Neurocomputing 282 (2018) 98–110

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Watch fashion shows to tell clothing attributes

Sanyi Zhang[a,b], Si Liu[b], Xiaochun Cao[b], Zhanjie Song[c,*], Jie Zhou[d]

[a] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
[b] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[c] School of Mathematics, Tianjin University, Tianjin 300354, China
[d] Department of Automation, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 23 February 2017
Revised 27 September 2017
Accepted 9 December 2017
Available online 13 December 2017

Communicated by Dr. M. Wang

Keywords:
Unsupervised learning
Clothing attribute
Triplet neural network
Semi-supervised
Fashion shows

ABSTRACT

In this paper, we propose a novel semi-supervised method to predict clothing attributes with the assistance of unlabeled data like fashion shows. To this end, a two-stage framework is built, i.e., the unsupervised triplet network pre-training stage that ensures frames in the same video having coherent representations while frames from different videos having larger feature distances, and a supervised clothing attribute prediction stage to estimate the value of attributes. Specifically, we first detect the clothes of frames in the collected 18,737 female fashion shows and 21,224 male fashion shows which contain no extra labels. Then a triplet neural network is constructed via embedding the temporal appearance consistency between frames in the same video and the representation gap in different videos. Finally, we transfer the triplet model parameters to multi-task clothing attribute prediction model, and fine-tune it with clothing images holding attribute labels. Extensive experiments demonstrate the advantages of the proposed method on two clothing datasets.

1. Introduction

With the popularity of online shopping especially for clothing, clothing related research [1,2], particularly clothing attribute prediction [3–9], has become a hot topic in the field of multimedia and computer vision. There are various related applications including: clothing recommendation [10,11], clothing retrieval [5,6], person re-identification [12,13], fashion parsing [14–20].

The general framework of existing work [3,21] for attribute learning is that they train the classifier with hand-crafted features [22,23]. However, the attribute prediction accuracy has been restricted by the limited discrimination of hand-crafted features.

To improve the discriminative feature representation, the deep convolutional neural network (DCNN) [24–28] is introduced for feature learning. The flowchart of attribute learning based on CNN is to train a powerful DCNN model with plenty of labeled images to learn a better feature representation from raw image for the attributes. More labeled images used, a better DCNN model maybe obtained. However, it is not an easy work to obtain large amounts of supervised data, especially for clothing attributes. So how to release humans from laborsome labeling is a meaningful problem. A natural input of vision system, which needs computer understanding the physical world, is video. Video itself provides temporal coherence that can be regarded as supervised cue for visual representation learning. But how to utilize this useful unsupervised information to reduce human labeling burden and learn a powerful feature is an open problem.

To address the above-mentioned challenges, in this paper, we propose a semi-supervised approach which first uses a large amount of unlabeled videos to pre-train a CNN model, and then fine-tunes the model with relatively small number of labeled images for clothing attribute prediction. As shown in Fig. 1, in the beginning, the clothing video datasets are collected from the Internet. Here, each video contains a single person walking on the catwalk. We extract keyframes from each video through appropriate choosing strategies e.g. uniformly extracting fixed frames, to generate the clothing pairs. Positive clothing pairs come from the same videos while negative pairs are selected from different videos. Fast R-CNN [29] approach is then applied to train a clothing detector with annotated clothing object bounding box location. Then, we design a triplet ranking ConvNet for training with unlabeled videos. The key motivation is that detected clothing should be similar with the one from same video, while dissimilar with the one from different videos. By exploring on this video context information, we use a triplet ranking loss while training to learn discriminative clothing feature. Finally, the unsupervised ConvNet information is transferred to the clothing attribute prediction task.

* Corresponding author.
E-mail addresses: zhangsanyi@tju.edu.cn (S. Zhang), liusi@iie.ac.cn (S. Liu), caoxiaochun@iie.ac.cn (X. Cao), zhanjiesong@tju.edu.cn (Z. Song), jzhou@tsinghua.edu.cn (J. Zhou).

https://doi.org/10.1016/j.neucom.2017.12.027
0925-2312/

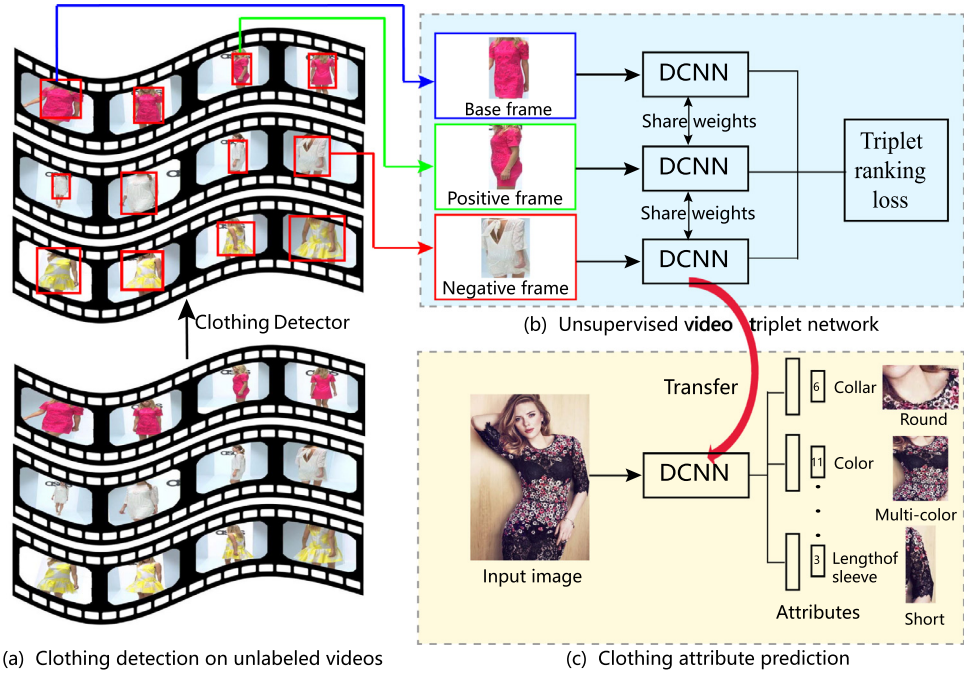


Fig. 1. Overview of our proposed approach. (a) We first detect the clothing in the unlabeled videos by the clothing detector. Then, (b) the triplet DCNN network is designed to train with detected clothing patches, which is based on the triplet ranking loss. (c) Adapting the triplet DCNN network for the initialization of clothing attribute prediction network and fine-tune the network with labeled attribute clothing images to predict the clothing attributes.

We evaluate the performance of the unsupervised ConvNet in two clothing attribute datasets, woman and man clothing. We also evaluate on two different deep ConvNet models based methods, AlexNet [8,25] and VGG16 [30,31]. Experimental results have certificated the effectiveness of the unsupervised trained model with video data.

The main contributions can be summarized as follows:

(a) We collect two new clothing video datasets for unsupervised learning. And two new clothing image datasets are annotated with 10 pre-defined clothing attributes. These datasets will be released to the public.

(b) A novel semi-supervised approach is proposed for clothing attribute learning, which makes full use of large scale unlabeled data. Concretely, two stages are designed, unsupervised triplet network to learn the pattern of discriminative clothing representation from unlabeled data and a multi-task framework to predict clothing attributes.

(c) We demonstrate the effectiveness of the video context. And it is also a good solution to alleviate costly human annotation.

The rest of this paper is organized as follows. In Section 2, we review the most relevant works on clothing attribute study and video context study. In Section 3 we introduce two new clothing video datasets. In Section 4 we present our method including clothing detection, triplet network and clothing attribute prediction. Section 5 gives both qualitative and quantitative experimental results and analyses. We conclude this paper in Section 6.

2. Related work

Here we review the related work from three aspects, clothing attribute study, unsupervised learning and video context study.

Clothing attribute study: Clothing attribute prediction can benefit various clothing relative applications. Chen et al. [4] proposed a fully automated system that describes the clothing appearance with a list of semantic attributes. And clothing attributes can be directly utilized in a novel application of dressing style analysis. Liu et al. [6] made use of labeled clothing attributes in the task of cross-scenario online shopping clothing retrieval. These traditional methods use hand-crafted feature to train attribute classifier. Recently, Chen et al. [8] proposed a deep domain adaptation approach describing people with fine-grained clothing attributes. Huang et al. [5] integrated attribute-guide learning into dealing cross-domain image retrieval problem. They designed a dual attribute-aware ranking network which combines attributes with visual information. Liu et al. [30,32] introduced a DeepFashion dataset which provided massive attribute annotation for multi-label attribute prediction. These methods all use labeled attributes images as an important clue in CNN training and can be applied in different applications. Our work is focused on how to make use of unlabeled clothing videos assist in improving the performance of multi-task clothing attribute prediction.

Unsupervised learning: Unsupervised representation learning has gained more attention in visual tasks. The goal of unsupervised learning is to learn a feature embedding which the distances of similar images are smaller than dissimilar images. Some researchers solve this problem through exploring from the spatial context information of images. [33] proposed to learn patch representation through judging a patch's position while randomly give one of eight spatial configuration. [34] proposed a context-free network (CFN) which is trained to solve Jigsaw puzzles as a pretext task. There are also some works utilizing video information, next we will give a brief introduction about this.

Video context study: Annotating images with proper tags often takes lots of time and effort. So researchers start to consider making use of thousands or millions of videos when training ConvNet. The sequential information of video owned is that the frames change smaller in the same video. The similarity of the frames in the same video is larger than frames from different videos. Some works prove that video-context information can help to improve the performance. Liang et al. [35] proposed computational baby learning framework which mimics baby learning process by learning with video contexts. Through the video contexts baby learning method can get better object detection results on Pascal VOC-07/10/12 object detection datasets. This method is based

Fig. 2. Some examples of woman clothing video dataset. Frames in every column come from one video.

on a pre-trained CNN model on ImageNet classification. An interesting work [36] proposed a method to learn manipulation action plans from unconstrained videos. Misra et al. [37] learned to discover multiple unknown objects from sparsely labeled videos. How to use the unsupervised video context information train a discriminative CNN model without using label information? Triplet ranking loss is a superior design to solve this problem. Triplet ranking loss can constrain the distance of the similar frames and dissimilar frames. It can be optimized by stochastic gradient descent (SGD) method. The triplet ranking loss has been used in many tasks, for example, image retrieval [38,39], face recognition [40], hashing code generation [41], object detection [42]. These works use triplet ranking loss as the final loss of the neural network's last layer. The most relevant work is [42], this paper proposed an unsupervised learning approach using 100K unlabeled videos to learn a visual representation. They designed a siamese-triplet network with triplet ranking loss to train a CNN model. They verified the effectiveness in the tasks of the object detection and surface normal estimation. Our work differs from their work in that we focus on how to utilize unsupervised video context for a specific domain, i.e., clothing attribute prediction.

3. Clothing video dataset

Two clothing video datasets are collected, i.e., woman and man clothing dataset which are composed of fashion shows. The number of woman clothing video set is 18,737 and a man clothing video set contains 21,224 videos. All videos are downloaded from the Internet website *asos*[1] which provides many videos showing clothes online. Generally, each video contains a single woman or man walking on the catwalk. The female walks from the backstage and then gradually move toward the front, turn around and go back. Some examples are shown in Figs. 2 and 3.

(a) Woman clothing video dataset: Several important statistic results including video duration and corresponding number of videos are shown in Table 1. We uniformly extract keyframes from woman videos. Considering the small variance of adjacent frames, we extract one frame from every 25 frames in each video. In the woman clothing video dataset, the duration is not long and the background is often fixed, but the important clues of videos are the multiple views of clothing (as shown in Fig. 2). To ensure each video keyframe containing clothing object, we would adopt different strategies uniformly extracting keyframes each video according to the duration. For the short videos, such as the ones in 7s, 8s, we extract frames from the beginning of videos. While for longer videos, the beginning frames are often background, we choose keyframes by skipping this duration. Finally we select seven frames from each video by eliminating the frames without clothing.

Table 1
The distribution of video duration in woman and man clothing video datasets.

Woman clothing video		Man clothing video	
Duration (s)	#videos	Duration (s)	#videos
≤ 9	2135	≤ 10	915
10	6626	11	4147
11	6656	12	15,558
≥ 12	3320	≥ 13	604
Total	18,737	Total	21,224

(b) Man clothing video dataset: For the man clothing videos, we also analyze the distribution of man clothing videos' durations in Table 1. According to man videos' content, we adopt a different strategy to extract keyframe. We use *ffmpeg*[2] to extract keyframes from man videos.

As the extracted keyframes may have unsatisfactory cases like containing no clothing object, we employ clothing detection algorithm which will be introduced in the following section.

4. Methodology

Our goal is to develop a semi-supervised deep model for improving the performance of clothing attribute prediction via the unlabeled videos. First, the clothing detector is trained to detect clothing in each keyframe based on fast R-CNN [29]. Second, based on these detected clothing, we generate triple pairs which consist of positive and negative pairs to train a triplet model. After that, we initialize the proposed clothing attribute prediction network via trained triplet network. Moreover, images with labeled attributes

[1] http://www.asos.com.

[2] https://www.ffmpeg.org/.

Fig. 3. Some examples of man clothing video dataset. Frames in every column come from one video.

are used to fine-tune the attribute learning network and evaluate the attribute prediction performance.

4.1. Clothing detection

We follow the object detection approach fast R-CNN [29] to get clothing bounding boxes. Fast R-CNN combines the advantages of R-CNN [43] and SPPnet [44], in terms of the fast speed and improved accuracy. According to the fast R-CNN framework, clothing proposals are first generated by selective search approach [45]. In our method, we revise the default fast mode for proposal generation with three color types (HSV, Lab and rgI) and three thresholds (k = 50, 100, 150). Then the extracted proposals are processed through a VGG16 model [31] for fast R-CNN training. Fast R-CNN has multiple-task losses, one is classification task and the other is object bounding box regression task. We predict the existence of clothing, the four-dimension coordinates of bounding box and the corresponding object confidence.

For our clothing detector training, we collect two new clothing datasets from the Internet, i.e., a woman clothing image dataset and a man clothing image dataset. Both datasets contain 13,500 clothing images with annotations of the upper-clothing bounding box coordinates. We adopt the same strategy for two clothing datasets as separating the clothing images into two subsets, 10,200 images for training and 3300 images for testing. Here, the object classes only contain two classes, namely clothing vs. background. Similar with fast R-CNN, we train the clothing detector with pre-trained VGG16 network [31] parameters initialization. The final detection average precision (AP) of the woman clothing detector in the test set is 90.61%, and the man clothing detector is 95%. We believe the detection accuracy is sufficient for the further processing.

Then we apply the trained clothing detector on clothing video frames. First, we use the selective search approach [45] to generate the video keyframes' object proposals. Then we feed these proposals into the clothing detection model and obtain the clothing object localization in the video frames. Because we care more about the detection precision instead of recall and the clothing detector is not perfect with 90.61% or 95% test average precision, we only keep the video frames according to object confidence larger than 0.95. In this way, certain false negatives are removed. However, considering the large amount of collected videos, the remaining pairs are sufficient to train a robust model. Fig. 4 shows some examples of two woman clothing videos' frames and the detected results of fast R-CNN. Fig. 5 shows some examples of two man clothing videos' frames and the fast R-CNN detected results.

4.2. Triplet network

In this part, we introduce an approach to train a triplet DCNN network which makes use of large scale of unlabeled video data from the web. A label-free neural network is constructed to force the similarity between two frames in the same video to be higher than those from different videos. To this end, the triplet network consists of three input frames, among which the first two belong to one video and the third one comes from another random selected video. We enforce that the first frame is closer to the second frame than the third frame. The three frames share the same network parameters. We adopt two popular deep ConvNet as the base network. One is based on the AlexNet [25] framework and the other is based on VGG16 [31] framework. We use the same convolutional layers of the AlexNet and VGG16, and then two full connected layers are followed. The neuron number of these two full connected layers are 4096 and 1024, respectively. The triplet ranking loss function is designed over the 1024 feature space. Fig. 6 shows an example of AlexNet framework with detailed network parameters.

For AlexNet [25] framework in our experiments, we take the input video frame resizing to 256×256×3 and extract 10 patches with 227 × 227 (the four corner patches and the center patch in original as well as their horizontal reflections). For VGG16 [31] framework, we take the input video frame resizing to 256 × 256 × 3 and extract 10 cropped patches with 224 × 224.

4.2.1. Triplet ranking loss function

We use X to indicate the video frame, $F(X)$ as the feature representation. Then we define the distance between two frames X^a and X^b based on cosine distance space as:

$$D\left(X^a, X^b\right) = 1 - \frac{X^a \cdot X^b}{\|X^a\| \left\|X^b\right\|} = 1 - \frac{\sum_{i=1}^{n} X_i^a \times X_i^b}{\sqrt{\sum_{i=1}^{n} \left(X_i^a\right)^2} \times \sqrt{\sum_{i=1}^{n} \left(X_i^b\right)^2}}. \quad (1)$$

A triple frames X_i, X_i^p, X_i^n, where X_i is ith frame from *video*, X_i^p is the frame from the same *video*, and X_i^n is randomly selected from other videos. The estimation of among three frames should meet the requirement:

$$D\left(X_i, X_i^p\right) + \alpha < D\left(X_i, X_i^n\right), \quad (2)$$

where α is a margin between distance of the positive pair and the negative pair.

Fig. 4. Example of two woman clothing videos' frames and the results of the fast R-CNN results. The first row and third row are the original woman video frames, the second row and fourth row are the fast R-CNN detected results.

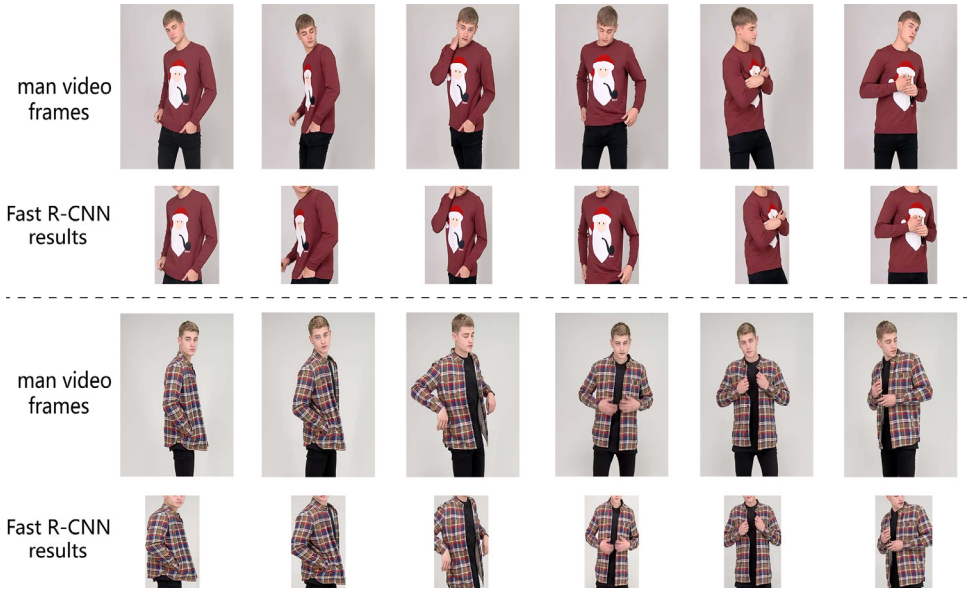


Fig. 5. Example of two man clothing videos' frames and the results of the fast R-CNN results. The first row and third row are the original man video frames, the second row and fourth row are the fast R-CNN detected results.

Then we define the triplet ranking loss function based on hinge loss as:

$$\ell\left(X_i, X_i^p, X_i^n\right) = \max\left\{0, D\left(X_i, X_i^p\right) - D\left(X_i, X_i^n\right) + \alpha\right\}. \tag{3}$$

Substituting Eq. (1) into Eq. (3), so the loss function can be written as:

$$\ell\left(X_i, X_i^p, X_i^n\right) = \max\left\{0, C_i \cdot C_i^n - C_i \cdot C_i^p + \alpha\right\}, \tag{4}$$

where $\frac{X_i}{\sqrt{\sum_{i=1}^n (X_i)^2}}$, $\frac{X_i^p}{\sqrt{\sum_{i=1}^n (X_i^p)^2}}$, $\frac{X_i^n}{\sqrt{\sum_{i=1}^n (X_i^n)^2}}$ are denoted as C_i, C_i^p, C_i^n.

In our experiment, we set $\alpha = 0.5$. Since the triplet ranking loss is convex, we solve it by stochastic gradient descent (SGD).

The gradients with respect to C_i, C_i^p, C_i^n are

$$\begin{aligned}
\frac{\partial \ell}{\partial C_i} &= \left(C_i^n - C_i^p\right) I_{condition>0} \\
\frac{\partial \ell}{\partial C_i^p} &= (-C_i) I_{condition>0} \\
\frac{\partial \ell}{\partial C_i^n} &= (C_i) I_{condition>0}
\end{aligned} \tag{5}$$

where $condition = C_i \cdot C_i^n - C_i \cdot C_i^p + \alpha$. If $condition > 0$ then $I = 1$, otherwise $I = 0$.

4.2.2. Triplet selection

In the triplet ConvNet training, it is important to generate the triplet pairs. We choose the frames in the same video as positive

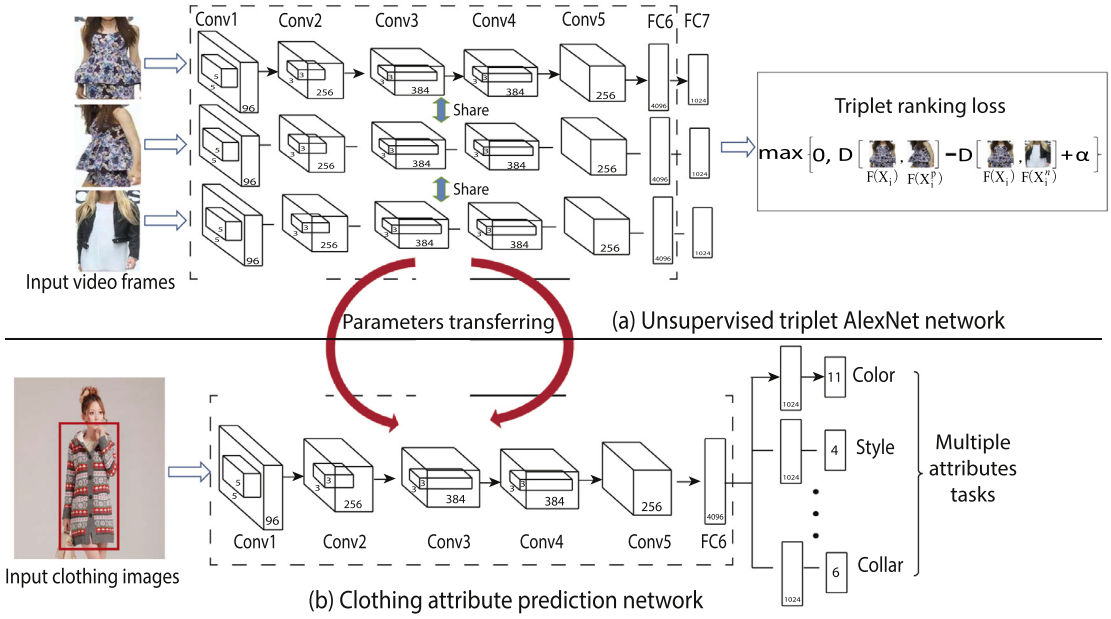


Fig. 6. The framework of proposed two-stage approach. (a) Unsupervised triplet network. Here we use AlexNet framework as a toy example. Three input frames go through three convolutional neural networks which share the same parameters. Three frames generate three output features $F(X_i)$, $F(X_i^p)$ and $F(X_i^n)$, the triplet ranking loss is calculated based on them. After the triplet CNN model is trained, it is transferred toward (b) clothing attribute prediction task. The attribute prediction network shares the five convolutional layers and first 4096 fully connected layer with the triplet network. Multiple tasks (totally 10 tasks) are defined. Each task corresponds to a specific attribute prediction task. Every task contains two fully connected layers, one has 1024 neurons and the other owns the number of corresponding attribute values. Every task takes the softmax cross-entropy loss.

pairs. The selection of the negative pair is critical for the discriminative training. Here we introduce two types of choosing negative pairs, i.e., random selection and hard negative selection.

(a) Random selection

We select the first frame from a video as the base reference frame. The positive frames are the remaining frames in the same video. For each positive pair in the mini-batch, we randomly select n sample frames as the third negative frames from different videos. In the experiment, we set the number of random sample frames n as 4.

(b) Hard negative selection

After 10 epochs of training, the randomly selected triplet pairs tend to be convergent. We can infer that the trained random triplet model has general ability of feature learning. We take hard negative mining strategy to generate new triplet pairs. The negative pairs which are close to the positive pairs are selected. We only perform optimization and learn on the hardest negative frames that the loss is highest in a mini-batch.

In order to increase the difficulty of training, for each positive pair, we select top n highest negative samples according to the loss function in Eq. (3) as the final negative frames. We finally compute our loss function with these n hard negative triplet pairs. We also use n as 4 in the experiment.

4.3. Adapting for clothing attribute prediction

After the triplet model is trained through the videos, we transfer this model to related tasks for better performance. In this paper, we adapt the pre-trained video model on clothing image attribute prediction task. We use the pre-trained triplet model for initialization, and change the learned video context information to the supervised clothing attribute prediction task.

For the clothing attribute prediction task, we design a multiple attribute prediction framework for fine-tuning. As Fig. 6 shows, we share the convolutional layers and the first fully connected layer parameters with the trained triplet ConvNet. Then we define two new fully connected layers for each specific attribute prediction task. The first fully connected layer is designed with 1024 neurons and the second one has an adequate number of neurons according to the defined kinds for each attribute. The loss is softmax cross-entropy loss. Thus this model simultaneously learns multiple attribute prediction tasks. Here, the two new layers are randomly initialized. The learning rate of the new layers is set higher than other layers because the parameters of these layers are randomly initialized.

5. Experiments

5.1. Experimental setting

Datasets: For training the fast R-CNN clothing detector, 13,500 woman clothing images and 13,500 man clothing images are collected and labeled with the upper clothing bounding box localization. The bounding box localization contains four-dimension coordinates with the left top point's position and the right bottom point's position. Both woman clothing images and man clothing images are divided into two subsets, 10,200 images for training and 3300 images for testing.

For training triplet network with videos, we obtain about 827,000 triplets from woman clothing videos and 856,000 triplets from man clothing videos. The number of triplets is exactly the same in the random negative training stage and hard negative training stage.

For clothing attribute prediction task, we have 10,200 images for training and 3300 images for testing. In total, 10 attributes are predicted, including color, style, collar, styleofcolor, styleofsleeve, lengthofsleeve, zip, belt, button and lengthofwhole. Each attribute has different values. As man and woman clothing have different values in some attributes, we define the attribute kinds according to their needs. The detailed woman and man clothes attribute kinds information is shown in Table 2. The numbers of images for some attributes are shown in Fig. 7.

Table 2
Clothing attribute values in woman and man clothing.

Attribute		Attribute values
Color	Woman and man clothing	Red, orange, yellow, green, blue, purple Black, white, gray, brown, multi-color
Collar		V-shape, round, pile collar, turndown collar/POLO Stand collar, irregular
Styleofcolor		Pure color, round dot, cell, irregular
Lengthofsleeve		Sleeveless, short, long
Zip		Exist, without
Belt		Exist, without
Button		Exist, without
Lengthofwhole		Median, short, long
Style	Woman clothing	Skinny, straight, loose, irregular
	Man clothing	Shirt, sweater, T-shirt, outwear, suit, tank, top, other
Styleofsleeve	Woman clothing	Normal sleeve, puff sleeve, shirt sleeve, pile sleeve, irregular
	Man clothing	Set-in sleeve, shirt sleeve, tight sleeve, irregular

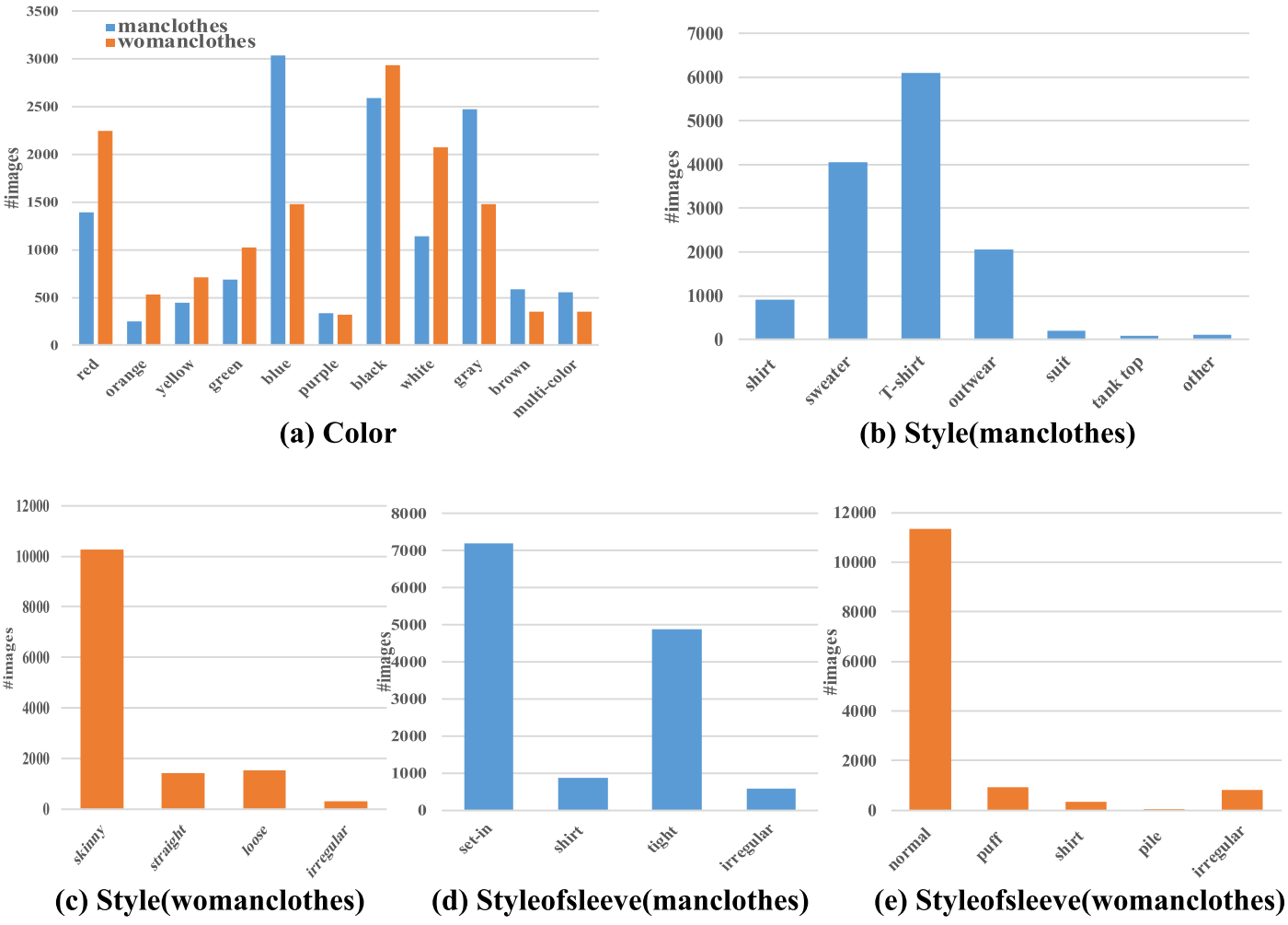


Fig. 7. Image number of some attributes in woman and man clothing image dataset. (a) Color attribute. (b) Style attribute of man clothing. (c) Style attribute of woman clothing. (d) Styleofsleeve attribute of man clothing. (e) Styleofsleeve attribute of woman clothing.

Baselines: We compare with some recent methods used in clothing attribute prediction task. (1) baseline_AlexNet [8]. The main network used in [8] is the standard AlexNet. (2) baseline_VGG16 [30]. The network adopted in [30] is named FashionNet, which has the same convolutional parts with VGG16 network. The FashionNet consists of three stages, i.e., predicting clothing landmarks, extracting local features via estimated clothing landmarks and fusing local and global features for category and clothing attribute prediction. It needs extra clothing landmarks and category annotations to train a powerful model. Here the annotations of landmarks and category are missed in our woman and man clothing datasets, so we ignore these parts and choose the core network VGG16 as comparisons. Both baseline_AlexNet and baseline_VGG16 methods are trained from scratch with woman clothing or man clothing images, respectively. Furthermore, we also add some pre-trained models to certificate the effectiveness of our unsupervised video model. (3) CFN_rec [34]. It employs the trained context-free network (CFN) model as initialization. The CFN model is trained through unsupervised learning which takes solving Jigsaw puzzles as a pretext task. (4) ImageNet_AlexNet and Ima-

geNet_VGG16. We also add experiments with supervised initialization, pre-trained on ImageNet dataset as comparisons. The AlexNet and VGG16 frameworks with ImageNet pre-trained model are denoted as ImageNet_AlexNet and ImageNet_VGG16, respectively.

First, we detect clothing in the labeled images using the trained clothing detector via our method (trained by fast R-CNN with 10,200 labeled bounding boxes). Then we train the deep network using the detected clothing regions. The network structure of the baseline is similar with that of AlexNet and VGG16 framework. We replace the last second fully connected layer with 1024 neurons and the output layer (holding 1000 neurons) with the number of attributes for each attribute prediction task. For all baselines, we use 10,200 labeled clothing attribute images for training and 3300 for testing.

For baseline_AlexNet method, the base learning rate is 0.01. We reduce the learning rate every 30 epochs and train for 90 epochs. All input images are resized to $227 \times 227 \times 3$.

For baseline_VGG16 method, the base learning rate is also 0.01, reduces the learning rate every 20 epochs and train for 90 epochs. All input images are resized to $224 \times 224 \times 3$. As a bad initialization cannot learn good representation due to the instability of gradient in deeper network, we follow [31] as using two stages to train the VGG16 baseline. In the first stage we train a shallow network named VGG11 (the network configuration comes from configuration A in [31]). When training VGG16, we initialize weight parameters of the first four convolutional layers and the first fully connected layer from the trained VGG11. We found that this strategy is not fit for man clothing attributes images. If we take this strategy into the man clothing attribute images, it cannot converge. So we initialize the all convolutional layers and the first fully connected layer from VGG11 when training VGG16 with man clothing attribute images.

Evaluation metrics: We use mean average precision (mAP) to measure the performance of the clothing attribute prediction. Average precision (AP) is defined as follows:

$$\mathrm{AP} = \sum_{i=1}^{N} w_i \times \frac{T_i}{P_i}, \tag{6}$$

where N denotes the number of attribute values, w_i is the weight ratio of the ith attribute kind in the testing set. T_i indicates the correct number of ith attribute kind prediction. P_i is the number of images which are predicted the ith attribute kind. Then mAP is defined as:

$$\mathrm{mAP} = \sum_{i=1}^{K} \frac{\mathrm{AP}_i}{K}, \tag{7}$$

where K is the number of attributes, AP_i denotes the ith attribute average precision.

5.2. Triplet network training details

In the pre-training on video context, the triplet AlexNet framework is shown in Fig. 6. 3 input frames go through 3 paths sharing same CNN network parameters. These inputs perform forward propagation and compute the triplet ranking loss based on the 1024 output feature space. Given a triple pair X_i, X_i^p, X_i^n, we will get three output feature $F(X_i)$, $F(X_i^p)$, $F(X_i^n)$ and compute the loss according to Eq. (3). The triplet VGG16 framework is the same to the triplet AlexNet, just own more convolutional layers.

Through the pre-processing clothing detection stage, we obtain about 122,000 woman clothing regions and 125,000 man clothing regions. For triplet AlexNet training, we start base learning rate ϵ from 0.0001 and set batch size B as 100. In the triplet ConvNet training, we first train 10 epochs with the fixed learning rate with the randomly selected triplet samples. Then we use the same learning rate to apply the hard negative mining triplets, reduce the learning rate every 20 epochs and train for 90 epochs. We note that the initialized bias should be set as 0.1.

For triplet VGG16 network training, we set the initial base learning rate $\epsilon = 0.01$, the batch size $B = 80$. In the triplet VGG16 ConvNet training, we adopt the same optimization strategy as triplet AlexNet training, the only difference is train 10 epochs in the hard negative mining stage. The convolutional weights are initialized from a normal distribution with the zero mean and 0.01 variance. The biases are initialized with zero. While training triplet VGG16 network, it also may have instability of gradient. We adopt the same initialization strategy training baseline_VGG16 model with woman clothing attribute images. All experiments are implemented with Caffe [24].

5.3. Clothing attribute prediction task

At the fine-tuning stage, we design a multiple attribute prediction network. Here we introduce an example of fine-tuning AlexNet framework. The basic AlexNet network configuration refers to Fig. 6. The parameters of convolutional layers and first 4096 fully connected layer are set as the pre-trained triplet CNN network. Then we add a new 1024 fully connected layer and a new fully connected layer whose neuron number is the number of attribute values. The five convolutional layers' parameters are initialized by the unsupervised video triplet AlexNet parameters. The new fully connected layers' parameters are randomly initialized.

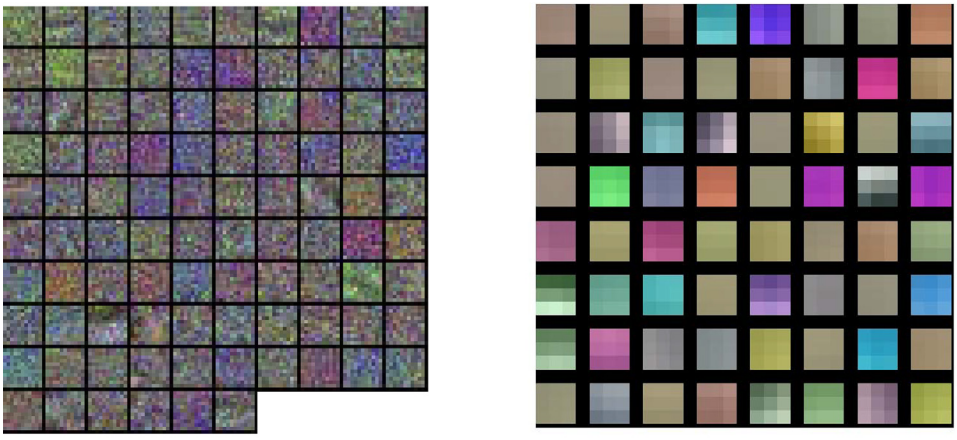
The VGG16 framework used in clothing attribute prediction task is similar with the AlexNet framework, all convolutional layers and the first fully connected layer are transferred from the unsupervised video triplet VGG16 network parameters. Then two new fully connected layers are followed, the neuron number is equivalent to the AlexNet framework in Fig. 6. The new fully connected layers' weights are initialized from a normal distribution with zero mean and 0.01 variance. The biases are initialized with zero.

We adopt a similar optimization strategy with the baseline method, the only difference is that set the base learning rate as $\epsilon = 0.001$. We adopt $\epsilon = 0.01$ in the baseline training in order to obtain a good performance.

We show the learned first convolutional layer filters in Figs. 8 and 9 based on AlexNet and VGG16 framework with woman and man clothing videos. We can observe that the unsupervised triplet CNN network can learn more colorful filters in the first convolutional layer. The top receptive fields of pool5 layer in our pre-trained AlexNet triplet ConvNet model is shown in Fig. 10. We can observe from Fig. 10 that the top receptive field often localize at the clothing object position in the video frames.

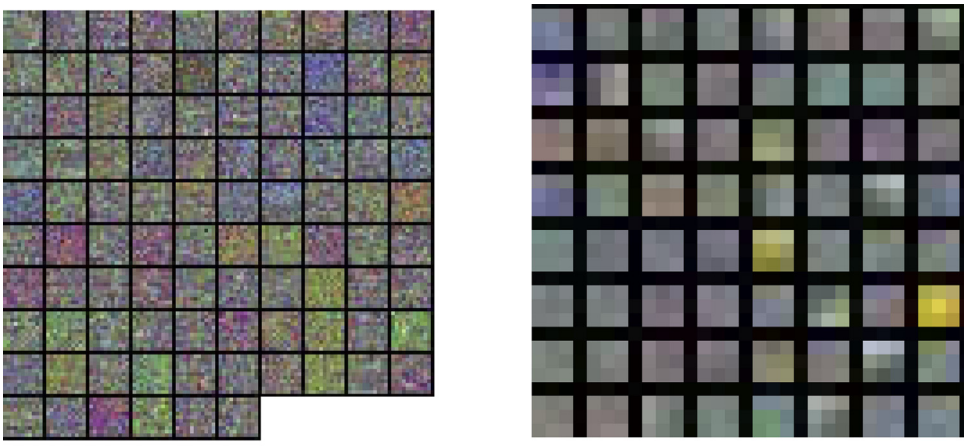
The mAP quantitative results are shown in Tables 3 and 4. The baseline_AlexNet method with no pre-trained model is indicated as baseline_AlexNet. Our AlexNet method is indicated as Our_AlexNet. The baseline VGG16 framework is indicated as baseline_VGG16 and Our_VGG16, respectively. We analyze the results in woman clothing attribute image dataset and man clothing attribute image dataset in the following.

Woman clothing image dataset: Table 3 shows the mAP results in woman clothing attribute image dataset. As to the baseline, we train the baseline_AlexNet and baseline_VGG16 network with 10,200 labeled clothing attribute images and separately obtain 71.09% mAP and 75.83% mAP. Fine-tuning on our unsupervised video triplet AlexNet, we get 4.91% and 1.5% higher mAP than the baseline_AlexNet and CFN_rec. Fine-tuning on our unsupervised video triplet VGG16 network, we obtain a 1.97% higher mAP than the baseline. Comparing to the same approach, we conclude that VGG16's performance is better than AlexNet (75.83% vs. 71.09% and 77.8% vs. 76%). Through Fig. 8, the unsupervised triplet network can learn more colorful information in first convolutional layer.



(a) Unsupervised AlexNet triplet network (b) Unsupervised VGG16 triplet network

Fig. 8. Visualization of the first convolutional layer filters learned with unlabeled woman clothing videos. (a) The learned conv1 layer filters of the unsupervised AlexNet triplet network. (b) The learned conv1_1 layer filters of the unsupervised VGG16 based triplet network. We can find that it learns more colorful information in this layer.



(a) Unsupervised AlexNet triplet network (b) Unsupervised VGG16 triplet network

Fig. 9. Visualization of the first convolutional layer filters learned with unlabeled man clothing videos. (a) The learned conv1 layer filters of the unsupervised AlexNet triplet network. (b) The learned conv1_1 layer filters of the unsupervised VGG16 based triplet network. We can find that this layer learns some colorful information.

Table 3
The mean average precision (mAP) results of woman clothing attribute prediction.

Method	Color	Style	Collar	Styleofcolor	Styleofsleeve	Lengthofsleeve	Zip	Belt	Button	Lengthofwhole	mAP
baseline_AlexNet [8]	38.44	70.41	37.13	75.71	78.81	87.98	89.96	84.59	68.71	79.19	71.09
CFN_rec[34]	55.7	69.54	43.7	80.01	76.07	89.7	91.95	86.67	73.84	77.84	74.5
Our_AlexNet	66.84	71.27	41.33	80.25	77.15	91.04	91.51	86.05	74.27	80.22	76
baseline_VGG16 [30]	44.75	72.86	47.93	81.67	78.43	92.13	92.39	87.4	80.47	80.25	75.83
Our_VGG16	64.79	70.96	49.87	83.87	78.66	91.86	91.97	86.37	79.81	79.84	77.8
ImageNet_AlexNet	**71.36**	71.27	51.46	83.6	78.38	92.65	91.75	88.11	78.77	81.44	78.88
ImageNet_VGG16	71.01	**77.31**	**64.68**	**87.97**	**79.47**	**95.73**	**95.39**	**90.33**	**89.91**	**84.07**	**83.59**

Table 4
The mean average precision (mAP) results of man clothing attribute prediction.

Method	Color	Style	Collar	Styleofcolor	Styleofsleeve	Lengthofsleeve	Zip	Belt	Button	Lengthofwhole	mAP
baseline_AlexNet [8]	52.04	77.73	69.48	74.89	79.61	94.92	87.14	99.71	79.01	96.49	81.1
CFN_rec[34]	61.59	80.29	71.55	79.77	80.74	95.23	88.67	99.58	82.51	97.32	83.73
Our_AlexNet	75.47	83.68	74.62	82.55	83.31	96.27	90.93	99.71	84.21	97.12	86.79
baseline_VGG16 [30]	52.18	83.38	76.66	79.88	83.52	95.53	93.1	99.71	87.03	97.05	84.8
Our_VGG16	67.8	84.52	76.33	81.86	84.51	96.33	93.2	99.78	87.92	97.41	87
ImageNet_AlexNet	74.37	86.59	81.91	83.9	86.22	96.61	92.97	99.82	88.11	97.35	88.79
ImageNet_VGG16	**76.21**	**90.91**	**88.8**	**87.79**	**90.11**	**97.72**	**96.37**	**99.82**	**94.05**	**97.87**	**91.97**

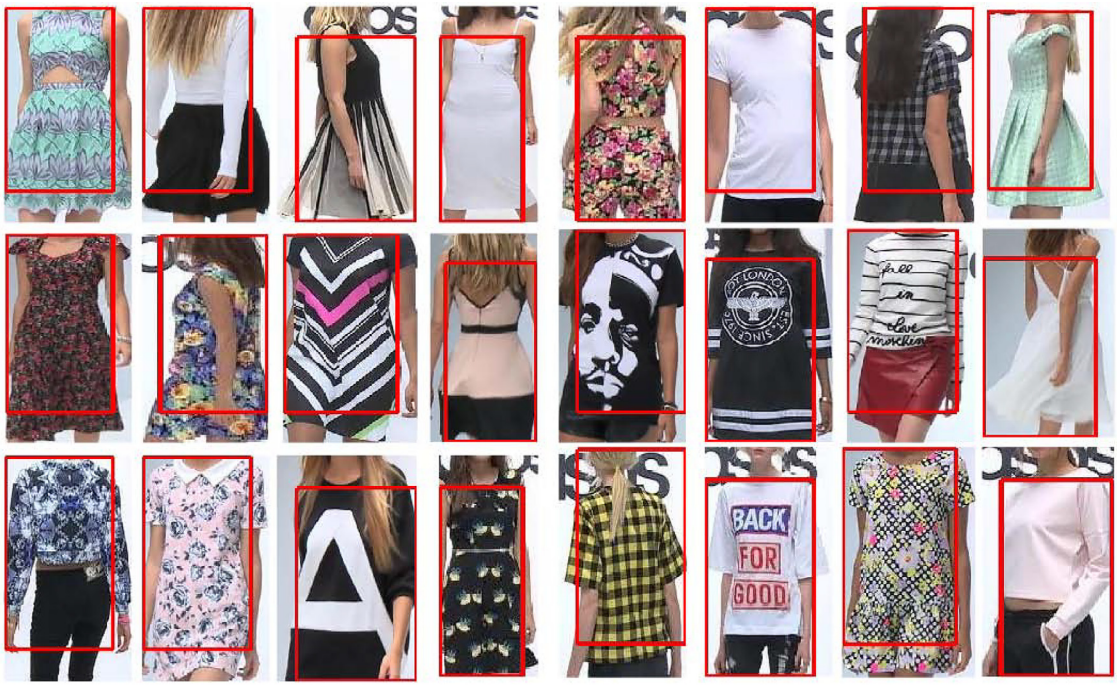


Fig. 10. Top receptive fields visualization of pool5 layer in our pre-trained triplet AlexNet network with clothing videos. The receptive fields are indicated by the red bounding boxes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

AlexNet		VGG	
baseline_AlexNet	our method	baseline_VGG16	our method
green	black	black	black
straight	skinny	skinny	skinny
POLO	round	POLO	round
irregular	pure color	pure color	pure color
normal sleeve	normal sleeve	normal sleeve	normal sleeve
long	long	long	long
without zip	without zip	without zip	without zip
without belt	without belt	without belt	without belt
exist button	without button	exist button	without button
median	median	median	median

AlexNet		VGG	
baseline_AlexNet	our method	baseline_VGG16	our method
blue	black	black	black
skinny	loose	skinny	loose
POLO	round	round	round
pure color	irregular	irregular	irregular
normal sleeve	normal sleeve	normal sleeve	normal sleeve
short	short	long	short
without zip	without zip	without zip	without zip
without belt	without belt	without belt	without belt
without button	without button	without button	without button
median	median	median	median

AlexNet		VGG	
baseline_AlexNet	our method	baseline_VGG16	our method
blue	brown	white	black
skinny	skinny	skinny	skinny
V-shape	V-shape	irregular	V-shape
pure color	irregular	irregular	irregular
normal sleeve	normal sleeve	normal sleeve	normal sleeve
long	short	short	short
without zip	without zip	without zip	without zip
without belt	without belt	without belt	without belt
exist button	without button	without button	without button
median	long	long	long

AlexNet		VGG	
baseline_AlexNet	our method	baseline_VGG16	our method
green	blue	blue	blue
skinny	skinny	loose	skinny
irregular	POLO	round	POLO
pure color	pure color	pure color	pure color
normal sleeve	normal sleeve	normal sleeve	normal sleeve
long	long	long	long
without zip	without zip	without zip	without zip
without belt	without belt	without belt	without belt
without button	exist button	exist button	exist button
median	median	median	median

Fig. 11. Woman clothing attribute prediction results. The bounding boxes with green color are the detected clothing. The attribute with red indicates that it is judged wrong. Note that the second image in the first row, the lengthofwhole attribute results are all wrong. In fact, all methods judge right. The reason of this error is because the groundtruth is wrong labeled with 'long' value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

So it helps the color and styleofcolor attributes to improve more performance. Other clothing attributes also obtain higher performance with our proposed method. It illustrates that using video context information learns more discriminative feature representation and can help improve the performance of the woman clothing attribute prediction task. We can find that the unsupervised methods can not surpass supervised ImageNet models, but the unsupervised methods can obtain closer scores to supervised methods. Though there is a gap between them, the unsupervised method is useful to alleviate human from labeling labor.

The qualitative clothing attribute prediction results are shown in Fig. 11. Compared with the baseline method, Fig. 11 shows that our approach is better than the baseline network. The color attribute can be judged by our method more precisely than the supervised method. These results also have been certified in mAP results.

Man clothing image dataset: Table 4 shows the mAP results in man clothing attribute image dataset. As to the baseline, we train baseline_AlexNet and baseline_VGG16 network with 10,200 labeled clothing attribute images and separately obtain 86.79% mAP and 87% mAP. Fine-tuning on our unsupervised video triplet AlexNet, we obtain 5.69% and 3.06% higher mAP than the baseline_AlexNet and unsupervised CFN_rec pre-trained model. Fine-tuning on our unsupervised video triplet VGG16 network, we also

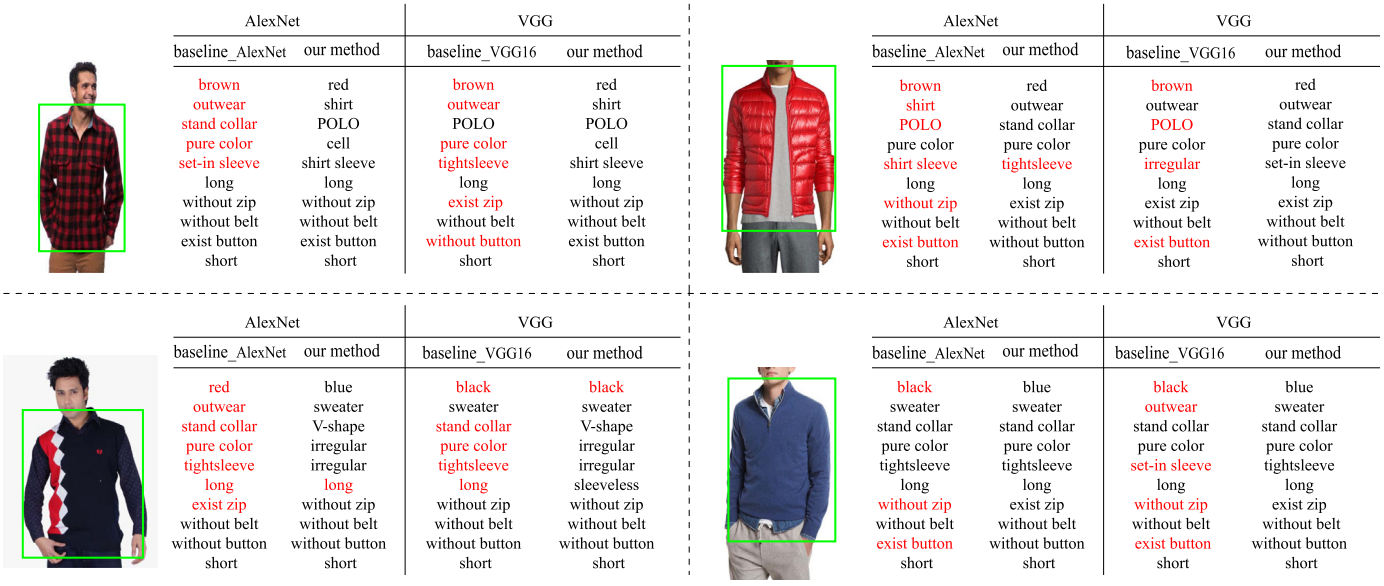


Fig. 12. Man clothing attribute prediction results. The bounding boxes with green color are the detected clothing. The attribute with red indicates it is judged wrong. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

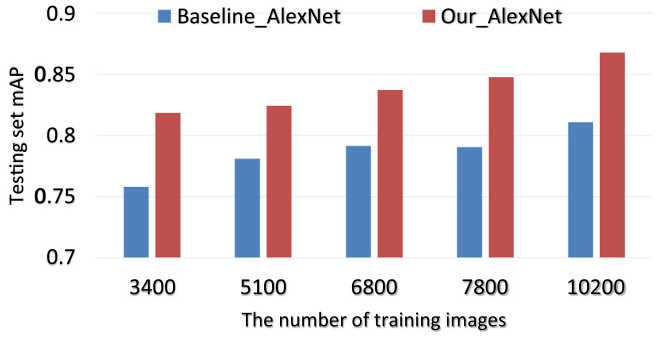


Fig. 13. The mAP performance comparison of different number of labeled training man clothing images with AlexNet model.

obtain a 2.2% improvement than the baseline. Similarly, only comparing the same approach we conclude that VGG16's performance is better than AlexNet (84.8% vs. 81.1% and 87% vs. 86.79%). We observe from Fig. 9 that the color attribute also helps to improve performance. The styleofcolor attribute obtains higher performance especially in AlexNet based method. The reason is that large number of unlabeled videos provide more color information and the trained model can also learn these while training. And many other clothing attributes can obtain better performance comparing with baseline approach. These results also illustrate that using video context information can help to improve the performance of the man clothing attribute prediction task. Through the experimental results in man clothing dataset, a similar conclusion can be concluded that the unsupervised methods can not surpass supervised ImageNet model based methods, but it can give a solution to reduce annotation labor.

The qualitative clothing attribute prediction results are shown in Fig. 12. Compared with the baseline method, Fig. 12 results show that our approach can obtain better attribute prediction results than baseline method.

Besides, we also certify the performance with different scale of labeled training images. We use man clothing images and AlexNet network for this experiment, the results are shown in Fig. 13. Through Fig. 13, we can observe that the more training images used, the higher mAP are obtained by the supervised method and our semi-supervised method in the testing set. The semi-supervised approach can obtain good results than the supervised approach. The mAP result of our method using 5100 training images can be close to the supervised method using 10,200 training images. This illustrates that unlabeled videos can help improve the performance of clothing attribute prediction.

6. Conclusion

In this paper, we explore a deep semi-supervised method which takes advantages of unlabeled videos to improve the performance of clothing attribute prediction task. First, triplet ranking loss is utilized to train an unsupervised network with video data via constraining the distance of similar and dissimilar frame pair. Then this triplet video network is transferred to learn a clothing attribute prediction model with clothing images whose attribute values are annotated. To that end, two new clothing video datasets and two new clothing image datasets are collected. We demonstrate the effectiveness of our proposed method in different CNN network (AlexNet and VGG16). We have shown the effectiveness of the video context and we believe that the unsupervised video learning algorithm is a good solution to alleviate human from labeling. In the future, we will explore some common prior information of clothing attribute in similar images, for example, the attribute value distribution, and utilize them to train more powerful model with unsupervised learning.

Acknowledgments

This work was supported by National Key Research and Development Plan (No. 2016YFB0800603), National Natural Science Foundation of China (No. 61332012, U1636214), Key Program of the Chinese Academy of Sciences (No. QYZDB-SSW-JSC003).

References

[1] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, R. Urtasun, Neuroaesthetics in fashion: modeling the perception of fashionability, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 869–877.

[2] S. Vittayakorn, K. Yamaguchi, A.C. Berg, T.L. Berg, Runway to realway: visual analysis of fashion, in: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV), 2015, pp. 951–958.
[3] L. Bourdev, S. Maji, J. Malik, Describing people: a poselet-based approach to attribute classification, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2011, pp. 1543–1550.
[4] H. Chen, A. Gallagher, B. Girod, Describing clothing by semantic attributes, in: Proceedings of European Conference on Computer Vision (ECCV), 2012, pp. 609–623.
[5] J. Huang, R.S. Feris, Q. Chen, S. Yan, Cross-domain image retrieval with a dual attribute-aware ranking network, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1062–1070.
[6] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, S. Yan, Street-to-shop: cross-scenario clothing retrieval via parts alignment and auxiliary set, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 3330–3337.
[7] T. Xiao, T. Xia, Y. Yang, C. Huang, X. Wang, Learning from massive noisy labeled data for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2691–2699.
[8] Q. Chen, J. Huang, R. Feris, L.M. Brown, J. Dong, S. Yan, Deep domain adaptation for describing people based on fine-grained clothing attributes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5315–5324.
[9] J. Shen, G. Liu, J. Chen, Y. Fang, J. Xie, Y. Yu, S. Yan, Unified structured learning for simultaneous human pose estimation and garment attribute classification, IEEE Trans. Image Process. 23 (11) (2014) 4786–4798.
[10] V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, N. Sundaresan, Large scale visual recommendations from street fashion images, in: Proceedings of ACM SIGKDD, 2014, pp. 1925–1934.
[11] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, S. Yan, Hi, magic closet, tell me what to wear!, in: Proceedings of ACM-MM, 2012, pp. 619–628.
[12] S.-Z. Chen, C.-C. Guo, J.-H. Lai, Deep ranking for person re-identification via joint representation learning, IEEE Trans. Image Process. 25 (5) (2016) 2353–2367.
[13] A. Li, L. Liu, K. Wang, S. Liu, S. Yan, Clothing attributes assisted person re-identification, IEEE Trans. Circuits Syst. Video Technol. 25 (5) (2015) 869–878.
[14] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, S. Yan, Deep human parsing with active template regression, IEEE Trans. Pattern Anal. Mach. Intell. 37 (12) (2015) 2402–2414.
[15] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, S. Yan, Fashion parsing with weak color-category labels, IEEE Trans. Multimedia 16 (1) (2014) 253–265.
[16] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, S. Yan, Matching-CNN meets KNN: quasi-parametric human parsing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1419–1427.
[17] Z. Hu, H. Yan, X. Lin, Clothing segmentation using foreground and background estimation based on the constrained Delaunay triangulation, Pattern Recognit. 41 (5) (2008) 1581–1592.
[18] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, R. Urtasun, A high performance CRF model for clothes parsing, in: Proceedings of Asian Conference on Computer Vision (ACCV), 2014, pp. 64–81.
[19] K. Yamaguchi, M. Kiapour, T. Berg, Paper doll parsing: retrieving similar styles to parse clothing items, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 3519–3526.
[20] S. Liu, X. Liang, L. Liu, K. Lu, L. Lin, X. Cao, S. Yan, Fashion parsing with video context, IEEE Trans. Multimedia 17 (8) (2015) 1347–1358.
[21] Z. Song, M. Wang, X.-S. Hua, S. Yan, Predicting occupation via human clothing and contexts, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2011, pp. 1084–1091.
[22] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 886–893.
[23] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110.
[24] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of ACM-MM, 2014, pp. 675–678.
[25] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of NIPS, 2012, pp. 1097–1105.
[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.
[27] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026–1034.
[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[29] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440–1448.
[30] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: powering robust clothes recognition and retrieval with rich annotations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1096–1104.
[31] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of ICLR, 2015.
[32] Z. Liu, S. Yan, P. Luo, X. Wang, X. Tang, Fashion landmark detection in the wild, in: European Conference on Computer Vision, Springer, 2016, pp. 229–245.
[33] C. Doersch, A. Gupta, A.A. Efros, Unsupervised visual representation learning by context prediction, in: International Conference on Computer Vision (ICCV), 2015.
[34] M. Noroozi, P. Favaro, Unsupervised learning of visual representations by solving jigsaw puzzles, in: ECCV, 2016.
[35] X. Liang, S. Liu, Y. Wei, L. Liu, L. Lin, S. Yan, Towards computational baby learning: a weakly-supervised approach for object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 999–1007.
[36] Y. Yang, Y. Li, Y. Aloimonos, Robot learning manipulation action plans by watching unconstrained videos from the world wide web, in: Proceedings of AAAI, 2015, pp. 3686–3693.
[37] I. Misra, A. Shrivastava, M. Hebert, Watch and learn: semi-supervised learning of object detectors from videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3593–3602.
[38] F. Zhao, Y. Huang, L. Wang, T. Tan, Deep semantic ranking based hashing for multi-label image retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1556–1564.
[39] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, Y. Wu, Learning fine-grained image similarity with deep ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1386–1393.
[40] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.
[41] H. Lai, Y. Pan, Y. Liu, S. Yan, Simultaneous feature learning and hash coding with deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3270–3278.
[42] X. Wang, A. Gupta, Unsupervised learning of visual representations using videos, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2794–2802.
[43] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580–587.
[44] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.
[45] J.R. Uijlings, K.E. van de Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, Int. J. Comput. Vision 104 (2) (2013) 154–171.

Sanyi Zhang received the B.E. and M.E. degrees in computer science from Taiyuan University of Technology, China. He is currently a Ph.D. candidate in the Tianjin University, China. His current research interests include computer vision and clothing attribute learning.

Si Liu is an associate professor in Institute of Information Engineering, Chinese Academy of Sciences. She used to be a research fellow at Learning and Vision Group of National University of Singapore. She received her Ph.D. degree from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences in 2012. Her research interests include computer vision and multimedia.

Xiaochun Cao received the B.E. and M.E. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA. He has been a professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, since 2012. He spent about 3 years with ObjectVideo Inc., as a research scientist. From 2008 to 2012, he was a professor with Tianjin University, Tianjin, China. He has authored and co-authored over 120 journal and conference papers. He is a fellow of the IET. He is on the Editorial Board of the IEEE Transactions of Image Processing. His dissertation was nominated for the University of Central Florida's university-level Outstanding Dissertation Award. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition.

Zhanjie Song was born in Hebei Province, China, in 1965. He received the Ph.D. degree from Nankai University in 2006. He was a postdoctoral fellow in signal and information processing, with School of Electronic and Information Engineering, Tianjin University. He is currently a professor with the School of Mathematics, a fellow of Visual Pattern Analysis Research Lab, and a vice-director with the Institute of TV and Image Information, all in Tianjin University. His current research interests are in sampling, approximation and reconstruction of random signals and random fields.

Jie Zhou received the B.S. and M.S. degrees both from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995. From then to 1997, he served as a postdoctoral fellow in the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a full professor in the Department of Automation, Tsinghua University. From 2015 to now, he is the Chair of Department of Automation, Tsinghua University. His research interests include pattern recognition, computer vision, and image processing. In recent years, he has authored more than 200 papers in peer-reviewed journals and conferences. Among them, more than 50 papers have been published in top journals and conferences (IEEE T-PAMI, T-IP, CVPR and ICCV). He received the National Outstanding Youth Foundation of China Award in 2012. He is a fellow of IAPR and associate editor of IEEE T-PAMI.