applied sciences

MDPI

Article

High-Accuracy Clothing and Style Classification via Multi-Feature Fusion

Xiaoling Chen [1], Yun Deng [1], Cheng Di [1,*], Huiyin Li [2,†], Guangyu Tang [2,†] and Hao Cai [2]

[1] School of Arts and Media, China University of Geosciences (Wuhan), Wuhan 430070, China
[2] School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China
* Correspondence: dicheng@cug.edu.cn
† These authors contributed equally to this work.

Abstract: In recent years, the online selection of virtual clothing styles has been used to explore and expand diversified personal aesthetics, and it is also an overall reform and challenge to the clothing industry. Under the condition of the existing clothing style categories, this paper puts forward a style classification method combining fine-grained and coarse-grained techniques. Furthermore, a new deep neural network is proposed, which can improve the robustness of recognition and avoid the interference of image background through the pan learning and the background learning of image features. In order to study the relationship between the fine-grained attributes of clothing and the whole style, firstly, the clothing types are learned to realize the pre-training of model parameters. Secondly, through the transfer learning of the first stage of the pre-training model parameters, the model parameters are fine-tuned to make them more suitable for identifying the coarse-grained style types. Finally, a network structure based on the dual attention mechanism is proposed to improve the accuracy of final identification by adding different attention mechanisms at different stages of the network to enhance the performance of network features. In the experiment, we collected 50,000 images of 10 clothing styles to train and evaluate the models. The results show that the proposed classification method can effectively distinguish clothing styles and types.

Keywords: superposition module; dual attention mechanism; convNeXt-SP; small-scale clothing data set

check for updates

Citation: Chen, X.; Deng, Y.; Di, C.; Li, H.; Tang, G.; Cai, H. High-Accuracy Clothing and Style Classification via Multi-Feature Fusion. *Appl. Sci.* **2022**, *12*, 10062. https://doi.org/10.3390/app121910062

Academic Editor: João M. F. Rodrigues

Received: 5 September 2022
Accepted: 29 September 2022
Published: 6 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the development of an intelligent information field and the gradual popularization of a high-speed bandwidth network (5G), more and more people have begun to pursue the online virtual shopping [1] experience to replace offline shopping. Clothing shopping has always been a manifestation of the human pursuit of beauty. Therefore, the realization of the virtual online clothing experience has gradually become a research hotspot in recent years [2,3]. The type of clothing has always been the first screening factor for clothing purchase [4–7], such as short sleeves, skirts, and jackets. The style of clothing such as gothic or baroque has also been considered by researchers in the field of clothing [8]; these are some basic issues in the field of modern intelligent clothing.

In the field of clothing classification [9], many researchers have conducted a lot of research. In combination with mechanical sensors, Willimon et al. [10] present a system for automatically extracting and classifying items in a pile of laundry. Using only visual sensors, the robot identifies and extracts items sequentially from the pile. The classification procedure relies upon silhouettes, edges, and other low-level image measurements of the articles of clothing. The accuracy rate of six kinds of clothing (pants, shorts, short-sleeve shirt, long-sleeve shirt, socks, or underwear) classification is 59%. Willimon et al. [11] present a novel approach for classifying items from a pile of laundry that exploits color, texture, shape, and edge information from 2D and 3D local and global information for each article of clothing using a Kinect sensor. They achieve a true positive rate of 90%. In terms

of new algorithms, Yamazaki et al. [12] propose a method of clothing classification using a single image. A set of Gabor filters is applied to an input image, and then, several image features that are invariant to translation, rotation and scale are generated. We propose the descriptions of the features focusing on clothing fabrics, wrinkles and cloth overlaps. Yan et al. [13] propose an improved CNN for clothing classification, adjust the structure of the original CNN model and increase the volume of the reel in the adjusted structure. Wang et al. [14] propose a knowledge-guided fashion network to solve the problem of visual fashion analysis, and they propose two important fashion grammars: (i) dependency grammar capturing kinematics-like relation, and (ii) symmetry grammar accounting for the bilateral symmetry of clothes. They introduce Bidirectional Convolutional Recurrent Neural Networks (BCRNNs) for efficiently approaching message passing over grammar topologies and producing regularized landmark layouts.

In the classification of clothing style [15–18], more and more people began to put their energy into the study. Sun et al. [8] propose one clothing style classification research algorithm based on a multi-core support vector machine (SVM) optimized visual word package model.

In our work, we have studied the network setup strategy for clothing style classification and the availability of similar learning features in similar learning tasks. The recognition accuracy of another task can be improved by transfer learning. The remainder of this paper is arranged as follows. Section 2 introduces the related work. Section 3.1 introduces the clothing category and style classification, while Section 3.6 introduces the lassification network of clothing types in detail. Section 4 analyzes the experimental results of the report, produces a summary and prospects for future improvement are outlined in Section 5.

2. Related Work

2.1. Clothes and Accessories Attributes

The attribute description of a garment is a unit of a garment [19]. Usually, a garment is composed of texture (e.g., palm, colorblock), fabric (e.g., leather, tweed), shape (e.g., crop, midi) and part (e.g., bow-F, fringed-H) [20]. The early attribute exploration method is to mine clothing images with fine-grained attribute labels from online shopping malls [21] to solve this problem [22], such as various shades of color (e.g., watermelon red, rosy red, purplish red), clothing types (e.g., down jacket, denim jacket), and patterns (e.g., thin horizontal stripes, houndstooth). Clothing can be located by a number of attributes. Conversely, it has long been realized to automatically classify multiple attributes from one article of clothing [23]. However, our work is not to identify the clothing attributes but to train the relationship parameters between the attributes and types of clothing during the clothing classification operation and then obtain a neural network with the ability of mining information, so that the neural network can deeply mine the internal fine-grained information of unmarked clothing and realize the target parameters of pre-fitting style classification according to the training.

2.2. Discover Style

The style types of clothing can be analyzed from different perspectives [8]. Different unique patterns (e.g., patterns and cats) can be used as a style, different materials (e.g., nylon and cotton) can be used as a style, and shapes can also be used as a style [24]. However, these classification methods cannot cover all clothing styles, because the same style of clothing also has different styles of pattern and material composition. The formation of a style is precisely the combination of fine-grained attributes [25,26], e.g., gothic, baroque or bohemian styles, which can be visually distinguished, but some styles of clothing are difficult to distinguish only by vision. For example, simple style, street style and Korean style are overlapped in some attributes. This undoubtedly increases the difficulty of classification [13]. Our task is to find out some popular styles of clothing in the world and train a model to match and retrieve the same style of clothing.

2.3. Attributes in Style

No matter from clothing stores or online shopping malls, the types of clothing styles can be divided into OL, sports, rural and other styles. These styles can be aggregated into a limited number of specific categories [27], and these coarse-grained styles can provide valuable fine-grained information [25,28]. That is to say, a style can contain several sub-attributes. Generally speaking, we can study the attributes of a garment from the perspective of upper body and lower body. The perspective of upper body can be divided into collar and sleeve; the perspective of lower body can be divided into skirt and pants. A collar can be divided into collar shape, neck line, etc.; a sleeve can be used as the attribute of sleeve length. We also add a total perspective (e.g., color, one-piece, mix) to avoid the loss of contact information between the part and the whole. This kind of relationship data between fine-grained and coarse-grained cannot be ignored, which is also our focus. Therefore, it is important to find a common set of attributes that fit all styles. According to the total set, we can deduce some elements to form a kind of style through a certain convolution neural network (CNN) mathematical model [29]. Therefore, it is our task to obtain the accurate classification style of this mathematical model.

3. Methods and Methodology

3.1. Clothing Multiclass Classification

We propose a convolutional neural network based on the clothing category data set to realize the classification of clothing categories. Let us start with (1) different kinds of clothing images being marked to distinguish the category of each image. (2) Then, the whole image is trained, and the parameters of the whole network are adjusted to make it infinitely close to the target parameters of classification. As shown in Figure 1, it is the overall structure of this network.

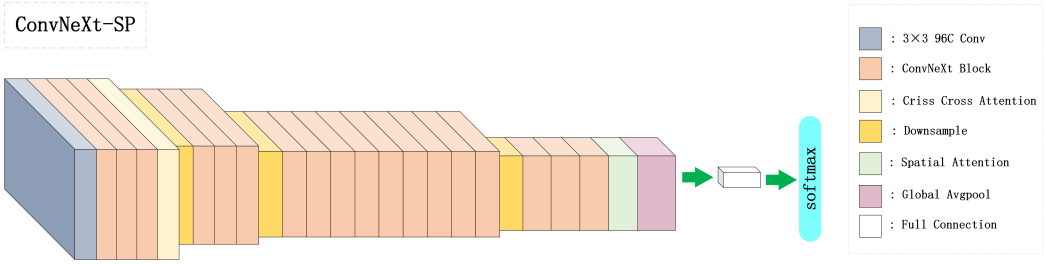

Figure 1. The total network structure of the CNN model with deep attributes.

3.2. Category Classification Network

The classification of clothing types is a basic problem in the field of modern intelligent clothing. The purpose of this paper is to calculate the category of clothing and the semantic attributes (e.g., color, collar type) of each parameter on the basis of removing the complex background. In the traditional situation, classification is realized by the combination of image feature extraction and classifier classification. For classifiers, there are two categories and multiple categories. The common classification algorithms are SVM (support vector machine) for linear and non-linear classification, e.g., LeNet [30], AlexNet [29], VGG [31], GoogLeNet [32] and ResNet [33] used in the field of deep convolution neural network. With the in-depth study of CNN, the emergence of these convolutional neural networks has gradually improved the accuracy of image classification; compared with the traditional machine learning algorithm, it also shows a better effect.

Due to the emergence of thousands of classification requirements for these large-scale data sets (e.g., Imagenet) [29], the demand for the feature expression ability of convolutional neural network algorithms is gradually higher; the deeper the network becomes, the larger the volume is. For the classification of our small-scale clothing data set, how to ensure the reasonable volume of the premise so that the network can maintain a good recognition rate and convergence ability and even improve its generalization identification ability has become our problem to solve.

3.3. Improved Network Architecture

Our classification network is improved from ConvNeXt network, which is the best performance feature extraction network available, and no complex structure is improved on the basis of ResNet. The details of the network parameter settings are based on some ideas from Transformer [34] or Swin Transformer [35]. ConvNeXt uses a lot of experimental data to prove which settings work well, and those that work well are retained, and the accuracy of classification by such settings alone exceeds that of Swin Transformer.

Because the volume of our data set is small and the classification type is only small, we used ConvNeXt-T to design a network suitable for the proposed clothing category classification. Table 1 shows the details of the proposed network parameters for clothing category classification.

Table 1. Network structure with input image size 224 × 224 × 3.

Layer	Output Size	Channel	Kernel Size/Stride
Convolution	56 × 56	96	4 × 4 / 4
ConvNeXt Block	56 × 56	96	$\begin{bmatrix} d7 \times 7/1 \\ 1 \times 1/1 \\ 1 \times 1/1 \end{bmatrix} \times 3$
Criss Cross Attention	56 × 56	96	-
Downsample	28 × 28	192	2 × 2 / 2
ConvNeXt Block	28 × 28	192	$\begin{bmatrix} d7 \times 7/1 \\ 1 \times 1/1 \\ 1 \times 1/1 \end{bmatrix} \times 3$
Downsample	14 × 14	384	2 × 2 / 2
ConvNeXt Block	14 × 14	384	$\begin{bmatrix} d7 \times 7/1 \\ 1 \times 1/1 \\ 1 \times 1/1 \end{bmatrix} \times 9$
Downsample	7 × 7	768	2 × 2 / 2
ConvNeXt Block	7 × 7	768	$\begin{bmatrix} d7 \times 7/1 \\ 1 \times 1/1 \\ 1 \times 1/1 \end{bmatrix} \times 3$
Spatial Attention	7 × 7	768	-
Global Avgpool	7 × 7	1024	7 × 7 /1
Full Connection	1 × 1	1000	-

The input size of the model is set as a 224 × 224 × 3 square image in the design. As shown in Table 1, in the first layer of the model, the images are preprocessed, using 4 × 4 convolutional kernels with a step size of 4 and a number of 96, so that the images are scaled down in the scale direction before processing and enriched and expanded with information in the number of channels, and Layer Normalization is performed to accelerate the convergence of training and improve the stability of network fluctuations. Then, the feature map will be fed into the ConvNeXT Block. There are four ConvNeXT Block groups in the whole network structure, where the number of ConvNeXT Blocks in each group is 3, 3, 9, and 3, respectively. A criss cross attention module is set up after the first ConvNeXT Block group to obtain criss cross attention weight feature maps. The second to fourth ConvNeXT Block groups are preceded by a downsample layer to adjust the size and number of channels of the feature map, and a spatial attention module is connected after the last ConvNeXT Block group to collect the attention weights of the feature map in the spatial dimension.After this, we use the global average pooling operator to pool the entire feature layer, making the spatial information of the model more representative, avoiding overfitting in this layer, reducing computational effort, and enhancing robustness

to changes in the input space. The last layer is the full connection layer of the model. The full connection dimension changes with the number of categories. Each ConvNext block uses a 7×7 convolution kernel with a step size of 1 and the number of input channels for deep separable convolution to learn image features while reducing computational effort. The number of channels is then adjusted using two consecutive 1×1 convolution kernels of step 1. The first 1×1 convolution kernel adjusts the number of feature map channels to four times the number of input channels, and the second 1×1 convolution kernel adjusts the number of feature map channels to the number of input channels. The 7×7 convolution is followed by Layer Normalization, and the second 1×1 convolution is followed by a corrected non-linear activation unit (GELU). The Layer Normalization layer is used to accelerate the convergence of training and increase the stability of network fluctuation. The non-linear activation unit is used to increase the non-linear relationship between each layer of the neural network, so that the prediction parameters can fit the real parameters and improve the precision degree.

3.4. Dual Attention Module

As shown in Figure 2, the first one is the criss cross attention module, which obtains the relationship of each pixel with the surrounding pixel points. The input feature map is passed through a 1×1 convolution kernel with a step size of 1 to obtain three feature maps—Q, K, and V—where the number of channels of Q, and K is compressed to $1/8$ the number of input channels to reduce the amount of subsequent operations. The Affinity operation is performed on Q, K to obtain the relationship between the pixel points in the feature map Q and the pixel points in the same column of the feature map K peers independent of the channel. Such an operation is performed for each position in Q, i.e., a new feature map of size $[(W + H - 1) \times W \times H]$, denoted as D, can be obtained.The softmax operation is performed on D to make the contribution of each position more explicit, and the feature map is obtained as A. The interrelationship between positions has been obtained after the operation as above.

The Aggregation operation is performed next, which involves the operation of two special evidence maps, A and V. $A_\mu \epsilon R^{W+H-1}$ denotes the feature vector at position μ in the feature map A, and ϕ_{μ_i} denotes the feature vector at layer i of feature map $(i = 1, 2, ..., C)$ at position μ in the feature map consisting of pixels in the same column as their peers, $\phi_{\mu\_i} \epsilon R^{W+H-1} (i = 1, 2, ..., W + H - 1)$. Multiplying A_μ and ϕ_{μ_i} vectors one by one and performing such an operation, we can obtain the criss cross attention weights with feature map size $C \times W \times H$. The obtained feature maps with the criss cross attention are subjected to an element by element summation operation with the original map.

The second one is the spatial attention module, which obtains the attention weights on the spatial dimension. The input feature value map is subjected to maximum pooling operation and average pooling operation to obtain two feature maps, respectively; then, the two feature maps are stitched together, and the number of channels is adjusted to 1 by a 1×1 convolution kernel with a step size of 1. to obtain the spatial attention weights. The obtained spatial attention weights are multiplied with the original map.

With the above two attention modules, the multi-scale feature extraction of the network training process can be enhanced to make the network performance superior, and the specific proof can be referred to the Experimental section. Compared with the traditional feature extraction methods, it has good anti-overfitting performance and good generalization ability without expanding the data, it can avoid the phenomenon of gradient divergence when the layers are deepened, and more importantly, it increases the expression of feature information.

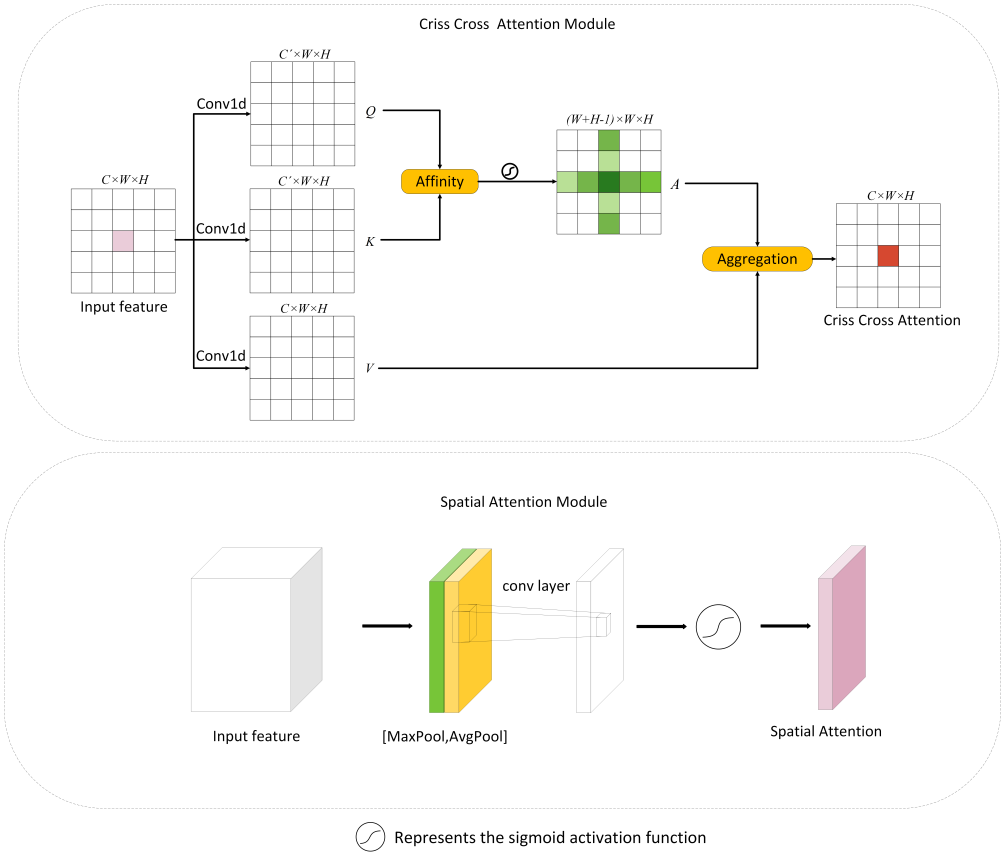

Figure 2. The attention module structure involved in this paper.

3.5. Multiclass Classification and Prediction

In the traditional sense, the classification tasks of a CNN network are mostly seen in single labels and single tasks. For example, the single label classification of a garment has only shirts, which can be summarized in the category of multiclassification. Let $\mathcal{L} = \{\lambda_1, \lambda_2, ..., \lambda_n\}$ be the limited sample space of classification (data set containing n images), $\mathcal{L}$, where a sample element $\lambda_i = \{(X^{(i)}, Y^{(i)})|X\epsilon\mathcal{X}, Y\epsilon\mathcal{Y}\}$, where $\mathcal{X} = \{x^{(1)}, x^{(2)}, ..., x^{(N)}\}$ denotes the sample space of all images in the data set, and N denotes the category classification of clothing in the data set. $\mathcal{Y} = \{y^{(n)}|y\epsilon\{0, 1\}\}$ denotes the sample space of label data in the data set.

At the end of the model, we use softmax classifier to classify the clothing categories. For the multi classification data set, we can use the average information entropy to evaluate its overall quality effect. The training capacity is not easy to converge, and the complexity of the data distribution in the data set is calculated, whether it is stable and uniform. If the data information is more complex and there are more types of different situations, then this information entropy is relatively large. On the contrary, the simpler the data information is, the less kinds of different situations appear; then, this information entropy will be relatively small. We mark the number of clothing types or styles in the data set as $(x_1, x_2, ..., x_n)$. The training probability of each species in the total data set is marked as $(p_1, p_2, ..., p_n)$. Then, we can obtain the formula:

$$H_{avg}(X) = -\frac{1}{N}\sum_{x=1}^{N} p(x_i) \cdot \log_2 p(x_i) \quad (1)$$

where X denotes the original input data set, and N denotes the number of types of tags contained in the data set. We set a threshold W. If the value of $H(x)$ is below the threshold, the image data marks are better, the distribution is regular, the quality is high and the convergence is easy. So, we can choose the training data set as the training data set first; otherwise, the error rate of the data set marking is high, and the content is complex.

Therefore, it is difficult to train the image data set with this kind of data set, and the effect is not good.

The input of softmax classifier is the output value $I = \{y_1, y_2, ..., y_n | n \epsilon N\}$ of the last layer of the network, where N is the number of channels in the last full connection layer. Softmax makes a complex weighted sum and non-linear processing of the input values and then obtains a probability distribution as the final output of the neural network:

$$\hat{y} = softmax(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_j}}, n = N \tag{2}$$

In this way, we can obtain the probability value $\hat{y}^{(i)}$ belonging to each category and then take the input one-hot coding label as the target probability $y'^{(i)}, y' \epsilon \{0, 1\}$; then, we can obtain the batch loss function L of cross entropy of each batch M:

$$\begin{aligned} L(y', \hat{y}) = -\frac{1}{M} \sum_{x}^{N} (y'(x) \log \hat{y}(x) \\ +(1 - y'(x)) \log(1 - \hat{y}(x))) \end{aligned} \tag{3}$$

3.6. Clothing Style Classification

In this paper, we transfer the parameters of the pre-trained clothing type model and propose a convolutional neural network based on the clothing style data set to realize the clothing style classification. First of all, (1) we mark the images of different styles of clothing to determine the information that needs attention. Then, (2) we train the labeled clothing image to realize the transfer learning and fine-tuning training of the model. Finally, (3) the network assistant of a pseudo twin structure is put forward to strengthen the diversity of the clothing style classification. It is the overall structure of the clothing style classification network.

3.7. Classification Network of Fashion Style

The classification of clothing style is a hot topic in the field of clothing. It consists of separating the styles of some clothing in the world with mathematical models. The purpose is to serve customers better. They can freely choose their favorite styles in daily clothing purchase or help designers in their style design. There are very fuzzy conditions to define the style of clothing in artificial classification. For example, in the national style, there are a wide range of areas to explore. There are many ethnic groups in the world, and there are very big differences between the costumes of different nationalities, which leads to the style having no rule to follow. For example, when two styles are close, that hinders the classification.

As shown in Figure 1, the clothing style classification network designed by us will have two forms. The first is to use the ConvNeXt-SP network designed by us to use the clothing category classification model for transfer learning and then carry out training, which can have a much better effect than non migration learning. The second is to use a dual attention mechanism network model to perform multi-task training. Here, a fully connected layer is added to the feature map obtained by the final global average pooling, which is then connected in the channel direction and then subjected to classification operations. This kind of network can have better performance in feature expression.

3.8. Transfer Learning Process

In our research, we found that there is a strong internal relationship between the classification of clothing types and styles, because there are many similar characteristics between them. Therefore, we can use the method of transfer learning [36] to assist us in the training of style classification model, so as to help the model converge faster and have higher accuracy. First, we train the models on large categories; that is, the models are trained in the field. We label the target domain of species as $D_s = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$, and the

distribution of its value domain is $P(\mathbf{x}_s)$. We label the auxiliary domain as $D_t = \{\mathbf{x}_j\}_{j=n+1}^{n+m}$, and the distribution of its value domain is $P(\mathbf{x}_s)$. The data distribution $P(\mathbf{x}_s)$ and $P(\mathbf{x}_t)$ of these two domains is different; that is, $P(\mathbf{x}_s) \neq P(\mathbf{x}_t)$. However, they share similar characteristics, so we can use transfer learning to obtain the knowledge of D_s in the target region with the help of D_t knowledge. Our source domain here is the DeepFashion [20] data set, while the target domain is the Style10 data set.

In the first stage, we trained the model of clothing classification, in which the model has learned a lot of clothing features. Then, based on these features, we find out the common feature representation between the source domain and the target domain, and we use these features to transfer knowledge to obtain our final clothing style classification model.

4. Experiments

4.1. Data Set and Augmentation

When training and testing the clothing category classification network model, we used DeepFashion, which is a high-quality and large-scale clothing data set opened by the Chinese University of Hong Kong. The data set was obtained by the DeepFashion website [37]. This data set contains 289,222 diverse clothes images from 46 different categories.

The DeepFashion data set classifies clothes of the same type with the same attributes. Each image maintains the original aspect ratio, and the length of the longest side is resized to 300 pixels. As shown in Table 2, all images of 46 categories are coded according to the category number. Nowadays, it is difficult to define and classify the styles of clothing in the field of clothing, and there is no authoritative organization to count the categories of clothing styles in the world. In the DeepFashion data set, there is an attribute named style, which contains several categories. However, the categories involved are too general to be representative (e.g., art, angeles, cat, etc.). If we assume that the style is s, when other styles of clothing contain s element, it will appear that the category cannot be specifically distinguished. The specific embodiment is that if we take cat as a style, all clothing containing cat will be classified into one category. For example, if cat is included in pastoral style, street style or sports clothing, it will be classified into one category. Obviously, this practice is not representative and unscientific. Therefore, we constructed a data set named Style10 to train and test our CNN model as well as to evaluate the performance of the proposed method. The Style10 data set contains a total of 10 clothing styles commonly seen today, including bohemian, camouflage, Chinese style, college, dress, ethnic, hip hop, professional, small fragrance and sports; each style contains almost 5000 images and 50,000 images in total. We resized the longest edge of each image to 224 pixels and labeled a portion of the image in order to remove distracting backgrounds using deep learning-based foreground extraction algorithms. The foreground extraction algorithm based on deep learning we are talking about here used DeepLabv3+ [38]. The steps are as follows: take 3000 images from the Style10 data set as the training data set for labeling, with only two categories of labels: foreground and background, where the foreground is the part of the person and the background is the other part, and use the labeled data set to train the DeepLabv3+ model. The trained model is applied to the entire Style10 data set to segment the foreground and background, and then using the segmented mask, it is easy to create images that retain only the foreground. Then, we randomly select 30,000 images as the train set, 10,442 images as the verification set and 9558 images as the test set.

Table 2. Extract of image name and label in DeepFashion data set.

ImageName	Category Name	Label
0000001.jpg	Skirts	0000010...000
0000002.jpg	Blouses_Shirts	0000000...010
0000003.jpg	Cardigans	0010000...000
0000004.jpg	Denim	0100000...000
0000005.jpg	Dresses	0000000...100
0000006.jpg	Jackets_Coats	0000100...000

Through the experimental experience of the researchers, it is found that a series of changes such as distortion, flipping, scaling, and color space transformation of the data set image can be used to expand the data, generating similar but different training samples. This can increase the training data but also prevent the problem of over fitting. Then, it is possible to rearrange the data set randomly to reduce the correlation between adjacent samples' sex. In this way, the robustness of the detection can be enhanced, and the generalization ability of the model can be improved, so that the convergence of the model can be accelerated.

Through rotation, horizontal flip, and shift, the operation expands the training set data, as shown in Figure 3.

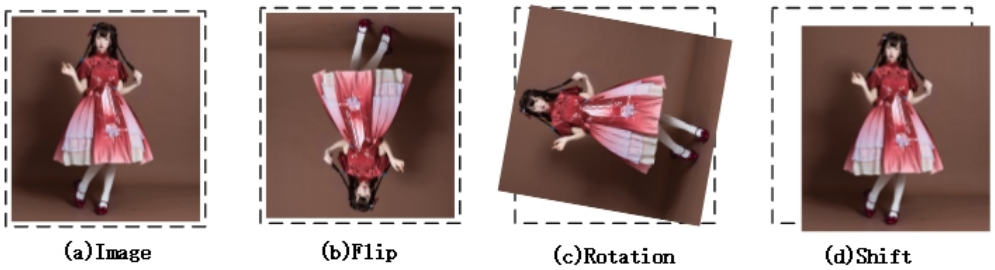

Figure 3. Data set augmentation.

By flipping the image angle or cutting it differently, the region to be detected can be matched with a model compatible with multi-angle scenes when training the model, which weakens the relevance of the object's position; similarly, the color space transformation can reduce the sensitivity of the model to the image color, that is, to weaken the relevance between color and goal.

4.2. Device and Training

The experiment is carried out in the following environment: the hardware device uses NVIDIA Tesla A100 GPU, the software platform uses Ubuntu 16.04 as the operating system, the programming language is Python 3.8, and the Pytorch 1.11.0 framework is used to realize our convolutional neural network.

4.3. Category Classification Network

The AdamW optimizer is used in multi-class classification network training. The learning rate will first go through the warm-up of an epoch, and the warm-up rules are as follows:

$$Lr = f * (1 - \alpha) + \alpha \tag{4}$$

where f is set to 0.001, and α changes with step. The specific change rules are as follows:

$$\alpha = Step_k \div (W \cdot s) \tag{5}$$

where W represents the total number of warm-up epochs set, which we set as 1, and S represents the total number of steps. After the warm-up of the epoch, we set the initial

learning rate at 0.0005 and trained 50 rounds. During the training, the learning rate was adjusted according to Cosine annealing.

After 50 rounds of training, our ConvNext-Tiny network with dual attention has an accuracy of 66.97% in the DeepFashion validation set; thus, it can be seen that the training of the model has a good convergence.

4.4. Style Classification Network

In our work, our style classification network comes from the migration and learning of a multi-label classification network. First, we train the original ConvNeXt-Tiny network on the Style10 data set for 100 rounds. The learning rate of the style classification network and the configuration of the optimizer are consistent with that of the category classification network. Then, we continued to train the ConvNeXt-Tiny network with dual attention on the Style10 data set for 100 rounds. Their accuracy on the validation set is 65.98% and 66.17%, separately. As shown in Figures 4 and 5, the effect of the confusion matrix for the original ConvNeXt-Tiny network and the ConvNeXt-Tiny network with dual attention was validated on the test set of the Style10 data set, respectively. It can be seen that the accuracy of the detection has been improved.

Finally, with the help of the model that trained 50 rounds on the DeepFashion data set, we used transfer learning to make the ConvNeXt-Tiny network with dual attention train 50 rounds on the Style10 data set. Finally, the accuracy rate of the verification set is 69.66%, and the network's ability to identify clothing style has been greatly improved. The effect of the confusion matrix validated on the test set for the ConvNeXt-Tiny network with dual attention performing transfer learning is shown in Figure 6, and it can be seen that the training of the model has a good convergence effect.

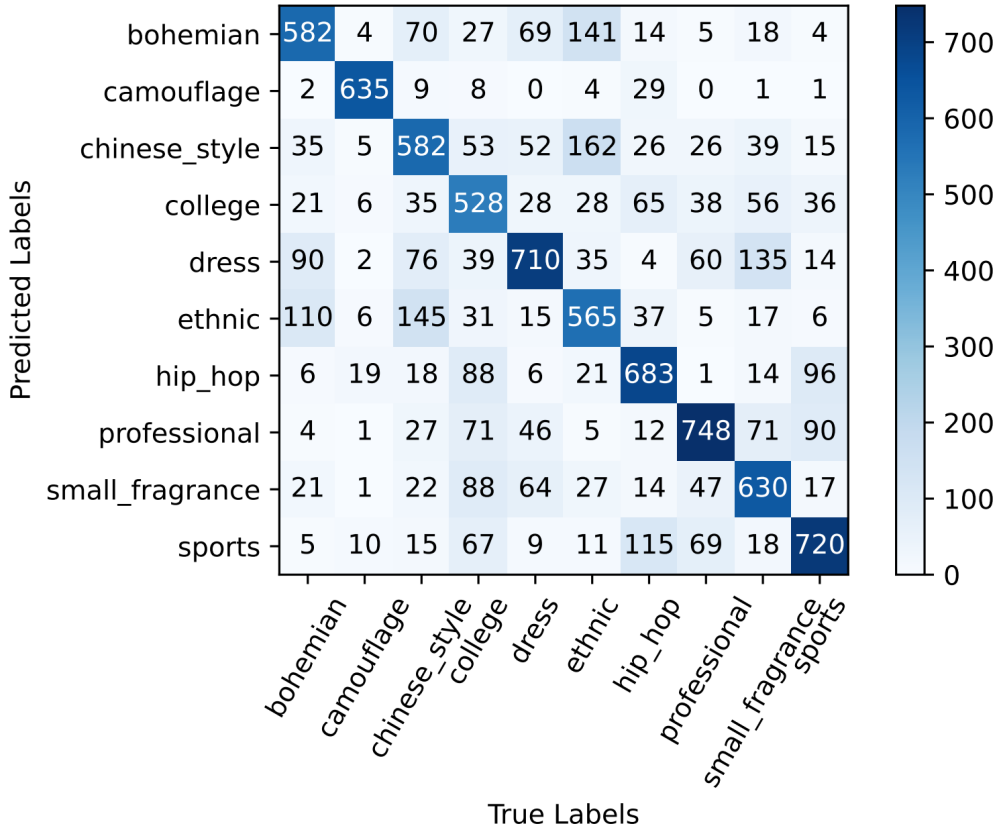

Figure 4. The confusion matrix ConvNeXt-Tiny network on the Style10 test set.

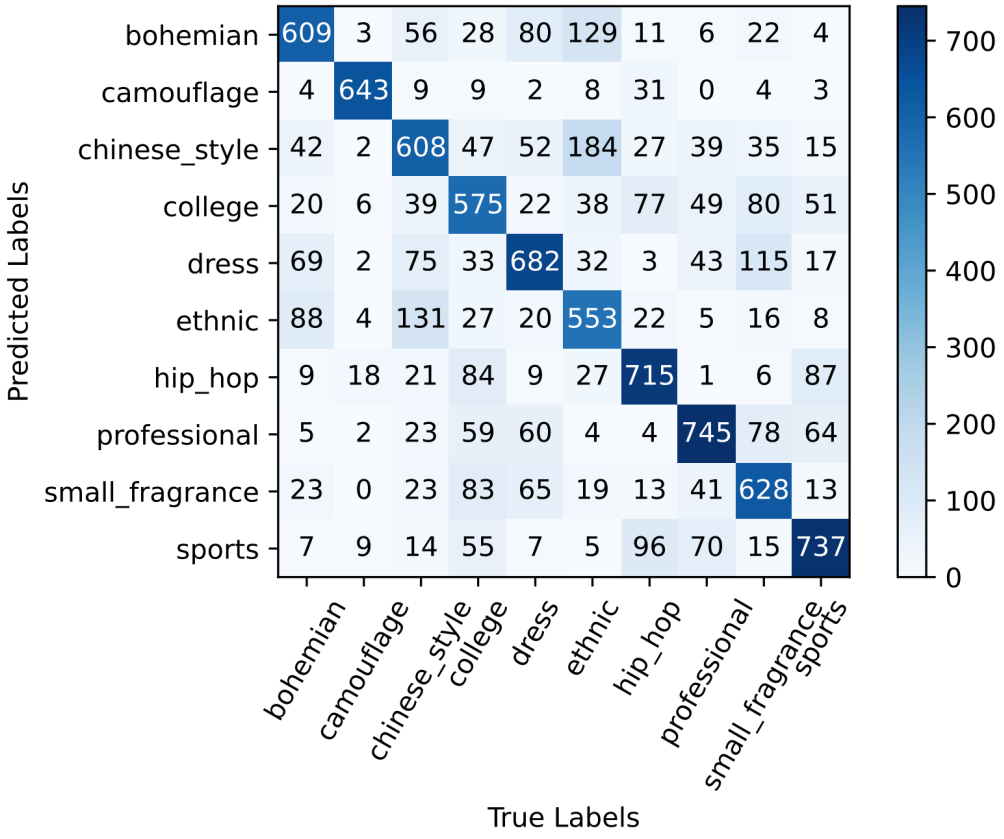

Figure 5. The confusion matrix of ConvNeXt-Tiny network with dual attention on the Style10 test set.

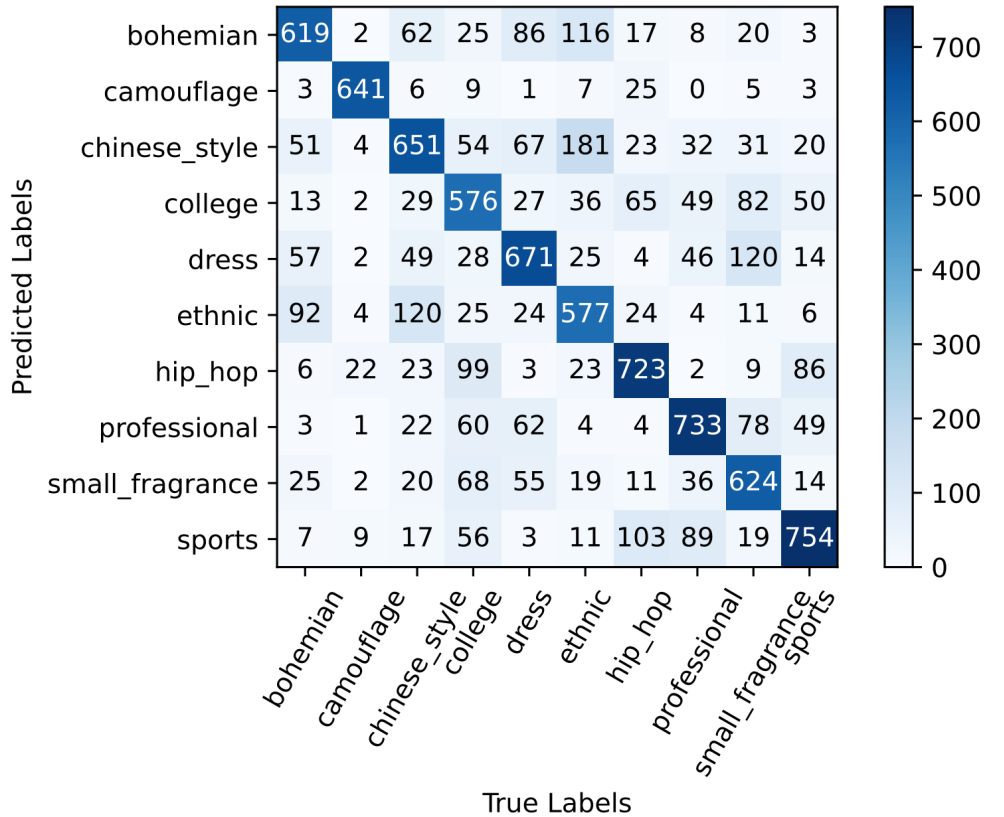

Figure 6. The confusion matrix of transfer learning of ConvNeXt-Tiny Network with dual attention on the Style10 test set.

We further compare our method with existing methods on the Style10 val set. Table 3 compares the accuracy of the different light-weight methods on the validation set and the running time of predicting an image separately. The results show that our method has a good balance of rate and time.

Table 3. Comparision of results on the Style10 val set.

Methods	Precision	CPU
VGG19	59.60%	1224 ms
GoogleNet	63.41%	**185 ms**
Resnet50	65.67%	549 ms
DenseNet201	65.59%	771 ms
SWIN-T	64.75%	744 ms
ConvNeXt-T(baseline)	65.98%	635 ms
Ours w/o transfer	66.17%	752 ms
Ours w/ transfer	**69.66%**	752 ms

4.5. Result

With the clothing category classification, we train the ConvNeXt network with dual attention with the DeepFashion data set. Within 25 epochs, our network converges rapidly and tends to continue to decline at the end of the training. The accuracy rate on the validation set is also gradually improved in the process of training, reaching the accuracy rate of about 66.97% at the end of the training. The reason why it does not exceed 0.8 is that the data classification in the DeepFashion data set is not clear enough, and there are often hybridized situations. We can see that each image is predicted correctly, but the confidence level is too low. Part of the reason is that the number of training iterations is too few and the data set is not regular enough. However, on the whole, the clothing classification network designed by us can converge well and has good recognition robustness after a short period of 50 rounds of training, so as to achieve the goal we expect. We saved the model trained on the DeepFashion data set for subsequent transfer learning.

We trained ConvNeXt-Tiny and ConvNext-Tiny with dual attention in the Style10 data set for 100 epochs. The validation set accuracy for ConvNeXt-Tiny was 65.98% and for ConvNeXt-Tiny with dual attention was 66.17%. The performance is improved with the addition of two attention modules. Then, we used the model trained for 50 rounds on DeepFashion and made ConvNext-Tiny with dual attention network train for 50 rounds on Style10 by means of transfer learning. It can be seen from Table 3 that there is a very obvious improvement. With the help of transfer learning, ConvNeXt-Tiny with dual attention achieved 69.66% accuracy on the Style10 validation set.

However, our proposed ConvNeXt network with dual attention still achieves a high accuracy in the actual detection. We have tested and verified 10 images of different clothing types. Below each image is the clothing category name predicted by the network and the confidence level of the category.

With the clothing style classification, we select Style10 data set to train the model and record the change of the validation set accuracy value during the training process. It can be seen that the ConvNeXt network with dual attention using transfer learning is fast and effective in terms of convergence speed and degree; it shows that our assumption of using clothing types to transfer training clothing style is correct, and it can be used to improve performance in different aspects in the future. In the second place is the ConvNeXt network with a dual attention classification network, which combines ConvNeXt network and dual attention. This network does not use transfer learning in the training process, but its effect is better than ConvNeXt network without transfer training.

From this, we can see that the purpose of using clothing types to transfer training clothing style can be achieved, and the effect is very ideal. It shows that the relationship between the fine-grained attributes of clothing and the whole style is close. As long as

the fine-grained attributes can be reasonably mined, an ideal result can be achieved. The excellent performance of ConvNeXt network with a dual attention structure also confirms the rationality of feature combination. It can be predicted that the effect of transfer learning on the ConvNeXt network with dual attention is better than that of the ConvNeXt network with dual attention that did not use transfer learning.

5. Conclusions

The classification of clothing categories and styles has always been a problem to be optimized in the field of intelligent clothing, which is closely related to people's demand for beauty appreciation and pursuit of the frontier, and it is also an industrial upgrading to change the old concept of the clothing industry. In online one-to-one services, we need to ensure an accurate classification of semantics and images to help users have a good service experience. In this paper, we propose a high accuracy classification method in the complex image background. We design ConvNeXt network with dual attention to improve the accuracy of clothing classification. In the process of model parameter training, the clothing attributes are refined gradually, which makes the parameters have strong clothing expression ability. Then, through transfer learning, the model parameters are fine-tuned, so that the fine-grained expression of parameters can be closer to the style attributes of clothing, so as to strengthen the internal relationship between categories and styles and achieve the goal of coarse-grained style classification. In order to achieve the requirements of more accurate prediction, we design a dual attention network structure to enhance the performance of the model. By incorporating different attention mechanisms in the shallow and deeper layers, respectively, we enhance the model control of global information, and the final image representation is enhanced. The results show that the proposed model and classification method can effectively predict the types and styles of clothing and exceed existing algorithms in models with similar volumes. For more classification requirements that may appear in the future, our model and method are also forward-looking and have flexible diversity regarding processing ability.

Author Contributions: Conceptualization, X.C. and Y.D.; methodology, C.D. and Y.D.; validation, H.L.; formal analysis, X.C. and H.L.; investigation, G.T.; resources, H.L. and Y.D.; data curation, G.T.; writing—original draft preparation, G.T. and H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Singh, M.; Katiyar, A.K. Consumers buying behavior towards Online Shopping—A study of Literature Review. *Asian J. Manag.* **2018**, *9*, 490–492. [CrossRef]
2. Changchenkit, C. The Social Media Exposure and Online Clothes Buying Behavior in Thailand. In *Proceedings of the International Academic Conferences*; International Institute of Social and Economic Sciences: London, UK, 1 January 2018.
3. Yang, F.; Li, X.; Huang, Z. Buy-online-and-pick-up-in-store Strategy and Showroom Strategy in the Omnichannel Retailing. In *Advances in Business and Management Forecasting*; Emerald Publishing Limited: York, UK, 6 September 2019.
4. Sun, L.; Aragon-Camarasa, G.; Rogers, S.; Stolkin, R.; Siebert, J.P. Single-Shot Clothing Category Recognition in Free-Configurations with Application to Autonomous Clothes Sorting. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2017), Vancouver, BC, Canada, 24–28 September 2017.
5. Hu, J.; Kita, Y. Classification of the category of clothing item after bringing it into limited shapes. In Proceedings of the 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), Seoul, Korea, 3–5 November 2015; pp. 588–594.
6. Chen, L.; Han, R.; Xing, S.; Ru, S. Research on Clothing Image Classification by Convolutional Neural Networks. In Proceedings of the 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Beijing, China, 13–15 October 2018.

7. Pan, H.; Wang, Y.; Liu, Q. A Part-Based and Feature Fusion Method for Clothing Classification. In Proceedings of the Pacific Rim Conference on Multimedia, Xi'an, China, 15–16 September 2016; Springer: Cham, Switzerland, 2016.
8. Sun, F.; Xu, P.; Ding, X. Multi-core SVM optimized visual word package model for garment style classification. *Clust. Comput.* **2019**, *22*, 1–7. [CrossRef]
9. Geršak, J. 1–Clothing classification systems. *Des. Cloth. Manuf. Process.* **2013**, 1–20.
10. Willimon, B.; Birchfield, S.; Walker, I. Classification of clothing using interactive perception. In Proceedings of the IEEE International Conference on Robotics & Automation, Shanghai, China, 9–13 May 2011.
11. Willimon, B.; Walker, I.; Birchfield, S. A new approach to clothing classification using mid-level layers. In Proceedings of the IEEE International Conference on Robotics & Automation, Karlsruhe, Germany, 6–10 May 2013.
12. Yamazaki, K.; Inaba, M. Clothing classification using image features derived from clothing fabrics, wrinkles and cloth overlaps. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013.
13. An, Y.; Rao-fen, W. Clothing Classification Method Based on CNN Improved Model Network. *Tech. Autom. Appl.* **2019**.
14. 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 CVPR, San Francisco, CA, USA, 17 December 2018.
15. Chen, K.; Yao, L.; Zhang, D.; Wang, X.; Chang, X.; Nie, F. A semisupervised recurrent convolutional attention model for human activity recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 1747–1756. [CrossRef] [PubMed]
16. Luo, M.; Chang, X.; Nie, L.; Yang, Y.; Hauptmann, A.G.; Zheng, Q. An adaptive semisupervised feature analysis for video semantic recognition. *IEEE Trans. Cybern.* **2017**, *48*, 648–660. [CrossRef]
17. Zhang, D.; Yao, L.; Chen, K.; Wang, S.; Chang, X.; Liu, Y. Making sense of spatio-temporal preserving representations for EEG-based human intention recognition. *IEEE Trans. Cybern.* **2019**, *50*, 3033–3044. [CrossRef] [PubMed]
18. Hong, Q.; Zhang, J. Introduction to Garment Style Classification. *J. Chengdu Text. Coll.* **2005**.
19. Bossard, L.; Dantone, M.; Leistner, C.; Wengert, C.; Gool, L.V. Apparel Classification with Style. In *Proceedings of the Asian Conference on Computer Vision-Volume Part IV*; Springer: Berlin/Heidelberg, Germany, 2012.
20. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
21. Kiapour, M.H.; Han, X.; Lazebnik, S.; Berg, A.C.; Berg, T.L. Where to Buy It: Matching Street Clothing Photos in Online Shops. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015.
22. Qiang, C.; Huang, J.; Feris, R.; Brown, L.M.; Yan, S. Deep Domain Adaptation for Describing People Based on Fine-Grained Clothing Attributes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 15 October 2015.
23. Li, G.; Feng, J. Unsupervised image clustering based on attribute method. *Microcomput. Its Appl.* **2012**.
24. Vaccaro, K.; Shivakumar, S.; Ding, Z.; Karahalios, K.; Kumar, R. The Elements of Fashion Style. In Proceedings of the 29th Annual ACM Symposium on User Interface Software and Technology, Tokyo, Japan, 16–19 October 2016; pp. 777–785. [CrossRef]
25. Seo, Y.; shik Shin, K. Image classification of fine-grained fashion image based on style using pre-trained convolutional neural network. In Proceedings of the IEEE International Conference on Big Data Analysis, Shanghai, China, 9–12 March 2018.
26. Liu, T.; Rubing Wang, J.C.; Han, S.; Yang, J. Fine-Grained Classification of Product Images Based on Convolutional Neural Networks. *Adv. Mol. Imaging* **2018**, *8*, 69. [CrossRef]
27. Hidayati, S. Clothing genre classification by exploiting the style elements. In Proceedings of the ACM Multimedia Conference, Nara, Japan, 29 October–2 November 2012.
28. Di, W.; Wah, C.; Bhardwaj, A.; Piramuthu, R.; Sundaresan, N. Style Finder: Fine-Grained Clothing Style Recognition and Retrieval. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013.
29. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
30. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
31. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Comput. Sci.* **2014**.
32. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *arXiv* **2014**, arXiv:1409.4842.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New York, NY, USA, 1 June 2016; pp. 770–778.
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
35. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 10012–10022.
36. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *3*, 1–40.

37. Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; Tang, X. Large-Scale Fashion (DeepFashion) Database. Available online: http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html (accessed on 18 June 2016).
38. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 7–14 September 2018; pp. 801–818.