applied sciences

MDPI

Article

EnCaps: Clothing Image Classification Based on Enhanced Capsule Network

Feng Yu [1,2], Chenghu Du [1], Ailing Hua [1], Minghua Jiang [1,2,*], Xiong Wei [1], Tao Peng [1,2] and Xinrong Hu [1,2]

[1] School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China; yufeng@wtu.edu.cn (F.Y.); duceh_lzy@163.com (C.D.); hal_wtu@163.com (A.H.); wx_wh@wtu.edu.cn (X.W.); pt@wtu.edu.cn (T.P.); hxr@wtu.edu.cn (X.H.)

[2] Engineering Research Center of Hubei Province for Clothing Information, Wuhan 430200, China

[*] Correspondence: minghuajiang@wtu.edu.cn

Abstract: Clothing image classification is more and more important in the development of online clothing shopping. The clothing category marking, clothing commodity retrieval, and similar clothing recommendations are the popular applications in current clothing shopping, which are based on the technology of accurate clothing image classification. Wide varieties and various styles of clothing lead to great difficulty for the accurate clothing image classification. The traditional neural network can not obtain the spatial structure information of clothing images, which leads to poor classification accuracy. In order to reach the high accuracy, the enhanced capsule (EnCaps) network is proposed with the image feature and spatial structure feature. First, the spatial structure extraction model is proposed to obtain the clothing structure feature based on the EnCaps network. Second, the enhanced feature extraction model is proposed to extract more robust clothing features based on deeper network structure and attention mechanism. Third, parameter optimization is used to reduce the computation in the proposed network based on inception mechanism. Experimental results indicate that the proposed EnCaps network achieves high performance in terms of classification accuracy and computational efficiency.

Keywords: clothing image classification; enhanced capsule network; spatial structure feature; attention mechanism; inception mechanism

check for updates

Citation: Yu, F.; Du, C.; Hua, A.; Jiang, M.; Wei, X.; Peng, T.; Hu, X. EnCaps: Clothing Image Classification Based on Enhanced Capsule Network. *Appl. Sci.* **2021**, *11*, 11024. https://doi.org/10.3390/app112211024

Received: 25 October 2021
Accepted: 18 November 2021
Published: 21 November 2021

1. Introduction

With the development of electronic commerce, internet shopping for clothing has become a common lifestyle [1–4]. Before the clothing information is uploaded to the online shopping mall, the category, texture, style, fabric, and shape of clothing should be labeled. The purchaser searches for suitable clothing by keyword retrieval. The manual label method may be very costly on a human level, and the correct labeling of clothing is based on personal judgment. The mistake of personal judgment is inevitable in the thousands of clothing updates. Furthermore, it is difficult to distinguish the fine-grained classification of clothing by personal judgment. Thus, the high-efficiency method of the clothing classification [5,6] is urgent in the rapid development of clothing shopping.

Clothing classification attracts a lot of attention in academic circles. The classification methods for clothing are usually divided into two categories. First, the traditional feature extraction methods for clothing classification can be also divided into two types, one is based on global shape features and global texture features [7], such as Fourier descriptors, geometric invariant distance, local binary patterns (LBP), etc., and another one is based on local feature methods that include scale-invariant feature transformation (SIFT), sped up robust features (SURF), histogram of oriented gradient (HOG), etc. [8–11]. The classification accuracy of traditional methods rely on the selective feature severely. In specific cases, the methods may achieve high-level accuracy with stable and conspicuous features. Generally, the clothing image is various, similar, and complex. The traditional methods can not extract robust features for classification. Second, with the technological development of convolutional neural network (CNN), CNN is widely used in clothing classification [12–14]. The method achieves better performance than the traditional methods. It extracts clothing features based on the deep network without manual setting, which can obtain robust clothing features [15,16]. The convolution network and polling operations are used to extract the clothing feature, and deeper networks may achieve better performance in general. However, the CNN only extracts the image features without a spatial relationship between different local features [17–20]. The traditional CNN can not breakthrough the bottleneck of classification accuracy.

Capsule network [21–24] is a novel CNN which is proposed to obtain the spatial relationship of different features, which extract the spatial relationship by the dynamic routing algorithm. The fundamental feature unit is expressed by the capsule, which is a vector. However, the traditional capsule network [21] is usually used in the MNIST database for handwritten digits recognition, and the size of input is defined as 28 × 28. With the advantage of the capsule network in spatial features, Ref. [25] proposes a novel multi-scale capsule network to extract the multi-scale feature. Ref. [26] proposes a subspace capsule network to exploit the idea of capsule networks to model possible variations in the appearance or implicitly-defined properties of an entity through a group of capsule subspaces instead of simply grouping neurons to create capsules. Ref. [27] proposes a multi-lane capsule networks that is a separable and resource efficient organization of capsule networks that allows parallel processing while achieving high accuracy at reduced cost. In a word, the traditional capsule network and improved capsule networks can not efficiently extract robust clothing feature, which are not suitable for the accurate classification of clothing images.

In order to further improve the accuracy of clothing image classification, we propose an enhanced feature capsule (EnCaps) network. The traditional convolutional neural network can not obtain the spatial location relationship. The capsule network is used to extract the spatial location relationship of different types of clothing. The original capsule network only has two convolution layers to extract the image feature. To extract the robust feature of clothing, we improve the original network and adopt the attention mechanism and deeper network in the proposed EnCaps network. Moreover, the inception mechanism is also fused in the EnCaps network to reduce the computation. The experimental results indicate that the proposed EnCaps network can achieve more accuracy and less computation.

The remainder of this paper is organized as follows: Section 2 introduces the methods and methodology of the proposed EnCaps network. Section 3 gives the experiments and results to indicate the validity of the proposed algorithm. The importance of proposed algorithm and future direction are discussed in Section 4. The conclusions are given in Section 5.

2. Materials and Methods

There are three key issues should be considered in the EnCaps network. First, the input image size of the traditional capsule network [21] is only 28 × 28, and we should increase the input size by improving the network structure to process more high-quality images—second, how to extract robust feature with efficient network structure; and, third, the vector indicates the capsule unit, but the processing unit of traditional CNN is pixels, which results in a more complex calculated amount. The parameter optimization strategy should be used in the proposed network. In order to solve the above issues, the EnCaps network is proposed with three novel strategies, and the method architecture is shown in Figure 1.

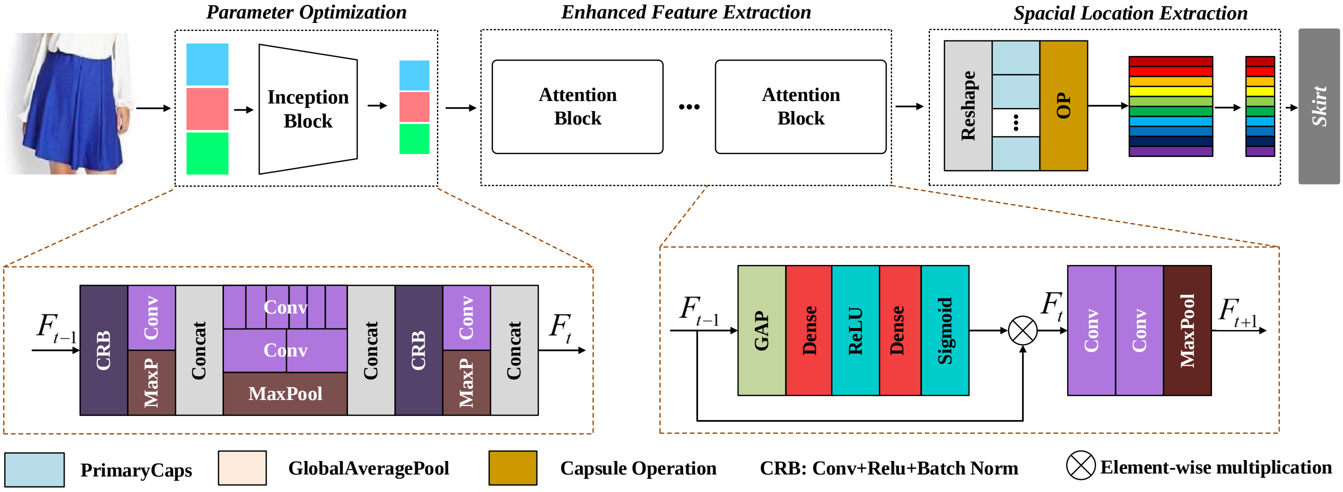


Figure 1. Overview of the proposed EnCaps. The network architecture consists of three main parts: (1) Parameter Optimization; (2) spatial structure Extraction; and (3) spatial structure Extraction.

2.1. Spatial Structure Extraction

Convolution is locally connected and parameter sharing, and as the layers of a convolutional network deepen, the network can learn more global contextual information and then use this information to make predictions. However, there is no available spatial information in the extracted features, which is one of the reasons for the prediction failure. First, the shape information of the extracted feature is important for object identification. The different types of clothing image have obvious structural features that can be used to classify. There are about seven common types of clothing profiles, such as 'A', 'H', 'X', 'T', 'Y', 'O', and 'V'. The feature of the clothing profile should enhance the classification accuracy. Thus, the spatial shape information is important for clothing classification. However, it is not enough to have profile information. The traditional convolutional neural network does not have the ability to analyze spatial information, which makes it difficult to distinguish between the two types of clothing with only slight differences. For example, skirt and dress, which are also in the shape of 'A', are extremely easy to be confused by the network due to no spatial information (localized spatial alignment information for clothing and body) being identified.

To complement the capability of spatial feature extraction, the capsule network [21] is introduced to process the features further. The capsule network extracts the structural features of objectives based on the capsule unit. The traditional capsule network is shown in Figure 2. The fundamental structure of the capsule network includes convolution layer, initial capsule layer, convolution capsule layer, and fully connected layer. Different from the traditional CNN, the feature vector v_i of objective replaces the scalar feature. The "prediction vectors" $\hat{u}_{j|i}$ from the capsules is obtained by Equation (1):

$$\hat{u}_{j|i} = W_{ij} u_i \tag{1}$$

where W_{ij} is the weight matrix of a certain capsule layer. i denotes the vector index of input capsules layer, and $j|i$ denotes the capsules index of PrimaryCapsules j corresponding to vector i.

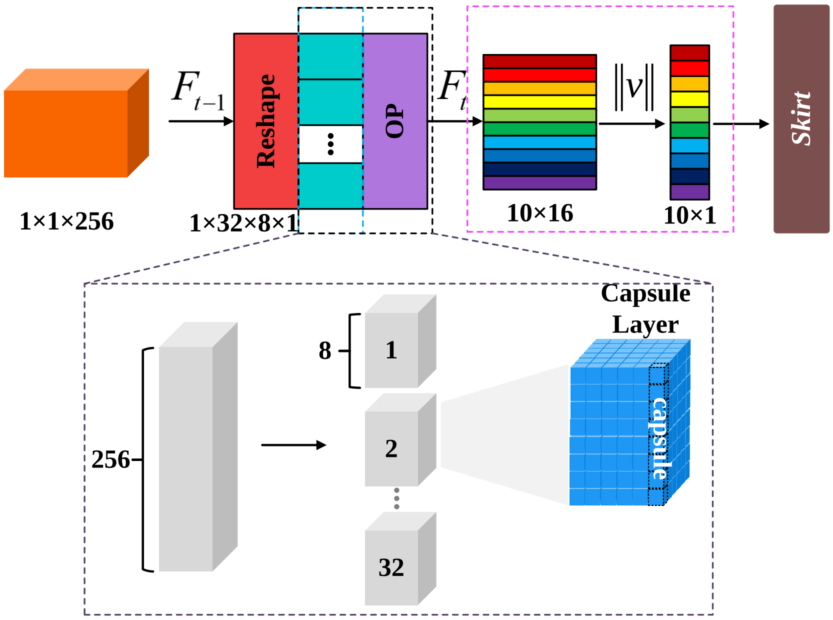


Figure 2. Overview of the traditional capsule network, where F_{t-1} is input feature, and F_t is output vector. It is used for spatial structure extraction.

Then, the high level feature $\hat{u}_{j|i}$ is processed by Equation (2) to realize the dynamic routing of features:

$$s_j = \sum_i c_{ij}\hat{u}_{j|i} \tag{2}$$

where the parameter c_{ij} indicates the routing probability from capsule i of L layer to capsule j of $L+1$. The computational formula of routing probability is represented by Equation (3):

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_j \exp(b_{ij})} \tag{3}$$

where b_{ij} is the prior probability from capsule i to capsule j, which is iteratively updated in the model training, and the initial value is 0.

Then, the parameter s_j is processed by the squashing function (Equation (4)) to obtain the $L+1$ layer capsule:

$$v_j = \frac{||s_j||^2}{1+||s_j||^2}\frac{s_j}{||s_j||} \tag{4}$$

Since capsule network allows multiple classifications to co-exist, the traditional cross-entropy loss can not be used directly; an alternative is the margin loss commonly used in SVM. The capsule mechanism is used at the end of EnCaps, and the use of margin loss is also a way to maintain the performance of the model. It can be expressed as Equation (5):

$$L_c = T_c max\left(0, m^+ - ||v_c||\right)^2 + \lambda(1-T_c)max\left(0, ||v_c|| - m^-\right)^2 \tag{5}$$

where c is a certain classification, T_c is the indicator for the classification ('1' indicates the presence of class c, '0' indicates the absence of class c), and m^+ is the upper bound that is predicting the existence of class c but not its true existence. m^- is the lower bound that is predicting that class c does not exist but does exist, and λ is the scale factor. If class c is existing, $||v_c||$ will not be less than 0.9, if class c does not exist, $||v_c||$ will not be greater than 0.1.

The input limitation of traditional capsule network is hardly used in clothing classification because the size of input image is only 28 × 28, and the input limitation of image size restrains the wide application of the capsule network. Thus, the image size of input network is improved in the EnCaps network for larger image size. The larger image is beneficial to obtain more feature information, which is useful for more accurate classification. In our proposed network, the size of 224 × 224 image is used, and the input size has increased by 64 times.

2.2. Enhanced Feature Extraction

The original capsule network only has two convolution layers, which can not extract the robust feature of objectives. In order to extract a more robust image feature, the enhanced feature extraction model is proposed with a deeper convolution network as shown in Figure 3.

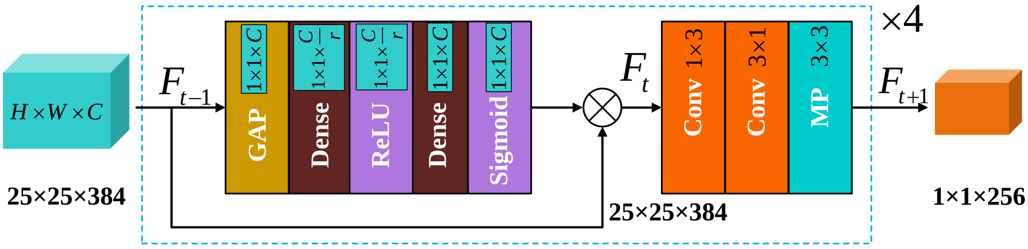


Figure 3. Overview of the enhanced feature extraction model, where F_{t-1} is the input feature, F_t is the middle feature, and F_{t+1} is the output feature of the enhanced feature extraction model.

In the proposed model, the deeper network structure and attention mechanism are used to extract robust features. The extracted 25 × 25 × 384 dimensional high-level feature map is extracted with a channel attention mechanism which ignores the irrelevant information and focuses on the key information in the image. The enhanced operation can be defined as:

$$F_t = F_{t-1} \cdot \sigma\Big(W_2 \delta\Big(W_1\big(\frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{t-1}(i,j)\big)\Big)\Big) \tag{6}$$

$$F_{t+1} = W_3^{1\times3}\Big(W_3^{3\times1}\big(M_2(F_t)\big)\Big) \tag{7}$$

where H denotes the height of the feature map, W denotes the width of the feature map, $W_.$ is the convolution operation, σ and δ are different activation functions, and $M_.$ is the max-pooling operation.

After a series of attentional enhancement operations, we reduce the size of the feature map from 25 × 25 × 384 to a one-dimensional vector of 1 × 1 × 256, and the detailed structure of stem module is shown in Table 1.

2.3. Parameter Optimization

In the early stage of feature extraction, the model is asked to use a more lightweight decoding network to extract more advanced features. In order to extract robust features with low computational cost, we use 1 × 3 and 3 × 1 convolution kernels to replace the 3 × 3 convolution kernel. It allows for a significant reduction in the size of the model without losing any information.

We refer to the stem module in the Inception network and improve it with three objectives: (1) to down-sample the images massively in the stem phase to reduce the aspect ratio, (2) to reduce the number of parameters in the feature map to make them more semantically informative, and (3) to reduce the size of the overall framework of stem to make the model more lightweight.

Table 1. The network structure of enhanced feature extraction model. 'k' represents 'kernel', 's' represents 'stride', and 'p' represents 'padding'.

Layer	EnCaps-Reduce	Shape
Input	/	N×384×25×25
Conv_5+ReLU+BN	SE_layer	N×384×25×25
	Conv-384(k:1,3;s:1,1;p:0,0)	N×384×25×25
	Conv-384(k:3,1;s:1,1;p:0,0)	N×384×25×25
	Max(k:2,2;s:2,2;p:1,1)	N×384×12×12
Conv_6+ReLU+BN	SE_layer	N×384×12×12
	Conv-512(k:1,3;s:1,1;p:1,1)	N×512×12×12
	Conv-512(k:3,1;s:1,1;p:1,1)	N×512×12×12
	Max(k:2,2;s:2,2;p:1,1)	N×512×6×6
Conv_7+ReLU+BN	SE_layer	N×512×6×6
	Conv-512(k:1,3;s:1,1;p:1,1)	N×512×4×4
	Conv-512(k:3,1;s:1,1;p:1,1)	N×512×4×4
	Max(k:2,2;s:2,2;p:1,1)	N×512×2×2
Conv_8+ReLU+BN	SE_layer	N×512×2×2
	Conv-256(k:1,2;s:1,1;p:1,1)	N×256×1×1
	Conv-256(k:2,1;s:1,1;p:1,1)	N×256×1×1
output		N×256

We use an asymmetric convolutional approach to optimize our network, which decomposes the 3 × 3 convolutional kernel into 3 × 1 and 1 × 3, and it allows the number of parameters to drop by 33% relatively while maintaining the same accuracy rate, as shown in Figure 4 in detail.

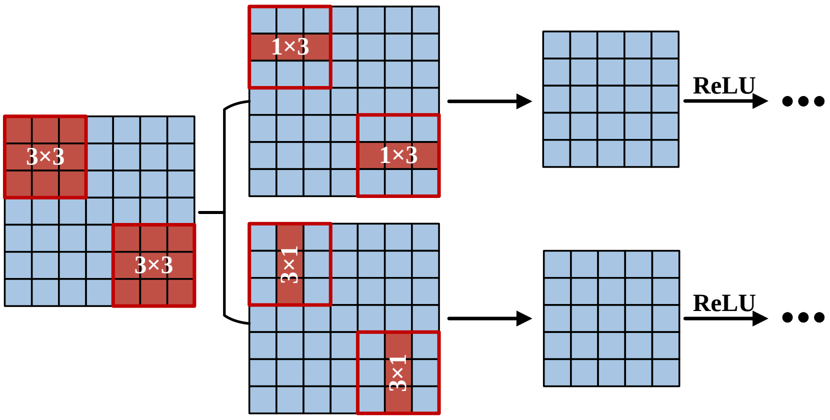


Figure 4. The details of asymmetric structural convolution operations.

As shown in Figure 5, we set up three consecutive 3 × 3 down-sampling layers and one fused down-sampling layer to convolve the image size from 224 × 224 × 3 to 53 × 53 × 160. The extracted features are further fused by three sets of asymmetric convolutional parallel structures to deepen the channel length to 51 × 51 × 352, and finally the dimensionality of feature map is increased to 25 × 25 × 384 via a continuous down-sampling layer and one fused down-sampling layer, and the detailed structure of the stem module is shown in Table 2.

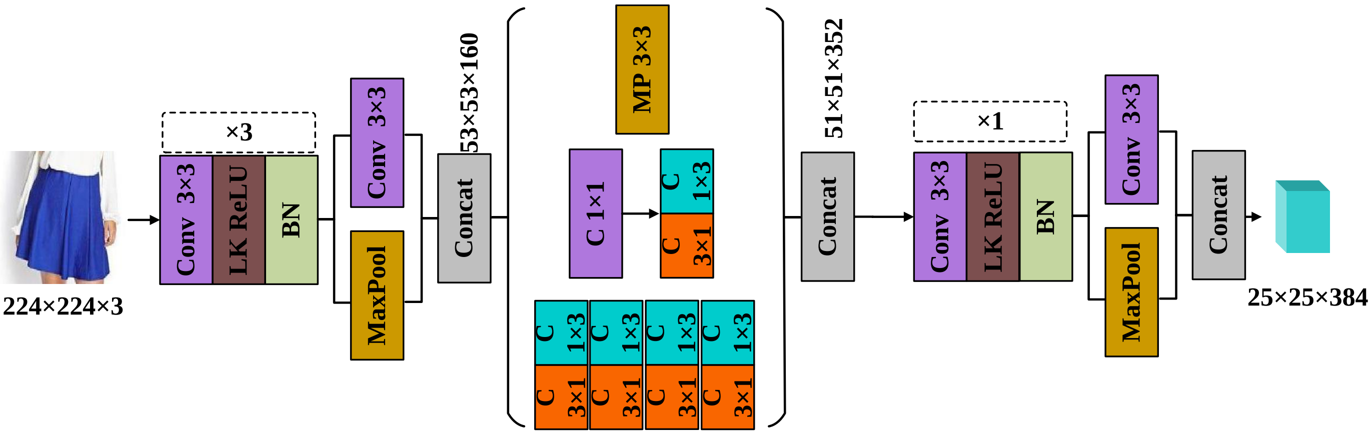


Figure 5. Overview of the Inception mechanism, which belongs to the parameter optimization. **C** denotes 2D convolution, MP denotes Maxpool, BN denotes batch normalization, and LK ReLU denotes Leaky ReLU.

Table 2. The network structure of enhanced feature extraction model. 'k' represents 'kernel', 's' represents 'stride', and 'p' represents 'padding'.

Layer		EnCaps-Stem		Shape
Input		/		N×3×224×224
Conv_1+ReLU+BN		Conv-32(k:3,3;s:2,2;p:1,1)		N×32×111×111
Conv_2+ReLU+BN		Conv-32(k:3,3;s:1,1;p:1,1)		N×32×109×109
Conv_3+ReLU+BN		Conv-64(k:3,3;s:1,1;p:1,1)		N×64×107×107
	Conv-96(k:3,3;s:2,2;p:1,1)		Max(k:3,3;s:2,2;p:1,1)	\
		Concatenate		N×160×53×53
	Conv-64(k:1,3;s:1,1;p:0,0)			\
	Conv-64(k:3,1;s:1,1;p:0,0)			
	Conv-64(k:1,3;s:1,1;p:0,0)			
	Conv-64(k:3,1;s:1,1;p:0,0)	Conv-64(k:1,1;s:1,1;p:0,0)		
	Conv-64(k:1,3;s:1,1;p:0,0)	Conv-96(k:3,1;s:1,1;p:1,1)	Max(k:3,3;s:1,1;p:1,1)	
	Conv-64(k:3,1;s:1,1;p:0,0)	Conv-96(k:1,3;s:1,1;p:1,1)		
	Conv-96(k:1,3;s:1,1;p:1,1)			
	Conv-96(k:3,1;s:1,1;p:1,1)			
		Concatenate		N×352×51×51
Conv_4+ReLU+BN		Conv-192(k:1,1;s:1,1;p:1,1)		N×352×51×51
	Conv-96(k:3,3;s:2,2;p:1,1)		Max(k:3,3;s:2,2;p:1,1)	\
output		Concatenate		N×384×25×25

3. Experiments and Results

3.1. Dataset

The dataset used for experiments is one part of the DeepFashion dataset [28], and it consists of 5000 images in 10 categories: blouse, cardigan, dress, hoodie, jeans, romper, short, skirt, tank, and tee, and each category contains 500 images of the corresponding clothing. The resolution of image in the dataset is 224 × 224. 4500 images of the whole dataset is used for training and another 500 images is used for testing. In addition, in order to enhance the generalization ability and robustness of the model, data enhancement operations including folding, random rotation, and random cropping are performed on the training samples. Figure 6 shows operations corresponding to data augmentation, where (b) represents flipping the original image from top to bottom or left to right, (c) represents rotating the original image at arbitrary angles, the excess area is clipped, and the missing area is filled with white pixels, and (d) represents cropping the original image randomly, and the missing area is filled with white pixels.

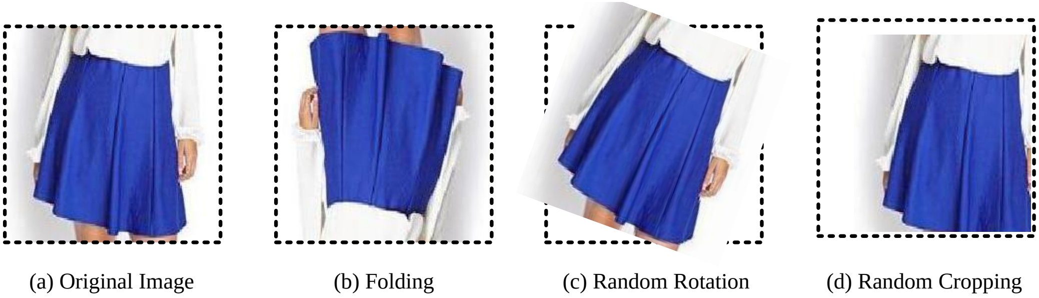

(a) Original Image (b) Folding (c) Random Rotation (d) Random Cropping

Figure 6. Data augmentation.

3.2. Evaluation Criterion

Accuracy, *precision*, *recall*, and F_1 evaluate the performance of the classification model, which can be expressed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{11}$$

where TP, FN, FP, and TN denote true positive, false negative, false positive, and true negative, respectively.

3.3. Experiment Platform Setting

The experiment is conducted on the Ubuntu 16.04 system with the Python language and the tensorflow framework. The hardware environment is equipped with Intel Gold 5118 CPU with 128 GB RAM and 32 GB Nvidia Tesla V100 GPU. By default, the Adam optimizer with $\beta 1 = 0.5$ and $\beta 2 = 0.999$ are used to train the model with 100 epochs, and the initial learning rate is set to 0.0003. The learning rate is automatically decayed by a factor of 0.1 when the validation loss is not significantly reduced.

3.4. The Comparison with Other Methods

The accuracy of convolutional neural network in image classification still depends on the number of samples, the data augmentation strategy, and so on. The problem of how to obtain the same performance with insufficient data annotation is also a concern. With less reliance on supervised learning and priori human annotation information, it is our goal to achieve better performance with smaller amounts of data. To demonstrate the superiority of EnCaps for small sample detection, we verify the accuracy of the 10 networks as the dataset is gradually incremented, as shown in Figure 7. The training set grows from 500 to 4500, the accuracy of EnCaps on the 500 validation set holds the leading value of 0.56 at first, and, as the dataset is incremented, the accuracy on the validation set is consistently higher than the rest of the networks, and maintains a very stable performance.

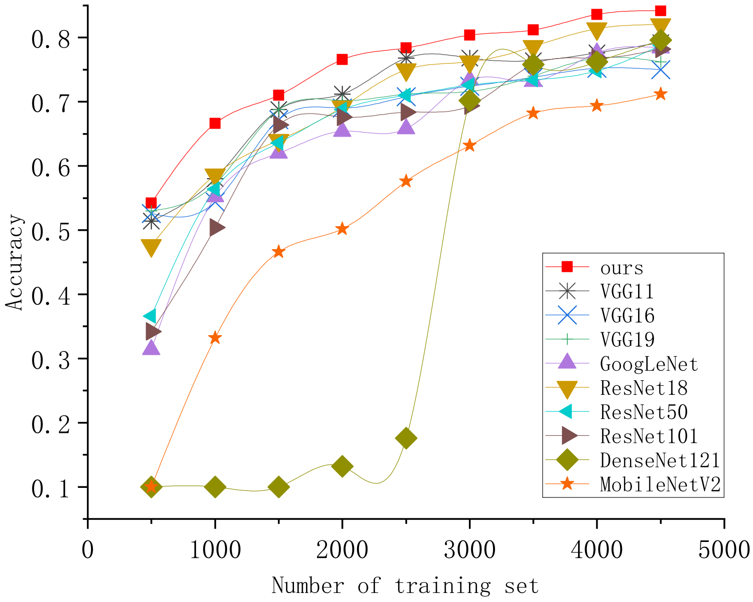


Figure 7. With the decrease of the number of training sets, the accuracy of each network in the verification set (500 pieces) is improved.

During the training, we use ablation experiments to verify the functions of attention mechanism and advanced feature extraction mechanism, respectively. First, we perform a validation of the validity of the channel attention module in the model. We examine the validation set loss and the validation set accuracy of the model, respectively, during the gradual increase of the attention module. As shown in Figure 8, as the attention module increases from 0 to 4, the validation loss of the network decreases and reaches a minimum that is 4. The validation accuracy increases and reaches a peak at the number of 4. As the number increases from 4 to 9, the validation accuracy gets smaller first, some performance is lost, and the validation loss bounces back, which in turn reduces the overall efficiency of the model.

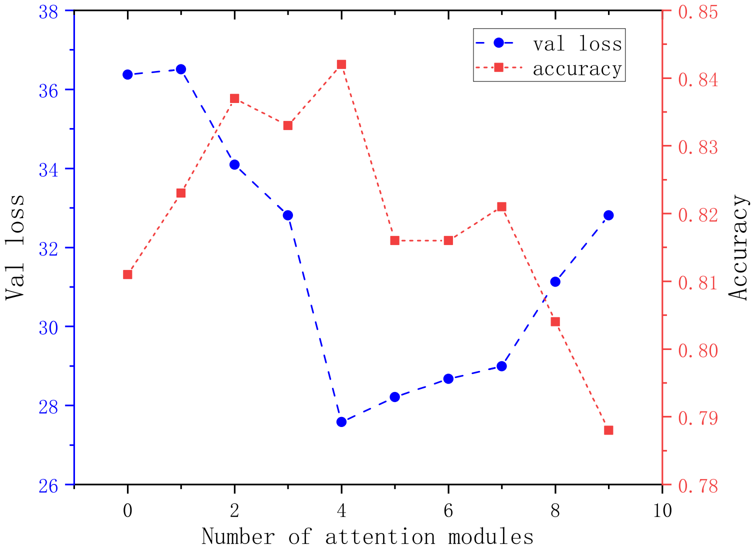


Figure 8. The relationship between the number of attention modules and network performance.

To further investigate the role of each module, we examine the accuracy and loss values on the validation set by removing the stem module, and the reduced module. There are three cases in experiments: (1) removal of the stem module, (2) removal of the reduced module, and (3) removal of the stem module and reduced module.

As shown in Table 3, when we remove the stem module, the accuracy on the validation set decreases from 0.842 to 0.682, and the validation loss value increases from 27.58 to 47.42. It can be intuitively learned that the effect of the stem module on our network is not only to reduce the amount of data, but also the contribution to the performance of the model by extracting more advanced semantic features of the images. When we remove the reduced module, the accuracy on the validation set drops from 0.842 to 0.778, and the validation loss value increases from 27.58 to 43.10. It is clear that the reduced module plays a bridging role in EnCaps, it performs further enhancement to the advanced feature of the stem module, and it normalizes the enhanced feature so that it can be more logical to access the final capsule module. The attention module in the reduced module succeeds in highlighting the focus weights in the feature map, and it has a side reaction to the importance of the stem module. When we remove both the stem and reduced modules, namely it indirectly uses the capsule module, the accuracy on the validation set decreases from 0.842 to 0.628, and the validation loss value increases from 27.58 to 52.76, and it demonstrates that the importance of both modules in the overall module for classification performance. Without these two modules, the number of network parameters and operations increases by a factor of 0.3, and the efficiency is plummeted.

Table 3. Ablation experiments of EnCaps.

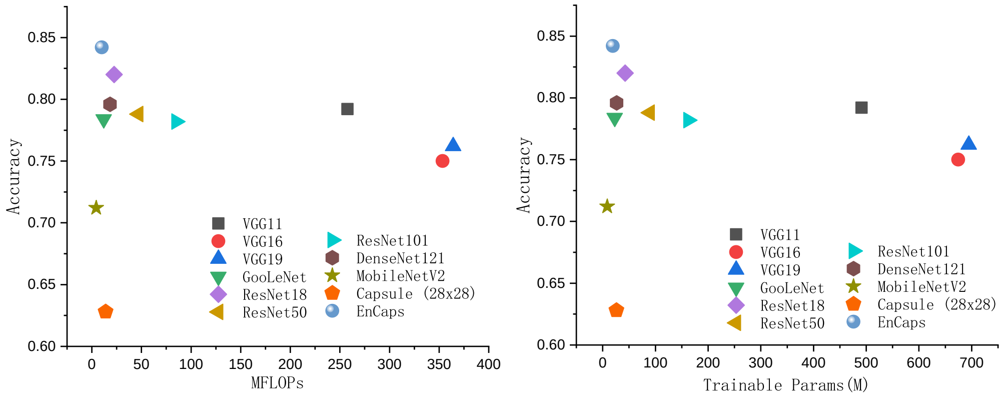
	Validation Accuracy	**Validation Loss**
w/o stem module	0.682	47.42
w/o reduced module	0.778	43.10
w/o stem & reduced module	0.628	52.76
EnCaps	0.842	27.58
Real	1	0

We compare our method with several stare-of-the-art methods including VGG [29], GoogLeNet [30], ResNet [31], DenseNet [32], MobileNetV2 [33], and Capsule network [21]. The quantitative evaluation results are shown in Table 4. We use the validation set to validate each network separately, the number of parameters, and the computational effort of MobileNetV2 are minimal, but the corresponding accuracy is also reduced due to the lack of computational effort. The accuracy of EnCaps, 0.842, is the highest of all the networks. The accuracy of the second highest is ResNet, whose value is 0.820, but with a higher number of parameters and operations than EnCaps. The accuracy of original Capsule network is 0.628, which is much lower than all other networks, and the number of parameters and operations is about 0.3 times higher than EnCaps network. In summary, EnCaps is able to obtain very high detection results in the field of classification, especially in clothing image classification, and maintains a lightweight model volume that remedies the shortcomings of traditional capsule network in classification.

We use visualization to compare the computation, number of parameters, and validation accuracy among 10 types of model, as shown in Figure 9, where the horizontal axis represents the MFLOPs (Million Floating-Point Operation Per Seconds) and the number of trainable parameters respectively, and the vertical axis represents the validation accuracy. By defining the graph, it can be seen that the graph is closer to the y-axis and further from the x-axis, which indicates the higher performance of the model. It can be seen that the EnCaps is at the highest point and relatively close to the y-axis, which means that our model is at the best performance level with a lightweight module and high accuracy.

Table 4. The comparison of calculation, parameter, and accuracy between different networks.

	MFLOPs	Trainable Params (M)	Top-1 Acc.
VGG11	257.592715	491.36	0.792
VGG16	353.619785	674.52	0.75
VGG19	364.23662	694.78	0.762
GoogLeNet	11.953082	22.83	0.784
ResNet18	22.363325	42.65	0.82
ResNet50	47.057184	89.75	0.788
ResNet101	85.041593	162.2	0.782
DenseNet121	18.396804	26.57	0.796
MobileNetV2	4.468249	8.53	0.712
Capsule (28 × 28)	13.863176	26.11	0.628
EnCaps	10.019512	19.13	0.842
Real	N/A	N/A	1



(a) (b)

Figure 9. The comparison of calculation, parameter, and accuracy between different networks. (**a**) the comparison of MFLOPs; (**b**) the comparison of trainable parameters.

From the previous conclusions, it is clear that the performance of ResNet-18 is outstanding among all 10 networks. Therefore, on the validation set, we examine the detection results of 500 pieces of clothing images with EnCaps and ResNet-18, respectively, and draw a confusion matrix that counts the number of detection results per image. As shown in Figure 10, the x-coordinate represents the predicted label, the y-coordinate represents the true label, and each square represents a count of the number of predicted results on the true label. It can be seen that seven EnCaps tests have a number of correct detections above 40, with a relatively low false detection rate, while only five of the ResNet-18 tests have a number of correct detections above 40, with a relatively high false detection rate. Some clothing that has different local features of the clothing leads to different types of clothing, and the ability of detected images at a fine granularity becomes a key basis for determining whether the model has high performance. EnCaps has a high recognition accuracy rate for the detection of different types of clothing, and it has a strong ability to screen fine-grained information to realize the highly robust.

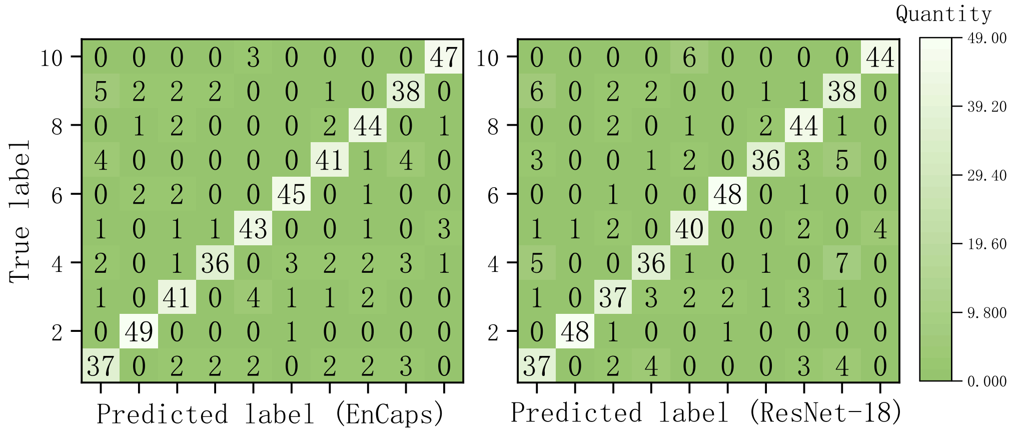


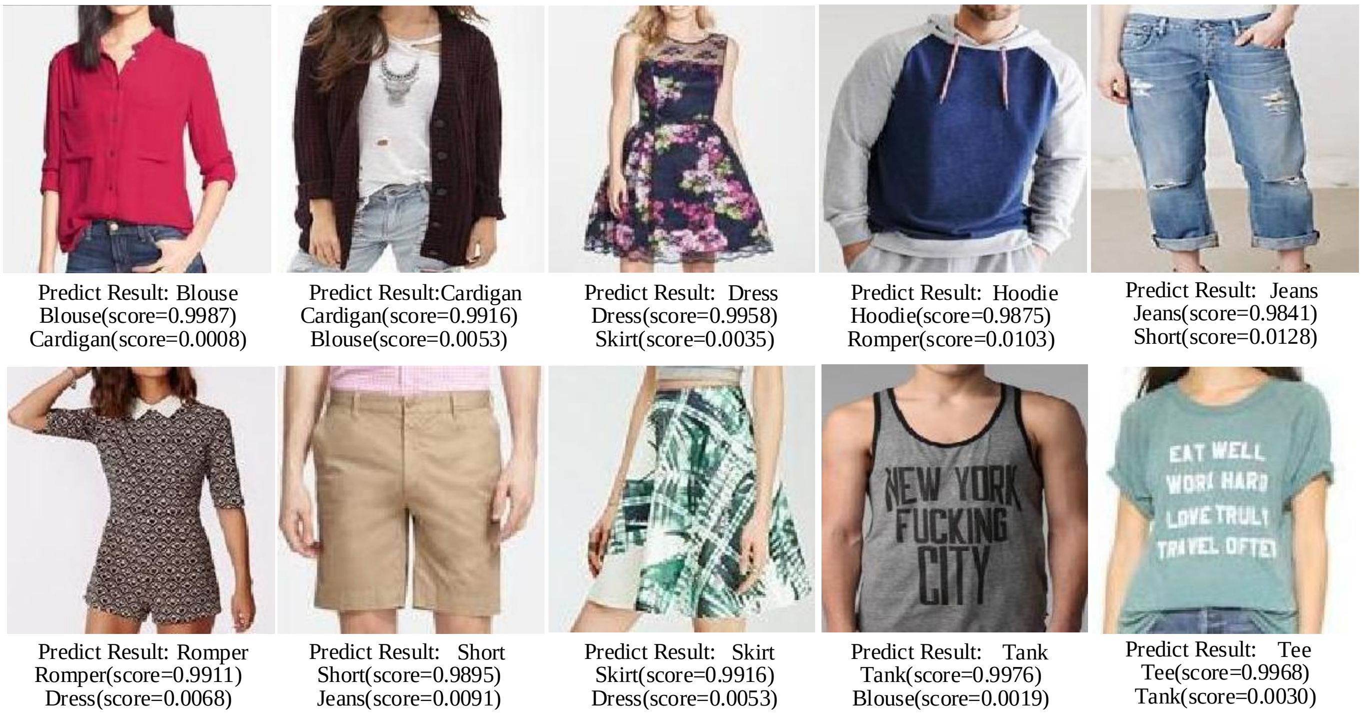
Figure 10. Performance comparison between EnCaps and ResNet-18 on the validation set.

In order to further validate the ability of EnCaps network for classifying clothing, we examine the accuracy, precision, recall, specificity, sensitivity, and F_1 metrics of the model on the validation set, as shown in Table 5, the metrics of EnCaps are at a relatively excellent level, and it can be seen that our proposed EnCaps has a great advantage over other neural networks both in terms of detection effectiveness and network volume.

Table 5. Performance of EnCaps in the accuracy, precision, recall, specificity, sensitivity, and F_1 metrics.

	Accuracy	**Precision**	**Recall**	**Specificity**	**Sensitivity**	F_1
Blouse	0.928	0.652	0.6	0.964	0.6	0.625
Jeans	0.991	0.925	0.98	0.991	0.98	0.951
Skirt	0.956	0.781	0.78	0.976	0.78	0.78
Cardigan	0.952	0.842	0.64	0.987	0.64	0.727
Dress	0.962	0.782	0.86	0.973	0.86	0.819
Short	0.978	0.882	0.9	0.987	0.9	0.891
Tee	0.962	0.816	0.8	0.98	0.8	0.808
Tank	0.968	0.815	0.88	0.978	0.88	0.846
Hoodie	0.948	0.731	0.76	0.969	0.76	0.745
Romper	0.984	0.904	0.94	0.989	0.94	0.922

According to the actual classification results of EnCaps, it demonstrates that EnCaps has high performance in the experiments. We randomly build 10 different types of clothing, test them individually, and obtain their classification results and the probability scores of the first two classification results. As shown in Figure 11, for some clothing images with only fine-grained distinctions, such as dress and romper, the EnCaps network extracts spatial features of the distribution of pixels between them to make a clear comparison and to be able to distinguish them. The proposed EnCaps network has the advanced feature extraction capability of traditional convolutional neural networks and the spatial structure perception capability of Capsule, so that our network is able to classify the indistinguishable images well.



Predict Result: Blouse
Blouse(score=0.9987)
Cardigan(score=0.0008)

Predict Result:Cardigan
Cardigan(score=0.9916)
Blouse(score=0.0053)

Predict Result: Dress
Dress(score=0.9958)
Skirt(score=0.0035)

Predict Result: Hoodie
Hoodie(score=0.9875)
Romper(score=0.0103)

Predict Result: Jeans
Jeans(score=0.9841)
Short(score=0.0128)

Predict Result: Romper
Romper(score=0.9911)
Dress(score=0.0068)

Predict Result: Short
Short(score=0.9895)
Jeans(score=0.0091)

Predict Result: Skirt
Skirt(score=0.9916)
Dress(score=0.0053)

Predict Result: Tank
Tank(score=0.9976)
Blouse(score=0.0019)

Predict Result: Tee
Tee(score=0.9968)
Tank(score=0.0030)

Figure 11. The effectiveness of the EnCaps network in detecting garment classification.

4. Discussion and Future Directions

To realize the more accurate classification of clothing images, we propose the EnCaps network, which uses spatial structure extraction, enhanced feature extraction, and parameter optimization to obtain spatial structure information and robust image feature of clothing images. The EnCaps network not only achieves high accuracy of classification, but also low parameter computation. The experimental results demonstrate the superiority of EnCaps network, which is attributed to the deeper network and optimal network structure. The accuracy and computational complexity are the key metrics that we should consider in the network design. The traditional classification network does not consider the spatial structure feature, and the feature may improve the classification accuracy based on the previous works. Thus, the concept of capsule network is fused into the designed network. The original capsule network is not suitable for clothing classification, and the proposed network is designed anew according to the demand of clothing classification. The input of image size is enlarged by modifying the input network, and more robust feature extraction is obtained by the deeper and more efficient Encaps network. Comparison with traditional capsule network, the network structure is designed anew, and the classification accuracy and efficiency are remarkably improved.

The classical deep learning network focuses on the image feature but spatial location relationship. The improvement of network depth and structure is usually considered to extract robust objective features, such as LeNet [34], AlexNet [35], VGGNet [29], GoogleNet [30], ResNet [31], DenseNet [32], MobileNet [33], YOLO [36–39], and so on. The improvement may be valid for the classical objection classification, such as pedestrian, vehicle, animal, and other objectives with obvious image features. Clothing classification belongs to the fine-grained classification which has inconspicuous features, and the classification methods based on image feature without spatial location feature for difficulty achieving impressive results. Maybe the EnCaps network and other similar networks, which can extract spatial location relationships, are the best choice for the fine-grained classification task.

From the experimental results in the clothing classification, two phenomena are worth discussing. First, the EnCaps network achieves the best performance in terms of top-1 accuracy among VGGNet, GoogleNet, ResNet, DenseNet, MobileNet, and the traditional Capsule network. Generally, the more complex and deeper network may obtain better performance, but EnCaps, which has the lowest computational cost besides MobileNet, which is a lightweight network for mobile devices, obtains the best accuracy. The phenomenon may demonstrate that the spatial location information plays an important part in the procedure of classification. Second, the EnCaps network obtains the best accuracy among all compared methods with a gradually increasing training set from beginning to end. The larger training set is used, and the more accurate performance will be achieved in general. In our experiment, the recognition model is trained by 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and 5000 samples, respectively. The EnCaps network obtains the best performance without exception, which may demonstrate that the spatial location information can boost the efficiency of training models with a small set of samples.

The development of artificial intelligent technology will change the nature of the clothing industry. In the future, more and more intelligent applications will be used for clothing. A system may identify your current clothing and recommend your favorite clothing. Clothing image classification is the fundamental technology in more complex applications, such as evaluation of clothing compatibility, clothing recommendation, and fashion trend prediction. Clothing image classification will be widely applied in the future clothing industry, and it may improve efficiency and convenience of clothing applications. The fine-grained classification method should be further studied for improving the accuracy.

5. Conclusions

A novel EnCaps network is proposed for clothing image classification. The proposed network adopts three strategies to obtain the spatial structure feature and robust image feature: (1) the spatial structure extraction model is proposed to obtain the spatial structure feature of clothing based on the improved capsule network, (2) enhanced feature extraction model is designed to obtain the robust image feature based on the deeper network structure and attention mechanism, and (3) the parameter optimization is used in the EnCaps network based on the inception mechanism. Experimental results indicate that the EnCaps network achieves the best comprehensive performance among classical deep learning networks, such as VGGNet, GoogleNet, ResNet, DenseNet, MobileNet, and the original capsule network. The accurate clothing classification network may be used in the clothing category marking, clothing commodity retrieval, and similar clothing recommendations. In the future work, the more efficient and robust network should be researched to obtain more accurate clothing classification.

Author Contributions: F.Y.: Conceptualization, Algorithm Implementation, Writing—Original Draft; C.D.: Methodology, Writing—Review and Editing; M.J.: Validation, Supervision; A.H.: Formal Analysis, Data Curation; X.W.: Investigation, Preparation of Experimental Equipment and Materials; T.P.: Visualization, Algorithm Design; X.H.: Project Administration, Funding Acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the following projects: Young Talents Program of the Scientific Research program of the Hubei education department of funder grant number No. Q20201709, research on the Key Technology of Flexible Intelligent Manufacturing of Clothing based on the Digital Twin of Hubei Key Research and Development program of funder grant number No. 2021BAA042, and the Open Topic of the Engineering Research Center of Hubei province for Clothing Information of funder grant number No. 900204.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported by the Young Talents Program of the Scientific Research program of the Hubei education department (Project No. Q20201709), research on the Key Technology of Flexible Intelligent Manufacturing of Clothing based on the Digital Twin of Hubei Key Research and Development program (Project No. 2021BAA042), and the Open Topic of the Engineering Research Center of Hubei province for Clothing Information (Project No. 900204).

Conflicts of Interest: We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, "EnCaps: Clothing Image Classification based on Enhanced Capsule Network".

References

1. Jiang, S.; Wu, Y.; Fu, Y. Deep bidirectional cross-triplet embedding for online clothing shopping. *Acm Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2018**, *14*, 1–22. [CrossRef]
2. Chen, Z.; Xu, Z.; Zhang, Y.; Gu, X. Query-free clothing retrieval via implicit relevance feedback. *IEEE Trans. Multimed.* **2017**, *20*, 2126–2137. [CrossRef]
3. Zhang, H.; Sun, Y.; Liu, L.; Wang, X.; Li, L.; Liu, W. ClothingOut: A category-supervised GAN model for clothing segmentation and retrieval. *Neural Comput. Appl.* **2020**, *32*, 4519–4530. [CrossRef]
4. Hidayati, S.C.; You, C.W.; Cheng, W.H.; Hua, K.L. Learning and recognition of clothing genres from full-body images. *IEEE Trans. Cybern.* **2017**, *48*, 1647–1659. [CrossRef]
5. Wang, W.; Xu, Y.; Shen, J.; Zhu, S.C. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4271–4280.
6. Zhang, S.; Song, Z.; Cao, X.; Zhang, H.; Zhou, J. Task-aware attention model for clothing attribute prediction. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 1051–1064. [CrossRef]
7. Zhang, Y.; Zhang, P.; Yuan, C.; Wang, Z. 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, Seattle, WA, USA, 13–19 June 2020; pp. 13538–13547.
8. An, L.; Li, W. An integrated approach to fashion flat sketches classification. *Int. J. Cloth. Sci. Technol.* **2014**, *26*, 346–366. [CrossRef]
9. Berg, T.L.; Ortiz, L.E.; Kiapour, M.H.; Yamaguchi, K. Parsing clothing in fashion photographs. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Providence, RI, USA, 16–21 June 2012.
10. Yang, X.; Yuan, S.; Tian, Y. Assistive clothing pattern recognition for visually impaired people. *IEEE Trans. Hum. Mach. Syst.* **2014**, *44*, 234–243. [CrossRef]
11. Chen, H.; Gallagher, A.; Girod, B. Describing clothing by semantic attributes. In *European Conference on Computer Vision*; Springer, Berlin, Heidelberg, 2012; pp. 609–623.
12. Zhan, H.; Yi, C.; Shi, B.; Lin, J.; Duan, L.Y.; Kot, A.C. Pose-normalized and appearance-preserved street-to-shop clothing image generation and feature learning. *IEEE Trans. Multimed.* **2020**, *23*, 133–144. [CrossRef]
13. Gao, Y.; Kuang, Z.; Li, G.; Luo, P.; Chen, Y.; Lin, L.; Zhang, W. Fashion Retrieval via Graph Reasoning Networks on a Similarity Pyramid. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, 1–15. [CrossRef]
14. Donati, L.; Iotti, E.; Mordonini, G.; Prati, A. Fashion Product Classification through Deep Learning and Computer Vision. *Appl. Sci.* **2019**, *9*, 1385. [CrossRef]
15. Kim, H.J.; Lee, D.H.; Niaz, A.; Kim, C.Y.; Memon, A.A.; Choi, K.N. Multiple-Clothing Detection and Fashion Landmark Estimation Using a Single-Stage Detector. *IEEE Access* **2021**, *9*, 11694–11704. [CrossRef]
16. Chen, W.; Huang, P.; Xu, J.; Guo, X.; Guo, C.; Sun, F.; Li, C.; Pfadler, A.; Zhao, H.; Zhao, B. POG: Personalized outfit generation for fashion recommendation at Alibaba iFashion. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2662–2670.
17. Zhang, S.; Liu, S.; Cao, X.; Song, Z.; Zhou, J. Watch fashion shows to tell clothing attributes. *Neurocomputing* **2018**, *282*, 98–110. [CrossRef]
18. Yang, Q.; Wu, A.; Zheng, W.S. Person re-identification by contour sketch under moderate clothing change. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**. [CrossRef] [PubMed]
19. Liu, L.; Zhang, H.; Ji, Y.; Wu, Q.J. Toward AI fashion design: An Attribute-GAN model for clothing match. *Neurocomputing* **2019**, *341*, 156–167. [CrossRef]
20. Huang, C.Q.; Chen, J.K.; Pan, Y.; Lai, H.J.; Yin, J.; Huang, Q.H. Clothing landmark detection using deep networks with prior of key point associations. *IEEE Trans. Cybern.* **2018**, *49*, 3744–3754. [CrossRef] [PubMed]
21. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3859–3869.
22. Hinton, G.E.; Sabour, S.; Frosst, N. Matrix capsules with EM routing. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.

23. Jia, B.; Huang, Q. DE-CapsNet: A Diverse Enhanced Capsule Network with Disperse Dynamic Routing. *Appl. Sci.* **2020**, *10*, 884. [CrossRef]
24. Zhang, Z.; Ye, S.; Liao, P.; Liu, Y.; Su, G.; Sun, Y. Enhanced Capsule Network for Medical image classification. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 1544–1547. [CrossRef]
25. Xiang, C.; Zhang, L.; Tang, Y.; Zou, W.; Xu, C. MS-CapsNet: A novel multi-scale capsule network. *IEEE Signal Process. Lett.* **2018**, *25*, 1850–1854. [CrossRef]
26. Edraki, M.; Rahnavard, N.; Shah, M. Subspace capsule network. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10745–10753.
27. Do Rosario, V.M.; Borin, E.; Breternitz, M. The multi-lane capsule network. *IEEE Signal Process. Lett.* **2019**, *26*, 1006–1010. [CrossRef]
28. Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; Tang, X. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1096–1104.
29. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
30. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
32. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
33. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
34. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
35. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
36. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
37. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
38. Farhadi, A.; Redmon, J. Yolov3: An incremental improvement. In *Computer Vision and Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 1804–2767.
39. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.