Clothes-Changing Person Re-identification with RGB Modality Only

Xinqian Gu[1,2], Hong Chang[1,2], Bingpeng Ma[2], Shutao Bai[1,2], Shiguang Shan[1,2], Xilin Chen[1,2]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
{xinqian.gu, shutao.bai}@vipl.ict.ac.cn, {changhong, sgshan, xlchen}@ict.ac.cn, bpma@ucas.ac.cn

Abstract

The key to address clothes-changing person re-identification (re-id) is to extract clothes-irrelevant features, e.g., face, hairstyle, body shape, and gait. Most current works mainly focus on modeling body shape from multi-modality information (e.g., silhouettes and sketches), but do not make full use of the clothes-irrelevant information in the original RGB images. In this paper, we propose a Clothes-based Adversarial Loss (CAL) to mine clothes-irrelevant features from the original RGB images by penalizing the predictive power of re-id model w.r.t. clothes. Extensive experiments demonstrate that using RGB images only, CAL outperforms all state-of-the-art methods on widely-used clothes-changing person re-id benchmarks. Besides, compared with images, videos contain richer appearance and additional temporal information, which can be used to model proper spatiotemporal patterns to assist clothes-changing re-id. Since there is no publicly available clothes-changing video re-id dataset, we contribute a new dataset named CCVID and show that there exists much room for improvement in modeling spatiotemporal information. The code and new dataset are available at: https://github.com/guxinqian/Simple-CCReID.

1. Introduction

Person re-identification (re-id) [12, 23, 54] aims to search the target person from surveillance videos across different locations and times. Most existing works [11, 19, 40] assume that pedestrians do not change their clothes in a short period of time. However, if we want to re-identify a pedestrian over a long period of time, the clothes-changing problem cannot be avoided. Besides, clothes-changing problem also exists in some short-time real-world scenarios, *e.g.*, criminal suspects usually change their clothes to avoid being identified and tracked. Due to the crucial role in intelligent surveillance system, clothes-changing person re-id [7, 49] has attracted increasing attention in recent years.
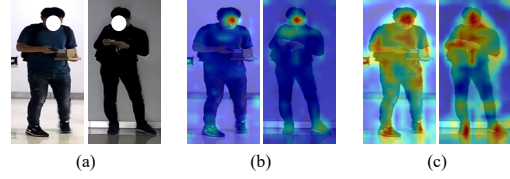
Humans can distinguish their acquaintances, even if

Figure 1. The visualization of (a) two original images, (b) the learned feature maps only with identification loss, and (c) the learned feature maps with identification loss and the proposed CAL. Note that all training settings of (b) and (c) are consistent except loss functions. (b) only highlights face as the clothes-irrelevant features, while (c) highlights more clothes-irrelevant features, *e.g.*, face, hairstyle, and body shape. (Since different samples of the same person in the training set mostly wear the same shoes, shoes are also highlighted.)

these acquaintances wear clothes that they have never seen before. The reason is that the human brain can decouple and utilize clothes-irrelevant features, *e.g.*, face, hairstyle, body shape, and gait. To avoid the interference of clothes, some clothes-changing re-id methods [6, 18] and gait recognition methods [4, 52] model body shape and gait from multi-modality inputs (e.g., skeletons [35], silhouettes [4], radio signals [7], contour sketches [49], and 3D shape [6]) or by disentangled representation learning [52]. However, multi-modality-based methods need additional models or equipment to capture multi-modality information, and learning disentangled representations is usually time-consuming.

Actually, the original RGB modality contains rich clothes-irrelevant information which is largely underutilized by the current methods. As for some clothes-changing re-id methods [6, 35], although they use a strong backbone (*i.e.* ResNet [15]) to extract features from the original images, without a properly designed loss function, the learned feature map only focuses on some simple clothes-irrelevant information, *e.g.*, face (see Fig. 1 (b)), while other crucial clothes-irrelevant information is omitted. As for most gait recognition methods [4, 13], they usually discard the original input videos and resort to other modality inputs, *e.g.*, silhouettes.

To better mine the clothes-irrelevant information in RGB

1

modality, in this paper, we propose *Clothes-based Adversarial Loss* (CAL). Specifically, we add a clothes classifier after the backbone of the re-id model and define CAL as a *multi-positive-class classification loss*, where all clothes classes belonging to the same identity are mutually positive classes. To the best of our knowledge, this is the first work that uses multi-positive-class classification to formulate multi-class adversarial learning. During training, minimizing CAL can force the backbone of the re-id model to learn clothes-irrelevant features by penalizing the predictive power of the re-id model w.r.t. different clothes of the same identity. With backpropagation, the learned feature map can highlight more clothes-irrelevant features, *e.g.*, hairstyle and body shape, compared with the feature map trained only with identification loss (see Fig. 1 (c)). Extensive experiments on widely used clothes-changing re-id benchmarks demonstrate that using RGB images only, CAL outperforms all state-of-the-art methods.

Most current clothes-changing person re-id works [35, 49, 50] mainly focus on image-based setting, where both query and gallery samples are images. However, in many real-world re-id scenarios, both query and gallery sets usually consist of lots of videos. Compared with images, videos contain richer appearance information and additional temporal information. It is more promising to learn proper spatiotemporal patterns from videos, *e.g.*, gait, which may be helpful for clothes-changing re-id. Since there is no publicly available dataset, we reconstruct a new Clothes-Changing Video person re-ID (CCVID) dataset from the raw data of a gait recognition dataset (*i.e.* FVG [52]) and provide fine-grained clothes labels. Extensive evaluations of state-of-the-art methods show that the utilization of richer appearance information and additional temporal information can boost the performance of clothes-changing person re-id significantly. We hope CCVID can inspire more clothes-changing video person re-id studies in the future.

2. Related Work

Clothes-changing person re-identification. The core problem to solve clothes-changing re-id is extracting clothes-irrelevant features. To this end, [52, 55] attempts to use disentangled representation learning to decouple appearance and structural information from RGB images, and considers structural information as clothes-irrelevant features. In contrast, other researchers attempt to use multi-modality information (*e.g.*, skeletons [35], silhouettes [18, 28], radio signals [7], contour sketches [49], or 3D shape [6]) to model body shape and extract clothes-irrelevant features. However, the training of disentangled representation learning is time-consuming, and multi-modality-based methods need additional models or equipment to extract multi-modality information. Besides, these methods do not fully excavate and utilize the clothes-

irrelevant features in the original images. In this paper, we propose a simple adversarial loss to decouple clothes-irrelevant features from the RGB modality.

Most current works [6, 18, 24, 26, 35, 49] mainly focus on image-based settings and only a few works [7, 10, 51] focus on video-based settings. Besides, there are some publicly available clothes-changing person re-id datasets, *e.g.*, PRCC [49], LTCC [35], and Celeb-reID [24, 26], but all of them are image-based datasets. In this paper, we contribute a large-scale clothes-changing video person re-id dataset and show that there exists much room to improve the spatiotemporal modeling for clothes-changing person re-id.

Video person re-identification. Most existing video person re-id methods [11, 19, 21, 22, 45] focus on clothes-consistent setting. Some research [5] demonstrates that appearance feature plays a more important role than motion feature in this setting. Though it is straightforward to distinguish a person through the appearance of his/her clothes in clothes-consistent setting, these deep person re-id models usually overfit clothes-relevant features and have limited application scenarios. In contrast, this paper focuses on clothes-changing setting and proposes a solution through learning clothes-irrelevant features.

Gait recognition. Gait recognition methods [4, 52] attempt to learn gait features from the video samples of persons. To avoid the interference of clothes, they usually discard the original RGB frames and model gait on skeletons [1, 8], silhouettes [4, 13], or disentangled representations [52]. As for clothes-changing person re-id, all kinds of clothes-irrelevant features besides gait are useful. Therefore, in this paper, we attempt to mine more clothes-irrelevant features from the original RGB modality.

Adversarial learning. Adversarial learning is first proposed in GAN [9] to force the generative model to generate realistic images. In recent years, it has been used in various tasks, *e.g.*, domain adaptation [32, 42], knowledge distillation [36], and representation learning [44]. Specifically, in [44], Wang *et al.* propose PAR to force the model to focus on discriminative global information by penalizing the predictive power of local features. Inspired by PAR, this paper proposes a Clothes-based Adversarial Loss to decouple clothes-irrelevant features. Although both PAR and the proposed method belong to multi-class adversarial learning, the motivation and formulation are different. We will discuss their differences in detail in Sec. 3.3.

3. Method

3.1. Framework and Notation

The framework of our method is shown in Fig. 2. In the framework, $g_\theta(\cdot)$ denotes the backbone with parameters θ and $C_\phi^{ID}(\cdot)$ denotes the identity classifier with parameters
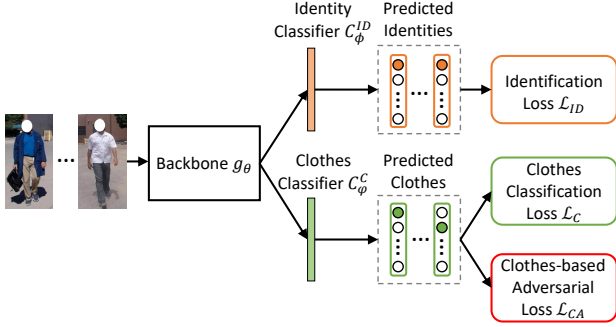
Figure 2. The framework of the proposed method. In each iteration, we first optimize the clothes classifier by minimizing $\mathcal{L}_C$. Then, we fix the parameters of the clothes classifier and minimize $\mathcal{L}_{ID}$ and $\mathcal{L}_{CA}$ to force the backbone to learn clothes-irrelevant features.

ϕ. Given a sample x_i, its identity label is denoted as y_i^{ID}, and its clothes label is denoted as y_i^C. Note that we define the clothes class as fine-grained identity class. All samples of the same identity are divided into different clothes classes belonging to this identity according to their clothes. The number of clothes classes is the sum of the number of suits of different persons. The annotation of such clothes labels is easy, since they only need to be labeled among all samples of the same person, and different persons do not share the same clothes label even if they wear the same clothes. Given a sample x_i with identity label y_i^{ID}, existing re-id methods [20, 40] define identification loss $\mathcal{L}_{ID}$ using the cross entropy between predicted identity $C_\phi^{ID}(g_\theta(x_i))$ and identity label y_i^{ID}, and train the re-id model by minimizing $\mathcal{L}_{ID}$.

As shown in Fig. 2, except for the identity classifier and the widely used identification loss, clothes classification loss $\mathcal{L}_C$ is used to train an additional clothes classifier. The proposed Clothes-based Adversarial Loss (CAL) $\mathcal{L}_{CA}$ is used to force the backbone to decouple clothes-irrelevant features. We will introduce CAL in detail in the next subsection.

3.2. Clothes-based Adversarial Loss

Existing clothes-changing person re-id [35, 49] and gait recognition methods [4, 8] do not make full use of the clothes-irrelevant information in RGB modality. In this paper, we propose CAL to force the backbone of the re-id model to mine clothes-irrelevant information by penalizing the predictive power of the re-id model w.r.t. clothes. To this end, we add a new clothes classifier $C_\varphi^C(\cdot)$ with parameters φ after the backbone. Each iteration in the training stage contains the following two-step optimization.

Training clothes classifier. In the first step, we optimize the clothes classifier by minimizing clothes classification loss $\mathcal{L}_C$ (the cross entropy loss between predicted clothes

$C_\varphi^C(g_\theta(x_i))$ and clothes label y_i^C). This process can be formulated as:

$$\min_\varphi \mathcal{L}_C(C_\varphi^C(g_\theta(x_i)), y_i^C). \tag{1}$$

When we denote $g_\theta(x_i)$ after l_2-normalization as f_i and denote the weights of j-th clothes classifier after l_2-normalization as φ_j, $\mathcal{L}_C$ can be expressed as:

$$\mathcal{L}_C = -\sum_{i=1}^{N} \log \frac{e^{(f_i \cdot \varphi_{y_i^C}/\tau)}}{\sum_{j=1}^{N_C} e^{(f_i \cdot \varphi_j/\tau)}}, \tag{2}$$

where N is the batch size, N_C is the number of clothes classes in the training set, and $\tau \in \mathcal{R}^+$ is a temperature parameter.

Learning clothes-irrelevant features. In the second step, we fix the parameters of the clothes classifier and force the backbone to learn clothes-irrelevant features. To this end, we should penalize the predictive power of re-id model w.r.t. clothes. A naive idea is defining $\mathcal{L}_{CA}$ opposite to $\mathcal{L}_C$ following [44], such that the trained clothes classifier cannot distinguish all kinds of clothes in the training set. In this way, we can get a widely used min-max optimization problem. However, since clothes class is defined as fine-grained identity class, penalizing the predictive power of re-id model w.r.t. all kinds of clothes will also reduce its predictive power w.r.t. identity, which is harmful to re-id (we will demonstrate this in Sec. 5.4). What we want to do is making the trained clothes classifier cannot distinguish the samples with the same identity and different clothes. So $\mathcal{L}_{CA}$ should be a *multi-positive-class classification loss*, where *all clothes classes belonging to the same identity are mutually positive classes*. For example, given a sample x_i, all clothes classes belonging to its identity class y_i^{ID} are defined as its positive clothes classes. Therefore, $\mathcal{L}_{CA}$ can be formulated as:

$$\mathcal{L}_{CA} = -\sum_{i=1}^{N}\sum_{c=1}^{N_C} q(c) \log \frac{e^{(f_i \cdot \varphi_c/\tau)}}{e^{(f_i \cdot \varphi_c/\tau)} + \sum_{j \in S_i^-} e^{(f_i \cdot \varphi_j/\tau)}}, \tag{3}$$

$$q(c) = \begin{cases} \dfrac{1}{K} & , c \in S_i^+ \\ 0 & , c \in S_i^- \end{cases}, \tag{4}$$

where S_i^+ (S_i^-) is the set of clothes classes with the same identity as (different identities from) f_i. K is the number of classes in S_i^+ and $q(c)$ is the weight of cross entropy loss for c-th clothes class. The positive class with the same clothes ($c = y_i^C$) and the positive classes with different clothes ($c \neq y_i^C$ and $c \in S_i^+$) have equal weight, *i.e.* $1/K$.

In a long-term person re-id system, both clothes-consistent re-id and clothes-changing re-id are equally important. When we maximize the dot product between f_i

and the proxy of the positive class with different clothes, the accuracy of clothes-changing re-id can be improved but the accuracy of clothes-consistent re-id may reduce. To improve the clothes-changing re-id ability of the model without reducing the clothes-consistent re-id accuracy heavily, Eq. (4) can be replaced by:

$$q(c) = \begin{cases} 1 - \epsilon + \dfrac{\epsilon}{K} & , c = y_i^C \\ \dfrac{\epsilon}{K} & , c \neq y_i^C \text{ and } c \in S_i^+ , \\ 0 & , c \in S_i^- \end{cases} \quad (5)$$

where $0 < \epsilon \leq 1$ is a hyper-parameter. When $\epsilon = 1$, Eq. (5) is equivalent to Eq. (4). Otherwise, the positive class with the same clothes has a bigger weight than the positive classes with different clothes.

In the meantime of optimizing CAL, the identity classifier is also optimized. Therefore, the optimization process of the second step is:

$$\min_{\theta,\phi} \mathcal{L}_{ID}(C_\phi^{ID}(g_\theta(x_i)), y_i^{ID}) + \mathcal{L}_{CA}(C_\varphi^C(g_\theta(x_i)), y_i^C). \quad (6)$$

Note that $\mathcal{L}_{ID}$ and $\mathcal{L}_{CA}$ have some affinity in learning clothes-irrelevant features. When we only use $\mathcal{L}_{ID}$ for training, the model tends to learn easy samples (with the same clothes) in the early stage of optimization and then learns to distinguish hard samples (with the same identity and different clothes) gradually. This is consistent with curriculum learning [2]. The objective of $\mathcal{L}_{CA}$ is to pull the features with the same identity closer, which is similar to $\mathcal{L}_{ID}$. Even though, we do not discard $\mathcal{L}_{ID}$ in Eq. (6). The reason is that only minimizing $\mathcal{L}_{CA}$ and forcing the model to distinguish hard samples in the early stage of optimization may lead to local optimum. On the contrary, we add $\mathcal{L}_{CA}$ for training after the first reduction of the learning rate in our experiments.

3.3. Discussion

Relations between CAL and PAR. The idea of CAL is inspired by PAR [44]. Both CAL and PAR belong to multi-class adversarial learning methods, but both motivation and formulation of these two methods are different. PAR defines the multi-class adversarial loss as negative cross entropy loss and forces the model to focus on discriminative global information by penalizing the predictive power of local features w.r.t. all classes. However, in this paper, since we define clothes class as fine-grained identity class, if we use the same formulation as PAR, *i.e.* negative cross entropy loss, to penalize the predictive power of re-id model w.r.t. all kinds of clothes, the predictive power of re-id model w.r.t. identify will also be reduced which is contrary to our target. Hence, we define CAL as a *multi-positive-class classification loss*, where all clothes classes belonging to the same identity are mutually positive classes. In other words,

Table 1. The statistics of our CCVID dataset and other video person re-id and clothes-changing person re-id datasets.

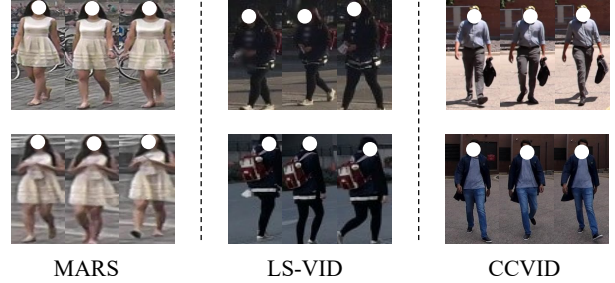| datasets	#identities	#sequences	#bboxes	changing clothes?
PRID	200	400	40,033	✗
iLIDS-VID	300	600	42,460	✗
MARS	1,261	19,608	1,191,003	✗
LS-VID	3,772	14,943	2,982,685	✗
Real28	28	-	4,324	✓
VC-Clothes	512	-	19,060	✓
LTCC	152	-	17,119	✓
PRCC	221	-	33,698	✓
Celeb-reID	1,052	-	34,186	✓
DeepChange	1,082	-	171,352	✓
LaST	10,860	-	224,721	✓
CCVID	226	2,856	347,833	✓

Figure 3. Two different video samples of the same identity on MARS, LS-VID, and CCVID datasets respectively. Only CCVID involves clothes changes.

we just want to make the trained clothes classifier cannot distinguish the samples with the same identity and different clothes. To the best of our knowledge, this is the first work that uses multi-positive-class classification to formulate multi-class adversarial learning. It is our main technical contribution.

Differences between CAL and label smooth regularization. To get a trade-off between clothes-changing re-id and clothes-consistent re-id accuracy, we set different weights for different positive classes in Eq. (5), but the weights of negative classes are still 0. In contrast, label smoothing regularization [41] sets the weights of negative classes to small non-zero values to avoid overfitting.

4. CCVID Dataset

As shown in Tab. 1 and Fig. 3, all existing publicly available video person re-id datasets (*i.e.* PRID [17], iLIDS-VID [45], MARS [53], and LS-VID [30]) do not involve clothes changes. Besides, existing publicly available clothes-changing person re-id datasets (*i.e.* Real28&VC-Clothes [43], LTCC [35], PRCC [49], Celeb-reID [26], DeepChange [48], and LaST [37]) only contain still images and do not involve sequence data. However, as analyzed in

the introduction, clothes-changing video re-id is closer to real-world re-id scenarios, and the abundant appearance information and additional temporal information in the video samples are helpful for clothes-changing re-id.

To provide a publicly available benchmark, we construct a Clothes-Changing Video person re-ID (CCVID) dataset from the raw data of a gait recognition dataset, *i.e.* FVG [52][1]. FVG dataset contains 2,856 sequences from 226 identities and each identity has 2∼5 suits of clothes. In the original FVG dataset, 1,620 sequences from 135 identities are collected in 2017 and 948 sequences from the other 79 identities are collected in 2018. There also are 12 persons whose sequences are collected both in 2017 and 2018. Since gait recognition methods usually use masked images, while re-id methods use the images after detection. So we reconstruct this dataset by performing detection [14] on the raw data. Since most frames of FVG only contain one person, we only detect the person with the highest score for each frame, and tracking algorithm is not required. The reconstructed CCVID dataset contains 347,833 bounding boxes. The length of each sequence changes from 27 to 410 frames, with an average length of 122. Besides, we also provide fine-grained clothes labels including tops, bottoms, shoes, carrying status, and accessories. For the convenience of evaluation, we re-divide the training and test sets to adapt to clothes-changing re-id. Specifically, 75 identities are reserved for training, and the remaining 151 identities are used for test. In the test set, 834 sequences are used as query set, and the other 1074 sequences form gallery set.

In Sec. 5.5, we will make a fair comparison between image-based setting and video-based setting on CCVID. Also, we will reproduce some state-of-the-art video person re-id and gait recognition methods on CCVID and compare their performance.

5. Experiments

5.1. Datasets and Evaluation Protocol

We mainly evaluate the proposed method on CCVID and two widely-used clothes-changing image person re-id datasets (*i.e.* PRCC [49] and LTCC [35]). The results on VC-Clothes [43], LaST [37], and DeepChange [48] are shown in supplementary materials. Top-1 accuracy and mAP are used as evaluation metrics and three kinds of test settings are defined as follows: (i) **general setting** (both clothes-changing and clothes-consistent ground truth samples are used to calculate accuracy), (ii) **clothes-changing setting** (abbreviated as **CC**. In this setting, only clothes-changing ground truth samples are used to calculate accu-

racy), and (iii) **same-clothes setting** (abbreviated as **SC** and also named **clothes-consistent setting**. In this setting, only clothes-consistent ground truth samples are used to calculate accuracy). For CCVID and LTCC, we report the accuracy for both general re-id and clothes-changing re-id. As for PRCC, following [49], the re-id accuracies in same-clothes setting and clothes-changing setting are reported.

5.2. Implementation Details

We use ResNet-50 [15] as the backbone of re-id model. To enrich the granularity, the last downsampling of ResNet-50 is removed. As for image-based datasets (*i.e.* LTCC and PRCC), following [25], we use global average pooling and global max pooling to integrate the output feature map of the backbone, and then concatenate them and use Batch-Norm [27] to normalize the image feature. Following [35], the input images are resized to 384×192. Random horizontal flipping, random cropping, and random erasing [56] are used for data augmentation. The batch size is set to 64. Each batch contains 8 persons and 8 images for each person. The model is trained by Adam [29] for 60 epochs and $\mathcal{L}_{CA}$ is used for training after the 25th epoch. The learning rate is initialized to $3.5e^{-4}$ and divided by 10 after every 20 epochs. τ in Eq. (3) is set to $1/16$ and ϵ in Eq. (5) is set to 0.1 by grid search on LTCC. The optimal parameter values are directly used for the other datasets without tuning.

As for the video-based dataset, *i.e.* CCVID, following [11], we use spatial max pooling and temporal average pooling to integrate the output feature map of the backbone and then use BatchNorm [27] to normalize the video feature. The frame lengths of different video samples are different. During training, the frame lengths of inputs should be equal and each frame would better be sampled with equal probability. Hence, for each original video, we randomly sample 8 frames with a stride of 4 to form a video clip. Each input frame is resized to 256×128 and only horizontal flip is used for data augmentation following [11]. Due to the limit of GPU memory, the batch size is set to 32 and each batch contains 8 persons and 4 video clips for each person. The model is trained by Adam [29] for 150 epochs and $\mathcal{L}_{CA}$ is used for training after the 50th epoch. The learning rate is initialized to $3.5e^{-4}$ and divided by 10 after every 40 epochs. In the test stage, each video sample is divided into a series of 8-frame clips with a stride of 4. The averaged feature of these clips is used as the representation of the original video for testing.

5.3. Comparison with State-of-the-art Methods

We compare the proposed CAL with three traditional re-id methods (*i.e.* HACNN [31], PCB [40], and IANet [20]) and six clothes-changing re-id methods (*i.e.* SPT+ASE [49], GI-ReID [28], CESD [35], RCSANet [25], 3DSL [6], and FSAM [18]) on LTCC and PRCC in Tab. 2. Note that these

---

Table 2. Comparison with state-of-the-art methods on LTCC and PRCC. 'sketch', 'sil.', 'pose', and '3D' represent the contour sketches, silhouettes, human poses, and 3D shape information, respectively.

method	modality	clothes label	extra training data	LTCC				PRCC			
				general		CC		SC		CC	
				top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
HACNN [31]	RGB			60.2	26.7	21.6	9.3	82.5	-	21.8	-
PCB [40]	RGB			65.1	30.6	23.5	10.0	99.8	97.0	41.8	38.7
IANet [20]	RGB			63.7	31.0	25.0	12.6	99.4	98.3	46.3	45.9
SPT+ASE [49]	sketch			-	-	-	-	64.2	-	34.4	-
GI-ReID [28]	RGB+sil.			63.2	29.4	23.7	10.4	80.0	-	33.3	-
CESD [35]	RGB+pose	✓		71.4	34.3	26.2	12.4	-	-	-	-
RCSANet [25]	RGB		✓	-	-	-	-	**100**	97.2	50.2	48.6
3DSL [6]	RGB+pose+sil.+3D	✓		-	-	31.2	14.8	-	-	51.3	-
FSAM [18]	RGB+pose+sil.			73.2	35.4	38.5	16.2	98.8	-	54.5	-
CAL	RGB	✓		**74.2**	**40.8**	**40.1**	**18.0**	**100**	**99.8**	**55.2**	**55.8**

Table 3. The ablation studies of CAL on CCVID, LTCC, and PRCC.

method	CCVID				LTCC				PRCC			
	general		CC		general		CC		SC		CC	
	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
baseline	78.3	75.4	77.3	73.9	65.5	29.4	28.1	11.0	99.8	97.9	45.6	43.3
w/ clothes classifier	58.8	55.8	46.2	45.6	62.3	31.0	21.9	10.9	99.5	99.5	33.1	37.4
CAL	**82.6**	**81.3**	**81.7**	**79.6**	**74.2**	**40.8**	**40.1**	**18.0**	**100**	**99.8**	**55.2**	**55.8**
CAL $(-\mathcal{L}_C)$	52.8	53.0	50.0	49.2	21.5	3.1	9.2	2.3	89.6	67.7	19.3	13.1
Triplet Loss [16]	81.5	78.1	81.1	77.0	71.8	37.5	34.7	16.6	100	99.8	48.6	49.7

clothes-changing re-id methods use information from different modalities to avoid the interference of clothes. Especially, 3DSL, FSAM integrate at least three modalities and the computational cost of these two methods is at least four times w.r.t. CAL. Besides, RCSANet uses additional clothes-consistent re-id data to enhance the performance in the same-clothes setting. Nevertheless, using RGB images only and without additional data, the proposed CAL outperforms all these methods consistently on both two datasets. This comparison can demonstrate the effectiveness of CAL.

Limitation. Although CAL achieves state-of-the-art performance without additional modalities and data, it needs clothes labels for adversarial learning. Fortunately, the annotation of such clothes labels is only among the samples of the same person and thus is easier than the annotation of identities. When clothes labels are unavailable in practice, the collection date can be used as pseudo clothes labels to train CAL, since the samples of the same person captured on different days have a high probability of wearing different clothes. We attempt this strategy on DeepChange dataset and demonstrate its effectiveness. The results are shown in supplementary materials. Besides, we will also try to use clustering algorithms to obtain pseudo clothes labels in the future.

5.4. Ablation Studies

The effectiveness of CAL. To verify the effectiveness of the proposed CAL, we reproduce a baseline method that only uses identification loss $\mathcal{L}_{ID}$ for training and all the other settings are consistent with CAL. As shown in Tab. 3, when we add a clothes classifier after the backbone of the baseline and retrain the backbone by minimizing clothes classification loss $\mathcal{L}_C$ and identification loss $\mathcal{L}_{ID}$, the re-id accuracy in the same-clothes setting is superior to the baseline, but the performance in clothes-changing setting is lower than the baseline. These results are reasonable, since minimizing $\mathcal{L}_C$ would force the backbone to learn clothes-relevant features. When $\mathcal{L}_C$ is only used to train the clothes classifier and $\mathcal{L}_{CA}$ is then used to train the backbone, CAL surpasses the baseline for a large margin in both general and clothes-changing setting. One possible explanation is that, with the help of CAL, the backbone is forced to learn clothes-irrelevant features and thus more robust against clothes changes.

Comparison between different formulations. As explained in Sec. 3.2, if we follow PAR [44] and define $\mathcal{L}_{CA} = -\mathcal{L}_C$, minimizing $\mathcal{L}_{CA}$ will penalize the predictive power of re-id model w.r.t. all kinds of clothes in the training set. Since the clothes label is defined as the fine-grained identity label, it will also reduce the predictive power of re-id model w.r.t. identity, which is harmful to re-id. To verify this, we compare CAL with CAL $(-\mathcal{L}_C)$ in Tab. 3. It can be seen that the accuracy of CAL $(-\mathcal{L}_C)$ is much lower than CAL and even lower than the baseline method in all general, clothes-changing, and the same-clothes settings.

Comparison with Triplet loss. We also compare CAL with the widely used metric learning loss, *i.e.* Triplet loss [16].

Table 4. Comparison on standard datasets without clothes-changing.

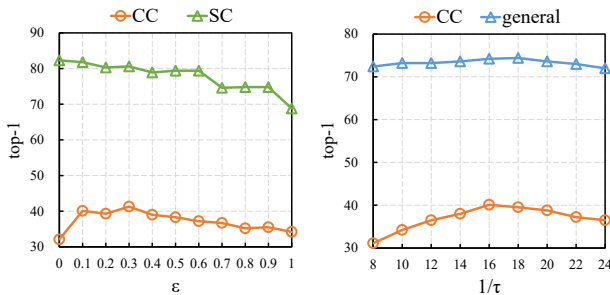| method	Market-1501		MSMT17	
	top-1	mAP	top-1	mAP
PCB [40]	93.8	81.6	68.2	40.4
IANet [20]	94.4	83.1	75.5	46.8
OSNet [57]	94.8	84.9	78.7	52.9
JDGL [55]	94.8	86.0	77.2	52.3
CircleLoss [39]	94.2	84.9	76.3	50.2
baseline	92.2	78.7	67.8	43.5
CAL	94.5	87.3	79.7	57.0
baseline (w/ triplet)	94.5	86.6	78.9	57.0
CAL (w/ triplet)	94.7	87.5	79.7	57.3

Figure 4. The top-1 accuracy of CAL with different ϵ and τ on LTCC. Note that the abscissa of the second subfigure is the $1/\tau$.

As shown in Tab. 3, Triplet loss is superior to the baseline, but CAL outperforms Triplet loss significantly, especially in clothes-changing setting. It is more likely that Triplet loss can only mine the hard cases in a mini-batch, while CAL uses a clothes classifier to save the proxies of all clothes classes in the training set and can mine clothes-irrelevant features in the global scope.

Results on standard person re-id benchmarks. When the test benchmarks do not involve clothes changes and one identity only has one clothes, the clothes classifier in our method will become an identity classifier and the proposed CAL will degenerate into cosine-similarity-based cross-entropy loss. We perform CAL on two standard re-id datasets, *i.e.* Market-1501 [54] and MSMT17 [47], and compare the results with baseline and some state-of-the-art methods in Tab. 4. It can be seen that CAL outperforms the baseline which only uses the original cross-entropy loss as supervision. When we combine these two methods with triplet loss [16], CAL still outperforms the baseline. Besides, CAL achieves comparable performance compared with state-of-the-art methods on these two benchmarks.

The influence of ϵ in CAL. By varying ϵ in Eq. (5), Fig. 4 shows the top-1 accuracy of CAL in the same-clothes setting and clothes-changing setting on LTCC. When ϵ is set to 0, CAL will degenerate into a clothes classification loss and constrain the backbone to learn clothes-relevant features. So, it achieves the lowest top-1 accuracy in clothes-changing setting. With the increase of ϵ, the weight of the

Table 5. Comparison with state-of-the-arts on CCVID. 'F' means only using the first frame for testing.

method	general		CC	
	top-1	mAP	top-1	mAP
baseline(F)	65.0	59.4	63.1	56.3
baseline	78.3	75.4	77.3	73.9
I3D [3]	79.7	76.9	78.5	75.3
+CAL	+1.5	+3.0	+2.3	+3.3
Non-Local [46]	80.7	78.0	79.3	76.2
+CAL	+2.9	+3.4	+3.7	+4.0
TCLNet [19]	81.4	77.9	80.7	75.9
+CAL	+1.3	+3.0	+1.4	+3.7
AP3D [11]	80.9	79.2	80.1	77.7
+CAL	+3.2	+2.0	+3.5	+2.3
GaitNet [52]	62.6	56.5	57.7	49.0
GaitSet [4]	81.9	73.2	71.0	62.1
CAL	**82.6**	**81.3**	**81.7**	**79.6**

positive class with the same clothes as f_i in Eq. (3) decreases gradually, so the top-1 accuracy in the same-clothes setting rate is generally decreasing. As for the accuracy in clothes-changing setting, it increases rapidly and then starts to oscillate, eventually tending to overfit. To get a trade-off between clothes-changing re-id accuracy and the traditional re-id accuracy in the same-clothes setting, we set ϵ to 0.1 for all other experiments.

The influence of temperature parameter τ. In general, the optimal temperature parameter τ is related to the number of classes in the training set. We show the experimental results with varying τ on LTCC in Fig. 4. The best performance is achieved when $\tau = 1/16$.

5.5. Further Analyses on CCVID

Image-based setting *vs*. video-based setting. To make a fair comparison between clothes-changing image re-id and clothes-changing video re-id, we reproduce the baseline method that uses all frames for training but only uses the first frame for testing (baseline(F) in Tab. 5) on CCVID. In the meantime, we also reproduce two classic temporal information modeling methods, *i.e.* I3D [3] and Non-Local [46], and two specially designed temporal information modeling methods for video re-id, *i.e.* TCLNet [19] and AP3D [11] on CCVID. Note that TCLNet is reproduced by the source code provided in the original paper. The implementaiton of I3D, Non-Local, and AP3D is based on the the source code of [11]. As shown in Tab. 5, compared with baseline(F), the baseline which uses all frames for testing achieves significant performance improvement (more than 13%). Compared with the baseline, these temporal information modeling methods can achieve further improvement. These comparisons show the superiority of video-based setting and suggest that there exists much room to improve in spatiotemporal modeling for clothes-changing re-id.

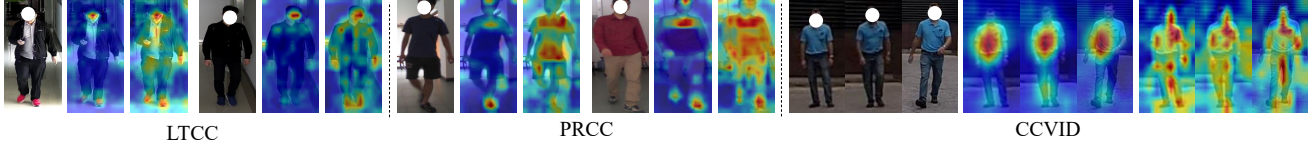**Gait recognition *vs*. clothes-changing video person re-**

Figure 5. The visualization of feature maps on LTCC, PRCC, and CCVID. In each triplet, the first column presents the original image/video frames. The second and third columns present the feature maps of the baseline method and CAL, respectively.

id. We also compare these temporal information modeling methods with two state-of-the-art gait recognition methods (*i.e.* GaitNet [52] and GaitSet [4]) on CCVID. Note that GaitNet and GaitSet model gait from silhouettes and disentangled representations, respectively. All these methods are reproduced by the source codes provided in their papers. As shown in Tab. 5, except for the top-1 accuracy of Gait-Set in the general setting, four temporal information modeling methods show grate advantage over two gait recognition methods. One possible explanation is that compared with silhouettes and disentangled representations, the original RGB modality provides more clothes-irrelevant information, which is helpful for clothes-changing re-id. Besides, the proposed CAL outperforms all these methods. When we combine these temporal information modeling methods with CAL, further improvement can be achieved. This comparison can demonstrate the effectiveness and generality of the proposed method.

5.6. Visualization

We visualize more feature maps of the baseline method and CAL in Fig. 5. On the two image-based datasets (*i.e.* PRCC and LTCC), the feature maps of the baseline method mainly focus on face, shoes, and shoulder. That is, these feature maps focus on clothes-relevant and clothes-irrelevant features which are all beneficial to re-identification: (1) As different samples of the same person in the training set mostly wear the same shoes, shoes are highlighted as key clothes-relevant features; (2) Face and the contour of shoulder are highlighted, which are a part of easy to learn clothes-irrelevant features. With the help of CAL, the feature maps highlight more clothes-irrelevant information, *e.g.*, hairstyle and body shape. As for the video-based dataset, *i.e.* CCVID, the feature maps of the baseline method mainly focus on the body regions. With the constraint of CAL, the learned features can highlight the regions of heads and describe the pose and body shape more clearly. Therefore, CAL can achieve higher clothes-changing re-id accuracy.

Discussion. A controversial issue is whether the high responses of CAL on the body regions are mainly caused by *the texture and color of clothes* or by *the body shape*. To verify this, we remove the top 2/7 (head regions) and bottom 1/7 (foot regions) of the testing images and only reserve the body regions to construct a quantitative experiment. The

Table 6. The results with only body regions as inputs on PRCC.

method	SC		CC	
	top-1	mAP	top-1	mAP
baseline	**96.7**	**92.6**	18.7	20.4
CAL	95.9	92.4	**24.5**	**25.4**

test results of the baseline method and CAL on PRCC are shown in Tab. 6. It can be seen that CAL is slightly inferior to the baseline in the same-clothes setting but still outperforms the baseline significantly when only body regions are reserved. Hence, we argue the high responses of CAL on the body regions are mainly due to clothes-irrelevant features (body shape), while less due to clothes-relevant features (the color and texture of clothes).

6. Conclusion

In this paper, we propose Clothes-based Adversarial Loss (CAL) for clothes-changing person re-id. During training, CAL forces the backbone of the re-id model to learn clothes-irrelevant features by penalizing its predictive power w.r.t. clothes. As a result, the learned backbone can better mine the clothes-irrelevant information from the original RGB modality and is more robust against clothes changes. Extensive experiments on the newly constructed CCVID and other related datasets demonstrate that CAL consistently improves over the baseline method by a large margin. Using RGB images only, it outperforms all state-of-the-art methods on these datasets. We hope CAL will become a commonly used loss of future clothes-changing person re-id methods.

Broader impacts. The proposed method can be used in the existing person re-id methods and boost the performance of clothes-changing re-id without additional data and multi-modality inputs. It makes long-term person re-id technology more practicable in intelligent monitoring systems and may inspire more valuable and innovative studies in the future. The potential negative impact lies in that surveillance data and person re-id datasets may cause privacy breaches. Hence, the collection process of these data should inform the persons who are contained in the collection and the utilization of these data should be regulated.
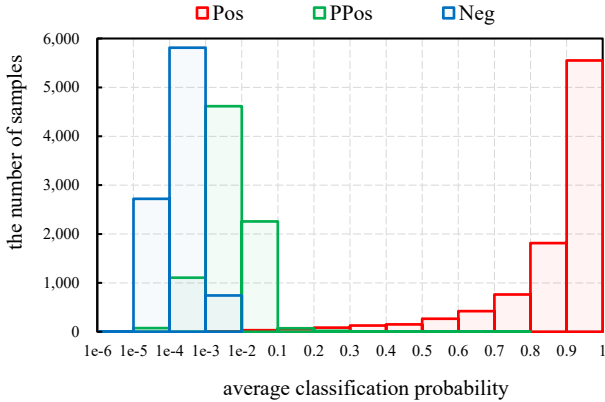
8

A. Convergence Analysis

Figure 6. The statistics of the average probability for three types of clothes classes on LTCC at the last training epoch. Best viewed in color.

To facilitate the convergence analysis, we define the classification probability of true positive clothes class (ground truth clothes class of f_i), pseudo positive clothes class (with the same identity but different clothes from f_i), and negative clothes class (with different identities from f_i) as p_{Pos}, p_{PPos}, and p_{Neg}. In the first step of optimization, minimizing $\mathcal{L}_C$ will maximize p_{Pos} and minimize p_{PPos} and p_{Neg}. In the second step, minimizing $\mathcal{L}_{CA}$ will maximize p_{Pos} and p_{PPos}, and minimize p_{Neg}. So the probability distribution of the convergence state should be $p_{Pos} \gg p_{PPos} \gg p_{Neg}$. To verify this, we count the average probability for these three types of clothes classes of all samples on LTCC at the last training epoch. As shown in Fig. 6, p_{Pos} of most samples converges to 0.6~1; p_{PPos} converges to 1e-4~0.1; and p_{Neg} converges to 1e-5~1e-2. The convergence results are consistent with our analysis.

B. Experiments on VC-Clothes

VC-Clothes [43] is a virtual dataset synthesized by GTA5. It contains 19,060 images from 512 identities and 4 cameras (scenes). Each identity has 1~3 suits of clothes and all samples of each identity captured by camera 2&3 must wear the same clothes. Hence, most current works [18, 43] report the results on the subset from camera 2&3 as the accuracy in the same-clothes (SC) setting. Besides, they report the results on the subset from camera 3&4 as the accuracy in clothes-changing (CC) setting. To make a fair comparison, we follow the settings in these works.

We compare the proposed CAL with two single-modality-based (RGB) re-id methods (*i.e.* MDLA [34] and PCB [40]) and three multi-modality-based re-id methods (*i.e.* Part-aligned [38], FSAM [18], and 3DSL [6]) on VC-Clothes in Tab. 7. It can be seen that using RGB images only, the proposed CAL outperforms the baseline and all

Table 7. Comparison with state-of-the-arts on VC-Clothes.

method	general (all cams)		SC (cam2&cam3)		CC (cam3&cam4)	
	top-1	mAP	top-1	mAP	top-1	mAP
MDLA [34]	88.9	76.8	94.3	93.9	59.2	60.8
PCB [40]	87.7	74.6	94.7	94.3	62.0	62.2
Part-aligned [38]	90.5	79.7	93.9	93.4	69.4	67.3
FSAM [18]	-	-	94.7	94.8	78.6	78.9
3DSL [6]	-	-	-	-	79.9	81.2
baseline	88.3	79.2	94.1	94.3	67.3	67.9
CAL	**92.9**	**87.2**	**95.1**	**95.3**	**81.4**	**81.7**

Table 8. The results in the same-clothes and clothes-changing settings for the data from all cameras on VC-Clothes.

method	SC (all cams)		CC (all cams)	
	top-1	mAP	top-1	mAP
baseline	94.5	93.9	74.2	66.5
CAL	**96.0**	**95.7**	**85.8**	**79.8**

Table 9. Comparison with state-of-the-art methods on LaST and DeepChange.

method	LaST		DeepChange	
	top-1	mAP	top-1	mAP
OSNet [57]	63.8	20.9	39.7	10.3
ReIDCaps [26]	-	-	39.5	11.3
BoT [33]	68.3	25.3	47.5	13.0
mAPLoss [37]	69.9	27.6	-	-
baseline	69.4	25.6	50.6	15.9
CAL	**73.7**	**28.8**	**54.0**	**19.0**

these state-of-the-art methods in general, the same-clothes, and clothes-changing settings. This comparison can demonstrate the effectiveness of CAL. Since these state-of-the-art methods do not report the accuracy in the same-clothes and clothes-changing settings on the whole dataset from all cameras, we only compare our method with the baseline in these settings in Tab. 8. The experimental results show CAL outperforms the baseline, especially in clothes-changing setting.

C. Experiments on LaST and DeepChange

LaST [37] and DeepChange [48] are two large-scale long-term person re-id datasets. LaST provides clothes labels of the training set, while DeepChange does not provide any clothes labels. We notice that the collection date of each image on DeepChange is given. Since the samples of the same person captured on different days have a high probability of wearing different clothes, we attempt to use the collection date as pseudo clothes labels to train CAL. To make a fair comparison, we set batch size as 64 following [37] and each batch contains 16 persons and 4 images for each person. On LaST, triplet loss [16] is used as the metric learning loss for both baseline and CAL (removing it will cause severe performance degradation). Finally, we report the performance in general setting to compare with

state-of-the-art methods. Especially, on DeepChange, we allow the true matches coming from the same camera but different tracklets as query following [48].

The comparisons with the baseline and state-of-the-art methods on LaST and DeepChange are shown in Tab. 9. It can be seen that the proposed CAL outperforms the baseline and these state-of-the-art methods significantly. Especially, on DeepChange, when the collection date is used as pseudo clothes label, CAL still works well. It demonstrates that the proposed method does not rely on accurate clothes annotations.

References

[1] Gunawan Ariyanto and Mark S. Nixon. Marionette mass-spring model for 3d gait biometrics. In *ICB*, 2012. 2

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009. 4

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 7

[4] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. GaitSet: Regarding gait as a set for cross-view gait recognition. In *AAAI*, 2019. 1, 2, 3, 7, 8

[5] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *ECCV*, 2020. 2

[6] 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 *CVPR*, 2021. 1, 2, 5, 6, 9

[7] Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, and Dina Katabi. Learning longterm representations for person re-identification using radio signals. In *CVPR*, 2020. 1, 2

[8] Yang Feng, Yuncheng Li, and Jiebo Luo. Learning effective gait features using lstm. In *ICPR*, 2016. 2, 3

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[10] Mengran Gou, Xikang Zhang, Angels Rates-Borras, Sadjad Asghari-Esfeden, Mario Sznaier, and Octavia Camps. Person re-identification in appearance impaired scenarios. In *BMVC*, 2016. 2

[11] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *ECCV*, 2020. 1, 2, 5, 7

[12] Xinqian Gu, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Temporal knowledge propagation for image-to-video person re-identification. In *ICCV*, 2019. 1

[13] Ju Han and Bir Bhanu. Individual recognition using gait energy image. *TPAMI*, 28(2):316–322, 2006. 1, 2

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 5

[16] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 6, 7, 9

[17] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011. 4

[18] 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 *CVPR*, 2021. 1, 2, 5, 6, 9

[19] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *ECCV*, 2020. 1, 2, 7

[20] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, 2019. 3, 5, 6, 7

[21] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *CVPR*, 2019. 2

[22] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. IAUnet: Global context-aware feature learning for person reidentification. *TNNLS*, 2020. 2

[23] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Feature completion for occluded person re-identification. *TPAMI*, 2021. 1

[24] 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. 2

[25] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing status awareness for long-term person re-identification. In *ICCV*, 2021. 5, 6

[26] 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. *TCSVT*, 30(10):3459–3471, 2020. 2, 4, 9

[27] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 5

[28] Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Xian-Sheng Hua, and Zhibo Chen. Cloth-changing person re-identification from a single image with gait prediction and regularization. *arXiv preprint arXiv:2103.15537*, 2021. 2, 5, 6

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[30] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *ICCV*, 2019. 4

[31] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 5, 6

[32] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018. 2

[33] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshops*, 2019. 9

[34] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *ICCV*, 2017. 9

[35] 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. 1, 2, 3, 4, 5, 6

[36] Zhiqiang Shen, Zhankui He, and Xiangyang Xue. Meal: Multi-model ensemble via adversarial learning. In *AAAI*, 2019. 2

[37] Xiujun Shu, Xiao Wang, Shiliang Zhang, Xianghao Zhang, Yuanqi Chen, Ge Li, and Qi Tian. Large-scale spatio-temporal person re-identification: Algorithm and benchmark. *arXiv preprint arXiv: 2105.15076*, 2021. 4, 5, 9

[38] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018. 9

[39] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, 2020. 7

[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 *ECCV*, 2018. 1, 3, 5, 6, 7, 9

[41] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 4

[42] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 2

[43] Fangbin Wan, Yang Wu, Xuelin Qian, and Yanwei Fu. When person re-identification meets changing clothes. In *CVPR Workshop*, 2020. 4, 5, 9

[44] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 2, 3, 4, 6

[45] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, 2014. 2, 4

[46] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 7

[47] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 7

[48] Peng Xu and Xiatian Zhu. Deepchange: A large long-term person re-identification benchmark with clothes change. *arXiv preprint arXiv:2105.14685*, 2021. 4, 5, 9, 10

[49] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *TPAMI*, 43(6):2029–2046, 2021. 1, 2, 3, 4, 5, 6

[50] 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. 2

[51] 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. *TMM*, 23:3562–3576, 2021. 2

[52] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *CVPR*, 2019. 1, 2, 5, 7, 8

[53] 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 *ECCV*, 2016. 4

[54] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 7

[55] 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. 2, 7

[56] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 5

[57] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, 2019. 7, 9