IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 4, 2020, accepted February 25, 2020, date of publication March 9, 2020, date of current version March 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2979164

Clothing Attribute Recognition Based on RCNN Framework Using L-Softmax Loss

JUN XIANG, TIANTIAN DONG, RURU PAN, AND WEIDONG GAO
Key Laboratory of Eco-textiles, Jiangnan University, Wuxi 214122, China

Corresponding author: Weidong Gao (gaowd3@163.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0309200, in part by the Fundamental Research Funds for the Central Universities under Grant JUSRP41804, and in part by the Postgraduate Research and Practice Innovation Program of Jiangnan University under Grant JNKY19_028.

ABSTRACT Due to the significant potential values in commercial and social applications, clothing image recognition has recently become a research hotspot, among which clothing attribute recognition is an important content. However, the large variations in the appearance and style of clothing and the image's complex forming conditions make the task challenging. Moreover, a generic treatment with deep convolutional neural networks cannot provide an ideal solution. Instead of using CNNs for classification, we proposed a novel approach based RCNN framework for the recognition task. Firstly, we apply the modified selective search algorithm to extract the region proposal. Then, the Inception-ResNet V1 model with L-Softmax is employed to represent images and identify their categories. After Soft-NMS, we use a simple neural network to correct the boundary of region box. To evaluate the performance of the framework, a dataset including about 100,000 shirt images was built. The experimental result show that our proposed framework achieved promising overall labeling rate, precision and recall of 87.77%, 73.59% and 83.84%. In addition, comparative experiments demonstrate the superiority of the proposed framework.

INDEX TERMS Image analysis, feature extraction, neural network, object detection, learning systems.

I. INTRODUCTION

With the popularity of e-commerce and the accumulation of large amounts of image data, how to extract useful information from huge datasets quickly becomes a research hotspot. As the largest category in the e-commerce industry, the clothing category is more urgently demanding this technology. By recognizing and analyzing the attribute related to fashion style, and combining the price of goods, sales volume and consumer reviews to build models, consumers' preferences and fashion trend can be predicted. The predicted results can not only assist sellers in making marketing plans and business decision, but also provide inspiration for designers. Clothing image attributes recognition, which is one of the most important work in the predicting tasks, is recently getting more and more attention in the computer vision community. This paper mainly studies clothing attribute recognition based on deep convolutional neural network [1]–[7].

Clothing attribute recognition is part of the wider task of visual object recognition. It poses significant challenges because of the rich clothing elements and variations in design features. Previous studies described clothing only by image-pixel-based low features, such as theme color distributions and textures, for the purpose of similarity search, without capturing the clothing features of visual appearance to recognize and classify the clothing attribute as well. To address these issues, this study proposed a RCNN-based [8] method to recognize the attributes of clothing, including collars, sleeves, front buttons, lengths and waists.

Beyond colors and patterns, style is an important dimension for describing clothing. However, the style of clothing is generally determined by some local attributes of the clothing such as the collars and sleeves. This work focus on a single category of clothing: women's shirts. As such, the category within this domain are fine-grained in nature, and they differ based on the presence of visual style elements or attributes. In this paper, a novel clothing attribute recognition technique is proposed to identify some labels about the style of clothing. The proposed system consists of three key modules. The first is to apply modified search selective algorithm [9] to generate category-independent region proposals that are used as candidate sets to the detector. The second module is a

The associate editor coordinating the review of this manuscript and approving it for publication was Imran Sarwar Bajwa.

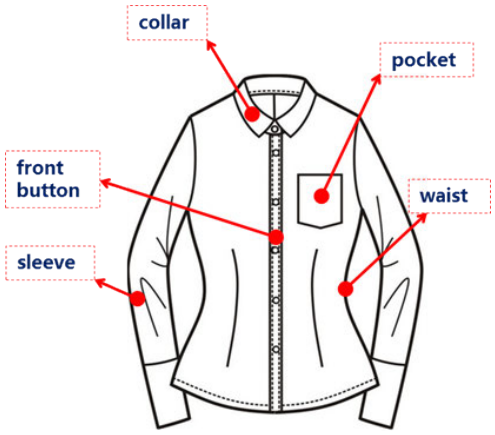


FIGURE 1. The attributes of woman shirt to be recognized.

deep sparse CNN [7] with L-Softmax loss [10] for extracting feature vectors of fixed-length from each proposed region and classify them in each category. The third module is for boundary correction of candidate regions.

The contributions of this paper can be summarized in two aspects. First, we improve the color similarity and texture similarity calculation methods in the search selective algorithm, which makes it more efficient to extract key part of the clothing. Second, we obtain a sparse convolutional neural network to extract features from candidate region and classify them with a classifier that use L-Softmax loss [10]. The proposed framework has multiple potential application, such as style-based navigation and retrieval, automatic style labeling of clothing images.

The rest of this paper is structured as follows. Section 2 review the related works in clothing image recognition. Section 3 first introduce the overview of the proposed framework, then present the specific steps of the proposed method. Experiments and comparative analyses are described in Section 4. Finally, section 5 conclude the whole study.

II. RELATED WORKS

Due to the increasingly large business value of the online-shopping and fashion industry, automatic clothing image analysis has received great attention. So there has been a great deal of studies in last few years on the subject of clothing recognition.

However, many research works focus on clothing segmentation, in the purpose of which is to solve the localization the clothing regions in the image. Some clothing extraction methods extracted clothing subjects by distinguishing the foreground and background. Hu *et al.* [11] proposed a clothing segmentation method using background and foreground estimation based on Constrained Delaunay Triangulation(CDT). Wu *et al.* [12] presented an automatic clothing extraction algorithm by combining efficient graph-based image segmentation and estimating the foreground and background of the image. Based on a large number of part detectors, Weber and Bauml *et al.* [13] proposed a segmentation approach which is able to separately segment a person's upper and lower clothing regions. Wang and Ai [14] proposed a layout model formulated as Markov Network which combines the blocking relationship to hunt for an approximately optimal clothing layout for group people.

In the past, clothing image retrieval were proposed to address the task of searching similar clothing offered by a query clothing image. To model the discrepancy between user photographs and clothing images from online shopping stores, street-to-shopping image retrieval task has been explored. Based on low-level feature descriptor for texture, shape and color, Vittayakorn *et al.* [15] presented a method to study outfit similarity on the runway and in the real-word settings. The works in [16] and [15], [17]–[19] focused on how to measure the similarity of clothing images, without considering specific instances that correspond to the unique features represented by clothing attributes. Recently, Li *et al.* [20] apply a multi-task learning framework for clothing image retrieval. This method utilizes a multi-task learning framework to guide the model to learn image representations for image retrieval. Chen and Liu [21] proposed a deep learning based method for clothing image retrieval and classification. The two methods which are based on deep learning achieve good performance on their database. This also proves the superiority of deep learning on large-scale dataset.

Further, clothing image recognition and analysis study were performed to represent and classify the clothing category. Li *et al.* [22] proposed a recognition framework that is based on multiple sources of features and ELM neural networks. Hidayati *et al.* [23] presented a novel approach for automatically classifying clothing genres using the visually differentiable style elements. Yang and Yu [24] proposed a video content analysis framework to locate the people's position, frame the clothing regions, and identify the category of the clothing. Hidayati *et al.* [25] also presented a method that automatically recognize visual style elements and ingredients for representing the trends of certain season fashion.

Different from the aforementioned recognition methods, this paper study the local attributes of clothing, such as the positioning and identification of the clothing collar. In the works of clothing local recognition and positioning, most approaches are limited to the level of pixel or super-pixel. Simo-Serra *et al.* [26] proposed a method to segment different clothing worn by a person by using a Conditional Random Field (CRF), which took the dependencies between human pose and clothing into account. After this, Chen *et al.* [27] and Yamaguchi *et al.* [28] earned style rules model which is based on CRF combined with clothing attributes. Chen proposed several attributes of clothing to describe the upper wear. They first estimated the pose of person in the image, and then extracted features from the specific regions for recognition. This method took pose variation into consideration, however, it was limited for the image of straight-frontal pose. Although having achieved certain success in clothing analysis, the aforementioned methods based on low-level feature depend heavily on feature extraction engineering which leads to their limitation.

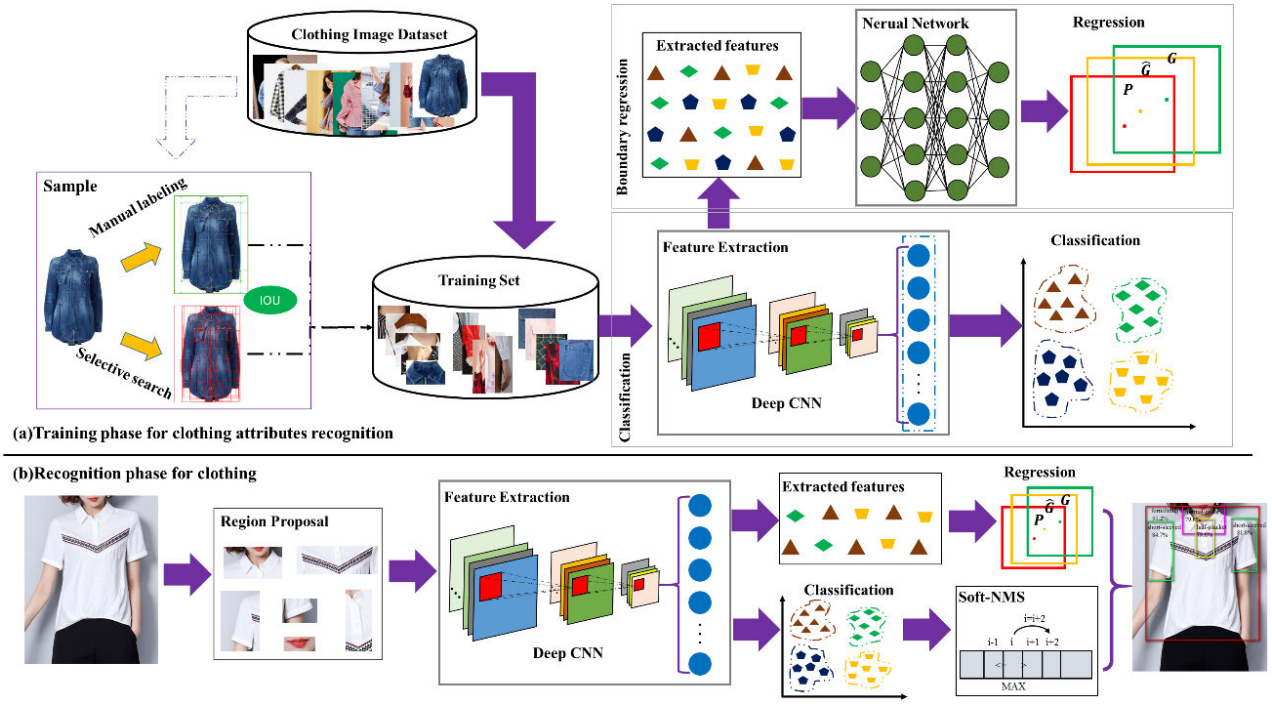


FIGURE 2. Overview for the proposed framework for clothing attributes recognition.

Recently, significant progress has been made on image analysis and recognition by moving from early low-level feature based algorithms to deep learning based frameworks. Deep representation has been widely employed in a variety of visual task, the most common of which are image classification [1], [3]–[7], object detection [8], [29], [30], image segmentation [31]–[34] pix-wise image labeling [32], [35] and human centric analysis [36], [37]. Sun and Liu [38] presented a methods based on Faster R-CNN and multi-task learning for clothing attribute recognition. Gu *et al.* [39] recently proposed an approach for clothes keypoints localization and attribute recognition based on prior knowledge of clothing and humans. The two approaches use the Mask-RCNN and Faster R-CNN object detection framework, which has a poor recognition effect on small target.

This study regards clothing attributes recognition as an object detection task, which includes regional positioning and classification. Unlike above method, this work considers more practical condition by training the model on the dataset with many pose. In addition, the proposed method describes the clothing attributes by extracting regional features.

III. METHODS

Figure 2 shows the overview of the proposed framework for clothing attributes recognition, which consists of the following four components: 1) region proposal; 2) feature extraction; 3) attribute category prediction; 4) boundary correction. A shirt image dataset named SAR which contains about 100,000 image of shirt from the online store, was built to train the recognition model. Firstly, we apply selective search algorithm to extract region proposal, and use the annotation data to build a classification models. Then, a deep sparse convolutional neural network based framework is employed for extracting feature. Further, L-SoftMax is used to predict the categories of regions according to the extracted features. Finally, a neural network is adopted for boundary correction. The figure shows the training phase and recognition phase for clothing attribute recognition of the proposed method. The four components in proposed framework for clothing image recognition will be describe in this section respectively.

A. REGION PROPOSAL

For proposed framework, the effect of the regional proposal will directly affect the accuracy of subsequent recognition. In common with RCNN [8], this study also use selective search (SS) algorithm to extract candidate region. SS algorithm was originally proposed by Uijlings *et al.* [9] for object recognition. First, a graph-based segmentation method is used to divide the image into many fragments, and then the small fragments are merged by the similarity between the regions to extract each region with possible object. The detailed steps of the SS algorithm for an input image are as follows,

1) Using the graph-based segmentation algorithm [40] to obtain a set of initial regions, denoted by $\boldsymbol{R} = \{r_1, r_2 \cdots r_n\}$.

2) Calculating similarities between neighbouring regions including color similarity, texture similarity, size similarity, and fit similarity, and then stored in the set $\boldsymbol{S}$;

3) Combining two neighbouring regions r_p and r_q with the highest similarity in the set $\boldsymbol{R}$ to form a new region r_t, and deleting the similarity between the neighbouring regions of r_p and r_q in the set $\boldsymbol{S}$;

4) Calculating the similarity between r_t and the neighbouring region, and store the value of the similarity into the set $\boldsymbol{S}$, and store the r_t in the set $\boldsymbol{R}$;

5) Iterating steps 3) and 4) until the set $\boldsymbol{S}$ is empty;

6) Filtering out possible candidate boxes by size.

Step 2 takes into account four local similarities in calculating the regional similarity, which are color similarity, texture similarity, size similarity, and fit similarity. In this study, we apply color moment to measure the color similarity between two regions. To enhance the generalization of color features, we calculate the color moments of multiple color components to represent color feature of the region. The color component include the R, G, B (RGB color space), L, a, b (Lab color space), H, S, V (HSV color space), and the weighted gray value denoted by I. Color moment contains first-order moment (mean), second-order central moment (variance), and third-order central moment (slope), which are calculated as follows,

$$\mu = \frac{1}{n}\sum_{i=1}^{n} h_i \tag{1}$$

$$\sigma = \left[\frac{1}{n}\sum_{i=1}^{n}(h_i - \mu)^2\right]^{\frac{1}{2}} \tag{2}$$

$$s = \left[\frac{1}{n}\sum_{i=1}^{n}(h_i - \mu)^3\right]^{\frac{1}{3}} \tag{3}$$

where μ indicates the first-order moment—mean, σ indicates second-order central moment—variance, s indicates the third-order central moment—slope; n represents the number of pixels contained in the region, and h_i denotes the color value at position i. By calculating the color moment of each channel, a 1×30 vector represented by $C_i = \{c_i^1, c_i^2 \cdots c_i^n\}$ can be obtained for each region. If c_p and c_q respectively represent the extracted color feature vectors of two neighbouring regions, the similarity between them is calculated as follows.

$$S_{color}(c_p, c_q) = \sqrt{\sum_{i=1}^{n}\left(c_p^i - c_q^i\right)^2} \tag{4}$$

The adjacent regions (r_p, r_q) are combined to form a new region r_t, and the color feature vector of the new region is represented by C_t.

$$C_t = \frac{size\left(C_p\right) \times C_p + size\left(C_q\right) \times C_q}{size\left(C_p\right) + size\left(C_q\right)} \tag{5}$$

This study apply Local Binary Pattern (LBP) [41] with a sampling number of 8 and a sampling radius of 1 to represent the image texture feature. For each color channel we extract a histogram using a bin size of 20. This leads to a texture histogram $T_i = \{t_i^1, t_i^2 \cdots t_i^n\}$ for each region ri with dimensionality n = 60 when three color channels (R, G and B in RGB color space) are used. Texture histograms are normalized using the L1-norm. Similarity is measured using histogram intersection. Texture histograms are efficiently propagated through the hierarchy.

$$s_{texture}(r_p, r_q) = \sum_{k=1}^{n} \min(t_p^k, t_q^k) \tag{6}$$

$$T_t = \frac{size\left(T_p\right) \times T_p + size\left(T_q\right) \times T_q}{size\left(T_p\right) + size\left(T_q\right)} \tag{7}$$

To encourage small regions to merge early, the size of the regions are involved in calculating the regional similarity. Giving the small regions more weight, can ensure that the image is multi-scale merged at each location. $S_{size}(r_p, r_q)$ is defined as the fraction of the image that r_p and r_q jointly occupy. In addition, the method take the degree of fit into account. To keep the measure fast, the proposed only apply the size of the regions and of the containing boxes. If we define B_{pq} to be the tight bounding box around r_p and r_q, $S_{fill}(r_p, r_q)$ can be denoted by formula 9.

$$s_{size}(r_p, r_q) = 1 - \frac{size(r_p) + size(r_q)}{size(image)} \tag{8}$$

$$s_{fill}(r_p, r_q) = 1 - \frac{size(B_{pq}) - size(r_p) - size(r_q)}{size(image)} \tag{9}$$

The final similarity measure is a combination of above four:

$$s(r_p, r_q) = a_c s_{color} + a_t s_{texture} + a_s s_{size} + a_f s_{fill} \tag{10}$$

where a $\in \{0,1\}$ denotes if the similarity measure is used or not.

B. FEATURE EXTRACTION

Early studies have demonstrated that the feature activations of CNNs included by input image can serve as image representation or visual signatures. The application of this mid-level feature or image representation indicates impressive improvement on the task of object recognition, image retrieval, image recognition, image analysis, image classification and others. Deeper convolutional neural network have been hotspot to large advances in image recognition and analysis performance in recently years. The Inception-Architecture and residual connection have achieved good performance with relatively low computational cost. The proposed framework mainly contains three schema—Stem, Inception-ResNet, Reduciton. The stem schema performs initial convolution on the input. The Inception-ResNet module abstract the image feature without changing the size of grid. The function of Reduction modules not only extracts and abstracts the feature, but also compresses the size of grid.

The proposed framework extract a 1792-dimensional feature vector from each region proposal using the TensorFlow [31] implementation of the Inception-ResNet-v1 described by Szegedy *et al.* [7]. Features are computed by forward propagating a mean-subtracted 299×299 RGB image. Table 1 shows the stem of Inception-ResNet-v1 network used in proposed framework. This stem convert the RGB image of $299 \times 299 \times 3$ into a $35 \times 35 \times 256$ grid through a convolutional

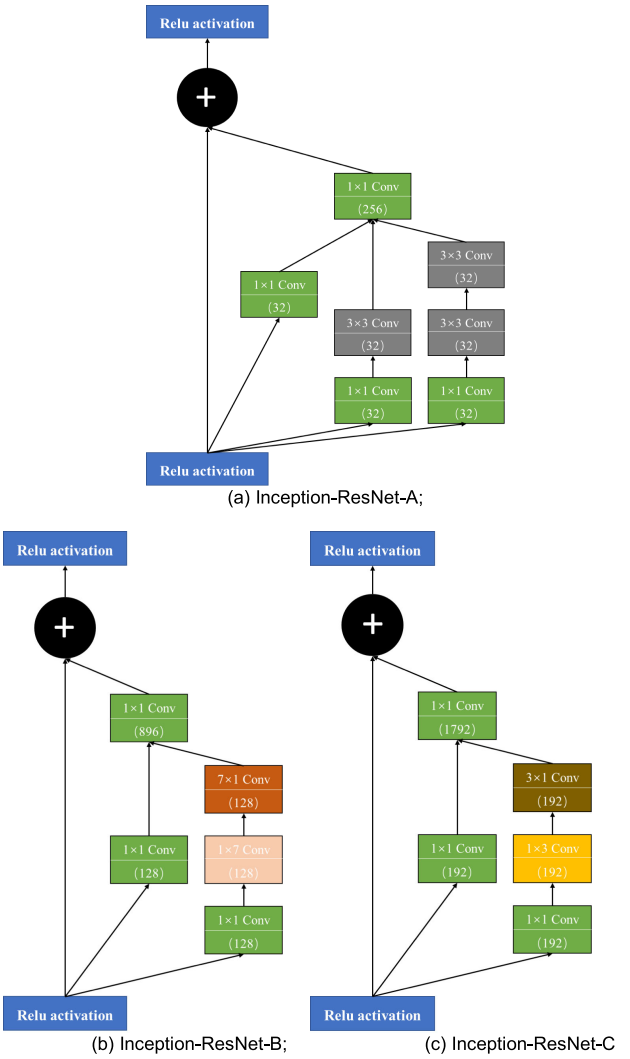


FIGURE 3. Three Inception-ResNet modules used in Inception-ResNet-V1.

network with 6 layers of convolution and 1 layer of Max-pooling. Figure 3 shows three Inception-ResNet modules used in the CNN network. The Inception-ResNet-A shown in Figure 3(a) is the schema for 35×35 grid with both inputs and outputs of 256 dimensions. Then, a dimension reduction module that is called Reduciton-A shown in Figure 4(a) is applied to reduce the size of grid from $35 \times 35 \times 256$ to $17 \times 17 \times 896$. After that, 10 Inception-ResNet-B modules shown in Figure 3(b) are used to abstract features for the grid of 17×17. Similarly, a dimension reduction schema called Reduction-B shown in Figure 4(b) is connected to reduce the size of grid from $17 \times 17 \times 896$ to $8 \times 8 \times 1792$. Continuously, 5 Inception-ResNet-C modules shown in Figure 3(c) are used for abstracting features for the grid of 8×8. Finally, the framework use an Average-Pooling layer to extract theme features.

To prevent model over-fitting and enhance adaptability during training, the proposed method apply a mechanism called LSR (Label Smoothing Regularization) [42] for

TABLE 1. The stem of Inception-ResNet-v1.

Type	Patch size/stride	Output size
Input	299×299×3	
Convolution	3×3/2	149×149×32
Convolution	3×3/1	147×147×32
Convolution	3×3/1	147×147×64
Maxpool	3×3/2	73×73×64
Convolution	1×1/1	73×73×80
Convolution	3×3/1	71×71×192
Convolution	3×3/2	35×35×256

TABLE 2. The outline of the network architecture.

Type	Patch size or remarks	Output size
Input	299×299×3	
Stem	As in table 1	35×35×256
5×Inception-ResNet-A	As in figure 3(a)	35×35×256
Reduction-A	As in figure 4(a)	17×17×896
10×Inception-ResNet-B	As in figure 3(b)	17×17×896
Reduction-B	As in figure 4(b)	8×8×1792
5×Inception-ResNet-C	As in figure 3(c)	8×8×1792
Average Pooling	8×8	1×1×1792

encouraging the model to be less confident. Consider a distribution over labels $u(k)$, independent of the training example x, and a smoothing parameter λ. For a training example with ground-truth label y, the label distribution $q(k \mid x) = \delta_{k,y}$ is replaced with $q'(k \mid x)$ described by formula 11, where $\delta_{k,y}$ is Dirac delta.

$$q'(k|x) = \lambda u(k) + (1-\lambda)\delta_{k,y} \tag{11}$$

$$\delta_{k,y} = \begin{cases} 0 & k = y \\ 1 & k \neq y \end{cases} \tag{12}$$

The $q'(k \mid x)$ is a mixture of the fixed distribution $u(k)$ and the original ground-truth distribution $q(k \mid x)$, with weights λ and $(1-\lambda)$, respectively. The author suggest to use the prior distribution over labels as $u(k)$. In this study, we assume that the data obeys uniform distribution $u(k) = 1/K$ where K indicates the number of classification categories. If we bring $u(k) = 1/K$ to equation (11):

$$q'(k|x) = \frac{\lambda}{K} + (1-\lambda)\delta_{k,y} \tag{13}$$

If we set $\lambda = 0$, $q'(k \mid x)$ is the one-hot encoding. In the study on shirt attributes recognition with $K = 19$ classes, we use $u(k) = 1/19$ and $\lambda = 0.1$.

C. ATTRIBUTES CATEGORY PREDICTION

The proposed framework employ Softmax classification to predict the class of input based on 1792-dimensional feature vector from the average-pooling layer. The Softmax is

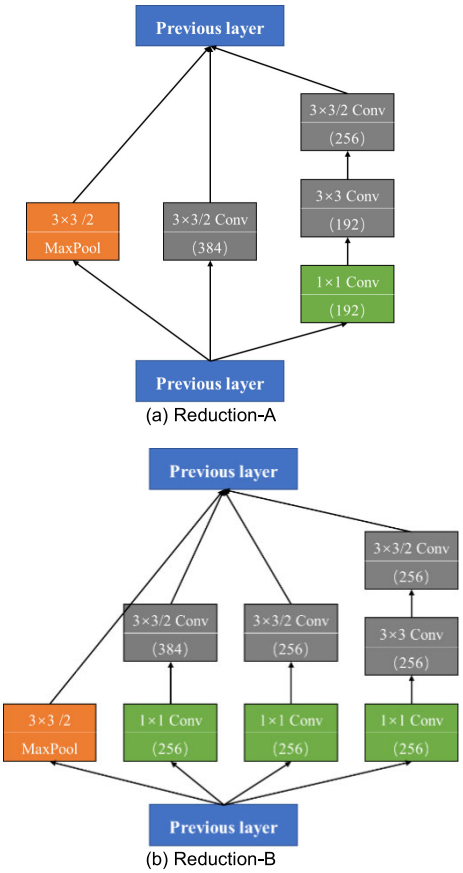


FIGURE 4. Two reduction modules used in Inception-ResNet-V1.

defined as follows:

$$S_i = \frac{e^{V_i}}{\sum_i^C e^{V_i}} \tag{14}$$

where V_i indicates the output of the front stage output unit, i represent the index of class, and S_i indicates the possibility that V_i belongs to C_i. The number of total class is C.

In convolutional neural networks (CNNs), cross-entropy loss with softmax is the most common used supervision component. Despite its popularity and simplicity, the component doesn't explicitly encourage discriminative learning of features. This study apply a generalized large-margin softmax (L-Softmax) loss [10] which explicitly encourages inter-class separability and intra-class compactness between learned features.

For binary classification, suppose x is a sample from class 0. To classify x correctly, the original softmax is forces $\mathbf{W}_0^T \cdot \mathbf{x} = \left\|\mathbf{W}_0^T\right\| \|\mathbf{x}\| \cos\theta_0 > \mathbf{W}_1^T \cdot \mathbf{x} = \left\|\mathbf{W}_1^T\right\| \|\mathbf{x}\| \cos\theta_1$. L-Softmax sets stringent restrictions as shown in formula 15.

$$\mathbf{W}_0^T \cdot \mathbf{x} \geq \left\|\mathbf{W}_\mathbf{0}^T\right\| \|\mathbf{x}\| \cos(m\theta_0) > \mathbf{W}_1^T \cdot \mathbf{x} \tag{15}$$

where m is a positive integer. As shown in above formula, $\mathbf{W}_0^T \cdot \mathbf{x} > \left\|\mathbf{W}_1^T\right\| \|\mathbf{x}\| \cos\theta_1$ and $\left\|\mathbf{W}_\mathbf{0}^T\right\| \|\mathbf{x}\| \cos(m\theta_0) > \mathbf{W}_1^T \cdot \mathbf{x}$ have to hold. Therefore, the new classification loss is a more restrictive requirement to correctly classify $\mathbf{x}$, by producing a more rigorous decision boundary for class 0.

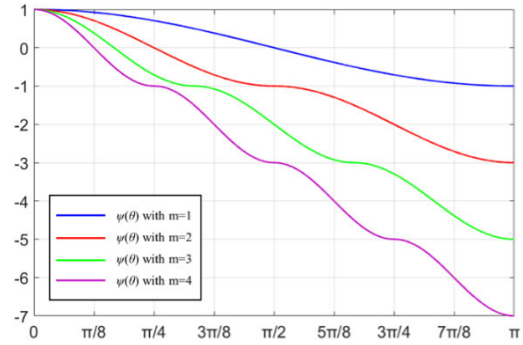


FIGURE 5. $\psi(\theta)$ for L-Softmax loss.

An input $\mathbf{x}$ with label y, after label smoothing, we can obtain its soft label $\{y_1, y_2 \ldots y_k\}$ with K labels. The Large-Margin Softmax loss is define as

$$L = \sum_{i=1}^{K} -\log \frac{e^{y_i \|\mathbf{W_i}\| \|\mathbf{x}\| \psi(\theta_i)}}{e^{\|\mathbf{W}_y\| \|\mathbf{x}\| \psi(\theta_y)} + \sum_{j \neq y} e^{\|\mathbf{W}_j\| \|\mathbf{x}\| \cos(\theta_i)}} \tag{16}$$

$$\psi(\theta) = (-1)^k \cos(m\theta) - 2k, \quad \theta \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}] \tag{17}$$

where m (previously mentioned) is a integer that is closely related to the classification margin. Larger m means the classification margin will be larger and the learning objective will be more difficult. Meanwhile $k \in [0, m-1]$ and k is an integer. Figure 5 show the $\psi(\theta)$ for L-Softmax loss with different m. When the m is equal 1, L-Softmax loss is the Softmax loss. The experimental results show that the model performs well when m = 4. So this study set $m = 4$.

For the convenience of forward and backward propagation, the proposer of L-Softmax loss suggests replacing $\cos\theta_j$ with $\frac{\mathbf{W}_j^T \cdot \mathbf{x}_i}{\left\|\mathbf{W}_j^T\right\| \|\mathbf{x}_i\|}$ and $\cos(m\theta_{y_i})$ with

$$\begin{aligned}\cos(m\theta_{y_i}) = {} & C_m^0 \cos^m \theta_{y_i} - C_m^2 \cos^{m-2} \theta_{y_i}(1 - \cos^2 \theta_{y_i}) \\ & + C_m^4 \cos^{m-4} \theta_{y_i}(1 - \cos^2 \theta_{y_i})^2 + \cdots \\ & + (-1)^n C_m^{2n} \cos^{m-2n} \theta_{y_i}(1 - \cos^2 \theta_{y_i})^n + \cdots\end{aligned} \tag{18}$$

where n is an integer and $2n \leq m$. The replacement is mainly used to get rid of θ. Then, we could perform derivation with respect of $\mathbf{W}$ and $\mathbf{x}$. It is also trivial to perform derivation with mini-batch input.

D. SOFT NON-MAXIMUM SUPPRESSION

As a vital part of the object detection pipeline, non-maximum suppression first sort all region boxes by their scores, than select the detection box with the maximum score by suppressing other boxes. The general method NMS would miss the object when an object region lies within the pre-defined overlap threshold. This study apply Soft-NMS [43] to do this

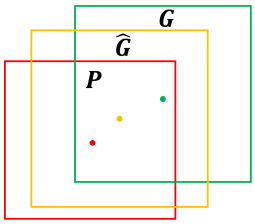


FIGURE 6. Schematic diagram of boundary correction.

work. The algorithm decays the detection scores of all objects as a continuous function of their overlap.

E. BOUNDARY CORRECTION

Boundary correction is a method to reduce the location errors. As shown in Figure 6, we assume that the green box $\boldsymbol{G}$ represents the group-truth bounding box, and the red box P represents the proposal bounding box. The purpose of boundary correction is to find a bounding box $\hat{\boldsymbol{G}}$ that is closer to the ground-truth box through a mapping relationship. However, the process of finding the mapping relationship is boundary-box regression. The proposed method apply a neural network with two fully connected layers. The input of boundary-box regression is the output 1792-dimensional feature vector from average-pooling layer. The components of regression include two kinds of transformations, translation transformation ($d_x(P)$ and $d_y(P)$) and scaling transformation ($d_w(P)$ and $d_h(P)$). The output of the algorithm is the location of prediction boundary box. The objective function of the regression can be described in formula 19.

$$d_*(P) = \mathbf{w}_*^T \cdot \mathbf{F}_P \tag{19}$$

where $\mathbf{w}_*^T$ indicates the set of parameters needed to learning, $* \in \{x, y, w, h\}$, $\mathbf{F}_P$ represent the feature vector of corresponding regions. Actually, the purpose of the regression is to minimize the gap between the predicted location and the ground-truth box location valued by (t_x, t_y, t_w, t_h). The loss function of the regression is described by

$$L_r = \sum_i^N (t_*^i - \mathbf{w}_*^T \cdot \mathbf{F}_{P^i})^2 \tag{20}$$

where N indicates the batch size of the training. The proposed method apply stochastic gradient descent algorithm to solve equation (21).

$$\mathbf{w}_* = \arg\min_{\mathbf{w}_*} \sum_i^N (t_*^i - \mathbf{w}_*^T \cdot \mathbf{F}_{P^i})^2 \tag{21}$$

IV. EXPERIMENTS

To optimize the parameters and strategies of selective search algorithm, we used number of boxes and MABO (which will be described in detail later) as the measure of region proposal. First, MABO and number of boxes at different scales and thresholds under a single strategy are compared to determine the optimization parameters. Then compare the results under different strategies (strategy of similarity summation) to determine the strategy used in this paper. In addition, due to your concerns, I found that the peer comparison experiments are lacking in the paper, so I added some comparative experiments, including YOLO [30], SSD [29], and the methods in your recommended papers.

A. DATASET

Although there are currently several open source clothing image databases, they are basically not suitable for this training task. To train the recognition model, a shirt image dataset named SAR with about 100,000 images is built. Most of the images in the dataset are collected from online stores (farfetch[1] and taobao[2]) and clothing factories, and the rest are from some commonly used image search engines (baidu[3]). Figure 7 show some sample images in the SAR. As shown in the figure, all images in dataset are dominated by shirt.

According to the common shirt type, the recognition task in the study contains identifying the circumference, sleeve type, sleeve length, waist, pockets, placket, and collar of shirt. Through the analysis of the images in the dataset, and combined with the classification of various attributes of the shirt, the specific classification under each dimension is determined as shown in Figure 8. Although the attributes classification of shirts are more detailed, considering the distribution of the images in the dataset, only a rough division in made here. We use the software named Label-Image to label the image. Otherwise, the principle of labeling is to use a minimum area to frame the target attribute region. Figure 8 also show the labeling samples for each attribute classification. Table 3 shows the distribution of the annotated data in each category. Although there is an imbalance between the numbers in each category, each of them exceeds 10,000, which can meet the training requirements.

B. THE DISCUSSION ON REGION PROPOSAL

The performance of selective search algorithm is evaluated by the overlap rate between the extracted region and labeled box. The overlap rate is represented by Intersection over Union (IoU) which is obtained by calculating the area ratio of the intersection and union of two region. This study employ the mean average best overlap (MABO) to evaluate the performance of selective search algorithm on all class. For a class $c \in C$, the ground-truth box is represented by $g_i^c \in \boldsymbol{G}^c$, the number of ground-truth is n, l_j represents the area extracted by the SS algorithm relative to this class, m is the number of classes, then MABO can be expressed as the following formula.

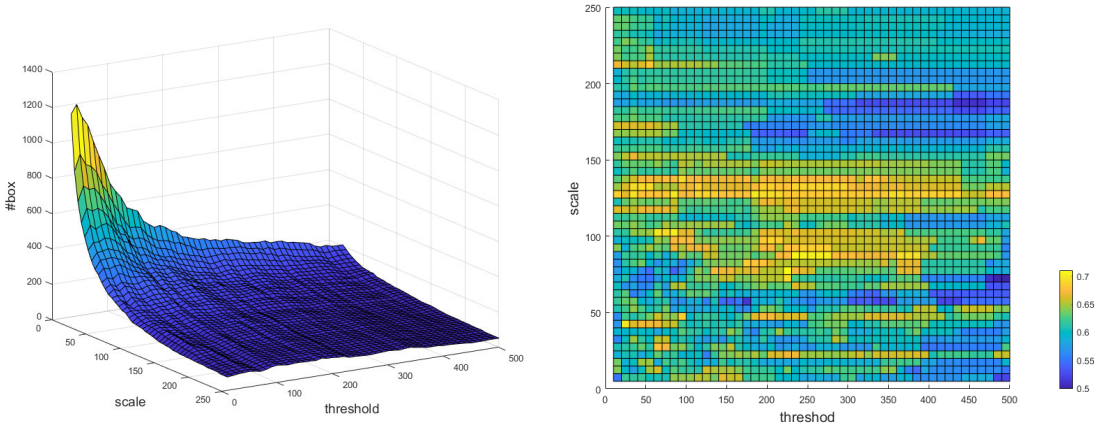
$$MABO = \frac{1}{m} \sum\nolimits_c^C \left(\frac{1}{n} \sum_{g_j^c \in \mathbf{G}^c} \max \left(\frac{g_j^c \cap l_j}{g_j^c \cup l_j} \right) \right) \tag{22}$$

[1] www.farfetch.com
[2] www.taobao.com
[3] image.baidu.com

FIGURE 7. Some sample images in the dataset.

FIGURE 8. The classification and labeling of shirt attributes.

Compared with the original version of the selective search algorithm, the proposed method employs color moment to describe image color features and LBP descriptor to describe texture features. For single strategy with four mentioned similarities, we control the extraction effect by optimizing the scale and threshold in the graph-based segmentation algorithm. Figure 9 show the region extraction effects with different scale and threshold. Figure 9(a) indicates that the smaller scale and threshold, the more region boxes are extracted. However, the more boxes do not represent the higher score. On the contrary, it will cause more computational waste. As shown in Figure 9(b), when the scale is about 85 or 130 and the threshold is about 250 or 380, the effects of region proposal is relatively good. To save the computational cost, the reminder of this paper set the scale $=130$ and threshold $= 380$.



(a) Effect on the number of boxes

(b) Effect on the MABO

FIGURE 9. The region extraction effect with different parameters by using single strategy.

TABLE 3. Distribution of annotated data.

Label Type		amount	Label Type		amount
Circumference	Skintight	17500	placket	Full-placket	22480
	Formfitting	39430		Half-placket	15360
	Loose	35110	Sleeve length	Sleeveless	12950
	Oversized	15620		Short-sleeved	21790
Waist shape	Waist-controlled	10210		Mid-sleeved	27290
Collar shape	Round-collar	18330		Long-sleeved	30000
	V-collar	18660	Sleeve shape	Horn-sleeved	12140
	Normal-collar	23620		Normal-sleeved	18280
	Peter-collar	17390	Pocket	pocket	13860

The author of SS algorithm demonstrate that using a variety of complementary grouping strategies can get a good quality set of object locations. As a full search for the best combination is computationally expensive, this experiment perform a greed search using the MABO only as optimization criterion, which is representative for the trade-off between the number of locations and its quality. From the resulting ordering we create five configurations as detailed in Table 4. Comparing the first row and second row in the table, the modified method has a certain improvement over the original algorithm in the proposal for the clothing attributes region. Moreover, it proves that the methods applying multiple strategies perform better than the methods applying single strategy. However, the more complex the combination strategy represent the more computational and time cost. Thus, the proposed framework takes a compromise as shown in the third row of the table.

The ABO score in each class is shown in Table 5. The modified selective search algorithm gains a good performance on the category of circumference, pocket and collar shape. For some classification such as placket, sleeve length where the features are not obvious, the scores are relatively low.

C. TRAINING FOR CLASSIFICATION MODEL

To enhance the generalization of data used to train the classification model, the images in the training-set consists of two parts, one part is manually labeled, and the other part is composed of the region extracted by the modified selective search, as shown in Figure 10. How to determine whether the extracted region is a positive or negative sample? This study selects positive and negative samples by setting two IOU-thresholds. The regions whose IOU score between the labeled regions is greater than the threshold T_1 are determined as positive samples. On the contrary, the regions whose IOU score between the labeled regions is lower than the threshold T_2 are determined as negative samples which is classified as background. This study set $T_1 = 0.7$ and $T_2 = 0.3$.

Since the model employed in this study requires the same size of input, the size of extracted region need to be normalized. The selected model is Inception-ResNet-V1 with an input size of $299 \times 299 \times 3$. As shown in Figure 11, the image normalization is a process of scaling. The scaling factor is determined by the longer side of the rectangle. The image scaling method applies bilinear interpolation.

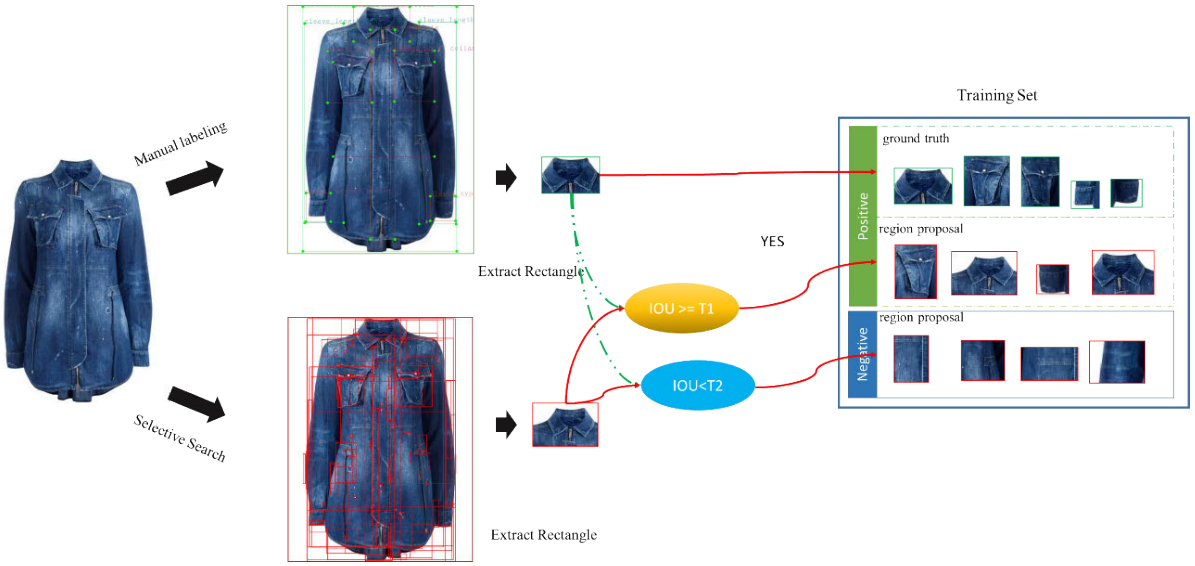


FIGURE 10. The composition of training-set.

TABLE 4. The selective search methods resulting from a greedy search.

(The second column (combination) indicates different configurations. Each cell in this column has three rows, where the first row represents a list of calculation methods for color similarity, the second column represents a list of calculation methods for texture similarity, and the third row represents a list of calculation methods for total similarity)

Type	Combination	strategies	MABO	#box
Single strategy	CM LBP C+T+S+F	1	0.695	275
	Hist(Lab) SIFT-Like C+T+S+F	1	0.674	289
Multiple strategies	CM LBP, SIFT-Like C+T+S+F, T+S+F	4	0.816	1084
	CM, Hist(Lab, HSV), LBP, SIFT-Like C+T+S+F, T+S+F	12	0.854	5692
	CM, Hist(Lab, RGB, HSV) LBP, SIFT-Like C+T+S+F, T+S+F, S+F, T	32	0.887	15168

The system's hardware environment is an HP workstation (Z840 TOWER: CPU–E5-2623 V4 @2.6GHz, Memory 64G) with a NVIDIA TITAN XP GPU (11G graphics memory). The model training is based on the framework named Tensorflow. To better initialize the model, we pre-train this deep neural network on ImageNet. When fine-tune the pre-trained model on our dataset, we set the learning rate with a lower level of 0.001. Moreover, the learning rate will gradually decay with a decay of 0.76 during training. To make the model converge faster and save computational cost, ADAM is used as the optimizer of the network. As shown in Figure 12(a), the specified learning rate is decayed by exponential decay. And Figure 12(b) shows the total loss changes during training. After 16 epochs of training, the model converges to a stable loss of approximately 0.4. The trained model's Top-1 mean accuracy on the verification set reached 82.32%. In order to test the effectiveness of the trained classification model, we created a test set (the number of images in each category is 100 and from different samples). Figure 13 show the distribution of the recognition result on verification set in each category. The trained model perform well on most category.

TABLE 5. The ABO score in each class.

class	ABO	class	ABO
tight	0.946	long-sleeved	0.576
formfitting	0.958	waist-controlled	0.682
loose	0.951	pocket	0.964
oversized	0.938	full-placket	0.573
horn-sleeve	0.879	half-placket	0.684
normal-sleeve	0.819	normal-collar	0.864
sleeveless	0.621	peter pan	0.859
short-sleeved	0.752	V-collar	0.884
mid-sleeved	0.776	round-collar	0.825

D. RESULTS AND ANALYSIS

After classification, the score of each extracted region is obtained. Then the model apply the Soft-NMS to output result by using the obtained score. Figure 14 show recognition result of two samples, which contain all the needed labels. We first perform experiments on the testing-set with about 10,000 images. There are three evaluation indicators for the recognition model, namely the labeling rate, precision and recall. Suppose the number of all tags in the dataset is m, the number of model output tags is n, and the number of correct tags in the model output is p. Thus, $\frac{n}{m}$, $\frac{p}{m}$ and $\frac{p}{n}$ represent the labeling rate, precision and recall. The verification results are reflected in Figure 15. The comprehensive labeling rate of the proposed model reaches 87.77%.

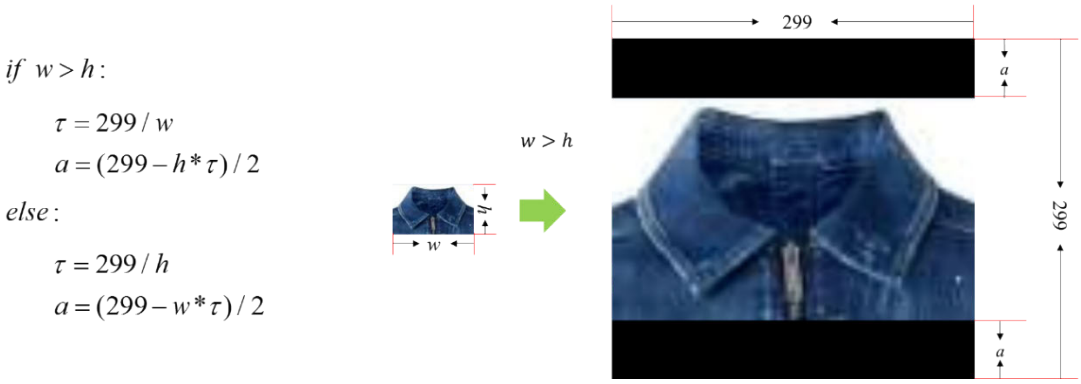


FIGURE 11. Image normalization process.

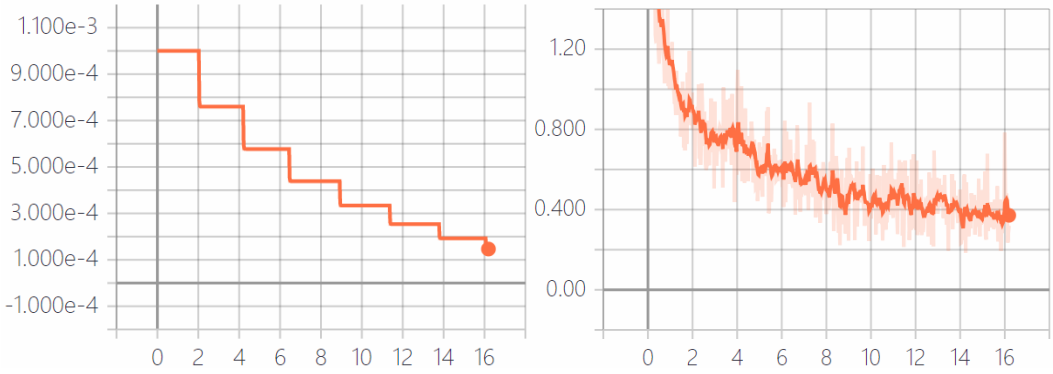


FIGURE 12. Learning rate and total loss curve during training.

TABLE 6. Comparison with other network framework.

Framework	With Softmax loss			With L-Softmax loss		
	Labeling rate	Precision	Recall	Labeling rate	Precision	Recall
AlexNet	73.19%	59.64%	81.49%	71.28%	60.10%	84.32%
VGG-16	79.34%	68.10%	85.83%	77.03%	67.22%	87.26%
Inception V1	75.68%	63.59%	84.02%	74.16%	64.13%	86.48%
Inception V2	78.93%	65.86%	83.44%	79.28%	67.82%	85.54%
Inception V3	80.68%	67.27%	83.38%	81.61%	68.12%	83.47%
Inception V4	83.38%	69.19%	82.98%	84.10%	70.15%	83.41%
Inception-ResNet V1	85.35%	71.66%	83.96%	87.77%	73.59%	83.84%
Inception-ResNet V2	86.31%	72.33%	83.80%	87.92%	74.00%	84.17%
ResNet-50	79.89%	69.63%	87.16%	80.15%	71.35%	89.02%

Comprehensive accuracy and recall rate reached 73.59% and 83.84% respectively.

The model performs better in the positioning and recognition of the circumference, because its visual features are more obvious. Moreover, people are not prone to great differences when labeling relevant feature regions. By analyzing the recognition results, there are three main reasons for missed detection and false detection: (1) The influence of the person's posture in the image is inevitable; (2) Selective search algorithm does not extract the relevant region; (3) The feature region of plain clothing are difficult to accurately label. Thus, the performance of sleeve length, waist shape and placket is relatively poor. When we perform the model on dataset including the image without people, the performance of the model has been significantly improved, and precision has reached 79.65%.

E. COMPARISONS

Since the feature extraction of image is crucial to the system, this study compare the recognition results and evaluation

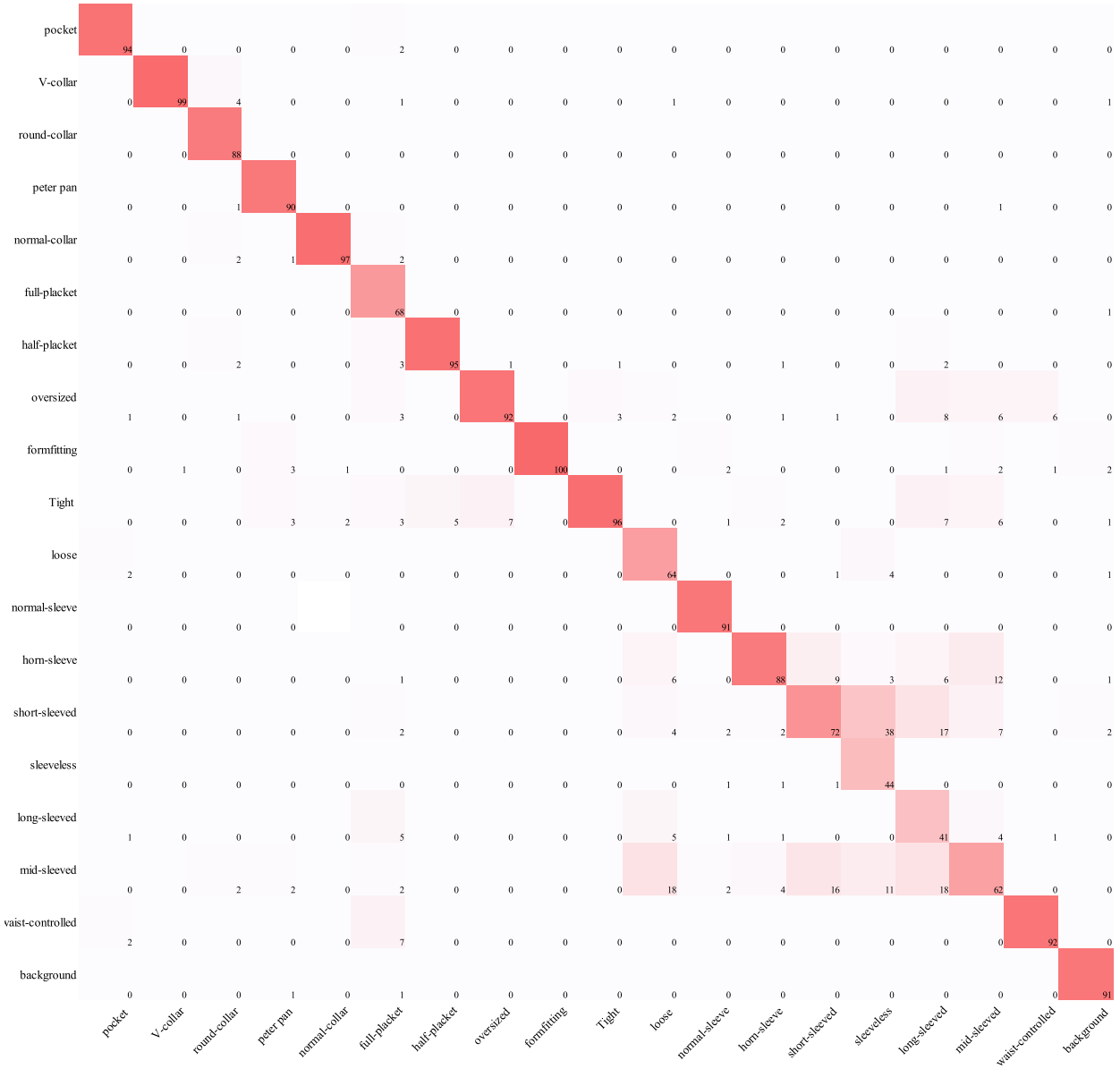
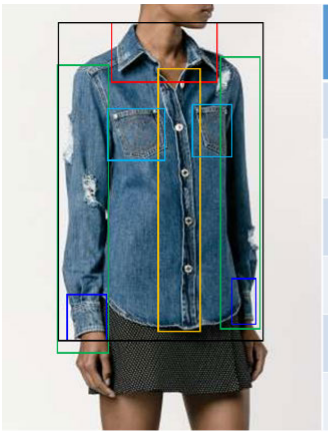


FIGURE 13. The distribution of the recognition result on the testing set. In the figure, the horizontal axis indicates the ground-truth label and the vertical axis indicates the recognition result.

metrics obtained by features from different network framework: AlexNet [1], VGG-16 [3], Inception V1 [4], Inception V2 [6], Inception V3 [5], Inception V4 [7], Inception-ResNet-V1 (proposed), Inception-ResNet-V2, ResNet-50 [2]. The comparative evaluation metrics include the three indicators mentioned previously, namely labeling rate, precision and recall.

The experimental results are shown in Table 5. VGG-16 and ResNet-50's performance is significantly better than AlexNet, because the size of VGG-16 and ResNet-50 model is higher than AlexNet. Although the Inception module increases the sparsity of the network and reduces the number of parameters of the model, the performance of the recognition is not lost to VGG. These results prove the superiority of this structure. The model size for Inception-ResNet V1 and Inception V3 are on one level, and Inception-ResNet V2 and Inception V4 are on one level. However, the former performs better than the latter. Moreover, the model with Inception-ResNet modules converge faster. As the author stated, the Inception-ResNet V2 model outperform Inception-ResNet V1, just by virtue of the increased model size. Nevertheless, the larger model size indicates more parameters and computation. For the task of clothing attributes recognition, Inception-ResNet V2 model takes more time than Inception-ResNet V1. Considering comprehensively, we propose to employ Inception-ResNet V1 as the model for classification. In further experiments, the models with L-Softmax loss gain a better recall than the corresponding models with Softmax loss. While the L-Softmax loss will increase the learning difficulty of the



type	value
Circumference	formfitting
Sleeve shape	normal-sleeve
Sleeve length	long-sleeved
Pocket	pocket
Placket	full-placket
Collar shape	normal-collar

type	value
Circumference	formfitting
Sleeve shape	horn-sleeve
Sleeve length	long-sleeved
Placket	full-placket
Collar shape	round-collar

FIGURE 14. Recognition result of two samples.

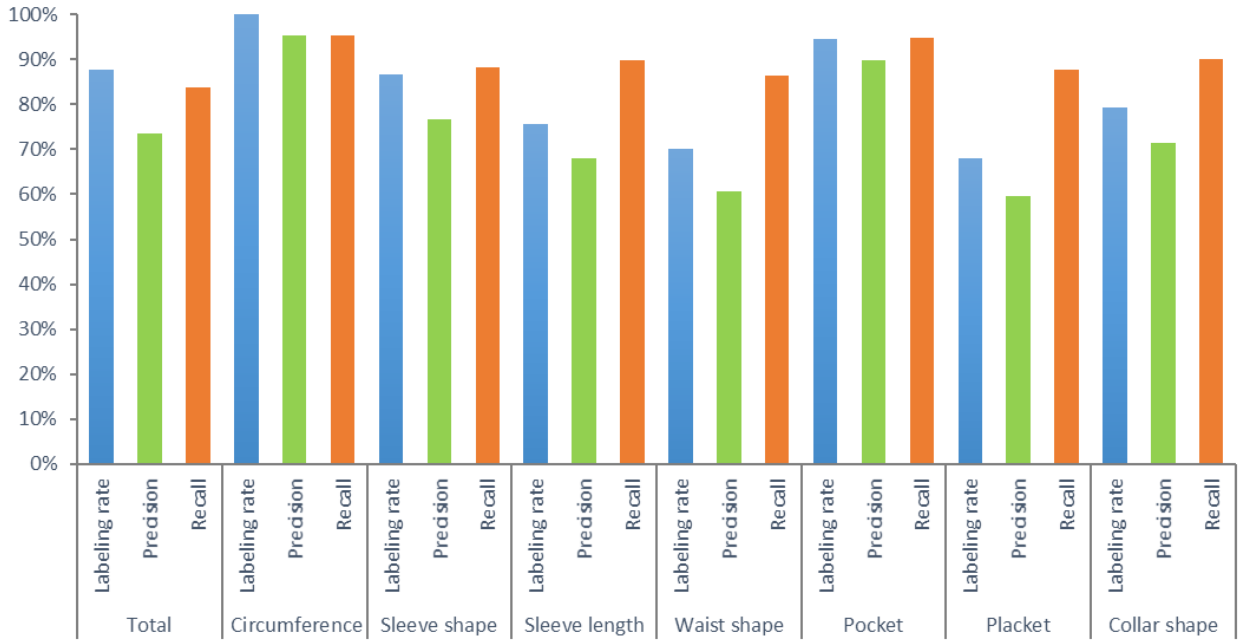


FIGURE 15. The verification results.

TABLE 7. Comparison with other methods.

Methods	Labeling rate	Precision	Recall
MFRCNN	90.50%	69.45%	76.74%
YOLO	76.18%	57.83%	75.91%
SSD	75.42%	59.62%	79.05%
Proposed	87.77%	73.59%	83.84%

model, so it has a significant improvement on the performance of sparse network.

The proposed framework is a typical two-stage object detection framework. The method in [38] (MFRCNN) apply the two-stage framework (Faster RCNN) which is based on a multi-task learning network. MFRCNN treats the identification of each part of the clothing as a task, so this method achieve a good labeling rate. However, its region proposal strategy leads to lower precision in the task of this paper. Compared with one-stage object detection framework like YOLO [30] and SSD [29], the advantage of the proposed framework is that the recognition accuracy is relatively high, and the disadvantage is that the computation cost is relatively expensive. However, it should be pointed out that YOLO and SSD are not sensitive enough to small objects, which is the main reason for poor performance.

V. CONCLUSION

In this paper, we address the task of recognizing clothing (shirt) attributes based on RCNN framework. Considering training time and accuracy comprehensively, we propose

a novel framework for automatically extracting the feature regions and recognizing the attributes of clothing. The framework first apply modified selective search algorithm to extract the region where targets may exist. Then the Inception-ResNet V1 network with L-Softmax is employed to extract the features of region and predict the category of it. After that, we use Soft-NMS to pick up the region with maximum score. Finally, a simple neural network is applied to correct the boundary of the selected region. Experimental result show that, the proposed framework has good performance on SAR with a labeling rate of 87.77%, a precision of 73.59% and a recall of 83.84%.

We are planning to improve current framework and implementation in several future directions. First, current framework needs to identify a large number of regions without targets each time. By drawing on the ideas of Faster-RCNN, we hope to solve the problem of slow speed caused by large number of candidate regions. Secondly, we also plan to apply the ideas of the proposed framework to the field of clothing retrieval. The key is the dimensionality reduction and coding of features.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[6] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[7] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.
[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[9] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.
[10] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. ICML*, 2016, p. 7.
[11] Z. Hu, H. Yan, and X. Lin, "Clothing segmentation using foreground and background estimation based on the constrained delaunay triangulation," *Pattern Recognit.*, vol. 41, no. 5, pp. 1581–1592, May 2008.
[12] X. Wu, B. Zhao, L.-L. Liang, and Q. Peng, "Clothing extraction by coarse region localization and fine foreground/background estimation," in *Advances in Multimedia Modeling*. Berlin, Germany: Springer, 2013, pp. 316–326.
[13] M. Weber and M. Bauml, "Part-based clothing segmentation for person retrieval," in *Proc. 8th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2011, pp. 361–366.
[14] N. Wang and H. Ai, "Who blocks who: Simultaneous clothing segmentation for grouping images," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1535–1542.
[15] S. Vittayakorn, K. Yamaguchi, A. C. Berg, and T. L. Berg, "Runway to realway: Visual analysis of fashion," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 951–958.
[16] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3330–3337.
[17] X. Wang and T. Zhang, "Clothes search in consumer photos via color matching and attribute learning," in *Proc. 19th ACM Int. Conf. Multimedia (MM)*, 2011, pp. 1353–1356.
[18] M. Mizuochi, A. Kanezaki, and T. Harada, "Clothing retrieval based on local similarity with multiple images," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 1165–1168.
[19] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3343–3351.
[20] R. Li, F. Feng, I. Ahmad, and X. Wang, "Retrieving real world clothing images via multi-weight deep convolutional neural networks," *Cluster Comput.*, vol. 22, no. S3, pp. 7123–7134, May 2019.
[21] J.-C. Chen and C.-F. Liu, "Deep net architectures for visual-based clothing image recognition on large database," *Soft Comput.*, vol. 21, no. 11, pp. 2923–2939, Jun. 2017.
[22] R. Li, W. Lu, H. Liang, Y. Mao, and X. Wang, "Multiple features with extreme learning machines for clothing image recognition," *IEEE Access*, vol. 6, pp. 36283–36294, 2018.
[23] S. C. Hidayati, C.-W. You, W.-H. Cheng, and K.-L. Hua, "Learning and recognition of clothing genres from full-body images," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1647–1659, May 2018.
[24] M. Yang and K. Yu, "Real-time clothing recognition in surveillance videos," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2937–2940.
[25] S. C. Hidayati, K.-L. Hua, W.-H. Cheng, and S.-W. Sun, "What are the fashion trends in new york?" in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 197–200.
[26] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun, "A high performance CRF model for clothes parsing," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 64–81.
[27] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 609–623.
[28] K. Yamaguchi, T. Okatani, K. Sudo, K. Murasaki, and Y. Taniguchi, "Mix and match: Joint model for clothing and attribute recognition," in *Proc. BMVC*, 2015, p. 4.
[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[31] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USEUIX Symp. Oper. Syst. Design Implement.*, 2016, pp. 265–283.
[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[33] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.
[34] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[35] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep image saliency computing via progressive representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1135–1149, Jun. 2016.
[36] L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, and L. Zhang, "A deep structured model with Radius–Margin bound for 3D human activity recognition," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 256–273, Jun. 2016.
[37] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, May 2016.
[38] Y. Sun and Q. Liu, "Attribute recognition from clothing using a faster R-CNN based multitask network," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 16, no. 2, Mar. 2018, Art. no. 1840009.

[39] Z. Gu, J. Zhang, Z. Pan, H. Zhao, and L. Zhang, "Clothes keypoints localization and attribute recognition via prior knowledge," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 550–555.
[40] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep. 2004.
[41] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[42] J.-P. Ainam, K. Qin, G. Liu, and G. Luo, "Sparse label smoothing regularization for person re-identification," *IEEE Access*, vol. 7, pp. 27899–27910, 2019.
[43] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS–improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5561–5569.

JUN XIANG received the master's degree in textile engineering from Jiangnan University. He is currently pursuing the Ph.D. degree in textile science and engineering with the School of Textile and Clothing, Jiangnan University, Wuxi, China. His research interests include image analysis, textile measurement, machine learning, and intelligent manufacturing.

TIANTIAN DONG received the bachelor's degree in clothing design and engineering from Qingdao University. She is currently pursuing the master's degree in textile science and engineering with the School of Textile and Clothing, Jiangnan University, Wuxi, China. She is interested in image analysis and functional fabric development.

RURU PAN received the B.S. and Ph.D. degrees in textile engineering from Jiangnan University, Wuxi, China, in 2005 and 2010, respectively, where he is currently a Professor with the School of Textile and Clothing. His current research interests include digital textile technology and digital image processing of textile.

WEIDONG GAO received the B.S. and M.S. degrees in textile engineering from the Wuxi Institute of Light Industry, Wuxi, China, in 1982 and 1985, respectively, and the Ph.D. degree in textile engineering from Donghua University, Shanghai, China, in 2011. He is currently a Full Professor with the School of Textile and Clothing, Jiangnan University, Wuxi. His current research interests include intelligent textile technology, intelligent weaving, and digital image processing of textile.