COCAS: A Large-Scale Clothes Changing Person Dataset for Re-identification

Shijie Yu*[1,2], Shihua Li*[3], Dapeng Chen[1], Rui Zhao[1], Junjie Yan[1], and Yu Qiao†[1]

[1]ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab,Shenzhen Institutes of Advanced Technology, Chinese Academy of Science
[2]University of Chinese Academy of Sciences, China
[3]Institute of Microelectronics of the Chinese Academy of Sciences

Abstract

*Recent years have witnessed great progress in person re-identification (re-id). Several academic benchmarks such as Market1501, CUHK03 and DukeMTMC play important roles to promote the re-id research. To our best knowledge, all the existing benchmarks assume the same person will have the same clothes. While in real-world scenarios, it is very often for a person to change clothes. To address the clothes changing person re-id problem, we construct a novel large-scale re-id benchmark named **ClOthes ChAnging Person Set** (COCAS), which provides multiple images of the same identity with different clothes. COCAS totally contains 62,382 body images from 5,266 persons. Based on COCAS, we introduce a new person re-id setting for clothes changing problem, where the query includes both a clothes template and a person image taking another clothes. Moreover, we propose a two-branch network named Biometric-Clothes Network (BC-Net) which can effectively integrate biometric and clothes feature for re-id under our setting. Experiments show that it is feasible for clothes changing re-id with clothes templates.*

1. Introduction

"On Tuesday, December 29, 2015, a white female suspect walked into the Comerica Bank, stating she was armed with a bomb and demanded money. The female suspect escaped with an undisclosed amount of cash. The video shows that the suspect run behind the laundromat, change clothes and flee north towards the I-94 Service Drive. "[1]

(a) A realistic scenario for clothes changing person re-id.

(b) Example of our clothes changing re-id setting

Figure 1. (a) shows a realistic case that a suspect with a black coat changed her clothes to a white coat. (b) shows our clothes changing re-id setting, where the man with white and black stripes T-shirt (the target image which is marked by a red box) is identified by the man with white T-shirt (the query image) and the white and black stripes T-shirt (the clothes template).

Person re-identification (re-id) plays more and more important roles in video surveillance systems with a wide range of applications. Previous re-id protocols [47, 46, 24, 22, 23, 39, 4, 6, 12, 2] assume the same person wears the same clothes, however, they cannot cope with the clothes changing case in the above news. As shown in Fig. 1a, the suspect wants to escape being arrested, thus she intentionally changes the clothes. The conventional re-id models [3, 34, 2, 47, 16, 33, 42] tend to fail for at least two reasons. First, these models are trained on identities with the same clothes. The clothes' appearance is statistically regarded as a kind of discriminative feature by the model. Furthermore, biometric traits like face and body shape are too weak to be learned, because they only take up a small part of the body image. On the other hand, learning a clothes-irrelevant per-

*Equally-contributed first authors(sj.yu@siat.ac.cn,lishihua@ime.ac.cn)
†Corresponding author (yu.qiao@siat.ac.cn)
[1]http://www.wjr.com/2016/01/06/woman-wanted-in-southwest-detroit-bank-robbery

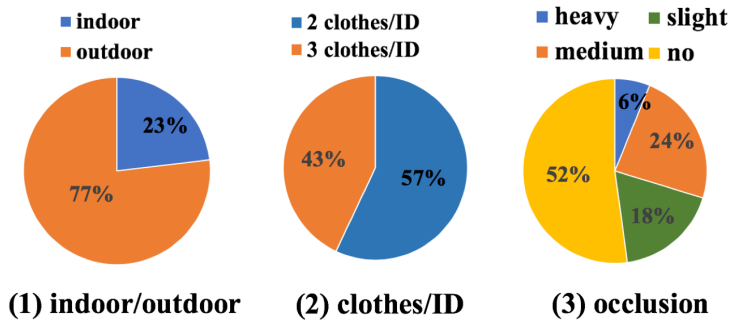

Figure 2. Statistical information of COCAS.

Figure 3. Instances in COCAS dataset.

Table 1. Comparing COCAS with public re-id datasets

Dataset	VIPeR[14]	ETHZ[32]	CUHK01[23]	CUHK03[24]	Market 1501[47]	Airport[18]	DukeMTMC[51]	COCAS
ID num	632	148	971	1,467	1,501	9,651	1,812	5,266
BBox num	1,264	8,580	3,884	14,096	32,668	39,902	36,411	62,382
Body img	hand	hand	hand	hand	DPM	ACF	hand	hand
Multi-Clot.	No	No	No	No	No	No	No	Yes

son re-id model is also difficult. The model can hardly be applied over the large-scale person image set if we only utilize unclear face and body shape information.

We instead consider an easier but practical clothes changing re-id problem as shown in Fig. 1b: a person image (target image) is searched by a clothes template image and an image of the same person wearing another clothes (query image). Compared with the current re-id setting, such setting makes use of clothes templates and still has a wide range of real applications. Take two examples, for finding the lost children/elders, the family just needs to provide a recent photo of the lost child/elder and an image of clothes the child/elder wears also. For clothes changed suspect tracking, the police can find more relevant images by a captured image and a clothes template described by a witness or left by the suspect.

To tackle the problem, we build a large-scale benchmark, named **C**l**O**thes **Ch**A**n**ging Person **S**et (COCAS). The benchmark contains 62,832 body images of 5,266 persons, and each person has 5∼25 images with 2∼3 clothes. For each person, we move the images of one kind of clothes into the gallery set, and these images are target images. We further find a clothes template from the internet, which is similar to the clothes in these target images of that person. All the remaining images are put into the query set. In fact, collecting such a large-scale person-related dataset is non-trivial. We will detailedly describe data collection, person and clothes association, privacy protection strategies and protocol definition in section 3. Biometric-Clothes Network (BC-Net) with two branches is also proposed to address clothes changing re-id. One branch extracts biometric characteristics of a person such as the face, body shape and hairstyle. The other branch extracts clothes feature, whose inputs are different for the query image and target image. A query image utilizes the clothes template for the branch to better match the target image, while a target image employs the clothes detection module to obtain a clothes image patch from itself.

In summary, our contributions are three-folder: (1) We define a kind of clothes changing re-id problem where the queries are composed of person images and clothes templates. (2) A novel large-scale dataset named COCAS is built for clothes changing re-id. (3) We propose BC-Net that can separate the clothes-relevant and clothes-irrelevant information, making the changing clothes re-id problem feasible by providing clothes templates of target images. Interesting ablation studies are conducted including examining how clothes appearance can influence the re-id. The performance of BC-Net indicates the clothes changing re-id is promising by employing the clothes templates.

2. Related Work

Re-ID Datasets. Most recent studies on person re-id are based upon data-driven methods [47, 16, 49, 33, 42], and there emerges a lot of person re-identification datasets. Among these datasets, VIPeR [14] is the earliest and most common dataset, containing 632 identities and 1,264 images captured by 2 cameras. ETHZ [32] and RAiD [10] contains 8,580 images of 148 identities and 6,920 images of 43 identities, respectively. One insufficiency of these datasets is that the data scale is too small so that they cannot fully support the training of deep neural network. Several large-scale datasets, including CUHK03 [24], Market1501 [47] and DukeMTMC [51], *etc.*, become popular along with the development of the deep neural networks. As the performance gain is gradually saturated on the above datasets, newly proposed datasets become larger and larger including

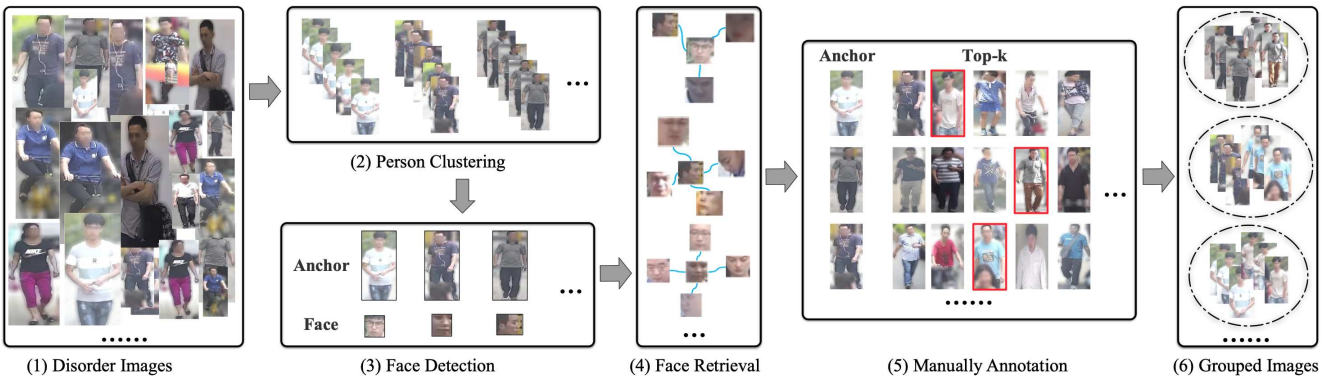

Figure 4. The pipeline of associating images of the same person. (1) The dataset with disordered body images. (2) Person image clustering. (3) Face detection and anchor image selection. (4) K-nearest neighbours searching by facial feature. (5) Manually annotation on the K-nearest neighbours to choose the truly matched person. The red boxes are examples to show the annotation results. (6) The final dataset, where each person has different types of clothes.

Airport [18], MSMT17 [39] and RPIfield [48].

Re-ID Methods. Early works on person re-id concentrated on either feature extraction [38, 28, 9, 11] or metric learning [19, 26, 5, 29]. Recent methods mainly benefits from the advances of CNN architectures, which learn the two aspects in an end-to-end fashion [7, 25, 20, 37, 42, 40]. Our work can be positioned as a kind of deep neural network. Typically, the re-id problem can be simply trained with the identification loss, by treating each person as a specific class [41]. With the deep similarity learning, person re-id is to train a siamese network with contrastive loss [35, 36], where the task is to reduce the distances between images of the same person and to enlarge the distances between the images of different persons. Several methods employed triplet loss to enforce the correct order of relative distances among image triplets, *i.e.*, the positive image pair is more similar than the negative image pair w.r.t. a same anchor image [8, 1]. Our method is different from the previous re-id methods, where the query includes both person image and clothes template. A two-branch network is employed to extract the biometric feature and clothes feature, supervised by both identification loss and triplet loss.

Clothes Changing in Re-ID. Clothes changing is an inevitable topic when it comes to re-id, but there are still less studies about it due to lack of large-scale realistic datasets. There are several related works on clothes changing re-id. Xue *et el.* [43] focused on re-id in photos based on the People in Photo Album (PIPA) dataset [45]. The work in [21] is based on a sub-dataset built from RAP [21], where 235 identities have changed clothes. Furthermore, Generative Adversial Network (GAN) is also applied to clothes changing. Jetchev *et el.* [17] introduced conditional analogy GAN to changed the person's clothes to target clothes. Zheng *et el.* [50] proposed to disentangle the person images into appearance and structure, then reassembled them to generate new person images with different clothes.

3. The COCAS Benchmark

COCAS is a large-scale person re-id benchmark that has different clothes for the same person. It contains 5,266 identities, and each identity has an average of 12 images. The images are captured in diverse realistic scenarios (30 cameras), including different illumination conditions (indoor and outdoor) and different occlusion conditions. The comparison with the existing re-id benchmark is shown in Tab. 1 and several instances fetched from COCAS dataset are shown in Fig. 3.

Data Collection. COCAS was collected in several commodity trading markets where we got permission to place 30 cameras indoors and outdoors. We recruited the people who did not mind being presented in the dataset (we promised to blur the facial region for personal privacy). As there was a high flow of people, a sufficient number of identities can be observed. As some people came to the markets every day and the data was collected in 4 different days, there were great chances to capture their images with different clothes. Clothes template images were acquired based on the collected person images. We first cropped the clothes patches from the person images by the a human parsing model, LIP [13], and searched the corresponding clothes templates by the image search engine from the Internet.

Person Association. Now we have collected the data we need, but how to associate the images of the same person with different clothes is non-trivial. It is awkward to annotate images one by one from such an enormous database of images. As shown in Fig. 4, the association has 4 main steps: *Person Clustering*: we cluster the similar person images based on the re-id feature, and manually remove the outliers images of different persons in the cluster. *Face Detection* [44]: we select one image as an anchor image from each cluster and detect the face images from the anchor images. *Face Retrieval*: we extract the facial feature by FaceNet [31] and search the top-k neighbouring anchor

images for each anchor image. *Manually Annotation*: we visualize the body images corresponding to the anchor images, and manually select the truly matched neighbouring images. Based on the association results, our dataset is arranged as follows. For each person, we select 2 or 3 different clothes where each type of clothes has 2∼5 images. Images of one kind of clothes are moved to the gallery set as the target images while other kinds are moved to query set as the query images. The partition as horizontal partition is illustrated in Fig. 5.

Privacy Protection. We blur the specific regions of the selected body images to protect the personal information, including the faces, time and locations. In greater details, MTCNN [44] has been used to get the bounding box of faces, and LIP [13] is also adopted to separate the background and body regions. We then apply the gaussian filter to blur both facial and background regions, and we call the blurred version desensitized COCAS. The experiments (section 5.2) show that the performance will drop a little if we use desensitized COCAS, but we believe the desensitized COCAS is still valuable. This is because the faces cannot be always clear and background should not be a discriminative factor for the realistic re-id problem. In this paper, most experiments are based on desensitized COCAS.

Variations. We explain the variation of COCAS. Their statistics are plotted in Fig. 2. (1) *Indoor/Outdoor*. We divide all the person images into two sets, according to the places they are captured, including 'indoor' (23%) and 'outdoor' (77%). The indoor and outdoor indicates different illumination conditions. (2) *Clothes/Person*. 2,264 identities (43%) have 3 different clothes and 3,002 identities (57%) have 2 different clothes. (3) *Occlusion*. A person image with occlusion means that the image is occluded by some obstacles like cars, trees or other persons. We also regard the case that the region of person is outside the image as a kind of occlusion. The images are categorized to four sets, including 'heavy occlusion' (6%), 'medium occlusion' (24%), 'slight occlusion' (18%) and 'no occlusion' (52%).

Protocols. Experimental protocols are defined as follows. Images of 2,800 persons are used for training, and the images of the remaining 2,466 persons are used for testing, which can be seen in Fig. 5. In testing, we take 15,985 images selected from the 2,466 persons as the query images, and take the other 12,378 images as the target images forming the gallery of testing set. We search the target images with both the query images and the clothes templates. Since a query image has multiple target images in the gallery set and CMC (Cumulative Matching Characteristic) curve can only reflect the retrieval precision of most similar target images. We additionally adopt mAP (mean Average Precision) that can reflect the overall ranking performance w.r.t. all the target images.

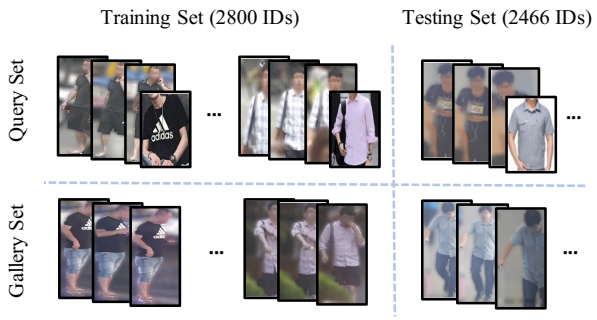

Figure 5. The partition of COCAS dataset. Vertical partition is to obtain training set and testing set according to person IDs. Horizontal partition divides COCAS into query set and gallery set according to clothes. Query set consists of query images and clothes templates while gallery set consists of target images.

4. Methodology

According to our protocol, we need to search the target image from the gallery set by a similar clothes template and the person's another image with different clothes. Intuitively, the biometric traits in the query image and the appearance of the clothes template are helpful to search the target image. Therefore, we propose the two branch Biometric-Clothes Network (BC-Net) : one branch extracts the person biometric feature, and the other extracts the clothes feature. The biometric feature branch takes the person image as the input, and employs a mask module to better exploit the clothes irrelevant information. The clothes feature branch takes the clothes image, either the clothes template or the detected clothes patch, as the input, to generate the clothes feature. The final person representation combines the biometric feature and the clothes feature.

4.1. Network Structure

BC-Net has two branches, aiming to extract biometric feature and clothes feature, respectively. The holistic architecture of BC-Net can be seen in Fig. 6.

Biometric Feature (BF) Branch. BF module takes a person image I^p as input and employs ResNet50 [15] as the backbone to yield feature maps $\mathbf{A}^p \in \mathbb{R}^{H\times W\times D}$, where H, W are the size of the feature map and D is the feature dimension. To better exploit clothes irrelevant feature from more specific regions of the person, we further design a mask module as demonstrated in Fig. 6. The module intends to emphasize the biometric feature while suppressing the feature of clothes and background. To obtain the mask $\mathbf{M}^p \in \mathbb{R}^{H\times W\times 1}$, $\mathbf{A}^p$ is first reduced to n-channel feature maps by three 1×1 convolution layers, and then each feature map is normalized by a softmax function, which takes all the $H \times W$ values as input vector. Max-pooling along channels is applied to reduce n-channel feature maps to 1-channel feature map, yielding the mask. Based on the mask

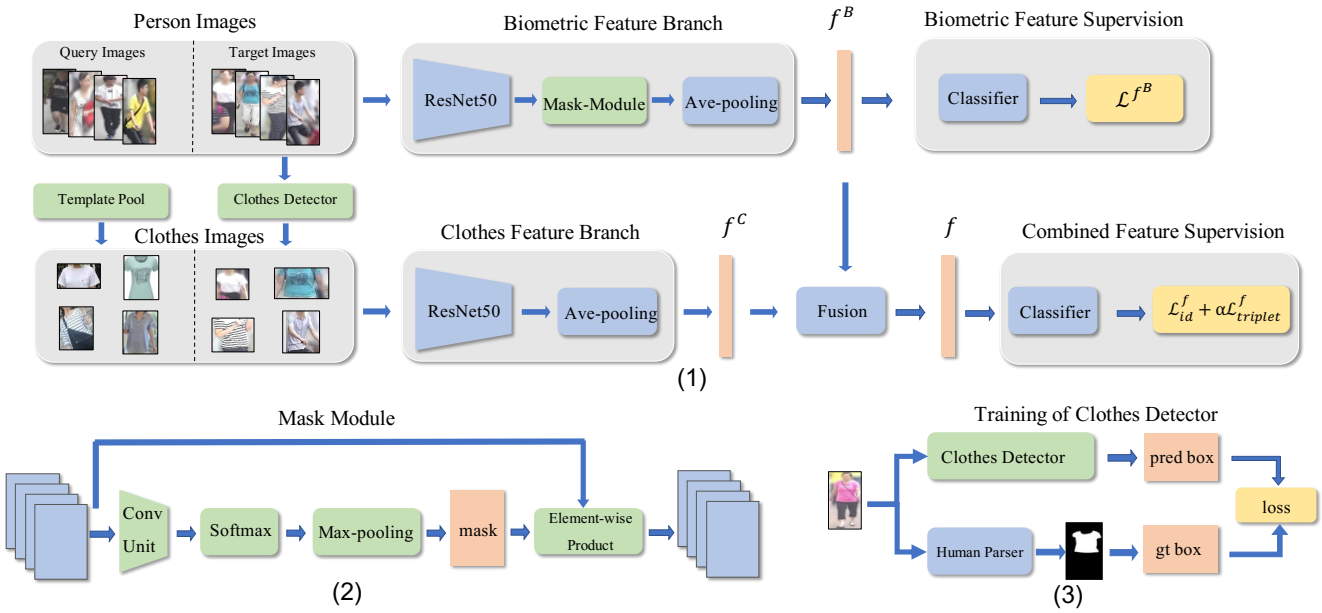

Figure 6. (1) The architecture of BC-Net. It contains two main branches including biometric feature branch and clothes feature branch. At the ends of these two streams, biometric feature and clothes feature are concatenated and then passed through a fully-connected layer to obtain the final feature with 256 dimensions. Note that the clothes detector based on faster RCNN is used to obtain clothes patches from target images. (2) The details of mask module. After convolution layers, the feature maps are normalized by softmax operation with each channels, then channel-wise max-pooling is applied to obtain the mask. At last, biometric feature is selected by an element-wise product between the mask and the input feature maps. (3) The training process of the clothes detector. LIP, a human parsing model, is applied to obtain the clothes bounding boxes of person images rather than manual annotation.

$\mathbf{M}^p$, the biometric feature $f^B \in \mathbb{R}^D$ is the obtained by the average pooling of the filtered feature map:

$$f_k^B = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} [\mathbf{A}_k^p \circ \mathbf{M}^p]_{i,j}, \tag{1}$$

where $\circ$ indicates the element-wise product, $\mathbf{A}_k^p \circ \mathbf{M}^p$ is the kth channel map of filtered feature maps and f_k^B is the kth element of f^B.

Clothes Feature (CF) Branch. CF branch tries to extract clothes related information. As our setting doesn't provide the clothes template for the target image, and we would like to process both query images and target images with the same network, thus a clothes detector is additionally employed for the target image. The clothes detector is based on Faster RCNN [30], which predicts the bounding box of the clothes from the target images. Either clothes template images or detected images are resized to the same size, and are fed into the CF branch as the input clothes image I^c. The CF branch also takes the ResNet50 as the backbone architecture, and outputs the clothes feature $f^C \in \mathbb{R}^D$ by average pooling over the feature maps $\mathbf{A}^c$.

The biometric feature f^B and corresponding clothes feature f^C are concatenated, then we estimate the feature vector $f \in \mathbb{R}^d$ by a linear projection:

$$f = \mathbf{W}[(f^B)^\top, (f^C)^\top]^\top + \mathbf{b}, \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{d \times 2D}$ and $\mathbf{b} \in \mathbb{R}^d$. f is finally normalized by its L_2 norm in both training and testing stages.

4.2. Loss Function

We employ both identification loss and triplet loss over the training samples. The nth training sample is indicated by $\mathcal{I}_n = \{I_n^p, I_n^c\}$, which consists of a person image and a clothes image. For a query person image, the clothes image is the clothes template describing the clothes in the target image. While for a target person image, the clothes image is the clothes image patch detected from itself.

The combined feature f_n can be regarded as a feature describing the target image, thus the conventional identification loss is employed for the combined features. Let $\mathcal{D} = \{\mathcal{I}_n\}_{n=1}^N$ indicate the training samples, we make use of the ID information to supervise the combined feature:

$$\mathcal{L}_{id}^f = -\frac{1}{N} \sum_{n=1}^{N} \sum_{l=1}^{L} y_{n,l} \log \left(\frac{e^{\mathbf{w}_l^\top f_n}}{\sum_{m=1}^{L} e^{\mathbf{w}_m^\top f_n}} \right), \tag{3}$$

where $\mathcal{D}$ has N images belonging to L persons. If the nth image belongs to the lth person, $y_{n,l} = 1$, otherwise $y_{n,l} = 0$. The parameters $\mathbf{w}_l$ associated with the feature embedding of the lth person.

We now define a distance function $d(\mathcal{I}_i, \mathcal{I}_j) = \|f_i - f_j\|_2^2$, and further employ triplet loss to optimize inter-sample re-

lationships. Let the ith triplet be $\mathcal{T}_i = (\mathcal{I}_i^a, \mathcal{I}_i^b, \mathcal{I}_i^c)$, and $\mathcal{I}_i^a$ is an anchor sample. $\mathcal{I}_i^b$ and $\mathcal{I}_i^a$ belong to same class while $\mathcal{I}_i^c$ and $\mathcal{I}_i^a$ are from different identities. The triplet loss is defined as:

$$\mathcal{L}_{triplet}^f = \frac{1}{N_{triplet}} \sum_{i=1}^{N_{triplet}} [d(\mathcal{I}_i^a, \mathcal{I}_i^b) + \eta - d(\mathcal{I}_i^a, \mathcal{I}_i^c)]_+, \quad (4)$$

where $N_{triplet}$ is the numbers the distance of positive pair to be smaller than the distance of negative pair at least by a margin η. The overall loss $\mathcal{L}^f$ on the combined feature f is the sum of the $\mathcal{L}_{id}^f$ and $\mathcal{L}_{triplet}^f$ defined as follows:

$$\mathcal{L}^f = \mathcal{L}_{id}^f + \alpha \mathcal{L}_{triplet}^f. \quad (5)$$

To better learn the biometric feature, we additionally impose an identification loss, denoted by $\mathcal{L}^{f^B}$.

4.3. Network Training

In BC-Net, the clothes detector and feature extractor are trained separately.

Clothes Detector Training. The clothes detector is based on Faster RCNN [30]. Instead of annotating the clothes bounding boxes manually, we employed LIP [13], an effective human parsing model. For each image in the training set, We utilize LIP to produce the clothes mask, then calculate two coordinates of the left-up corner and right-bottom corner as the ground truth bounding box. Stochastic gradient descent (SGD) is applied with momentum 0.9 for 30 epochs. 4 GPUs are employed for detector training and each GPU is set with a batch size of 12.

Feature Extractor Training. We employ SGD to optimize the feature extractor with a momentum of 0.9. The optimization lasts for 100 epochs, and the initial learning rate is 0.00035, which is further decayed to 0.00005 after 40 epochs. 4 GPUs are used for training, and the batch size of each GPU is set to 32, *i.e.*, 32 person images and 32 corresponding clothes images. The 32 samples are about 8 persons and each person has 4 samples. For triplet loss, we take each sample as anchor sample, and choose the farthest positive sample and the closet negative sample to compose a triplet.

5. Experiments

In the experiment, we first apply the current state-of-the-art approaches to the COCAS as well as other person re-id datasets only considering the identity information. Then we demonstrate how our method can improve the performance on COCAS by employing the clothes template and other post-processing strategies. Extensive ablation studies are conducted to evaluate the effectiveness of different components in our method.

Implementation Details. The input person images are resized to 256×128 and the input clothes templates are resized to 128×128. Random cropping, flipping and erasing are used for data augmentation. The margin η in Eq. 4 is set to 0.3, The loss balance weights of α is set to 1.0.

5.1. Overall Results

Learning with only ID labels. First, we treat COCAS as a common person re-id dataset with only ID labels, *i.e.*, without clothes templates. To highlight the dataset difference, we also incorporate Market1501[47] and DukeMTMC[51] for comparison. All the datasets follow the standard training and testing partition protocol. Without employing additional clothes templates, our method treats all the images equally by detecting the clothes image patch from the original images and feeding them to the clothes feature branch. The results of several state-of-the-art (SOTA) and ours are shown in Fig. 7. It can be seen that our method can perform equally well with SOTAs on existing datasets, and all the methods obtain inferior results without utilizing the clothes templates.

Learning with provided clothes templates. We now involve the clothes templates for training. In particular, the query image takes the provided the template for the clothes branch and the target image utilizes the detected clothes patch. After training, we obtain the combined feature f, which is further normalized by its L_2 norm. Compared with the feature only trained with ID labels, the combined feature significantly improves the results even the similarity is measured by the Euclidean distance. As shown in Fig. 8, it achieves 37.6% and 39.9% mAP and top-1 gains, respectively. We further study the effectiveness of two different similarity measuring schemes, *i.e.*, the metric learning method (XQDA) [27] and the re-ranking method (RR) [52]. Results in Tab. 2-2,3,4 show that XQDA and RR are effective and complementary. XQDA and RR improve the Euclidean feature distance by 8% and 10.4% mAP, and their combination achieves 21.7% mAP gains.

5.2. Ablation Study

In this section, we try to figure out what information is crucial to the clothes changing re-id. We also investigate various factors that can significantly influence the accuracy, including loss functions and clothes detector.

5.2.1 Performance Analysis

Biometric Feature v.s. Clothes Feature. To evaluate the effectiveness of the biometric feature and the clothes feature, we construct two variants for comparison. One only utilizes the biometric feature, and sets the clothes feature before fusion to be zero. The other utilizes the clothes feature in a similar manner. As shown in Tab. 2-9,10, only

No.	Experiment Setup						Performance			
	mask	feature	loss	clothes detector	dataset	metric	mAP	top-1	top-5	top-10
1.	*w/*	*BF+CF*	$\mathcal{L}^f + \mathcal{L}^{f^B}$	*faster RCNN*	*desensitized*	*Euclid*	46.8	49.3	64.0	71.4
2.	*w/*	*BF+CF*	$\mathcal{L}^f + \mathcal{L}^{f^B}$	*faster RCNN*	*desensitized*	*Euclid+RR*	54.8	53.9	60.7	69.0
3.	*w/*	*BF+CF*	$\mathcal{L}^f + \mathcal{L}^{f^B}$	*faster RCNN*	*desensitized*	*XQDA*	57.2	59.4	74.7	81.8
4.	*w/*	*BF+CF*	$\mathcal{L}^f + \mathcal{L}^{f^B}$	*faster RCNN*	*desensitized*	*XQDA+RR*	68.5	66.3	72.9	79.9
5.	*w/*	*BF+CF*	$\mathcal{L}^f + \mathcal{L}^{f^B}$	*faster RCNN*	*raw*	*Euclid*	52.8	55.3	69.5	76.1
6.	*w/*	*BF+CF*	$\mathcal{L}^f + \mathcal{L}^{f^B}$	*faster RCNN*	*raw*	*Euclid+RR*	63.7	62.3	68.0	76.2
7.	*w/*	*BF+CF*	$\mathcal{L}^f + \mathcal{L}^{f^B}$	*faster RCNN*	*raw*	*XQDA*	65.1	67.0	**80.0**	**85.7**
8.	*w/*	*BF+CF*	$\mathcal{L}^f + \mathcal{L}^{f^B}$	*faster RCNN*	*raw*	*XQDA+RR*	**75.4**	**73.3**	77.9	84.5
9.	*w/*	*BF*	$\mathcal{L}^f + \mathcal{L}^{f^B}$	*faster RCNN*	*desensitized*	*Euclid*	12.2	12.4	20.2	25.2
10.	*w/*	*CF*	$\mathcal{L}^f + \mathcal{L}^{f^B}$	*faster RCNN*	*desensitized*	*Euclid*	28.7	27.6	45.0	55.3
11.	*w/*	*BF+CF*	$\mathcal{L}^f$	*faster RCNN*	*desensitized*	*Euclid*	32.7	33.7	50.6	60.3
12.	*w/*	*BF+CF*	*w/o triplet loss*	*faster RCNN*	*desensitized*	*Euclid*	42.8	44.8	59.2	66.5
13.	*w/o*	*BF+CF*	$\mathcal{L}^f + \mathcal{L}^{f^B}$	*faster RCNN*	*desensitized*	*Euclid*	43.6	45.8	60.6	67.9
14.	*w/*	*BF+CF*	$\mathcal{L}^f + \mathcal{L}^{f^B}$	*None*	*desensitized*	*Euclid*	39.5	41.0	55.7	63.4

Table 2. Evaluation of our method on the COCAS dataset. We study the influence of mask, different features, loss function, clothes detector, desensitization, and different similarity metrics. Top-1, 5, 10 accuracies and mAP(%) are reported. *BF* and *CF* denote the biometric feature and the clothes feature respectively. The combined feature is denoted by *BF+CF*.

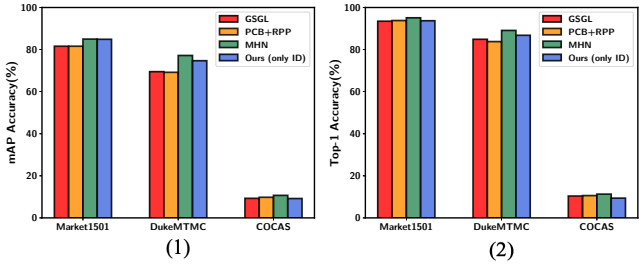

Figure 7. Training different datasets with SOTA methods and ours. (1) and (2) show the results of different methods on different methods in terms of the mAP and top-1 accuracy. Involving SOTA methods are GSGL [3], PCB+RPP [34] and MHN [2].

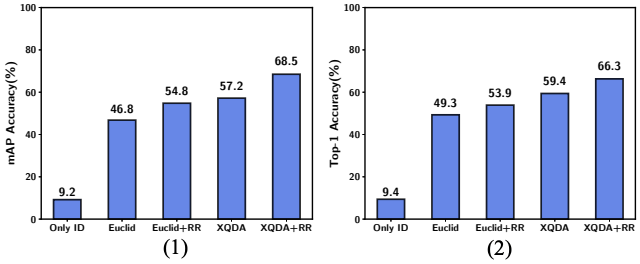

Figure 8. Training with provided clothes templates. A significant gap between using only ID and using provided clothes templates is shown above. We also demonstrate the effectiveness of XQDA[27] and re-ranking[52].

employing biometric feature or clothes feature leads to inferior results, whose mAP drops 34.6% and 18.1%, respectively. Note that the results of clothes feature are better than biometric feature, which indicates the clothes appearance is more important. Besides, the biometric feature is indispensable and complementary to the clothes feature, the final performance significantly boosts when combining the two features together. Fig. 9 demonstrates several retrieval results generated by the three features. It can be seen that the biometric feature is independent with the clothes appearance and the combined feature can actually achieve better performance.

Mask Module. To better obtain the biometric feature, the mask module is employed in the biometric feature branch. Quantitatively, the mask module improves mAP from 43.6% to 46.8% and top-1 from 45.8% to 49.3% in Tab. 2-1,13. We also visualize the mask over the original image in Fig. 10, which indicates the mask mainly focuses on the facial and the joint regions. Although the facial region is desensitized in COCAS, it still serves as an important biometric clue. Meanwhile, the joint regions are potentially related to the pose or the body shape of a person.

Influence of Desensitization. In COCAS, we have obscured the faces and backgrounds of all images for privacy protection. As the facial region conveys important biometric information, we also train BC-Net with raw COCAS. The results can be seen in Tab. 2-5,6,7,8. Compared with the results of desensitized COCAS in Tab. 2-1,2,3,4, it improves about 6% $\sim$ 9% mAP when using the same similarity metric, indicating the desensitization actually weakens the biometric information. Nevertheless, the facial information is still helpful as has been analyzed in the mask module.

Figure 9. Examples of retrieval results by clothes feature, biometric feature and the combined feature, respectively. Green box means the images belong to the same identity with the query images, while the red box indicates the incorrect retrieved results. As shown above, the images retrieved via clothes feature have similar clothes with templates, *e.g.*, red clothes find red clothes. While if only using biometric feature, the images that have similar clothes or body shape will be found. Combined feature is effective to find the images which have both characteristics of query images and clothes templates.

Figure 10. Visualization of masks. (1) shows the masks of the same person with different clothes. (2) shows the masks which are generated from images with different situations, including front, sideways, looking down, occlusion, back, *etc.*

5.2.2 Design Choices and Alternatives

Loss Function. As described in sec 4.3, BC-Net is trained by the loss functions over both biometric feature and the combined feature. We construct two variants. The first removes the loss imposed over the biometric feature, *i.e.*, training network only with $\mathcal{L}^f$. The second removes the triplet loss term in $\mathcal{L}^f$, *i.e.*, training network with the loss of $\mathcal{L}^f_{id} + \mathcal{L}^{f^B}$. The results are reported in Tab. 2-11,12. Without $\mathcal{L}^f_{triplet}$, the performance decreases 4.0% mAP and 4.5% top-1. While without $\mathcal{L}^{f^B}$, the mAP drops sharply from 46.8% to 32.7% and the top-1 accuracy drops from 49.3% to 33.7%. The results show that $\mathcal{L}^{f^B}$ is crucial to better extract the fine-grain biometric feature and filter irrelevant features out.

Clothes Detector. In BC-Net, we should first train the clothes detector, then use it to train the holistic network. To evaluate whether the clothes detector is necessary, we simply remove the clothes detector. If the person images are target images, the person images will be directly fed into both BF branch and CF branch. As the results shown in Tab. 2-14, without clothes detector, our method achieves 39.5% of mAP and 41.0% top-1, which drops 7.3% and 8.3% respectively. The clothes detector potentially removes the influence of other regions, such as the background or the trousers.

6. Conclusion

We have introduced a new person re-id benchmark considering the clothes changing problem, where each query is composed of a person image and a clothes template image. The benchmark contains over 60k images from 5,266 persons, where each identity has multiple kinds of clothes. For this re-id setting, we proposed the Biometric-Clothes Network, which can extract the biometric feature and the clothes feature, separately. Experiments have shown that traditional re-id methods perform badly when meeting clothes changing. While our method works well by utilizing the clothes templates. The proposed setting and solution is promising in tracking suspects and finding lost children/elders in real-world scenarios.

References

[1] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, 2018. 3

[2] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *ICCV*, 2019. 1, 7

[3] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *CVPR*, 2018. 1, 7

[4] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, June 2016. 1

[5] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, 2016. 3

[6] Dapeng Chen, Zejian Yuan, Gang Hua, Nanning Zheng, and Jingdong Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *CVPR*, 2015. 1

[7] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *CVPR*, 2017. 3

[8] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016. 3

[9] Etienne Corvee, Francois Bremond, Monique Thonnat, et al. Person re-identification using spatial covariance regions of human body parts. In *AVSS*, 2010. 3

[10] Abir Das, Anirban Chakraborty, and Amit K Roy-Chowdhury. Consistent re-identification in a camera network. In *ECCV*, 2014. 2

[11] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 3

[12] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 1

[13] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017. 3, 4, 6

[14] Douglas Gray, S. Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. 2007. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[16] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *CVPR*, 2018. 1, 2

[17] Nikolay Jetchev and Urs Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *ICCV Workshop*, 2017. 3

[18] Srikrishna Karanam, Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, and Richard J. Radke. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets, 2016. 2, 3

[19] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 3

[20] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017. 3

[21] D. Li, Z. Zhang, X. Chen, and K. Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. In *TIP*, 2019. 3

[22] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *CVPR*, 2013. 1

[23] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012. 1, 2

[24] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 2

[25] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 3

[26] Zhen Li, Shiyu Chang, Feng Liang, Thomas S Huang, Liangliang Cao, and John R Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013. 3

[27] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 6, 7

[28] Bingpeng Ma, Yu Su, and Frédéric Jurie. Bicov: a novel image representation for person re-identification and face verification. In *BMVC*, 2012. 3

[29] Alexis Mignon and Frédéric Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 3

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 5, 6

[31] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 3

[32] W.R. Schwartz and L.S. Davis. Learning Discriminative Appearance-Based Models Using Partial Least Squares. In *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing*, 2009. 2

[33] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018. 1, 2

[34] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 1, 7

[35] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 3

[36] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016. 3

[37] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016. 3

[38] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. Shape and appearance context modeling. In *ICCV*, 2007. 3

[39] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*. 1, 3

[40] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 3

[41] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017. 3

[42] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, 2018. 1, 2, 3

[43] Jia Xue, Zibo Meng, Karthik Katipally, Haibo Wang, and Kees van Zon. Clothing change aware person identification. In *CVPR Workshop*, 2018. 3

[44] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016. 3, 4

[45] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *CVPR*, 2015. 3

[46] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, and Shengjin Wang. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 1

[47] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 2, 6

[48] Meng Zheng, Srikrishna Karanam, and Richard J. Radke. Rpifield: A new dataset for temporally evaluating person re-identification. In *CVPR Workshop*, 2018. 3

[49] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *ICCV*, 2015. 2

[50] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019. 3

[51] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 2, 6

[52] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 6, 7