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

Shijie Yu∗1,2, Shihua Li∗3, Dapeng Chen1, Rui Zhao1, Junjie Yan1, and Yu Qiao†1

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

∗Equally-contributed first authors(sj.yu@siat.ac.cn, lishihua@ime.ac.cn)
†Corresponding author (yu.qiao@siat.ac.cn)
http://www.wjr.com/2016/01/06/woman-wanted-in-southwest-detroit-bank-robbery
Table 1. Comparing COCAS with public re-id datasets

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ID num</td>
<td>632</td>
<td>148</td>
<td>971</td>
<td>1,467</td>
<td>1,501</td>
<td>9,651</td>
<td>1,812</td>
<td>5,266</td>
</tr>
<tr>
<td>BBox num</td>
<td>1,264</td>
<td>8,580</td>
<td>3,884</td>
<td>14,096</td>
<td>32,668</td>
<td>39,902</td>
<td>36,411</td>
<td>62,382</td>
</tr>
<tr>
<td>Body img</td>
<td>hand</td>
<td>hand</td>
<td>hand</td>
<td>hand</td>
<td>DPM</td>
<td>ACF</td>
<td>hand</td>
<td>hand</td>
</tr>
<tr>
<td>Multi-Clot.</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Figure 2. Statistical information of COCAS.

Figure 3. Instances in COCAS dataset.

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...
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 (indoors and outdoors) 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.

3. The COCAS Benchmark

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 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.

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 \(P\) as input and employs ResNet50 [15] as the backbone to yield feature maps \(A^p \in \mathbb{R}^{H \times W \times D}\), 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 \(M^p \in \mathbb{R}^{H \times W \times 1}\), \(A^p\) is first reduced to n-channel feature maps by three \(1 \times 1\) 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...
Clothes Detector
Classifier

Ave-pooling

Element-wise

on Faster RCNN is deployed for the target image. The clothes detector is based on the same network, thus a clothes detector is additionally employed to process both query images and target images with the clothes template for the target image, and we would like to obtain the clothes bounding boxes of person images rather than manual annotation.

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

$$f^B_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} [A^k \circ M^k]_{i,j}, \quad (1)$$

where $\circ$ indicates the element-wise product, $A^k \circ M^k$ is the $k$th channel map of filtered feature maps and $f^B_k$ is the $k$th element of $f^B$.

**Clothes Feature (CF) Branch.** CF branch tries to extract clothes related information. As we set 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 $A^C$.

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:

$$f = W[(f^B)^\top, (f^C)^\top]^\top + b, \quad (2)$$

where $W \in \mathbb{R}^{d \times 2D}$ and $b \in \mathbb{R}^d$. $f$ 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 $n$th training sample is indicated by $\mathcal{I}_n = \{I^p_n, I^c_n\}$, 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:

$$\mathcal{L}_{id}^f = -\frac{1}{N} \sum_{n=1}^{N} \sum_{l=1}^{L} \frac{y_{n,l} \log \left( \frac{c w_i^\top f_n}{\sum_{m=1}^{L} c w_m^\top f_n} \right)}{\sum_{m=1}^{L} c w_m^\top f_n}, \quad (3)$$

where $\mathcal{D}$ has $N$ images belonging to $L$ persons. If the $n$th image belongs to the $l$th person, $y_{n,l} = 1$, otherwise $y_{n,l} = 0$. The parameters $w_i$ associated with the feature embedding of the $l$th 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 $i$th triplet be $T_i = (I_i^a, I_i^b, I_i^c)$, and $I_i^a$ is an anchor sample. $I_i^b$ and $I_i^c$ belong to the same class while $I_i^e$ and $I_i^n$ are from different identities. The triplet loss is defined as:

$$
L_{\text{triplet}}^f = \frac{1}{N_{\text{triplet}}} \sum_{i=1}^{N_{\text{triplet}}} \left[ d(I_i^a, I_i^b) + \eta - d(I_i^a, I_i^c) \right]_+, \quad (4)
$$

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

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

To better learn the biometric feature, we additionally impose an identification loss, denoted by $L_{\text{id}}^f$.

### 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 $256 \times 128$ and the input clothes templates are resized to $128 \times 128$. Random cropping, flipping and erasing are used for data augmentation. The margin $\eta$ in Eq. 4 is set to 0.3, The loss balance weights of $\alpha$ 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 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
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 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

**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.](image)

![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[32].](image)

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

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% ~ 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.
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 $\mathcal{L}^t$, i.e., training network with the loss of $\mathcal{L}^f_{id} + \mathcal{L}^f_{attr}$. The results are reported in Tab. 2-11,12. Without $\mathcal{L}^t_{triplet}$, the performance decreases 4.0% mAP and 4.5% top-1. While without $\mathcal{L}^f_{attr}$, 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 $\mathcal{L}^f_{attr}$ 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

[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


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


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