CASIA-E: A Large Comprehensive Dataset for Gait Recognition

Chunfeng Song, *Member, IEEE*, Yongzhen Huang, *Senior Member, IEEE*, Weining Wang, and Liang Wang, *Fellow, IEEE*

Abstract—Gait recognition plays a special role in visual surveillance due to its unique advantage, *e.g.*, long-distance, cross-view and non-cooperative recognition. However, it has not yet been widely applied. One reason for this awkwardness is the lack of a truly big dataset captured in practical outdoor scenarios. Here, the "big" at least means: (1) huge amount of gait videos; (2) sufficient subjects; (3) rich attributes; and (4) spatial and temporal variations. Moreover, most existing large-scale gait datasets are collected indoors, which have few challenges from real scenes, such as the dynamic and complex background clutters, illumination variations, vertical view variations, *etc.* In this article, we introduce a newly built big outdoor gait dataset, called CASIA-E. It contains more than one thousand people distributed over near one million videos. Each person involves 26 view angles and varied appearances caused by changes of bag carrying, dressing and walking styles. The videos are captured across five months and across three kinds of outdoor scenes. Soft biometric features are also recorded for all subjects including age, gender, height, weight, and nationality. Besides, we report an experimental benchmark and examine some meaningful problems that have not been well studied previously, *e.g.*, the influence of million-level training videos, vertical view angles, walking styles, and the thermal infrared modality. We believe that such a big outdoor dataset and the experimental benchmark will promote the development of gait recognition in both academic research and industrial applications.

Index Terms—Deep learning, gait dataset, gait recognition, soft biometrics

1 Introduction

Biometric recognition is a hot topic in pattern recognition, and many kinds of biometric features are developed like face, iris and fingerprint, whose raw data are mainly static images [9], [13], [14], [39], [41], [44], [56]. Gait is a kind of behavioral biometric feature whose raw data are walking videos. It is probably the only biometric feature for person identification at a long distance (*e.g.*, more than 50 meters with high-definition cameras), and is a good choice for non-cooperative person identification due to its unique advantage of cross-view recognition, playing a critical role in applications like video surveillance [37], [71].

A general framework of gait recognition usually consists of three components: 1) extracting silhouettes of pedestrians, 2) generating feature representation, and 3) designing similarity measurement. Most studies of gait recognition focus on the second step, and various feature representations are developed like Gait Energy Image(GEI) [17], Gait Entropy Image(GEnI) [2], Gait Flow Image(GFI) [29] and Chrono Gait Image(CGI) [55], which are illustrated in the upper part of Fig. 1. Recent researches show that deep learning makes great progress in gait feature representation. For example, Wu *et al.* [63] find that convolutional neural network (CNN) based features outperform traditional manually designed features in terms of cross-view recognition by about 30%. Some learned feature representations based on CNN are visualized in the bottom part of Fig. 1.

However, even with the special advantage and much progress in cross-view recognition, gait recognition has not yet been widely applied. One reason is that there is no truly big dataset. As a previous backbone of the newly built dataset, CASIA-B [75] is a widely used large-scale cross-view gait dataset, which contains 13,640 videos and 124 people, built ten years ago. With limited gait data, it is really difficult to learn an effective model like the successful cases in face recognition and object&scene classification, where many big datasets, *e.g.*, WebFace [67], MegaFace [27], ImageNet [42], COCO [33] and SUN [64], are built and utilized to design complex models. Generally, video-based gait recognition needs much more samples than face recognition and object&scene classification for model training. However, the number of samples on CASIA-B is much less than datasets for image-based visual tasks. Besides CASIA-B, current datasets that contain more subjects and video sequences are the OU-ISIR series, which are built by the Osaka University. For example, the OU-ISIR LP-Age [65] consisting of more than 60 thousand subjects, is the largest one among all publicly released datasets with respect to the number of subjects.

- *Chunfeng Song, Weining Wang, and Liang Wang are with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Center for Excellence in Brain Science and Intelligence Technology (CEBSIT), Institute of Automation, Chinese Academy of Sciences (CASIA), University of Chinese Academy of Sciences (UCAS), Beijing 100190, China.*
  E-mail: {chunfeng.song, weining.wang, wangliang}@nlpr.ia.ac.cn.
- *Yongzhen Huang is with the School of Artificial Intelligence, Beijing Normal University (BNU), Beijing 100875, China.*
  E-mail: huangyongzhen@bnu.edu.cn.

Manuscript received 13 Mar. 2022; accepted 11 June 2022. Date of publication 15 June 2022; date of current version 3 Feb. 2023.
This work was supported in part by the National Natural Science Foundation of China under Grants 61721004, U1803261, 61633021, and 62006231, and in part by the Shandong Provincial Key Research and Development Program under Grant 2019JZZY010119.
(Corresponding author: Liang Wang.)
Recommended for acceptance by H. Kjellstrom.
Digital Object Identifier no. 10.1109/TPAMI.2022.3183288

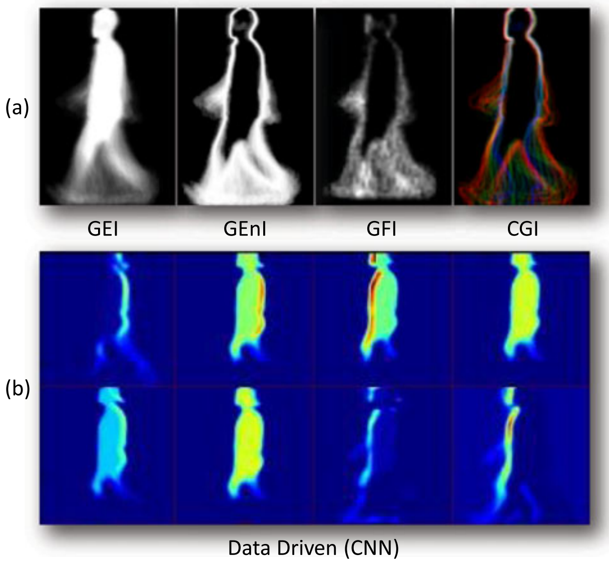


Fig. 1. Visualization of several widely used feature representations of gait. (a) Traditional gait templates and (b) data-driven gait features learned by deep CNN models.

Despite great contributions of the above datasets, they are built under constrained indoor scenes and have few variations for each subject, which greatly limit the practical algorithm evaluation in real challenging outdoor scenes. To promote the development of gait recognition and its application, it is really time to build a truly big outdoor dataset! How to understand the "big" of a gait dataset? At least the following four aspects should be taken into account.

- Although the current state-of-the-art methods [6], [63] largely raise the accuracy record of cross-view gait recognition, their used networks are not deep enough due to the potential over-fitting problem when training with limited videos. It is desired to collect *huge amount of gait videos* to avoid over-fitting and pursue higher recognition accuracy. In addition, a big dataset with *enough categories* is necessary to obtain rich and effective feature representations.
- Even with Big Data in terms of the number of people and videos, it is necessary to collect *various attributes* for each people in order to mimic real environment. To this end, we need a dataset to contain various status of pedestrians, *e.g.*, view angles, walking styles, bag carrying and dressing.
- Another important factor for a big gait dataset is the *variation of space and time* because people appear in different scenes and seasons in real world. The change of scenes will influence backgrounds and then the quality of the extracted silhouettes. The change of seasons will influence dressing and then the shape of silhouettes.
- Finally, it is better to contain multiple kinds of *soft biometric features* like age, gender, height and weight. These soft biometric features are helpful for other related gait analysis and examination.

To our knowledge, currently there is no such a dataset satisfying all of the above-mentioned requirements. Some datasets involve changes in attributes of each person but are limited in the number of subjects and view angles, e.g., CMU MoBo [15] and USF [44]. Some datasets contain sufficient subjects but limited changes in attributes, e.g., OU-ISIR LP [23]. In addition, some recently published large-scale datasets [50], [65] are collected indoors which have no challenges from real scenes, such as the dynamic and complex background clutters, illumination variations, vertical view variations, *etc*. More analysis on the existing gait datasets will be provided in the next section. Without such a big outdoor dataset, it is really difficult to train an effective model which is expected to be generalized well for complex real-world scenarios, *e.g.*, a road full of pedestrians with different views, dressing, carrying bags or not and walking styles.

In this paper, we introduce a newly built outdoor gait dataset called CASIA-E. More than one thousand volunteers contribute to this dataset, distributed over near one million videos, nearly 50 times larger than the previous one. It includes 26 views and rich changes for each pedestrian, *e.g.*, bag carrying, dressing and walking styles. Also, this dataset records soft biometric features like age, gender, height and weight. The videos are captured over five months and in different scenes, one of which is "dynamic outdoor" with complex backgrounds including moving cars, bikes, cloud and multiple pedestrians. This dataset takes a meaningful step forward to exploit complex and powerful models (*e.g.*, deep neural networks) for gait recognition. We also evaluate several basic algorithms on this dataset with respect to most attributes and their combinations, providing an experimental benchmark and obtaining some meaningful findings.

The organization of this paper is as follows. In Section 2, we analyze related work of this paper, including previous gait datasets and popular algorithms. In Section 3, we explain how to build CASIA-E, including its layout and statistical analysis. Experimental benchmark and results are provided in Section 4. Finally, Section 5 concludes the paper and points out future directions.

2 Related Work

In this section, we discuss two related aspects: datasets and gait feature representations.

2.1 Dataset

A good dataset is critical for training gait recognition algorithms and evaluating their performance. Generally speaking, there are four dimensions to analyze the variety of current gait datasets.

- The number of subjects. This factor is undoubtedly important. OU-ISIR LP-Age [65], consisting of more than 60 thousands subjects, is the largest one in all publicly released datasets. Other datasets are generally with only hundreds or less.
- The number of videos. To learn robust representations, it is necessary to capture enough samples for each subject. Take CASIA-B [75] as an example, although it has only 124 subjects, it includes 13,640 videos, more than most gait datasets. This means that each subject involves more than one hundred videos, which is a critical factor to learn good representations for gait recognition.

TABLE 1
Summary of the Existing Gait Datasets' Attributes

Dataset	#Subjects	#Sequences	#Views	Environment	Other factors
MIT [45]	25	210	1	Static indoor	None
CMU MoBo [15]	25	600	6	Static indoor	Speed, bag carrying, surface incline
Georgia Tech [25]	15	268	-	Static outdoor	-
	18	20	-	Static indoor	-
UMD HID 1 [26]	25	100	1	Static outdoor	None
UMD HID 2 [10]	55	222	2	Static outdoor	None
SOTON Small [47]	12	-	3	Static indoor	Bag carrying, dressing, shoes
SOTON Large [47]	115	2,128	2	Static indoor/outdoor	None
SOTON Multimodal [40]	>300	$>5{,}000$	12	Static indoor	-
SOTON Temporal [43]	25	2,280	12	Static indoor	-
USF [44]	122	1,870	2	Static outdoor	Bag carrying, surface, shoes
OU-ISIR A [36]	34	612	1	Static indoor	Walking speed
OU-ISIR B [36]	68	2,764	1	Static indoor	Dressing
OU-ISIR C [36]	200	5,000	25	Static indoor	None
OU-ISIR D [36]	185	370	1	Static indoor	Gait fluctuation
OU-ISIR LP [23]	**4,007**	7,842	2	Static indoor	None
OU-ISIR MVLP [50]	**10,307**	288,596	14	Static indoor	None
OU-ISIR LP-Age [65]	**63,846**	63,846	1	Static indoor	Age
TUM GAID [21]	305	3,370	1	Static indoor	Bag carrying, shoes
WOSG [11]	155	684	8	Dynamic outdoor	Illumination
CASIA-A [71]	20	240	3	Static outdoor	None
CASIA-B [75]	124	13,640	11	Static indoor	Bag carrying, dressing
CASIA-C [71]	153	1,530	1	Static outdoor	Bag carrying, walking speed
CASIA-E(Ours)	**1,014**	**778,752**	**26**	**Multiple outdoor**	**Bag carrying, walking style, dressing and soft biometric features**

The most remarkable attributes are marked in blue.

- The number of views. In the real scenes of using gait recognition, the view change is very common, and thus a dataset should contain different views of gait videos for learning cross-view recognition models. For example, OU-ISIR C [36] has 25 views, OU-ISIR MVLP [50] has 14 views.
- The variety of subjects and scenes. Here the variety of subjects is mainly caused by the change of silhouettes due to different dressing, bags and walking styles, which are designed by various datasets more or less [11], [15], [21], [36], [44], [47]. Besides, to better simulate the real scenes, some datasets are built in the outdoor environment, but most of them are with simple and static backgrounds.

In Table 1, we list the main attributes of the existing datasets related to gait recognition. Besides these datasets, there are also some similar gait datasets [8], [35], [73] that are related to this work. Although so many datasets have been built, most of them only consider a part of potential attributes. Some datasets have rich changes in dressing, walking and bag carrying but not enough subjects and views, *e.g.*, CMU-MoBo [15] with 25 subjects and 6 views, SOTON-Small [47] with 12 subjects and 3 views, USF [44] with 122 subjects and 2 views, and TUM [21] with 305 subjects and only 1 view. Some datasets contain many subjects but limited changes in other attributes, *e.g.*, OU-ISIR LP [23] is composed of 4,007 subjects but with only 2 views in a static indoor scene, totally with around 8,000 videos. Its extended version OU-ISIR LP-Age [65] has an amazing number of subjects but with only 1 video for each subjects. The most relevant one is OU-ISIR MVLP [50], which has both large numbers of subjects and views, with roughly 288 thousand indoor videos. Recently, a model-based large-scale dataset named OUMVLP-Pose [1] is proposed for multi-view gait recognition with human pose-based cues, which is also built upon OU-ISIR MVLP [50]. Though above three datasets contribute greatly to gait recognition with so many subjects and views, however, they are collected in static indoor scenes, which have few challenges from real scenes, such as the dynamic background clutters and illumination variations.

To promote the gait recognition into practical application, it is necessary to build a new outdoor gait dataset that contains these diverse attributes as many as possible. This is our main motivation of designing and establishing CASIA-E. We make our best efforts to cover the important factors of a big gait dataset. It contains 1,014 subjects, which is not as big as that of OU-ISIR MVLP [50] but each subject involves more than 700 videos and in total 778,752 videos, nearly 3 times more than OU-ISIR MVLP [50] and 100 times more than OU-ISIR LP [23]. In addition, we design 26 views (a combination of 13 horizontal and 2 vertical views), three appearance changes for each subject (bag carrying, dressing and walking styles), and three different outdoor backgrounds, including complex scenes with moving cars, bikes and other pedestrians.

2.2 Gait Feature Representation

Over the past 20 years, various classic approaches have been proposed [5], [52], [53], [54], [57], [68] to address the problems involved in gait recognition like feature representations [2], [17], [29], [55], model and metric/similarity learning [3], [4], [59], [66], of which designing feature representation for gait recognition attracts most attention. Here, we focus on feature representations which are more related to gait recognition. Other factors like model and metric/similarity learning are not specific for gait recognition because these factors also exist in other areas like face recognition and general object classification. Accordingly, in this part, we only focus on gait feature representation, which is also mainly evaluated in the experimental section.

2.2.1 Manually Designed Feature Representation

Feature-based gait recognition methods often extract features from binary silhouettes [24], [58]. Specific gait templates are then produced through properly dealing with a sequence of silhouettes over a gait cycle. Some classic and widely used templates include Gait Energy Image (GEI) [17], Gait Entropy Image (GEnI) [2], Gait Flow Image (GFI) [29] and Chrono Gait Image (CGI) [55]. Examples of these templates are shown in Fig. 1. A recent empirical study [23] by Iwama *et al.* shows that GEI, despite of its simplicity, is the most stable and effective feature for gait recognition. In addition, there are some works trying to enhance GEI features. For example, Wang *et al.* apply Principal Component Analysis (PCA) [58] over GEI for cross-view gait recognition. Guan *et al.* [16] try to process GEI with Linear Discriminant Analysis (LDA). The method presented in [22] compresses GEI with Locality Preserving Projection (LPP).

2.2.2 Deep Learning-Based Feature Representation

Due to the success of deep learning in various computer vision tasks [18], [19], [28], [32], [34], [48], [49], [62], [74], some data-driven methods are introduced to gait recognition. These methods usually learn better feature representations. Recently, Wu *et al.* [63] propose a deep two-stream convolutional neural network (TS-CNN) based framework to learn similarities between pairs of GEIs for cross-view gait recognition, achieving the state-of-the-art performance. Another famous CNN-based method GEI-Net [46] directly learns gait representations from GEIs and then corresponds to identities. Following these works, there are numbers of deep learning-based methods [20], [30], [70] are proposed. GaitGAN [69] and TS-GAN [60] introduce the generative adversarial networks for cross-view gait recognition. A recent review [51] has evaluated the influence of different input and output architectures. Besides, some researchers propose to use deep convolutional neural networks [6], [61] to learn gait representations from original silhouettes rather than GEIs. Recently, a new study that focuses on the temporal part-level gait feature learning has been proposed [12]. In the experimental part, we choose the original GEI gait feature and the deep learning method for evaluation. Considering the super large amounts of gait sequences (778 k) in our dataset, it is hard to implement all SOTA benchmark methods. Among them, the GEI-Net [46] is the simplest yet popular model which can represent the CNN based models, while TS-CNN [63] is the first one to introduce the pairwise metric learning for gait recognition and is widely cited as one of current SOTA methods. Therefore, we select them as the default deep learning-based methods for evaluation.

According to our experience, with enough data and well-tuned training process, deep learning-based feature representation usually outperforms manually designed feature representation in terms of recognition. However, deep learning-based feature representation generally needs a large scale of data to achieve good performance and costs more time in both training and testing stages. Besides, it relies on the skills of network design and training. That is, if the network is not well designed or sufficiently trained, the performance will largely deteriorate.

3 Building CASIA-E

CASIA-E is the newest member of the CASIA gait family that has been built over the past 17 years. It is much larger than the previous members, containing 1,014 persons distributed in varying outdoor scenes with different horizontal views, vertical views, dressing and walking styles. For each person, there are 768 videos on average. The default resolution of the video is 1920×1080 pixels at 25 FPS. In the rest of this section, we will introduce CASIA-E in detail including layout, scenes, view, dressing and walking styles, and provide the statistical description of subjects like age, gender, height and weight.

3.1 Layout

In the real environment where gait recognition is applied, a person may be registered in one view angle but required to be recognized in another view angle. Therefore, it is necessary to design a camera array to cover view changes as many as possible. Besides, the height of cameras to be fixed is not controllable. Usually, the height changes from one meter (indoor) to four meters (outdoor) according to the environment.

It is not possible naturally to cover all horizontal and vertical views. In CASIA-E, we use a array with eight cameras to capture 26 views: 13 horizontal (from 0 ° to 180 ° at an interval of 15 °) by 2 vertical views (1.2 m and 3.5 m). Due to the real space limitation, the distance of cameras to the walking pedestrian is roughly 4 meters for Scene#1, and meters for Scene#2 and Scene#3. In addition, the walking line of Scene#1 is limited to 5 meters, while Scene#2 and Scene#3 are roughly 7.5 meters. The layout is illustrated in Fig. 2. For a large-scale standard dataset, it is necessary to set rules for volunteers to ensure the quality and consistency of the whole dataset. When subjects walk along Line-1, four cameras capture the view angles of 45°, 60°, 75° and 90°. Along Line-2, cameras capture 135°, 120°, 105° and 90°. Along Line-3, cameras capture 0°, 15°, 30° and 45°. Along Line-4, cameras capture 180°, 165°, 150° and 135°. In total, we get 13 horizontal views in total (45°, 90° and 135° appear twice). And finally, we get 26 views (13 horizontal views by 2 vertical views).

3.2 Scenes

CASIA-E contains three outdoor scenes, *i.e.*, simple static background (Scene#1), complex static background (Scene#2) and complex dynamic background (Scene#3), as shown in Fig. 5. Scene#1 and Scene#2 are relatively simpler whereas Scene#3 is very complex, containing moving cars, bicycles, pedestrians under dynamic backgrounds with cloud moving, weather change, time change (from morning to afternoon) and season change (from May to September, crossing three seasons of Beijing, *i.e.*, Spring, Summer and Autumn). The lighting at different scenes and capturing time vary greatly, which usually exist in surveillance systems. For the

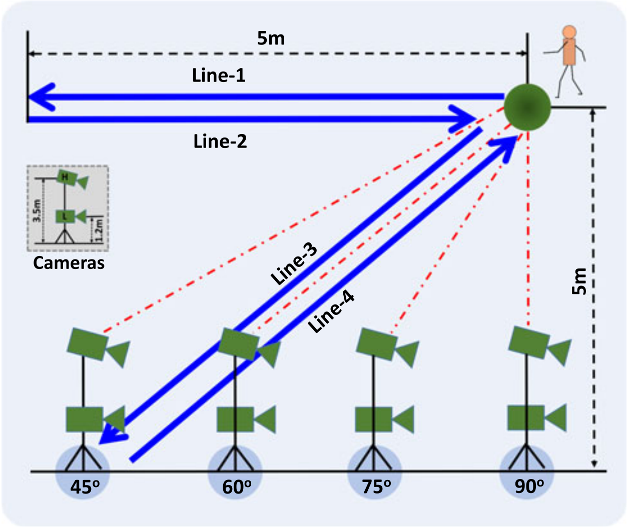


Fig. 2. The layout of cameras. A set of two cameras are fixed in a tripod. In each set, the low and high cameras are with different vertical views. The four blue solid lines illustrate the walking routes of subjects. Please see text for more details.

same person in different scenes, the length of gait sequence is also different, which is caused by the distance of cameras and walking lines. For example, a gait sequence with 90° and L-height in Scene#1, Scene#2 and Scene#3 contains roughly 100, 150 and 200 frames, respectively.

In addition, to explore the influence of different modalities on gait recognition, we also collect a dataset with the *thermal infrared camera*. It contains 270 subjects with two walking directions. We add this subset as a part of the CASIA-E dataset, with a new scene which has not been evaluated in gait recognition, i.e., the thermal infrared scene, as shown in Fig. 14.

3.3 View, Dressing, and Walking Style

View, dressing and walking style are three main factors influencing the performance of gait recognition.

As explained in Section 3.1, for each subject, we design 13 horizontal views and 2 vertical views, and in total 26 views. According to our investigation (see Table 1), this dataset involves the most views in all gait datasets. Three kinds of dressing are taken into account in all scenes: NM, CL and BG. Following CASIA-B, NM means "normal" gait, CL means "clothing change" based on NM, and BG means "bag carrying" based on NM. In the collection of the dataset, we have asked all individuals to prepare a coat and bag of themselves for capturing. In case of that some individuals may forget taking such props, we also provide more than ten kinds of different clothing and bags for them. In this way, we could make the dressing conditions as rich as possible. Consequently, there are two sequences for each dressing condition. These dressing conditions are common in practice. Besides, in Scene#3, we add an extra walking case: stopping 1∼2 seconds when walking. This requirement is to mimic subjects' action of receiving a call or looking around. Thus, there are two kinds of walking styles in Scene#3. Fig. 3 shows some examples. In addition, we have captured

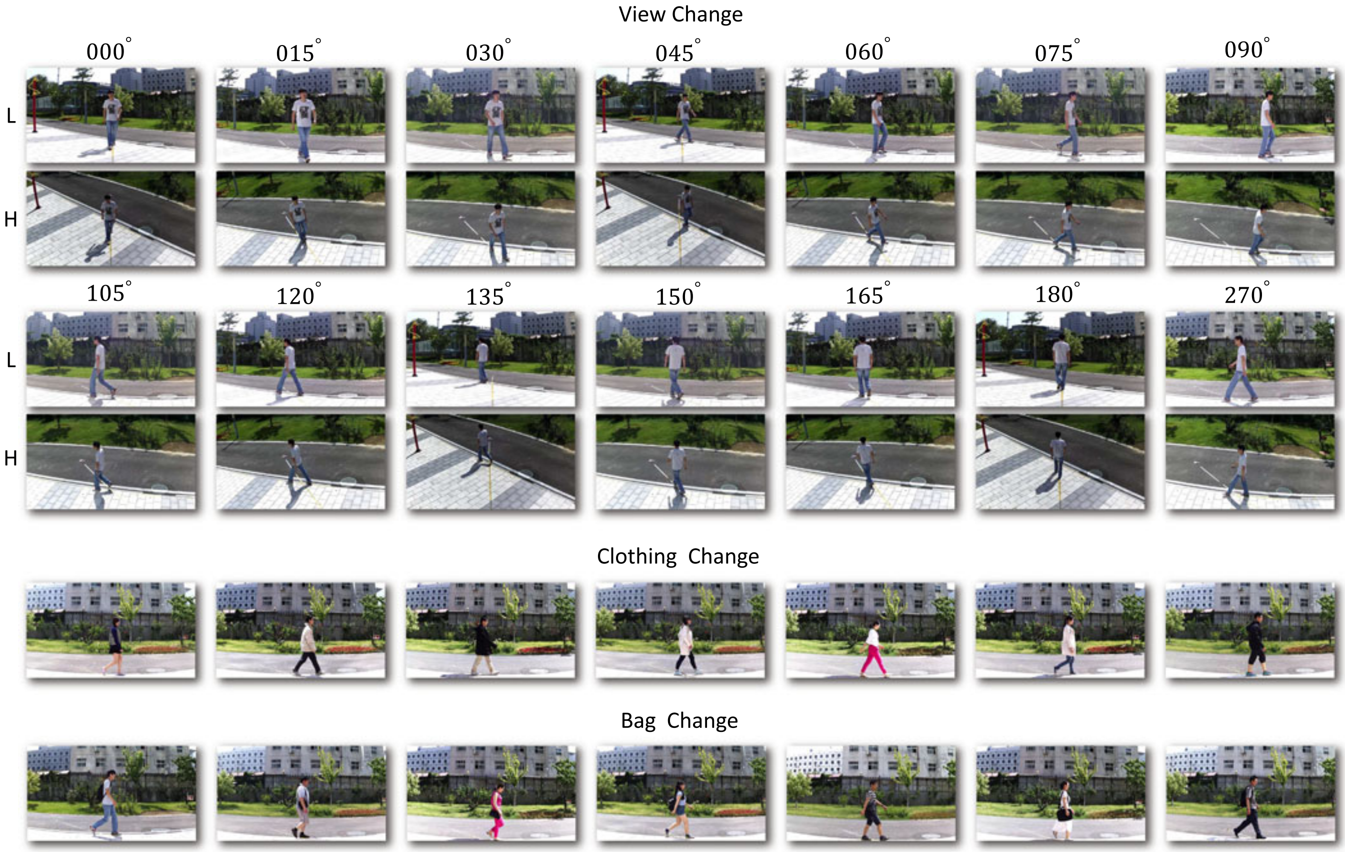


Fig. 3. Example images of CASIA-E, reflecting view changes (horizontal and vertical views), clothing changes and bag changes from top to down. "L" and "H" indicate videos captured by the low and the high cameras, respectively.

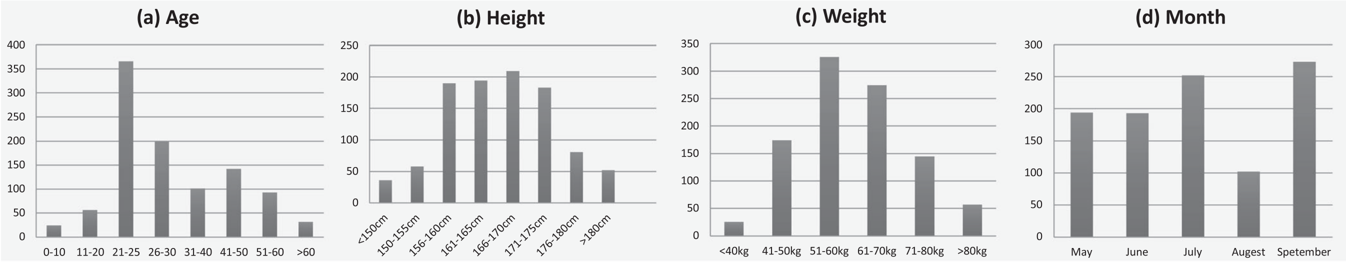


Fig. 4. Distributions of subjects with respect to age, height, weight, and month.

videos in unconstraint conditions, in which multiple people are walking freely. These videos would be a part of our new dataset.

3.4 Statistics of Attributes

CASIA-E contains 1,014 subjects with big variations in age, gender and body shape. We provide statistical analysis below.

- *Statistics of age.* The age of subjects varies from 3 to 85 years old. There are 32 subjects older than 60 years old and 25 subjects younger than 10 years old. As shown in Fig. 4a, most subjects are distributed from 21 to 60 years old. Though the most ideal distribution is the uniform distribution, the age distribution is consistent with the practical case of most monitoring systems.
- *Statistics of gender.* The ratio of male subjects and female subjects is strictly 1:1. That is, 507 males and 507 females exactly.
- *Statistics of height.* The height of subjects varies from 104 cm to 195 cm. As shown in Figs. 4b, most subjects are with the height from 150 cm to 180 cm.
- *Statistics of weight.* The weight of subjects varies from 15 kg to 105 kg. As shown in Figs. 4c, most subjects are with the weight from 51 kg to 70 kg.
- *Statistics of month.* Collecting the whole dataset takes 5 months (from May to September) as described in Fig. 4d, involving three seasons, *i.e.*, Spring, Summer and Autumn. The data domains under different seasons are different greatly, in which the model adaptation ability could be explored. Thus, CASIA-E can be used to study the cross-domain gait recognition. For example, we could use the data captured in Spring for training, and use the data captured in Summer or Autumn for testing. It does not mean that each individual is collected from time to time, while the whole dataset has sufficient diversities in terms of capturing seasons/months.

4 Evaluation on CASIA-E

CASIA-E is so big a dataset for gait recognition that it is expected to provid its experimental benchmark. On one hand, it is possible to find some interesting and meaningful conclusions with such a big dataset. On the other hand, it would be convenient for other researchers to evaluate and compare their proposed algorithms of gait recognition on CASIA-E. To this end, we provide an extensive experimental study on CASIA-E.

In this section, we first introduce the experimental setting and then report the experimental performance of some traditional and newly developed algorithms, followed with the studies of each attribute, including training samples, dressing, scene and walking style. Then, we focus on cross-view related evaluation, as well as the combination of cross-view and cross-clothing change, *i.e.*, taken them into account together for joint evaluation. These two factors often appear in practical applications.

4.1 Experimental Setting

Dataset Partition. There are in total 1,014 people on CASIA-E, among which ID001-ID500 are used for training, ID501-ID614 for validation, and ID615-ID1014 for testing. Some gait recognition methods (especially deep learning-based algorithms) usually need a validation set to obtain a good training model. If there is no such need, all samples in the validation set can be used for training. For each subject, there are 384 kinds of variants, (3 scenes + 1 walking-style) × (14 horizontal views + 2 repetitive views) × (2 vertical views) × (3 wearing conditions: NM/CL/BG). For each variant, we capture 2 sequences, among which the first one is in the gallery and the second is for probe. For a set of testing videos, there are at least two sequences, each of which is with specific view, dressing and scene. The probe and gallery videos have already been divided in the dataset, and we report the average recognition rates (ACC) in terms of different ranking orders.

Data Pre-Processing Pipeline. All GEI maps of videos are already provided in CASIA-E, which are extracted via the following steps. The pipeline is shown in Fig. 6.

First, for a given image from videos, the person is detected with the ACF detector and the image in the bounding box would be cropped into a person image. Since the original frame has a resolution of 1920×1080, the height of cropped boxes vary from 360-540 pixels corresponding to their body heights.

Second, we use the deep learning-based segmentation algorithm [62] to obtain the silhouette from the image. Note some state-of-the-art segmentation methods [7], [18], [32] mostly focus on the common semantic segmentation tasks which usually contain multiple categories, while the human segmentation only has one. In addition, these methods usually need a large-scale dataset for training, e.g., MSCOCO [33] dataset, which further limit their application in the human segmentation task. We show some segmentation examples of our method which is slightly adjusted based on the CNN model [62] in Fig. 9. For a comprehensive comparison, we also evaluate several state-of-the-art methods, including the instance segmentation method Mask R-CNN [18], the semantic segmentation method DeeplabV3+ [7], as well as the human parsing method JPPNet [31]. It is encouraging that the pre-trained segmentation methods perform really well on our dataset, despite some negligible errors. Note that all these methods have not been fine-tuned on our dataset, it would be possible for them to learn better segmentation with sufficient human segmentation data. For instance, we could finetune the pre-trained model (e.g., Mask-RCNN) on the COCO dataset with only one *person* class. Another possible solution resorts to the domain adaptation learning, with which the domain gap between the segmentation training data and the gait data could be narrowed. Among above pre-trained models, the JPPNet method that has been trained on a human parsing dataset performs best, showing the great potential as the candidate for robust gait silhouette extraction. In most conditions, our method could get satisfied silhouettes, while it is really difficult in Scene#3 which has more complex dynamic objects in the background. For each frame of silhouettes, we calculate the distance between the top and the bottom pixels as the height of silhouette.

Third, we draw a rectangle box whose center is the gravity of the silhouette. The height of the rectangle box is calculated in the second step. The aspect ratio of the rectangle box is 11/16. And then we crop off the region within the rectangle and resize it into a fixed size (88×128), so as to generate the normalized silhouettes.

Finally, we use the normalized silhouettes from the whole video to compute its GEI. This is a little different from the classic GEI generator that uses a gait cycle [58] in this step because we empirically find that using the whole video works comparably better in most experiments. In fact, the gait cycle segmentation is difficult in outdoor scenes, which is mainly due to the unstable detection and semantic segmentation. Meanwhile, it should be noted that without the limit of the gait cycle, it will be more convenient to calculate GEI in practice. As shown in Fig. 7, the quality of GEI of H and L views are different, e.g., the L view is more stable, while the H view contains noises due to its flat visual angles.

4.2 Baseline Evaluation

In the baseline experiment, we choose three kinds of feature representations with different numbers of training samples and ranking orders for evaluation. The first one is the original GEI features [17]. As we discussed in Section 2, although many kinds of feature representations are manually designed, the GEI features are very stable and effective for gait representation. The second and the third ones are GEI-Net [46] and the Deep Two-Stream CNN (TS-CNN for short) with the MT model used in Wu *et al.*'s work [63], whose structures are shown in Fig. 8. These two methods are the representatives of deep learning-based algorithms for gait recognition, and achieve good performance.

In order to fit the large-scale dataset, we slightly adjust the TS-CNN model to get better performance. First, we double the filters in the C1 layer to catch more local information. Second, we enhance the C3 layer, *i.e.*, change the filter size of C3 to 3×3 and add a 1,024-dimensional fully-connected layer (FC4) after C3. The TS-CNN model receives pairs of GEIs as its inputs with a balanced strategy. That is, the numbers of positive and negative samples are almost equivalent. We label a pair as 1 if the gallery and the probe are with the same ID, else as 0. The network is trained using back-propagation with the cross-entropy loss, and the weights are updated with a mini-batch size of 128. The initial weights in each layer are from a zero-mean Gaussian distribution with standard deviation 0.01. The bias terms in convolutional and fully connected layers are initialized as zeros. The learning rate starts at 0.01, and is reduced to 0.001 when ACC on the validation set stops increasing.

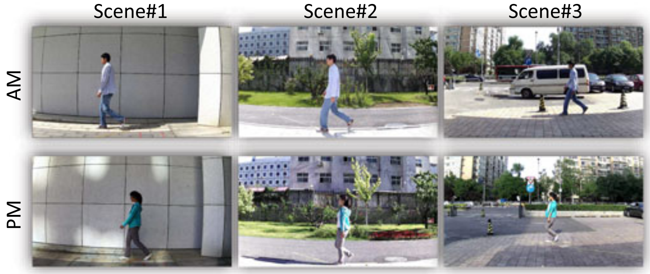


Fig. 5. Example images of the three outdoor scenes of CASIA-E. From left to right are simple static background, complex static background and complex dynamic background, respectively. Note that the lighting condition is varying at different time, e.g., capturing at AM and PM would meet quite different lighting directions, which naturally increases the proposed dataset's difficulty and complexity.

The baseline condition is defined as: "Scene#1"+ "view angle: 90"+ "normal dressing"+ "normal pose". The number of training subjects in the baseline evaluation changes from one hundred to five hundreds, and in Fig. 10 we report the ACC with top 1, 3, 5 and 10 retrieval when using different numbers of subjects for training.

For GEI+PCA, the number of training samples is insensitive probably because that PCA has limited capability to fit a very large dataset. Therefore, simply increasing the number of training samples has little contribution to enhancing the recognition accuracy. In other words, the scale of datasets will not obviously influence the performance for this kind of features.

For TS-CNN and GEI-Net, ACC becomes better as the number of training samples increases, which reflects the value of a big gait dataset for deep learning-based methods. Deep Two-Stream CNN performs better than GEI-Net probably due to the two-stream metric learning network structure and well-controlled over-fitting (using pairs of GEIs as inputs).

In the rest of experiments, we mainly report the results with TS-CNN using 500 people for training because its good performance in the baseline evaluation.

4.3 Cross-Scene Evaluation

In this subsection, we analyze the influence of scenes. The differences of three scenes are summarized as follows.

- The complexity of three scenes is different. As shown in Fig. 5, Scene#1 is simple, with clear background, and Scene#2 is more complex than Scene#1 because the background is full of other objects like trees, grass, buildings and their shadow. Scene#3 is very challenging not only because the background is full of additional objects like cars, buses, bicycles, street lamps and cloud but also because the background is

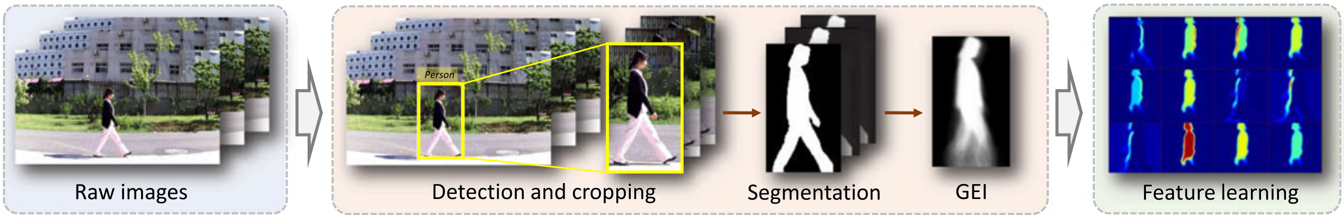


Fig. 6. The pipeline of data pre-processing. The raw images are first cropped according to the detection bounding boxes. Then the person images is segmented into binary silhouettes to produce the GEIs for final feature learning.

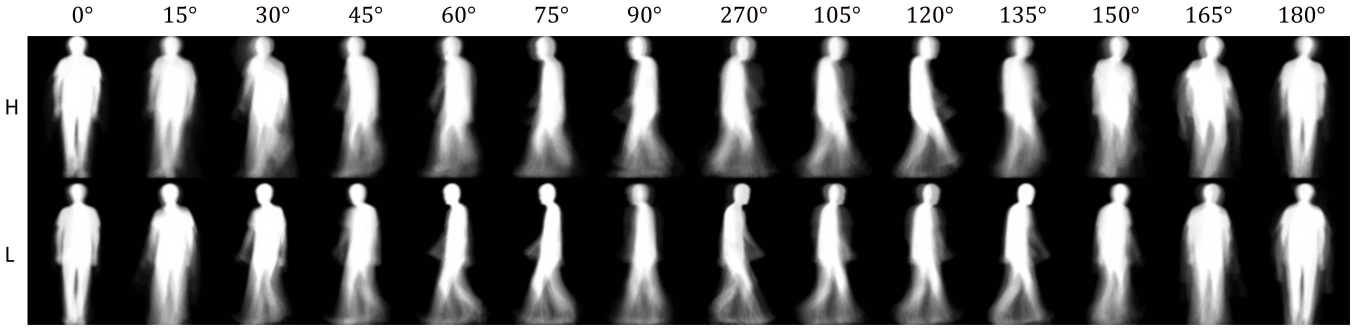


Fig. 7. Examples of GEIs with different vertical and horizontal views. “L” and “H” indicate videos captured by the low and the high cameras, respectively.

not static (with some moving objects like cars, bicycles and pedestrians), which potentially influences the quality of the extracted silhouettes.

- The setup of cameras in three scenes is different. It is hard to keep the setup of cameras (*e.g.*, angle and location) strictly the same in three scenes. Besides, the distance between cameras and subjects is four meters in Scene#1 due to the limited space and five meters in both Scene#2 and Scene#3.
- The quality of videos in three scenes is different, which is mainly caused by illumination. In Scene#1,

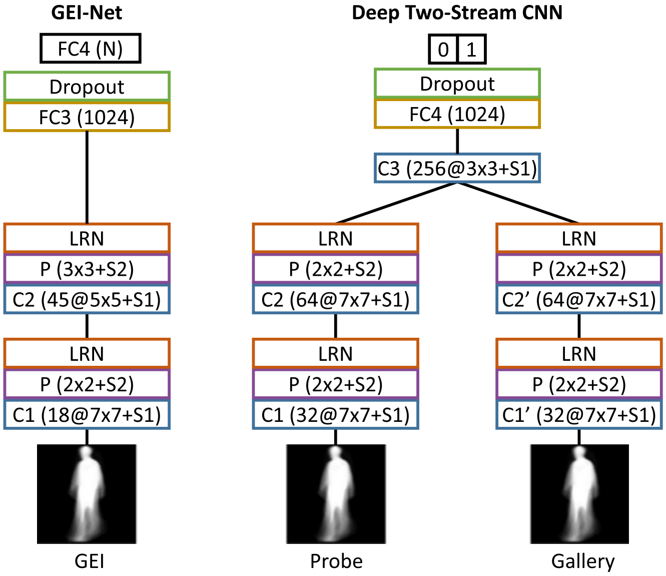


Fig. 8. The structures of GEI-Net and Deep Two-Stream CNN model. C: convolution layers; FC: fully-connected layers; P: pooling layer; LRN: local response normalization layer; S: stride of convolution or pooling; C1 (32@7×7+S1) means that C1 is a convolutional layer with 32 filters. Each filter is with a size of 7× 7, and the stride S is 1.

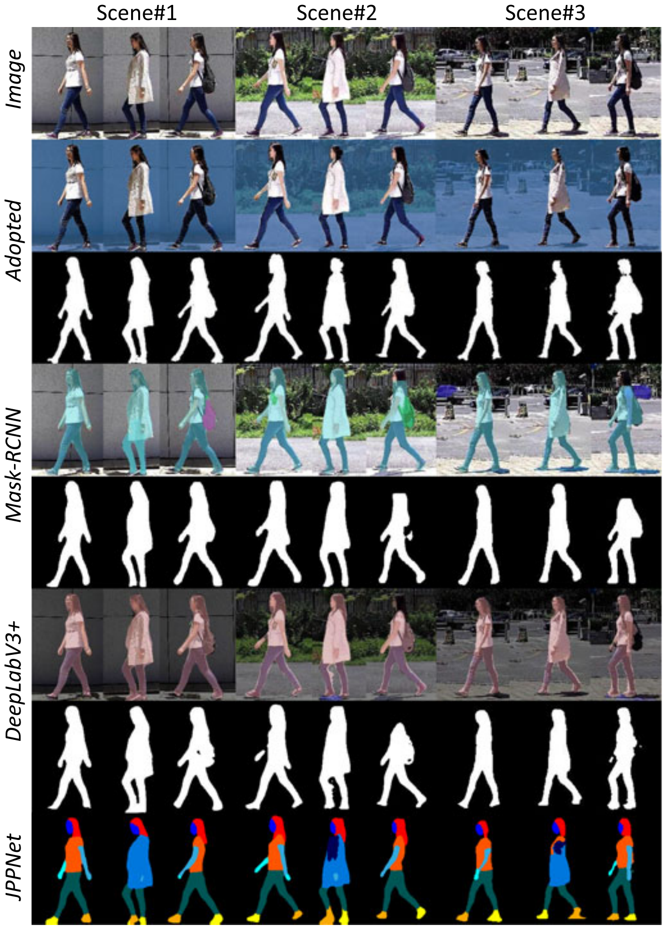


Fig. 9. Examples of segmented masks in three outdoor scenes. The adopted method is compared with Mask R-CNN [18], DeeplabV3+ [7], and JPPNet [31]. As Scene#3 contains dynamic and complex objects in the background, the segmentation is slightly worse than the other two scenes.

TABLE 2
Performance (%) of Deep TS-CNN in Cross-Scene Evaluation

Gallery	Scene#1			Scene#2			Scene#3		
Probe	Scene#1	Scene#2	Scene#3	Scene#1	Scene#2	Scene#3	Scene#1	Scene#2	Scene#3
ACC@top1	92.19	75.75	52.14	77.58	97.75	66.75	59.95	72.50	79.85
ACC@top10	97.99	94.50	76.32	95.72	99.00	85.64	84.38	88.50	92.19

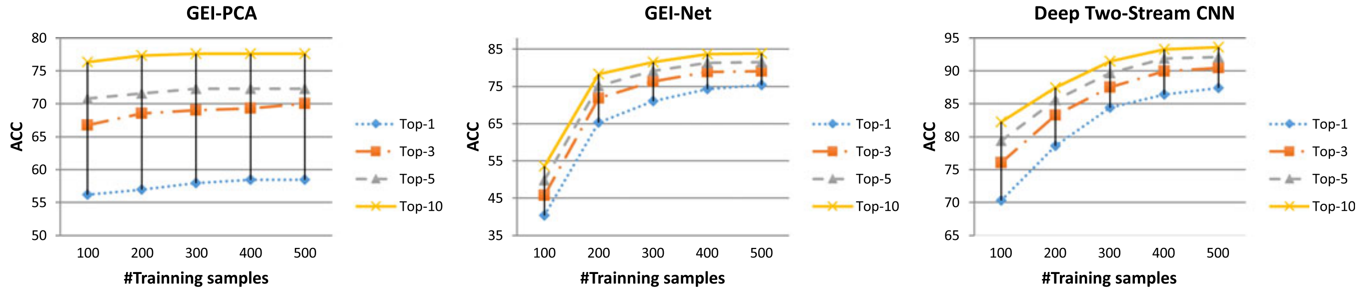


Fig. 10. Baseline performance (%) of different feature representations on CASIA-E in Scene#1.

illumination is well controlled due to its simplicity. But in Scene#2 and Scene#3, it is not easy to maintain a stable illumination environment. The sunshine and the shadow would change the brightness of subjects. In addition, when objects from different scenes are with similar color to the clothing of subjects, the quality of the extracted silhouettes will deteriorate.

We report the cross-scene experimental results in Table 2. Videos from these three scenes are used for training and testing. From the experimental results, we can draw the following conclusions.

1) The algorithm is sensitive to scenes. For example, when the gallery and probe samples are both from Scene#1, ACC@top1 is 92.19%. This metric decreases to 75.75% once the probes are from Scene#2, and decreases to 52.14% when the probes are from Scene#3. Similar results can be observed for other cases of cross-scene evaluation. The complexity of scenes influences the quality of silhouettes. This experiment indicates that complex scenes is a big challenge to gait recognition because it causes bad silhouettes.
2) The average performance in three scenes is: Scene#2 > Scene#1 > Scene#3. It is a little strange that ACC with the gallery from Scene#2 is higher than that from Scene#1, even though Scene#1 is relatively simpler. This is probably because the videos captured in Scene#1 is shorter due to the limited space between cameras and subjects. With less silhouettes in each video captured in Scene#1, the quality of GEI in this scene is not as good as that in Scene#2 because we use the average image of all silhouettes to generate the GEI.

The challenges of cross-scene recognition remind us to exploit more stable solutions in addressing the problem of extracting accurate silhouettes. One way is to enhance the object segmentation algorithms. Another way is to decrease the reliance on the quality of silhouettes by jointly learning segmentation and recognition.

Readers may find that ACC in Table 2 is higher than that in Fig. 10 with the same number of training people. This is because in the cross-scene experiment, we use videos from three scenes for training, which is approximately three times of those in the baseline experiment with videos only from Scene#1 for training.

4.4 Cross-Walking-Style Evaluation

As there are few gait recognition related works about the walking-style, whereas it does exist in real-world application and usually affect the producing of traditional GEIs. Cross-walking in this paper specially means walking with a short stop. Besides the view noise introduced in the short stopping, the main challenge of cross-walking-style is the pose variations. With a stopping action, the pose sequence would be highly different from normal walking actions, and hence it could result in large intra-subject variations. Therefore, in this work, we introduce such a new challenging condition which often happens in real life and explore its influence. In Scene#3, subjects are asked to stop one time at any location when walking from one point to another point. We show some examples in Fig. 11. In this subsection, we analyze the influence of such a walking style in Scene#3, and report the cross-walking results in Table 3.

The experimental results show that the change of walking style decreases the performance from 2.83% (minimal change) to 8.53% (maximal change) for ACC@top1, and from 2.04% (minimal change) to 3.51% (maximal change) for ACC@top10, which is not so sensitive as cross-scene. This result is not difficult to understand. The GEI feature used in CASIA-E is the accumulation of a set of silhouettes. Therefore, "stop" or "non-stop" will not greatly influence the final representation of the GEI feature.

4.5 Cross-Dressing Evaluation

In this subsection, we analyze the influence of dressing change. Videos from three scenes are used for training and testing.

The results are listed in Table 4, from which we find that bags and dressing will severely decrease ACC. Here, NM means "normal," BG means "carrying bag" and CL means "clothing change". For example, comparing with NM-to-NM, ACC in NM-to-BG decreases 10.3% at ACC@top1 and

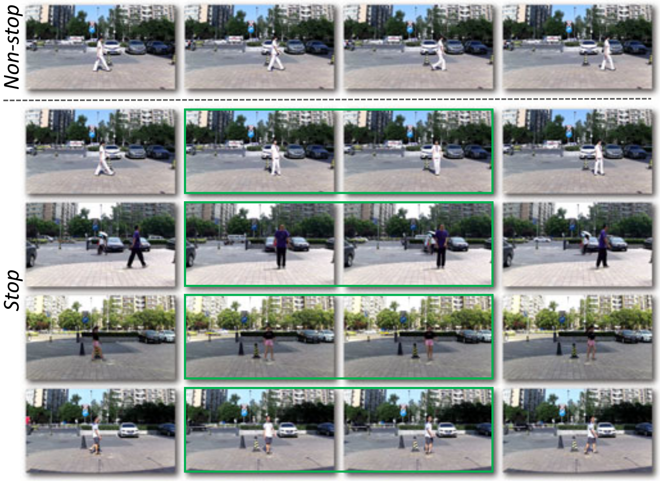


Fig. 11. Example videos of cross-walking-style. The first row denotes a gait sequence without stop, and the rest rows are examples of gait sequences with stop.

TABLE 3
Performance (%) of TS-CNN in Cross-Walking-Style Evaluation

Gallery	Stop		Non-stop	
Probe	Stop	Non-stop	Stop	Non-stop
ACC@top1	88.13	79.60	76.26	79.09
ACC@top10	96.71	93.20	90.15	92.19

4.9% at ACC@top10 on average; and ACC in NM-to-CL decreases 11.2% at ACC@top1 and 6.2% at ACC@top10 on average. From these results, we find that cross-dressing gait recognition is a very challenging problem. CASIA-E involves rich variations about cross-dressing as shown in Fig. 3. It contains more than ten categories of gait videos with changes of bags and clothing.

Both cross-scene and cross-dressing change the silhouettes of subjects. It is interesting that cross-scene has more negative influence although cross-dressing will change the silhouettes more. This is probably because that human and machines have different principles on prior knowledge for gait recognition. Cross-scene will change silhouettes more in a "global" way. However, cross-dressing (caused by bag and clothing change) obviously changes silhouettes more in a "local" way. Machines are not good at global visual perception, and thus machine algorithms are more sensitive to cross-scene recognition. In contrast, after learning a lot of cross-dressing samples, algorithms can know the importance of each part of silhouettes, so as to decrease the influence of clothing change.

TABLE 5
Performance (%) of TS-CNN in Cross-Vertical-View Evaluation

Gallery		H		L	
Probe		H	L	H	L
ACC@top1	Scene#1	89.14	75.82	74.49	91.18
	Scene#2	93.25	68.50	71.25	93.50
	Scene#3	91.18	45.59	55.42	76.57
ACC@top10	Scene#1	97.22	93.95	98.49	98.49
	Scene#2	97.75	91.50	93.50	97.25
	Scene#3	91.41	75.32	84.13	92.70

H: Video captured by the high camera. L: Video captured by the low camera.

4.6 Cross-View Evaluation

4.6.1 Cross-Vertical-View

In this subsection, we study the influence of cross-vertical-view, and report the results in Table 5. We train our model with both the H and L samples. The difference of vertical views is caused by the height of the fixed cameras. As shown in Fig. 2, there are two kinds of height, *i.e.*, 3.5 m and 1.2 m, corresponding to two vertical view angles. Videos from three scenes are used for training and testing.

It is clearly reflected in Table 5 that the vertical view is a sensitive factor to ACC. For example, compared with the non-cross-vertical-view data in scene#1 shown in Tables 2 (92.19%) and 4 (94.71%), ACC with cross-vertical-views decreases about 17% and 19%, respectively.

Generally speaking, cross-vertical-view is also a kind of global change of silhouettes (see Fig. 3). According to our analysis in the last subsection, this kind of variation has great influence on the performance of machine algorithms. Even when the vertical views of cameras only change a little, the silhouette will change a lot in global. However, human visual system is not very sensitive to this kind of global change. This is consistent with our intuition that people seldom fails to recognize a subject no matter the subject stands at a platform with the height of 1.2 m or 3.5 m.

Besides the study in [38], there are no studies on cross-vertical-view gait recognition in the recent literature. The

TABLE 4
Performance (%) of TS-CNN in Cross-Dressing Evaluation

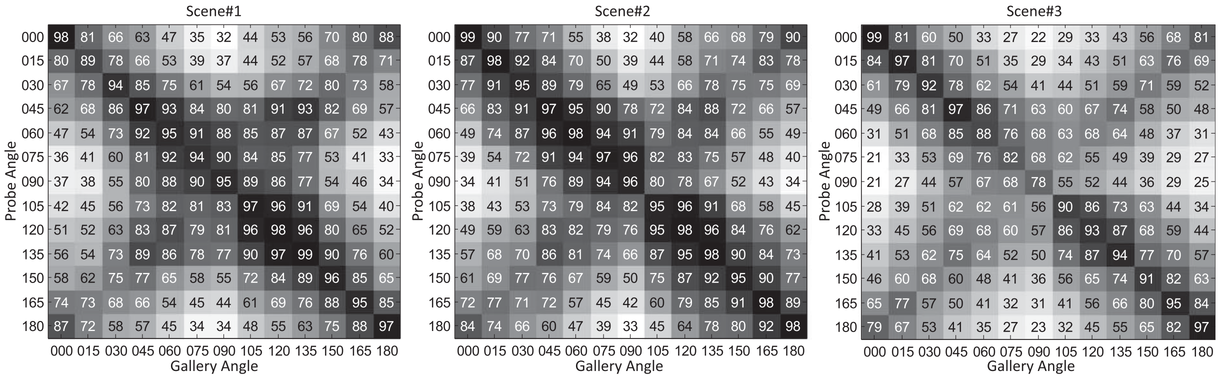
Gallery		NM			BG			CL		
Probe		NM	BG	CL	NM	BG	CL	NM	BG	CL
ACC@top1	Scene#1	94.71	86.84	85.39	85.39	94.94	81.36	81.36	72.15	94.16
	Scene#2	95.75	89.09	89.50	89.50	98.99	83.00	83.00	73.60	98.23
	Scene#3	81.36	64.91	63.22	63.22	82.96	49.12	49.12	45.61	78.14
ACC@top10	Scene#1	98.74	95.19	94.96	94.96	98.23	94.46	94.46	91.65	97.21
	Scene#2	98.25	96.45	95.25	95.25	99.49	95.00	95.00	93.91	99.24
	Scene#3	94.21	84.96	82.37	82.37	93.73	73.80	73.80	73.43	90.70

NM: normal. BG: carrying bags. CL: clothing change.

TABLE 6
Comparison of Cross-Horizontal-View Evaluation for ACC@top1, Excluding the Identical-View Cases

Gallery		0° ∼ 180°													
Probe		0°	15°	30°	45°	60°	75°	90°	105°	120°	135°	150°	165°	180°	Mean
TS-CNN	Scene#1	58.04	59.75	67.64	75.99	72.27	64.72	63.06	70.86	76.93	78.00	73.08	67.20	57.26	68.06
	Scene#2	59.38	68.52	72.45	79.85	74.39	67.51	61.02	67.57	77.78	80.93	73.14	70.62	63.33	70.50
	Scene#3	46.51	56.44	61.12	63.69	57.86	50.51	45.29	53.05	59.05	61.60	60.34	57.08	51.18	55.67
GaitSet [6]	Scene#1	76.38	81.33	86.86	88.02	85.26	82.45	83.17	85.85	86.71	87.28	87.04	86.07	79.01	84.26
	Scene#2	79.76	87.42	88.22	89.05	88.95	86.82	83.58	85.85	88.75	89.19	88.10	87.91	84.25	86.76
	Scene#3	74.78	84.98	86.64	86.86	85.21	80.48	78.53	84.79	86.53	87.03	85.97	84.30	76.15	83.25

Models are trained and evaluated in three scenes, respectively.



Scene#1

	000	015	030	045	060	075	090	105	120	135	150	165	180
000	98	81	66	63	47	35	32	44	53	56	70	80	88
015	80	89	78	66	53	39	37	44	52	57	68	78	71
030	67	78	94	85	75	61	54	56	67	72	80	73	58
045	62	68	86	97	93	84	80	81	91	93	82	69	57
060	47	54	73	92	95	91	88	85	87	87	67	52	43
075	36	41	60	81	92	94	90	84	85	77	53	41	33
090	37	38	55	80	88	90	95	89	86	77	54	46	34
105	42	45	56	73	82	81	83	97	96	91	69	54	40
120	51	52	63	83	87	79	81	96	98	96	80	65	52
135	56	54	73	89	86	78	77	90	97	99	90	76	60
150	58	62	75	77	65	58	55	72	84	89	96	85	65
165	74	73	68	66	54	45	44	61	69	76	88	95	85
180	87	72	58	57	45	34	34	48	55	63	75	88	97

Scene#2

	000	015	030	045	060	075	090	105	120	135	150	165	180
000	99	90	77	71	55	38	32	40	58	66	68	79	90
015	87	98	92	84	70	50	39	44	58	71	74	83	78
030	77	91	95	89	79	65	49	53	66	78	75	75	69
045	66	83	91	97	95	90	78	72	84	88	72	66	57
060	49	74	87	96	98	94	91	79	84	84	66	55	49
075	39	54	72	91	94	97	96	82	83	75	57	48	40
090	34	41	51	76	89	94	96	80	78	67	52	43	34
105	38	43	53	73	79	84	82	95	96	91	68	58	45
120	49	59	63	83	82	79	76	95	98	96	84	76	62
135	57	68	70	86	81	74	66	87	95	98	90	84	73
150	61	69	77	76	67	59	50	75	87	92	95	90	77
165	72	77	71	72	57	45	42	60	79	85	91	98	89
180	84	74	66	60	47	39	33	45	64	78	80	92	98

Scene#3

	000	015	030	045	060	075	090	105	120	135	150	165	180
000	99	81	60	50	33	27	22	29	33	43	56	68	81
015	84	97	81	70	51	35	29	34	43	51	63	76	69
030	61	79	92	78	62	54	41	44	51	59	71	59	52
045	49	66	81	97	86	71	63	60	67	74	58	50	48
060	31	51	68	85	88	76	68	63	68	64	48	37	31
075	21	33	53	69	76	82	68	62	55	49	39	29	27
090	21	27	44	57	67	68	78	55	52	44	36	29	25
105	28	39	51	62	62	61	56	90	86	73	63	44	34
120	33	45	56	69	68	60	57	86	93	87	68	59	44
135	41	53	62	75	64	52	50	74	87	94	77	70	57
150	46	60	68	60	48	41	36	56	65	74	91	82	63
165	65	77	57	50	41	32	31	41	56	66	80	95	84
180	79	67	53	41	35	27	23	32	45	55	65	82	97

Fig. 12. Heat maps of cross-horizontal-view evaluation for ACC@top1. From left to right, they are the heat maps of Scene#1, 2, and 3.

performance shown in Table 5 is bad. This is probably because we only design two kinds of height of fixed cameras in the vertical direction, and thus lack sufficient cross-vertical-view data. In future, more attention should be paid to this problem.

4.6.2 *Cross-Horizontal-View*

In this section, we report the influence of horizontal views on gait recognition. Videos from three scenes are used for training and testing. We follow the rule used in [63]: fix the probe view angle and report ACC while the gallery view angle varies. Besides the TS-CNN method with well processed GEIs, we also evaluate the influence of cross-horizontal view with a recent frame-based state-of-the-art method GaitSet [6], which could provide a novel aspect of learning gait features directly from raw silhouettes. The comparison results in three scenes are shown in Table 6 (ACC@top1). For TS-CNN, we also report the view-to-view heat maps in Fig. 12 (ACC@top1). From the results of Table 6 and Fig. 12, we can draw the following conclusions:

1) For both the GEI-based TS-CNN and the raw silhouette-based GaitSet [6], the average accuracies in Scene#1 and Scene#2 are higher than that in Scene#3. Scene#3 is a very complex outdoor scene, and some silhouettes are not well segmented, thus severely influencing the performance. This indicates that the quality of silhouette segmentation is critical to cross-view gait recognition, even much more critical than other evaluations like cross-scene and cross-dressing.
2) In the comparison, the overall performance of GaitSet is better than TS-CNN, which indicates that GaitSet could extract more gait knowledge from set-wise silhouettes. In different scenes, the performance of GaitSet is rather stable, indicating that learning from raw inputs could avoid the drawbacks of noisy frames and length-varying sequences under different scenes.
3) In the view-to-view heat maps, there is a clear "X" shape in all scenes due to: (1) the matching in the same view performs well; and (2) the accuracy of symmetric matching is also relatively high, *e.g.*, 0-to-180, 15-to-165. Once the cross-view is over 15 degrees, the performance will obviously decrease. Especially, 0-to-90 matching performs worst because it severely changes the silhouette appearances.

4.7 Cross-View and dressing Evaluation

Cross-horizontal-view and cross-dressing are probably two most common problems of gait recognition, which usually happen at the same time. So we hope to analyze the influence of these two problems together.

The experimental results of cross-view&dressing with TS-CNN are reported in Fig. 13. Only videos from scenes#1 are used for training and testing here. The average accuracy of "cross-view&NM-BG," "cross-view&NM-CL" and "cross-view&BG-CL" is 64.81%, 57.87% and 50.57% for ACC@top1, respectively. And the overall average accuracy of these three evaluation cases is nearly 4%-15% lower than the performance with only cross-view evaluation, which

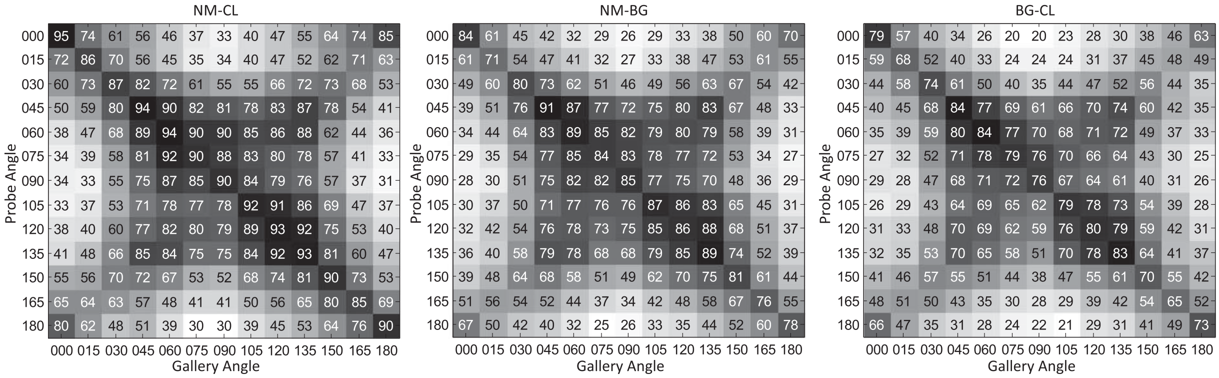


Fig. 13. Heat maps of cross-view&dressing evaluation for ACC@top1 in Scene#1. From left to right, they are cross-view results together with cross-dressing of NM-to-BG, NM-to-CL and CL-to-BG, respectively.

indicates that cross-view&dressing together is a more difficult problem.

As have been evaluated in the last subsection, GaitSet [6] has shown advantages for the cross-view gait recognition. In a recently held competition of Human Identification at a Distance (HID-2020)[1], by IAPR Technical Committee on Biometrics (TC4), a series of methods based on GaitSet have been adopted and achieved advanced performance, as shown in Table 7. The dataset used in this competition is a subset of CASIA-E, in which only 10 sequences for each subject are randomly selected, including varying views, clothing and scenes. Despite the great challenges introduced by varying views, lighting and clothing conditions in real scenes, the Rank-1 method has reached a 60.3% accuracy, which indicates the great potential of GaitSet-based methods and shows that there is still much room of improvement for practical gait recognition. Therefore, we hope our full CASIA-E dataset could further advance the research of large-scale gait recognition in real challenging outdoor scenes.

4.8 Thermal Infrared Subset Evaluation

To explore the influence of different modalities on gait recognition, we evaluate the thermal infrared subset. It contains 270 subjects with two walking directions, i.e., walking from left or right. All thermal infrared videos are captured with thermal infrared camera with a resolution of 480P@25FPS. We set the first 135 subjects as the training set, the left 135 subjects as the testing set. For testing, one walking sequence is in the gallery and the other is as probe. To apply the gait recognition method on this subset, we first detect and crop the person image with simple HOG detector from the raw images, then implement the average pooling on the whole sequences to calculate the final representation, as shown in Fig. 14. As such a representation is very similar to GEI, we name it as Gait Thermal Image (GTI).

With the well processed GTIs, we evaluate this subset with the adopted TS-CNN and report the results in Table 8. It is not surprising that we have not achieved high recognition accuracy in such a small subset, due to the challenging thermal infrared scene. We also evaluate the frame-based GaitSet [6] method on this subset, and find that it could achieve surprising improvement from raw thermal infrared frames. This observation further shows the effectiveness of learning directly from the raw gait images with deep learning methods, especially for the noisy data. Unlike the common RGB image, the thermal infrared image introduces more challenges, such as the irregular temperature-sensitive distribution, the low image quality due to the limitation of the capturing device, as well as the small dataset scale which is hard to support training models that need large amounts of data. Noticed that TS-CNN takes the intermediate product GTIs as inputs, while GaitSet randomly selects group-wise thermal infrared frames to build its inputs. In such a condition, the sampling strategy of GaitSet is more flexible for mining discriminative gait features and filtering

TABLE 7
Competition Results of IAPR Technical Committee on Biometrics (TC4-HID2020)

Rank	Method	Accuracy
Baseline	Original GaitSet [6]	49.9
First	+Global and Local Features	63.0
Second	+Temporal Proposal Module	54.1
Third	+Multi-grid Spatial and Temporal Feature Fusion	53.4
Fourth	+Micro-motion Capture Module	51.5

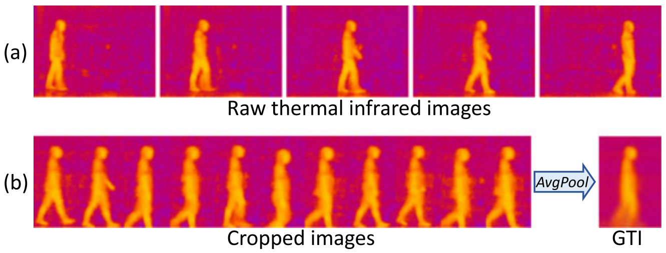


Fig. 14. Thermal infrared examples. (a) is the raw images and (b) is the pipeline of producing gait templates, i.e., the Gait Thermal Image (GTI).

TABLE 8
Performance (%) of TS-CNN and GaitSet [6] in the Thermal Infrared Subset Evaluation

Gallery → Probe	L → R				R → L			
ACC@	Top1	Top3	Top5	Top10	Top1	Top3	Top5	Top10
TS-CNN	62.86	74.29	79.52	83.93	61.43	71.43	77.62	82.86
GaitSet [6]	85.71	94.29	95.71	100.00	82.86	95.71	95.71	97.14

[1] IAPR TC4-HID2020 Competition, http://hid2020.iapr-tc4.org

TABLE 9
Summary of Experimental Benchmark

Name	Experiment	Remarks
CASIA-E-1	Baseline in Section 4.2	#Training samples, 3 kinds of features
CASIA-E-2	Cross-scene in Section 4.3	Scene #1, #2, #3
CASIA-E-3	Cross-walking-style in Section 4.4	Non-stop, stop
CASIA-E-4	Cross-dressing in Section 4.5	NM (normal) BG (carrying bags) CL (clothing change)
CASIA-E-5	Cross-vertical-view in Section 4.6.1	Two vertical views
CASIA-E-6	Cross-horizontal-view in Section 4.6.2	13 horizontal views
CASIA-E-7	Cross-view&dressing in Section 4.7	13 horizontal views × (NM, BG, CL)
CASIA-E-8	Thermal Infrared in Section 4.8	A subset of new modality

the unnecessary clutters directly from the raw data. As an important subset of CASIA-E, the thermal infrared scene is helpful when the common camera is unavailable such as in the night, or in some confidential conditions.

4.9 Summary of Experimental Benchmark

As the final part of this section, we summarize all experiments in Table 9. It would be convenient for readers to discuss and quickly locate each part of the experimental benchmark, also convenient for comparison.

5 Conclusion and Future Work

In this paper, we have introduced CASIA-E, a newly built gait recognition dataset, with huge amount of videos, sufficient subjects, rich attributes, spatial and temporal variations, and multiple soft biometric features. We have also provided an experimental benchmark on CASIA-E. The experiments involve the main features of CASIA-E, including the number of training samples, scenes, horizontal views, vertical views, walking styles and dressing. This should be the first experimental study with so many training videos (half a million), about 50 times more than the previous one. In addition, some situations have seldom been studied before, *e.g*, vertical views and walking styles. We believe that such a big dataset as well as its experimental benchmark will promote the application of gait recognition in real-world scenarios.

The dataset and the source code corresponding to the experiments will be released. Some future studies are pointed out as follows.

- Some readers may note that the soft biometric features are also collected but we did not conduct the related experiments. This is because there is not enough space to study all these soft biometric features collected in CASIA-E, including age, gender, height and weight. However, it does not mean that these soft biometric features are unimportant. Partly, one previous research [72] of us has somewhat explored the potentials that the soft biometric features could supply by CASIA-E. Therefore, they are very critical for many practical applications, *e.g.*, gait-based gender judgment and gait-based age prediction, worthy a new paper with a comprehensive study in future.
- Second, more experimental factors could be considered, *e.g.*, domain adaptation of gait recognition, false alarm rate, over-fitting and generalization of models. These factors have not attracted enough attention in previous researches of gait recognition, and each of them is worthy being investigated. It is expected this dataset could provide a new platform for researchers to evaluate their gait representation learning methods, especially on challenging outdoor scenes.
- Finally, we believe that there is still much space to extend the CASIA-E dataset. For example, we did not pay much attention to designing enough vertical view change when capturing the gait videos. Once we started the project of CASIA-E, it is difficult to change the schedule because everything in this complex project had been arranged. Such samples will be helpful to enhance the performance of cross-vertical-view recognition.

Acknowledgments

The first two authors contributed equally to this work. The authors would like to thank Prof. T. Tan for his valuable support and contribution to this work.

References

[1] W. An *et al.*, “Performance evaluation of model-based gait on multi-view very large population database with pose sequences,” *IEEE Trans. Biom., Behav., Ident. Sci.*, vol. 2, no. 4, pp. 421–430, Oct. 2020.

[2] K. Bashir, T. Xiang, and S. Gong, “Gait recognition using gait entropy image,” in *Proc. 3rd Int. Conf. Imag. Crime Detection Prevention*, 2009, pp. 1–6.

[3] A. F. Bobick and A. Y. Johnson, “Gait recognition using static, activity-specific parameters,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. I–I.

[4] I. Bouchrika and M. S. Nixon, “Model-based feature extraction for gait analysis and recognition,” in *Proc. Int. Conf. Comput. Vis./Comput. Graph. Collaboration Techn. Appl.*, 2007, pp. 150–160.

[5] N. V. Boulgouris and Z. X. Chi, “Gait recognition using radon transform and linear discriminant analysis,” *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 731–740, Mar. 2007.

[6] H. Chao, Y. He, J. Zhang, and J. Feng, “Gaitset: Regarding gait as a set for cross-view gait recognition,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8126–8133.

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.

[8] R. Chereshnev and A. Kertész-Farkas, “HuGaDB: Human gait database for activity recognition from wearable inertial sensor networks,” in *Proc. Int. Conf. Anal. Images, Soc. Netw. Texts*, 2017, pp. 131–141.

[9] P. Connor and A. Ross, “Biometric recognition by gait: A. survey of modalities and features,” *Comput. Vis. Image Understanding*, vol. 167, pp. 1–27, 2018.

[10] N. Cuntoor, A. Kale, and R. Chellappa, "Combining multiple evidences for gait recognition," in *Proc. Int. Conf. Multimedia Expo.*, 2003, pp. III–113.

[11] B. DeCann, A. Ross, and J. Dawson, "Investigating gait recognition in the short-wave infrared (SWIR) spectrum: Dataset and challenges," in *Proc. Biometric Surveill. Technol. Hum. Activity Identification X. Int. Soc. Opt. Photon.*, 2013, Art. no. 87120J.

[12] C. Fan *et al.*, "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14 213–14 221.

[13] D. Gafurov, "A survey of biometric gait recognition: Approaches, security and challenges," in *Proc. Annu. Norwegian Comput. Sci. Conf.*, 2007, pp. 19–21.

[14] W. E. L. Grimson, "Gait analysis for recognition and classification," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2002, pp. 155–162.

[15] R. Gross and J. Shi, "The CMU motion of body (mobo)database," Robotics Institute, Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-RI-TR-01–18, 2001.

[16] Y. Guan, C.-T. Li, and F. Roli, "On reducing the effect of covariate factors in gait recognition: A classifier ensemble method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1521–1528, Jul. 2015.

[17] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2980–2988.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[20] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task GANs for view-specific feature learning in gait recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 1, pp. 102–113, Jan. 2019.

[21] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll, "The TUM gait from audio, image and depth (GAID) database: Multimodal recognition of subjects and traits," *J. Vis. Commun. Image Representation*, vol. 25, no. 1, pp. 195–206, 2014.

[22] S. Huang, A. Elgammal, J. Lu, and D. Yang, "Cross-speed gait recognition using speed-invariant gait templates and globality-locality preserving projections," *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 10, pp. 2071–2083, Oct. 2015.

[23] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 5, pp. 1511–1521, Oct. 2012.

[24] N. Jia, V. Sanchez, C.-T. Li, and H. Mansour, "On reducing the effect of silhouette quality on individual gait recognition: A feature fusion approach," in *Proc. Int. Conf. Biometrics Special Int. Group*, 2015, pp. 1–5.

[25] A. Y. Johnson and A. F. Bobick, "A multi-view method for gait recognition using static body parameters," in *Proc. Int. Conf. Audio Video-Based Biometric Person Authentication*, 2001, pp. 301–311.

[26] A. Kale, N. Cuntoor, and R. Chellappa, "A framework for activity-specific human identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. IV-3660–IV-3663.

[27] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4873–4882.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[29] T. H. Lam, K. H. Cheung, and J. N. Liu, "Gait flow image: A silhouette-based gait representation for human identification," *Pattern Recognit.*, vol. 44, no. 4, pp. 973–987, 2011.

[30] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait recognition via semi-supervised disentangled representation learning to identity and covariate features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13 306–13 316.

[31] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 871–885, Apr. 2018.

[32] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5168–5177.

[33] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Eur. Conf. Comput. Vis.*, 2014.

[34] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2015, pp. 740–755.

[35] Y. Makihara *et al.*, "Gait collector: An automatic gait data collection system in conjunction with an experience-based long-run exhibition," in *Proc. Int. Conf. Biometrics*, 2016, pp. 1–8.

[36] Y. Makihara *et al.*, "The OU-ISIR gait database comprising the treadmill dataset," *IPSJ Trans. Comput. Vis. Appl.*, vol. 4, pp. 53–62, 2012.

[37] D. Muramatsu, Y. Makihara, H. Iwama, T. Tanoue, and Y. Yagi, "Gait verification system for criminal investigation," *IPSJ Trans. Comput. Vis. Appl.*, vol. 5, pp. 747–748, 2013.

[38] D. Muramatsu, A. Shiraishi, Y. Makihara, M. Z. Uddin, and Y. Yagi, "Gait-based person recognition using arbitrary view transformation model," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 140–154, Jan. 2015.

[39] M. S. Nixon and J. N. Carter, "Automatic recognition by gait," *Proc. IEEE Proc. IRE*, vol. 94, no. 11, pp. 2013–2024, Nov. 2006.

[40] M. S. Nixon, B. H. Guo, S. V. Stevenage, E. S. Jaha, N. Almudhahka, and D. Martinho-Corbishley, "Towards automated eyewitness descriptions: Describing the face, body and clothing for recognition," *Vis. Cogn.*, vol. 25, pp. 524–538, 2016.

[41] C. Prakash, R. Kumar, and N. Mittal, "Recent developments in human gait research: Parameters, approaches, applications, machine learning techniques, datasets and challenges," *Artif. Intell. Rev.*, vol. 49, no. 1, pp. 1–40, 2018.

[42] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[43] S. Samangooei, J. D. Bustard, M. Nixon, and J. Carter, "On acquisition and analysis of a dataset comprising of gait, ear and semantic data," *Multibiometrics Hum. Identification*," pp. 277–301, 2011.

[44] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.

[45] G. Shakhnarovich, L. Lee, and T. Darrell, "Integrated face and gait recognition from multiple views," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. I–I.

[46] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," in *Proc. Int. Conf. Biometrics*, 2016, pp. 1–8.

[47] J. D. Shutler, M. G. Grant, M. S. Nixon, and J. N. Carter, "On a large sequence-based human gait database," *Appl. Sci. Soft Comput.*, vol. 24, pp. 339–346, 2004.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[49] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[50] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 4, pp. 1–14, 2018.

[51] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "On input/output architectures for convolutional neural network-based cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2708–2719, Sep. 2019.

[52] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.

[53] I. R. Vega and S. Sarkar, "Statistical motion model based on the change of feature relationships: Human gait-based recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1323–1328, Oct. 2003.

[54] D. K. Wagg and M. S. Nixon, "On automated model-based extraction and analysis of gait," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2004, pp. 11–16.

[55] C. Wang, J. Zhang, L. Wang, J. Pu, and X. Yuan, "Human identification using temporal information preserving gait template," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2164–2176, Nov. 2012.

[56] J. Wang, M. She, S. Nahavandi, and A. Kouzani, "A review of vision-based gait recognition methods for human identification," in *Int. Conf. Digit. Image Comput., Techn. Appl.*, 2010, pp. 320—327.

[57] L. Wang, H. Ning, T. Tan, and W. Hu, "Fusion of static and dynamic body biometrics for gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 2, pp. 149–158, Feb. 2004.

[58] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1505–1518, Dec. 2003.

[59] L. Wang, G. Zhao, N. Rajpoot, and M. S. Nixon, "Special issue on new advances in video-based gait analysis and applications: Challenges and solutions," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 40, no. 4, pp. 982–985, Aug. 2010.

[60] Y. Wang, C. Song, Y. Huang, Z. Wang, and L. Wang, "Learning view invariant gait features with two-stream GAN," *Neurocomputing*, vol. 339, pp. 245–254, 2019.

[61] Z. Wu, Y. Huang, and L. Wang, "Learning representative deep features for image set analysis," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1960–1968, Nov. 2015.

[62] Z. Wu, Y. Huang, L. Wang, and T. Tan, "Early hierarchical contexts learned by convolutional networks for image segmentation," in *Proc. 22nd Int. Conf. Pattern Recognit.*, 2014, pp. 1538–1543.

[63] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, Feb. 2017.

[64] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3485–3492.

[65] C. Xu, Y. Makihara, G. Ogi, X. Li, Y. Yagi, and J. Lu, "The OU-ISIR gait database comprising the large population dataset with age and performance evaluation of age estimation," *IPSJ Trans. Comput. Vis. Appl.*, vol. 9, no. 24, pp. 1–14, 2017.

[66] C. Yam, M. S. Nixon, and J. N. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognit.*, vol. 37, no. 5, pp. 1057–1072, 2004.

[67] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*.

[68] J. H. Yoo, D. Hwang, K. Y. Moon, and M. S. Nixon, "Automated human recognition by gait using neural network," in *Proc. 1st Workshops Image Process. Theory, Tools Appl.*, 2008, pp. 1–6.

[69] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh, "GaitGAN: Invariant gait feature extraction using generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 532–539.

[70] S. Yu, H. Chen, Q. Wang, L. Shen, and Y. Huang, "Invariant feature extraction for gait recognition using only one uniform model," *Neurocomputing*, vol. 239, pp. 81–93, 2017.

[71] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. 18th Int. Conf. Pattern Recognit.*, 2006, pp. 441–444.

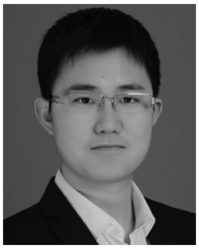
[72] Y. Zhang, Y. Huang, L. Wang, and S. Yu, "A comprehensive study on gait biometrics using a joint CNN-based method," *Pattern Recognit.*, vol. 93, pp. 228–236, 2019.

[73] Y. Zhang, G. Pan, K. Jia, M. Lu, Y. Wang, and Z. Wu, "Accelerometer-based gait recognition by sparse representation of signature points with clusters," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1864–1875, Sep. 2014.

[74] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1556–1564.

[75] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, "Robust view transformation model for gait recognition," 2011, pp. 2073–2076.

Chunfeng Song (Member, IEEE) received the BEng degree from the QiLu University of Technology, in 2012, the MEng degree from North China Electric Power University, in 2016, and the PhD degree from the University of Chinese Academy of Sciences (UCAS), in 2020. Since 2020, he has joined the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA) as an assistant professor. He has published more than 20 conference and journal papers such as *IEEE Transactions on Image Processing*, CVPR, ECCV, AAAI, BMVC, and PR. His current research interests include person identification, image segmentation, and unsupervised learning.



Yongzhen Huang (Senior Member, IEEE) received the BE degree from the Huazhong University of Science and Technology, in 2006 and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2011. He is currently an associate professor with the School of Artificial Intelligence, Beijing Normal University. His research interests include pattern recognition, computer vision and machine learning, and has published one book and more than 80 papers at international journals and conferences such as *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *International Journal of Computer Vision*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Systems, Man, and Cybernetics*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Circuits and Systems for Video Technology*, CVPR, ICCV, ECCV, NIPS, and AAAI.

Weining Wang received the BE degree from North China Electric Power University, China, in 2015 and the PhD degree from Institute of Automation, Chinese Academy of Sciences (CASIA), in 2020. Since 2020, she has joined the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA) as an assistant professor. Her research interests include machine learning, pattern recognition, and computer vision.

Liang Wang (Fellow, IEEE) received the BEng and MEng degrees from Anhui University, in 1997 and 2000, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2004. From 2004 to 2010, he was a research assistant with Imperial College London, United Kingdom, and Monash University, Australia, a research fellow with the University of Melbourne, Australia, and a lecturer with the University of Bath, United Kingdom, respectively. Currently, he is a full professor of the Hundred Talents Program with the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published in highly ranked international journals, such as *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *IEEE Transactions on Image Processing*, and leading international conferences, such as CVPR, ICCV, and ICDM.