Supplementary: 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

Contents

1. Detailed Network Structures of GI-ReID

GI-ReID, as a image-based cloth-changing ReID framework, with gait information as assistance, consists of an auxiliary Gait-Stream and a mainstream ReID-Stream. ReID-Stream can be arbitrary commonly-used network architectures, such as ResNet [6], and also can be some ReID-specific network architectures, such as PCB [23], OSNet [33]. Thus, in this section, we mainly introduce/describe the detailed network architecture of Gait-Stream which contains two key parts, GSP module for gait information prediction/augmentation and GaitSet [2] for gait features extraction.

Architecture of Gait Sequence Prediction (GSP) Module: The proposed GSP module consists of a feature encoder E, a decoder D, a position embedder P, and a feature aggregator A.

(1). Encoder E. The encoder E is a CNN with four Conv. layers (filter size = 4×4 and stride = 2). The number of filters is increased from 64 $\rightarrow$ 512. Each Conv. layer is followed by a batch-normalization (BN) layer [9] and a rectified linear unit (ReLU) activation function [18]. In the end, a 100-dimensional feature is obtained through a fully connected (FC) layer.

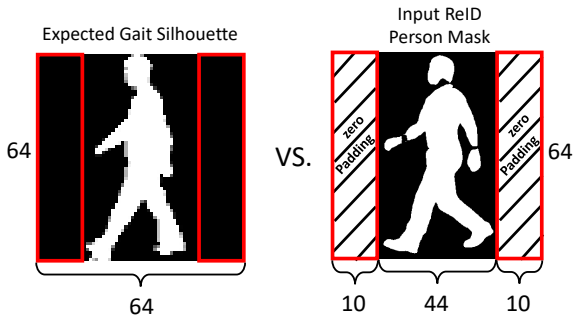

Figure 1. We apply "resize+zero_padding" in the person masks (right) when fine-tuning GSP module on the ReID-specific datasets, because the raw gait training data (left) typically have the height-width ratio of (1:1), which is important/necessary for training GSP to get satisfactory gait prediction results.

Note that, when pre-training GSP module on the gait-specific datasets following [2], the input gait silhouette of encoder E has a size of $1 \times 64 \times 64$ (height-width ratio is 1:1). We use CASIA-B [2] as training dataset. On the ReID-specific datasets, since the input person images usually have a height-width ratio of 2:1 (*e.g.*, height-256, width-128), we need leverage an operation of "resize+zero_padding" to handle such training data gap, which is pivotal for GSP's accurate gait sequence prediction. For better understanding, we vividly visualize such process in Figure 1.

(2). Position Embedder P and Feature Aggregator A. To reduce the gait prediction ambiguity and difficulty of GSP, a position embedder P and a feature aggregator A are introduced to integrate the prior information of input middle frame index into the prediction process of GSP. The position embedder P has a similar structure to that of the encoder E, but with one more FC layer to regress the 1D position label

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

$\widetilde{p}$. The feature aggregator A 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. With respect to the architecture of A, it is implemented only by a FC layer, which aims to regress to the aggregated 100-dimension feature $f_S^{\widetilde{p}} \in \mathbb{R}^{100}$ from the **101-dimension** concatenated vector of the raw encoded feature $f_S \in \mathbb{R}^{100}$ and the embedded middle position prior vector $\widetilde{p} \in \mathbb{R}^1$.

(3). Decoder D. The structure of the decoder D is symmetrical to that of the encoder E. A FC layer along with reshaping is first employed to convert the input 100D feature into the same size as the last feature output of the encoder E, and then four DeConv. layers are used for up-sampling. A sigmoid activation function is applied in the end, and outputs the gait predictions with a size of $N \times 64 \times 64$, where each channel indicates a predicted gait frame of final results.

Architecture of GaitSet: GaitSet [2] is a classic set-based gait recognition network, which takes a set of silhouettes/gait frames as input. After obtaining features from each input silhouette independently using a CNN, *set pooling* is applied to merge features over frames into a set-level feature. This set-level feature is then used for discrimination learning via *horizontal pyramid mapping* (HPM), which aims to extract features of different spatial locations on different scales. We recommend seeing more details from their original paper [2].

2. Training Details of our GI-ReID

Phase-1: Pre-training for GaitSet. The input is a set of aligned silhouettes in size of 64×44. The silhouettes are directly provided by the datasets and are aligned based on methods in [24]. The set cardinality in the training is set to be 30. Adam is chosen as an optimizer. The number of scales S in HPM is set as 5. The margin in separate triplet loss $\mathcal{L}_{tri}^{sep}$ [2] is set as 0.2. The mini-batch is composed of $P = 16$ and $N = 8$ (P, N respectively mean the number of person identities and input gait frames). We set the number of channels in $C1$ and $C2$ as 32, in $C3$ and $C4$ as 64 and in $C5$ and $C6$ as 128 (following [2]). The learning rate is set to be 1×10^{-4}, and the model is trained for 80 epochs.

Phase-2: Joint Training for GSP module and GaitSet. After pre-training GaitSet, we jointly train the proposed gait sequence prediction (GSP) module and GaitSet for Gait-Stream. Specifically, during the joint-training, we also reuse CASIA-B dataset for effective gait prediction training. Following [7], a batch is formed by first randomly sampling P identities. For each identity, we sample N continuous gait frames as the ground-truth gait sequence. Then the batch size is $B = P \times N$. We set $P = 4$ and $N = 8$ (*i.e.*, batch size $B = P \times N = 32$. As presented in the main manuscript, we use the middle one of the ground-truth gait sequence (*i.e.*, the fourth one when $N = 8$) as input for GSP training. We first optimize GSP with the proposed position loss $\mathcal{L}_{position}$ and prediction loss $\mathcal{L}_{pred}$ (loss balance is set as 1:1) for 80 epochs, which enables GSP to output reasonable predicted gait sequence results. We train GSP with Adam optimizer [11] with a initial learning rate of 5×10^{-4}. We optimize the Adam optimizer with a weight decay of 1×10^{-4}. The learning rate is decayed by a factor of 0.1 at 40 epoch.

Algorithm 1 Training Process of GI-ReID

Input: gait dataset $\mathcal{G}$ (*e.g.*, CASIA-B [2]), ReID dataset $\mathcal{R}$ (*e.g.*, LTCC [21]). Learning rate is simply denoted as η. The entire GI-ReID framework consists of GSP module GSP_θ, GaitSet (GS) GS_ϕ, SC constraints related FC layers SC_ψ, and ReID-Stream backbone $ReID_\omega$.
Output: inference ReID vector r.
Phase-1: Pre-training for GaitSet
for $epoch = 1$ **to** 80 **do**
  Sample $P \times N = 16 \times 8$ samples from gait training set $\mathcal{G}$.
  $\mathcal{L}_{total} = \mathcal{L}_{tri}^{sep}$ // *Use the separate triplet loss as objective function [2].*
  $\phi = \phi - \eta\nabla_\phi \mathcal{L}_{tri}^{sep}$ // *Update GaitSet (GS) GS_ϕ.*
end for
Phase-2: Joint Training for GSP module and GaitSet
for $epoch = 1$ **to** 80 **do**
  Sample $P \times N = 4 \times 8$ samples from gait training set $\mathcal{G}$.
  $\mathcal{L}_{total} = \mathcal{L}_{position} + \mathcal{L}_{pred}$ // *Use the proposed position loss and prediction loss as objective functions.*
  $\theta = \theta - \eta\nabla_\theta \mathcal{L}_{total}$ // *Warm up GSP module GSP_θ.*
end for
for $epoch = 1$ **to** 160 **do**
  Sample $P \times N = 4 \times 8$ samples from gait training set $\mathcal{G}$.
  $\mathcal{L}_{total} = \mathcal{L}_{position} + \mathcal{L}_{pred} + \mathcal{L}_{tri}^{sep}$ // *Use the position loss, prediction loss , and separate triplet loss as objective functions.*
  $(\theta, \phi) = (\theta, \phi) - \eta\nabla_{(\theta,\phi)} \mathcal{L}_{total}$ // *Jointly update GSP module GSP_θ and GaitSet (GS) GS_ϕ.*
end for
Phase-3: Joint Training for Gait-Stream and ReID-Stream
for $epoch = 1$ **to** 240 **do**
  Sample $P \times N = 10 \times 8$ samples from ReID training set $\mathcal{R}$.
  $\mathcal{L}_{total} = 0.1 * \mathcal{L}_{position} + 0.1 * \mathcal{L}_{pred} + 0.1 * \mathcal{L}_{tri}^{sep} + \mathcal{L}_{cla} + \mathcal{L}_{tri}^{HM} + 0.5 * \mathcal{L}_{MMD} + 0.5 * \mathcal{L}_{recon.}$ // *Total objective functions consists of the position loss, prediction loss, separate triplet loss (for Gait-Stream), and the classification loss, triplet loss (with hard-mining, HM) (for ReID-Stream), and the MMD loss, reconstruction loss (SC constraints).*
  $(\theta, \phi, \psi, \omega) = (\theta, \phi, \psi, \omega) - \eta\nabla_{(\theta,\phi,\psi,\omega)} \mathcal{L}_{total}$ // *Jointly update GSP module GSP_θ, GaitSet (GS) GS_ϕ, SC constraints related FC embedding layers SC_ψ, and ReID-Stream backbone $ReID_\omega$.*
end for

After warming up the GSP module for 80 epochs, we jointly train GSP and GaitSet for extra 160 epochs with initial learning rate as 5×10^{-4}. We also use Adam optimizer [11] for optimization with a weight decay of 1×10^{-4}, the learning rate is decayed by a factor of 0.5 at 40, 80, and 120 epochs. When jointly training GSP and GaitSet, excluding the GSP-related position loss $\mathcal{L}_{position}$ and prediction loss $\mathcal{L}_{pred}$, we further use *separate triplet loss* $\mathcal{L}_{tri}^{sep}$ for training, which is introduced in details in GaitSet [2], and we also set the loss weight as 1.0 for this supervision.

Phase-3: Joint Training for Gait-Stream and ReID-Stream. When we jointly training Gait-Stream and ReID-Stream on the ReID datasets, Gait-Stream is also fine-tuned/learnable. Since the full gait sequence ground truth are not available for ReID-specific datasets, we adjust the original prediction loss $\mathcal{L}_{pred}$ in GSP by only calculating L1 distance between the single input person mask and the middle frame result of the entire predicted gait sequence.

On the large-scale cloth-changing datasets VC-Clothes [25], LTCC [21], and PRCC [29], we set training batch size as $B = 80 = P \times N = 10 \times 8$. Both of Gait-Stream (including GSP and GaitSet) and ReID-Stream use Adam optimizer [11] for optimization, where the initial learning rate for Gait-Stream is 1×10^{-5}, for ReID-Stream is 5×10^{-4}. We optimize two Adam optimizers for Gait-Stream and ReID-Stream with a weight decay of 1×10^{-5} for total 240 epochs. The learning rate is decayed by a factor of 0.1 at 80 and 160 epochs for ReID-Stream, while no learning rate decay for Gait-Stream. For the losses usage, we adopt the widely-adopted classification loss $\mathcal{L}_{cla}$ [5,23], and triplet loss with batch hard mining $\mathcal{L}_{tri}^{HM}$ [7]) as basic optimization objectives for ReID-Stream training, and we set these two loss weights as 1.0. Besides, for the Gait-Stream related losses, including $\mathcal{L}_{position}, \mathcal{L}_{pred}, \mathcal{L}_{tri}^{sep}$, we set all their loss weights as 0.1. For the semantics consistency (SC) constraints related FC embedding layers, we merge their learnable parameters into ReID-Stream's optimization, and set the balance weights for MMD loss $\mathcal{L}_{MMD}$ and reconstruction penalty $\mathcal{L}_{recon.}$ as 0.5. The pseudo code of the entire training process of our GI-ReID is given in Algorithm 1.

3. Details of Datasets

We use one widely-used video ReID dataset MARS [32], and four image-based cloth-changing ReID datasets Real28 [25], VC-Clothes [25], LTCC [21], PRCC [29] to perform experiments. In Table 1, we present the detailed information about these ReID datasets.

Table 1. Brief introduction/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

MARS is a popular dataset for video-based person ReID. There are 20,715 track-lets come from 1,261 pedestrians who are captured by at least 2 cameras. We use the train/test split protocol defined in [32].

Real28, VC-Clothes, LTCC and PRCC are all newly released image datasets for cloth-changing ReID [21, 25, 29].

Real28 is a **small** real-scenario dataset, which is collected in 3 different days (with different clothing) by 4 cameras. It consists of totally 4,324 images from 28 different identities with 2 indoor scenes and 2 outdoors. Similar to [25], since the size of Real28 is not big enough for training deep learning models, we just use it for evaluation. There are totally 336 images in the query and 3,988 images in the gallery.

VC-Clothes is a **synthetic** dataset where images are rendered by the Grand Theft Auto V (GTA5). It has 512 identities, 4 scenes (cameras) and on average 9 images/scenes for each identity and a total number of 19,060 images. Following [25], we equally split the dataset by identities: 256 identities for training and the other 256 for testing. We randomly chose 4 images per person from each camera as query, and have the other images serve as gallery images. Eventually, we get totally 9,449 images in the training, 1,020 images as queries and 8,591 others in the gallery.

LTCC is a **large-scale real-scenario** cloth-changing dataset, which contains 17,138 person images of 152 identities. On average, there are 5 different clothes for each cloth-changing person, with the numbers of outfit changes ranging from 2 to 14. Following [21], we split the LTCC dataset into training and testing sets. The training set consists of 77 identities, where 46 people have cloth changes and the rest of 31 people wear the same outfits during the recording. Similarly, the testing set contains 45 people with changing clothes and 30 people wearing the same outfits.

PRCC is also a **large-scale real-scenario** cloth-changing dataset, recently published by Yang *et al*. [29]. It consists of 221 identities with three camera views *Camera A*, *Camera B*, and *Camera C*. Each person in Cameras A and B is wearing the same clothes, but the images are captured in different rooms. For Camera C, the person wears different clothes, and the images are captured in a different day. The images in the PRCC dataset include not only clothing changes for the same person across different camera views but also other variations, *e.g*. changes in illumination, occlusion, pose and viewpoint. In summary, nearly 50 images exists for each person in each camera view. Therefore, approximately 152 images of each person are included in the dataset, for 33,698 images in total.

Following [29], we split the PRCC dataset into a training set and a testing set. The training set consists of 150 people, and the testing set consists of 71 people, with no overlap between them in terms of identities. The testing set was further divided into a gallery set and a probe set. For each identity in the testing set, we chose one image in *Camera view A* to form the gallery set for a single-shot matching. All images in *Camera views B* and *Camera C* were used for the probe set. Specifically, the person matching between Cam-

era views A and B was performed without clothing changes, whereas the matching between Camera views A and C was cross-clothes matching. The results were assessed in terms of the cumulated matching characteristics, specifically, the Rank-1, Rank-10, and Rank-20 matching accuracy.

4. Experimental Results of Different Settings

Experimental Setups. As we described in the main manuscript, 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, (3) Imitated Cloth-Changing Video ReID. In the main manuscripts, we have presented all the results related to the most challenging setting of (1) real cloth-changing image ReID. The rest results about (2)(3) are shown here. Baseline means the model that only ingests RGB images.

2) General Video ReID. In this setting, we use a general video ReID dataset MARS for experiments. This dataset has no cloth-changing cases. This group of experiments aims to verify two things: 1) gait could benefit ReID even without clothes variations. 2) extracting gait feature from video is easier than that from image, or said, exploiting gait feature in the image-based CC-ReID is more challenging. Since MARS itself contains continuous video frames/clips and human gait masks [1], we don't need GSP to additionally predict gait sequence, so we discard it for simplicity.

3) Imitated Cloth-Changing Video ReID. We still use MARS as dataset to perform experiments in this setting. But the difference is that we *imitate* cloth-changing cases for the images with the same identity through a data augmentation strategy—body-wise color jitter (*i.e.*, randomly change the brightness, contrast and saturation of the human body region in an person image) for training. This group of experiments aims to show that gait information could alleviate the ReID interference caused by clothes changing. GSP module is also removed in this setting.

Results of General Video ReID. Table 2 shows the results. We observe that: 1) Thanks to the leverage of gait characteristics through the proposed Gait-Stream (GS), *Baseline + GS (concat)* and *Baseline + GS + SC* outperform *Baseline* by 1.07%/1.29% in mAP respectively, which demonstrates that gait information indeed benefits ReID. 2) We find that *Baseline + GS + SC* further outperforms *Baseline + GS (concat)* by 0.22% in mAP. This result validates the superiority of our gait utilization manner (*i.e.*, regularization), which makes ReID-Stream not only robust to the gait estimation error, but also computationally efficient (Gait-Stream is not needed in the inference).

[1] https://pan.baidu.com/s/16ZrlM1f_1_T-eZHmQTTkYg.

Table 2. Performance (%) comparison on the general video ReID dataset MARS [32]. **GS** refers to Gait-Stream and **SC** refers to semantics consistency constraints. Note that 'concat' means concatenating ReID vector r and gait vector g together for ReID. The backbone is ResNet-50.

Methods	MARS	
	mAP	Rank-1
Baseline	79.12	87.34
Baseline + GS (concat)	80.19	88.16
Baseline + GS + SC (ours)	**80.41**	**88.32**

Results of Imitated Cloth-Changing Video ReID. To prove that gait indeed could alleviate clothes variation issue, we *imitate* cloth-changing cases for MARS (denoted as CC-MARS). In Table 3, we observe that 1) Disturbed by the synthetic clothing change, *Baseline* suffers from large degradation, 68.52% on CC-MARS vs. 79.12% on raw MARS in mAP. 2) With the assistance of gait, *Baseline+GS (concat)* and *Baseline+GS+SC* improve *Baseline* near 5.0% in mAP. 3) On CC-MARS, the gait 'concat' scheme shows a little superiority than ours. We analyse that's because the 'concat' could help ReID more explicitly, especially when meeting changing clothes. But, the 'concat' scheme needs maintain Gait-Stream in the inference, leading extra computational cost. 4) As video ReID datasets, it is relatively easy to extract gait features on MARS/CC-MARS.

Table 3. Performance (%) comparison on the imitated (using color jitter) cloth-changing video ReID dataset, termed as CC-MARS. The ReID backbone is ResNet-50.

Methods	CC-MARS	
	mAP	Rank-1
Baseline	68.52	72.31
Baseline + GS (concat)	**73.46**	**79.34**
Baseline + GS + SC (ours)	73.13	79.15

5. Comparison with State-of-the-Arts (Complete version)

To save space, we only present the latest approaches in the main manuscripts, and here we show comparisons with more approaches and more evaluation settings on LTCC (Table 4) and PRCC datasets (Table 5).

From the comparison results on PRCC that are shown in Table 5, we observe that 1) Although person ReID with no clothing change (*i.e.*"Same Clothes" in the Table 5) is not the purpose in our work, our method GI-ReID can still achieve an accuracy of 85.97% in Rank-1, which is better than that of all hand-crafted features with metric learning methods and most deep learning methods. 2) When the input images are RGB images without clothing changes, Alexnet [14], VGG16 [22], HA-CNN [15], and PCB [23] all achieve good performance, but they have a sharp per-

Table 4. Performance (%) comparisons of our GI-ReID and other competitors on the cloth-changing dataset LTCC [21]. 'Standard' and 'Cloth-changing' respectively mean the standard setting and cloth-changing setting as mentioned in our main manuscript. '(Image)' or '(Parsing)' represents that the input data is the person image or the body parsing image. '†' means the setting that only identities with clothes changing are used for training.

Methods	Standard		Cloth-changing		Standard†		Cloth-changing†	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
LOMO [16] + KISSME [12]	26.57	9.11	10.75	5.25	19.47	7.37	8.32	4.37
LOMO [16] + XQDA [16]	25.35	9.54	10.95	5.56	22.52	8.21	10.55	4.95
LOMO [16] + NullSpace [31]	34.83	11.92	16.45	6.29	27.59	9.43	13.37	5.34
ResNet-50 (Image) [6]	58.82	25.98	20.08	9.02	57.20	22.82	20.68	8.38
ResNet-50 (Parsing) [6]	19.87	6.64	7.51	3.75	18.86	6.16	6.28	3.46
PCB (Parsing) [23]	27.38	9.16	9.33	4.50	25.96	7.77	10.54	4.04
ResNet-50 + Face [28]	60.44	25.42	22.10	9.44	55.37	22.23	20.68	8.99
PCB [23]	65.11	30.60	23.52	10.03	59.22	26.61	21.93	8.81
HACNN [15]	60.24	26.71	21.59	9.25	57.12	23.48	20.81	8.27
MuDeep [20]	61.86	27.52	23.53	10.23	56.99	24.10	18.66	8.76
Face [28]	60.44	25.42	22.10	9.44	55.37	22.23	20.68	8.99
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 [21])	–	–	26.15	12.40	–	–	25.15	11.67
LTCC-shape + Gait-Stream (ours)	–	–	**28.86**	**14.19**	–	–	26.41	**13.26**

Table 5. Performance (%) comparisons of our GI-ReID and other competitors on the cloth-changing dataset PRCC [29]. "RGB" means the inputs of the model are RGB images; "Sketch" means the inputs of the model are contour sketch images

Methods	Cameras A and C (Cross-Clothes)			Cameras A and B (Same Clothes)		
	Rank-1	Rank-10	Rank-20	Rank-1	Rank-10	Rank-20
LBP [19] + KISSME [13]	18.71	58.09	71.40	39.03	76.18	86.91
HOG [4] + KISSME [13]	17.52	49.52	63.55	36.02	68.83	80.49
LBP [19] + HOG [4] + KISSME [13]	17.66	54.07	67.85	47.73	81.88	90.54
LOMO [17] + KISSME [13]	18.55	49.81	67.27	47.40	81.42	90.38
LBP [19] + XQDA [17]	18.25	52.75	61.98	40.66	77.74	87.44
HOG [4] + XQDA [17]	22.11	57.33	69.93	42.32	75.63	85.38
LBP [19] + HOG [4] + XQDA [17]	23.71	62.04	74.49	54.16	84.11	91.21
LOMO [17] + XQDA [17]	14.53	43.63	60.34	29.41	67.24	80.52
Shape [1]	11.48	38.66	53.21	23.87	68.41	76.32
LNSCT [27]	15.33	53.87	67.12	35.54	69.56	82.37
Alexnet [14] (RGB)	16.33	48.01	65.87	63.28	91.70	94.73
VGG16 [22] (RGB)	18.21	46.13	60.76	71.39	95.89	98.68
HA-CNN [15] (RGB)	21.81	59.47	67.45	82.45	98.12	99.04
PCB [23] (RGB)	22.86	61.24	78.27	**86.88**	98.79	99.62
Alexnet [14] (Sketch)	14.94	57.68	75.40	38.00	82.15	91.91
VGG16 [22] (Sketch)	18.79	66.01	81.27	54.00	91.33	96.73
HA-CNN [15] (Sketch)	20.45	63.87	79.58	58.63	90.45	95.78
PCB [23] (Sketch)	22.48	61.07	77.05	57.36	92.12	96.72
SketchNet [30] (Sketch+RGB)	17.89	43.70	58.62	64.56	95.09	97.84
Face [26]	2.97	9.85	13.52	4.75	13.40	45.54
Deformable Conv. [3]	25.98	71.67	85.31	61.87	92.13	97.65
STN [10]	27.47	69.53	83.22	59.21	91.43	96.11
RCSANet [8]	31.60	–	–	–	–	–
PRCC-contour [29]	34.38	77.30	88.05	64.20	92.62	96.65
+ Gait-Stream (ours)	36.19	79.93	91,67	–	–	–
Baseline (ResNet-50)	22.23	61.08	76.44	75.81	97.34	98.95
GI-ReID (ResNet-50)	33.26	75.09	87.44	78.95	97.89	99.11
Baseline (OSNet)	28.70	72.34	85.89	83.68	98.24	99.26
GI-ReID (OSNet)	**37.55**	**82.25**	**93.76**	85.97	**98.82**	**99.72**

formance drop when a clothing change occurs, illustrating the challenge of person ReID when a person dresses differently. Therefore, the application of existing person ReID methods is not straightforward in this scenario. In contrast, our GI-ReID that leverages gait information is beneficial to learn the clothing invariant feature, which makes our method achieve satisfactory performance 37.55% in Rank-1 even under the cloth-changing scenario.

6. Study on Failure Cases (Limitations)

As we described in the main manuscript, since the existed large difference on the capture viewpoint and environment between gait and ReID training data, the predicted results of gait sequence prediction (GSP) module are not so accurate when occlusion, partial, multi-person, *etc*, existed in the person images. As shown in Figure 2, GSP gives unsatisfactory gait prediction results, where large estimation errors exist in the predicted gait frames, which will hurt the ReID performance. That is why we **indirectly** use gait prediction results in a two-stream knowledge regularization manner, which makes our GI-ReID robust/less sensitive to these failure cases.

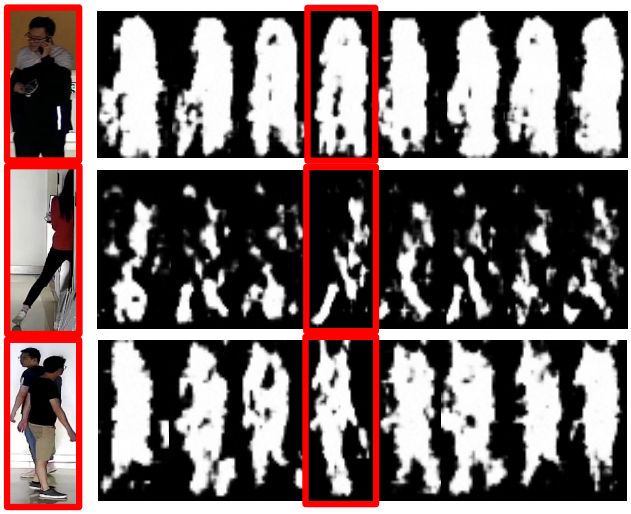

Figure 2. Failure cases of gait sequence prediction (GSP).

7. Social Impact

Positive. In this paper, we propose to utilize human unique gait to address the cloth-changing ReID (CC-ReID) problem from a single image. A novel gait-involved two-stream framework GI-ReID is introduced for image-based CC-ReID. To our best knowledge, this paper is the first attempt to take gait as a regulator with a Gait-Stream (discarded in the inference), to encourage the cloth-agnostic representation learning of image-based ReID-Stream. This is very important for both of academic community and industry, and it is also valuable and meaningful to bridge the gap between the fast-developing ReID algorithms and practical applications.

Negative. Due to the urgent demand of public safety and increasing number of surveillance cameras, person ReID is imperative in intelligent surveillance systems with significant research impact and practical importance, but this task also might raise questions about the risk of leaking private information. On the other hand, the data collected from the surveillance equipments or downloaded from the internet may violate the privacy of human beings. Therefore, we appeal and encourage research that understands and mitigates the risks arising from surveillance applications.

References

[1] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE TPAMI*, 24(4):509–522, 2002. 5

[2] 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. 1, 2

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

[4] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. 2005. 5

[5] 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. 3

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 5

[7] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2, 3

[8] 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. 5

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ICML*, 2015. 1

[10] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015. 5

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 2, 3

[12] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. 2008. 5

[13] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295. IEEE, 2012. 5

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural net-

works. *Communications of the ACM*, 60(6):84–90, 2017. 4, 5

[15] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 4, 5

[16] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 5

[17] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 5

[18] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 1

[19] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996. 5

[20] 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. 5

[21] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. *WACV*, 2020. 2, 3, 5

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 5

[23] 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. 1, 3, 4, 5

[24] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1):4, 2018. 2

[25] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *CVPRW*, pages 830–831, 2020. 3

[26] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515. Springer, 2016. 5

[27] 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. 5

[28] Jia Xue, Zibo Meng, Karthik Katipally, Haibo Wang, and Kees van Zon. Clothing change aware person identification. In *CVPRW*, pages 2112–2120, 2018. 5

[29] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *TPAMI*, 2019. 3, 5

[30] 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. 5

[31] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identificatio. In *CVPR*, 2016. 5

[32] 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. 3, 4

[33] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, et al. Omni-scale feature learning for person re-identification. *ICCV*, 2019. 1