An In-depth Exploration of Person Re-identification and Gait Recognition in Cloth-Changing Conditions

Weijia Li[1,4*], Saihui Hou[2,3*], Chunjie Zhang[1,4†], Chunshui Cao[3], Xu Liu[3],
Yongzhen Huang[2,3†], Yao Zhao[1,4]

[1] Institute of Information Science, Beijing Jiaotong University
[2] School of Artificial Intelligence, Beijing Normal University, [3] WATRIX.AI
[4] Beijing Key Laboratory of Advanced Information Science and Network, Beijing, 100044, China

21125203@bjtu.edu.cn, housaihui@bnu.edu.cn, cjzhang@bjtu.edu.cn,
{chunshui.cao, xu.liu}@watrix.ai, huangyongzhen@bnu.edu.cn, yzhao@bjtu.edu.cn

Abstract

The target of person re-identification (ReID) and gait recognition is consistent, that is to match the target pedestrian under surveillance cameras. For the cloth-changing problem, video-based ReID is rarely studied due to the lack of a suitable cloth-changing benchmark, and gait recognition is often researched under controlled conditions. To tackle this problem, we propose a Cloth-Changing benchmark for Person re-identification and Gait recognition (CCPG). It is a cloth-changing dataset, and there are several highlights in CCPG, (1) it provides 200 identities and over 16K sequences are captured indoors and outdoors, (2) each identity has seven different cloth-changing statuses, which is hardly seen in previous datasets, (3) RGB and silhouettes version data are both available for research purposes. Moreover, aiming to investigate the cloth-changing problem systematically, comprehensive experiments are conducted on video-based ReID and gait recognition methods. The experimental results demonstrate the superiority of ReID and gait recognition separately in different cloth-changing conditions and suggest that gait recognition is a potential solution for addressing the cloth-changing problem. Our dataset will be available at https://github.com/BNU-IVC/CCPG.

1. Introduction

With the rapid development of video devices, more and more surveillance systems are taken into real applications and play an essential role in protecting the security of our society. However, along with the significantly growing

---
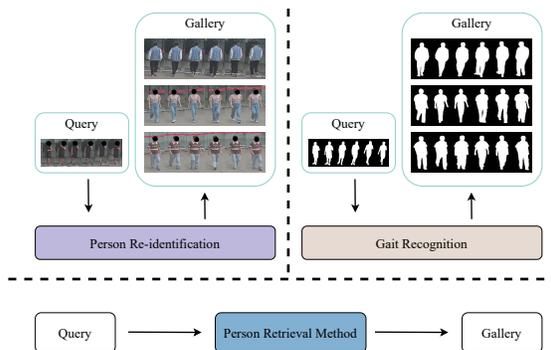*Equal contribution.
†Corresponding authors.

Figure 1. The targets of person re-identification and gait recognition are consistent, to search the target person from the gallery. The main difference is that gait recognition normally requires pedestrian segmentation.

amount of data, traditional social security is facing two dilemmas: data storage and inefficient monitoring. Due to artificial intelligence progress and diverse demand in social security, many security technologies have come into our sights, such as face recognition, person re-identification (ReID), and gait recognition. These technologies reduce data size vastly and improve efficiency in social security.

Video-based ReID plays a vital role in surveillance video analysis, and it intends to match the probe sequence of the specific person from gallery sequences in surveillance systems by learning features of multiple frames [5, 27, 43]. Compared with image-based ReID [10, 21, 36], video-based ReID provides both appearance and rich temporal information. Previous video-based ReID methods [9, 12, 19, 20, 33, 36, 38] focus on the cloth-consistent setting, where people do not change their clothes in the short term. However, in the real world, clothes-changing situations are everywhere. Due to its practical application in social security, cloth-

changing ReID has gotten more attention in these years. Another technology is gait recognition. Gait is a unique biometric characteristic that describes the walking patterns of one person [23] and contains spatial statics and temporal dynamics in the walking process. Compared with other biometric features, gait is hard to disguise and can work in long distances [13], so it has a potential for social security. Popular gait recognition datasets [28,37] are collected under control conditions, and some real-world benchmarks [41, 44] come into our sights in these years. Many gait recognition methods [4, 7, 17, 18, 24] are mainly developed based on these datasets. Several challenges are studied explicitly, such as carrying bags and cloth-changing. However, the research on gait recognition in real scenarios is insufficient. So gait recognition in the wild has attracted increasing attention.

In real scenarios, the final targets of ReID and gait recognition are consistent, that is, to recognize the target person across cameras, as shown in Fig. 1. Previous works [3, 39] conduct experiments to compare the performance of video-based ReID and gait recognition on video-based ReID datasets, which are cloth-unchanged. However, dressing variation in practical applications is a significant problem, especially the finer cloth-changing problem, *e.g.*, changing a whole fit, changing only tops, and changing only pants. Little work specializes in these cloth-changing conditions, and no comparison experiment has been conducted between video-based ReID and gait recognition. The main reason is the lack of such a cloth-changing benchmark for comparison of cloth-changing conditions. So it is time to build such a cloth-changing dataset for video-based ReID and gait recognition. First of all, we should take three aspects into consideration. **(1) RGB data**. A benchmark should provide RGB data for comparison between video-based ReID and gait recognition at least, because RGB data is the basic research data for them. **(2) Clothes statuses**. People usually change their clothes daily, and a dataset should contain different clothing statuses close to daily life. **(3) Collection environment**. The raw data should not be collected under strict restrictions, which makes the dataset similar to real-world applications as much as possible since our final target is to satisfy social security demands in the real world. To our best knowledge, no such public dataset could satisfy all the requirements mentioned above. CCVID [8] only contains the front views of the pedestrians, rather than diverse views. GREW [44] lacks RGB data and finer clothes changing.

Therefore, in this paper, we propose a benchmark, **C**loth-**C**hanging benchmark for **P**erson re-identification and **G**ait recognition (**CCPG**). Especially, the CCPG dataset has several important features. **(1)** It contains 200 subjects wearing many different clothes and over 16,000 sequences, and the RGB data is available. **(2)** Subjects contain seven different clothes statuses, whose cloth-changing ways include changing a whole fit, changing tops, changing pants, and carrying bags, and these abundant types of dressing are supported for cloth-changing comparison. In addition, more accurate tracking results are provided and reduce the noise in the dataset, because precise human detection guarantees the following recognition mission. **(3)** Our raw data is collected indoors and outdoors, and many challenges are considered, including viewing angles, occlusions, illumination changes, etc., which make it more similar to the real world.

Taking advantages of our new cloth-changing benchmark, we conduct a series of systematic experiments between video-based ReID and gait recognition, aiming to compare their performance in different dressing settings. First, leading video-based ReID methods are performed on the CCPG dataset in different cloth-changing settings, which indicates that they are sensitive to the appearance features and still have room to improve on cloth-changing problems. Second, popular gait recognition methods are implemented on the CCPG dataset and achieve higher performance in some cloth-changing settings. In addition, gait recognition methods are as good as ReID in partial cloth-changing situations, *e.g.*, only changing tops and only changing pants. Third, we compare the performance of these pedestrian retrieval methods and analyze experimental results in different cloth-changing settings. The key to addressing the cloth-changing problem is to exploit the cloth-invariant features, and these experimental results demonstrate the SOTA gait recognition methods have more potential for cloth-changing problems.

In summary, our main contributions are summarized as follows:

- First, we present a brand new cloth-changing benchmark, named CCPG. Compared with others, our benchmark provides finer cloth-changing conditions. We hope it could help study the cloth-changing problem for video-based ReID and gait recognition. Importantly, it should be emphasized that our data collection obtains permission from authorities and adult volunteers for research purposes.

- Second, with the motivation to study the performance of video-based ReID and gait recognition methods under different cloth-changing conditions, comprehensive experiments are conducted on them, and the results show the better performance of gait recognition on CCPG, and suggest that gait recognition is a potential solution for solving cloth-changing problems.

2. Related Work

In this section, we discuss video-based ReID and gait recognition in aspects of datasets and methods.

Table 1. Statistics of our proposed dataset and popular video-based ReID datasets, which are ranked in publication time.

Dataset	IDs	Tracklets	Views	Cloth-Changing
PRID-2011 [11]	200	400	2	✗
iLIDS-VID [32]	300	600	2	✗
MARS [42]	1,261	20,715	6	✗
Duke-Video [35]	1,812	4,832	8	✗
Duke-Tracklet [15]	1,788	12,647	8	✗
LPW [26]	2,731	7,694	4	✗
LS-VID [14]	3,772	14,943	15	✗
CCVID [8]	226	347,833	1	✓
CCPG	200	16,566	10	✓

2.1. Video-based ReID

2.1.1 Datasets

Most video-based ReID datasets [8,11,14,15,26,32,35,42] are for short-term scenarios, *i.e.*, clothes unchanged. We compare these datasets shown in Tab. 1 and analyze four main datasets. (1) An early dataset, iLIDS-VID [32], consists of 300 subjects walking in an airport. Its main challenges are similar clothes, illumination, complex backgrounds, and severe occlusions. (2) MARS [42] includes 1,261 IDs and over 20k tracklets. However, indoor scenes are rarely taken into account. (3) LPW [14] contains 7,694 tracklets and over 590k images of three scenes, which is large-scale and large-age-span. However, dressing variation is not taken into consideration. (4) CCVID [8] is reconstructed from Front View Gait dataset (FVG) [40], which is the first cloth-changing video-based ReID dataset. However, it only contains front views of subjects and does not contain indoor scenes.

2.1.2 Methods

Many video-based ReID methods [9, 12, 19, 20, 33, 36, 38] perform well on short-term (cloth-consistent) datasets. PSTA [33] exploits frame-level features from different terms of temporal information and models spatial-temporal features for ReID. BiCnet-TKS [12] argues that images

at original resolution contain detailed visual cues, down-sampled images consist of long-range contexts, and it can model spatial-temporal information well through the two different resolution branches. Considering that temporal appearance misalignment is unavoidable and 3D convolution may destroy the appearance representation of video clips, AP3D [9] introduces APM module to align spatial information and utilize 3D convolution to aggregate temporal information. PiT [38] proposes a multi-direction and multi-scale pyramid in transformer and integrates information on patches under different-direction division strategies.

2.2. Gait Recognition

2.2.1 Datasets

There are many gait recognition dataset [25, 28, 29, 31, 37, 40, 41, 44], and a comparison of these datasets is shown in Tab. 2. Here we discuss three main datasets. (1) CASIA-B [37] is a classic dataset for gait recognition, which contains clothes variation and different walking types. (2) OU-MVLP [28] is a much larger dataset that includes 10,307 identities and over 288k sequences. However, these two are collected in static indoors, which leads to limited applicability in real scenarios. (3) GREW [44] is a large-scale dataset collected in the wild, and contains many challenges. But it does not provide the RGB-type data, which means that a comparison experiment for video-based ReID and gait recognition can not be implemented. And the combination of GREW and CASIA-B is not enough to cover daily clothes types.

2.2.2 Methods

According to data types, gait recognition methods can be roughly classified into two categories, model-based methods and appearance-based methods.
Model-based Methods. These methods [17, 30] aim to use other body information (*e.g.*, key points and skeleton) to generate more discriminative gait features that are more robust to view-changing and cloth-changing. PoseGait [17]

Dataset	IDs	Tracklets	Views	Environment	Data Tpye	Cloth-Changing
CASIA-A [31]	20	240	3	Indoor	RGB	✗
CASIA-B [37]	124	13,640	11	Indoor	RGB, Silh.	✓
CASIA-C [29]	153	1,530	1	Outdoor	Infr., Silh.	✗
OU-MVLP [28]	10,307	288,596	14	Indoor	Silh.	✗
FVG [40]	226	2,856	1	Outdoor	RGB	✓
GREW [44]	26,345	128,671	882	Outdoor	Silh., Flow, Pose	✓
Gait3D [41]	4,000	25,309	39	Indoor	Silh, Pose, Flow	✗
CASIA-E [25]	1,014	778,752	26	Outdoor	Silh., Infr	✓
CCPG	200	16,566	10	Indoor & Outdoor	Silh, RGB	✓

Table 2. Statistics of our proposed dataset and popular gait recognition datasets, which are ranked in publication time. "Silh." and "Infr." mean silhouette and infrared.

exploits the human pose features designed on 3D coordinates of joints of the human body, and CNN extracts gait representation from the pose features. GaitGraph [30] extracts spatial-temporal representation for gait recognition, using human skeleton structure based on GCN. However, these methods above are sensitive to the resolution of inputs because it is not easy to get the body information from low-resolution inputs.

Appearance-based Methods. These methods aim to extract discriminative features mainly based on the silhouettes. Depending on the input type of silhouettes, these methods are divided into three categories: template-based methods, set-based methods, and sequences-based methods. (1) *template-based methods*. GEINet [24] uses a CNN architecture to extract gait features through a single image, *e.g.*, GEI. (2) *sequences-based methods*. Sequences-based methods use frames of each sequence rather than a template image to extract motion information. 3D convolutions are used to capture spatio-temporal features at fixed video length [34]. GaitPart [7] argues that each part of the human body has its independent features. GaitGL [18] uses 3D CNN to extract global and local information. (3) *set-based methods*. GaitSet [4] regards a whole sequence as an unordered set that is immune to frame permutations, and extracts the gait features from the unordered sets.

3. The CCPG Dataset

3.1. Overview of CCPG

CCPG is a cloth-changing benchmark, and it contains 200 subjects wearing seven different clothes, walking across outdoor and indoor scenes, and includes over 16k sequences. Moreover, without strict collection requirements, it brings other challenges, such as diverse views and occlusion. Importantly, our entire data collection is permitted by the authority and volunteers. In the rest of this section, we will introduce the CCPG dataset in detail, including data collecting, data processing, and dataset statistics.

3.2. Data Collection

The raw videos of CCPG are collected in an idle factory building in a school. Our team is authorized to place cameras there, and we recruit 200 volunteers to participate in our data-collecting program. Also, we promise to protect their privacy at any time. At the data collection site, we place two cameras outside and eight cameras inside. As for the height of the cameras, one is about 2.7 meters, and the others are 3 meters. Considering the realistic collection constraints, we still make our dataset similar to the real world as much as possible.

Walking Route. To imitate walking patterns of people in the real world, we design a simple walking route for data collecting, and the detail of it is shown in Fig. 2. A volunteer stands at the start point to begin the collecting process, and this volunteer walks from outside to inside under the two cameras (*Cam.1, Cam.2*), which record outdoor videos. Then this volunteer goes into the innermost square area via four cameras (*Cam.3, Cam.4, Cam.5, Cam.6*). Backgrounds changing situation happens in these four cameras. At last, this volunteer is asked to walk around counterclockwise in the square area via (*Cam.7, Cam.8, Cam.9, Cam.10*). Specifically, we set two boxes in front of two cameras (*Cam.7, Cam.9*), and we can get partial occlusion sequences, which can be used for further occlusion analysis.

Cloth-changing Situations. The dressing is an essential element that influences the performance of video-based ReID methods. Therefore, we design a cloth-changing plan to simulate daily dressing, containing four dressing situations happening in everyday life, which are changing bottoms, changing coats, changing the whole clothes, and changing clothes with a bag. Therefore, we build a large clothes pool consisting of many coats, pants, and bags. As for tops, there are many different colors (*e.g.*, white, light blue, and red) and various types (*e.g.*, shirts, hoodies, and long coats), and the amount of them is 13. For bottoms, eight pairs of pants are provided, such as shorts, trousers, and jeans. We argue that carrying a bag happens frequently, so five different bags are provided, which are a yellow satchel, a beige satchel, a gray handbag, a green backpack, and a gray backpack. During the data collecting, volunteers are asked to select some clothes randomly in this clothes pool according to requirements and get different outfit combinations.

In our plan, there are seven different outfits for each volunteer, and each volunteer changes into a new set of clothes after completing a walking route. *For the first time*, volunteers wearing their own clothes (U0D0) and finish the route. *For the second and third time*, they replace their tops and pants in turn, *i.e.*, they get a whole new outfit (U1D0, U1D1). *For the fourth and fifth time*, they change into a whole new outfit twice (U2D2, U3U3). *For the sixth*
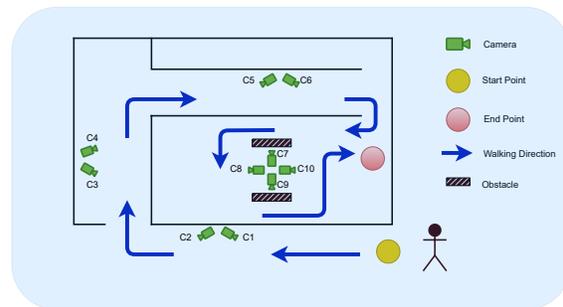

Figure 2. The layout is our designed route. Volunteers start from the bottom right, walk along arrows, and end at the endpoint. Ten cameras are fixed indoors and outdoors, and record the cross-scene walking process. "C1" to "C10" mean "*Cam.1*" to "*Cam.10*".

time, they change back to their original pants (U0D3). *For the last time*, they change back their tops and carry a bag that they randomly select in the clothes pool (U0D0BG). In summary, seven different dressings are close to everyday life, *i.e.*, changing tops only, changing pants only, changing a whole set of outfits, or changing clothes with carrying a bag.

3.3. Data Processing

From raw videos to an available ReID dataset, this requires detection tracklets generation and data annotation [36]. To generate accurate and continuous bounding boxes for each subject, we use YOLO [2] to realize the mission of human detection. Considering that irrelevant people may walk indoors and outdoor cameras capture some passersby, we clean the processed data and discard those useless sequences.

Face information is one of the ReID features and makes a difference in the short-distance recognition mission. Meanwhile, considering the COVID-19 pandemic [6], people have to wear a mask every day, which may cause difficulty extracting facial information for ReID. In addition, we notice that the shoes of subjects are not changed in some cloth-changing image-based ReID datasets, which happens in our dataset due to collection restrictions. So shoes become an element in cloth-changing ReID we need to study. Therefore, we need to mask these regions in gray, with the target to explore how they work when clothes are changed, and we discuss the realistic meaning of these actions next.

(1) Gray facial area. Facial area contains ID-relevant and cloth-irrelevant information that can work when clothes are changed. If the facial area is filled in gray, it is similar to a common situation that people wearing the same style of mask. At the same time, this version data could help us study the role of face information when clothes are changed.

(2) Gray shoes area. Shoes should be cloth-relevant features, but they are unchanged in our dataset due to realistic restrictions. The appearance information on the area of the shoes can become cloth-relevant features for each identity. It is not realistic, because people usually change their clothes along with their shoes. To explore the influence of the unchanged shoes, we fill the shoes area in gray, making appearance information on the shoes area be ID-irrelevant. We argue that only appearance of shoes is remove, and other information on the area of shoes is maintained, such as shape information and motion information. The version of data could help on exploring the ReID models performance without shoes appearance.

(3) Gray facial and shoes areas. It means facial and shoes areas are filled in gray which is close to a situation that people change their shoes and wear masks. This version of data helps us study the ability of models to exploit cloth-invariant features without the assistance of face and shoes.

Hence, we implement the human parsing method [16] on this dataset to differentiate face and shoes areas, which are two parts of human parsing. For gait recognition, we extract human segmentation images on our dataset by performing U-Net [22], and the segmentaion results with low quality are removed by annotators. Finally, we reconstruct our dataset, and obtain four RGB version datasets and a silhouette dataset, and the examples of these datasets are shown in Fig. 4.

- **RGB.** Under the guidance of human parsing, we mask facial area, shoes area, and both facial and shoes areas. Because the parsing area may be out of human bounds sometimes, we use segmentation results to rectify the boundaries of parsing. In all, we get four subsets of RGB datasets, completed RGB named CCPG-A, RGB w/o face named CCPG-B, RGB w/o shoes named CCPG-C, and RGB w/o face&shoes named CCPG-D.

- **Silhouette.** Human segmentation is implemented, and we get a subset for gait recognition mission, named CCPG-G.

3.4. Dataset Statistics

CCPG contains 200 subjects with variations in clothes, gender, and body shape, and we provide several statistical analyses below, as shown in Fig. 5.

(1) Statistics of gender: the ratio of males and females is nearly 3:2. That is, 122 males and 78 females to be precise.

(2) Statistics of sequences: For sequence length, most sequences length ranges from 50 to 180, and a few sequences are distributed in the region that ranges from 50 to 18.

(3) Statistics of sequences under cameras: For cameras, the average length of sequences in *Cam.1* have the maximum number, and sequences in *Cam.3.* have the minimum number of frames.
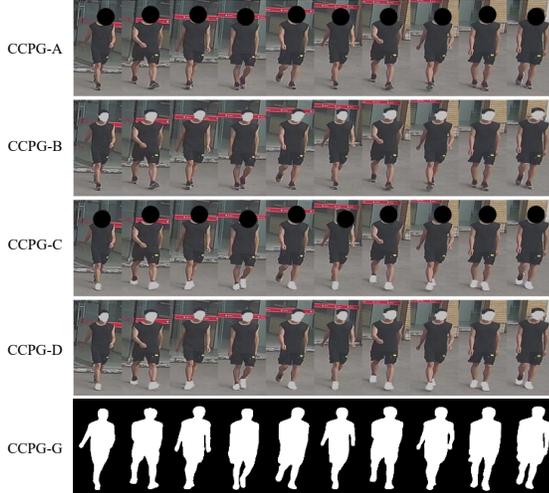

Figure 3. The visualization of cloth-changing situations. Two row images are two different persons in different clothes. From left to right are U0D0, U1D0, U1D1, U2D2, U3U3, U0D3, U0D0BG. "U", "D" and "BG" mean "Up", "Down" and "Bag"

Figure 4. Four RGB version datasets and one silhouette version dataset.
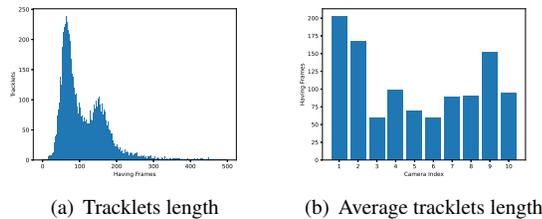

(a) Tracklets length   (b) Average tracklets length

Figure 5. Statistics of CCPG. (1) the distribution of tracklets length, (2) the average length of tracklets under each camera.

4. Experiments On CCPG

CCPG is a brand new cloth-changing benchmark, and contains many types of dressing variations, making it is possible to study video-based ReID and gait recognition methods in cloth-changing conditions. In this section, we do these things, (1) we introduce our evaluation protocols of CCPG, (2) we introduce methods for experiments, (3) we conduct experiments of video-based ReID and gait recognition, and report the performance of these methods. Then, we analyze the experimental results with visualization of heatmaps.

4.1. Evaluation Protocols

In ReID evaluation, a cross-camera search mode is adopted commonly, *i.e.*, query and gallery captured by different cameras [42], and we follow the previous mode.

For dataset division, the training set contains the first hundred subjects (ID: 0-99), and the test set contains the remaining subjects (ID: 100-199). To evaluate performance in different settings, we adopt top-1 accuracy and mAP (mean

Average Precision). Inspired by the partition settings [4], we proposed three cloth-changing settings below:

(1) Cloth-changing setting (**CL-Full**): for each subject, whole outfits are changed. Sequences of U0D0 and U0D0BG make up the query, and sequences of U1D1, U2D2, and U3D3 make up the gallery.

(2) Ups-changing setting (**CL-UP**): for each subject, only tops are changed. Sequences of U3D3 are used for the query, and sequences of U0D3 are kept in the gallery.

(3) Pants-changing setting (**CL-DN**): for each subject, only pants are changed. Sequences of U1D0 are used for the query, and sequences of U1D1 are kept in the gallery.

4.2. Comparison Methods on CCPG

In order to conduct comparison experiments, representative video-based ReID methods and gait recognition methods are selected.

Video-based ReID. For video-based ReID, we select four representative models, *i.e.*, AP3D [9], PSTA [33], BiCnet-TKS [12] and PiT [38], which achieve high performance on MARS [42], iLIS-VID [32] and LS-VID [14] benchmarks. In addition, the resolution of images in sequences is set to 256*128 as default, and other implementation details are consistent with default configurations.

Gait Recognition. For gait recognition, four popular and public methods are chosen here, *i.e.*, OGBase (provided by OpenGait[1]), GaitSet [4], GaitPart [7], GaitGL [18], and AUG-OGBase, and we adopt the implementations in Open-Gait program. The configurations for OGBase, GaitSet, GaitPart, and GaitGL are consistent with the default configuration provided in OpenGait program. AUG-OGBase is modified from OGBase according to our practical experience, as shown in Tab. 3. Compared with the default configuration of OGBase, we add BN layers after the convolution layers, change the batch size to 16*16, set milestones as [30,000, 60,000], and set the total iteration to be 80,000. Additionally, the resolution of the input silhouettes is 128*88 in our gait recognition experiments.

Block	Layer	In_C	Out_C	Kernel_Size	Stride	Padding
Block_1	Conv_1_BN	1	32	5	2	2
	Conv_2_BN	32	32	3	1	1
	Max_Pooling	-	-	2	2	0
Block_2	Conv_3_BN	32	64	3	1	1
	Conv_4_BN	64	64	3	1	1
	Max_Pooling	-	-	2	2	0
Block_3	Conv_5_BN	64	128	3	1	1
	Conv_6_BN	128	128	3	1	1
	Conv_7_BN	128	256	3	1	1
	Conv_8_BN	256	256	3	1	1

Table 3. The structure of our AUG-OGBase.

---

[1]https://github.com/ShiqiYu/OpenGait

(a) ReID on CCPG-A and CCPG-B VS gait recognition. (shoes not masked)

Dateset	Method	CL-Full		CL-UP		CL-DN	
		top-1	mAP	top-1	mAP	top-1	mAP
CCPG-A	AP3D	**90.1**	60.7	89.2	71.3	**96.2**	76.5
	BiCnet-TKS	87.5	60.5	90.4	73.7	90.8	76.4
	PSTA	89.5	**66.6**	92.5	80.0	93.0	80.3
	PiT	87.6	65.3	92.2	**80.7**	94.3	**80.8**
CCPG-B	AP3D	86.7	60.1	89.3	77.2	87.2	74.6
	BiCnet-TKS	84.2	57.9	87.0	73.0	90.8	76.8
	PSTA	**88.2**	**65.3**	91.2	**79.3**	92.3	**79.4**
	PiT	85.1	60.1	**92.7**	78.0	**92.8**	78.4
CCPG-G	OGBase	78.4	44.5	82.3	58.3	86.0	59.3
	GaitSet	77.7	46.4	83.5	59.6	83.2	61.4
	GaitPart	77.8	45.5	84.5	63.1	83.3	60.1
	GaitGL	69.1	27.0	75.0	37.1	77.6	37.6
	AUG-OGBase	**84.7**	**52.9**	**88.4**	**67.5**	**89.4**	**67.9**

(b) ReID on CCPG-C and CCPG-D VS gait recognition. (shoes masked)

Dateset	Method	CL-Full		CL-UP		CL-DN	
		top-1	mAP	top-1	mAP	top-1	mAP
CCPG-C	AP3D	68.4	31.4	72.8	50.2	**86.4**	58.9
	BiCnet-TKS	**68.6**	37.6	**76.3**	**59.9**	79.2	60.9
	PSTA	66.3	**39.0**	74.4	59.3	86.2	**68.2**
	PiT	60.7	35.2	67.2	58.3	82.4	67.2
CCPG-D	AP3D	55.1	27.3	60.4	49.0	80.1	63.3
	BiCnet-TKS	**64.5**	36.9	72.3	59.8	78.7	62.3
	PSTA	62.6	**37.6**	**73.8**	**60.2**	83.9	67.8
	PiT	57.1	30.8	68.4	55.4	79.1	65.3
CCPG-G	OGBase	78.4	44.5	82.3	58.3	86.0	59.3
	GaitSet	77.7	46.4	83.5	59.6	83.2	61.4
	GaitPart	77.8	45.5	84.5	63.1	83.3	60.1
	GaitGL	69.1	27.0	75.0	37.1	77.6	37.6
	AUG-OGBase	**84.7**	**52.9**	**88.4**	**67.5**	**89.4**	**67.9**

Table 4. Comparison results between video-based ReID and gait recognition.

4.3. Experimental Results and Analysis

In this section, we conduct experiments for video-based ReID methods on the CCPG-A, CCPG-B, CCPG-C, and CCPG-D datasets, and gait recognition methods on the CCPG-G dataset. We report the comparison results in Tab. 4, and give the analysis in the following subsections.

4.3.1 ReID on CCPG-A vs Gait Recognition

Given the results of other ReID experiments [36, 38], ReID methods are able to achieve high performance in cloth-consistent datasets. From our observation in Tab. 4, these video-based ReID methods still perform excellently on the CCPG-A dataset, but the CCPG-A is a cloth-changing dataset. Hence, we visualize the feature maps of them in Fig. 6(a). It is obvious that video-based ReID models trained on the CCPG-A are more likely to concertrate on the areas of face and shoes, which are indeed cloth-irrelevant information except on the area of shoes. But in the real world, people usually change their clothes along with shoes, and faces are not captured easily due to wearing a mask and bad camera angles. Moreover, for the cloth-changing pedestrian retrieval mission, it is the cloth-irrelevant information of human body that builds discriminative features for each person. Hence, we speculate that the performance of video-based ReID methods may be deteriorated in our CCPG dataset, especially when appearance information on the area of the face and shoes is removed.
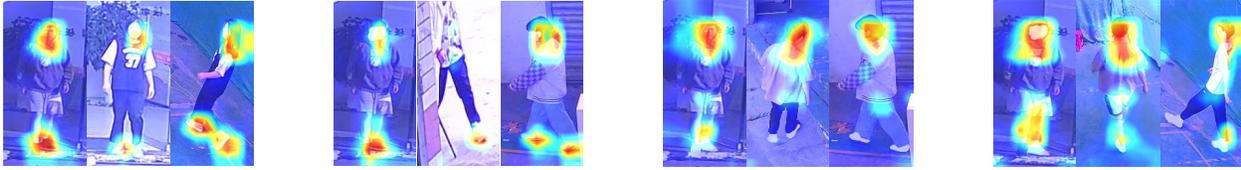
4.3.2 ReID on CCPG-B vs Gait Recognition

Due to the long-distance surveillance and various angles, facial information can not be captured well, and the video-based ReID experiments on CCPG-B are similar to some realistic situations. In Tab. 4, we notice that video-based methods get a performance degradation in the CL-Full set-

ting. From the visualization results as shown in Fig. 6(b), these video-based ReID models still focus on the information on the areas of head and shoes, and overfit on the areas of shoes specially. Without the facial information, video-based ReID models still learn information from the area of the head, we speculate that the area of the head contains some unique information, such as hairstyle and head shape, and models can generate discriminative features with their assistance. Moreover, we notice that performance on CL-DN is always over CL-UP, and the main reson is that the tops area is larger than the pants area. Tops-unchanged maintains more cloth-consistent information compared with pants-unchanged. Additionally, similar to the situation in the CCPG-A dataset, shoes are not changed when clothes are changed in the CCPG-B dataset, and video-based ReID models focus more information on the area of the shoes. Hence, we argue that it is still insufficient to reflect the true performance of video-based ReID methods in real cloth-changing scenarios.

4.3.3 ReID on CCPG-C vs Gait Recognition

In the CCPG-C dataset, only appearance information on the area of shoes is masked, but other information on the area is maintained, like shape information. From the results shown in Tab. 4, we notice that video-based ReID methods drop largely without ID-relevant information on the area of shoes, which is similarly to a situation that pedestrians change their shoes. With the assistance of the heatmaps in Fig. 6(c), we notice that video-based ReID models focus primarily on the area of head-shoulder and motion information on the area of legs. Therefore, when shoes information is not ID-relevant along with clothes-changing, we argue that these models focus on the features on the areas mentioned before, which are cloth-irrelevant. Importantly, these ReID methods are not so good as gait recognition methods in exploiting the cloth-irrelevant features.

(a) Video-based ReID on CCPG-A.　　(b) Video-based ReID on CCPG-B.　　(c) Video-based ReID on CCPG-C.　　(d) Video-based ReID on CCPG-D.

Figure 6. Visualization of heatmaps in video-based ReID.

4.3.4　ReID on CCPG-D vs Gait Recognition

Compared with previous experiments, our video-based ReID experiments on the CCPG-D dataset are more similar to the realistic situations. For the results reported in Tab. 4, gait recognition methods commonly perform better than video-based ReID methods. As the feature visualization maps in Fig. 6(d), we notice ReID models more focus on motion information on the areas of the human body and head-shoulder. Significantly, we also find that gait recognition methods achieve better rank-1 accuracy in the CL-DN setting, but lower on mAP than video-based ReID. We have the following conjecture: gait recognition methods are sensitive to view variation, and they can easily find the results for similar angles in the gallery, but miss those positive samples with bigger angle differences. It causes higher rank-1 accuracy but lower mAP for gait recognition.

4.4. Experimental Summary

Equipped with the CCPG dataset, we design three types of cloth-changing situations which are close to daily dressing, and conduct comprehensive experiments for video-based ReID and gait recognition on the cloth-changing problem. Due to the appearance information on the area of the shoes, we argue that the performance of video-based ReID on the CCPG-C and CCPG-D datasets can reflect the real performance on the cloth-changing problem, which is comparable with gait recognition. Based on the experimental results, we have the following conclusions. (1) In the CL-Full and CL-UP settings, gait recognition methods can surpass the video-based ReID methods, and it suggests that gait recognition has more potential on addressing the cloth-changing problem. (2) For certain partial cloth-changing situations, video-based ReID still has good performance, especially in the CL-DN setting. But the results indicates that the performance of video-based ReID methods decreases with increasing levels of cloth-changing on the human body, which means that these video-based ReID methods are fragile to appearance variations. The above contents further strengthen our point of view that, compared with video-based ReID, gait recognition is a more promising solution for addressing the cloth-changing problem.

4.5. Discussion and Further Work

Gait recognition is a potential approach for addressing the cloth-changing problem, and it may be helpful in boosting the performance of ReID. One possible way is to add silhouettes as another input modality, and then combine RGB and silhouettes to learn a shared feature space. Another is utilizing knowledge distillation to encourage the features to approximate its countpart extracted by a pre-trained gait model. As for the pedestrian occlusion, we can take keypoints confidence of pose estimation into consideration, which may help tackle this problem.

5. Conclusion

In this paper, we devote to exploring the different cloth-changing conditions for video-based ReID and gait recognition. Considering the limited work in real scenarios, we do the following things. First, We build a new cloth-changing benchmark named CCPG. It contains 200 subjects in seven different clothes walking across outdoor and indoor scenes, and has over 16,000 sequences, as well as other challenges in real applications. Second, comprehensive experiments are conducted to compare the performance of video-based ReID and gait recognition on the CCPG dataset, and the results indicate gait recognition has greater development potential over video-based ReID on the cloth-changing problem. We hope our work can provide a reference for future work on the cloth-changing problem in the real world. The codes in MindSpore [1] come soon at https://gitee.com/chunjie-zhang/ccpg-cvpr2023.

6. Acknowledgement

References

[1] HUAWEI MindSpore. http://www.mindspore.cn/.

[2] Akansha Bathija and Grishma Sharma. Visual object detection and tracking using yolo and sort. *International Journal of Engineering Research Technology*, 8(11), 2019.

[3] Zhigang Chang, Zhao Yang, Yongbiao Chen, Qin Zhou, and Shibao Zheng. Seq-masks: Bridging the gap between appearance and gait modeling for video-based person re-identification. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2021.

[4] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8126–8133, 2019.

[5] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE transactions on pattern analysis and machine intelligence*, 40(2):392–408, 2017.

[6] Sir John Daniel. Education and the covid-19 pandemic. *Prospects*, 49(1):91–96, 2020.

[7] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14233, 2020.

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

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

[10] Fabian Herzog, Xunbo Ji, Torben Teepe, Stefan Hörmann, Johannes Gilg, and Gerhard Rigoll. Lightweight multi-branch network for person re-identification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1129–1133. IEEE, 2021.

[11] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.

[12] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2014–2023, 2021.

[13] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European conference on computer vision*, pages 382–398. Springer, 2020.

[14] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3958–3967, 2019.

[15] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European conference on computer vision (ECCV)*, pages 737–753, 2018.

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

[17] 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.

[18] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14648–14656, 2021.

[19] Jiawei Liu, Zheng-Jun Zha, Wei Wu, Kecheng Zheng, and Qibin Sun. Spatial-temporal correlation and topology learning for person re-identification in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4370–4379, 2021.

[20] Jiawei Liu, Zheng-Jun Zha, Xierong Zhu, and Na Jiang. Co-saliency spatio-temporal interaction network for person re-identification in videos. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1012–1018. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[21] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3750–3759, 2019.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[23] Sudeep Sarkar, P Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin W Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE transactions on pattern analysis and machine intelligence*, 27(2):162–177, 2005.

[24] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *2016 international conference on biometrics (ICB)*, pages 1–8. IEEE, 2016.

[25] Chunfeng Song, Yongzhen Huang, Weining Wang, and Liang Wang. Casia-e: A large comprehensive dataset for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2022.

[26] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[27] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018.

[28] 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):1–14, 2018.

[29] Daoliang Tan, Kaiqi Huang, Shiqi Yu, and Tieniu Tan. Efficient night gait recognition based on template matching. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 1000–1003. IEEE, 2006.

[30] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2314–2318. IEEE, 2021.

[31] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE transactions on pattern analysis and machine intelligence*, 25(12):1505–1518, 2003.

[32] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European conference on computer vision*, pages 688–703. Springer, 2014.

[33] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid spatial-temporal aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12026–12035, 2021.

[34] Thomas Wolf, Mohammadreza Babaee, and Gerhard Rigoll. Multi-view gait recognition using 3d convolutional neural networks. In *2016 IEEE international conference on image processing (ICIP)*, pages 4165–4169. IEEE, 2016.

[35] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5177–5186, 2018.

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

[37] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 441–444. IEEE, 2006.

[38] Xianghao Zang, Ge Li, and Wei Gao. Multi-direction and multi-scale pyramid in transformer for video-based pedestrian retrieval. *IEEE Transactions on Industrial Informatics*, 2022.

[39] Shaoxiong Zhang, Yunhong Wang, Tianrui Chai, Annan Li, and Anil K Jain. Realgait: Gait recognition for person re-identification. *arXiv preprint arXiv:2201.04806*, 2022.

[40] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4710–4719, 2019.

[41] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20228–20237, 2022.

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

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

[44] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. Gait recognition in the wild: A benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14789–14799, 2021.