Neural Computing and Applications (2023) 35:13021–13035
https://doi.org/10.1007/s00521-020-05528-9

S.I.: DEEP SOCIAL COMPUTING

Transferring fashion to surveillance with weak labels

Qi Zheng[1,2,3] · Zheng He[1,2] · Chao Liang[1,2,3] · Jun Chen[1,2,3] · Chia-Wen Lin[4] · Dapeng Tao[5]

Received: 23 March 2020 / Accepted: 10 November 2020 / Published online: 23 November 2020

Abstract
In this paper, we address the problem of automatic clothing parsing in surveillance images using the information from user-generated tags, such as "jeans" and "T-shirt." Although clothing parsing has achieved great success in the fashion domain, it is quite challenging to parse target under practical surveillance conditions due to the presence of complex environmental interference, such as that from low resolution, viewpoint variations and lighting changes. Our method is developed to capture target information from the fashion domain and apply this information to a surveillance domain by weakly supervised transfer learning. Most target tags convey strong location information (e.g., "T-shirt" is always shown in the upper region), which can be used as weak labels for our transfer method. Both quantitative and qualitative experiments conducted on practical surveillance datasets demonstrate the effectiveness of the proposed surveillance data enhancing method.

Keywords Clothing parsing · Transfer learning · Weakly supervised learning

1 Introduction

Clothing parsing aims to label specific items on the level of pixels. Though clothing parsing is a relatively new research area in sensing data, it has attracted increasing levels of attention across numerous fields on humans, ranging from person tracking [1–3], body shape estimation [4] and content-based image retrieval [5] to fashion images parsing [6]. Target parsing in fashion has received the most attention [7–11] because fashion images are often depicted in stable illumination with a uniform viewpoint, a high resolution and sharp edges.

Compared to its fashion applications, target parsing applied under surveillance camera has rarely been studied. Some related work only focuses on extracting an entire target region [12], providing labels on the image-level [13, 14], or recognizing items through a rough region [15, 16], diverging from the conditions of genuine pixel-level clothing parsing. Associated difficulties are twofold. The first difficulty is the related to complex environmental interferences, such as lighting changes, viewpoint variations, low resolution and motion blur. The second difficulty is related to the endless and painstaking pixel-level labeling work required to manage innumerable images. Despite these difficulties, clothing parsing in a surveillance environment is of great importance because it can be used as an

✉ Zheng He
hezheng@whu.edu.cn

Qi Zheng
zhengq@whu.edu.cn

Chao Liang
cliang@whu.edu.cn

Jun Chen
chenj.whu@gmail.com

Chia-Wen Lin
cwlin@ee.nthu.edu.tw

Dapeng Tao
dapeng.tao@gmail.com

1 National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China

2 Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China

3 Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China

4 Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan, R.O.C.

5 School of Information Science and Engineering, Yunnan University, Kunming 650504, Yunnan, China

implicit cue signaling a person identity, location and even occupation, which are all key intelligence clues used in a security system.

Given the successes of clothing parsing achieved in the fashion domain, it is natural to conceptualize clothing parsing in surveillance environments with the aid of numerous fashion images. However, directly using fashion images as training data to parse surveillance data remains challenging due to intrinsic domain differences between surveillance and fashion environments. Some instances of failure, such as labels covering on wrong parts or missing human parts, from MIT dataset are illustrated in Fig. 1.

Some researchers have attempted to apply transfer learning to cross-domain attribute recognition. The transfer learning aims to store knowledge (as a "trained model" in this work) gained while solving one problem and applying it, which needs to be updated, to a different but relevant problem [19]. Along this line of reasoning, some early works [16, 20, 21] used CNN-based domain adaptation networks to jointly model data from two domains or employed graph models with latent variables to update model. However, such feature adaptation methods mainly focus on attribute transfers. In terms of clothing parsing, they are inefficient because semantic segmentation can suffer due to the complexity of the high-dimensional features of visual cues, including appearance, shape and context.

We note that the segmentation region structures of parsing results share numerous similarities across the two different domains. As Fig. 2 shows, even when there is a large appearance gap between the two domains, the corresponding gap is smaller in parsing results containing rich information on the shape, location and area. Based on this observation, we explored the relationship between a parsing model and a data instance structure, instead of focusing on similar feature representations. The classic instance-based transfer learning [22] assumed that certain parts of the data in the source domain can be used to re-weight

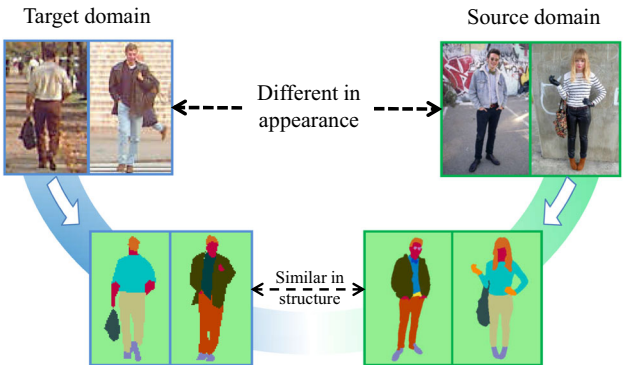


Fig. 2 While the appearance of images always varies, the corresponding structures of the parsing results are similar because clothing is always found within a corresponding spatial region, and all segments are always integrated into a similar human shape

models, originally trained in the source domain, in the target domain. We broaden the assumption that certain portions of the data of the target domain are also of central importance to transferring when they are compatible with a trained model from a source domain. In this work, we refer to certain portions of the data that have parsing results with high confidence values calculated from structure evaluations as *fine data*, and we refer to the remaining data as *unfit data*. With the help of *fine data*, an iterative instance-based transfer scheme is introduced to improve the parsing of surveillance images, as illustrated in Fig. 3. We first consider a fashion domain as the source domain and a surveillance domain as the target domain. Then, an initial parsing model is trained on the fashion domain. Next, the model is used to parse surveillance images with stripe constraint, and the *fine data* are utilized to update a previously learned parsing model. Our quantitative and qualitative experiments demonstrate that the proposed method can be applied to a new surveillance domain.

This paper is an extension of our previous conference paper [23]. Compared to our previous work, we present an improved clothing parsing approach and a new data

Fig. 1 Some failed parsing results derived from an MIT surveillance dataset [17] with a clothing parsing model directly trained on the Fashionista dataset [18]. Parsing results for pedestrians with wrong parts, such as (**a**), represent more than 15% of the dataset. Parsing results for pictures with wrong labels, such as (**b**) and (**c**), account for roughly 28%

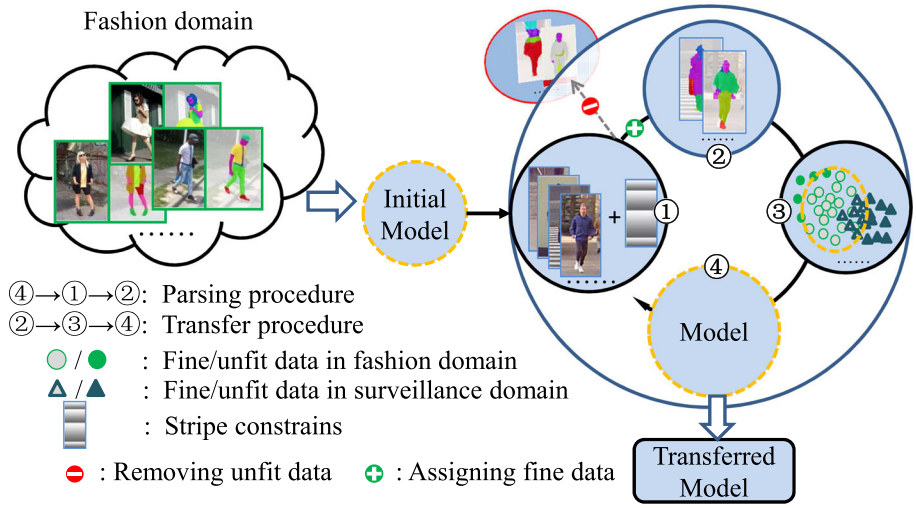


Fig. 3 Iterative instance-based transfer scheme. Given an initial model trained on a fashion dataset, surveillance images (in circle 1) are parsed. Then, the fine surveillance parsing results (in circle 2) using fine fashion data are selected by our selection strategy (in circle 3) to update the parsing model (in circle 4), and then, the model is used to parse the surveillance images. Finally, the iteration is stopped with data confidence converged (related description will be presented in Sect. 4)

selection module, and we present expanded experiments to evaluate the performance of the model. The major contributions of this work are as follows:

- We address clothing parsing problems of surveillance environments without the use of pixel-level labels.
- We propose a novel transfer learning method and design an evaluation strategy to explore useful information for pedestrian parsing problems from both domains.
- Our experiments show the flexibility of the proposed method, which can be incorporated into any segmentation algorithm and which can derive impressive results from mainstream fashion and surveillance datasets.

The remainder of this paper is organized as follows: Sect. 2 presents related work, Sect. 3 introduces the clothing parsing approach, Sect. 4 describes the technical details of the transferring scheme, Sect. 5 represents our experimental results, and discussion, and Sect. 6 concludes the work.

2 Related work

2.1 Clothing parsing

Clothing parsing is an attractive computer vision tool developed in recent years that is important for enabling many applications and for developing useful representations. Different from some similar work like face parsing [24, 25] focusing on face components or street scene parsing [26] focusing on outdoor stuff, clothing parsing mainly considers the parsing work on clothing and body parts.

Early work is mainly based on low-level features [6–8, 18, 27, 28]. One of the first approaches developed by Chen et al. [27] modeled clothing as a grammar of sketch templates to match input images. Then, a representative work from Yamaguchi et al. [18] studied human pose estimations for sequentially attributing labels and for refining clothing parsing using a retrieval-based approach [7]. A similar approach developed by Simo-Serra et al. [6] further explored the shape and location priors for garments and achieved improved results. Later work [28] tried to introduce mid-level semantics to facilitate clothing parsing. Traditional hand-crafted pipelines often combine low-level features with graphical models, which need to be carefully designed.

The development of deep learning in recent years has highlighted new ways to address various problems [29–33] with high-level features and has achieved considerable success in the area of clothing parsing. For example, Yang et al. [34] extended the output of a fully convolutional neural network (FCN) to infer clothing contexts from superpixels, and Tangseng et al. [9] trained a FCN architecture with a side-branch network and CRF postprocessing to achieve impressive results. Meanwhile, He et al. [35] adopted a lightweight multiscale network to achieve the fastest levels of parsing performance. However, these approaches mainly focused on ways to parse clothing in fashion environments. Here, we address clothing parsing problems related to surveillance to explore ways to address complex environmental interference with insufficient training data. The superresolution technology [36–38] shows a possible way to enhance surveillance image but still suffer from insufficient labels. [39] proposed a robust semi-supervised learning method based on maximum correntropy criterion. The proposed method could effectively capture the negative influence of noisy labels from

complex and very high dimension image data which inspired us to use weak labels to deal with insufficient labels problem.

2.2 Domain adaptation

A key challenge facing our strategy concerns bridging the domain gap between the fashion and surveillance domains. Domain adaptation problems, which are often encountered in transfer learning, have been widely studied in the realm of computer vision [22, 40–46] and data mining [47, 48]. However, most transfer methods only address this problem in relation to different datasets of a single scene. Dong et al. [49] noted that spatial information could benefit the domain adaption performance in various remote sensing scenes. But those scenes are still quite different from ours. For fashion scenes, some nonparametric methods [7, 50, 51] have been proposed and successfully applied to existing datasets and newly annotated images. Recently, some deep learning architectures have also shown transfer characteristics, such as [52] has used the CNN to deal with clothing classification transferring problem directly. And [53] proposed a deep domain adaptation method by matching the discriminative embedding between source domain and target domain effectively with pseudo labels, which obviously improved the domain adaptation performance. Some other methods rely on retraining the last few layers of a network with samples from the target domain or on combining region matching with CNN matching from original and target images [54, 55]. In regard to surveillance scenes, some works [56, 57] have tried to identify shared features to build connections between two datasets. Other works [58–60] have adapted transfer learning processes by using external data or by exploring intra-latent information to improve the performance of target data.

Unlike the above works, which only focus on single scenes, some CNN-based methods [20, 61], have been developed to address domain adaptation problems encountered between fashion scenes and real-life images. However, target domain conditions (primarily photographs of humans taken with mobile phones from a close distance) remain superior to those of the surveillance domain. While similar studies of different scenes have used generative adversarial networks (GANs) [62, 63], such works mainly address street scene parsing problems of the synthetic and real domains. The first work to transfer semantic representations between the fashion and surveillance domains is [16] which only uses coarse segmentation as a latent variable for person re-identification and searching. Compared to the above works, we aim to apply transfer learning to generate delicate clothing parsing results from surveillance environments. Specifically, we use instance-based transfer learning to transfer clothing parsing from the fashion domain to the surveillance domain. Our approach follows the principles of self-paced learning [64, 65] by selecting samples iteratively. However, our sample selection criteria are based on the connections between two different domains. Moreover, while self-paced learning is applied to fully supervised settings with labeled samples, our approach does not apply labels to all target samples.

3 Clothing parsing approach

In this section, we describe the method of parsing clothing from practical surveillance environments and introduce our clothing parsing approach.

3.1 Overview of the clothing parsing approach

Clothing parsing can be viewed as a labeling problem in which each pixel of an image is assigned a semantic label that can be selected from the background, from images of skin or hair, or from a large set of clothing items, e.g., boots, tights and sweaters. However, images of some garment items, such as “ring” or “bracelet,” cannot be parsed in the surveillance domain due to the low quality of such images. These items are removed from our framework. Table 1 lists the garment items we use.

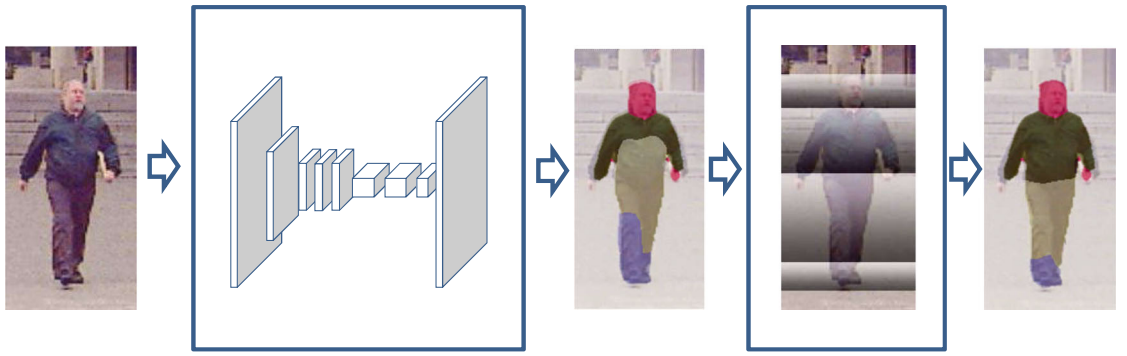
Our clothing parsing approach involves two steps as Fig. 4 shows. For the first step, segmentation algorithms are used to obtain initial clothing regions. To prove the effectiveness of our method, we adopt a classic parsing architecture, the fully convolutional network (FCN) [66] as our backbone. In the second step, stripes are leveraged to constrain the segmentation results from previous step. The positioning of specific item is often fixed within the same stripe, as pedestrians typically walk and stand upright in surveillance images, rendering it possible to refine the results.

3.2 Clothing parsing backbone

All parsing models are available for the clothing parsing backbone. Here, we use the classic FCN (Fig. 5) as our target parsing backbone. The FCN is an end-to-end training that has achieved remarkable results in many areas of computer vision. According to experiments reported in [9], the FCN model also performs well when applied to a fashion dataset. As an extension of the CNN, the FCN transforms the fully connected layers of the CNN into convolutional layers to allow the classification net to output a heatmap and applies a spatial loss to produce an efficient machine for end-to-end dense learning. However, spatial information is fuzzy after a couple convolutional layers. To address this problem, FCN defines a “skip” architecture to

Table 1 We use 44 garment items for different body parts taken from a fashion dataset as our experimental labels and omit labels that cannot be parsed

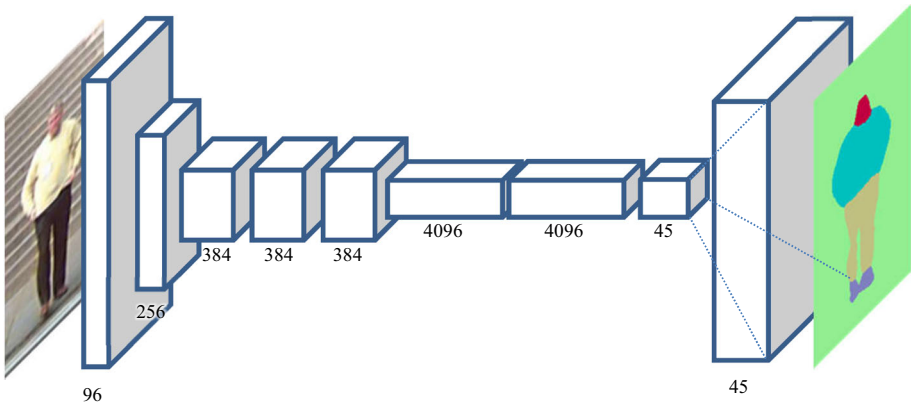
Top region	Hair, hat, sunglasses
Upper region	Blazer, blouse, bodysuit, bra, cape, cardigan, coat, gloves, jacket, jumper, scarf, shirt, sweater, sweatshirt, T-shirt, tie, top, vest
Lower region	Jeans, leggings, panties, pants, shorts, skirt, stockings, tights
Bottom region	Boots, clogs, flats, heels, loafers, sandals, shoes, sneakers, socks
Other	Bag, belt, dress, romper, skin, suit



(a) Input Image **(b)** Parsing Model **(c)** Initial Result **(d)** Stripe Constraint **(e)** Output Result

Fig. 4 Overview of our clothing parsing approach involving two steps (as illustrated in the boxes). The first step can be applied using any parsing model. The second step, the stripe constraint, is used to refine the initial results derived from surveillance images

Fig. 5 Overview of clothing parsing model. We first train an end-to-end fully convolutional networks with 45 classes. Then given a pedestrian image, we put it into the network to output a human item prediction per pixel



combine coarse, high-layer information with fine, low-layer information. Further information on this point is given in [66]. Here, we alter the last output layer by increasing the number of classes involved from 21 to 45 (including the background class) to adapt the model to our surveillance dataset.

3.3 Stripe constraint

The stripe constraint is the postprocessing step after the clothing parsing backbone is determined and is used to refine the parsing results. Each item label corresponds to one stripe, and these "stripes" divide each pedestrian in an image into a set of horizontal regions. From a collection of surveillance datasets, we found that a change in the positioning of the surveillance camera has little impact on the stripe proportions of pedestrians, as they stand or walk upright, and their clothing is always typical in a common surveillance scene. Therefore, we can apply a mask around each stripe to remove false predictions. A stripe mask only considers a vertical pixel with two hinge functions defined as follows:

$$f_n(y) = \begin{cases} \max\{0, 1 - (t_n - y)/h_n\}, & y < t_n \\ 1, & t_n \le y \le b_n \\ \max\{0, 1 - (y - b_n)/h_n\}, & y > b_n. \end{cases} \quad (1)$$

This is shown in Fig. 6b, where y is the pixel position in the vertical direction and t_n, b_n denote the top line and bottom line of the n-th stripe, respectively. And h_n denotes the stripe height, which is calculated as

$$h_n = b_n - t_n. \quad (2)$$

The hyperparameters t_n and b_n are obtained from the stripe proportion parameters Γ_{t_n} and Γ_{b_n} and from the target image height H by

$$\begin{aligned} t_n &= \Gamma_{t_n} \cdot H, \\ b_n &= \Gamma_{b_n} \cdot H, \end{aligned} \quad (3)$$

where Γ_{t_n} and Γ_{b_n} are obtained from the statistics for a corresponding pedestrian region of a fashion dataset. Note that we set the mask of some items (like skin) to 1 without considering t_n and b_n, as these items may appear anywhere in the image. The stripe mask reserves predictions for the stripe and weakens predictions oriented farther from the stripe. Then, results are refined through the elementwise multiplication of the stripe mask and the corresponding original probability map of the last step. Figure 6 illustrates the stripe constraint processing step.

4 Surveillance adaptation

In this section, we discuss the technical details of transferring a fashion-based model to a surveillance dataset with weak labels.

4.1 Overview of the surveillance adaptation approach

Images given in fashion datasets with pixel-level labels are usually applied under ideal imaging conditions. However, it is unreliable to directly use a model trained to the fashion domain to predict pixel-level labels in the surveillance domain as illustrated in Fig. 1. To bridge this gap, we design a novel instance-based transfer learning method for surveillance adaptation. Here, the surveillance (target) domain is quite different from the fashion (source) domain, while their parsing tasks are similar. Thus, our transfer method involves transductive transfer learning, which aims to improve the learning of the target predictive function [22]. We expand on the original assumptions of instance-based transfer learning to use certain parts (referred to as fine data, in this paper) of the data from both domains to learn the target domain, and we design a confidence evaluation to find the fine data. It is worth noting that the fine data in both domains build a special link between the fashion and surveillance domains. For the target domain, we consider data with high confidence as fine data. For the source domain, we explore data similar to the fine data in the target domain. Then the fine data which consist of source images with ground truth and target images with high quality labels, are set as a new training dataset to update the parsing model. For domain adaptation, we propose an iterative parsing method. We initialize the model based on the fashion dataset. Then, the fine data and the parsing model are iteratively updated with the inner-loop of the proposed method. The proposed procedure is outlined by Algorithm 1.

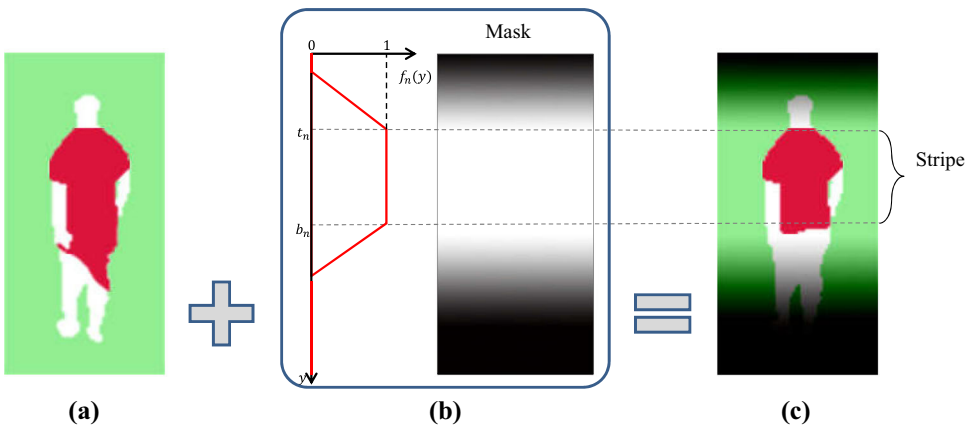


Fig. 6 Overview of the stripe constraint step. The stripe mask function, shown as a red line, constrains the parsing label, shown in the lighter area, and prevents the generation of false results for the darker area

Algorithm 1 Surveillance Adaptation

Input: Surveillance dataset D_s; Fashion dataset D_f

Output: Parsing model M

Initialize M by training D_f
Initialize fine surveillance dataset G_s as an empty set
Initialize fine fashion dataset $G_f = D_f$
repeat
    Generate surveillance parsing results
    Select fine surveillance data to update G_s (Algorithm 2)
    Select fine fashion data to update G_f (Algorithm 2)
    Update M by training $\{G_f, G_s\}$
    Compute the mean confidence C_{mean} of surveillance parsing results
until C_{mean} converge.
return M

4.2 Confidence evaluation

Confidence evaluation plays a significant role in the instance-based transfer learning model, which produces the criteria for fine data selection processing (in Sect. 4.3). We combine global and local information to evaluate the confidence level, which is defined as

$$C = c_g \cdot c_l, \tag{4}$$

where c_g and c_l are used to evaluate the image segmentation result performance globally and locally.

The first part of the equation c_g is the confidence value of the deep classification model (CNN-m) [67]:

$$c_g = p(y^j = 1|x, \theta), \tag{5}$$

where $p(\cdot)$ denotes the class label probabilities of the softmax part of the CNN-m, j represents that the image is from fine data or unfit data, y^j is the prediction for label j, x is the input data, θ is the parameter of the CNN-m. In this phase, we mainly focus on evaluating pedestrian profile performance. We collect pedestrian profile images from the fashion datasets and the PPSS (Pedestrian Parsing in Surveillance Scenes Dataset) [68] as positive samples and failed parsing segmentation results derived from our initial model trained on fashion datasets as negative samples, as shown in Fig. 7. Only two prediction labels (positive or negative) are included in the model, which could be considered a binary classification problem to explore the parameters of the CNN-m. A parsing result with a high global confidence c_g denotes the achievement of relatively good results for the pedestrian profiles and vice versa. Most images included in the fashion datasets only provide a frontal pose. To enrich the profiles of other poses to train the CNN-m, we also use the PPSS as an auxiliary database, as it contains images with pedestrians in various poses and shares similar human profiles with other datasets such as experimental datasets MIT and PRID (see Fig. 8), which is instrumental in calculating generic pedestrians in surveillance scenes with confidence.

The second part of the equation c_l denotes the confidence level, which only considers the local positioning and area of each label. It ensures that hair always appears in the upper section of an image, that the area ratio of sunglasses does not cover too much of an image, etc. We define the confidence level as

$$c_l = \min\{f(D_p(\Delta p)), f(D_a(\Delta a))\}, \tag{6}$$

where $f(\cdot)$ is the hyperbolic tangent used to normalize the score between 0 and 1, and where

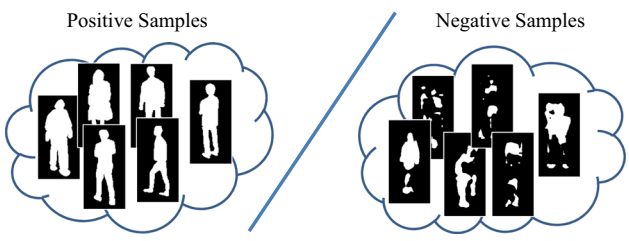


Fig. 7 Positive samples are set as pedestrian profile images from the fashion datasets and the PPSS. Negative samples are collected from failed parsing segmentation results parsed by our initial clothing parsing model

$$D_p(\Delta p) = \frac{W^2 + H^2}{\max(\Delta p)^2} - 1, \tag{7}$$

$$D_a(\Delta a) = \frac{W \cdot H}{\max(\Delta a)} - 1, \tag{8}$$

denote the position score and the area score, respectively. Here, W and H are the width and height of the input image, respectively, $\max(\Delta p)$ denotes the max Euclidean distance between each predicted regional centroid (the arithmetic mean positioning of all points in the region) and the corresponding mean position obtained from all the labels from the fashion datasets, while $\max(\Delta a)$ is the maximum difference observed from the region across all labels. A higher degree of confidence denotes better parsing results at the label level.

4.3 Data selection

Data selection (steps 6 and 7 of Algorithm 1) involves two selection parts. The first is selecting the fine data in the surveillance dataset. The second is selecting the fine data in the fashion dataset following the first part.

First, each sample of the surveillance dataset is assigned to the fine dataset if its C is greater than a threshold α (we set α to 0.5 in our experiments; more information is given in Sect. 5.2) or is otherwise assigned to the unfit dataset. After generating the fine data of the surveillance domain, we extract features from the last convolutional layer to calculate the mean feature F_{mean}. We then assign the fine fashion data with features similar to those of F_{mean} to the fine dataset. Then, the combined fine dataset is used to train a better parsing model. Algorithm 2 illustrates this data selection procedure.

The combined fine data are used to form a new model, which then selects new data.

Algorithm 2 Data Selection

Input: Last fine fashion dataset G_f; Surveillance dataset D_s; Last parsing model M

Output: Updated fine fashion dataset G_f; Fine surveillance dataset G_s

Initialize G_s as an empty set
Initialize $C = 0$
for each image I_s in D_s **do**
    Compute parsing output $L = M(I_s)$
    Compute output confidence $C = c_g \cdot c_l$
    if $C > \alpha$ **then**
        Put I_s into G_s
    end if
end for
Compute mean feature of fine surveillance data F_{mean}
for each image I_f in G_f **do**
    Compute feature F_f
    Compute feature distance $Dist = |F_f - F_{mean}|$
end for
Sort distances
Put top ninety percent I_f ordered by distance into a temp set G_{f_temp} and assign G_{f_temp} to G_f
return G_f, G_s

Fig. 8 Various poses with different orientations are reflected in the PPSS, which covers poses found in the MIT and PRID datasets. Here, we illustrate differences observed between profiles in eight orientations across the three datasets

Fig. 9 Our test samples. We label images with items at the pixel level, unlike labels used for the human-parsing related dataset

5 Experiments

In the following section, we describe experiments conducted on several datasets to investigate the performance of our surveillance adaptation approach.

5.1 Datasets

5.1.1 Training datasets

The Fashionista [18], CCP [8], MIT [17] and PRID [69] datasets are used in our experiment. Note that the first two datasets are used as source data, while the last two datasets are used as target data. Fashionista includes 685 photographs with pixel-level annotations denoted by 53 different clothing items and 3 additional labels. CCP includes 2098 high-resolution fashion photographs with significant human/clothing variations reflecting a wide range of styles, accessories, and garments. Fashion images mainly show individuals positioned in a frontal standing pose. The MIT contains 888 pictures of pedestrians with 65 attributes [70], including age, gender and non-clothing labels. These attributes have been manually reduced to 44 attributes. We use 763 pictures during domain adaptation and 25 pictures for parameter C tuning. The PRID is captured with two different static surveillance camera views. It includes 1134 images of people (1034 for retraining). The PPSS dataset is used to train and explore pedestrian profile scores as a part of the judgment model.

5.1.2 Test datasets

For each target dataset, our measurements involve 100 images for testing. There are no datasets related to clothing parsing in the surveillance domain. The data that are the closest match were collected from the human parsing work,

Table 2 Clothing parsing average image accuracy performance

Dataset	Fa-MIT	Fa-PRID	CCP-MIT	CCP-PRID
Background	0.733	0.686	0.733	0.686
ToT [66]	0.790	0.747	0.790	0.747
AdaptSegNet [63] + Item	0.821	0.773	0.813	0.763
Original	0.743	0.689	0.747	0.699
Original + Item	0.798	0.754	0.796	0.750
Transferred	0.770	0.746	0.763	0.748
Transferred + Item	**0.873**	**0.818**	**0.847**	**0.790**

The bold values indicate the best results

Table 3 Clothing parsing average foreground accuracy performance

Dataset	Fa-MIT	Fa-PRID	CCP-MIT	CCP-PRID
ToT [66]	0.254	0.210	0.254	0.210
AdaptSegNet [63] + Item	0.375	0.302	0.321	0.292
Original	0.199	0.112	0.179	0.123
Original + Item	0.328	0.244	0.294	0.241
Transferred	0.298	0.242	0.267	0.266
Transferred + Item	**0.548**	**0.404**	**0.435**	**0.389**

The bold values indicate the best results

Table 4 Ablation study

Dataset	Fa-MIT	Fa-PRID	CCP-MIT	CCP-PRID
Without SS	0.373	0.279	0.341	0.272
Without FS [23]	0.538	0.385	0.423	0.358
Without SC	0.446	0.333	0.363	0.309
Complete	**0.548**	**0.404**	**0.435**	**0.389**

Clothing parsing average foreground accuracy for four levels of the process. SS, FS, and SC represent surveillance data selection, fashion data selection and stripe constraint, respectively

The bold values indicate the best results

but they are still quite different. Therefore, we manually label these images with the same tags as those of the fashion domain at the pixel level (shown in Fig. 9).

5.2 Confidence threshold

The parameter C controls the scale and the quality of the retraining dataset. To tune its threshold α, we use a small random sample of 25 MIT images as a validation set. Then, we test the value of C from 0.1 to 0.9 by a step of 0.1, with 0.5 denoting optimal performance.

5.3 Clothing parsing accuracy

We measure the performance of labeling in terms of 3 metrics: average image accuracy, average foreground accuracy and average IoU values. The image accuracy measures the pixel accuracy of an entire image, while the foreground accuracy disregards background pixels. The average IoU measures the mean Intersection over Union (also called the Jaccard Index) score of all garments.

Four types of measurements are applied to different datasets in our experiments: from the Fashionista dataset to the MIT dataset (Fa-MIT), from the Fashionista dataset to the PRID dataset (Fa-PRID), from the CCP dataset to the MIT dataset (CCP-MIT) and from the CCP dataset to the PRID dataset (CCP-PRID). We also evaluate our models with the following 4 settings: by parsing images from the original model trained on the fashion dataset (Original), by parsing images from the original model with given item labels (Original + Item), by parsing images from the transferred model without given item labels (Transferred), and by parsing images from the transferred model with given item labels (Transferred + Item). We also compare our method to the most relevant segmentation domain adaptation model AdaptSegNet [63], which is a state-of-the-art method for synthetic-to-real urban scene domain adaptation.

Table 2 compares average levels of image accuracy. The most frequent labels found in our images are background labels. Simply predicting all regions as backgrounds (Background) results in a reasonably strong level of accuracy (73.3% and 68.6% for each surveillance dataset). We also evaluate the parsing model with our test dataset alone (train on the test (ToT)) by tenfold cross-validation and by the AdaptSegNet model. The results of

Table 5 Average IoU performance for clothing parsing performance

Dataset	Fa-MIT	Fa-PRID	CCP-MIT	CCP-PRID
ToT [66]	0.184	0.141	0.184	0.141
AdaptSegNet [63] + Item	0.250	0.198	0.221	0.155
Original	0.082	0.074	0.104	0.058
Original + Item	0.235	0.189	0.195	0.111
Transferred	0.115	0.096	0.118	0.122
Transferred + Item	**0.315**	**0.248**	**0.276**	**0.216**

The results of the transferred method substantially outperform the original results

The bold values indicate the best results

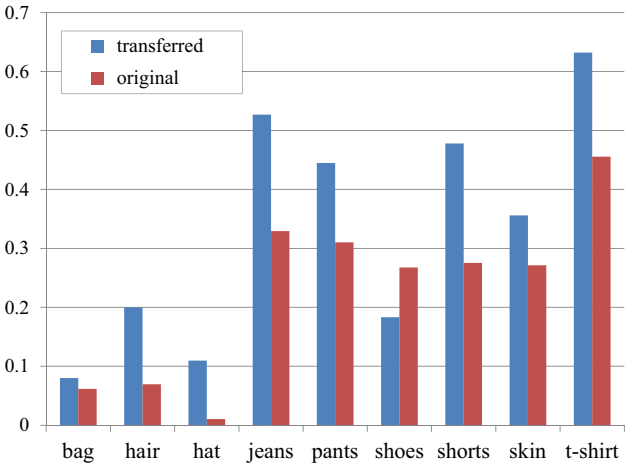


Fig. 10 IoUs of the main items. Red bars denote the IoU accuracy of the original model, and blue bars denote the IoU accuracy after transfer

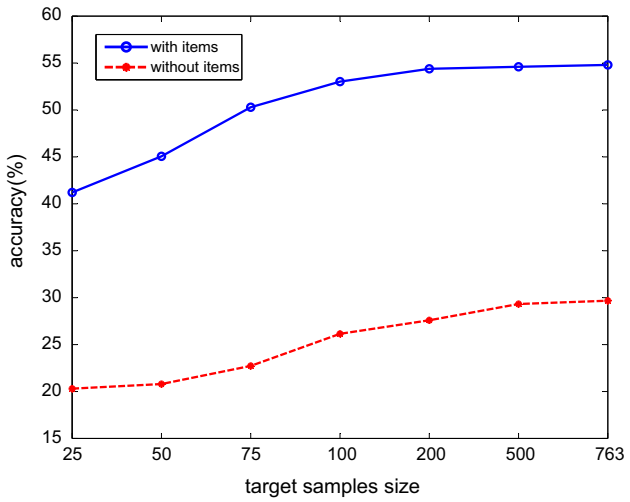


Fig. 11 The parsing accuracy of the MIT dataset when size varies. The red and blue lines indicate parsing performance with and without given item labels, respectively

these evaluations are inferior to the Background results and are used as a baseline for comparison. Inevitably, a full parsing problem with all 44 garment possibilities is quite challenging to address. Thus, the transferred clothing parsing method without given item labels (garment metadata) is less effective than that using given item labels, but it still performs better than the Background. The transferred predictions show an efficient improvement relative to the original predictions, improving the accuracy levels by 87.3%, 81.8%, 84.7% and 79.0% and drastically outperforming the AdaptSegNet results.

We further evaluate the performance for the foreground in images, as Table 3 shows. The image accuracy is based more heavily on all pixels, while the foreground accuracy mainly focuses on the body parts of pedestrians. A foreground evaluation is more challenging due to a lack of pixels confirmed from the background, leading to lower values of image accuracy. Nevertheless, our method improves foreground accuracy levels significantly. From the table, we can see that the accuracy of the transferred results with item labels roughly increases by a factor of 1.6 compared to the original results with given item labels.

To analyse the effectiveness of our method, we conduct ablation study with investigating the effect of removing certain modules. As reported in Table 4, the results of the complete process are superior to the results generated without surveillance data selection (without SS) or fashion data selection (without FS). The latter selection strategy is described in our previous work. From these results, we find that the improvement observed mainly results from the selected surveillance data. When we remove the stripe constraint module (without SC), performance drops approximately 20% compared to the complete approach.

Table 5 shows the average IoU performance for the 44 garments in the same datasets. Here, we also consider the performance of the FCN model trained on the test and the performance of the AdaptSegNet model as our baseline. Our method achieves approximately a 50–70% improvement with respect to the ToT values of the MIT and PRID datasets and an improvement of 25–35% relative to the AdaptSegNet results. Figure 10 plots the IoU scores for the major items (the 10 items most frequently found in the MIT test set) in the original model compared to the updated model. Though some items that cover a small area, like

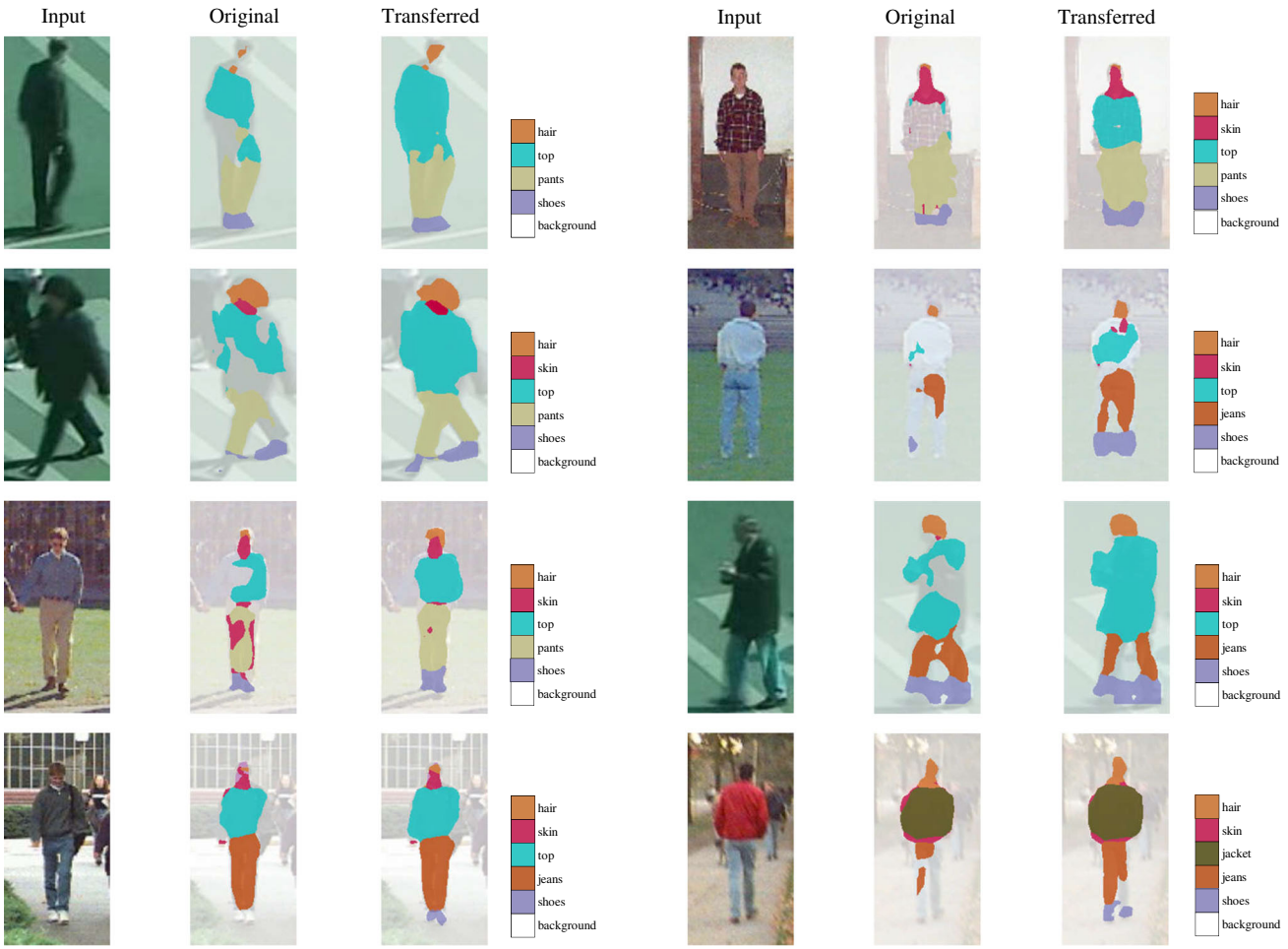


Fig. 12 Transferred parsing results (right column) compared to the corresponding original results (middle column)

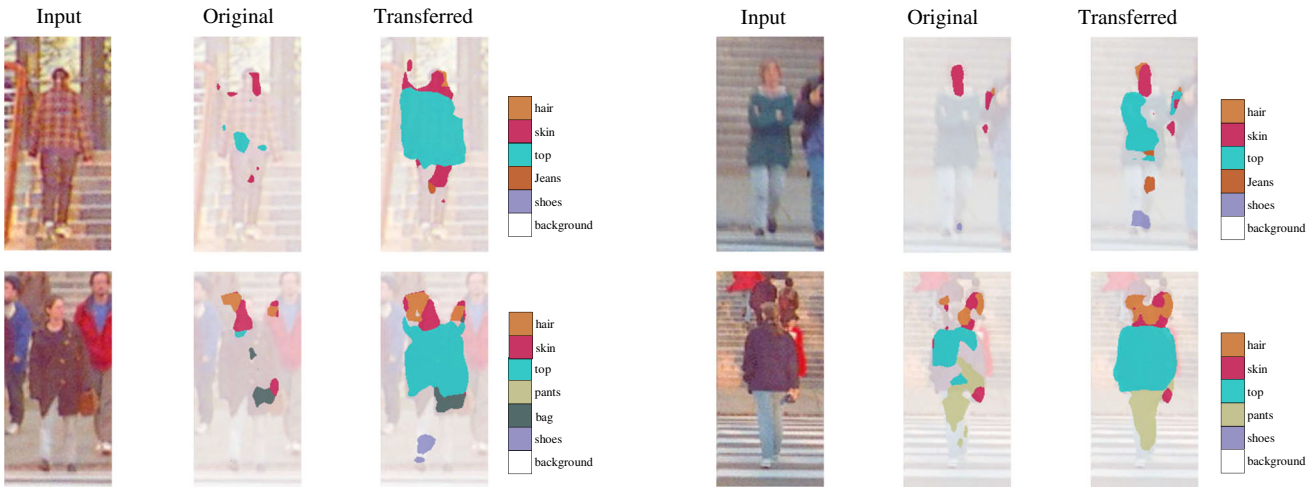


Fig. 13 Failure cases, which are still superior to the original results

shoes, exhibit worse performance after transferring, our model outperforms the original for several items, especially for major foreground items, such as T-shirts, jeans, shorts, and pants. This result leads to a significant boost in foreground accuracy levels.

In Fig. 11, we also explore the influence of target dataset scale by adjusting the MIT surveillance samples size gradually. We observe that the accuracy is more sensitive if images are with given items. It shows an upward trend until 500 target images. And given the item labels, it is impressive that even in a small target samples (only 100 images), it also obtains a considerable increase (54.8% vs. 41.2%).

5.4 Qualitative evaluation

Our work mainly focuses on the surveillance domain. Thus, we show some clothing parsing results from the MIT and PRID datasets. Figure 12 shows final parsing results compared to the original results. Our method can successfully parse clothing at a challenging resolution, illumination level, and contrast ratio. It can also manage various orientations and complex backgrounds relatively well.

Failure cases are illustrated in Fig. 13. Results can deteriorate under the following scenarios: (a) when several persons appear in a single image; (b) when some items (e.g., garments and backgrounds) share a similar appearance; and (c) when illumination conditions are poor.

6 Conclusions and future work

The proposed method makes it possible to parse clothing in a surveillance dataset lacking pixel-level labels. The core idea of our method is as follows: to use instance-based transfer learning methods to transfer a fashion-trained model to a surveillance model using only weak labels. Our experiments demonstrate that our method is effective. The proposed algorithm is simple, and its effects are significantly promoted. Here, we mainly focus on parsing clothing from full-body images. However, in real conditions, pedestrians may only show parts of their bodies. In future work, we plan to address this problem to render our method applicable to more complex surveillance environments.

Funding Funding was provided by National Natural Science Foundation of China (Grand Nos. 61872277, 61862015, U1611461, U1736206, 61876135, 61872362, 61671336, 61801335 and 61671332), National Key R&D Program of China (Grand No. 2017YFC0803700), Technology Research Program of Ministry of Public Security (Grand No. 2016JSYJA12), Hubei Province Technological Innovation Major Project (Grand Nos. 2016AAA015, 2017AAA123 and 2018AAA062), Nature Science Foundation of Hubei Province (Grand Nos. 2018CFA024 and 2019CFB472), Nature Science Foundation of Jiangsu Province (Grand No. BK20160386).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Li A, Liu L, Wang K, Liu S, Yan S (2015) Clothing attributes assisted person reidentification. IEEE Trans Circuits Syst Video Technol 25(5):869–878
2. Wang Z, Hu R, Liang C, Yu Y, Jiang J, Ye M, Chen J, Leng Q (2016) Zero-shot person re-identification via cross-view consistency. IEEE Trans Multimed 18(2):260–272
3. Ye M, Liang C, Yu Y, Wang Z, Leng Q, Xiao C, Chen J, Hu R (2016) Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. IEEE Trans Multimed 18(12):2553–2566
4. Yang J, Franco J-S, Hétroy-Wheeler F, Wuhrer S (2016) Estimation of human body shape in motion with wide clothing. In: Proceedings of the European conference on computer vision (ECCV), Amsterdam, The Netherland, pp 439–454
5. Weber M, Bauml M, Stiefelhagen R (2011) Part-based clothing segmentation for person retrieval. In: Proceedings of the IEEE international conference on advanced video and signal based surveillance (AVSS), Klagenfurt, Austria, pp 361–366
6. Simo-Serra E, Fidler S, Moreno-Noguer F, Urtasun R (2014) A high performance CRF model for clothes parsing. In: Proceedings of the Asian conference on computer vision (ACCV), Singapore, pp 64–81
7. Yamaguchi K, Kiapour MH, Berg TL (2013) Paper doll parsing: retrieving similar styles to parse clothing items. In: Proceedings of the IEEE international conference on computer vision (CVPR), Portland, OR, USA, pp 3519–3526
8. Yang W, Luo P, Lin L (2014) Clothing co-parsing by joint image segmentation and labeling. In: Proceedings of the IEEE conference on computer vision pattern recognition (CVPR), Columbus, OH, USA, pp 3182–3189
9. Tangseng P, Wu Z, Yamaguchi K (2017) Looking at outfit to parse clothing. https://arxiv.org/abs/1703.01386
10. Bourdev L, Maji S, Malik J (2011) Describing people: a poselet-based approach to attribute classification. In: Proceedings of the IEEE conference on international conference on computer vision (ICCV), Barcelona, Spain, pp 1543–1550
11. Guan P, Freifeld O, Black MJ (2010) A 2D human body model dressed in Eigen clothing. In: Proceedings of the European conference on computer vision (ECCV), Heraklion, Crete, Greece, pp 285–298
12. Gallagher AC, Chen T (2008) Clothing cosegmentation for recognizing people. In: Proceedings of the IEEE conference on computer vision pattern recognition (CVPR), Anchorage, AK, USA, pp 1–8
13. Layne R, Hospedales TM, Gong S, Mary Q (2012) Person re-identification by attributes. In: Proceedings of the British machine vision conference (BMVC), Guildfor, UK, p. 8
14. Koestinger M, Hirzer M, Wohlhart P, Roth PM, Bischof H (2012) Large scale metric learning from equivalence constraints. In: Proceedings of the IEEE conference on computer vision pattern recognition (CVPR), Providence, Rhode Island, pp 2288–2295
15. Yang M, Yu K (2011) Real-time clothing recognition in surveillance videos. In: Proceedings of the IEEE international conference on image process. (ICIP), Brussels, Belgium, pp 2937–2940
16. Shi Z, Hospedales TM, Xiang T (2015) Transferring a semantic representation for person re-identification and search. In:

Proceedings of the IEEE conference on computer vision pattern recognition (CVPR), Boston, MA, USA, pp 4184–4193
17. Oren M, Papageorgiou C, Sinha P, Osuna E, Poggio T (1997) Pedestrian detection using wavelet templates. In: Proceedings of the IEEE conference on computer vision pattern recognition (CVPR), San Juan, Puerto Rico, pp 193–199
18. Yamaguchi K, Kiapour MH, Ortiz LE, Berg TL (2012) Parsing clothing in fashion photographs. In: Proceedings of the IEEE conference on computer vision pattern recognition (CVPR), Providence, Rhode Island, pp 3570–3577
19. West J, Ventury D, Warnick S (2007) Spring research presentation: a theoretical foundation for inductive transfer. Brigham Young University, College of Physical and Mathematical Sciences
20. Chen Q, Huang J, Feris R, Brown LM, Dong J, Yan S (2015) Deep domain adaptation for describing people based on fine-grained clothing attributes. In: Proceedings of the IEEE conference on computer vision pattern recognition (CVPR), Boston, MA, USA, pp 5315–5324
21. Xiao T, Xia T, Yang Y, Huang C, Wang X (2015) Learning from massive noisy labeled data for image classification. In: Proceedings of the IEEE conference on computer vision pattern recognition (CVPR), Boston, MA, USA, pp 2691–2699
22. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
23. Zheng Q, Chen J, Liang C, Fang W, Jing X, Hu R (2017) Transferring clothing parsing from fashion dataset to surveillance. In: Proceedings of the IEEE international conference on acoustics speech and signal processing (ICASSP), New Orleans, LA, USA, pp 1667–1671
24. Lin J, Yang H, Chen D et al (2019) Face parsing with RoI Tanh-Warping. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 5654–5663
25. Chu W, Hung WC, Tsai YH et al (2019) Weakly-supervised caricature face parsing through domain adaptation. In: Proceedings of the IEEE international conference on image processing (ICIP). IEEE, pp 3282–3286
26. Zhang P, Liu W, Lei Y et al (2020) RAPNet: residual atrous pyramid network for importance-aware street scene parsing. IEEE Trans Image Process 29:5010–5021
27. Chen H, Xu ZJ, Liu ZQ, Zhu SC (2006) Composite templates for cloth modeling and sketching. In: Proceedings of the IEEE conference on computer vision on pattern recognition (CVPR), vol 1, New York, NY, USA, pp 943–950
28. Wang F, Zhao Q, Yin B, Xu T (2016) Parsing fashion image into mid-level semantic parts based on chain-conditional random fields. IET Image Process 10(6):456–463
29. Chen D, Tang Y, Zhang H, Wang L, Li X (2019) Incremental factorization of big time series data with blind factor approximation. IEEE Trans Knowl Data Eng
30. Ke H, Chen D, Shi B, Zhang J, Liu X, Zhang X, Li X (2019) Improving brain E-health services via high-performance EEG classification with grouping Bayesian optimization. IEEE Trans Serv Comput. https://doi.org/10.1109/TSC.2019.2962673
31. Chen D, Hu Y, Wang L, Zomaya AY, Li X (2016) H-parafac: hierarchical parallel factor analysis of multidimensional big data. IEEE Trans Parallel Distrib Syst 28(4):1091–1104
32. Ruan W, Liu W, Bao Q, Chen J, Cheng Y, Mei T (2019) Poinet: Pose-guided ovonic insight network for multi-person pose tracking. In: Proceedings of the 27th ACM international conference on multimedia. ACM, pp 284–292
33. Liu F, Xue S, Wu J et al (2020) Deep learning for community detection: progress, challenges and opportunities
34. Yang L, Rodriguez H, Crucianu M, Ferecatu M (2017) Fully convolutional network with superpixel parsing for fashion web image segmentation. In: International conference on multimedia modeling (MMM), Reykjavík, Iceland, pp 139–151
35. He Y, Yang L, Chen L (2017) Real-time fashion-guided clothing semantic parsing: a lightweight multi-scale inception neural network and benchmark. In: AAAI conference on artificial intelligent (AAAI), San Francisco, CA, USA
36. Zhou L, Wang Z, Luo Y, Xiong Z (2020) Separability and compactness network for image recognition and superresolution. IEEE Trans Neural Netw Learn Syst 30(11):3275–3286
37. Yi P, Wang Z, Jiang K, Shao Z, Ma J (2020) Multi-temporal ultra dense memory network for video super-resolution. IEEE Trans Circuits Syst Video Technol 30(8):2503–2516
38. Jiang K, Wang Z, Yi P, Wang G, Lu T, Jiang J (2019) Edge-enhanced GAN for remote sensing image superresolution. IEEE Trans Geosci Remote Sens 57(8):5799–5812
39. Du B, Tang X, Wang Z et al (2019) Robust graph-based semisupervised learning for noisy labeled data via maximum correntropy criterion. IEEE Trans Cybern 49(4):1440–1453
40. Fernando B, Habrard A, Sebban M, Tuytelaars T (2013) Unsupervised visual domain adaptation using subspace alignment. In: Proceedings of the IEEE international conference on computer vision (ICCV), Sydney, NSW, Australia, pp 2960–2967
41. Saenko K, Kulis B, Fritz M, Darrell T (2010), Adapting visual category models to new domains. In: Proceedings of the European conference on computer vision (ECCV), Heraklion, Crete, Greece, pp 213–226
42. Bian W, Tao D, Rui Y (2012) Cross-domain human action recognition. IEEE Trans Syst Man Cybern B 42(2):298
43. Razavian AS, Azizpour H, Sullivan J, Carlsson S (2014) Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, pp 512–519
44. Hoffman J, Tzeng E, Donahue J, Jia Y, Saenko K, Darrell T (2014) One-shot adaptation of supervised deep convolutional models. In: Proceedings of the international conference on learning representations (ICLR), Banff, Canada
45. Li X, Zhang L, Du B et al (2017) Iterative reweighting heterogeneous transfer learning framework for supervised remote sensing image classification. IEEE J Sel Top Appl Earth Observ Remote Sens 10(5):1–14
46. Zhang Z, Zhao Y, Wang Y et al (2013) Transferring training instances for convenient cross-view object classification in surveillance. IEEE Trans Inf Forens Sec 8(10):1632–1641
47. Wu J, Zhu X, Zhang C et al (2014) Bag constrained structure pattern mining for multi-graph classification. IEEE Trans Knowl Data Eng 26(10):2382–2396
48. Wu J, Pan S, Zhu X et al (2015) Boosting for multi-graph classification. IEEE Trans Cybern 45(3):416–429
49. Dong Y, Liang T, Zhang Y, Du B (2020) Spectral-spatial weighted kernel manifold embedded distribution alignment for remote sensing image classification. IEEE Trans Cybern
50. Liang X, Lin L, Yang W, Luo P, Huang J, Yan S (2016) Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval. IEEE Trans Multimed 18(6):1175–1186
51. Liu S, Feng J, Domokos C, Xu H, Huang J, Hu Z, Yan S (2014) Fashion parsing with weak color-category labels. IEEE Trans Multimed 16(1):253–265
52. Elleuch M, Mezghani A, Khemakhem M et al (2019) Clothing classification using deep CNN architecture based on transfer learning. In: International conference on hybrid intelligent systems. Springer, Cham, pp 240–248
53. Wang Z, Du B, Guo Y (2020) Domain adaptation with neural embedding matching. IEEE Trans Neural Netw Learn Syst 31(7):2387–2397

54. Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Columbus, OH, USA, pp 1717–1724
55. Liu S, Liang X, Liu L, Shen X, Yang J, Xu C, Lin L, Cao X, Yan S (2015) Matching-CNN meets KNN: quasi-parametric human parsing. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA, USA, pp 1419–1427
56. Xu X, Gong S, Hospedales T (2013) Cross-domain traffic scene understanding by motion model transfer. In: Proceedings of the 4th ACM/IEEE international workshop on Analysis and retrieval of tracked events and motion in imagery stream (ARTEMIS), Barcelona, Spain, pp 77–86
57. Rajagopal AK, Subramanian R, Ricci E, Vieriu RL, Lanz O, Ramakrishnan KR, Sebe N (2014) Exploring transfer learning approaches for head pose classification from multi-view surveillance images. Int J Comput Vis 109(1):146–167
58. Zheng WS (2012) Transfer re-identification: From person to set-based verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Providence, Rhode Island, pp 2650–2657
59. Li W, Zhao R, Wang X (2012) Human reidentification with transferred metric learning. In: Proceedings of the Asian conference on computer vision (ACCV), Daejeon, Korea, pp 31–44
60. Mckenna S (2015) Cross-scenario transfer person re-identification. IEEE Trans Circuits Syst Video Technol 26(8):1447–1460
61. Dong Q, Gong S, Zhu X (2017) Multi-task curriculum transfer deep learning of clothing attributes. In: IEEE winter conference on applications of computer vision (WACV), Santa Rosa, CA, USA, pp 520–529
62. Hoffman J, Tzeng E, Park T, Zhu JY, Isola P, Saenko K, Efros AA, Darrell T (2017) Cycada: Cycle-consistent adversarial domain adaptation. arXiv:1711.03213
63. Tsai YH, Hung WC, Schulter S, Sohn K, Yang MH, Chandraker M (2018) Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Salt Lake City, Utah, USA
64. Kumar MP, Packer B, Koller D (2010) Self-paced learning for latent variable models. In: Proceedings of the international conference on neural information processing system (NIPS), British Columbia, Canada, pp 1189–1197
65. Jiang L, Meng D, Yu SI, Lan Z, Shan S, Hauptmann A (2014) Self-paced learning with diversity. Montreal, Quebec, Canada, pp 2078–2086
66. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA, USA, pp 3431–3440
67. Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: delving deep into convolutional nets. In: Proceedings of the British machine vision conference (BMVC), Nottingham, UK
68. Luo P, Wang X, Tang X (2013) Pedestrian parsing via deep decompositional network. In: Proceedings of the IEEE international conference on computer vision (ICCV), Sydney, NSW, Australia, pp 2648–2655
69. Hirzer M, Beleznai C, Roth PM, Bischof H (2011) Person re-identification by descriptive and discriminative classification. In: Scandinavian conference on image analysis (SCIA), Ystad, Sweden, pp 91–102
70. Deng Y, Luo P, Loy CC, Tang X (2014) Pedestrian attribute recognition at far distance. In: Proceedings of the ACM international conference on multimedia (ACM MM), Orlando, FL, USA, pp 789–792