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

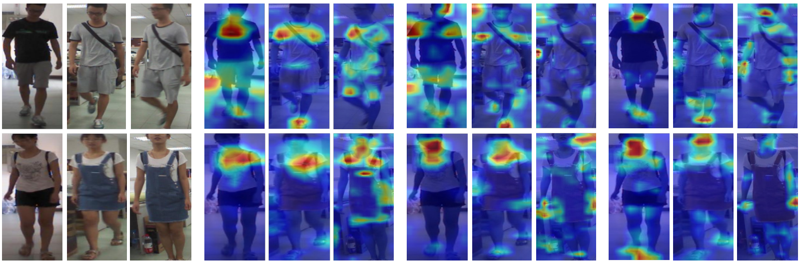
Abstract

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



(a) (b) (c) (d)

Figure 1: The visualizations of the activation maps learned by ResNet-50 in (b), our CRE module in (c) and our both CRE and BSGA modules in (d). (a) shows six original images of two persons. We can see that (b) always highlights some clothes-relevant features, (c) highlights head and limbs but occasionally some background noise, while (d) highlights more discriminative features, e.g., face, and body shape.

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 $\mathrm{X} \in \mathbb{R}^{H\times W\times 3}$, we obtain its semantic segmentation map $\mathrm{S} \in \mathbb{P}^{H\times W}(\mathbb{P} = \{0,1,2,3,4,5,6\})$

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 $f_e(\mathrm{X}, \mathrm{S})$ can be expressed as:

$$\mathrm{X}_{ijk} = 0, if : \mathrm{S}_{ij} = \{2,3,4\}, \tag{1}$$

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.

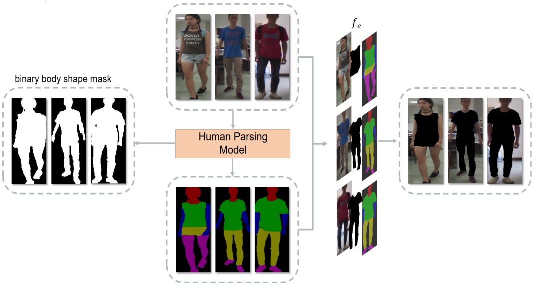


Figure 2: The structure of the Clothes-Relevant information Erasure(CRE) module.

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 $\mathrm{X}_{stage2} \in \mathbb{R}^{512\times h\times w}$. Then we utilize a dimension reduction operation to get the feature map $\mathrm{X}_f \in \mathbb{R}^{2\times h\times w}$. $\mathrm{M}_x \in \{0,1\}^{1\times H\times W}$ 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 $\mathrm{X}_c \in \mathbb{R}^{3\times h\times w}$. X_c can be calculated as follows:

$$\begin{gathered} \mathrm{X}_f = f^{7\times 7}(\mathrm{X}_{stage2}), \\ \mathrm{X}_c = [f^{7\times 7}(\mathrm{M}_x^{'}); \mathrm{X}_f], \end{gathered} \tag{2}$$

where $f^{7\times 7}(\cdot)$ represents a convolution operation with the filter size of 7×7. $\mathrm{M}_x^{'}$ denotes the body shape mask after downsampling. Then we need to use X_c to learn a simple attention map $\mathrm{A}_x \in \mathbb{R}^{1\times h\times w}$ in the following way:

$$\mathrm{A}_x = f^{7\times 7}(\mathrm{X}_c). \tag{3}$$

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:

$$\mathrm{A}_f = \phi_a(\mathrm{A}_x, \mathrm{M}_x^{'}) = \sigma(\mathrm{A}_x \odot f^{7\times 7}(\mathrm{M}_x^{'})), \tag{4}$$

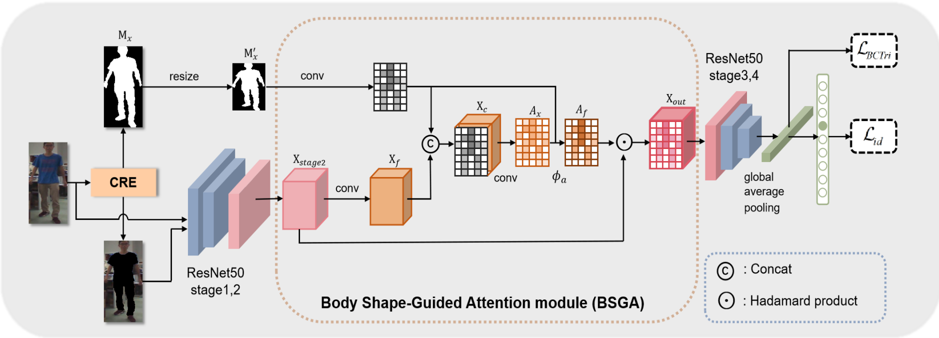


Figure 3: Illustration of our framework. The original image and the generated image obtained by the CRE module are fed into the ResNet-50 backbone. The feature map output after stage2 and the binary body shape mask obtained by CRE are sent into the BSGA module together. The feature map output by BSGA is sent into the later stage for feature extraction. Finally, we calculate the loss function with the output feature vector of the network.

where $\odot$ denotes the Hadamard product, σ 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_{stage2} to get the output feature map X_{out}. This process can be described as follow:

$$\mathrm{X}_{out} = \mathrm{X}_{stage2} \odot \mathrm{A}_f. \tag{5}$$

We send the weighted feature map X_{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:

$$\mathcal{L}_{id} = -\frac{1}{2M}\sum_{i}^{2M}\log(p(y_i|x_i)). \tag{6}$$

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,

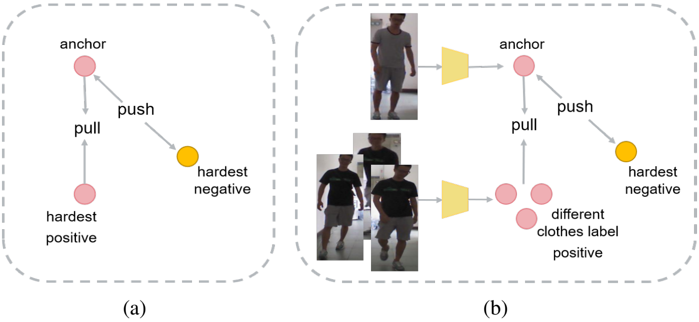


Figure 4: Two triplet mining methods. (a) is Batch Hard mining. A triplet contains an anchor, the hardest positive sample and the hardest negative sample. (b) is the Batch Cross-clothes mining we propose in this paper. A triplet contains an anchor, a positive sample of different clothes label and the 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^n 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:

$$\mathcal{L}_{BCTri} = \sum_{n=1}^{N} \max(m + d(x_i, x_j^n) - d(x_i, x_k), 0), \tag{7}$$

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:

$$\mathcal{L} = \mathcal{L}_{id} + \lambda \mathcal{L}_{BCTri}, \tag{8}$$

where λ 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 [43] 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 [54] 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 [45] 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 [12] which is initialized with ImageNet [4] 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 256×128. m in $\mathcal{L}_{BCTri}$ is set to 0.3. Random horizontal flipping, random cropping, and random erasing [61] 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 [22] 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 1: Comparison with state-of-the-art methods on VC-Clothes.

method	ref	general(all cams)		SC(cam2&cam3)		CC(cam3&cam4)	
		top-1	mAP	top-1	mAP	top-1	mAP
MDLA [36]	ICCV2017	88.9	76.8	93.4	93.9	59.2	60.8
PCB [41]	ECCV2018	87.7	74.6	94.7	94.3	62.0	62.2
Part-aligned [39]	ECCV2018	90.5	79.7	93.9	93.4	69.4	67.3
TransReID [13]	ICCV2021	90.5	80.1	95.1	94.5	70.0	71.8
FSAM [15]	CVPR2021	-	-	94.7	94.8	78.6	78.9
3DSL [2]	CVPR2021	-	-	-	-	79.9	81.2
PS [58]	SPL2021	93.1	84.9	94.7	92.9	82.4	80.3
CAL [11]	CVPR2022	92.9	87.2	**95.1**	**95.3**	81.4	81.7
ours w/o BSGA		94.2	85.7	94.9	93.7	83.5	81.7
ours		**94.4**	**88.2**	94.9	94.4	**84.5**	**84.3**

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.

method	ref	SC		CC	
		top-1	mAP	top-1	mAP
HACNN [27]	CVPR2018	82.5	-	21.8	-
PCB [41]	ECCV2018	99.8	97.0	41.8	38.7
IANet [16]	CVPR2019	99.4	98.3	46.3	45.9
TransReID [13]	ICCV2021	97.3	95.9	47.1	49.3
SPT+ASE [54]	TPAMI2019	64.2	-	34.4	-
GI-ReID [21]	CVPR2022	80.0	-	33.3	-
RCSANet [19]	ICCV2021	100	97.2	50.2	48.6
3DSL [2]	CVPR2021	-	-	51.3	-
FSAM [15]	CVPR2021	98.8	-	54.5	-
PS [38]	SPL2021	99.2	96.6	61.1	58.3
CAL [11]	CVPR2022	**100**	**99.8**	55.2	55.8
ours w/o BSGA		98.8	96.0	59.7	56.8
ours		99.6	97.3	**61.8**	**58.7**

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

method	ref	top-1	top-5	top-10
se-resnext [18]	CVPR2018	16.7	25.5	31.2
senet [18]	CVPR2018	18.2	25.2	30.6
PCB [41]	ECCV2018	16.9	25.6	30.6
MGN [44]	ACM MM2018	18.8	28.8	33.0
TransReID [13]	ICCV2021	21.2	32.4	37.6
PS [38]	SPL2021	18.8	**35.2**	**45.5**
LSD [53]	IVC2021	16.4	27.9	34.8
baseline		20.6	29.1	33.3
baseline+BSGA		22.1	30.9	36.7
baseline+CRE		**23.0**	33.6	40.0

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.

method				PRCC				VC-Clothes			
				SC		CC		general		CC	
baseline	*CRE*	*BSGA*	$\mathcal{L}_{BCTri}$	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
✓			✓	99.5	98.4	47.2	44.9	91.6	80.9	74.7	73.4
✓	✓		✓	98.8	96.0	59.7	56.8	94.2	85.7	83.5	81.7
✓		✓	✓	**99.8**	**98.4**	48.4	46.7	92.1	81.5	76.1	75.0
✓	✓	✓		99.4	97.5	60.0	56.3	93.6	85.4	80.2	79.7
✓	✓	✓	✓	99.6	97.3	**61.8**	**58.7**	**94.4**	**88.2**	**84.5**	**84.3**

Table 5: Comparison with existing attention modules in clothes-changing setting on PRCC and VC-Clothes.

method	PRCC (CC)		VC-Clothes (CC)	
	top-1	mAP	top-1	mAP
CRE (ours)	59.7	56.8	83.5	81.7
+CBAM [48]	59.4	57.6	82.5	82.3
+SE [18]	58.7	57.4	84.1	82.1
+CIA [16]	60.9	57.8	83.9	82.7
+ECA [46]	59.5	55.8	82.4	81.1
+PSA [31]	58.4	55.2	82.7	82.6
+BSGA (ours)	**61.8**	**58.7**	**84.5**	**84.3**

Table 6: Analytical experiment results in clothes-changing setting on PRCC. $\mathcal{L}_{Tri}$ is Batch Hard mining triplet loss [14] which is widely used in Re-ID tasks.

ours	$\mathcal{L}_{Tri}$	$\mathcal{L}_{BCTri}$	top-1	top-5	top-10	mAP
✓			60.0	66.8	70.1	56.3
✓	✓		61.1	66.6	68.6	56.1
✓		✓	**61.8**	**67.9**	**70.7**	**58.7**

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.

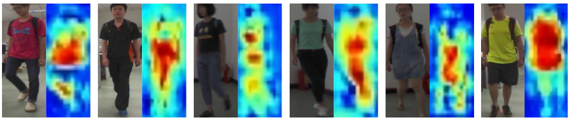


Figure 5: Visualization of the original images and learned attention maps.

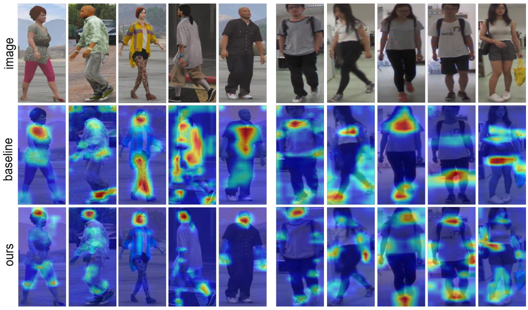


Figure 6: The visualization of heat maps on PRCC (right) and VC-Clothes (left). The highlights are the parts that the model pays more attention to.

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

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3908–3916, 2015.

[2] Jiaxing Chen, Xinyang Jiang, Fudong Wang, Jun Zhang, Feng Zheng, Xing Sun, and Wei-Shi Zheng. Learning 3d shape feature for texture-insensitive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8146–8155, 2021.

[3] 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 *Proceedings of the iEEE conference on computer vision and pattern recognition*, pages 1335–1344, 2016.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[5] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015.

[6] Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, and Dina Katabi. Learning longterm representations for person re-identification using radio signals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10699–10709, 2020.

[7] Xing Fan, Wei Jiang, Hao Luo, and Mengjuan Fei. Spherereid: Deep hypersphere manifold embedding for person re-identification. *Journal of Visual Communication and Image Representation*, 60:51–58, 2019.

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2360–2367, 2010. doi: 10.1109/CVPR.2010.5539926.

[9] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008.

[10] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *European Conference on Computer Vision*, pages 228–243. Springer, 2020.

[11] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1060–1069, 2022.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15013–15022, October 2021.

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[15] Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10513–10522, 2021.

[16] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9317–9326, 2019.

[17] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Feature completion for occluded person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[19] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing status awareness for long-term person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11895–11904, 2021.

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[21] Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14278–14287, 2022.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2288–2295. IEEE, 2012.

[24] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[25] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[26] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[27] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018.

[28] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2019.

[29] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015.

[30] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7):3492–3506, 2017.

[31] Huajun Liu, Fuqiang Liu, Xinyi Fan, and Dong Huang. Polarized self-attention: towards high-quality pixel-wise regression. *arXiv preprint arXiv:2107.00782*, 2021.

[32] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[33] Lianyang Ma, Xiaokang Yang, and Dacheng Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE transactions on image processing*, 23(8):3656–3670, 2014.

[34] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1363–1372, 2016.

[35] Bryan James Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, Q Mary, et al. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010.

[36] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5399–5408, 2017.

[37] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[38] Xiujun Shu, Ge Li, Xiao Wang, Weijian Ruan, and Qi Tian. Semantic-guided pixel sampling for cloth-changing person re-identification. *IEEE Signal Processing Letters*, 28:1365–1369, 2021.

[39] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 402–419, 2018.

[40] 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 *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018.

[41] 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 *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018.

[42] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *European conference on computer vision*, pages 135–153. Springer, 2016.

[43] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 830–831, 2020.

[44] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018.

[45] Kai Wang, Zhi Ma, Shiyan Chen, Jinni Yang, Keke Zhou, and Tao Li. A benchmark for clothes variation in person re-identification. *International Journal of Intelligent Systems*, 35(12):1881–1898, 2020.

[46] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Supplementary material for 'eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Seattle, WA, USA*, pages 13–19, 2020.

[47] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

[48] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[49] Wangmeng Xiang, Jianqiang Huang, Xianbiao Qi, Xiansheng Hua, and Lei Zhang. Homocentric hypersphere feature embedding for person re-identification. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1237–1241. IEEE, 2019.

[50] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1249–1258, 2016.

[51] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *European conference on computer vision*, pages 1–16. Springer, 2014.

[52] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2119–2128, 2018.

[53] Ehsan Yaghoubi, Diana Borza, Bruno Degardin, and Hugo Proença. You look so different! haven't i seen you a long time ago? *Image and Vision Computing*, 115:104288, 2021.

[54] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2029–2046, 2019.

[55] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

[56] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *2014 22nd international conference on pattern recognition*, pages 34–39. IEEE, 2014.

[57] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[58] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European conference on computer vision*, pages 868–884. Springer, 2016.

[59] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1367–1376, 2017.

[60] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*, pages 3754–3762, 2017.

[61] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

[62] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.