Clothing Status Awareness for Long-Term Person Re-Identification

Yan Huang[1,3] Qiang Wu[1] JingSong Xu[1] Yi Zhong[2] ZhaoXiang Zhang[3]
[1] School of Electrical and Data Engineering, University of Technology Sydney, Australia
[2] School of Information and Electronics, Beijing Institute of Technology, China
[3] Institute of Automation, Chinese Academy of Sciences, Beijing China.
huangyan.750@outlook.com, {Qiang.Wu,JingSong.Xu}@uts.edu.au, yi.zhong@bit.edu.cn

Abstract

Long-Term person re-identification (LT-reID) exposes extreme challenges because of the longer time gaps between two recording footages where a person is likely to change clothing. There are two types of approaches for LT-reID: biometrics-based approach and data adaptation based approach. The former one is to seek clothing irrelevant biometric features. However, seeking high quality biometric feature is the main concern. The latter one adopts fine-tuning strategy by using data with significant clothing change. However, the performance is compromised when it is applied to cases without clothing change. This work argues that these approaches in fact are not aware of clothing status (i.e., change or no-change) of a pedestrian. Instead, they blindly assume all footages of a pedestrian have different clothes. To tackle this issue, a Regularization via Clothing Status Awareness Network (RCSANet) is proposed to regularize descriptions of a pedestrian by embedding the clothing status awareness. Consequently, the description can be enhanced to maintain the best ID discriminative feature while improving its robustness to real-world LT-reID where both clothing-change case and no-clothing-change case exist. Experiments show that RCSANet performs reasonably well on three LT-reID datasets.

1. Introduction

Person re-Identification (re-ID) is to confirm the identity of a person in visual traces. Based on the scale of time gaps when footages are captured, there are two different scenarios for person re-ID: 1) Short-Term re-ID (ST-reID) and 2) Long-Term re-ID (LT-reID). The first scenario normally addresses the time gap of several minutes. In this case, we can safely assume that a person does not change clothing. This scenario is investigated first by the research community and produces many state-of-the-art works with encouraging performance [15, 41, 39, 45, 12, 10, 11]. In recent years, more researches have focused on LT-reID. In LT-reID, the time gap between two footages can be several days or even longer. Therefore, besides significant challenges of non-human factors (*i.e.*, image resolution, illumination, *etc.*) which can be observed in ST-reID, there are significant challenges caused by human factor (*i.e.*, clothing and dressing accessory) in LT-reID. Clothing change in particular is very common in LT-reID, although it may not necessarily change in some circumstances. Such changes dramatically constrain the performance of methods for ST-reID when they are applied to LT-reID directly.

In order to tackle the challenge shown in LT-reID, various approaches are reported although they are still on a preliminary level. These approaches can be categorized into two types: 1) **biometrics-based approach** and 2) **data adaptation based approach.**

Biometrics-based approaches strive to avoid clothing-related information that is deemed not stable in LT-reID. One type of these methods [36, 32, 37, 26, 21, 34, 17] explore biometric features for LT-reID such as motion, body contour/shape, and face. Theoretically, the biometric feature should be robust. However, it heavily relies on high-quality footages. For example, to obtain people motion features, it needs to successfully extract the human body from cluttered backgrounds and to track the human body throughout the entire video period. Due to limitations of image segmentation, tracking, and body part occlusion, it is implausible to guarantee the reliability of motion feature from footage [36]. Another type of biometrics-based approach introduces additional depth information using RGB-D camera. The depth information does provide another source of helpful information (*e.g.*, 3D biometric information) for LT-reID [1, 19, 6, 30]. However, it introduces extra complexity to the camera setup. In addition, due to its limited sensing distance, it is still far from real practice.

Instead of using biometric information, data adaptation approaches attempt to use fine-tuning mechanisms [9, 13] by elevating the performance of models pretrained on a ST-reID dataset (*e.g.*, Market-1501 [39]) before training LT-reID data with diverse clothing-change cases. Through

deliberately designed network architectures and loss functions, a model is able to tolerate clothing change to a certain extent. This method also has certain limitations. It expects a pretrained model to be adjusted on parameters by using clothing-change data. However, it does not explicitly consider the actual clothing status (*i.e.*, change or no-change) of each individual during training. That is, the method always tends to learn clothing irrelevant ID features from a certain amount of clothing-change data. Therefore, the core problem in the data adaptation approach is that it simply feeds the input footage into a complex model. However, due to the lack of clothing status awareness, methods may cause sub-optimal performance when they handle the no-clothing-change case. The essence of a LT-reID method should correctly sense the clothing status of each individual in order to dynamically regularize ID features. These regularized ID features should be able to tackle clothing-change cases if any without sacrificing their discrimination ability when no-clothing-change cases also exist in LT-reID.

In light of the above discussion, this paper proposes a Regularization via Clothing Status Awareness Network (RCSANet). RCSANet decouples ID discriminative feature learning and clothing status awareness learning into two separate processes in the early part of the network. In the later part of RCSANet, ID features can be regularized through a proposed Feature Regularization Module (FRM) in order to encourage them are more consistent when a person wears the same clothes. Such a regularization process is achieved by embedding the ability of clothing status awareness into FRM. In this way, RCSANet does not sacrifice performance on no-clothing-change cases while still maintaining its performance on clothing-change cases. It is noteworthy that the proposed RCSANet only requires ID labels throughout the entire training process. It does not require extra clothing type annotations.

Contributions of this paper can be summarized in threefold: 1) Unlike existing biometrics-based approaches and data adaptation based approaches, this paper proposes a novel clothing status aware LT-reID solution. 2) The proposed RCSANet explicitly builds up a clothing status awareness learning process, which is used to enhance the robustness of ID features for handling both clothing-change case and no-clothing-change case in LT-reID. 3) Extensive experiments are conducted to demonstrate the effectiveness of our RCSANet on three LT-reID benchmarks where both clothing-change case and no-clothing-change case exist.

2. Related Works

2.1. Biometrics-based Approach

Biometric traits have been studied for LT-reID, including motion [36], body contour/shape [32, 21, 17], and face [34, 26]. **1) Motion.** Zhang *et al.* [36] extract motion features to associate different persons for LT-reID. However, to extract robust motion features, a complete motion cycle is normally required, making it difficult to be applied to image-based LT-reID scenario. **2) Body Contour/Shape.** Yang *et al.* [32] use body contours for LT-reID. To achieve this, a person only can change her/his clothing moderately with a similar thickness, which is confined to a limited LT-reID application scenario. Qian *et al.* [21] employ a pose detector to localize body joints which are used for learning body shape features based on spatial relationships between joints. However, the body shape is sensitive towards shooting angles of cameras, which may not be effective in the real world. **3) Face.** Wan *et al.* [26] and Yu *et al.* [34] learn face features for LT-reID. The face only takes up a small part of the body region, and it is not always available when the image quality or camera view (*e.g.*, back view) is poor.

RGB-D images provided by depth camera (*e.g.*, Kinect) are used in existing LT-reID works for biometric features extraction. Barbosa *et al.* [1] propose extracting 3D soft-biometric features for LT-reID using RGB-D images provided by Kinect. Munaro *et al.* [19] transform point clouds of persons to a standard pose. The transformed point clouds are used for composing 3D models for LT-reID to eliminate impacts caused by clothing change. Haque *et al.* [6] leverage raw depth video data as training inputs, and propose a recurrent attention model that re-identifies persons by focusing on small, discriminative body regions to tackle clothing change. However, depth sensor is hard to be widely deployed for real-world LT-reID due to the complexity of camera setup and the limitation on sensing distance.

2.2. Data Adaptation Based Approach

Huang *et al.* [9] introduce a two-step fine-tuning framework to explore ID discriminative features for LT-reID. The model is pretrained on a ST-reID dataset and then fine-tuned on LT-reID dataset with diverse clothing-change cases. Huang *et al.* [13] propose a ReIDCaps network to learn ID discriminative features by employing capsule layers [23]. Similar to [9], the backbone of ReIDCaps is also pretrained on a ST-reID dataset. Both methods more or less attempt to normalize features learned from a ST-reID dataset to mitigate impacts caused by clothing change through a fine-tuning mechanism. However, these methods actually do not explicitly learn the clothing status awareness in LT-reID. Existing LT-reID datasets (*e.g.*, CelebreID [13] and PRCC [32]) do have certain amount of cases where there is no clothing change in both training and test sets. However, they are just regarded as disturbances during model training since the loss term and network architecture are deliberately designed for clothing-change cases. To tackle such an issue, our RCSANet introduces clothing status awareness learning to enhance the learned ID discriminantive features by a regularization strategy for LT-reID.

To the best of our knowledge, ReIDCaps [13] made a preliminary attempt to learn the clothing status awareness by using vector capsule neurons [23]. This method makes an implicit assumption that the orientation of vector neurons (with 24-dim) in the classification layer should be able to automatically sense the clothing status of each individual. However, the original definition of capsule neuron holds a view that the orientation of capsule is to represent different types of properties (*e.g.*, pose, deformation, texture, *etc.*) [23]. Without using any explicit constraint, the implicit assumption (*i.e.*, automatically sense clothing status) in ReIDCaps [13] is hard to be satisfied since clothing status just belongs to one of the properties. Unlike [13], our RCSANet explicitly builds up the clothing status awareness learning without using complex capsule neurons.

3. Method

As illustrated in Fig. 1, RCSANet involves an Inter-Class Enforcement (ICE) stream and an Intra-Class appearance Regularization (ICR) stream. **As a minor contribution**, the ICE stream is a well-designed baseline model for LT-reID by maximizing inter-class differences to learn ID discriminative features. **As the major contribution**, the ICR stream is used to regularize features learned by ICE through clothing status awareness for each individual to encourage the ID features can be used for both clothing-change and no-clothing cases in real-world LT-reID.

3.1. The ICE stream

As illustrated in Fig. 1, during training, given N IDs and K images for each ID in a mini-batch, the ICE stream learns ID discriminative features for each image. Our ICE stream is based on the classic ID-discriminative Embedding (IDE) [40] network. That is, a CNN backbone followed by an identification loss (L_{id}) to separate inter-class embeddings. On top of IDE we introduce a Mixed Pooling Module (MPM) to learn both extreme and smooth ID features from the output of backbone. In addition, a clothing insensitive triplet loss (L_{cit}) is introduced to reduce impacts caused by intra-class clothing change.

MPM. As shown in Fig. 1, an image I_n^k (k-th image belongs to n-th ID) is fed into the backbone of ICE to obtain the output feature map which will be fed into MPM. In MPM, we first equally partition feature maps to two parts horizontally (*i.e.*, upper and lower) since clothes of upper and lower body parts are normally inconsistent. Then, both Global Max Pooling (GMP) and Global Average Pooling (GAP) are conducted on every part. The ID feature is achieved by a concatenation operation as follows:

$$f_{id}(I_n^k) = \Big[f_{up}^a(I_n^k), f_{up}^m(I_n^k), f_{low}^a(I_n^k), f_{low}^m(I_n^k)\Big]^T, \quad (1)$$

where a is short for GAP and m is short for GMP (*e.g.*, $f_{up}^a(I_n^k)$=GAP($I_n^k[:, H/2:, :]$), $f_{low}^m(I_n^k)$=GMP($I_n^k[:, :H/2, :]$)). The procedure of our MPM is nonparametric.

The rationale behind utilization of MPM. In LT-reID, the surveillance area normally covers a large scale [13]. GMP adopted in MPM is able to extract extreme features on useful body regions but also on some unique but useless background information (*e.g.*, a yellow bin appears in quite a few areas). In contrast, GAP adopted in MPM learns more smooth features. However, it takes all information (body region + background) into account and results in an average value which may be trivial and not suitable to person re-ID matching. The combination of them demonstrate a better performance in our experiments (refer to Sec. 4.6).

Identification Loss. To separate inter-class embeddings, a classifier (from FC1 to FC2, refer to Fig. 1) is used in ICE followed by L_{id}:

$$L_{id} = \mathbb{E}\Big[- log\big(p(n|I_n^k)\big)\Big] \quad (2)$$

where $p(n|I_n^k))$ is the predicted probability of I_n^k belonging to ID n.

Clothing Insensitive Triplet Loss. L_{cit} is adopted to reduce the impact caused by intra-class clothing change. That is, to minimize the gap amongst features of the same person even the clothing is changed dramatically. In the meantime, L_{cit} maximizes the gap amongst features of different persons even the clothing is similar. Let I_α, I_ρ, I_η be a mined hard triplet within a training mini-batch. That is, I_α and I_ρ (I_η) are images belonging to the same (different) ID but with the furthest (closest) distance. Here, I_α represents I_n^k for convenience. We use the Weighted Regularization Triplet (WRT) [28, 33] loss for this task because WRT does not need extra hyperparameters to control the margin of positive and negative pairs:

$$\begin{gathered} L_{cit} = log\big(1 + e^{w_\alpha^\rho \cdot d_{\alpha\rho} - w_\alpha^\eta \cdot d_{\alpha\eta}}\big), \\ w_\alpha^\rho = \frac{e^{d_{\alpha\rho}}}{\sum_{\rho_1,\rho_2 \in \mathbb{P}} e^{d_{\rho_1\rho_2}}}, w_\alpha^\eta = \frac{e^{d_{\alpha\eta}}}{\sum_{\eta_1,\eta_2 \in \mathbb{N}} e^{d_{\eta_1\eta_2}}} \end{gathered} \quad (3)$$

where d represents Euclidean distance, $\mathbb{P}$ and $\mathbb{N}$ are the corresponding positive set and negative set to I_α in a mini-batch, respectively.

3.2. The ICR stream

Since existing LT-reID methods handle input images regardless of the clothing status, the ICR stream is proposed to tackle this issue. ICR receives images to extract their appearance features (*i.e.*, f_{ap}) which are used to regularize f_{id} learned from ICE. As illustrated in Fig. 1, in ICR, layers before the output of MPM are entirely the same as ICE. The difference is that in order to extract f_{ap}, the CNN backbone used in ICR is pretrained using a ST-reID dataset. With the pre-training, significant dressing information can be learned by f_{ap} since clothing is the most important feature to differentiate persons in ST-reID. In addition, there

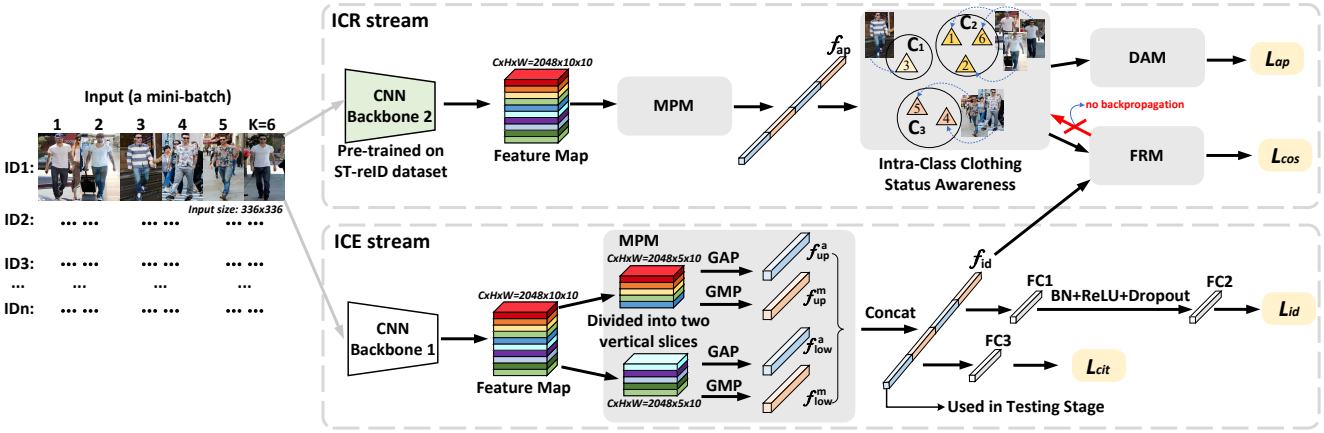


Figure 1. Overview of RCSANet. MPM, FRM, and DAM are marked in gray. Loss functions are marked in yellow. FC and BN are short for fully-connected layer and batch normalization layer, respectively. In ICR, the triangle represents f_{ap} of images.

are two modules involved in ICR. A Feature Regularization Module (FRM) (see Sec. 3.2.1) is proposed to regularize f_{id} to be more consistent when a person wears the same clothes. A Domain Adaptation Module (DAM) (see Sec. 3.2.2) that reduces the domain gap between parameters pretrained on ST-reID data and LT-reID data.

3.2.1 FRM

During training, FRM is proposed to regularize $f_{id}(I_n^k)$ (see Eq. 1) by embedding clothing status awareness. To this end, ICR extracts f_{ap} for the K input images belonging to an ID n in a mini-batch at each iteration. Then, an unsupervised clustering method is employed to separate the K f_{ap} into N_C clusters (*e.g.*, C_1, C_2, and C_3 when N_C=3, see Fig. 1). The clustering result is used to sense clothing status (change or no-change) for a person. For instance, samples belonging to the same cluster are regarded as similar clothing (no-change). Conversely, samples belonging to different clusters are regarded as clothing changes. Two similar clothes may be confused with each other due to the limited capability of unsupervised clustering. However, such an approach can still sense clothing status as long as two clothes are unnecessarily exactly the same. In order to regularize $f_{id}(I_n^k)$, we dynamically calculate $f_{ap}^{sim}(I_n^k)$ and $f_{ap}^{dis}(I_n^k)$ of $f_{id}(I_n^k)$ in each training iteration according to clustering results. f_{ap}^{sim} (f_{ap}^{dis}) is used to minimize (maximize) the similarity between $f_{id}(I_n^k)$ and its similar (dissimilar) clothing, which are respectively calculated as follows:

$f_{ap}^{sim}(I_n^k)$ is the cluster centroid where $f_{ap}(I_n^k)$ belongs to. For instance, if $f_{ap}(I_n^k)$ belongs to cluster C_i:

$$f_{ap}^{sim}(I_n^k) = \overline{\sum f_{ap}(I_n^{k'}) \in C_i}, \tag{4}$$

where $k' \in [1, K]$, for K images belonging to this ID. If the cluster C_i only contains $f_{ap}(I_n^k)$, $f_{ap}^{sim}(I_n^k)$=$f_{ap}(I_n^k)$.

$f_{ap}^{dis}(I_n^k)$ is calculated by searching appearance feature which is furthest from $f_{ap}(I_n^k)$ in the K images belonging to the same ID n:

$$\begin{aligned} f_{ap}^{dis}(I_n^k) &= f_{ap}(I_n^{k^*}), \\ k^* &= \underset{k' \in [1,K]}{argmax} \Big[d\big(f_{ap}(I_n^k), f_{ap}(I_n^{k'})\big) \Big], \end{aligned} \tag{5}$$

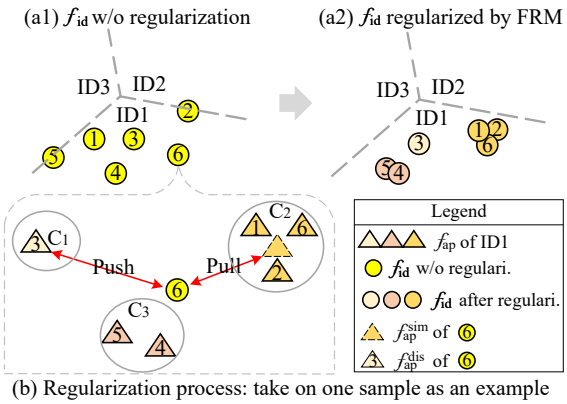
where $d(,)$ is used to measure feature similarity (Euclidean distance is adopted). Given k-th image, k^* is the image index when the distance between $f_{ap}(I_n^k)$ (feature of k-th image) and $f_{ap}(I_n^{k'})$ (features of k'-th image where $k' \in [1, K]$ could be any one of the K images) is the largest.

Finally, $f_{ap}^{sim}(I_n^k)$ and $f_{ap}^{dis}(I_n^k)$ are used to regularize $f_{id}(I_n^k)$ by minimizing the following cosine loss (L_{cos}):

$$\begin{aligned} L_{cos}^{sim} &= 1 - cos\Big[f_{id}(I_n^k), f_{ap}^{sim}(I_n^k)\Big], \\ L_{cos}^{dis} &= max\Big[0, cos\big(f_{id}(I_n^k), f_{ap}^{dis}(I_n^k)\big) - \xi\Big], \\ L_{cos} &= L_{cos}^{sim} + L_{cos}^{dis}, \end{aligned} \tag{6}$$

where $\xi \in [-1, 1]$ is a margin parameter. Given $f_{id}(I_n^k)$ learned by ICE, L_{cos} is used to regularize $f_{id}(I_n^k)$. In this process, L_{cos}^{sim} pulls $f_{id}(I_n^k)$ close to $f_{ap}^{sim}(I_n^k)$, while L_{cos}^{dis} enforces $f_{id}(I_n^k)$ away from $f_{ap}^{dis}(I_n^k)$. As shown in Fig. 2, a sample (*e.g.*, $f_{id}(I_n^6)$) is pulled towards its cluster centroid (*i.e.*, $f_{ap}^{sim}(I_n^6)$) by minimizing L_{cos}^{sim} to encourage $f_{id}(I_n^6)$ remaining in the cluster with the same/similar clothing. Conversely, $f_{id}(I_n^6)$ is pushed away from the sample of the most dissimilar clothing (*i.e.*, $f_{ap}(I_n^3)$) by minimizing L_{cos}^{dis}. By minimizing both losses, the network can be aware of the clothing change of a subject.

Fig. 2 intuitively illustrates how the regularization process works and why it is useful to LT-reID when both clothing-change and no-clothing-change cases exist. As



(b) Regularization process: take on one sample as an example

Figure 2. A schematic overview of FRM. (a1) and (a2) illustrate f_{id} of ID1. Dotted lines represent boundaries between different IDs. (b) illustrates how FRM regularizes $f_{id}(I_n^6)$ with f_{ap} belonging to ID1. Number 1-6 represent the corresponding relationship of feature points amongst (a1), (a2), and (b).

shown in Fig. 2 (a1) that, without regularization, some samples may lie between two different IDs since they are almost indistinguishable. This is because clothing information that is significant for person re-ID is no longer reliable in LT-reID. These information are easily overwhelmed when f_{id} is learned under certain constraints (*e.g.*, L_{cit} in ICE) and with diverse clothing-change data. Moreover, in Fig. 2 (a1), it can see that regardless of clothing change or not, features spread in the space. It does not guarantee the same description when two images of the same clothing are fed into LT-reID model. Therefore, in our FRM, clothing information (*i.e.*, f_{ap}) is re-picked up and used to regularize f_{id} by embedding the clothing status awareness for each individual (see Fig. 2 (b)). For images belonging to the same person ID, the clothing status awareness is obtained by using an unsupervised clustering strategy. After regularization, samples that are indistinguishable can be gradually 'hauled back' to their affiliated ID during training (see Fig. 2 (a2)). Such a regularization process enforces consistency of f_{id} when a person wears the same clothing. Finally, by combining ICE and ICR, f_{id} can best differentiate person IDs when both clothing-change case (mainly contributed by ICE) and no-clothing-change case (mainly contributed by ICR) exist.

3.2.2 DAM

After ICR pre-training (see Sec. 3.2), all parameters in ICR are frozen when ICR and ICE are trained together. That is, parameters in ICR are not updated by minimizing L_{cos} (refer to Fig. 1). Therefore, for each image, its f_{ap} does not change during the entire training stage. However, due to the domain gap between the pre-trained data and LT-reID data, f_{ap} extracted from LT-reID data may not be robust enough. Inspired by the feature-based weight initialization [38], we add a DAM module in ICR to reduce this domain gap. There is only one fully-connected layer with N_C neurons in DAM. N_C is the number of clusters for the K images belonging to n-th ID. An intra-class appearance classification loss (L_{ap}) is employed after the fully-connected layer to classify f_{ap} according to the clustering result:

$$L_{ap} = -log\frac{e^{W_{C_j}^T \cdot f_{ap}(I_n^k)}}{\sum_{i=1}^{N_C} e^{W_{C_i}^T \cdot f_{ap}(I_n^k)}}, \tag{7}$$

where C_j is the cluster label of image I_n^k. $W \in \mathbb{R}^{d_f \times N_C}$ represents parameters of the fully-connected layer in DAM. $W_C \in \mathbb{R}^{N_C}$ is the C-th column of W. d_f is the feature dimension of f_{ap}. Following [38], we use the mean feature of each cluster to initialize W to ensure the convergence of network training. Without DAM, the ability of extracting f_{ap} for LT-reID images is learned from ST-reID data only. Using f_{ap}, the clustering result may be compromised if the domain gap is large (can be caused by large variations of background or illumination). DAM enforces the ICR stream to distinguish different types of clothing from LT-reID images by using the clustering result during training. By doing so, the parameters of ICR stream can be updated. Consequently, the clustering result can be strengthened.

3.3. Optimization

In training, the ICE and ICR streams are jointly optimized. The total objective can be formulated as a weighted sum of following losses:

$$L_{total} = L_{id} + L_{cit} + \lambda_{cos} \cdot L_{cos} + L_{ap}, \tag{8}$$

where λ_{cos} is weight to control the importance of L_{cos}. Since L_{cos} is used for feature regularization, we set a small weight λ_{cos}=0.1 to ensure the learning of f_{id} is stable. L_{ap} is not involved until $0.5\times N_{iter}$ training iterations, where N_{iter} is the total number of iterations. That is, before the training is half done, all parameters of backbone in ICR are frozen (*i.e.*, DAM is not involved).

4. Experiments

4.1. Implementation Details

Network Details. We implement our method with Pytorch [20]. Following the previous LT-reID approach [13], the ImageNet-trained [22] DenseNet-121 [8] is used as the backbone in ICE and ICR (without parameters sharing). As in [9, 13], the backbone of our ICR is pretrained on the ST-reID dataset Market-1501 [39] to learn the ability of extracting appearance/clothing features (*i.e.*, f_{ap}) for LT-reID data. We respectively set the number of neurons in FC1, FC2, and FC3 layers (refer to Fig. 1) to 512, N_{id}, and 2048, where N_{id} is the number of training ID. The non-parametric

DBSCAN [5] algorithm is used for the clustering task in ICR since it does not require one to specify the number of clusters. We set the minimum number of samples for each cluster to 1 in DBSCAN.

Training Details. The SGD is used for training with momentum 0.9 and weight decay 5e-4. In ICE, we initially set the learning rate to 1e-2 for FC1 and FC2, and 1e-3 for the backbone and FC3. In the beginning, all parameters in ICR are frozen until training is half done (refer to Sec. 3.2.2). When ICR starts to update, we set the learning rate of backbone and DAM to 1e-5 and 1e-4, respectively. For each training mini-batch, we set N=12, K=6. The training is finished after 72 epochs. All input images are resized to 336×336 with random horizontal flipping.

4.2. Datasets for Evaluations

Three LT-reID datasets have been released and available for evaluation, including Celeb-reID [13], Celeb-reID-light [9], and PRCC [32]: **Celeb-reID** [13] uses street snapshots of celebrities acquired from the Internet. There are 34,186 images with 1,052 IDs. The probability of clothing change for a person is 70% on average. That is, both clothing-change case (70%) and no-clothing-change case (30%) coexist in the training and test sets of Celeb-reID. **Celeb-reID-light** [9] is a light version of Celeb-reID. The difference is that, both training and testing sets of Celeb-reID-light only contain clothing-change cases (*i.e.*, a person will not wear the same clothing twice), making it a pure clothing-change LT-reID dataset. **PRCC** [32] is acquired under three camera views. This dataset is specifically built for body contour feature extraction to overcome the change of clothing by enforcing a person to wear clothes of similar thickness. For each person, images acquired by camera A and B (C) are without (with) clothing change. Following [32], clothing-change evaluation and no-clothing-change evaluation are respectively conducted on PRCC.

4.3. Inference and Evaluation Criterion

In testing, the ICR stream is not involved since f_{ap} extracted from ICR is regarded as an auxiliary function to improve the discriminative ability of f_{id} extracted from ICE. Therefore, only f_{id} (refer to Eq. 1) extracted from ICE is used as the final person description for inference. Following existing LT-reID works [13, 32], both rank-n accuracy and mean Average Precision (mAP) are reported.

4.4. Result on Celeb-reID and Celeb-reID-light

We compare our RCSANet with seven ST-reID methods and two LT-reID methods on Celeb-reID and Celeb-reID-light. Results are shown in Tab. 1. So far, only [9, 13] are LT-reID methods that report performance on two datasets. ReIDCaps^{fg} [13] means using *fine-grained* body parts to train and test the model (refer to [13]). In addition to directly use the whole image (*e.g.*, RCSANet), we also report the result of our method based on the fine-grained body parts learning strategy (*i.e.*, RCSANet^{fg}). Our method outperforms all other methods. For instance, our performance surpasses the best ST-reID method DG-Net by 15.2% and 23.1% in terms of rank-1 accuracy on two datasets, respectively. More performance improvements on Celeb-reID-light demonstrate the proposed method can be better applied to an absolute clothing-change task. When compared with the State-Of-The-Art (SOTA) LT-reID method with the same backbone employed (*i.e.*, ReIDCaps [13] with DenseNet-121 backbone pretrained on Market-1501), our method still achieves significant performance gains over two datasets. Note that, the ICE stream, which is used to extract f_{id} for inference, is not even pretrained with any

Table 1. Comparison with SOTA methods on Celeb-reID and Celeb-reID-light (%). 'R-1' is short for rank-1 accuracy.

Method	Celeb-reID		Celeb-reID-light	
	mAP	R-1	mAP	R-1
methods designed for ST-reID				
Two-Stream [44]	7.8	36.3	-	-
MLFN [3]	6.0	41.4	6.3	10.6
HACNN [16]	9.5	47.6	11.5	16.2
Part-Aligned [24]	6.4	19.4	-	-
PCB [25]	8.2	37.1	-	-
MGN [27]	10.8	49.0	13.9	21.5
DG-Net [43]	10.6	50.1	12.6	23.5
methods designed for LT-reID				
2SF-BPart [9]	-	-	14.0	26.8
ReIDCaps [13]	9.8	51.2	11.2	20.3
ReIDCaps^{fg} [13]	**15.8**	**63.0**	**19.0**	**33.5**
RCSANet (**Ours**)	11.9	55.6	16.7	29.5
RCSANet^{fg} (**Ours**)	**17.5**	**65.3**	**24.4**	**46.6**

Table 2. Comparison with SOTA methods on PRCC (%). 'C-C' is short for 'Clothing-Change'. As in [32] 'Sketch' means the inputs of the model are contour sketch images.

Method	No-C-C		C-C	
	mAP	R-1	mAP	R-1
methods designed for ST-reID				
LOMO+XQDA [18]	-	29.4	-	14.5
Face [29]	-	4.8	-	3.0
Shape Context [2]		23.9		11.5
LNSCT [31]		35.5		15.3
PCB [25]	-	86.9	-	22.9
PCB(Sketch) [25]	-	57.4	-	22.5
HACNN [16]	-	82.5	-	21.8
HACNN(Sketch) [16]	-	58.6	-	20.5
SketchNet [35]	-	64.6	-	17.9
methods designed for LT-reID				
ASENet [32]+STN [14]	-	59.2	-	27.5
ASENet [32]+SPT [32]	-	64.2	-	**34.4**
RCSANet (**Ours**)	**96.6**	**99.6**	31.5	31.6
RCSANet^{fg} (**Ours**)	**97.2**	**100.0**	**48.6**	**50.2**

extra re-ID dataset. Only ICR is pretrianed using Market-1501 but does not involved in testing (refer to Sec. 4.1). This comparison demonstrates that our method is the best to tackle clothing-change cases (*i.e.*, Celeb-reID-light) while also robust to the scenario where both clothing-change case and no-clothing-change case coexist (*e.g.*, Celeb-reID).

4.5. Result on PRCC

We compare our RCSANet with seven ST-reID methods and one LT-reID methods on PRCC. The result is shown in Tab. 2. ASENet [32] is specifically designed for the PRCC dataset, which directly uses body contour as the input. It is observed that, our RCSANet achieves the best performance in both clothing-change evaluation and no-clothing-change evaluation. Compared with ASENet, our RCSANet improves the rank-1 accuracy in no-clothing-change evaluation by more than 30%. In addition, our RCSANet^{fg} achieves 100% in terms of rank-1 accuracy in no-clothing-change evaluation, which significantly outperforms all ST-reID methods in this comparison. Amongst these ST-reID methods, as reported in [32], PCB (proposed in 2018s) achieves the best performance in the no-clothing-change evaluation on PRCC (*e.g.*, rank-1: 86.9%). Some recent published ST-reID methods may achieve competitive performance in this no-clothing-change evaluation compared with our RCSANet. However, since they are not designed for the LT-reID task, our experiments just report the performance of seven ST-reID methods mentioned in [32]. In the clothing-change evaluation, our method also outperforms all the other methods by a large margin when the fine-grained body parts learning strategy is adopted. It can observe that, without using body parts, the performance of our method is lower than ASENet (*i.e.*, Rank-1: 31.6% *vs.* 34.4%). This is because the PRCC dataset is specifically built for body contour feature extraction by enforcing a person to wear clothes of similar thickness [32]. Therefore, ASENet which uses the body contour as input is more suitable for the characteristic of PRCC dataset. When conducting no-clothing-change evaluation, the body contour is no longer competitive compared with RGB images which contain more useful appearance information, resulting in inferior performance (*e.g.*, Rank-1: 64.2%).

4.6. Quantitative Evaluations

Ablation Study of our RCSANet is given in Tab. 3. Since the main contribution of RCSANet is to handle both clothing-change cases and non-clothing-change cases, the ablation study is mainly conducted on the Celeb-reID dataset. This is because as a typical LT-reID dataset Celeb-reID contains both cases collected under uncontrolled environment without special constraints on persons (*e.g.*, a person does not need to wear clothes of similar thickness as in PRCC [32]). Compared with Celeb-reID, the Celeb-reID-light dataset does not contain no-clothing-change cases for each individual. Therefore, Celeb-reID is most suitable for the ablation study. The ablation study is divided into five parts: **1)** Evaluating the effectiveness of L_{id} and L_{cit} in ICE; **2)** Evaluating MPM in ICE by using different combinations of pooling layers and slices; **3)** Evaluating modules in ICR when both ICR and ICE are trained together. **4)** Evaluating features for inference. **5)** The usage of different backbones (*i.e.*, ResNet-50 [7] *vs.* DenseNet-121 [8]).

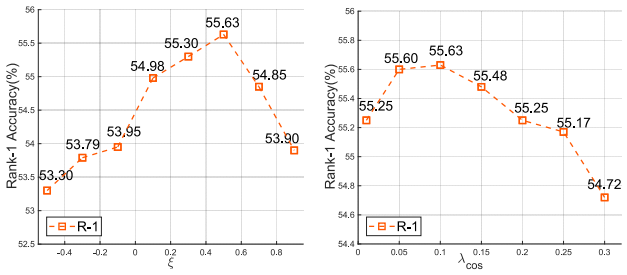
Table 3. Ablation study of proposed methods (%). Celeb-reID is used in this experiment. 'w/o' is short for 'without'.

Methods	mAP	R-1
1) Evaluate L_{id} and L_{cit}: Only train ICE (using MPM)		
w/o. L_{id}	7.8	41.4
w/o. L_{cit}	9.6	49.9
L_{id}+L_{cit}	10.3	52.0
2) Evaluate MPM in ICE: Only train ICE (L_{id}+L_{cit})		
w/o. division, GAP only	9.4	50.2
w/o. division, GMP only	9.2	49.9
w/o. division, GAP+GMP (similar to [42])	9.6	50.6
2 horizontal slices, GAP+GMP	9.2	50.8
2 vertical slices, GAP only	9.8	51.4
2 vertical slices, GMP only (LMP [4])	9.6	51.0
4 vertical slices, GAP+GMP	9.9	51.6
2 vertical slices, GAP+GMP (**Our MPM**)	10.3	52.0
3) Evaluate modules in ICR: Train full ICE with ICR		
w/o. DAM and L_{cos}^{sim} + L_{cos}^{dis}	10.3	52.0
w/o. DAM and L_{cos}^{sim}	10.8	53.0
w/o. DAM and L_{cos}^{dis}	10.5	52.5
w/o. DAM only	11.7	54.8
ICE + ICR final (**Ours**)	**11.9**	**55.6**
4) Evaluate features for inference: Train full ICE with ICR		
only use f_{ap} extracted from ICR	2.1	3.3
f_{ap} + f_{id}	3.3	16.4
only use f_{id} extracted from ICE (**Ours**)	**11.9**	**55.6**
5) ResNet-50 *vs.* DenseNet-121: Train full ICE with ICR		
backbone: ResNet-50	11.0	54.9
backbone: DenseNet-121 (**Ours**)	**11.9**	**55.6**

1) It is observed that without L_{id} or L_{cit}, the performance drops from 52.0% to 41.4% or 49.9% in rank-1 accuracy, which demonstrates that the combination of L_{id} and L_{cit} can enhance the discrimination ability of f_{id} in ICE.

2) Without slice division, the performance is always lower than MPM regardless of the types of pooling layer used (*e.g.*, 52% *vs.* 50.6% in terms of rank-1 accuracy). Other types of combinations also cannot outperform MPM in the comparison. Two vertical slices with GMP (*i.e.*, LMP) has been used in [4] for the classic cross-domain person re-ID task. However, it is still lower than our MPM in terms of rank-1 accuracy by 1% (*i.e.*, 52.0% *vs.* 51.0%).

3) The FRM which consists of L_{cos}^{sim} and L_{cos}^{dis} is the bridge between ICE and ICR. Without L_{cos}^{sim} and L_{cos}^{dis}, ICE and ICR are trained separately. Therefore, the performance is the same as training ICE only (*i.e.*, mAP: 10.3% and rank-



(a) Change ξ. (b) Change λ_{cos}.

Figure 3. Hyperparameters analysis for ξ and λ_{cos}.

1: 52.0%). In addition, the performance drops from 55.6% to 53.0% or 52.5% when L_{cos}^{sim} or L_{cos}^{dis} is removed. This results show that two losses should be joint optimized to achieve the best performance. To reduce the domain gap, we introduce DAM in ICR. Without using DAM (*i.e.*, parameters of backbone in ICR are frozen), the performance can drop about 1% in rank-1 accuracy.

4) As show in Tab. 3, when we change to use f_{ap} for inference, a sharp performance drop is observed. This is because f_{ap} is used to regularize f_{id} during RCSANet training to encourage f_{id} can handle both clothing-change and no-clothing-change cases. f_{ap} does not have the ability to distinguish different IDs since there is no supervision information provided in ICR (refer to Sec. 3.2 for details). Apparently, when the ranking list of f_{ap} and f_{id} are directly combined (*i.e.*, $f_{ap} + f_{id}$ in Tab. 3), the performance is also lower than only using f_{id}.

5) The backbone we choose for the proposed RCSANet is DenseNet-121 (see Sec. 4.1) which has demonstrated its effectiveness in existing LT-reID work (*i.e.*, ReIDCaps [13]). ResNet-50 [7] is also widely used in the traditional ST-reID scenario. Therefore, we replace DenseNet-121 with ResNet-50 in order to testify the influence caused by using different backbones. It can be seen that the performance is slightly lower than DenseNet-121 when ResNet-50 is adopted (*e.g.*, mAP: 11.0% *vs.* 11.9%). For the selection of backbones, other alternatives may achieve better performance. This work just provides a brief comparison between DenseNet-121 and ResNet-50.

Hyperparameters Analysis is illustrated in Fig. 3. There are two hyperparameters in our RCSANet, including $\xi \in [-1, 1]$ in Eq. 6 and λ_{cos} in Eq. 8. **1)** By varying ξ, the result is shown in Fig. 3 (a). ξ can affect the regularization performance of ID features when L_{cos} is adopted between ICR and ICE. It can be seen that the performance decreases when ξ <0.5 or >0.5. Therefore, we choose ξ=0.5 in Eq. 6 that achieves the best performance. **2)** λ_{cos} is an important parameter to control the ability of ICR for ID feature regularization. The result of varying λ_{cos} is shown in Fig. 3 (b). The performance can gradually degrade when

Table 4. Performance evaluation by changing training data (%). 'Cel' is short for 'Celeb-reID'. 'X→ Y' means training is conducted on X and testing is conducted on Y.

Methods	Cel→Cel		Cel-light→Cel	
	mAP	R-1	mAP	R-1
SOTA(ReIDCaps [13])	9.8	51.2	6.5	41.8
RCSANet (**Ours**)	**11.9**	**55.6**	**11.1**	**50.4**

λ_{cos} >0.1 since a large weight for λ_{cos} may lead to over-regularization. The performance also slightly drops about 0.4% when we change λ_{cos} from 0.1 to 0.01. A small λ_{cos} may reduce its effect on ID feature regularization.

Sensitivity to The Number of No-Clothing-Change Cases in Training Data. Intuitively, if a model is able to tackle no-clothing-change cases in LT-reID, the training data should contain a certain number of cases where a person does not change clothing. However, this is hard to be guaranteed in LT-reID where a person has a great chance to change clothing. In order to show the impact caused by training data, this experiment uses Celeb-reID-light (only contains clothing-change cases) for training and Celeb-reID (both clothing-change case and no-clothing-change case coexist) for testing. The two datasets have a small domain gap since the former is a subset of the latter [13]. Tab. 4 shows results. If ReIDCaps uses Celeb-reID-light for training, the model suffers from a catastrophic performance degradation (*i.e.*, Rank-1: from 51.2% to 41.8%). The result of our method (*i.e.*, from 55.6% to 50.4%) is much better than ReIDCaps. This experiment shows that in order to handle no-clothing-change cases in LT-reID, our method does not heavily rely on the number of no-clothing-change cases in training data. This is because the clothing status awareness in our method has the ability to explicitly learn clothing information even clothing is always changed within each IDs.

5. Conclusion

This paper proposes handling the LT-reID issue by embedding the clothing status awareness learning in our RCSANet. During RCSANet training, the robustness of ID discriminative features learned from RCSANet for LT-reID can be further improved using a proposed feature regularization process (*i.e.*, FRM) via checking the status of clothing of each individual person, aiming to effectively handle both the clothing-change cases and no-clothing-change cases. Compared with existing LT-reID approaches, the proposed RCSANet is able to effectively tackle LT-reID issue when both clothing-change cases and no-clothing-change cases exist, making it more solid and robust to real-world LT-reID applications. Experiments on three benchmarks demonstrate that our RCSANet consistently brings substantial improvement to LT-reID accuracy under various situations.

References

[1] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. Re-identification with rgb-d sensors. In *ECCV*, 2012.

[2] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE TPAMI*, 2002.

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

[4] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

[6] Albert Haque, Alexandre Alahi, and Li Fei-Fei. Recurrent attention models for depth-based person identification. In *CVPR*, 2016.

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

[8] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[9] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In *IJCNN*, 2019.

[10] Yan Huang, Qiang Wu, JingSong Xu, and Yi Zhong. Sbsgan: Suppression of inter-domain background shift for person re-identification. In *ICCV*, 2019.

[11] Yan Huang, Qiang Wu, Jingsong Xu, Yi Zhong, and Zhaoxiang Zhang. Unsupervised domain adaptation with background shift mitigating for person re-identification. *Springer IJCV*, 2021.

[12] Yan Huang, Jingsong Xu, Qiang Wu, Zhedong Zheng, Zhaoxiang Zhang, and Jian Zhang. Multi-pseudo regularized label for generated data in person re-identification. *IEEE TIP*, 2018.

[13] Yan Huang, Jingsong Xu, Qiang Wu, Yi Zhong, Peng Zhang, and Zhaoxiang Zhang. Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *IEEE TCSVT*, 2019.

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *NeurIPS*, 2015.

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

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

[17] Yu-Jhe Li, Zhengyi Luo, Xinshuo Weng, and Kris M Kitani. Learning shape representations for person re-identification under clothing change. In *WACV*, 2020.

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

[19] Matteo Munaro, Alberto Basso, Andrea Fossati, Luc Van Gool, and Emanuele Menegatti. 3d reconstruction of freely moving persons for re-identification with a depth sensor. In *ICRA*, 2014.

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[21] 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 *ACCV*, 2020.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[23] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *NeurIPS*, 2017.

[24] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018.

[25] 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.

[26] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *CVPRW*, 2020.

[27] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACMMM*, 2018.

[28] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR*, 2019.

[29] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

[30] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Robust depth-based person re-identification. *IEEE TIP*, 2017.

[31] Xiaohua Xie, Jianhuang Lai, and Wei-Shi Zheng. Extraction of illumination invariant facial features from a single image using nonsubsampled contourlet transform. *Pattern Recognition*, 2010.

[32] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE TPAMI*, 2019.

[33] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *arXiv preprint arXiv:2001.04193*, 2020.

[34] Shijie Yu, Shihua Li, Dapeng Chen, Rui Zhao, Junjie Yan, and Yu Qiao. Cocas: A large-scale clothes changing person dataset for re-identification. In *CVPR*, 2020.

[35] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *CVPR*, 2016.

[36] Peng Zhang, Qiang Wu, Jingsong Xu, and Jian Zhang. Long-term person re-identification using true motion from videos. In *WACV*, 2018.

[37] Peng Zhang, Jingsong Xu, Qiang Wu, Yan Huang, and Xianye Ben. Learning spatial-temporal representations over walking tracklet for long-term person re-identification in the wild. *IEEE TMM*, 2020.

[38] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *ICCV*, 2019.

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

[40] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[41] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011.

[42] Zhedong Zheng, Minyue Jiang, Zhigang Wang, Jian Wang, Zechen Bai, Xuanmeng Zhang, Xin Yu, Xiao Tan, Yi Yang, Shilei Wen, et al. Going beyond real data: A robust visual representation for vehicle re-identification. In *CVPRW*, 2020.

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

[44] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM TOMM*, 2017.

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