNetv2[29], DenseNet201[18], ResNet34[14], EfficientNet-b0[35] and ConvNext-tiny[25], to conduct experiments on the processed pure clothing dataset with texture features.

Model	Accuracy
GoogleNet[34]	0.684
MobileNetv2[29]	0.69
DenseNet201[18]	0.626
ResNet34[14]	0.694
ConvNext-tiny[25]	0.646
EfficientNet-b0[35]	0.700

Table 5. Classification experiments on clothing images with texture features.

From Table 5, we can see that the classification results of using the segmented clothing image with textures is not as good as if we have extracted both features separately. We conjecture that the result of segmentation also affects the classification accuracy. So we use two feature extraction networks to learn two types of features of the image sepa-

rately. First, the shape features of the clothing are learned more intently using the shape extraction module, and the features are extracted directly from the original image by the texture extraction module on the other side. Then, feature fusion of the features learned from these two separately can retain the important features of the original image in a greater extent.

4.5. Improvement by Shape Extraction Module

We chose EfficientNet-b0 as the network for texture feature extraction and shape feature extraction. We improve the feature extraction capability of the clothing shape extraction stream according to the characteristics of the clothing shape. We are improving the network in following three ways:

The number of the network's stages. Based on the data in Table 4, we can see that deeper networks do not work well, and we try to further reduce the number of stages of the network to improve the accuracy of the network.

The receptive field of the network. Based on the characteristics of the clothing shape, we guess whether increasing the receptive field of the network will also improve the accuracy of the network. So, we try two methods, one is to change the original convolutional kernels in the EfficientNet-b0 network from 5×5 to 7×7, and the other is to change the depth-separable convolutional order according to the settings in the ConvNext network and then uses 7×7 convolutional kernels.

Attentional Mechanisms. The dimensionality reduction method used in SE-attention is not conducive to the weight learning of channel attention in feature maps. So we use different attention mechanism methods to try to find out a more suitable attention mechanism for clothing shape feature learning.

Shape Extraction	Acc	Model Size
EfficientNet-b0	0.724	15.63M
Reduce one stage	0.728	12.24M
Reduce two stage	0.718	4.02M
k7x7; change order	0.698	15.11M
k7x7	0.720	16.21M
k7x7; reduce one stage; change order	0.700	11.74M
k7x7; reduce one stage	0.730	12.82M
Replace SE with CBAM	0.716	41.99M
Replace SE with ECA	0.726	13.19M
Reduce one stage; replace SE with ECA	0.730	10.22M
k7x7; reduce one stage; replace SE with ECA (our method)	**0.732**	10.80M

Table 6. The Ablation Experiment.

The ablation experiments performed according to our conjectures and improvements are shown in Table 6. According to the experimental results, it is found that by changing the number of layers of the network, the size of the convolution kernel, and the attention mechanism, a network with fewer parameters and more accurate accuracy can be realized.

Parts	Composition				
EfficientNet-b0	√	√	√	√	√
Reduce one stage		√	√	√	√
Replace SE with ECA			√		√
k7×7				√	√
Accuracy	0.732	0.738	0.740	0.740	**0.746**

Table 7. The variation of accuracy after the fusion of our improved shape extraction network and texture extraction network.

It can be seen from Table 6 that using the 7×7 convolution kernel directly will be better than using the 7×7 convolution kernel after changing the order according to ConvNext, but it is not as good as the original effect of EfficientNet. When the number of stages of EfficientNet reaches the optimum, using a 7×7 convolution kernel will increase the accuracy by 0.6%. The use of ECA-attention can bring a little improvement whether it is on the original EfficiNet or the adjusted network. Finally, the accuracy of our shape extraction network increase from the original 72.4% to 73.2% after three improvements are made.

We fuse the features extracted from the improved shape extraction network and the texture extraction network to form our TSFFNet. Table 7 shows the variation of accuracy after the fusion of our improved shape extraction network and texture extraction network.

As can be seen from Table 7, improving the accuracy of the shape extraction module can also further improve the accuracy of the network for fashion image classification. With the improvements we have made, the accuracy of our TSFFNet has reached 74.6% from the initial 73.2%, an improvement of 1.4%.

4.6. Results on other Dataset

Since the current fashion dataset has little mark information, although ModaNet[47] has street images with masks of single person, the data information is too complex for our dataset to do classification, so we still use the relevant information provided in Deepfashion2 for the training of clothing acquisition.

To test the effectiveness of our model, the trained segmentation weights are applied directly on the Clothing1M and Deepfashion1[24] dataset, and then the features of the obtained clothing images and fashion images are fused to perform classification validation. The results of our model with other mainstream classification networks are shown in Table 8.

From the results in Table 8, we can see that although our clothing image acquisition module only uses 5500 images

Model	Accuracy	
	Clothing1M	Deepfashion1
GoogLeNet	0.66	0.605
DenseNet201	0.643	0.641
ESPNetv2	0.679	0.658
MobileNetv2	0.689	0.662
Resnet34	0.719	0.679
EfficientNet-b0	0.721	0.675
ConvNext-tiny	0.697	0.660
Our Method (TSFFNet)	**0.732**	**0.689**

Table 8. The Results of our model on other Dataset.

for network training, it can also achieve better results than other mainstream networks in the Clothing1M and Deepfashion1 dataset.

However, since Clothing1M and Deepfashion1 does not provide mask information and most of the images in the dataset contain multiple garments, the single clothing style classification we trained did not achieve a high improvement. At the same time, as there is no open source code for recent papers on clothing classification and the available classification data are based on the Deepfashion1 dataset, we are unable to do comparative experiments with existing clothing classification networks on the same dataset for the time being. Therefore, we chose to use the mainstream network for our comparison experiments.

5. Conclusion

In this paper, we point out that starting from the properties of the clothing itself can better help classify the clothing. Take the clothing style classification task as an example, most clothing categories can be distinguished from the shape, such as tops, bottoms, skirts, etc. And we find that compared to extract the clothing features together, targeted enhancements are more helpful in improving the accuracy of clothing classification. In particular, after our experimental verification, the accuracy of clothing classification is more dependent on the shape characteristics of clothing. Therefore, we propose a network that can individually enhance shape and texture features: a texture-shape feature fusion network(TSFFNet). At the same time, we have further improved the network according to the characteristics of the clothing style dataset, so that the network can classify clothing images lighter and more accurately. The experimental results demonstrate that our TSFFNet improves the feature representation of clothes images and achieves comparable performance to the other methods. After the design of the network, the experiment verifies that our network can achieve 74.6% accuracy on the fashion dataset, which gains 12.4% improvement over using the best mainstream classification network alone.

Acknowledgement

This work was supported by national natural science foundation of China (No. 62202346), Hubei key research and development program (No.2021BAA042), open project of engineering research center of Hubei province for clothing information (No. 2022HBCI01), Wuhan applied basic frontier research project (No. 2022013988065212), MIIT's AI Industry Innovation Task unveils flagship projects (Key technologies, equipment, and systems for flexible customized and intelligent manufacturing in the clothing industry), and Hubei science and technology project of safe production special fund (Scene control platform based on proprioception information computing of artificial intelligence).

References

[1] Z. Al-Halah, R. Stiefelhagen, and K. Grauman. Fashion forward: Forecasting visual style in fashion. In *Proceedings of the IEEE international conference on computer vision*, pages 388–397, 2017. 1, 3

[2] J. Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 3

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 4

[4] H. Cho, C. Ahn, K. Min Yoo, J. Seol, and S.-g. Lee. Leveraging class hierarchy in fashion classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 2, 3

[5] C. Corbiere, H. Ben-Younes, A. Ramé, and C. Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 2268–2274, 2017. 1, 3

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 3

[7] L. De Divitiis, F. Becattini, C. Baecchi, and A. D. Bimbo. Disentangling features for fashion recommendation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2022. 1, 3

[8] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5337–5345, 2019. 2, 4, 6

[9] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 5

[10] S. Goenka, Z. Zheng, A. Jaiswal, R. Chada, Y. Wu, V. Hedau, and P. Natarajan. Fashionvlp: Vision language transformer

for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14105–14115, 2022. 1, 3

[11] K. Grauman. Computer vision for fashion: From individual recommendations to world-wide trends. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 3–3, 2020. 1, 3

[12] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE international conference on computer vision*, pages 3343–3351, 2015. 2

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7, 8

[15] Y. Hou, E. Vig, M. Donoser, and L. Bazzani. Learning attribute-driven disentangled representations for interactive fashion retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12147–12157, 2021. 1, 3

[16] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[17] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE international conference on computer vision*, pages 1062–1070, 2015. 2

[18] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014. 7, 8

[19] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015. 3

[20] B. Kolisnik, I. Hogan, and F. Zulkernine. Condition-cnn: A hierarchical multi-label fashion image classification model. *Expert Systems with Applications*, 182:115195, 2021. 1, 2

[21] Z. Kuang, Y. Gao, G. Li, P. Luo, Y. Chen, L. Lin, and W. Zhang. Fashion retrieval via graph reasoning networks on a similarity pyramid. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3066–3075, 2019. 1, 3

[22] P. Li, Y. Li, X. Jiang, and X. Zhen. Two-stream multi-task network for fashion recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3038–3042. IEEE, 2019. 1, 3

[23] L. Liao, X. He, B. Zhao, C.-W. Ngo, and T.-S. Chua. Interpretable multimodal retrieval for fashion products. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1571–1579, 2018. 1, 3

[24] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 1, 2, 3, 6, 9

[25] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 7, 8

[26] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 3

[27] U. Mall, K. Matzen, B. Hariharan, N. Snavely, and K. Bala. Geostyle: Discovering fashion trends and events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 411–420, 2019. 1, 3

[28] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9190–9200, 2019. 7, 8

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 7, 8

[30] M. Shajini and A. Ramanan. An improved landmark-driven and spatial–channel attentive convolutional neural network for fashion clothes classification. *The Visual Computer*, 37(6):1517–1526, 2021. 1, 2, 3

[31] Z. Shen. Mining sustainable fashion e-commerce: social media texts and consumer behaviors. *Electronic Commerce Research*, pages 1–23, 2021. 1

[32] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 6

[33] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 4

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 7, 8

[35] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 5, 7, 8

[36] M. Tan and Q. V. Le. Efficientnetv2: Smaller models and faster training. *arXiv preprint arXiv:2104.00298*, 2021. 7, 8

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[38] Y. Wan, C. Yan, B. Zhang, and G. Zou. Learning image representation via attribute-aware attention networks for fashion classification. In *International Conference on Multimedia Modeling*, pages 69–81. Springer, 2022. 1

[39] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu. Supplementary material for 'eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and*

Pattern Recognition, IEEE, Seattle, WA, USA, pages 13–19, 2020. 5

[40] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4271–4280, 2018. 1, 2, 3

[41] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3

[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3

[43] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015. 2, 6

[44] W. Yang, P. Luo, and L. Lin. Clothing co-parsing by joint image segmentation and labeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3182–3189, 2014. 2

[45] S. Zhang, Z. Song, X. Cao, H. Zhang, and J. Zhou. Task-aware attention model for clothing attribute prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1051–1064, 2019. 3

[46] Y. Zhang, P. Zhang, C. Yuan, and Z. Wang. Texture and shape biased two-stream networks for clothing classification and attribute recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13538–13547, 2020. 1, 2, 3, 5

[47] S. Zheng, F. Yang, M. H. Kiapour, and R. Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1670–1678, 2018. 2, 9

[48] X. Zhu and M. Bain. B-cnn: branch convolutional neural network for hierarchical classification. *arXiv preprint arXiv:1709.09890*, 2017. 1, 2