Cloth-Changing Person Re-identification from A Single Image with Gait Prediction and Regularization

Xin Jin[1,2]*, Tianyu He[2], Kecheng Zheng[1], Zhiheng Yin[3], Xu Shen[2], Zhen Huang[1], Ruoyu Feng[1],
Jianqiang Huang[2], Zhibo Chen[1]†, Xian-Sheng Hua[2]†
[1]University of Science and Technology of China, [2]Alibaba Cloud Computing Ltd.
[3]University of Michigan
{jinxustc,zkcys001,hz13,ustcfry}@mail.ustc.edu.cn, yzhiheng@umich.edu, chenzhibo@ustc.edu.cn
{timhe.hty,shenxu.sx,jianqiang.hjq,xiansheng.hxs}@alibaba-inc.com

Abstract

*Cloth-Changing person re-identification (CC-ReID) aims at matching the same person across different locations over a long-duration, e.g., over days, and therefore inevitably has cases of changing clothing. In this paper, we focus on handling well the CC-ReID problem under a more challenging setting, i.e., just from a single image, which enables an efficient and latency-free person identity matching for surveillance. Specifically, we introduce **G**ait recognition as an auxiliary task to drive the **I**mage **ReID** model to learn cloth-agnostic representations by leveraging personal unique and cloth-independent gait information, we name this framework as **GI-ReID**. GI-ReID adopts a two-stream architecture that consists of an image ReID-Stream and an auxiliary gait recognition stream (Gait-Stream). The Gait-Stream, that is discarded in the inference for high efficiency, acts as a regulator to encourage the ReID-Stream to capture cloth-invariant biometric motion features during the training. To get temporal continuous motion cues from a single image, we design a Gait Sequence Prediction (GSP) module for Gait-Stream to enrich gait information. Finally, a semantics consistency constraint over two streams is enforced for effective knowledge regularization. Extensive experiments on multiple image-based Cloth-Changing ReID benchmarks, e.g., LTCC, PRCC, Real28, and VC-Clothes, demonstrate that GI-ReID performs favorably against the state-of-the-art methods.*

1. Introduction

Person re-identification (ReID) aims at identifying a specific person across cameras, times, and locations. Abundant approaches have been proposed to address the challenging geometric misalignment among person images caused by diversities of human poses [43, 48, 66], camera viewpoints [24, 50, 62], and style/scales [25, 26]. These methods usually inadvertently assume that both query and gallery images of the same person have the *same clothing*. In general, they perform well on the trained short-term datasets but suffer from significant performance degradations when testing on a long-term collected ReID dataset [45,52,57,59]. Because large clothing variations occur over long-duration among these datasets, which seriously hinders the accuracy of ReID. For example, Figure 1(a) shows a realistic wanted case [1] where a suspect that captured by surveillance devices at different times/locations changed her coat from black to white, which makes ReID difficult, especially when she wears a mask and the captured images are of low quality.

Figure 1. (a) shows a realistic wanted case that a suspect changed her coat from black to white for hiding. (b) reveals that the gait of person could help ReID, especially when the identity matching meets the cloth-changing challenge (All faces in the images are masked for anonymization).

In recent years, to handle the cloth-changing ReID (CC-ReID) problem, some studies have contributed some new datasets where clothing changes are commonplace (*e.g.*,

*This work was done when he was visiting Alibaba as a research intern.
†Corresponding author.

[1]Information comes from https://www.wjr.com/2016/01/06/woman-wanted-in-southwest-detroit-bank-robbery/

Celebrities-reID [19, 21], PRCC [57], LTCC [45], Real28 and VC-Clothes [52]). They also propose some new algorithms that could learn cloth-agnostic representations for CC-ReID. For instance, Yang *et al*. [57] propose a contour-sketch-based network to overcome the moderate cloth-changing problem. Similarly, Qian *et al*. [45], Li *et al*. [32], and Hong *et al*. [17] all use body shape to tackle the CC-ReID problem. However, no matter of using a contour sketch or body shape, all these methods are prone to suffer from the estimation error problem. Because the single-view contour/shape inference (from 2D image) is extremely difficult due to the vast range of possible situations, especially when people wear thick clothes in winter. Besides, these contour-sketch-based or shape-based methods only focus on extracting *static spatial cues* from persons as extra cloth-agnostic representations, the rich *dynamic motion information* (*e.g*., gait, implied motion [28]) are often ignored.

In this paper, we explore to leverage the unique gait features that imply dynamic motion cues of a pedestrian to drive a model to learn cloth-agnostic and discriminative ReID representations. As shown in Figure 1(b), although it is hard to identify the same person when he/she wears different clothes, or to distinguish the different persons when they wear similar/same clothes, we can still leverage their unique/discriminative gaits to achieve correct identity matching. It is because that gait, as a unique biometric feature, has the superior invariance compared with other easy-changing appearance characteristics, *e.g*., face, body shape, contour [36,63]. Besides, gait can be authenticated at a long distance even with low-quality camera imaging.

Unfortunately, existing gait-related studies mainly rely on large video sequences [3, 8]. Capturing videos requires time latency and saving videos needs a large hardware storage cost, which are both undesirable for the real-time ReID applications. Even the recent work [55] first attempts to achieve gait recognition from a single image, how to leverage gait feature to handle CC-ReID problem from a single image is still under-studied and this task is more challenging due to the potential viewpoint-variations and occlusions.

In this paper, we propose a **G**ait-assisted **I**mage-based **ReID** framework, termed as GI-ReID, which could learn cloth-agnostic ReID representations from a single image with the gait feature assistance. GI-ReID consists of a main image-based ReID-Stream and an auxiliary gait recognition stream (Gait-Stream). Figure 2 shows the entire framework. The Gait-Stream aims to regularize the ReID-Stream to learn cloth-agnostic features from a single RGB image for effective CC-ReID. It is discarded in the inference for the high efficiency. Since the comprehensive gait features extraction typically needs a gait video sequence as input [3,8], we introduce a new Gait Sequence Prediction (GSP) module for Gait-Stream to approximately forecast continuous gait frames from a single input query image, which enriches the learned gait information. Finally, to encourage the main ReID-Stream's efficient learning from Gait-Stream, we further enforce a high-level Semantics Consistency (SC) constraint for the same person over two streams' features. We summarize our main contributions as follows:

- We specially aimed at handling the challenging cloth-changing issue for image ReID to promote practical applications. A Gait-assisted Image-based cloth-changing ReID (GI-ReID) framework is proposed. As a regulator, the Gait-Stream in GI-ReID can be removed in the inference without sacrificing ReID performance. This reduces the dependency on the accuracy of gait recognition, making our method computationally efficient and robust.
- A well-designed Gait Sequence Prediction (GSP) module makes our method effective in the challenging image-based ReID scenarios. And, a high-level semantics consistency (SC) constraint enables an effective regularization over two streams, enhancing the distinguishing power of ReID-Stream under the cloth-changing setting.

With the gait prediction and regularization, GI-ReID achieves a state-of-the-art performance on the image-based cloth-changing ReID. It is also general enough to be compatible with the existing ReID-specific networks, except ResNet-50 [13], we also use OSNet [69], LTCC-shape [45], and PRCC-contour [57] as our baselines for evaluation.

2. Related Work

2.1. Person Re-identification

General ReID. Without cloth-changing cases, the general ReID has achieved a great success with the deep learning. It includes exploring fine-grained pedestrian feature descriptions [9, 51, 53, 69], and addressing spatial misalignment caused by (a) different camera viewpoints [24, 49], (b) different poses [10, 43, 48], (c) semantics inconsistency [26, 65], (d) occlusion/partial-observation [14, 39, 68, 70], *etc*. These methods rely substantially on static spatial texture information. However, when person ReID meets changing clothes, the texture information is not so reliable since it changes significantly even for the same person. Compared to static texture, the gait information, as a discriminative biometric modality, is more consistent and reliable.

Cloth-Changing ReID. Considering the wider application range and greater practical value of Cloth-Changing ReID (CC-ReID), more and more studies pay their attention to solve this challenging problem. Huang *et al*. [19, 21] propose to use vector-neuron capsules [46] to perceive cloth changes of the same person. Yang *et al*. [57], Qian *et al*. [45]/ Li *et al*. [32], Yu *et al*. [59]/Wan *et al*. [52] propose to leverage contour sketch, body shape, face/hairstyle to assist ReID under the cloth-changing setting, respectively. Nevertheless, these methods usually suffer from estimation error due to the difficulty of obtaining external cues (*e.g*., body shape, face, *etc*). Besides, they also ignore the explo-

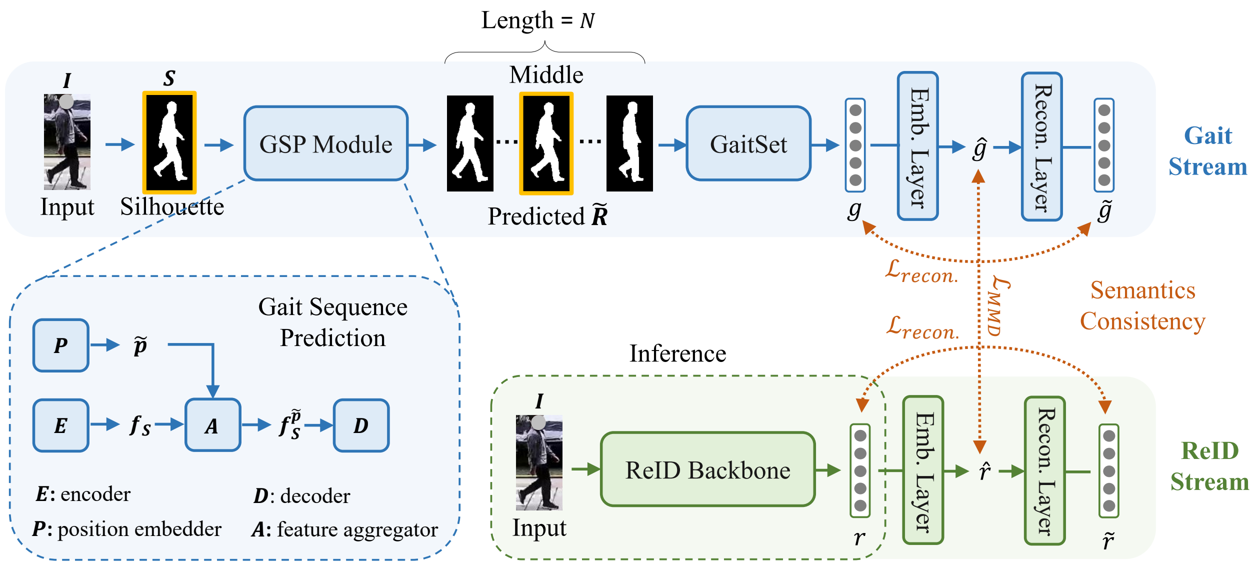

Figure 2. Overview of the proposed GI-ReID, which consists of *ReID-Stream* and *Gait-Stream*, they are jointly trained with a high-level Semantics Consistency (SC) constraint. The *Gait-Stream* plays the role of a regulator to drive *ReID-Stream* to learn cloth-agnostic representations from a single image, and it is **discarded** in the inference for computational efficiency. Gait Sequence Prediction (GSP) module aims at predicting gait frames from an image. GaitSet [3] is responsible for extracting discriminative gait features.

Table 1. Differences between Gait Recognition and CC-ReID.

Task	Gait Recognition	Cloth-Changing Person ReID
Data format	Gait Energy Image (GEI) / Sequence set of silhouette / Video sequences	**Discontinuous** RGB images across cameras
Datasets	USF, CASIA-B, OU-ISIR, OU-MVLP, *etc*.	COCAS, PRCC, LTCC, Real28, VC-Clothes, *etc*.
Unsolved problems	1) Viewing angles (*e.g*., frontal view); 2) Occlusion, body incompleteness; 3) Cluttered/complex background;	Clothes variation

ration of discriminative dynamic motion cues, like gait.

FITD [63] solves the cloth-changing ReID problem based on true motion cues of videos. Our work differs from FITD for at least three perspectives: 1). FITD uses motion information derived from dense trajectories (optical flow), which requires continuous **video sequences**. Our GI-ReID handles cloth-changing ReID from **a single image** with gait prediction and regularization, which is more challenging and practical. 2). FITD directly uses human motion cues to complete ReID, which relies on the accurate motion prediction and may suffer from estimation errors. Our GI-ReID just takes the gait recognition task as a regulator to drive the main ReID model to learn cloth-independent features, which makes our method less sensitive to gait estimation errors. 3). FITD only characterizes temporal motion patterns for ReID, ignoring other distinguishable local spatial cues, like personal belongings (*e.g*., backpacks). Our GI-ReID not only explores dynamic gait cues, but also learns from raw RGB images, leading more comprehensive features.

2.2. Gait Recognition and Prediction

Gait recognition [2, 3, 7, 8, 30, 36, 37, 40, 55] directly uses gait sequence for identity matching, which is also cloth-independent, but different from our work and cannot be directly applied into image-based cloth-changing ReID. We clarify the differences between two tasks in detail in Table 1: this paper focuses on image-based cloth-changing ReID where large viewpoint variations, occlusion, and complex environments make gait recognition failed. And, these gait sequence based methods are not optimal for the image-based CC-ReID. Thus, we just take the gait recognition as an auxiliary regularization to drive ReID model to learn cloth-agnostic representations, which makes our method robust to the recognition errors. Moreover, the gait representations can be grouped into model-based [31, 33, 42] and appearance-based [2, 3, 8]. The first one relies on human pose, while the latter relies on silhouettes. We use silhouette as gait representation for simplicity and robustness.

Gait Prediction from a single frame, or said, the field of video frame prediction (*i.e*., motion prediction) has been widely studied and achieved a great success [12, 18, 35, 41, 55], which verifies the feasibility of our work. This task is very challenging, that's why we carefully design the gait sequence prediction module while indirectly using the prediction results in a robust regularization manner to help cloth-changing ReID.

3. Proposed GI-ReID Framework

GI-ReID framework aims to fully exploit the unique human gait to handle the cloth-changing challenge of ReID just depending on a single image. Figure 2 shows the flowchart of the entire framework. Given a single person image, its silhouette (*i.e*., mask) will be first extracted as input to the Gait-Stream using semantic segmentation methods, such as PointRend [27]. With the proposed gait sequence prediction (GSP) module, we could predict a gait sequence with more comprehensive gait information, which is then fed into the subsequent recognition network (GaitSet [3]) to extract discriminative gait features. Through a high-level semantics consistency (SC) constraint, the cloth-independent Gait-Stream acts as a regulator to encourage the main ReID-Stream to capture cloth-agnostic features from a single RGB image. We discuss the details of each component in the following sections.

3.1. The Auxiliary Gait-Stream

Gait-Stream is composed of two parts: Gait Sequence Prediction (GSP) module and the pre-trained gait recognition network (GaitSet [3]). GSP is designed for gait information augmentation. Then, GaitSet extracts cloth-independent and discriminative motion feature cues from augmented gait to guide/regularize ReID-Stream's training.

Gait Sequence Prediction (GSP) Module: GSP module aims to predict a gait sequence that contains continuous gait frames. This module is related to the general video frame prediction task (*i.e.*, frame interpolation and extrapolation studies [12, 18, 35, 41]), and gait sequence prediction can be deemed as a "gait frame synthesis" process.

As shown in Figure 2, GSP is based on an auto-encoder architecture [6] with feature encoder E and decoder D. In order to reduce the prediction ambiguity and difficulty (*e.g.*, given a dangling arm, it is hard to guess whether it will rise or fall in the next frame), we manually integrate an extra prior information of **middle frame index** into the inner learned feature through a position embedder P and a feature aggregator A. Intuitively, the **middle frame index** means that the input gait silhouette corresponds to the middle result of the predicted gait sequence. Such prior knowledge aims to drive GSP module to predict the adjacent walking statuses *before and after* the current input walking status so as to reduce prediction ambiguity.

(1). Encoder. Given a silhouette input S, the encoder E aims to extract a dimension-shrinked compact feature:

$$f_S = E(S). \tag{1}$$

Specific/detailed network structures (including other components in Gait-Stream) can be found in **Supplementary**.

(2). Position Embedder and Feature Aggregator. Considering the prediction ambiguity [38], we introduce a *middle frame input principle*, which assumes that the input silhouette always corresponds to the **middle** one of the predicted gait sequence. During the GSP training, we take the gait frame in the middle position of the ground truth gait sequence as input to GSP, and use a one-dimensional vector $p \in \mathbb{R}^1$, to denote such *position label*. Given a ground truth gait sequence with N frames, the position label $p_{mid} \in \mathbb{R}^1$ of the input middle gait is defined as $p_{mid} = N//2$ which indicates the relative position relationship of input frame to the entire sequence. For convenience, we convert position label to one-hot vector to calculate loss. In formula, the position embedder P works as:

$$\widetilde{p} = P(S) \in \mathbb{R}^1, \quad \mathcal{L}_{position} = ||\widetilde{p} - p_{mid}||_2^2, \tag{2}$$

where we compare the embedded position output $\widetilde{p}$ with the ground truth p_{mid} to construct a *position* loss $\mathcal{L}_{position}$. P is to build a mapping between input and middle position.

Feature aggregator A, implemented by a fully connected layer, is inserted between the encoder and the decoder to convert the raw encoded features f_S into *middle-position-aware* features $f_S^{\widetilde{p}}$ by taking the embedded middle position information $\widetilde{p}$ into account for the following decoder, which explicitly tells the decoder that we need to predict the gait statuses *before and after* the current input middle gait status, and thus reduces prediction ambiguity for the predicted results. This feature aggregation process is formulated as:

$$f_S^{\widetilde{p}} = A([\,f_S, \widetilde{p}\,]), \tag{3}$$

where $[\cdot]$ means a simple concatenation.

(3). Decoder. We feed the aggregated feature $f_S^{\widetilde{p}}$ into the decoder D, which has a symmetrical structure to that of the encoder E, to predict the gait sequence with a pre-defined fixed number of frames N. Such process is formulated as,

$$\widetilde{R} = D(f_S^{\widetilde{p}}) \in \mathbb{R}^{N*h*w}, \quad \mathcal{L}_{pred.} = ||\widetilde{R} - GT||_2^2, \tag{4}$$

where (h, w) denotes the $(height, width)$ of predicted gait frames, same as the input silhouette image. A prediction loss $\mathcal{L}_{pred.}$ is calculated to ensure the predicted gait sequence results is consistent with the ground truth (GT).

Gait Feature Extraction: The predicted gait sequence $\widetilde{R}$ is fed into the pre-trained GaitSet [3] to learn discriminative and cloth-independent gait feature g. GaitSet is a set-based gait recognition model that takes a set of silhouettes as an input and aggregates features over frames into a set-level feature, which is formulated as $g = GaitSet(\widetilde{R})$. More details are presented in **Supplementary**.

3.2. The Main ReID-Stream

The backbone of the ReID-Stream could be any of the off-the-shelf networks, such as commonly-used ResNet-50 [13], ReID-specific PCB [51], MGN [53], and OSNet [69]. And, we use the widely-adopted classification loss [9, 51], and triplet loss with batch hard mining [15]) on the ReID feature vector r as basic optimization objectives for training. The feature r is finally used for reference.

3.3. Joint Learning of Two Streams

Due to the potential rough silhouette extraction and the gait sequence prediction errors of GSP module, it is very difficult to directly exploit the gait information alone to complete effective ReID. Experimentally, we have attempted to conduct CC-ReID with only the predicted gait sequence $\widetilde{R}$ as input, and found this scheme failed to deliver good results (see ablation study for more details). Therefore, to exploit the cloth-independent merits of the gait information while avoiding the above-mentioned issues, we propose to jointly train Gait-Stream and ReID-Stream through a high-level semantics consistency (SC) constraint, where gait characteristics is taken as a regulator to drive the cloth-agnostic feature learning of ReID-Stream. Note that the SC constraint is also not needed in the inference.

Semantics Consistency (SC) Constraint. SC constraint is essentially related to the common feature learning works, such as knowledge distillation [16], mutual learning [64], and knowledge amalgamation [58]. Our SC constraint differs from them mainly in two perspectives: 1). SC is to encourage a high-level common feature learning from two modalities (dynamic gait and static RGB image). 2). SC ensures information integrity for each stream/modality.

The details of the SC constraint are shown in Figure 2. The learned gait feature g of Gait-Stream and ReID feature r of ReID-Stream are first transformed to a common and interactable space, via an embedding layer: $\hat{r} = Emb.(r)$ and $\hat{g} = Emb.(g)$, where $\hat{r}$ and $\hat{g}$ have the same feature dimensions. Then, we enforce the transformed features $\hat{r}$ and $\hat{g}$ to be closed to each other by minimizing the Maximum Mean Discrepancy (MMD) [11]. MMD is a distance metric to measure the domain mismatch for probability distributions. We use it to measure the high-level semantics discrepancy between the transformed features $\hat{r}$ and $\hat{g}$, and minimize it to drive ReID-Stream to pay more attention to cloth-independent gait biometric. An empirical approximation to the MMD distance of $\hat{r}$ and $\hat{g}$ is simplified as follows:

$$\mathcal{L}_{MMD} = \|\mu(\hat{g}) - \mu(\hat{r})\|_2^2 + \|\sigma(\hat{g}) - \sigma(\hat{r})\|_2^2, \quad (5)$$

where $\mu(\cdot), \sigma(\cdot)$ denotes the mean, variance calculation functions for the transformed features $\hat{r}$ and $\hat{g}$.

To avoid the information lost caused by feature regularization with SC constraint, we further enforce a reconstruction penalty to ensure that the transformed features $\hat{g}$ and $\hat{r}$ could be recovered to original versions. Specifically, we reconstruct the original output features through a *Recon.* layer (implemented by FC layer): $\widetilde{r} = Recon.(\hat{r})$ and $\widetilde{g} = Recon.(\hat{g})$, and calculate the corresponding reconstruction loss as follows:

$$\mathcal{L}_{recon.} = \|\widetilde{g} - g\|_2^2 + \|\widetilde{r} - r\|_2^2. \quad (6)$$

Training Pipeline. The whole training process of the proposed GI-ReID consists of three stages: 1). Pre-training GaitSet [3] for gait feature extraction. 2). Joint Training for the proposed gait sequence prediction (GSP) module and GaitSet in Gait-Stream on gait-related datasets. 3). Joint Training for Gait-Stream and ReID-Stream on CC-ReID-related datasets. More details are provided in **Supplementary**, including pseudo code and loss balance strategy.

4. Experiment

4.1. Datasets, Metric and Experimental Setups

Datasets Details. We use four recent cloth-changing ReID datasets Real28 [52], VC-Clothes [52], LTCC [45], PRCC [57], and one general video ReID dataset MARS [67] (to highlight the difficulty and necessity of image-based CC-ReID) to perform experiments. Table 2 gives a brief information and comparison of these ReID datasets. More detailed introductions can be found in **Supplementary**.

Table 2. Brief introduction and comparison of datasets.

	MARS	Real28	VC-Clothes	LTCC	PRCC
Category	Video	Image	Image	Image	Image
Photo Style	Real	Real	Synthetic	Real	Real
Scale	Large	Small	Large	Large	Large
Cloth Change	No	Yes	Yes	Yes	Yes
Identities	1,261	28	512	152	221
Samples	20,715	4,324	19,060	17,138	33,698
Cameras	6	4	4	N/A	3
Usage	Train&Test	Test	Train&Test	Train&Test	Train&Test

Evaluation Metrics. We use the cumulative matching characteristics (CMC) at Rank-1/-10/-20, and mean average precision (mAP) to evaluate the performance.

Experimental Setups. We build **three** kinds of different experiment settings to comprehensively validate the effectiveness of gait biometric for person ReID, and also validate the rationality/superiority of the proposed gait prediction and regularization in our GI-ReID framework: (1) Real Cloth-Changing Image ReID, (2) General Video ReID, and (3) Imitated Cloth-Changing Video ReID. In the main manuscripts, to save space and highlight core contributions of our paper, we only present the results related to the most challenging setting of (1) Real Cloth-Changing Image ReID. The rest results about (2)(3) are in **Supplementary**.

For (1) Real Cloth-Changing Image ReID, we employ real image-based cloth-changing datasets Real28 [52], VC-Clothes [52], LTCC [45], and PRCC [57] for experiments to validate the effectiveness of GSP module, SC constraint, and also compare our GI-ReID with SOTA cloth-changing ReID methods. In this setting, GSP module and GaitSet are both first pre-trained on gait-specific datasets CASIA-B [3] and then fine-tuned on the CC-ReID datasets with the SC constraint $\mathcal{L}_{MMD}$&$\mathcal{L}_{recon.}$ and the ReID supervisions. ResNet-50 [13], OSNet [69], LTCC-shape [45], and PRCC-contour [57] are taken as ReID backbone for comparisons.

4.2. Ablation Study

Baseline means the model that only ingests RGB images.

Table 3. Performance (%) comparison on the real image-based cloth-changing datasets Real28, VC-Clothes, LTCC. GS-GSP means Gait-Stream (GS) with gait sequence prediction (GSP) module. The ReID backbone is ResNet-50. 'Standard' is the setting where the images in the test set with the same identity and camera view are discarded when computing mAP/Rank-1 [45].

Methods	Real28		VC-Clothes		LTCC (Standard)	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Baseline	4.1	6.7	49.1	53.7	23.2	55.1
+ GS (concat)	6.8	7.9	52.3	58.9	26.5	60.0
+ GS-GSP (concat)	10.1	10.8	**59.0**	63.7	28.8	**64.5**
+ GS-GSP + SC (ours)	**10.4**	**11.1**	57.8	**64.5**	**29.4**	63.2

Results of Real Cloth-Changing Image ReID. We conduct ablation experiments on three cloth-changing datasets Real28, VC-Clothes, and LTCC. Real28 is too small for training, so we train model on VC-Clothes and only test on Real28 [52]. In Table 3, we see that 1) All Gait-Stream (GS) related schemes achieve obvious gains (over 2.7% in mAP) over *Baseline*, which demonstrates the effectiveness of using gait to handle cloth-changing issue. 2) With the well-designed GSP module, *Baseline+ GS-GSP (concat)* outperforms the ablated scheme *Baseline+GS (concat)* by 3.3%/6.7%/2.3% in mAP on Real28/VC-Clothes/LTCC, which demonstrates the effectiveness of gait sequence prediction (GSP) on gait information augmentation. Note that *Baseline+GS (concat)* just uses Gait-Stream (GS) but removes GSP, where we duplicate the only available single person silhouette as input to GaitSet. 3) Semantics consistency (SC) performs well in the cloth-changing settings, it helps our scheme GI-ReID achieve the best performance on the most evaluation cases while saving computational cost by discarding Gait-Stream in the inference.

Table 4. Performance (%) comparison on the cloth-changing dataset LTCC. Such experiment aims to show that our GI-ReID can bring gains because of the exploration of gait information, rather than simply introducing silhouettes (*i.e*., human masks). The ReID backbone is ResNet-50.

Methods	LTCC (Cloth-Changing)	
	mAP	Rank-1
Baseline	8.10	19.58
Silhouette-ReID	7.04	17.92
GI-ReID (ours)	**10.38**	**23.72**

Improvement Comes From Gait Prediction, Not Silhouettes Usage. We believe that our GI-ReID could successfully address the cloth-changing ReID problem from a single image is indeed because it effectively leverages the gait prediction, instead of the introduction of human silhouettes (*i.e*., masks). To prove that, we additionally design a scheme of *Silhouette-ReID* that directly takes the person RGB-Silhouette pair as input to ReID model (following [4, 47]), and compare it with our GI-ReID on the cloth-changing ReID dataset LTCC. ResNet-50 is taken as ReID backbone for all schemes for comparison fairness. As shown in Table 4, we found that *Silhouette-ReID* is even inferior to the baseline scheme *Baseline (ResNet-50)* by 1.06% in mAP under the cloth-changing setting. We analyze that directly using silhouette to remove the background clutters in pixel-level will make ReID model pay more attention on the foreground objects' appearance/clothes color information, which is unexpected and unreliable for cloth-changing ReID, and thus leads to a performance drop.

Study on Directly Using Gait Recognition Methods for Cloth-Changing ReID. As we have discussed in the related work, directly using the algorithms of gait recognition for solving cloth-changing ReID problem is not optimal, especially in the image-based CC-ReID scenarios. Experimentally, we compare the proposed GI-ReID with two popular pure gait recognition works, GaitSet [3] and PA-GCR [55]. GaitSet needs a set/sequence of person silhouettes as input, but recently-released cloth-changing ReID datasets are image datasets that lack of continuous frames for the same person. Thus, we duplicate the only available single one person silhouette to a set as input to approximately apply GaitSet into image-based CC-ReID task. As shown in Table 5, these pure gait recognition works of GaitSet [3] and PA-GCR [55] are both inferior to the baseline scheme *Baseline (ResNet-50)* in mAP under the cloth-changing setting, which indicates that simply using gait biometric for person matching can not work well for cloth-changing ReID, our gait prediction and regularization idea performs better for handling CC-ReID, especially for the image-based CC-ReID.

Table 5. Performance (%) comparison on the cloth-changing dataset LTCC. Such experiment aims to show that these pure gait recognition works can not work well for cloth-changing ReID. The ReID backbone is ResNet-50.

Methods	LTCC (Cloth-Changing)	
	mAP	Rank-1
Baseline	8.10	19.58
GaitSet [3]	2.14	7.22
PA-GCR [55]	3.36	9.01
GI-ReID (ours)	**10.38**	**23.72**

4.3. Design Choices in Our GI-ReID Framework

We study the different design choices in our GI-ReID framework. We train and test model on the real large-scale cloth-changing ReID dataset LTCC [45].

Influence of the Length N of Predicted Gait Sequence. As shown in Eq-(4) of Sec. 3.1, the output of GSP $\widetilde{R} \in \mathbb{R}^{N*h*w}$ is a sequence with N predicted gait frames. We study the influence of length N w.r.t the ReID performance. Table 6a shows that when $N = 8$, our GI-ReID gets the best performance, achieving a good trade-off between gait prediction error and gait information augmentation.

Is 'Middle Frame Input Principle' Necessary? As described in Eq-(2) of GSP in Sec. 3.1, we employ a position embedder P and a feature aggregator A to set up a *middle frame input principle* to reduce the gait prediction am-

Table 6. Study on the different design choices in the (a)(b) GSP module, and (c) SC constraint of our GI-ReID framework. 'Cloth-Changing' setting means that the images with same identity, camera view and clothes are discarded during the testing.

(a) Study on the gait prediction length N.

Methods	LTCC			
	Standard		Cloth-Changing	
	mAP	Rank-1	mAP	Rank-1
Baseline	23.2	55.1	8.1	19.6
N=4	26.9	59.2	8.9	21.7
N=6	28.2	61.9	9.8	22.6
N=8 (ours)	**29.4**	**63.2**	**10.4**	**23.7**
N=10	28.4	63.1	10.4	22.8
N=12	27.7	60.8	10.0	22.5

(b) Study on the input gait position p in GSP.

Methods	LTCC			
	Standard		Cloth-Changing	
	mAP	Rank-1	mAP	Rank-1
Baseline	23.2	55.1	8.1	19.6
Arb.	27.1	59.5	9.2	20.5
BEGN	28.4	61.2	9.8	22.0
END	28.1	61.5	9.5	22.4
Mid. (ours)	**29.4**	**63.2**	**10.4**	**23.7**

(c) Study on the used losses in SC constraint.

Methods	LTCC			
	Standard		Cloth-Changing	
	mAP	Rank-1	mAP	Rank-1
Baseline	23.2	55.1	8.1	19.6
w/ $\mathcal{L}_{MSE}$	27.5	61.0	9.0	21.4
w/o $\mathcal{L}_{recon.}$	28.3	62.7	9.6	22.9
ours	**29.4**	**63.2**	**10.4**	**23.7**

Table 7. Study on the different ReID inference strategies.

Methods	LTCC			
	Standard		Cloth-Changing	
	mAP	Rank-1	mAP	Rank-1
Baseline	23.2	55.1	8.1	19.6
$\widetilde{R}$	8.6	21.1	4.3	9.9
$\hat{r}+\hat{g}$	**29.8**	**64.0**	**10.9**	**24.4**
$\widetilde{r}+\widetilde{g}$	28.9	63.2	9.7	23.1
$\widetilde{r}$	28.1	60.8	9.1	21.3
r (ours)	29.4	63.2	10.4	23.7

biguity and difficulty. Here we compare several schemes to show the necessity of such design. ***Arb.***: we remove position embedder P, feature aggregator A, position loss $\mathcal{L}_{position}$ for GSP, and take the gait silhouette at *arbitrary* position as input for training. ***BEGN*** and ***END***: we respectively take the gait stance at the *beginning* and the *end* position as input to predict gait sequence during the GSP training. In Table 6b, the scheme ***Mid. (ours)*** that uses the gait frame at middle position for gait sequence prediction achieves the best performance, outperforming *Arb.* by 2.3% in mAP in the standard setting, which reveals that predicting the gait statuses **before and after** the input middle gait status indeed could reduce prediction difficulty/ambiguity.

Why Use MMD for Regularization? For the SC constraint, we shrink the gap between the embeded ReID vector $\hat{r}$ and gait vector $\hat{g}$ by minimizing MMD through $\mathcal{L}_{MMD}$. We study this design in Table 6c and find that when replacing $\mathcal{L}_{MMD}$ with $\mathcal{L}_{MSE}$, the performance of *w/* $\mathcal{L}_{MSE}$ drops nearly 2.0% in mAP. That's because MMD loss is a distribution-level constraint and could better enforce the high-level semantics consistency between dynamic motion gait features and static spatial ReID features. MSE loss is an element-wise constraint, and not so suitable to coordinate two modalities of motion gait and RGB feature.

Is Reconstruction Penalty Necessary? When removing $\mathcal{L}_{recon.}$ in Eq-(6), as shown in Table 6c, the shceme *w/o* $\mathcal{L}_{recon.}$ is inferior to ours by 1.1%/0.8% in mAP in the two settings, which demonstrates that avoiding information lost caused by feature regularization could enhance the final ReID performance of our GI-ReID framework.

Which One for ReID Inference? We compare several cases of using (1) predicted gait sequence $\widetilde{R}$, (2) aligned features fusion $\hat{r}+\hat{g}$, (3) reconstructed features fusion $\widetilde{r}+\widetilde{g}$, and (4) reconstructed ReID vector $\widetilde{r}$ for ReID inference. Table 7 shows that 1) Directly using the predicted gait sequence $\widetilde{R}$ for CC-ReID failed to get satisfactory results, this also indicates that these gait recognition works [3, 7, 8] are not optimal for CC-ReID. 2) Using the well-aligned features fusion $\hat{r}+\hat{g}$ achieves the best performance, outperforming ours by 0.4%/0.5% in mAP in the two settings, but this scheme still needs Gait-Stream in the inference. 3) Using the reconstructed ReID vector $\widetilde{r}$ for inference suffers from information lost and is inferior to ours by 1.3% in mAP in the both two settings. 4) Our scheme that using the regularized ReID vector r achieves the second best performance while saving the computation costs brought by Gait-Stream.

Figure 3. Six predicted gait sequences *vs.* realistic gait samples.

4.4. More Analysis, Visualization and Insights

To further prove that the proposed gait sequence prediction (GSP) module can actually predict unique human motion features, and the achieved improvements of GI-ReID indeed come from gait information, not from using additional gait-related datasets or person silhouette images, here we provide more analysis and visualization results. For example, when testing the gait-based recognition performance using GaitSet [3] on the predicted human gait sequences $\widetilde{R}$ generated by GSP module (see **Supplementary** for details), it can achieve a competitive 62.4% in Rank-1 on CASIA-B.

Gait Sequence Prediction Visualization. Figure 3 further shows 6 groups of gait prediction results (left) and 2 groups of realistic gait samples from CASIA-B dataset [3] (right). Compared to the realistic gait samples, the predicted gait results (*i.e.*, the outputs of GSP) have the reasonable continuous movements, *e.g.*, swing arms and opening/closing legs. The gait-stream could learn the discriminative dynamic clues from these predicted gait results, like *the walking stride*, *the left-right swinging range of arms*, *the opening/closing angle of legs*, *etc*, (see red circles in Figure 3).

Feature Map Visualization. To better understand how our GI-ReID works, we visualize the intermediate activation feature maps of *Baseline* and our GI-ReID for comparison

Table 8. Performance (%) comparisons of our GI-ReID and other competitors on the cloth-changing datasets LTCC [45] and PRCC [57]. '†' means that only identities with clothes changing are used for training. More results are presented in **Supplementary**.

(a) Comparison results on LTCC.

Methods	Standard		Cloth-changing		Standard†		Cloth-changing†	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
LOMO [34] + NullSpace [61]	34.83	11.92	16.45	6.29	27.59	9.43	13.37	5.34
ResNet-50 + Face [56]	60.44	25.42	22.10	9.44	55.37	22.23	20.68	8.99
PCB [51]	65.11	30.60	23.52	10.03	59.22	26.61	21.93	8.81
HACNN [29]	60.24	26.71	21.59	9.25	57.12	23.48	20.81	8.27
MuDeep [44]	61.86	27.52	23.53	10.23	56.99	24.10	18.66	8.76
Baseline (ResNet-50)	55.14	23.21	19.58	8.10	54.27	21.98	19.14	7.74
GI-ReID (ResNet-50, ours)	63.21	29.44	23.72	10.38	61.39	27.88	22.59	9.87
Baseline (OSNet)	66.07	31.18	23.43	10.56	61.22	27.41	22.97	9.74
GI-ReID (OSNet, ours)	**73.59**	**36.07**	28.11	13.17	**66.94**	**33.04**	**26.71**	12.69
Baseline (LTCC-shape [45])	–	–	26.15	12.40	–	–	25.15	11.67
LTCC-shape + Gait-Stream (ours)	–	–	**28.86**	**14.19**	–	–	26.41	**13.26**

(b) Comparison results on PRCC.

Methods	Cross-clothes		
	Rank-1	Rank-10	Rank-20
Shape [1]	11.48	38.66	53.21
LNSCT [54]	15.33	53.87	67.12
HACNN [29]	21.81	59.47	67.45
PCB [51]	22.86	61.24	78.27
SketchNet [60]	17.89	43.70	58.62
Deformable [5]	25.98	71.67	85.31
STN [22]	27.47	69.53	83.22
RCSANet [20]	31.60	–	–
PRCC-contour [57]	34.38	77.30	88.05
+ Gait-Stream (ours)	36.19	79.93	91,67
Baseline (ResNet-50)	22.23	61.08	76.44
GI-ReID (ResNet-50)	33.26	75.09	87.44
Baseline (OSNet)	28.70	72.34	85.89
GI-ReID (OSNet)	**37.55**	**82.25**	**93.76**

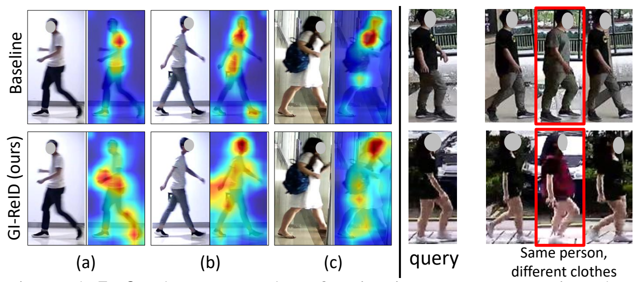

Figure 4. **Left:** three examples of activation maps comparison between baseline and our GI-ReID, which shows GI-ReID not only focuses on people's clothes, but also pay attention to the holistic human gait and local face; **Right:** Top-3 ranking list of GI-ReID for two query images on Real28. GI-ReID could identify the same person with different clothes based on the assistance of gait.

following [23,25,69]. On the left of Figure 4, we show three examples of activation maps on Real28, and we observe that the feature maps of *Baseline* have high response mainly on person's clothes. In contrast, the activation features of our GI-ReID not only have high response on person's clothes, but also cover the holistic human body structure (gait) and local face information (robust to cloth changing).

4.5. Comparison with State-of-the-Arts

The study on cloth-changing ReID is relatively rare [19, 21, 32, 45, 52, 57, 59], and most of them have not released source codes, even the dataset [59]. We compare our GI-ReID with multiple general ReID algorithms, including PCB [51], HACNN [29], MuDeep [44], and specific cloth-changing ReID methods LTCC-shape [45], PRCC-contour [57], RCSANet [20]. In Table 8, we observe that 1) Thanks to the cloth-independent gait characteristics, our scheme GI-ReID (OSNet) achieves the best performance on PRCC, outperforming the second best PRCC-contour [57] by 3.17% in Rank-1 in the cross-clothes setting. 2) The proposed Gait-Stream, as a kind of regularization, could benefit other methods, *e.g.*, LTCC-shape [45]. We find that the scheme of *LTCC-shape + Gait-Stream* could further obtain 1.79%/1.59% gain in mAP on LTCC. 3) For two cloth-changing settings of LTCC, our scheme GI-ReID (ResNet-50) both achieve obvious gains (2.28%/2.13% in mAP) over the Baseline (ResNet-50), which totally-fair results indicate that GI-ReID could handle clothes changes and learn identity-relevant features. 4) Our method is compatible with existing ReID networks, *e.g.*, built upon the strong ReID-specific network OSNet [69], GI-ReID (OSNet) further achieves gains than GI-ReID (ResNet-50).

4.6. Failure Cases Analysis

Due to the large difference on the capture viewpoints and environments between gait and ReID training data, the predicted gait results of GSP are not always perfect (see **Supplementary**) when occlusion, partial, multi-person, *etc*, existed in the person images, which may affect the CC-ReID performance. That is why we indirectly use gait predictions in a knowledge regularization manner, which makes GI-ReID robust and not sensitive to these failure cases.

5. Conclusion

In this paper, we propose to utilize human unique gait to address the cloth-changing ReID problem from a single image. A novel gait-involved two-stream framework GI-ReID is introduced, which takes gait as a regulator with a Gait-Stream (discarded in the inference), to encourage the cloth-agnostic representation learning of image-based ReID-Stream. To facilitate the gait utilization, a gait sequence prediction (GSP) module and a high-level semantics consistency (SC) constraint are further designed. Extensive experiments on multiple CC-ReID benchmarks demonstrate the effectiveness and superiority of GI-ReID.

6. Acknowledgements

This work was (partially) supported by the National Key R&D Program of China under Grant 2020AAA0103902, Alibaba Innovative Research (AIR) program, and NSFC under Grant U1908209, 61632001, 62021001.

References

[1] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE TPAMI*, 24(4):509–522, 2002. 8

[2] Cassandra Carley, Ergys Ristani, and Carlo Tomasi. Person re-identification from gait using an autocorrelation network. In *CVPRW*, pages 0–0, 2019. 3

[3] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, volume 33, pages 8126–8133, 2019. 2, 3, 4, 5, 6, 7

[4] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *ECCV*, pages 734–750, 2018. 6

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 8

[6] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 4

[7] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, and Ahmed Bouridane. Gait recognition for person re-identification. *The Journal of Supercomputing*, pages 1–20, 2020. 3, 7

[8] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *CVPR*, June 2020. 2, 3, 7

[9] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, volume 33, pages 8295–8302, 2019. 2, 4

[10] Yixiao Ge, Zhuowan Li, Haiyu Zhao, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NeurIPS*, 2018. 2

[11] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 5

[12] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *CVPR*, pages 11474–11484, 2020. 3, 4

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 4, 5

[14] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *CVPR*, 2018. 2

[15] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 4

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5

[17] 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*, pages 10513–10522, 2021. 2

[18] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *NeurIPS*, pages 517–526, 2018. 3, 4

[19] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In *IJCNN*, pages 1–8. IEEE, 2019. 2, 8

[20] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing status awareness for long-term person re-identification. In *ICCV*, pages 11895–11904, 2021. 8

[21] 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*, 2019. 2, 8

[22] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015. 8

[23] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Global distance-distributions separation for unsupervised person re-identification. *ECCV*, 2020. 8

[24] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. In *AAAI*, 2020. 1, 2

[25] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, pages 3143–3152, 2020. 1, 8

[26] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *AAAI*, 2020. 1, 2

[27] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, pages 9799–9808, 2020. 4

[28] Zoe Kourtzi and Nancy Kanwisher. Activation in human mt/mst by static images with implied motion. *Journal of cognitive neuroscience*, 12(1):48–55, 2000. 2

[29] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 8

[30] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. In *CVPR*, pages 13309–13319, 2020. 3

[31] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end model-based gait recognition. In *ACCV*, 2020. 3

[32] Yu-Jhe Li, Zhengyi Luo, Xinshuo Weng, and Kris M Kitani. Learning shape representations for clothing variations in person re-identification. *arXiv preprint arXiv:2003.07340*, 2020. 2, 8

[33] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020. 3

[34] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 8

[35] Yu-Lun Liu, Yi-Tung Liao, Yen-Yu Lin, and Yung-Yu Chuang. Deep video frame interpolation using cyclic frame generation. In *AAAI*, volume 33, pages 8794–8802, 2019. 3, 4

[36] Zheng Liu, Zhaoxiang Zhang, Qiang Wu, and Yunhong Wang. Enhancing person re-identification by integrating gait biometric. *Neurocomputing*, 168:1144–1156, 2015. 2, 3

[37] Yasushi Makihara, Atsuyuki Suzuki, Daigo Muramatsu, Xiang Li, and Yasushi Yagi. Joint intensity and spatial metric learning for robust gait recognition. In *CVPR*, pages 5705–5715, 2017. 3

[38] Simone Meyer, Oliver Wang, Henning Zimmer, Max Grosse, and Alexander Sorkine-Hornung. Phase-based frame interpolation for video. In *CVPR*, pages 1410–1418, 2015. 4

[39] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, pages 542–551, 2019. 2

[40] Daigo Muramatsu, Akira Shiraishi, Yasushi Makihara, Md Zasim Uddin, and Yasushi Yagi. Gait-based person recognition using arbitrary view transformation model. *TIP*, 24(1):140–154, 2014. 3

[41] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *CVPR*, pages 1701–1710, 2018. 3, 4

[42] Mark Nixon et al. Model-based gait recognition. 2009. 3

[43] Xuelin Qian, Yanwei Fu, Wenxuan Wang, et al. Pose-normalized image generation for person re-identification. In *ECCV*, 2018. 1, 2

[44] Xuelin Qian, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Leader-based multi-scale attention deep architecture for person re-identification. *TPAMI*, 2019. 8

[45] 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. *WACV*, 2020. 1, 2, 5, 6, 8

[46] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *NeurIPS*, pages 3856–3866, 2017. 2

[47] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018. 6

[48] Chi Su, Jianing Li, Shiliang Zhang, et al. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017. 1, 2

[49] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. *arXiv preprint arXiv:1812.02162*, 2018. 2

[50] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*, pages 608–617, 2019. 1

[51] 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*, pages 480–496, 2018. 2, 4, 8

[52] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *CVPRW*, pages 830–831, 2020. 1, 2, 5, 6, 8

[53] Guanshuo Wang, Yufeng Yuan, Xiong Chen, et al. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, pages 274–282, 2018. 2, 4

[54] Xiaohua Xie, Jianhuang Lai, and Wei-Shi Zheng. Extraction of illumination invariant facial features from a single image using nonsubsampled contourlet transform. *Pattern Recognition*, 43(12):4177–4189, 2010. 8

[55] Chi Xu, Yasushi Makihara, Xiang Li, Yasushi Yagi, and Jianfeng Lu. Gait recognition from a single image using a phase-aware gait cycle reconstruction network. eccv 2020. 2, 3, 6

[56] Jia Xue, Zibo Meng, Karthik Katipally, Haibo Wang, and Kees van Zon. Clothing change aware person identification. In *CVPRW*, pages 2112–2120, 2018. 8

[57] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *TPAMI*, 2019. 1, 2, 5, 8

[58] Jingwen Ye, Yixin Ji, Xinchao Wang, Kairi Ou, Dapeng Tao, and Mingli Song. Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more. In *CVPR*, pages 2829–2838, 2019. 5

[59] 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*, pages 3400–3409, 2020. 1, 2, 8

[60] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. Sketchnet: Sketch classification with web images. In *CVPR*, pages 1105–1113, 2016. 8

[61] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identificatio. In *CVPR*, 2016. 8

[62] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016. 1

[63] Peng Zhang, Qiang Wu, Jingsong Xu, and Jian Zhang. Long-term person re-identification using true motion from videos. In *WACV*, pages 494–502. IEEE, 2018. 2, 3

[64] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, 2018. 5

[65] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, et al. Densely semantically aligned person re-identification. In *CVPR*, 2019. 2

[66] Haiyu Zhao, Maoqing Tian, Shuyang Sun, et al. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017. 1

[67] 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*, pages 868–884. Springer, 2016. 5

[68] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *ICCV*, 2015. 2

[69] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, et al. Omni-scale feature learning for person re-identification. *ICCV*, 2019. 2, 4, 5, 8

[70] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *ICME*, pages 1–6. IEEE, 2018. 2