Clothing Landmark Detection Using Deep Networks With Prior of Key Point Associations

Chang-Qin Huang, *Member, IEEE*, Ji-Kai Chen, Yan Pan, Han-Jiang Lai, Jian Yin, and Qiong-Hao Huang

***Abstract*—This paper considers a problem of landmark point detection in clothes, which is important and valuable for clothing industry. A novel method for landmark localization has been proposed, which is based on a deep end-to-end architecture using prior of key point associations. With the estimated landmark points as input, a deep network has been proposed to predict clothing categories and attributes. A systematic design of the proposed detecting system is implemented by using deep learning techniques and a large-scale clothes dataset containing 145 000 upper-body clothing images with landmark annotations. Experimental results indicate that clothing categories and attributes can be well classified by using the detected landmark points, which are associated with regions of interest in clothes (e.g., the sleeves and the collars) and share robust learning representation property with respect to large variances of human poses, nonfrontal views, or occlusion. A comprehensive performance evaluation over two newly released datasets is carried out in this paper, showing that the proposed system with deep architecture for clothing landmark detection outperforms the state-of-the-art techniques.**

***Index Terms*—Image recognition, neural networks, object detection.**

I. Introduction

With the ever growing market of the online clothing shopping, category or attribute recognition of clothes has become a fundamental technique in many practical applications, e.g., attribute-similar clothes retrieval [10], [14], [36], [37], outfit recommendation [32], fashion style recognition [16], occupation recognition [30], and people identification by clothing cues in surveillance videos [6]. Many methods for clothing category or attribute recognition have been proposed (e.g., [5], [6], [14], [21], [33], [36], and [40]).

In the last few years, we are witnessing dramatic progress in deep convolutional networks. An emerging stream for clothing-related recognition is the deep learning methods, which use deep convolutional neural networks to simultaneously learn feature representations and classify clothing-related images. However, most of these methods do not consider large variances in human poses, occlusion, and nonfrontal views in the clothing images, which may degrade their performance. Very recently, Liu *et al.* [24] proposed a deep learning method that learns clothing features by jointly predicting landmark locations and clothing attributes. The estimated landmarks are used to pool or gate the learned feature maps, making the feature representations more robust to clothing deformation/occlusion. Despite the method in [24] has achieved good performance, it estimates the landmark locations by simple regression on the coordinates of landmark points, resulting in the estimated landmark points being inaccurate.

In this paper, we take a step further over [24] in improving clothes category and attribute prediction via landmark localization. Specifically, we propose a novel method for landmark localization based on a deep end-to-end architecture using part affinity fields (PAFs) [3]. This architecture has four building blocks.

1) An input image is encoded to discriminative feature representations via a pretrained convolutional subnetwork based on the VGG-16 model [31].
2) On top of this convolutional subnetwork, we construct two deconvolution layers to make each of the output deconvolution feature maps to be in the same size as an input image.
3) On top of the deconvolution feature maps, there are two branches: the first branch is a convolution layer to predict an initial probabilistic map (for each of the landmarks) that represents the probabilities of a landmark point's position within the input image, and the second one is a convolutional layer to output feature maps that capture the associations of landmark points.
4) By using element-wise summation to merge an initial probabilistic map and its corresponding feature map of associations, we obtain the final probabilistic map of a specific landmark, followed by a cross-entropy loss function defined on this probabilistic map.

With the estimated landmark points as input, we propose a deep network that can predict clothing categories

Manuscript received February 28, 2018; revised May 9, 2018 and June 11, 2018; accepted June 21, 2018. This work was supported in part by the National Science Foundation of China under Grant 61370229, Grant 61772567, Grant 61370021, Grant U1611264, and Grant 61602530, in part by the S&T Projects of Guangdong Province under Grant 2015A030401087, Grant 2016B010109008, and Grant 2015B010110002, in part by GDUPS, and in part by the Guangzhou Science Technology and Innovation Commission under Grant 201704030079. This paper was recommended by Associate Editor D. Wang. *(Corresponding author: Yan Pan.)*

C.-Q. Huang was with the School of Information Technology in Education, South China Normal University, Guangzhou 510631, China. He is now with the Department of Educational Technology, Zhejiang Normal University, Jinhua, Zhejiang 321004, China (e-mail: cqhuang@zju.edu.cn).

Q.-H. Huang is with the School of Information Technology in Education, South China Normal University, Guangzhou 510631, China.

J.-K. Chen, Y. Pan, H.-J. Lai, and J. Yin are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: panyan5@mail.sysu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2018.2850745

and attributes (e.g., sleeve types and collar types). The landmark points are used to locate the regions of interest in clothes (e.g., the regions of the sleeves, the collars, or the whole clothes), making the learned feature representations more robust to varying human poses, nonfrontal views, or occlusion.

In order to facilitate the research of clothing landmark localization, we release a dataset with 145 000 upper-body clothing images cropped from online shopping websites, each image is annotated with ten landmark points (i.e., neck, collarband, and left and right shoulder/elbow/wrist/waist). The landmark annotations in this dataset are in a finer level than either the 6-landmark annotations in [24] or the bounding box annotations. To the best of our knowledge, this is so far the largest dataset for clothing landmark detection. In addition, we release a dataset for clothing category and attribute recognition. This dataset contains 25 500 images in eight categories, each image is annotated with a category and two attributes (the collar type and the sleeve type).

We conduct extensive evaluations of clothing landmark detection and clothing category and attribute classification on publicly available or our released datasets. The results show the following.

1) The proposed landmark detection method outperforms the baseline methods on our released dataset for clothing landmark detection.
2) The proposed approach for clothing category and attribute classification has superior performance gains over the state-of-the-art baselines.

The main contributions of this paper are summarized as follows.

1) We develop a deep learning method for landmark localization, which uses carefully designed stack of convolution/deconvolution layers to obtain initial probabilistic maps of landmark locations, and then uses associations of the locations/orientations of landmarks to improve the prediction accuracy of landmark points.
2) With the estimated landmark points, we propose a deep network for clothing category and attribute classification. The landmark points are used to locate the regions of interest in clothes, making the learned features more robust to a large number of variances of human poses, nonfrontal views, or occlusion.
3) We propose a 10-landmark configuration for upper-body clothes, which provides a finer level of location information than the landmark configuration (or the bounding box annotations) in the existing methods.
4) We release a dataset with 145 000 upper-body clothing images, each image being annotated with ten landmark points. To the best of our knowledge, this is so far the largest dataset for clothing landmark detection. We also release a dataset for clothing category and attribute recognition, each image is annotated with the clothing category, the collar type, and the sleeve type. These datasets can be beneficial to the research of clothes landmark localization and clothing category or attribute prediction.

II. Related Works

A. Clothing Landmark Detection Methods

Landmark detection has shown its success in various applications, e.g., face alignment and human pose estimation. To handle the possibly large variances in human poses in clothing images, clothing landmark detection, a task of predicting the key points defined on the clothes, has received considerable attention in recent years [24], [25]. The predicts landmark locations can be used to obtain more robust feature representations of clothes, making more accurate predictions in clothing-related tasks (e.g., classifying clothing categories and attributes).

Several clothing landmark detection methods have been proposed in the literature. Liu *et al.* [24] released a dataset for clothing landmark detection, which defines six key points in each of the upper-body clothing images. Liu *et al.* [25] also proposed a cascade three-stage method to predict clothing landmarks. They showed that the detected clothing landmarks can be used to align the clothing images and improve the performance of clothing category and attribute recognition. However, these methods (e.g., [24] and [25]) have not consider the possible associations of landmark points, which may degrade their performance.

In this paper, we propose an end-to-end deep architecture that first uses carefully designed stack of convolution/deconvolution layers to obtain initial probabilistic maps of landmark locations, and then incorporates the possible associations of landmark points to refine these probabilistic maps and obtain more accurate predictions of landmark points. In addition, different from the 6-point landmark configuration for upper-body clothes in [24] and [25], we propose a 10-point landmark configuration which provides finer level location information. We also release a large-scale dataset with 145 000 annotated images for 10-point landmark detection.

B. Clothing Category and Attribute Recognition Methods

In recent years, clothing category and attribute recognition has attracted considerable research interest in computer vision and multimedia communities. Clothing category and attribute recognition has been widely used in attribute-similar clothes retrieval [10], [14], [36], [37], outfit recommendation [32], fashion style recognition [16], occupation recognition [30], and people identification by clothing cues in surveillance videos [6].

The early methods (e.g., [5], [21], and [33]) for clothing category and attribute recognition usually adopt a pipeline of two steps.

1) Extracting hand-crafted visual features (e.g., HOG [8], color histogram) from images.
2) Learning classifiers for clothing-related recognition based on these features. However, the limited expressive power of the hand-crafted visual features may degrade the performance of these methods.

In the last few years, dramatic progress has been made in deep convolution networks. Approaches based on deep networks have achieved state-of-the-art performance on image classification [13], [17], [31], [34], [38], object detection [17], [34], and

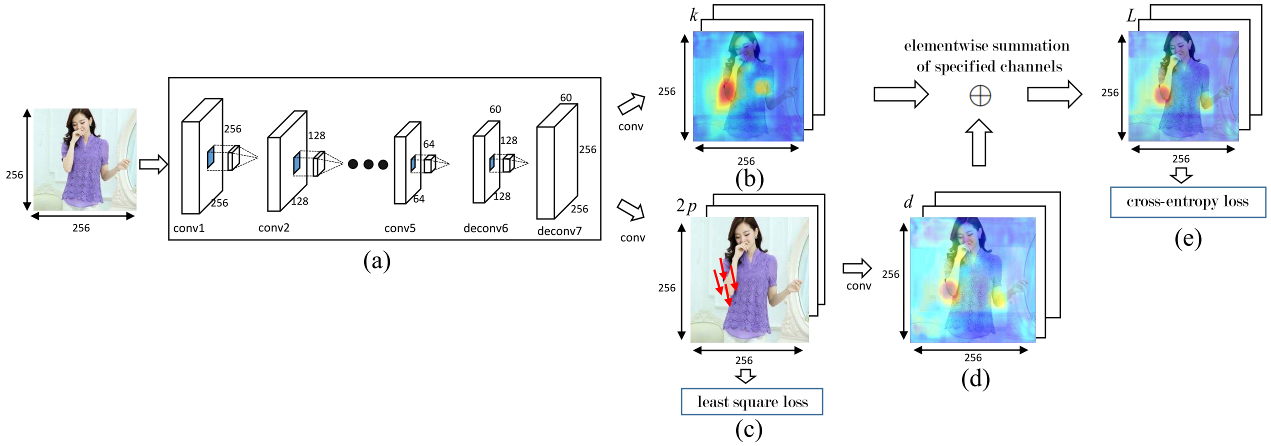


Fig. 1. Overview of the proposed deep network for clothing landmark detection. The proposed architecture is a fully convolutional network. (a) Convolution subnetwork based on the VGG-16 model by removing the fully connected layers and the last three max-pooling layers, followed by a stack of two deconvolution layers to upsample the output deconvolution feature map to the same size as the input image. (b) Convolution layer on top of the deconvolution feature maps. This convolution layer outputs k initial probabilistic maps. The ith map represents the estimated initial probabilities of the ith landmark point's position in the input image ($i = 1, 2, \ldots, k$). (c) Convolution layer that we call "PAF layer." In this layer, we use a least square loss to incorporate the prior of landmark pairs' associations that preserve the relative locations and directions of landmarks. (d) Convolution layer that outputs feature maps in the same size of the initial probabilistic maps in (c). (e) By conducting element-wise summation on the initial probabilistic maps in (c) and the feature maps in (d), we obtain the refined probabilistic maps for landmarks. These refined probabilistic maps are used as input to the cross-entropy loss function for landmarks.

other recognition tasks (e.g., [18], [20], [35], [39], and [41]). The success of deep learning methods for images is mainly due to their power of automatically learning effective image representations. For the clothing category and attribute recognition, an emerging stream is the deep learning methods that use deep convolutional neural networks to simultaneously learn more expressive feature representations and clothing-related classifiers from images. Chen *et al.* [6] addressed the problem of describing people via predicting clothing attributes by deep convolutional networks. Veit *et al.* [36] proposed a deep architecture to learn and predict how fashionable a person looks on an image and give outfit recommendations to the users. Huang *et al.* [14] proposed a dual attribute-aware ranking network for cross-scenario clothing retrieval, in which the retrieval feature representations are driven by clothing attribute learning. Xiao *et al.* [40] proposed a deep architecture that learns clothes classifiers with only a limited number of clean labels and millions of easily obtained noisy labels. Dong *et al.* [9] developed a deep learning framework capable of model transfer learning from well-controlled shop clothing images collected from Web retailers to in-the-wild images from the street and formulated a novel multitask curriculum transfer deep learning method to explore multiple sources of different types of Web annotations with multilabeled fine-grained attributes. Most of these methods, however, do not consider the occlusion and various human poses in the clothing images.

III. Proposed Approach for Clothing Landmark Detection

Landmark detection for upper-body clothes contains two steps. The first step is to locate the positions of upper-body clothes by a detector (e.g., locating a piece of upper-body clothing by a bounding box). The second step is to learn a mapping function to map an input bounding-box image I to k coordinates $l_1, l_2, \ldots, l_k$ of landmark points, where l_i represents the coordinate of the ith landmark point.

A. Locating the Clothes

In the first step, we train a detector to identify the positions of clothes in images. Specifically, we collect 20 000 upper-body clothing images, and then manually annotate the position of clothes in each image by bounding boxes. After that, we train a detection model by using an object detection method (e.g., Fast-RCNN [29] and YOLO [28]). Here, we use YOLO as the object detection method, which is a leading approach of object detection.

With the learned detection model, for an image used for clothing landmark detection, we use this detection model to identify the bounding box that locates the upper-body clothes in the image. Then, this bounding box is resized to an image of 256 × 256, which is used as input to the following landmark detection step.

B. Deep Architecture for Landmark Detection

In this paper, we propose a deep architecture for clothing landmark detection. As shown in Fig. 1, the proposed architecture has four parts.

1) A convolutional subnetwork of stacked convolution layers to capture the feature representation of an input image.
2) A stack of deconvolution layers that outputs resolution-preserving deconvolution feature maps, each is in the same size of the input image.
3) An additional convolution layer on top of the deconvolution feature maps, where this convolution layer is as input to the loss function.
4) Two branches on top of the deconvolution feature maps, where the first branch is a convolution layer to predict initial probabilistic maps, each represents the

TABLE I
CONFIGURATIONS OF THE CONVOLUTIONAL SUBNETWORK BASED ON VGG-16. THE INPUT SIZE IS 256×256

type	filter size / stride	output size
convolution	3×3 / 1	$256 \times 256 \times 64$
convolution	3×3 / 1	$256 \times 256 \times 64$
max pooling	2×2 / 2	$128 \times 128 \times 64$
convolution	3×3 / 1	$128 \times 128 \times 128$
convolution	3×3 / 1	$128 \times 128 \times 128$
max pooling	2×2 / 2	$64 \times 64 \times 128$
convolution	3×3 / 1	$64 \times 64 \times 256$
convolution	3×3 / 1	$64 \times 64 \times 256$
convolution	3×3 / 1	$64 \times 64 \times 256$
convolution	3×3 / 1	$64 \times 64 \times 512$
convolution	3×3 / 1	$64 \times 64 \times 512$
convolution	3×3 / 1	$64 \times 64 \times 512$
convolution	3×3 / 1	$64 \times 64 \times 512$
convolution	3×3 / 1	$64 \times 64 \times 512$
convolution	3×3 / 1	$64 \times 64 \times 512$
deconvolution	4×4 / 2	$128 \times 128 \times 60$
deconvolution	4×4 / 2	$256 \times 256 \times 60$
convolution	3×3 / 1	$256 \times 256 \times 10$

probabilities of a landmark point's position within the input image, and the second one is a convolutional layer to output feature maps that preserve the associations of landmark points.

By using element-wise summation to merge an initial probabilistic map and its corresponding feature map of associations, we obtain the final probabilistic map of a specific landmark, followed by a cross-entropy loss function defined on this probabilistic map. In the following, we will present the details of these parts, respectively.

1) Convolutional Subnetwork: As shown in Fig. 1(a), we use a convolutional subnetwork with multiple convolution layers to capture a discriminative feature representation of the input images. This subnetwork is based on the architecture of VGG-16 [31], where we remove the three fully connected layers. Since spatial information is important for the landmark detection task, in order to keep sufficient spatial information, we also remove the last three max-pooling layer in VGG-16. After that, this subnetwork becomes a fully convolutional network that converts an input image to a set of feature maps that capture the discriminate image features.

For self-containness, Table I shows the configurations of the convolutional subnetwork for the input size 256×256. Note that each of the convolution layers is followed by a rectification activation layer which is omitted in Table I. In training, we use the pretrained VGG-16 [31] model to initialize the weights in this subnetwork.

2) Deconvolution Layers: Preserving sufficient spatial information in images is one of the crucial factors to estimate accurate positions of landmark points in landmark detection. Suppose the size of an input image is $h \times w$ (e.g., $h = w = 256$). Since the above-mentioned convolution subnetwork has two max-pooling layers, the size of each output feature map is $(h/4) \times (w/4)$ (e.g., 64×64) is too small to capture sufficient spatial information. Inspired by the practice in object segmentation [26], we add two deconvolution layers on top of the convolution subnetwork, so that each of the output feature maps that are outputs of the last deconvolution layer is in the same size as an input image (e.g., 256×256). These feature maps in large size can provide sufficient spatial information for the following steps to estimate the positions of landmark points.

3) Initial Probabilistic Maps for Landmarks: There are two branches on top of the deconvolution layers. The first branch generates k probabilistic maps, each corresponds to a landmark. Each of these maps represents the initial probabilities of a landmark point's position within the input image, which will be refined in the following steps.

Specifically, on top of the deconvolution layer, we add a convolution layer that outputs k feature maps, each in the size $h \times w$. That is, each of these feature maps is in the same size of an input image. The ith feature map corresponds to the ith landmark point ($i = 1, 2, \ldots, k$). We denote $F^{(i)}$ as the ith feature map. Then we convert $F^{(i)}$ to a probabilistic map $G^{(i)}$, where $G^{(i)}$ is defined as

$$G_{s,t}^{(i)} = \frac{F_{s,t}^{(i)}}{\sum_{s=1}^{h} \sum_{t=1}^{w} F_{s,t}^{(i)}} \tag{1}$$

where $F_{s,t}^{(i)}$ or $G_{s,t}^{(i)}$ represents the element in the sth row and tth column in $F^{(i)}$ or $G^{(i)}$ ($s = 1, 2, \ldots, h$; $t = 1, 2, \ldots, w$), respectively. We can verify that $G_{s,t}^{(i)} \geq 0$ and $\sum_{s=1}^{h} \sum_{t=1}^{w} G_{s,t}^{(i)} = 1$. The value of $G_{s,t}^{(i)}$ can be regarded as the predicted probability that the coordinate of the ith landmark point is (s, t).

4) Refining Probabilistic Maps: Each of the initial probabilistic maps represents the estimated probability distributions of a specific landmark, where these maps are independent to each other. However, there exist some associations of the relative locations and directions between two clothing landmarks in a pair. For example, as shown in Fig. 1(c), for a pair of the landmark at the shoulder and the landmark the corresponding elbow, their relative locations and orientations must be in a suitable range. Very recently, Cao *et al.* [3] proposed to use PAFs for multiperson pose estimation, where PAFs are nonparametric representations that encode the locations and orientations of landmark pairs, so as to capture the associations between landmarks. Given a set of detected (initial) landmark points, a PAF plays the role of a confidence metric to measure the association of a landmark pair, i.e., whether this landmark pair is in reasonable relative positions or directions. Specifically, PAFs construct a set of 2-D vector fields to encode the location and orientation of landmark pairs, each vector field can be regarded as a landmark-to-landmark association.

Inspired by the idea of PAFs in [3], in the proposed architecture for clothing landmark detection, we incorporate the prior information of landmark pairs' associations in training, making the estimated landmarks to preserve their relative locations and directions.

On top of the deconvolution maps, the second branch is a convolution layer which we call the PAF layer, where we incorporate the prior of landmark pairs' associations. The association of a landmark pair is represented by a 2-D vector that encodes the direction from one landmark (in the pair) to the other, which can be regarded as a kind of prior of relative locations/directions of human body parts. Fig. 1(c) shows an example of the direction starts from the landmark of the left shoulder to the one of the left elbow. Following [3], in our upper-body landmark configuration, we chose ten landmark pairs and define a specific 2-D vector for each of these pairs.

Specifically, the associations of landmark pairs are represented by a set of c tensors $\mathbf{P} = (\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_c)$, each tensor $\mathbf{P}_l \in \mathcal{R}^{h\times w\times 2}$ $(l = 1, 2, \ldots, c)$ corresponds to a specific landmark pair, where h and w equals to the height and width of an input image. Note that each of these tensors consist of two feature maps in the size $h \times w$. Suppose $\mathbf{P}_l$ $(l = 1, 2, \ldots, c)$ is the tensor corresponds to a chosen pair of landmarks whose coordinates are $L^{(i)}$ and $L^{(j)}$ $(i, j \in \{1, 2, \ldots, k\})$. We define a vector $\mathbf{v}$ as

$$\mathbf{v} = \left(L^{(j)} - L^{(i)}\right) / \left\| L^{(j)} - L^{(i)} \right\|_2 \tag{2}$$

which represents the 2-D unit vector in the direction of the landmark pair. For any coordinate p in an input image, we define

$$\mathbf{P}_l^{(p)} = \begin{cases} \mathbf{v} & \text{if } 0 \le v \cdot \left(p - L^{(i)}\right) \le \left\| L^{(j)} - L^{(i)} \right\|_2 \\ & \text{and } \left| v_\perp \cdot \left(p - L^{(i)}\right) \right| \le \epsilon \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{3}$$

where ϵ is a predefined distance in pixels, $\mathbf{v}_\perp$ represents the vector perpendicular to $\mathbf{v}$. By applying (3) to all of the coordinate in an input image, we can obtain the tensor P_l in the size $h \times w \times 2$.

In order to incorporate $\mathbf{P} = (\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_c)$ in training, we define a least square loss on the PAF layer. Suppose the output feature maps of the convolution layer in the second branch is $\mathbf{O} = [\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_c]$, where $\mathbf{O}_l \in \mathcal{R}^{h\times w\times 2}$ $(l = 1, 2, \ldots, c)$ and $\mathbf{O} \in \mathcal{R}^{h\times w\times 2c}$ represents the concatenation of $\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_c$. The least square loss is defined as

$$\mathcal{L}_{ls} = \frac{1}{c} \sum_{l=1}^{c} \|\mathbf{O}_l - \mathbf{P}_l\|_2^2. \tag{4}$$

On top of the PAF layer, we add an additional convolution layer that outputs k feature maps, each has the same size as an initial probabilistic map in the first branch. Then, as shown in Fig. 1(b), (d), and (e), with these feature maps and the initial probabilistic maps, the proposed architecture conducts element-wise summation and outputs the refined probabilistic maps for landmarks in Fig. 1(e).

5) Cross-Entropy Loss: In clothing landmark detection, the landmark points annotated by human are used as ground-truth. Note that the points nearby a human annotated landmark points can also be candidates of ground-truth. For example, suppose the coordinate of a human annotated landmark point is (x, y), then the points $(x + 1, y)$, $(x - 1, y)$, $(x, y + 1)$, $(x, y - 1)$, $(x + 1, y + 1)$, $(x + 1, y - 1)$, $(x - 1, y + 1)$, or $(x - 1, y - 1)$ can also be reasonable candidates of ground-truth. To address this issue, the proposed network does not directly predict the coordinates of landmark points, but predict a heat map that reflects the relative positions of landmark points in an input image.

Specifically, for the ith ground-truth landmark point with human annotated coordinate (x_i, y_i), we construct a weight map $U^{(i)}$ that is defined as

$$U_{s,t}^{(i)} = \max(0, 1 - \delta \max(|s - x|, |t - y|)) \tag{5}$$

where $0 \le \delta \le 1$ is a hyper-parameter that controls the weight decay, $s = 1, 2, \ldots, h$, $t = 1, 2, \ldots, w$. The meaning of $U_{s,t}^{(i)}$ is that, if neither the offset $|s-x|$ or $|t-y|$ is larger than $(1/\delta)$, the weight of the coordinate (s, t) is linearly decayed with the offset increasing; otherwise the weight of the coordinate (s, t) is 0. In our experiments, we empirically set $\delta = 0.8$.

After that, we also convert $U^{(i)}$ to a probabilistic map $W^{(i)}$, where $W^{(i)}$ is defined as

$$W_{s,t}^{(i)} = \frac{U_{s,t}^{(i)}}{\sum_{s=1}^{h} \sum_{t=1}^{w} U_{s,t}^{(i)}} \tag{6}$$

where $s = 1, 2, \ldots, h$, $t = 1, 2, \ldots, w$. The value of $W_{s,t}^{(i)}$ can be regarded as the probability that the ith landmark is at the position (s, t).

Suppose the refined probabilistic maps are $G^{(i)}$ $(i = 1, 2, \ldots, k)$. Equipped with $G^{(i)}$ and $W^{(i)}$, we use a cross-entropy loss function to measure the discrepancy between $G^{(i)}$ and $W^{(i)}$ for $i = 1, 2, \ldots, k$. Specifically, in training the proposed network, we minimize the following loss function defined by:

$$\mathcal{L}_{ce} = -\frac{1}{k} \sum_{i=1}^{k} \ell\left(G^{(i)}, W^{(i)}\right) \tag{7}$$

where $\ell(G^{(i)}, W(i))$ is defined as

$$\ell\left(G^{(i)}, W^{(i)}\right) = \sum_{s=1}^{h} \sum_{t=1}^{w} W_{s,t}^{(i)} \log\left(G_{s,t}^{(i)}\right). \tag{8}$$

6) Combination of Loss Functions: The overall loss function is defined as

$$\mathcal{L}_{\text{comb}} = \lambda \mathcal{L}_{\text{ls}} + \mathcal{L}_{\text{ce}} \tag{9}$$

where λ is a tradeoff parameter for the loss functions. In our experiments, we empirically set $\lambda = 1$.

IV. Proposed Method for Clothing Category and Attribute Classification

We seek to improve the performance of clothing category and attribute classification by the help of clothing landmark detection. With the estimated landmark points as input, we propose a deep architecture that can predict the categories and attributes (e.g., the sleeve types and the collar types) of upper-body clothes. The landmark points are used to locate the areas of the sleeves, the collars or the whole clothing, making the learned feature representation more robust to large variances of human poses, nonfrontal views, or occlusion.

Here we focus on the task of identifying the categories, sleeve types and collar types of upper-body clothes. The proposed architecture can also be extended to predicting other clothing attributes. For an upper-body clothing image with ten estimated landmark points, we first identify three parts of interest: 1) the collar part; 2) the left arm part; and 3) the right arm part. These parts can provide local information that

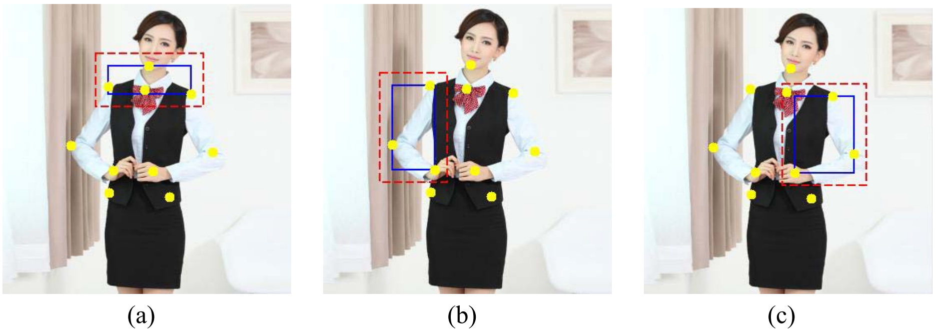


Fig. 2. Examples of cropped rectangle based on the given localized landmarks. The blue rectangles denote the minimum rectangle that covers the specified landmarks. The red rectangles denote the extended ones. (a) Collar. (b) Left arm. (c) Right arm.

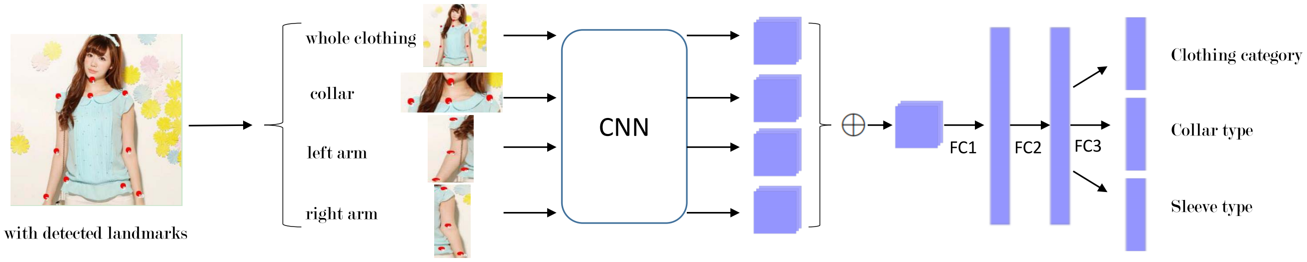


Fig. 3. Illustration of the network for clothing category classification. ⊕ denotes element-wise mean average pooling.

is related to the collar type, the sleeve types, and the clothing categories.

Specifically, as shown in Fig. 2(a), given four landmark points related to the collar part, we find the minimum rectangle [see the blue rectangle in Fig. 2(a)] that covers these landmark points. Then, we enlarge this rectangle by extending its width and height by 40%, respectively, with the original rectangle at the center position [see the red rectangle in Fig. 2(a)]. Similarly, as shown in Fig. 2(b) [or Fig. 2(c)], given three landmark points related to the left (or right) arm part, we also find the minimum rectangle that covers these landmark points, and then enlarge this rectangle by extending its width and height by 40%, respectively.

Next, we will present the proposed deep architecture for clothing category and attribute classification. For an input upper-body clothing image, we extract the rectangles of the whole image, the extended rectangle of the collar part, the left arm part, and the right arm part, respectively, as four inputs. Then we resize these inputs to the same size as the input image (e.g., $h \times w$), before using them as inputs to the proposed deep architecture. As shown in Fig. 3, through the proposed architecture, each of the four inputs is converted to a feature vector as the input's representation, respectively, via a convolutional subnetwork based on VGG-16 [31]. Then, the four feature vectors are averaged to one vector as the feature representation of the whole image. On top of this feature representation, we construct two fully connected layer, followed by three sibling vectors corresponding to the discriminative representation of the collar type, the sleeve type, and the clothing category, respectively. Finally, softmax loss functions are defined on these sibling vectors.

V. Released Datasets

In order to evaluate the proposed methods in this paper, we collect and release a large-scale dataset for landmark detection in upper-body clothes and a dataset for clothing category and attribute prediction.

Online shopping websites are a widely used source for constructing clothing-related datasets. The proposed datasets contain diverse upper-body clothing images that are crawled from a famous online shopping website[1] in China. These images are in various poses and viewpoints. We exclude the images in duplication, low resolution, low quality, or weird width/height ratios.

A. Dataset for Clothing Landmark Detection

Very recently, Liu *et al.* [24] released the DeepFashion datasets for clothing landmark detection, in which each upper-body clothing image is annotated with six landmark. In this paper, we propose a 10-landmark configuration for upper-body clothes. In order to evaluate the proposed 10-point landmark configuration and the proposed landmark detection method, we collect and release FCLD, a dataset for fine-level clothing landmark detection. This dataset contains 145 000 upper-body clothes, each is annotated with ten landmarks, which has more landmark-annotated images and provides a finer level of location information in the landmark annotations than DeepFashion [24].

Specifically, we define ten landmark points for a piece of upper-body clothing, where these points are located in crucial positions that can describe the structure of the clothing. Fig. 4 shows some examples of the landmark annotations.

[1] www.jd.com

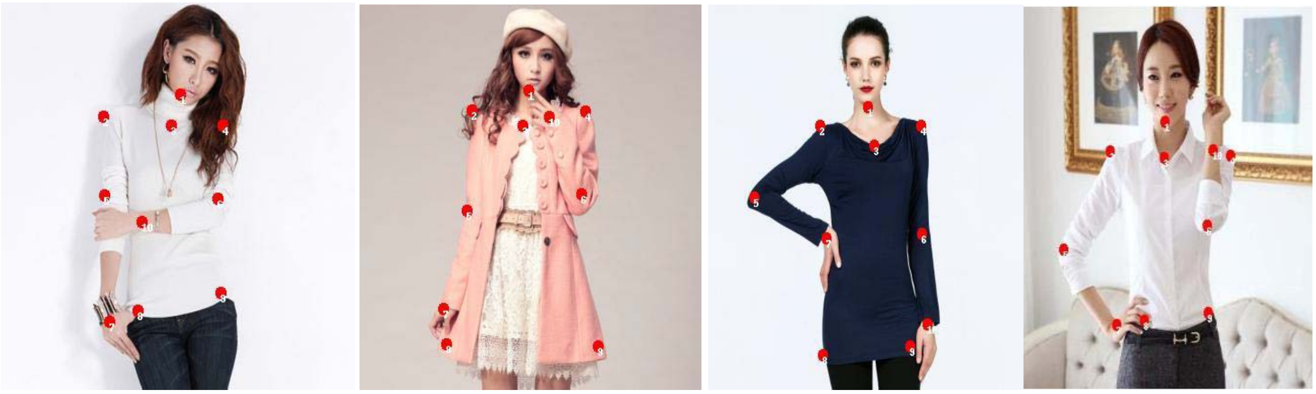


Fig. 4. Examples of clothing landmarks (upper and bottom collar end, left and right shoulder, left and right elbow, left and right cuff or hand, and left and right end of the bottom hem) in the FCLD dataset.

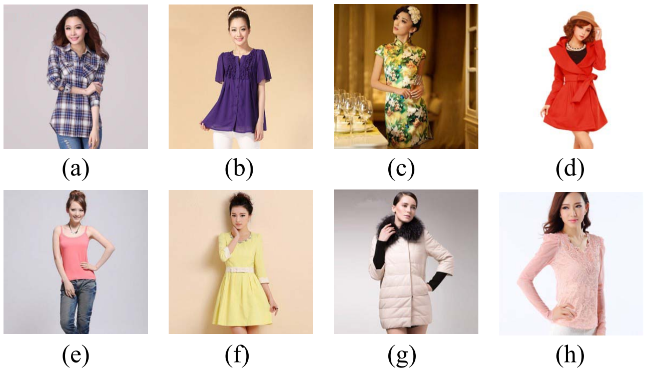


Fig. 5. Examples of clothes in each clothing category in the CCAP dataset. (a) Blouse. (b) T-shirt. (c) Cheongsam. (d) Coat. (e) Camisole. (f) Dress. (g) Fur coat. (h) Bottoming shirt.

The ten landmark points correspond to the upper/bottom end of the collar, the left/right shoulder, the left/right elbow, the left/right cuff, and the left/right end of the bottom hem, respectively. These landmark points can provides fine level of location information in the clothing image. For example, with these landmark points, one can easily locate the collar or the left/right arms. Such fine-level location information can be used to improve the feature extraction from clothing images.

B. Dataset for Clothing Category and Attribute Prediction

We also collect and release CCAP, a dataset for clothing category and attribute prediction. We collect 25 500 upper-body clothing images, each is annotated with its clothing category, collar type, and sleeve type. Each image belongs to one of eight clothing categories (i.e., blouse, T-shirt, cheongsam, coat, camisole, dress, fur coat, and bottoming shirt). Fig. 5 shows some example clothes in different categories. There are nine collar types for these images. Example clothes in different collar types are shown in Fig. 6. For the sleeve types, here we focus on the length of sleeves rather than their patterns. The collected images are divided in three sleeve types, i.e., sleeveless, short-sleeve, and long-sleeve. Example clothes in different sleeve types are shown in Fig. 7.

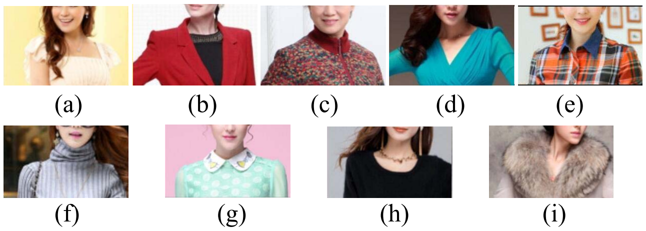


Fig. 6. Examples of collar types in the CCAP dataset. (a) Square neck collar. (b) Notched collar. (c) Stand-up collar. (d) V-shape collar. (e) Spread collar. (f) Turtleneck collar. (g) Peter pan collar. (h) Round collar. (i) Fur collar.

Fig. 7. Examples of sleeve types [(a) sleeveless, (b) short sleeved, and (c) long sleeved] in the CCAP dataset.

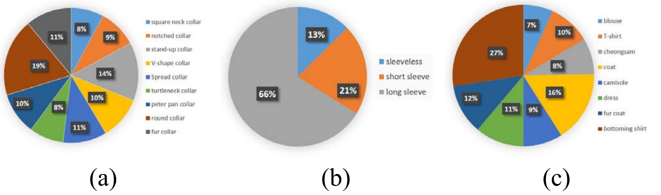
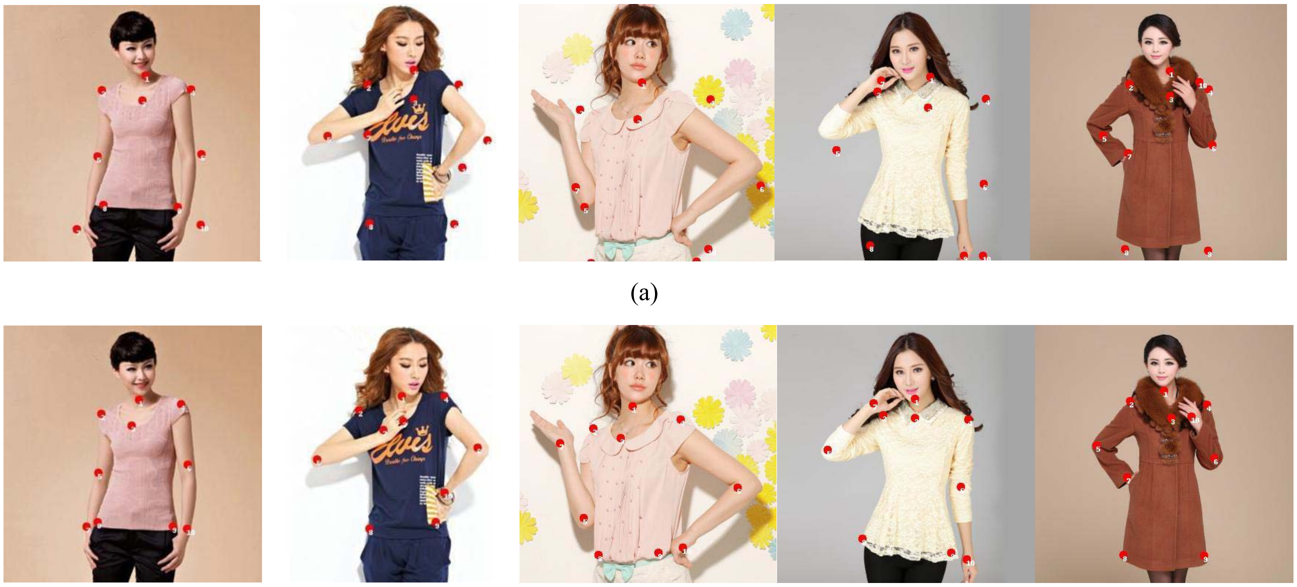


Fig. 8. Distributions of categories/attributes in the CCAP dataset. (a) Collar types. (b) Sleeve types. (c) Clothing category.

We manually annotate each of the 25 500 images with its clothing category, collar type, and sleeve type. The distributions of these categories and attributes are shown in Fig 8. Then we randomly split the whole dataset into a training set and test set, approximately with a ratio 5:1, i.e., 21 000 images in the training set, and 4500 images in the test set.

VI. Experiments Results

In this section, we evaluate the proposed clothing landmark detection method and the clothing category and attribute prediction method on both publicly available and our collected FCLB and CCAP datasets. The results show the following.



(a)

(b)

Fig. 9. Example results of landmark detection on the FLCD dataset. (a) Example results of the baseline method. (b) Example results of the proposed method.

1) For 10-point clothing landmark detection, the proposed method with carefully designed deep networks and the prior of landmarks' associations outperforms the baseline methods.
2) For clothing category and attribute prediction, the proposed method, which uses estimated landmark points to obtain more robust feature representations, shows superior performance gains over the baseline methods.

The proposed methods and the baseline methods used in this paper are implemented based on the open source Caffe [15] framework. In all of these methods, the network parameters are initialized with the pretrained VGG-16 [31] model. We set the learning rate in fine tuning to be 0.0001, and the momentum to be 0.9.

A. Experiments for Landmark Detection

1) Evaluation Metric: Following the practice of landmark detection in other tasks (e.g., face alignment [19]), we use the *normalized mean error* as the evaluation metric in the clothing landmark detection experiments. The normalized mean error is defined as

$$\text{mean error} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{\left\|L_i^{(j)} - L_i^{\hat{(j)}}\right\|_2^2}{k \times D_i} \tag{10}$$

where $L_i^{(j)}$ represents the coordinate of the jth ground-truth (i.e., human annotations) landmark point in the ith test image, $L_i^{\hat{(j)}}$ represents the coordinate of the jth predicted landmark point in the ith test image, k is the number of landmark points, n is the number of test images, and D_i represents a normalized factor defined as Euclidean distance between left-shoulder landmark point and the right-shoulder landmark point in the ith test image.

TABLE II
COMPARISON RESULTS WITH RESPECT TO NORMALIZED MEAN ERROR ON THE FCLD DATASET

method	mean error
FashionNet [24]	0.115
our method	0.060

We carefully reimplement the landmark detection method proposed in [24] as the baseline method in comparison.[2] This baseline method adopts a network structure similar to VGG-16 [31], where the last fully connected layer (i.e., fc7 in VGG-16) is followed by a regression loss that measures the mean square errors between the predicted landmark coordinate and the manually annotated landmark coordinates.

In the experiments, we use the same training/test split in both the baseline method and the proposed method. The comparison results with respect to normalized mean errors on the FCLD dataset are shown in Table II. As can be seen, the proposed clothing landmark detection method achieves a mean error of 0.060, which shows superior performance gains against the baseline method whose mean error is 0.115. Fig. 9 shows some example results of clothing landmark detection.

B. Experiments for Category and Attributes Classification

1) Results on CCAP Dataset: Here, we consider the tasks of classifying clothing categories, collar types, and sleeve types. Following the practice in object classification (e.g., [17], [31], and [34]), for each task, we use the top-1 accuracy as the evaluation metric.

We implement three baseline methods. Since the proposed network for clothing category and attribute classification is based on VGG-16 [31], for a fair comparison, the baseline

[2] The network structure of the FashionNet method in [24] has three branches. In our implementation, we only use the branch for landmark locations, and remove the landmark visibility and the other two branches.

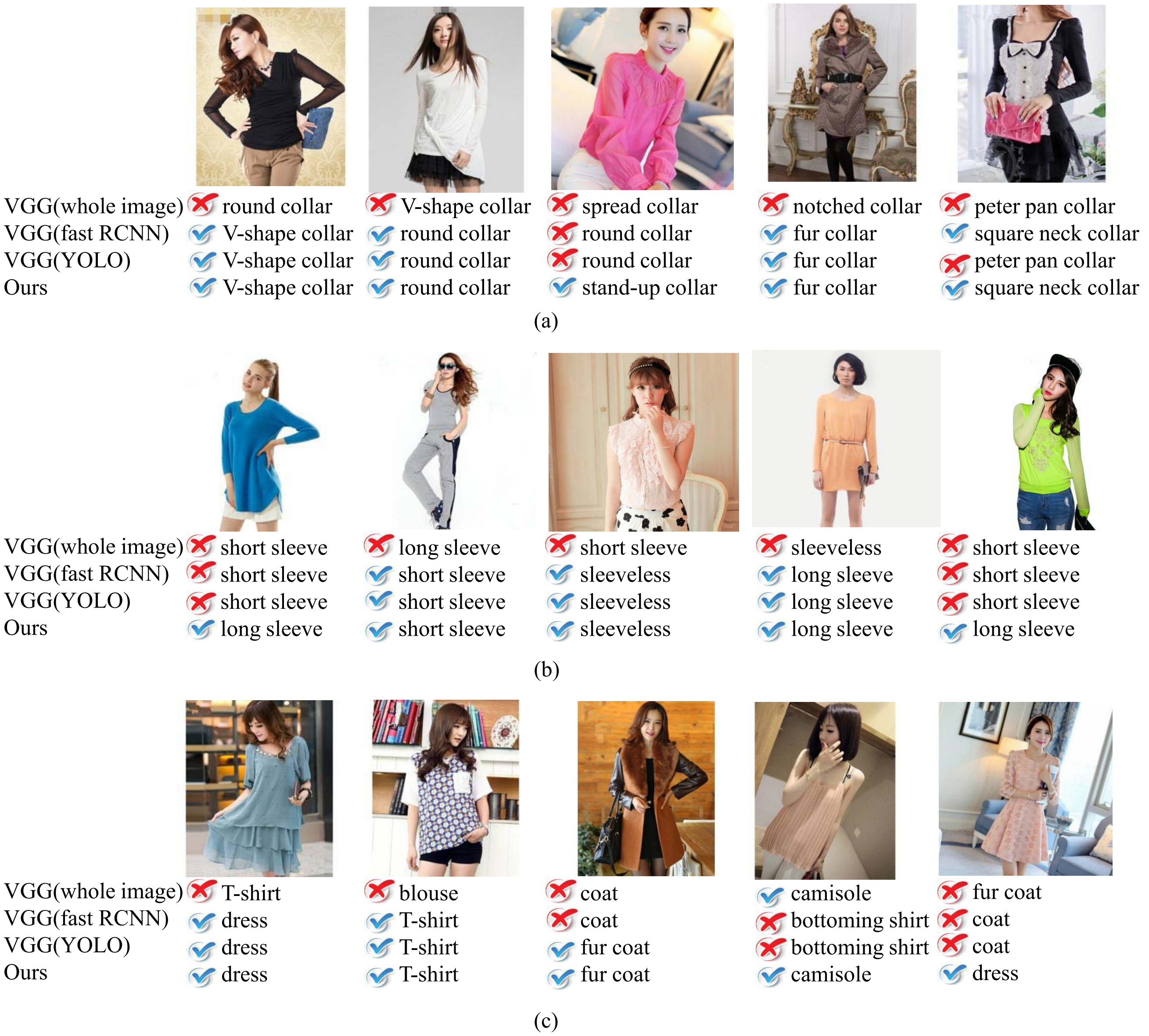

VGG(whole image) ✗ round collar ✗ V-shape collar ✗ spread collar ✗ notched collar ✗ peter pan collar
VGG(fast RCNN) ✓ V-shape collar ✓ round collar ✗ round collar ✓ fur collar ✓ square neck collar
VGG(YOLO) ✓ V-shape collar ✓ round collar ✗ round collar ✓ fur collar ✗ peter pan collar
Ours ✓ V-shape collar ✓ round collar ✓ stand-up collar ✓ fur collar ✓ square neck collar
(a)
VGG(whole image) ✗ short sleeve ✗ long sleeve ✗ short sleeve ✗ sleeveless ✗ short sleeve
VGG(fast RCNN) ✗ short sleeve ✓ short sleeve ✓ sleeveless ✓ long sleeve ✗ short sleeve
VGG(YOLO) ✗ short sleeve ✓ short sleeve ✓ sleeveless ✓ long sleeve ✗ short sleeve
Ours ✓ long sleeve ✓ short sleeve ✓ sleeveless ✓ long sleeve ✓ long sleeve
(b)
VGG(whole image) ✗ T-shirt ✗ blouse ✗ coat ✓ camisole ✗ fur coat
VGG(fast RCNN) ✓ dress ✓ T-shirt ✗ coat ✗ bottoming shirt ✗ coat
VGG(YOLO) ✓ dress ✓ T-shirt ✓ fur coat ✗ bottoming shirt ✗ coat
Ours ✓ dress ✓ T-shirt ✓ fur coat ✓ camisole ✓ dress
(c)

Fig. 10. Some example results of clothing category and attribute prediction on the CCAP dataset. (a) Example results on collar classification. (b) Example results on sleeve classification. (c) Example results on clothing category classification.

methods are also based on VGG-16. First, we construct a VGG-16 network for clothing category and attribute classification, where we replace the 1000-node fully connected layer in VGG-16 by the three sibling layers as shown in Fig. 3. For the first baseline, we use the original clothing images (without bounding box detection) to finetune this VGG-16 network and then conduct classification. For the second baseline, we use Fast-RCNN [29] to detect bounding boxes that contain clothes, resize each of the clips in the bounding boxes to 256 × 256 (e.g., resize the longer edge to 256 and resize the shorter edge proportionally, the rest region is padded by black pixels), use these clips to finetune the VGG-16 network and conduct classification. Similarly to the second baseline, for the third baseline, we use YOLO [28] to detect bounding boxes that contain clothes, resize each of the clips in the bounding boxes to 256 × 256, use these clips to finetune the VGG-16 network and conduct classification.

The compared results on the CCAP dataset are presented in Table III. We can see that the proposed method performs better than the baseline methods on all of the three classification tasks. For example, in collar type classification, the proposed method achieves a top-1 accuracy of 86.0%, indicating a relative increase of 5.4% over the second best baseline. Fig. 10 shows some example results on the CCAP dataset. These results verify that introducing the information of estimated landmark can help to improve the performance of clothing category and attribute classification.

2) Experiments on DeepFashion Dataset: We also evaluate the proposed method and the baselines on the publicly available benchmark for clothing category and attribute classification in DeepFashion dataset [24]. This benchmark dataset consists of 289 222 clothing images in total, including 209 222 images for training, 40 000 images for validation, and 40 000

TABLE III
COMPARISON RESULTS OF CLOTHING CATEGORY AND ATTRIBUTE CLASSIFICATION ON THE CCAP DATASET

model	collar	sleeve	clothing
VGG (whole image)	79.4	94.0	78.4
VGG (fast RCNN)	80.8	94.2	79.0
VGG (YOLO)	81.6	94.2	79.7
our method	**86.0**	**95.2**	**81.2**

TABLE IV
COMPARISON RESULTS ON THE DEEPFASHION DATASET

model	top-1	top-3	top-5
FashionNet [24]	N/A	82.58	90.17
VGG (whole image)	67.97	86.89	92.89
VGG (fast RCNN)	67.52	86.53	92.68
VGG (YOLO)	67.60	86.67	92.72
our method	**70.59**	**88.60**	**94.09**

images for test. For the proposed method and the baseline methods, we use the same train/test split as in [24].

We use the two baseline methods based on VGG-16 as in Section VI-B1. In addition, we use the FashionNet method proposed in [24] as a baseline method.[3] Following the practice in object classification (e.g., [17], [31], and [34]), for each task, we use the top-1 accuracy, top-3 accuracy, and top-5 accuracy as the evaluation metrics.

As we can see in Table IV, the comparison results show that the proposed method outperforms the baseline methods by a clear margin. For example, the top-3 accuracy of the proposed method is 88.60%, which indicates a relative increase of 2.2% over the second best baseline. The proposed method achieves 94.09% with respect to the top-5 accuracy, which shows a relative increase of 1.5% over the second best baseline. These results verify the effectiveness of the proposed method.

VII. CONCLUSION

In this paper, we developed an intelligent system based on deep neural networks to predict ten landmark points in a piece of upper-body clothing. With these predicted landmark points, one can locate the regions of interest in clothes (e.g., the regions of the sleeves, the collars, or the whole clothes), which help in classifying clothing categories and attributes. The landmark points predicted from the deep learner model can be regarded as a set of feature representations of the regions of interest. It is interesting and beneficial to learn that deep learning techniques can extract "good" features for robust modeling. It should be pointed out that our released datasets, including a large-scale dataset with 145 000 upper-body clothing images for clothing landmark detection, and a dataset with 25 500 images for clothing category and attribute prediction, can be employed as benchmark datasets for further studies.

It is believed that such a detecting system is useful and valuable for real-world applications in clothing industry. Therefore, further researches on this topic should be stressed by looking at the system performance improvements and the implementation aspects.

REFERENCES

[1] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Proc. Eur. Conf. Comput. Vis.*, Zürich, Switzerland, 2014, pp. 584–599.

[2] L. Bossard *et al.*, "Apparel classification with style," in *Proc. Asian Conf. Comput. Vis.*, Daejeon, South Korea, 2012, pp. 321–335.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 7291–7299.

[4] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *Proc. Eur. Conf. Comput. Vis.*, Zürich, Switzerland, 2014, pp. 109–122.

[5] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. Eur. Conf. Comput. Vis.*, Florence, Italy, 2012, pp. 609–623.

[6] Q. Chen *et al.*, "Deep domain adaptation for describing people based on fine-grained clothing attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 5315–5324.

[7] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 3286–3293.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, 2005, pp. 886–893.

[9] Q. Dong, S. Gong, and X. Zhu, "Multi-task curriculum transfer deep learning of clothing attributes," in *Proc. IEEE Win. Conf. Appl. Comput. Vis. (WACV)*, Santa Rosa, CA, USA, 2017, pp. 520–529.

[10] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan, "Style finder: Fine-grained clothing style detection and retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Portland, OR, USA, 2013, pp. 8–13.

[11] R. Girshick, "Joint cascade face detection and alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[13] C. Huang *et al.*, "Large-scale semantic Web image retrieval using bimodal deep learning techniques," *Inf. Sci.*, vols. 430–431, pp. 331–348, Mar. 2018.

[14] J. Huang, R. S. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1062–1070.

[15] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014. [Online]. Available: https://arxiv.org/abs/1408.5093

[16] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg, "Hipster wars: Discovering elements of fashion styles," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 472–488.

[17] 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.

[18] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 3270–3278.

[19] H. Lai *et al.*, "Deep recurrent regression for facial landmark detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 5, pp. 1144–1157, May 2018.

[20] H. Lai, P. Yan, X. Shu, Y. Wei, and S. Yan, "Instance-aware hashing for multi-label image retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2469–2479, Jun. 2016.

[21] K. Laenen, S. Zoghbi, and M. F. Moens, "Cross-modal search for fashion attributes," in *Proc. KDD Workshop Mach. Learn. Meets Fashion*, 2017, pp. 1–10.

[22] X. Liang *et al.*, "Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1175–1186, Jun. 2016.

[23] S. Liu *et al.*, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 3330–3337.

[24] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 1096–1104.

[3] Note that the source code of FashionNet is not publicly available. For the clothing category and attribute classification benchmark in DeepFashion, since we use the same train/test split as that in [24], we directly cite the results of FashionNet from [24].

[25] Z. Liu, S. Yan, X. Wang, and X. Tang, “Fashion landmark detection in the wild,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 229–245

[26] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 3431–3440.

[27] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[28] 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.*, Las Vegas, NV, USA, 2016, pp. 779–788.

[29] G. Ross, “Fast R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2016, pp. 229–245.

[30] M. Shao, L. Li, and Y. Fu, “What do you do? Occupation recognition in a photo via social context,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, 2013, pp. 3631–3638.

[31] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[32] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun, “Neuroaesthetics in fashion: Modeling the perception of fashionability,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 869–877.

[33] Z. Song, M. Wang, X.-S. Hua, and S. Yan, “Predicting occupation via human clothing and Contexts,” in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, 2011, pp. 1084–1091.

[34] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1–9.

[35] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 1701–1708.

[36] A. Veit *et al.*, “Learning visual clothing style with heterogeneous dyadic co-occurrences,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 4642–4650.

[37] X. Wang, and T. Zhang, “Clothes search in consumer photos via color matching and attribute learning,” in *Proc. Int. Conf. Multimedea*, Scottsdale, AZ, USA, 2011, pp. 1353–1356.

[38] Y. Wei *et al.*, “Cross-modal retrieval with CNN visual features: A new baseline,” *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

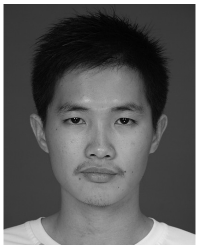
[39] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, “Supervised hashing for image retrieval via image representation learning,” in *Proc. AAAI Conf. Artif. Intell.*, Quebec City, QC, Canada, 2014, pp. 2156–2162.

[40] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, “Learning from massive noisy labeled data for image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 2691–2699.

[41] J. Yu, X. Yang, F. Gao, and D. Tao, “Deep multimodal distance metric learning using click constraints for image ranking,” *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4014–4024, Dec. 2017.

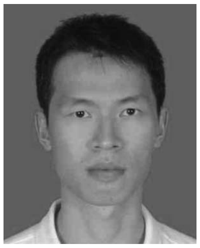
Chang-Qin Huang (M’17) received the Ph.D. degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2005. He is currently a Professor with South China Normal University, Guangzhou, China. He was a Visiting Scientist with Zhejiang University. His current research interests include semantic information retrieval, big data in education, and service-oriented computing.

Dr. Huang was a recipient of the Pearl River Scholarship. He is a member of CCF and ACM.



Ji-Kai Chen received the B.S. degree in electronic engineering from Sun Yat-sen University, Guangzhou, China, in 2014, where he is currently pursuing the master’s degree with the School of Data and Computer Science.

His current research interests include deep learning methods and computer vision.



Yan Pan received the B.S. degree in information science and the Ph.D. degree in computer science from Sun Yat-sen University, Guangzhou, China, in 2002 and 2007, respectively.

He is currently an Associate Professor with Sun Yat-sen University. His current research interests include hashing methods, machine learning algorithms, learning to rank, and computer vision.

Dr. Pan was a recipient of the object categorization task in PASCAL VOC 2012 Challenge. He has served as a reviewer for several conferences and journals.

Han-Jiang Lai received the B.S. degree in software engineering and the Ph.D. degree in computer software and theory from Sun Yat-sen University, Guangzhou, China, in 2009 and 2013, respectively.

He is currently an Associate Professor with the School of Data and Computer Science, Sun Yat-sen University. His current research interests include hashing methods, computer vision, optimization algorithms, and learning to rank.

Jian Yin received the B.S., M.S., and Ph.D. degrees from Wuhan University, Wuhan, China, in 1989, 1991, and 1994, respectively, all in computer science.

In 1994, he joined Sun Yat-sen University, Guangzhou, China, where he is currently a Professor with the School of Information Science and Technology. He has published over 100 refereed journal and conference papers. His current research interests include data mining, artificial intelligence, and machine learning.

Dr. Yin is a Senior Member of China Computer Federation.

Qiong-Hao Huang received the master’s degree from the School of Computer Science, South China Normal University, Guangzhou, China, where he is currently pursuing the Ph.D. degree with the School of Information Technology in Education.

His current research interests include cloud computing, big data processing, and machine learning algorithms.