Learning Clothes-irrelevant Cues for Clothes-Changing Person Re-identification

Jingyi Mu
mjy0421@njust.edu.cn
Yong Li
yong.li@njust.edu.cn
Jun Li
junli@njust.edu.cn
Jian Yang
csjyang@njust.edu.cn

PCA Lab
Nanjing University of Science and Technology
Nanjing, China

Abstract

Person re-identification (re-ID), aiming to match a target person in a series of cross-camera images, is a challenging problem when people change their clothes. Essentially, different clothes are used to learn distinctive features usually, resulting in a failed identification. To mitigate this issue, we propose a Clothes-Relevant information Erasure (CRE) module to drive the model to adaptively learn clothes-irrelevant cues by utilizing the person’s semantic information to eliminate the clothes-relevant features. Furthermore, we introduce a Body Shape-Guided Attention (BSGA) module so that the model can learn richer and more discriminative features. Compared to the state-of-the-art baselines, the experimental results on three benchmark datasets show the effectiveness and superiority of our method in the clothes-changing re-ID task.

1 Introduction

Person re-identification (re-ID), as a cross-camera tracking task, is a technique that exploits computer vision algorithms to match a target person in cross-camera images or video sequences. Most existing works [10, 16, 17, 27, 40, 41, 44] follow in clothes-unchanging re-ID datasets [25, 57, 60], assuming that a person’s clothes will not change in a short period of time. One significant reason is that clothes-relevant information (e.g. color, style, part details and texture) can provide crucial discriminative features for identifying when facing challenging factors, such as different viewpoints, varying low-image resolutions, illumination changes and unconstrained poses. However, these clothes-relevant features are easy to result in a failed identification when one person’s clothes change in a long period of time and different persons wear the same or similar clothes.

To overcome this problem, some clothes-changing re-ID methods [2, 6, 37, 54] introduce some additional side information to learn the expected clothes-irrelevant features, e.g. contour sketches [54], 3D shapes [2], radio signals [6], and skeletons [37]. At this point, the person’s biological information, such as body shape, gait, hairstyle, face, etc., should play
an important role in identification. However, these models are not comprehensive enough to explore these biological patterns, the learned features are usually not discriminative enough for re-ID under clothes-changing conditions. For example, two different people may have similar body shapes, but their hairstyles and faces are very different, so it is difficult to distinguish them if we merely learn the body shape features. In addition, as shown in Figure 1(b), a strong backbone ResNet-50 [12] is not robust to the changing clothes as it pays more attention to the whole body and the details of the clothes. These observations drive us to extract enough yet more clothes-irrelevant features for robust re-ID.

In this work, our goal is to develop a light-weight and accurate model for clothes-changing re-ID. Firstly, we propose a Clothes-Relevant information Erasure (CRE) module. Specifically, we utilize a pre-trained human parsing model [24] to obtain the semantic information of the human body and identify the clothing parts of the human in the original image, including upper clothes, pants, dress, skirt, belt, bag and scarf. Then we erase these parts and retain the other clothes-irrelevant parts. The erased images and the original images are put together into the network for training. As shown in Figure 1(c), the model pays more attention to the head and limbs after eliminating the interference of changing clothes.

Furthermore, to address the background noise, we present a Body Shape-Guided Attention (BSGA) module into the network for learning more discriminative features. Specifically, the binary body mask is used to guide the learning of the spatial attention map so that the model can learn more discriminative spatial features adaptively. At the same time, the model will pay more attention to rich features of the human body, such as the body shape (see Figure 1(d)). In summary, the main contributions of this work are as follows:

- We propose a Clothes-Relevant information Erasure (CRE) module to erase the clothes-relevant information on the original images and put the generated images into the network for training so that the model can learn more biological features which are clothes-irrelevant.

- We further present a Body Shape-Guided Attention (BSGA) module into the network so as to enable the model to learn richer and more discriminative features.

- Our approach not only achieves superior performance on benchmark clothes-changing re-ID datasets, but also introduces a small number of negligible parameters.
2 Related Work

Person Re-identification. Early re-ID methods focus on two key tasks: One is feature representation learning [8, 9, 29, 34], another one is deep metric learning [23, 33, 35, 47, 51]. In the era of deep learning, most methods use convolutional neural networks to extract discriminative deep features and the models are optimized by using siamese loss [1, 26, 30, 42, 52, 56], triplet loss [3, 5] or cross-entropy loss [7, 49, 50, 58, 59]. Recently, transformer-based re-ID [13] has gradually become a new powerful baseline. However, these methods are based on the premise that a person’s clothes remain unchanged, and the identification depends more on the clothes-relevant features. These methods easily suffer from the interference of changing clothes, resulting in wrong discrimination.

Clothes-changing Person Re-identification. Some methods already exist to process the clothes-changing re-ID task [2, 11, 15, 19, 37, 38, 54]. Yang et al. [54] utilizes a spatial polar transformation to extract angle-specific features. Qian et al. [37] uses shape distillation to eliminate the identity-irrelevant clothing cues. Hong et al. [15] learns fine-grained discriminative mask with the guidance of identities and extracts fine-grained body shape features. Gu et al. [11] uses adversarial learning to mine the clothes-irrelevant features. Shu et al. [38] generates semantic-guided pixel-sampling images to make the re-ID model automatically exploit clothes-irrelevant cues. Inspired by [28, 38], we also use the human semantic maps to obtain the clothing position information. Instead of introducing the pixel bank, we directly cover the clothing parts on the original images so as to reduce the time and spatial cost. In fact, the model occasionally learns some background clutters in the process of learning clothes-irrelevant features, which will lead to failed identification. Moreover, there are rich biological features of the human body. Therefore, our BSGA module utilizes a body shape mask to guide the learning of the attention map so that the model can learn richer and more discriminative features.

3 Method

In this section, we first introduce the CRE module, which targets to obtain the generated images after erasing clothes-relevant information and the binary body shape masks. Then we elaborate our proposed BSGA module, which is embedded directly between the second and third stage in the ResNet-50 backbone. Finally, we describe the loss function to guide the feature representation learning.

3.1 The Clothes-Relevant information Erasure (CRE) module

Given an input original RGB image of a person, we first utilize the human parsing model [24] to segment the semantic parts of the human body. The model divides the human body into 18 parts: background, hat, hair, sunglasses, upper clothes, skirt, pants, dress, belt, left shoe, right shoe, face, left leg, right leg, left arm, right arm, bag, scarf. Since we only need to identify the clothes-relevant semantic parts, we merge the parts for convenience. For instance, the hat, hair, sunglasses, and face are merged into the overall head part. Finally, we divide an original person image into 7 parts: background, head, tops, bottoms, dress, arm, lower limbs.

The framework of CRE is shown in Figure 2. Given an original person RGB image $X \in \mathbb{R}^{H \times W \times 3}$, we obtain its semantic segmentation map $S \in \mathbb{P}^{H \times W}$ ($\mathbb{P} = \{0, 1, 2, 3, 4, 5, 6\}$)
through the human body parsing model [24]. \( H \) and \( W \) denote the height, and width of the image. 0 to 6 represent the labels of background, head, tops, bottoms, dress, arm, lower limbs respectively. \( f_e(\cdot) \) denotes the mapping from the original image to the generated image. Simply, our clothes-relevant information erasure operation \( f_e(X, S) \) can be expressed as:

\[
X_{ijk} = 0, \text{if } S_{ij} = \{2, 3, 4\},
\]

where \( k \in \{0, 1, 2\}, i \in [0, H), j \in [0, W) \).

The erased images lack diversity and would cause overfitting if we only use them for training. Thus, we exploit both the erased images and the original images as the augmented training dataset. Note that the erased images share the same identity annotation as the original images. In our CRE module, we can also obtain the binary body shape masks for use by the BSGA module, which will be introduced in the next section.

### 3.2 The Body Shape-Guided Attention (BSGA) module

Considering the computational complexity, we only insert the BSGA module at one bottleneck of the model where the downsampling of feature maps occurs. Therefore, we embed it directly between ResNet-50 stage2 and stage3. The output feature map of ResNet-50 backbone stage2 is \( X_{\text{stage}2} \in \mathbb{R}^{512 \times h \times w} \). Then we utilize a dimension reduction operation to get the feature map \( X_f \in \mathbb{R}^{2 \times h \times w} \). \( M_x \in \{0, 1\}^{1 \times H \times W} \) is the binary body shape mask obtained by the CRE module, where \( H, W \) indicate the height, width of the input image. First, we apply bilinear interpolation to downsample \( M_x \) to \( \{0, 1\}^{1 \times h \times w} \). We combine \( M_x \) and \( X_f \) at the channel level to obtain a new feature map \( X_c \in \mathbb{R}^{3 \times h \times w} \). \( X_c \) can be calculated as follows:

\[
\begin{align*}
X_f &= f_{7 \times 7}(X_{\text{stage}2}), \\
X_c &= [f_{7 \times 7}(M_x^\prime); X_f],
\end{align*}
\]

where \( f_{7 \times 7}(\cdot) \) represents a convolution operation with the filter size of \( 7 \times 7 \). \( M_x^\prime \) denotes the body shape mask after downsampling. Then we need to use \( X_c \) to learn a simple attention map \( A_x \in \mathbb{R}^{1 \times h \times w} \) in the following way:

\[
A_x = f_{7 \times 7}(X_c).
\]

At this point, \( A_x \) is a spatial attention map generated based on the input feature map and the body shape mask. The original feature map also makes a certain contribution in this process, so \( A_x \) still focuses on some background clutter. To better eliminate the distracting noise and enable the model to learn more discriminative features, our final spatial attention map \( A_f \) is computed as:

\[
A_f = \phi_a(A_x, M_x^\prime) = \sigma(A_x \odot f_{7 \times 7}(M_x^\prime)), \tag{4}
\]

![Figure 2: The structure of the Clothes-Relevant information Erasure (CRE) module.](image)
where \( \odot \) denotes the Hadamard product, \( \sigma \) is the sigmoid activation. In the final generated spatial attention map, there will be higher scores in the regions of the body shape and lower scores in the background clutter regions. It will drive the model to pay more attention to the shape of the human body regions on the input feature map. Eventually, there is a spatial weighting operation between \( A_f \) and the input feature map \( X_{\text{stage2}} \) to get the output feature map \( X_{\text{out}} \). This process can be described as follow:

\[
X_{\text{out}} = X_{\text{stage2}} \odot A_f.
\]  

We send the weighted feature map \( X_{\text{out}} \) to ResNet-50 backbone stage3 as input. In this process, we only introduce a few parameters into convolution kernels, and the number of parameters is so small that it can almost be ignored. At the same time, except for several convolution operations, only two spatial weighting operations are involved, resulting in a negligible computational cost.

### 3.3 Loss Function

#### Identity Loss. For a mini-batch of size \( M \), the generated images after erasing clothes-relevant information are put into the network together with the original images. Therefore, the size of the batch becomes \( 2M \). For a sample \( x_i \) with label \( y_i \), \( p(x_i) \) is the predicted probability, and encoded as \( p(y_i|x_i) \) with a softmax function. The identity loss is then computed by the cross-entropy:

\[
L_{id} = -\frac{1}{2M} \sum_{i}^{2M} \log(p(y_i|x_i)).
\]  

#### Pairwise loss. Triplet loss treats the re-ID model training process as a retrieval ranking problem \([55]\). Its goal is to minimize the distance between the positive pair and maximize the distance between the negative pair, which is widely used in re-ID tasks. In our approach,
Figure 4: Two triplet mining methods. (a) is Batch Hard mining. A triplet contains an anchor, the hardest positive sample and the hardest negative sample. (b) is the Batch Cross-clothes mining we propose in this paper. A triplet contains an anchor, a positive sample of different clothes label and the hardest negative sample (In a mini-batch, there may be more than one positive sample which has different clothes label from the anchor).

since each person has more than one kind of clothes, to fully extract clothes-irrelevant features, we mine triplets through the following method named **Batch Cross-clothes mining**:

In a mini-batch, $x_i$ is an anchor, $x_j$ is a positive sample with the same identity but different clothes label and $x_k$ is the negative sample closest to $x_i$. In this way, we can sufficiently minimize the distance between the positive pair of different clothes without introducing a large number of simple triplets. For an anchor, the triplet loss is defined as:

$$L_{BCTri} = \sum_{n=1}^{N} \max(m + d(x_i, x_j) - d(x_i, x_k), 0),$$

where $d(\cdot)$ measures the distance between two samples. In this paper, we use cosine distance as the distance measure. $N$ represents the number of positive samples with the same identity as $x_i$ but different clothing in a mini-batch. One anchor keeps $N$ triplets.

It should be noted that the feature vectors between the batch norm layer [20] and the classifier are utilized to calculate the Batch Cross-clothes triplet loss. Our method is simple to calculate, only using identity loss and pairwise loss without introducing additional loss functions. The total loss can be denoted as:

$$\mathcal{L} = \mathcal{L}_{id} + \lambda \mathcal{L}_{BCTri},$$

where $\lambda$ value is usually set as 1.

4 Experiments

In this section, we thoroughly evaluate our CRE and BSGA. First, we introduce the datasets and evaluation protocols. Then we describe the implementation in detail. We further compare our method with the state-of-the-arts. To explore the efficacy of our method, we perform the ablation study and visual inspections.
4.1 Datasets and Settings

VC-Clothes is a virtual dataset synthesized by GTA5 with 4 scenes. It contains 19060 images from 512 identities. The same person under Camera 2 and Camera 3 wears the same clothes, while the same person under Camera 3 and Camera 4 wears different clothes.

PRCC contains 33,698 images with 221 identities, all of which are real scenarios captured by 3 different cameras. A person wears the same clothes under camera A and camera B, while different clothes under camera A and camera C.

NKUP is captured by 15 cameras installed on the university campus. It contains 9,738 images from 107 identities. Each person in this dataset wears two or three types of clothing. The same person wears different kinds of clothes in query and gallery.

Evaluation Protocol. Mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC top-k) are standard metrics used to measure re-ID performance. Three kinds of test settings are defined as follows: (1) clothes-changing setting (CC). The samples with the same identity, camera view, and clothes are discarded to calculate accuracy. (2) same-clothes setting (SC). The samples with the same identity and camera view are discarded but retain the same clothes samples to calculate accuracy. (3) general setting. The samples with the same identity and camera view are discarded and both cross clothes samples and the same clothes samples are used to calculate accuracy.

4.2 Implementation Details

We use ResNet-50 which is initialized with ImageNet pre-trained model as the backbone of re-ID model. We remove the last downsampling of ResNet-50 to enrich the granularity. Following [32], the input images are resized to $256 \times 128$. $m$ in $\mathcal{L}_{BCTri}$ is set to 0.3. Random horizontal flipping, random cropping, and random erasing are used for data augmentation. The batch size is set to 32. In each batch, we randomly sample 8 identities and 4 instances for each person. The model is trained by Adam for 60 epochs, and the learning rate is initialized to 0.0003 and divided by 10 after every 20 epochs.

4.3 Comparison with State-of-the-art Methods

<table>
<thead>
<tr>
<th>method</th>
<th>ref</th>
<th>general(all cams)</th>
<th>SC(cam2&amp;cam3)</th>
<th>CC(cam3&amp;cam4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>top-1 mAP</td>
<td>top-1 mAP</td>
<td>top-1 mAP</td>
</tr>
<tr>
<td>MDLA [36]</td>
<td>ICCV2017</td>
<td>88.9</td>
<td>76.8</td>
<td>93.4</td>
</tr>
<tr>
<td>PCB [41]</td>
<td>ECCV2018</td>
<td>87.7</td>
<td>74.6</td>
<td>94.7</td>
</tr>
<tr>
<td>Part-aligned [39]</td>
<td>ECCV2018</td>
<td>90.5</td>
<td>79.7</td>
<td>93.9</td>
</tr>
<tr>
<td>TransReID [13]</td>
<td>ICCV2021</td>
<td>90.5</td>
<td>80.1</td>
<td>95.1</td>
</tr>
<tr>
<td>FSAM [15]</td>
<td>CVPR2021</td>
<td>-</td>
<td>-</td>
<td>94.7</td>
</tr>
<tr>
<td>3DSL [2]</td>
<td>CVPR2021</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PS [38]</td>
<td>SPL2021</td>
<td>93.1</td>
<td>84.9</td>
<td>94.7</td>
</tr>
<tr>
<td>CAL [11]</td>
<td>CVPR2022</td>
<td>92.9</td>
<td>87.2</td>
<td>95.1</td>
</tr>
<tr>
<td>ours w/o BSGA</td>
<td></td>
<td>94.2</td>
<td>85.7</td>
<td>94.9</td>
</tr>
<tr>
<td>ours</td>
<td></td>
<td>94.4</td>
<td><strong>88.2</strong></td>
<td>94.9</td>
</tr>
</tbody>
</table>

Comparison on VC-Clothes. In Table 1, we summarize the results of our method together with other competitive methods. Our method surpasses all previous methods in both general setting and clothes-changing setting. In particular, in clothes-changing setting, our
Table 2: Comparison with state-of-the-art methods on PRCC.

<table>
<thead>
<tr>
<th>method</th>
<th>ref</th>
<th>SC</th>
<th>CC</th>
<th>SC</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HACNN [17]</td>
<td>CVPR2018</td>
<td>82.5</td>
<td>-</td>
<td>21.8</td>
<td>-</td>
</tr>
<tr>
<td>PCB [41]</td>
<td>ECCV2018</td>
<td>99.8</td>
<td>97.0</td>
<td>41.8</td>
<td>38.7</td>
</tr>
<tr>
<td>IANet [16]</td>
<td>CVPR2019</td>
<td>99.4</td>
<td>98.3</td>
<td>46.3</td>
<td>45.9</td>
</tr>
<tr>
<td>TransReID [40]</td>
<td>ICCV2021</td>
<td>97.3</td>
<td>95.9</td>
<td>47.1</td>
<td>49.3</td>
</tr>
<tr>
<td>SPT+ASE [21]</td>
<td>TPAMI2019</td>
<td>64.2</td>
<td>-</td>
<td>34.4</td>
<td>-</td>
</tr>
<tr>
<td>GI-ReID [44]</td>
<td>CVPR2022</td>
<td>80.0</td>
<td>-</td>
<td>33.3</td>
<td>-</td>
</tr>
<tr>
<td>RCSANet [12]</td>
<td>ICCV2021</td>
<td>100</td>
<td>97.2</td>
<td>50.2</td>
<td>48.6</td>
</tr>
<tr>
<td>3DSL [19]</td>
<td>CVPR2021</td>
<td>-</td>
<td>-</td>
<td>51.3</td>
<td>-</td>
</tr>
<tr>
<td>FSAM [15]</td>
<td>CVPR2021</td>
<td>98.8</td>
<td>-</td>
<td>54.5</td>
<td>-</td>
</tr>
<tr>
<td>PS [20]</td>
<td>SPL2021</td>
<td>99.2</td>
<td>96.6</td>
<td>61.1</td>
<td>58.3</td>
</tr>
<tr>
<td>CAL [22]</td>
<td>CVPR2022</td>
<td>100</td>
<td>99.8</td>
<td>55.2</td>
<td>55.8</td>
</tr>
<tr>
<td>ours w/o BSGA</td>
<td></td>
<td>98.8</td>
<td>96.0</td>
<td>59.7</td>
<td>56.8</td>
</tr>
<tr>
<td>ours</td>
<td></td>
<td>99.6</td>
<td>97.3</td>
<td>61.8</td>
<td>58.7</td>
</tr>
</tbody>
</table>

Table 3: Comparison with state-of-the-art methods on NKUP.

<table>
<thead>
<tr>
<th>method</th>
<th>ref</th>
<th>top-1</th>
<th>top-5</th>
<th>top-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>se-resnext [18]</td>
<td>CVPR2018</td>
<td>16.7</td>
<td>25.5</td>
<td>31.2</td>
</tr>
<tr>
<td>se-ResNet [18]</td>
<td>CVPR2018</td>
<td>18.2</td>
<td>25.2</td>
<td>30.6</td>
</tr>
<tr>
<td>PCB [41]</td>
<td>ECCV2018</td>
<td>16.9</td>
<td>25.6</td>
<td>30.6</td>
</tr>
<tr>
<td>MGN [17]</td>
<td>ACM MM2018</td>
<td>18.8</td>
<td>28.8</td>
<td>33.0</td>
</tr>
<tr>
<td>TransReID [40]</td>
<td>ICCV2021</td>
<td>21.2</td>
<td>32.4</td>
<td>37.6</td>
</tr>
<tr>
<td>PS [21]</td>
<td>SPL2021</td>
<td>18.8</td>
<td>35.2</td>
<td>45.5</td>
</tr>
<tr>
<td>LSD [22]</td>
<td>IVC2021</td>
<td>16.4</td>
<td>27.9</td>
<td>34.8</td>
</tr>
</tbody>
</table>

method outperforms the second best method by a margin of around 2-3 pp. (percentage point) w.r.t. mAP and top-1 accuracy.

**Comparison on PRCC.** We make comparisons on our approach with four traditional re-id methods and seven clothes-changing re-id methods on PRCC in Table 2. We can note that our method outperforms all other methods in clothes-changing setting. It should be noted that in these methods, GI-ReID, 3DSL and FSAM all integrate other modalities into the model. However, the performance of our method in same-clothes setting is not the best. The main reason is that our method pays more attention to clothes-irrelevant features, such as head and limbs. Although these features have certain discriminative power, clothes-relevant cues contribute more to identification when clothes remain unchanged. Therefore, the performance of our method is reduced but still at a high level.

**Comparison on NKUP.** Table 3 shows the comparative experimental results of our approach with other re-ID methods on NKUP. NUKP has two major challenges: (1) The dataset masks the face information of the persons. (2) The image resolution is low, and the views change greatly. It can be observed that the performance is higher than the baseline after adding our CRE module and BSGA module into the network. In the meantime, the top-1 accuracy of our CRE outperforms all other methods.

### 4.4 Ablation Study

We use the image-based re-ID method [16] based on ResNet-50 as our baseline. In order to verify the validity of our modules, we implement the analysis on our method by combining different modules on PRCC and VC-Clothes.

Table 4: The ablation studies of our method on PRCC and VC-Clothes.

<table>
<thead>
<tr>
<th>method</th>
<th>PRCC</th>
<th>VC-Clothes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SC</td>
<td>CC</td>
</tr>
<tr>
<td>baseline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>99.5</td>
<td>98.4</td>
</tr>
<tr>
<td>✓</td>
<td>98.8</td>
<td>96.0</td>
</tr>
<tr>
<td>✓</td>
<td>99.8</td>
<td>98.4</td>
</tr>
<tr>
<td>✓</td>
<td>99.4</td>
<td>97.5</td>
</tr>
<tr>
<td>✓</td>
<td>99.6</td>
<td>97.3</td>
</tr>
<tr>
<td>✓</td>
<td>61.8</td>
<td>58.7</td>
</tr>
</tbody>
</table>

Table 4: The ablation studies of our method on PRCC and VC-Clothes.

<table>
<thead>
<tr>
<th>method</th>
<th>PRCC</th>
<th>VC-Clothes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SC</td>
<td>CC</td>
</tr>
<tr>
<td>baseline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>99.5</td>
<td>98.4</td>
</tr>
<tr>
<td>✓</td>
<td>98.8</td>
<td>96.0</td>
</tr>
<tr>
<td>✓</td>
<td>99.8</td>
<td>98.4</td>
</tr>
<tr>
<td>✓</td>
<td>99.4</td>
<td>97.5</td>
</tr>
<tr>
<td>✓</td>
<td>99.6</td>
<td>97.3</td>
</tr>
<tr>
<td>✓</td>
<td>61.8</td>
<td>58.7</td>
</tr>
</tbody>
</table>
Table 5: Comparison with existing attention modules in clothes-changing setting on PRCC and VC-Clothes.

<table>
<thead>
<tr>
<th>method</th>
<th>PRCC (CC)</th>
<th>VC-Clothes (CC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>top-1 mAP</td>
<td>top-1 mAP</td>
</tr>
<tr>
<td>CRE (ours)</td>
<td>59.7</td>
<td>83.5</td>
</tr>
<tr>
<td>+CBAM [48]</td>
<td>59.4</td>
<td>82.5</td>
</tr>
<tr>
<td>+SE [18]</td>
<td>58.7</td>
<td>84.1</td>
</tr>
<tr>
<td>+CIA [16]</td>
<td>60.9</td>
<td>83.9</td>
</tr>
<tr>
<td>+ECA [46]</td>
<td>59.5</td>
<td>82.4</td>
</tr>
<tr>
<td>+PSA [31]</td>
<td>58.4</td>
<td>82.7</td>
</tr>
<tr>
<td>+BSGA (ours)</td>
<td>61.8</td>
<td>84.5</td>
</tr>
</tbody>
</table>

Effectiveness of CRE. As shown in Table 4, we integrate only the CRE module to the baseline and achieve very high performance in both general setting and clothes-changing setting. In particular, in clothes-changing setting, it outperforms the baseline by a large margin of around 12 pp. and around 9 pp. on PRCC and VC-Clothes, respectively. This owes much to our CRE module, which eliminates clothes-relevant information and drives the model to learn more clothes-irrelevant features.

Effectiveness of BSGA. We only add the BSGA module and binary masks from CRE to the baseline and summarize the experimental results in the third row of Table 4. The performance of our BSGA is higher than the baseline in all three settings of both datasets. In detail, it surpasses the baseline in general setting by approximately 0.6 pp. and exceeds the baseline by 1 to 2 pp. in clothes-changing setting. Since the performance in same-clothes setting is almost 100%, the performance improvement is insignificant. Thanks to the body shape mask, background noise is suppressed to a certain extent. BSGA is an attention mechanism at the feature level, so its performance improvement is not as significant as CRE.

The existing attention modules are proposed to enhance important features and suppress unnecessary ones. However, there is no direct guidance for this process, making these methods easily produce unreliable attentions. Specifically for clothes-changing re-ID, our BSGA generates the attention maps guided by binary body shape masks, which could directly locate body parts and are more reliable. Some body information, such as hairstyle, body shape and limbs, can also be crucial discriminative features. We also make comparisons with some existing attention modules and our approach achieves the best performance (see Table 5).

Effectiveness of $\mathcal{L}_{BCTri}$. We make a comparison between Batch Hard mining triplet loss $\mathcal{L}_{Tri}$ and our Batch Cross-clothes mining triplet loss $\mathcal{L}_{BCTri}$. The experimental results are shown in Table 6. The performance of ours is higher than that using Batch Hard triplet loss for clothes-changing re-ID tasks. The principal reason is that our approach closes the distance of the samples with the same identity and different clothes, while Batch Hard mining only closes the distance between the anchor and the hardest positive sample. Batch Cross-clothes mining is beneficial for fully extracting clothes-irrelevant features.

We integrate the CRE module and the BSGA module into the network. As shown in the fifth row of Table 4, our method achieves 61.8% and 58.7% w.r.t. top-1 and mAP on PRCC in clothes-changing setting. Besides, it achieves 84.5% and 84.3% w.r.t. top-1 and mAP on VC-Clothes.

4.5 Visualization

For further analysis, we visualize the learned attention maps in the BSGA module in Figure 5. We can see from the rough outline of attention maps that these attention maps highlight
the human body parts and suppress background clutters.

To visually observe the final performance of our approach, we use class activation mapping [62] to visualize the heat maps of the input images, and the results are shown in Figure 6, from which we can see which part of an image contributes more to the final output of the model.

We can observe that: (1) The baseline method pays more attention to clothes-relevant features, such as color, texture, style, etc. (2) Our method pays more attention to the clothes-irrelevant parts, such as the face, feet, legs, arms, and key points of contour which are important discriminative features. Besides, our model is rarely affected by background noise. These heat maps make it more intuitive that our method learns richer clothes-irrelevant features, which are beneficial for the clothes-changing re-ID task.

5 Conclusion

In this paper, we propose a Clothes-Relevant information Erasure (CRE) module and a Body Shape-Guided Attention (BSGA) module for clothes-changing person re-id. We erase the clothes-relevant information from the original images so that the model can adaptively explore clothes-irrelevant cues. We further utilize the body shape mask to guide the learning of the attention map, which makes the model focus on richer and more discriminative features. Thorough experiments on three clothes-changing re-ID benchmarks demonstrate the performance advantages of our method.

6 Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62102180, U1713208, 62072242, the Natural Science Foundation of Jiangsu Province under Grant Nos. BK20210329. Note that the PCA Lab is associated with Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, Nanjing University of Science and Technology.
References


