A Comprehensive Survey on Deep Gait Recognition: Algorithms, Datasets, and Challenges

Chuanfu Shen, Shiqi Yu, *Member, IEEE*, Jilong Wang, George Q. Huang, *Fellow, IEEE*, and Liang Wang, *Fellow, IEEE*

***Abstract*—Gait recognition aims to identify a person at a distance, serving as a promising solution for long-distance and less-cooperation pedestrian recognition. Recently, significant advances in gait recognition have achieved inspiring success in many challenging scenarios by utilizing deep learning techniques. Against the backdrop that deep gait recognition has achieved almost perfect performance in laboratory datasets, much recent research has introduced new challenges for gait recognition, including robust deep representation modeling, in-the-wild gait recognition, and even recognition from new visual sensors such as infrared and depth cameras. Meanwhile, the increasing performance of gait recognition might also reveal concerns about biometrics security and privacy prevention for society. We provide a comprehensive survey on recent literature using deep learning and a discussion on the privacy and security of gait biometrics. This survey reviews the existing deep gait recognition methods through a novel view based on our proposed taxonomy. The proposed taxonomy differs from the conventional taxonomy of categorizing available gait recognition methods into the model- or appearance-based methods, while our taxonomic hierarchy considers deep gait recognition from two perspectives: deep representation learning and deep network architectures, illustrating the current approaches from both micro and macro levels. We also include up-to-date reviews of datasets and performance evaluations on diverse scenarios. Finally, we introduce privacy and security concerns on gait biometrics and discuss outstanding challenges and potential directions for future research.**

***Index Terms*—Gait recognition, deep learning, representation learning, biometrics security and privacy.**

Received 16 April 2024; revised 28 June 2024 and 5 September 2024; accepted 11 October 2024. Date of publication 25 October 2024; date of current version 27 March 2025. This work was supported in part by the Shenzhen International Research Cooperation Project under Grant GJHZ20220913142611021, and in part by the National Natural Science Foundation of China under Grant 62476120 and Grant 61976144. This article was recommended for publication by Associate Editor D. Muramatsu upon evaluation of the reviewers' comments. *(Corresponding author: Shiqi Yu.)*

Chuanfu Shen is with the Department of Data and Systems Engineering, The University of Hong Kong, Hong Kong, SAR, China, and also with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: noahshen@connect.hku.hk).

Shiqi Yu is with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: yusq@sustech.edu.cn).

Jilong Wang is with the Department of Automation, University of Science and Technology of China, Hefei 230026, China, and also with the New Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jilongw@mail.ustc.edu.cn).

George Q. Huang is with the Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, SAR, China (e-mail: gq.huang@polyu.edu.hk).

Liang Wang is with the New Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wangliang@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TBIOM.2024.3486345

I. Introduction

GAIT recognition is a biometrics application that aims to identify pedestrians by their walking patterns [1], [2]. It can be viewed as a vision-based person retrieval problem with the objective of human identification from gait sequences captured by visual cameras. One significant advantage of gait recognition is its ability to perform human identification at a distance, making it suitable for low-resolution and long-distance scenarios [3]. In the context of the COVID-19 pandemic [4], where traditional surveillance systems relying on face, iris, or fingerprint may be deficient. Therefore gait recognition emerges as a preferred solution. Additionally, gait recognition offers the benefit of requiring less active cooperation from individuals. These unique characteristics make gait recognition significantly potential for diverse applications, including surveillance, forensics, and healthcare.

Although gait recognition research shortly started three decades ago [2], [5], the field of study has continuously advanced and expanded over the years. To our best knowledge, we categorize the evolution of gait recognition research into three stages. *The initial stage*, in the early 1990s [5], focused on exploring the feasibility of human identification at a distance. The early approaches showed reasonable performance, but they were evaluated on small-scale benchmarks with a limited number of subjects, *e.g.*, ten at most in [3], [6].

The second stage emerged from DARPA Human Identification at a Distance (HumanID) program [1], [7], [8], which not only promoted techniques but also introduced valuable datasets. At this time, the methods gradually formed two categories: appearance-based and model-based. The appearance-based methods [9], [10], [11] directly exploit shape information from gait representations like silhouettes, unlike model-based methods [12], [13], [14] that explicitly model a deformable human body to represent individuals. Additionally, datasets began to include over a hundred subjects [15], [16], [17], and they started to explore factors like view variants [1], [18] and appearance-changing scenarios [3], [15]. With the promising evaluation performance in this stage, gait recognition demonstrated the feasibility and potential for further exploration.

Then the advent of the deep learning era in gait recognition powers *the third stage*, which is distinguished from the previous period in three main aspects: *(i)* Deep learning techniques have revolutionized gait recognition by allowing the learning of abstract and discriminative gait features directly

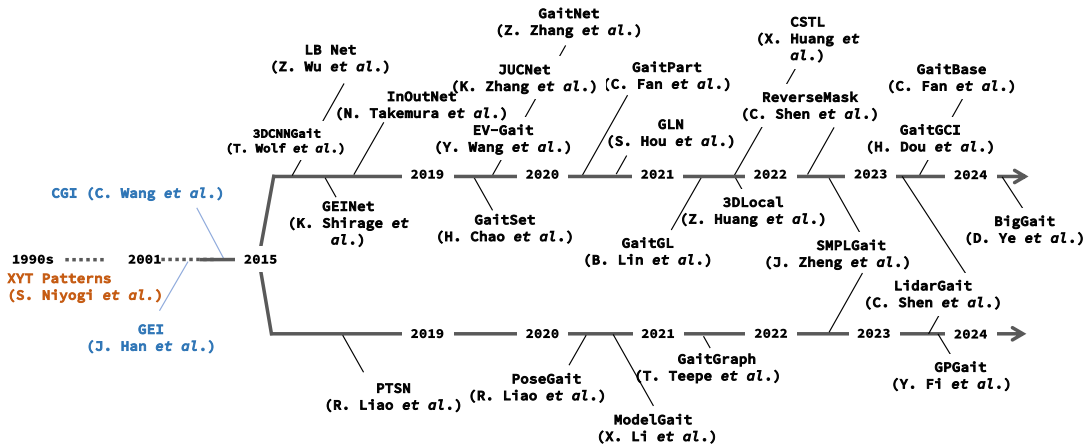


Fig. 1. Milestone of gait recognition approaches. The reference in orange represents methods at the early stage, while references in blue indicate handcrafted feature-based methods. The references in black are representative works in the deep learning era. We categorise appearance-based methods 3DCNNGait [19], GEINet [20], LBNet [21], InOutNet [22], GaitSet [23], [24], EV-Gait [25], JUCNet [26], GaitNet [27], GaitPart [28], GLN [29], GaitGL [30], 3DLocal [31], CSTL [32], ReverseMask [33], GaitGCI [34], GaitBase [35], BigGait [36] on the top branch. The bottom branch presents some representative model-based methods, such as PoseGait [37], PTSN [38], ModelGait [39], GaitGraph [40], and GPGait [41]. LidarGait [42] and SMPLGait [43] contain both model- and appearance-based features, showing conventional taxonomy is insufficient for the rapidly developing deep gait recognition.

from input data, eliminating the need for expert knowledge and manual feature engineering. Furthermore, these deep feature-based methods [20], [21], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55] consistently outperform traditional approaches that rely on hand-crafted features, such as gait energy images [9] and gait history images [56]. *(ii)* Deep gait recognition continuously refreshes the scale record of pedestrians. The state-of-the-art methods can make satisfying results in datasets with hundreds of subjects. For example, the overall recognition accuracy is beyond 93% [33] under a very difficult appearance-changing setting on CASIA-B [15]. When evaluating scenario with thousand of subjects, the cutting-edge methods can gain 98.3% [57] rank-1 accuracy on OUMVLP [58] dataset. For evaluating outdoor scenarios with over 5 thousand subjects, deep models [34], [59] can surprisingly achieve over 70% rank-1 accuracy on GREW [60]. *(iii)* With the advancement of gait recognition research in addressing challenges related to viewpoints and clothing, deep learning has opened up new possibilities for tackling even more complex problems such as in-the-wild gait recognition [42], [43], multi-modal recognition [61], end-to-end recognition [62], and unsupervised recognition [63].

Due to the rapid increase in the number of deep gait recognition methods [64], [65], [66] and the great progress made in recent years, as shown in Fig. 1, we are motivated to develop this survey of deep learning for gait recognition. Though there are some outstanding surveys on deep gait recognition, this paper is distinguished from the existing surveys in four aspects. First of all, our survey is up-to-date and comprehensively covers state-of-the-art modalities, algorithms, datasets, and challenges. Meanwhile, recent surveys [64], [67] have comprehensively reviewed advances using deep learning for gait recognition. Despite the fact that they provide insights into the technical aspects of deep gait recognition methods, one downside of the existing survey papers is that they do not cover remarkable achievements made in 2022 and 2023, notably by the appearance of datasets and challenges, such as gait recognition in the wild [60], no label [63], cloth-changing settings [68], and 3D space [42]. Secondly, the existing surveys [64], [67], [69], [70] follow conventional taxonomy categorizing deep gait recognition methods into the aspect of either the neural networks [64], [67] (*e.g.*, CNN, RNN, GCN) or the used representations [69], [70] (*i.e.*, model-based, appearance-based). However, the conventional taxonomy struggles to effectively classify new deep gait recognition methods, for example, PoseMapGait [71] and LidarGait [42] utilizing both appearance- and model-based characteristics. To this end, we propose a novel taxonomy with two dimensions, *i.e.*, deep representation and neural architectures, to provide other scholars with a systematic understanding of deep learning techniques for gait recognition. Lastly, we provide the performance evaluations of gait recognition in four scenarios in detail, *i.e.*, gait recognition in the cross-view setting, in the wild, in the cloth-changing setting, and in the 3D space. The comprehensive comparisons help to keep pace with the developments of diverse methods. Additionally, this paper discusses the neglected but significant topic of biometric security and privacy, addressing potential threats and challenges for exploration in the future.

To provide an exclusive taxonomy for deep gait recognition helping to obtain a deeper understanding, this paper proposes a novel taxonomy on deep gait recognition from two perspectives: deep representations learning techniques and deep architectures designs. We explain the proposed taxonomy in detail: (1) for deep representations learning techniques in gait recognition, the deep representations learning methods can be diversely analyzed from six perspectives rather than conventional model-based/free perspectives. Specifically, we comprehensively discuss deep representations learning in gait recognition from comparisons of 2D/3D, template/frames, shuffled/ordered, global/local, single/multi-scale, and long/short-term. (2) Considering deep architectures designs, the deep architectures are divided into *discriminative* and *generative* models. In our humble opinion, this taxonomy provides a reasonable classification for the existing gait

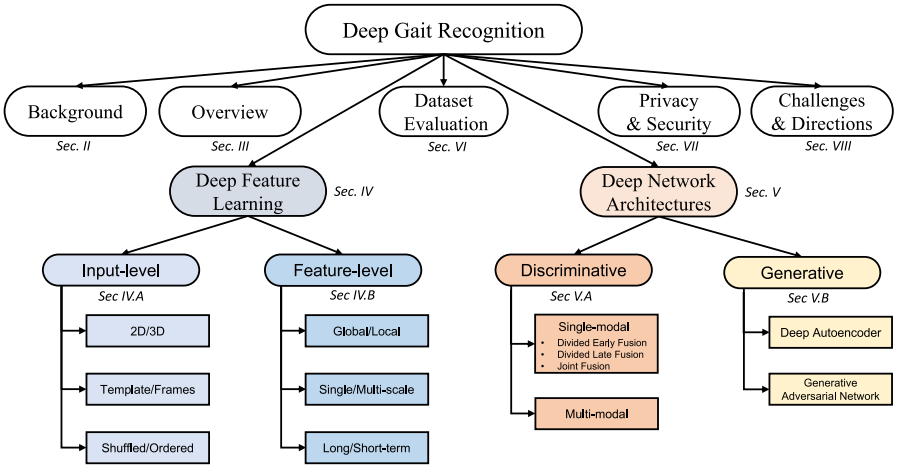


Fig. 2. Overall structure of this survey paper and our proposed taxonomy of the existing deep learning-based methods for gait recognition. Our proposed taxonomy, as shown in the figure, considers deep gait recognition from the perspectives of feature learning and network architecture. It provides both micro and macro scopes to enhance the understanding of deep learning-based gait recognition. Specifically, Section IV focuses on the micro-scope of methods, exploring existing methods based on deep gait feature learning. Section V is presented from a more macro perspective, examining deep models architecture in gait recognition. Other fields of study also adopt a similar taxonomy by categorizing deep models into either discriminative or generative approaches.

recognition papers, and gives readers a big picture of deep gait recognition.

The rest of the survey is organized as illustrated in Fig. 2. We first introduce the background of gait recognition in Section II and overview the main components of deep learning methods in Section III. Section IV reviews existing deep gait recognition methods according to the feature learning perspective. Section V reviews current deep gait recognition from the deep network architectures perspective. Datasets and evaluations are presented in Section VI. The security and privacy concerns are introduced in Section VII, followed by challenges and suggested directions in Section VIII. The last section concludes the paper.

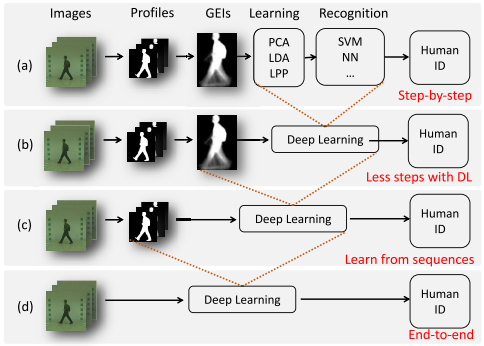


Fig. 3. Four typical workflows on deep gait recognition. Figure adapted by [78].

II. Background of Gait Recognition

Preliminaries: Gait recognition is generally established into two modes: identification and verification. Verification is a one-to-one comparison used to confirm the identity, whereas identification is a one-to-many comparison used to retrieve identity from an ID gallery [72]. Unless otherwise stated, gait recognition in this survey refers to the vision-based pedestrian identification problem from sequential gait streams, which are obtained from visual sensors such as surveillance cameras, depth sensors, and Lidar sensors.

Generally, the recognition tasks are typically categorized into two trends: *open-set* and *closed-set* recognition, which is called *subject-independent* and *subject-dependent* settings in [64]. In the closed-set setting, both the training and testing phases include different samples from the same identities. However, in the open-set setting, the testing phase is tasked with recognizing a set of unseen subjects that have not been encountered yet in the training phase. The closed-set recognition has been widely studied past two decades [3], while the more realistic and challenging open-set recognition has gained increasing interest in recent years [67].

As shown in Fig. 3 (a), a gait recognition system mainly consists of five steps:

1) *Gait Data Acquisition*: Capturing pedestrian walking sequences from various cameras, including RGB, infrared, and depth cameras, *etc*. Most gait datasets are collected in either laboratory environments under controlled conditions [15], [42], [58], or in-the-wild environments with complex background [43], [60], [68].
2) *Data Annotation*: Pedestrian detection and tracking algorithms are used to extract the of-the-interest sequences. Then these sequences are labeled with detailed attributes such as identity, view, clothing, carrying condition, *etc*.
3) *Gait Representation Generation*: To transform the raw data into gait representations such as gait silhouettes, skeletons, and optical flows. It is an essential step to maintain gait representation and to eliminate irrelevant features like the background and texture of clothes [62].
4) *Feature Learning*: Conducting a learning model to effectively represent the underlying structure of the raw gait data by a compact and abstract feature, which is a crucial step for developing a gait recognition system. While handcrafted features have been extensively studied for over two decades [56], deep learning-based feature extraction has become a widely studied paradigm in recent literature [64].

TABLE I
A COMPREHENSIVE ANALYSIS OF DIFFERENT INPUT MODALITIES

Data Type	Capture Devices	Resolution	Cost	Computation	Operational Range	Limitations
RGB Image	Camera	+++	+	None	+++	Sensitive to viewpoint and illumination
Silhouette	Camera	+++	+	++	+++	Require accurate segmentation
Gait Template	Camera	+++	+	++	+++	Lost micro-motion
Optical Flow	Camera	+++	+	+	++	Require subject moving
Skeleton	Diverse	+	+	++	++	Require accurate estimation
3D Human Mesh	Diverse	+	+	+++	+	Require accurate estimation
Depth Image	Depth camera	+++	++	None	+	Unsuitable for outdoor conditions
Infrared Image	Infrared camera	++	+	None	+++	Less visual details
Event Stream	Event camera	+++	++	None	+++	Cannot perceive static objects
Point Cloud	LiDAR sensor	++	+++	None	+++	High cost

The cost refers to the monetary cost required for the acquisition devices, while the computation denotes the computational cost of transforming the raw data into the targeted representation.

5) *Pedestrian Retrieval*: To retrieve the identity of a given query sequence, the trained model extracts feature representations from the query and gallery sequences and calculates their similarity. The similarity scores are used to rank the gallery sequences, resulting in a ranking list to evaluate the performance of the learning model [73].

A. Various Input Data

The environmental conditions in real-world scenarios are complex and changing, necessitating the use of various data modalities to achieve robust and reliable performance. To facilitate a better understanding of the different types of data, we have summarized their advantages and limitations in Table I. Next, we will introduce each modality specifically.

• *RGB Image:* RGB images contain a sufficient amount of information to identify different subjects. However, the texture and color biases on clothing appearance [74] are introduced when learning algorithms apply directly to RGB images. If RGB images are rarely taken as the input directly, the algorithm will recognize different subjects wrongly according to their appearance [68]. In gait recognition, clothing is regarded as a kind of variation, and gait recognition should be robust to different clothing styles. Many recent methods [27], [39], [62] take RGB images as input, extracting representations robust to clothing and texture and achieving outstanding results. We believe that RGB images have great potential for gait recognition since rich information has not been fully utilized.

• *Silhouette:* A silhouette is a binary mask of a human body by removing the background and maintaining the human foreground. Human silhouettes were primarily obtained by background subtraction at an early age [15], [17], [75], [76]. At the same time, the advanced segmentation methods [62], [77], [78] based on deep learning can provide much better quality human silhouettes than background subtraction. Human body silhouettes still contain an informative appearance even though color and texture are removed, but internal body structure information is partially lost in silhouettes. Besides, silhouettes are easy to be affected by clothing and camera views [79]. Because of efficiency and simplicity, silhouettes were the most popular gait data in the past 20 years [23], [28], [78], [80], [81].

• *Gait Template:* Although silhouettes are efficient and simple, a sequence of silhouettes is high-dimensional. Before deep learning was widely deployed for visual recognition, it was not easy to extract features from sequential silhouettes using traditional methods like SVM [82], [83] and Boosting [84]. Han and Bhanu proposed Gait Energy Image (GEI) [9], in which average cyclic silhouettes sequence into a single gait template, and such denoising processing aims to be robust for incomplete silhouettes. GEI contains comparable information to sequential silhouettes, and its data dimensionality is much less. Despite its simplicity, the GEI template is robust to many variations and achieved great success in gait recognition [56]. Besides GEI, some other similar features have also been proposed. They are Motion Energy Image [11], Gait History Image [11], Gait Entropy Image [85], Chrono-gait image [10], Gait Moment Image [86], *etc*. Even though deep models can handle the high data dimensionality of sequential silhouettes, some recent methods [20], [87] still prefer GEI because of its low computational cost and robustness on noise.

• *Optical Flow Image:* Gait recognition is a task to identify a subject via its walking patterns. Therefore, the motion information is significant but hardly described from silhouettes. Optical flow images can be utilized to provide more motion information than silhouettes. Castro et al. also demonstrates that optical flow images can achieve state-of-the-art performance [88]. However, the computational cost for optical flow images is relatively high, and it is also very challenging to obtain optical flow images of high quality. Recent deep learning-based FlowNet [89] and its successors [90], [91], can achieve relatively better optical flow images and might improve gait recognition accordingly.

• *Body Skeleton:* Many methods employ body structures to extract gait motion [12], [13], [14]. The gait recognition methods based on skeleton should be more robust to view and clothing variations than those based on silhouettes. However, it is not easy to extract high-accuracy human body models at the moment. Human pose estimation has achieved encouraging precision via deep learning in recent years. Those human pose estimation methods include but are not limited to DeepPose [92], OpenPose [93] and HR-Net [77]. Then gait recognition with human body models has returned back research of interest [38], [40], [41], [94], and many datasets [42], [60], [95], [96] with pose annotations presented to advance model-based gait recognition.

• *Human Mesh:* Mesh is a type of 3D representation that consists of a collection of vertices and polygons to define the exact shape of an object [97], [98]. Compared to skeletons, the human mesh can provide more structural information. There are various human mesh recovery methods [99], [100] to construct a complete 3D body model. ModelGait [39] fine-tunes a mesh recovery model on a gait dataset and distinguishes different subjects via extracted structural parameters, showing the promising performance of utilizing human mesh as auxiliary supervision information. Gait recognition based on body meshes will be an exciting topic in the future with the improvement of human body mesh estimation accuracy.

• *Depth Image:* Unlike color images, depth images can provide a 3D structure of bodies since each pixel value is the distance between the object and the camera. The low-priced depth cameras like Kinect [101] provide the possibility for gait recognition using depth images. In [102], traditional GEI is compared with depth-based templates such as Depth-GEI, DGHEI, and GEV, and experiments show that depth templates can achieve better performance. A comprehensive review on gait recognition with depth images can be found in [103], introducing public depth datasets and most methods with depth images. Depth image-based gait recognition has a primary challenge in that a depth camera can only capture data in a range of 10 meters. Besides, the active infrared light from depth cameras will decrease dramatically with the distance and can also be disturbed by sunlight. For those reasons, gait recognition with depth images is difficult to deploy into an outdoor system to capture gait from a distance.

• *Dynamic Event Stream:* Event stream cameras can capture high-speed movements without blurs. The dynamic vision sensors can capture microsecond-level pixel intensity changes as events by a class of neuromorphic devices. By converting the event stream into image-like representations [25], CNN-based methods can contribute to extracting discriminative features from this data modality. Event streams may provide much more promising performance from their ability to capture dynamic fine-grained motion. In the literature, EV-Gait [25] is the first work on dynamic vision sensors for gait recognition. It achieved nearly 96% recognition accuracy in a real-world setting and comparable performance with state-of-the-art RGB-based gait recognition methods on the CASIA-B benchmark. However, more studies on this new sensor are needed, and it is great potential for event cameras to deploy gait recognition systems in the future.

• *Point Clouds:* Point clouds are typically produced by IoF sensors like LiDAR, which is capable of facilitating outdoor gait recognition [42], [61], [104] with precise 3D information. LiDAR sensor provides not only robust gait representation in many challenging scenarios such as poor illumination and occlusion, but it can also perform gait recognition at a large range of distances. Point clouds have been preferred in recent years because they provide precise 3D geometry. LidarGait [42] demonstrates that LiDAR-based gait recognition can outperform traditional camera-based methods by a large margin in challenging conditions like poor illumination and cross-view scenarios. Besides, point-based gait recognition shall be in favor of privacy-preserving scenarios such as nursing homes, and it is potentially superior to camera-based methods in biometrics protection with less sensitive information.

B. Feature Learning for Gait Recognition

The gait feature learning methods can be generally categorized into handcrafted feature-based methods (Section II-B1) and deep feature-based methods (Section II-B2).

1) Handcrafted Feature-Based Gait Recognition: Gait recognition methods can be broadly categorized into two groups: model-based and appearance-based methods, depending on whether they explicitly model the structure of the human body [3]. **Model-based** methods are divided into structural and motion models. *Structural models* [105], [106], [107] use static body parameters like stride length, while *motion models* focus on dynamic features like phase-weighted magnitude spectra [108] or Fourier description [2]. Some approaches combine both for improved accuracy [109]. **Appearance-based** methods extract gait features directly from input data without modeling the human body, dominating the field due to their efficiency. For instance, Collins et al. [80] proposed a silhouette-based method, while Wang et al. [110] applied principal component analysis to reduce feature dimensionality. Later on, due to the widespread adoption of gait templates, appearance-based methods dominate the field of gait recognition.

2) Deep Feature-Based Gait Recognition: Deep feature learning has emerged as the dominant approach for gait recognition, revolutionizing how representations are extracted. While traditional handcrafted feature-based methods rely on prior knowledge and expertise to design descriptors, deep feature learning methods focus on network architecture and loss functions. With deep learning, neural networks automatically extract gait features through a series of stacked layers. As shown in Fig. 3, early methods in this area overlooked temporal information within gait sequences and heavily relied on gait templates [9], [10]. However, recent works [78], [111] have made significant strides in extracting frame-level features directly from a sequence of silhouettes. The typical pipeline of deep gait recognition involves front-ground segmentation, gait alignment, and feature extraction. Lastly, some recent works [39], [62] have explored all-in-one models that learn all necessary steps from input to output.

III. A Glimpse of Deep Gait Recognition

Deep gait recognition methods generally comprise three main components: a backbone for feature extraction, a bottleneck for spatial-temporal feature aggregation, and a head for representation mapping.

A. Backbone Networks

The backbone of a deep gait recognition model serves as the feature extractor, transforming input data into deep abstract representations. Traditional gait recognition methods used hand-crafted features like gait templates [9], [85], [112]. However, with the emergence of deep learning, learned

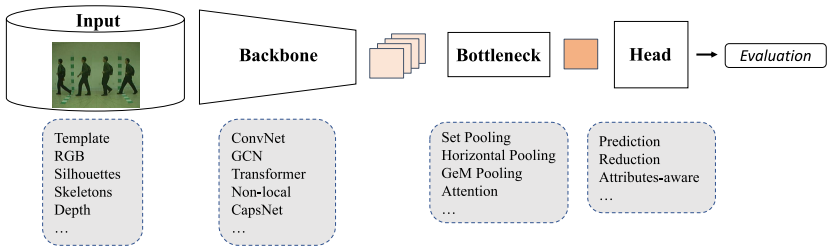


Fig. 4. The typical workflow of deep gait recognition models.

features from deep neural networks have shown significant improvements over manually designed features.

Convolutional Neural Networks (CNNs) are the seminal backbone networks in deep gait recognition models. While early works used 2D CNNs and achieved better results than handcrafted features [27], [113], [114], they lacked effectiveness in capturing temporal information. To address this, more advanced backbones [19], [30] with 3D convolutional layers were introduced to extract robust spatiotemporal features for improved identification.

Recent literature has explored gait-specific backbone networks to enhance performance further. For example, focal convolutions [28], [30] were proposed to capture fine-grained gait patterns from horizontally partitioned inputs, and 3D local convolutions aimed to obtain component-specific information from the pedestrian body adaptively [31]. Additionally, graph convolution networks have been adopted for models taking skeletons as input, effectively extracting structural and dynamic information [40].

Notably, a significant difference between gait recognition and other visual recognition tasks lies in the use of relatively shallow backbone networks (ranging from 4 to 8 layers). For instance, a ResNet-like backbone with only 9 convolutional layers as proposed in GaitBase [35] has been widely adopted as the feature encoder for silhouettes. Many visual recognition tasks exploit very deep models with tens or hundreds of layers. We analyze this phenomenon mainly because gait recognition often utilizes silhouettes and skeletons, which contain lower information entropy compared to RGB images or videos. This observation highlights the importance of using the original RGB modality to take advantage of deeper models for improved precision.

B. Bottleneck Networks

Bottleneck networks are built for aggregating dynamic gait information and enhancing the discriminative features via temporal modeling and spatial manipulation, respectively.

Early approaches often overlooked the importance of designing bottleneck networks, relying on fully connected layers for feature reduction and global feature learning. However, these methods were sensitive to noise and could overfit on poorly segmented data, leading to significant performance degradation [20], [21], [114]. Therefore, Horizontal Pyramid Pooling [24], Patch Pyramid Mapping [115], and Generalized-mean Pooling [30] were proposed to capture fine-grained gait cues, which leveraged partial features to prevent overfitting. Hou et al. [29] introduced feature lateral learning, where the inherent feature pyramid utilizes multiple-scale features to enhance gait representations. Besides, Huang et al. [32] measured the importance of different parts across frames, which exploited the most discriminative parts and generated more robust spatio-temporal representations. Considering temporal representations modeling, Set Pooling [24] proposed various feature pooling strategies along the time dimension. The Micro-motion Capture Module [28] and Adaptive Temporal Aggregation [32] made use of attention mechanism to extract gait patterns in the long short-term manner, and recurrent neural networks [116] are also able to perform adaptive temporal representation from sequential inputs.

C. Head Networks

Head networks following the bottleneck are optional but serve specific purposes. In object detection, deep models typically have two heads for object recognition and bounding box localization [117]. Similarly, gait recognition can employ various head networks for multi-task gait recognition.

The verification head [21] and identification head [29] are the most commonly used heads in deep gait models. While identification heads achieve satisfying precision with a strong backbone and bottleneck, it was noted that their feature dimensionality may not be compact enough for practical applications. To address this, the compact block [29], [31] was introduced to reduce representation dimensionality and memory usage. Moreover, other head networks have been proposed, such as quality-aware [118], view-aware [119], condition-aware [120], and gender-aware [121], offering contributions for multi-task gait recognition.

D. Loss Functions

The loss functions are designed to measure the similarity between samples in the embedding space. By minimizing the loss function during training, the deep networks learn to map similar samples to nearby points and dissimilar samples to faraway points in the embedding space. Gait recognition conducts deep metric learning using many commonly used losses, including cross-entropy, contrastive, and triplet losses.

Triplet loss [122] is widely used in recent state-of-the-art methods [35]. Instead of using pairs, this loss takes distance triplets (*anchor*, *positive*, *negative*) as input. The loss pulls the positive samples close to the anchor and pushes the negative away from the anchor. In order to prevent the features from converging into a small space, the distance between the

anchor-negative pair should be at least one margin m farther than that of the anchor-positive pair.

$$L_{tri} = \frac{1}{N_{tp+}} \sum_{\substack{a,p,n \\ y_a=y_p\neq y_n}} max(m + d(a,p) - d(a,n), 0),$$

where N_{tp+} denotes the number of triplets of non-zero loss terms in a mini-batch, a, p, n stand for anchor, positive and negative, respectively. $d(a,p)$ and $d(a,n)$ denotes the distance between anchor-positive and anchor-negative respectively. The hyper-parameter m in contrastive loss refers to the margin.

Other losses: The aforementioned loss functions achieve significant performance for gait recognition, but these losses are also widely used in many tasks like image classification, face recognition, and person re-identification. In the context of human-centric tasks, many challenges such as cross-view and changing appearance have not been well solved. Toward solving challenges of the deep gait recognition, many work has designed gait-specific loss functions including, Angle Center Loss [111], Quintuplet Loss [26], and View Loss [119], [123].

E. Evaluation

Evaluating the performance of gait recognition algorithms depends on: evaluation *metrics* and *protocols*.

In gait recognition, the choice of evaluation metrics depends on the specific recognition modes being considered. In the one-to-one verification mode, the performance is typically evaluated using operating characteristic curves (ROC), which provide a visual representation of the trade-off between the true positive rate and the false positive rate. Besides, the one-to-many identification mode employs several primary metrics, including Cumulative Matching Characteristics (CMC) curves [15], [21], mean Average Precision (mAP) [68], [73], and mean Inverse Negative Penalty (mINP) [43], [124]. However, on the other hand, the numberial metric are far behind the demands of developing modern deep gait models because deep learning methods lack of explainability. Fortunately, recent advancements in explainable AI like t-SNE [32] and activation visualizations [33], now offer qualitative insights in order to enhance model interpretability and guide the architectural design.

Evaluation protocols may vary across different datasets, as each dataset often emphasizes specific aspects or challenges of gait recognition. A dataset may include multiple evaluation settings, resulting in multiple evaluation protocols. Since a survey by Hou et al. has provided a comprehensive study on evaluation protocols for gait recognition, we recommend referring to [73] for more detailed information and a more thorough understanding.

IV. Deep Representations Learning

In deep gait recognition, representation learning involves the extraction of abstract deep gait descriptors using deep networks. In this section, deep gait feature learning will be discussed from input-level (Section IV-A) and feature-level (Section IV-B) perspectives.

A. From Input-Level Perspective

Deep models learn different gait representations given different input data. Depending on the characteristics of input data, such as ordered frames or shuffled frames, deep models can capture different aspects of gait information. For instance, when ordered frames are used, deep models can effectively capture dynamic motion, while micro-motion features may be neglected when shuffled frames are used. In this section, we explore deep representation learning from an input-level perspective, focusing on the following three aspects.

1) 2D/3D Representation Learning: 2D representation learning extracts geometric information from data captured by 2D visual sensors [2], [3], [67]. With the simplicity and efficiency of the 2D representation, 2D representations have dominated gait recognition for over 30 years. The commonly used 2D representations range from images, silhouettes, skeletons, and optical flows. Among them, silhouettes are the most popular 2D representation because silhouettes are easy to get and have precise body shape information. Using silhouettes as gait representation, many work [35], [78], [125] achieve satisfying performances in both indoor and outdoor environments. However, silhouettes lack representing inter-frame motion information, which inspires methods using optical flows to exploit spatial-temporal cues explicitly [88]. Besides, model-based methods also utilize 2D skeletons [37], [40] as input, trying to learn better view-invariant features, but model-based methods suffer from the precision of the pose estimation model [60]. Therefore, more research [27], [39], [62] focuses on disentangling gait-unrelated and gait-related features from RGB images.

3D representation learning aims to learn features from 3D data captured from either 2D-to-3D estimation models or advanced sensors. 3D representation is preferred because of its outstanding performance in handling viewpoint changes compared to 2D representation [42]. Estimation-based 3D representation, such as 3D skeletons [37], [93], [126] and 3D human mesh [43], [99], [100], can enhance the performance of gait recognition. However, estimation-based 3D representation is sensitive to many factors such as illumination and resolution, which limit its performance in challenging real-world scenarios. To address this issue, more recent research has focused on capturing depth images of human bodies [102], [103], [127] from depth cameras. However, depth cameras are only feasible indoors and cannot be used when the distance to pedestrians is over 10 meters. In recent years, Shen et al. [42] have used LiDAR sensors to obtain precise 3D structures of the human body and demonstrated LiDAR-based gait recognition with robust cross-view performance, outperforming 2D representation-based methods by a large margin.

2) Template/Frames-Based Representation Learning: Representation learning can be categorized into *frames-based* methods and *template-based* methods, according to whether utilizing a whole sequence of frames as input or not.

Template-based methods receive a gait template as input for gait feature extraction. These methods were popular before the deep learning era when traditional learning methods like SVM [82] and Boosting [84] struggled to extract effective

features from high-dimensional data. Therefore, researchers designed various gait templates [9], [10], [11], [85], [128], and these templates achieved high accuracy in many datasets [56]. For example, Gait Energy Image (GEI) represents the average of cyclic silhouettes, preserving static information while partially losing gait motion information. As the most popular gait template, GEI [9] is widely used all the time because it is computationally efficient and temporally robust. However, with the advancement of deep learning, many recent methods [32], [78], [81], [124], [129] have focused on frames-based representation learning to extract dynamic motion features for better performance.

Frames-based methods utilize all frames of gait sequences as input, enabling the extraction of fine-grained spatial information from diverse modalities such as silhouettes [32], [78], [129], body skeletons [40], SMPL [39], [43], and 3D point clouds [42], [61], [104]. These methods also leverage sequential inputs to model gait motions using various temporal modeling techniques, including set-based [23], long short-term [81], shift-based [124], and attentive modeling [32]. By combining spatial and temporal information, frames-based methods enhance recognition robustness and accuracy, making them a prominent trend in deep gait recognition.

3) Shuffled/Ordered Representation Learning: Shuffled learning utilizes unordered sequences as model input to aggregate set-based temporal information, making it robust to frame permutations and extendable to cross-scenes scenarios. This approach has been widely used in deep gait recognition [23], [114]. By regarding gait as a set of gait silhouettes, set-based shuffled learning is robust to scenarios when the input samples contain discontinuous frames or have a frame rate different from the training dataset [24]. Although ordered representation learning has shown outstanding performance in in-the-lab datasets, models using shuffled inputs outperform those using ordered inputs in outdoor datasets [42], [43], [60]. This is because people walk at varying speeds and routes in the real world, unlike the in-the-lab dataset setting. However, recent research [59], [130] has demonstrated that incorporating residual learning and increasing the number of layers can improve recognition accuracy on in-the-wild datasets, even when using ordered inputs for temporal feature extraction.

Ordered learning involves taking sequences in their natural order for motion modeling. For tasks that are based on sequences in computer vision, using ordered frames is the most straightforward and natural way to model fine-grained motion [131]. Ordered representation learning models temporal features via Recurrent Neural Networks (RNNs) [132], Convolution Neural Networks (CNNs) [113], Long Short-term Memory Networks (LSTM) [44], or Transformer Networks [59]. This ordered motion learning is effective for in-the-lab datasets that contain gait sequences only in walking status [23], [28], [30]. However, the use of temporal learning from ordered inputs has underperformed when applied to in-the-wild datasets [43], [60]. This is due to two possible reasons, as shown by the results reported in GREW [60], Gait3D [43], and SUSTech1K [42]. Firstly, people may walk at diverse speeds and routes in real-world scenes [133]. Secondly, the quality of gait representations in the inputs might be low, disrupting dynamic motion learning. To address this, recent studies have introduced deep residual learning [59] and continuous frame sampling [124] to capture gait motion patterns from ordered inputs in the real-world scenario.

B. From Feature-Level Perspective

In this part, we discuss deep representation learning from a feature-level perspective.

1) Global/Local Representation Learning: Global representation learning refers to extracting holistic features from the entire human body. Early gait recognition methods based on deep learning used global representation learning. For example, Hossain and Chetty [134] introduced deep learning into gait recognition by using several layers of convolution to extract deep gait features from the GEI templates. Another example is that Wu et al. [21] employed a 3-layer CNN to process GEIs of two gait sequences from two branches, utilizing verification mode to determine whether the two samples belong to the same person. Moreover, global representation can also be learned from human skeletons estimated from images by pose estimation models such as OpenPose [93], or from 3D point clouds [104] gained by LiDAR sensors. These model-based feature learning methods, as demonstrated in studies [38], [135], [136], have shown robust performance in cross-view recognition. However, gait recognition is an instance-level recognition task that relies on subtle differences to distinguish people. Therefore global representation is suboptimal to methods that explicitly learn features from multiple local regions.

Local representation learning focuses on capturing fine-grained features by exploiting spatial representations from specific regions of human bodies. In the traditional era, Liu et al. [137] highlighted the significance of different body parts having distinct shapes and moving patterns. Recent works have proposed effective methods for learning local representations. For example, Fan et al. [28] introduced GaitPart, utilizing focal convolutions on horizontally partitioned body parts. Shen et al. [33] addressed the limitation of hard partitioning and proposed ReverseMask, randomly generating masks to force CNNs to learn local features. Horizontally Pyramid Pooling, employed in methods like [23], [28], has become a widely used technique for separating holistic features into multiple horizontal strips to capture local patterns. Local features can also be extracted from pose-driven regions of interest (RoI) [138], attention regions of appearance [139], body components through parsing [137], or patch-level approaches [140]. Research [30], [111] has demonstrated that local representation learning-based methods can achieve better results.

In conclusion, global features focus more on holistic information, whereas local features focus on partial information. Global features contain more coarse information, while local features contain fine-grained information. GaitGL [30] combines the two kinds of features for better performances, which is a reasonable solution for learning better gait representations.

2) Single/Multi-Scale Representation Learning: Single-scale representation learning involves the process of learning representations at a fixed scale, where a single level of abstraction is used to represent the features of the input data. This method has been extensively employed in traditional machine learning algorithms such as XYT template [5] and gait in eigenspace [141]. Additionally, single-scale representation learning has been widely used in various visual tasks in the deep learning era due to its implementation simplicity and low computational cost. For instance, GEINet [20] takes GEIs at a holistic scale as input and extracts deep features for recognition. Other methods [21], [22], [114] adopted single-scale learning as well.

*Multi-scale representation learning*involves learning representations from multiple levels to capture different scales of features. This approach has been widely adopted in object detection and person re-identification [142], [143] and is effective for gait recognition because it enhances the discriminative capabilities of different human body parts. One example of this is the use of horizontal pyramid pooling [142] in GaitSet [23], [24], inspired by multi-scale learning in person re-identification. This method has been employed in many subsequent works, such as GaitGL [30], GaitPart [28], and CSTL [32]. Additionally, GLN [29] merges the features from different stages in a top-down manner to enhance the gait representations with multi-scale robustness.

3) Long/Short-Term Representation Learning: Long-term representation learning extracts human dynamic motion at a temporally large scale. At the input level, template-based methods use a sequential set of gait images to construct a gait template [56]. At feature-level long-term representation learning, long-term features can be captured by temporal pooling [23], [24], temporal attention [32], or LSTM [144]. While GaitSet [23] and GLN [29] focus on global set-level temporal feature extraction, but they neglect inter-frame dependency modeling. To address this, GaitGL [30] employed multiple stacked layers of 3D convolutions to model inter-frame features. Differently, CSTL [32] proposes a feature selective pooling module that preserves the most discriminative spatial local features into a final representation. As experiments presented in outdoor datasets [42], [43], [60], long-term temporal cues perform robustly in complex scenarios where human walks in diverse routes and cameras capture gait sequence from different viewpoints.

Short-term representation learning involves extracting gait micro-motion within the adjacent frames. Current literature proposed various strategies for modeling gait local patterns, including LSTMs [44], 1D convolutions [28], [32], [145], 3D convolutions [19], and temporal shift [124]. LSTMs, a type of recursive neural network, are commonly used in the field of speech recognition and handwriting recognition for sequential signals and have also been widely used in gait recognition, such as PTSN [38], and Zhang's GaitNet [116]. GaitPart [28] extracted local temporal clues by 1D convolutions and subsequently aggregated them into compact features with inter-frame dependency. Recent methods [30], [123] using 3D convolutions extract spatial-temporal features simultaneously, enhancing the ability to capture fine-grained gait dynamics with discriminativeness. However, while short-term representation learning has achieved outstanding results on laboratory datasets, these in-the-lab datasets capture gait sequences with humans walking continuously in limited viewpoints. When conducted on in-the-wild datasets, methods with temporal convolutions to extract short-term gait cues underperform methods based on long-term representations [42], [43].

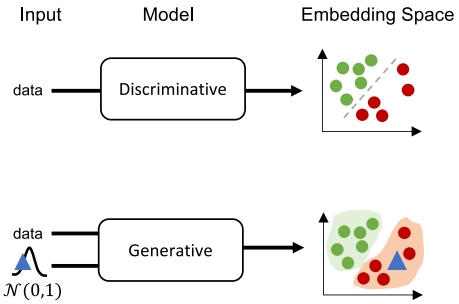


Fig. 5. Illustration of the discriminative and generative model.

V. Deep Models for Gait Recognition

In this section, we categorize deep gait models into discriminative and generative models. As illustrated in Fig. 5, discriminative models are trained to learn the classification boundary between different identities, while generative models do not learn the decision boundary but model the underlying distribution of the data, enabling the generation of new samples from the given data.

A. Discriminative Model

The discriminative models are the methods of directly learning decision boundaries to recognize human gait. To learn discriminative gait representations, these models can be divided into two categories, including the single-modal model and the multi-modal model.

1) Single-Modal Architecture: Learning deep representation from a single modality is a classical and most commonly used method in many visual recognition tasks. The single-modal models dominate the study of deep gait recognition because these methods are often more computationally efficient than multi-modal models. As gait recognition involves spatiotemporal modeling to preserve appearance and motion features, we summarize single-modal models on deep gait recognition into three classes according to the position where spatiotemporal information is fused together in the deep networks. These three classes of single-modal models are, namely, (1) divided early fusion models, referring to models in which spatial information extraction is followed by temporal modeling. (2) divided late fusion models, capturing spatial information and then aggregating temporal motion. (3) joint fusion models, which consider spatiotemporal modeling simultaneously. The single-modal gait recognition methods also follow the same development process, as detailed in three following aspects.

Divided early fusion models refer to the category of methods that fuse gait inputs at the very initial stage and then extract gait representation with spatiotemporal

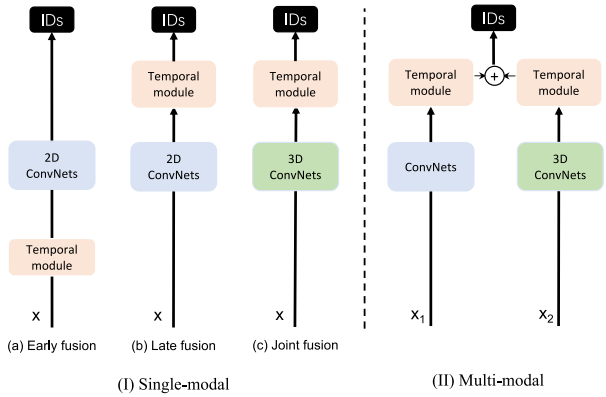


Fig. 6. Illustration of four types of models in discriminative architecture.

information as illustrated in Fig. 6(I). This architecture is also called template-based architecture in other literature. As many works [20], [21], [78] indicate that divided early fusion architecture obstructs deep feature extraction spatially and temporally, this architecture had been adopted in many early methods and has still been utilized in many recent works [146] because of its computational friendly.

Among various template-based methods, gait energy image [9] is the one that cannot be neglected from the history of gait recognition. By averaging gait silhouettes along the temporal dimension, GEI showed its simplicity and efficiency in cross-view datasets with over 10,000 subjects [58], making it become the most dominant template among traditional gait recognition methods. Therefore GEI continued to be the first choice of data modality to learn gait features when deep learning first meets gait recognition. Particularly, Shiraga et al. [20] proposed a CNN-based model, named GEINet, which composes only two sequential triplets of convolution, pooling, and normalization layers, along with two subsequent fully connected layers. GEINet utilized CNN on GEI, and marginally outperformed the first deep gait recognition method paper [134] extracting features from silhouette frames by a Restricted Boltzmann Machine.

Regarding gait-based human identification in verification mode, Wu et al. [21] proposed a 3D CNN network with triplets of continuous gait silhouettes as the input. They also proposed LBNet to learn paired filters on a pair of GEIs at the initial input layer and to output a similarity of two input GEIs. Furthermore, a comprehensive discussion was proposed by Takemura et al. [22], deeply investigating deep gait recognition in verification versus identification mode with contrastive loss and triplet ranking loss, respectively. The study found that triplet ranking loss could effectively improve cross-view accuracy from consideration of the intra-subject spatial displacement caused by view-invariant. Although the aforementioned divided early fusion models significantly improved performance, these methods [20], [21], [134] based on GEI still suffer from performance degradation when evaluated on the clothes-changing subsets. Therefore, Yao et al. [147] introduced a novel gait template called skeleton gait energy image. This method could perform better in an unconstrained environment with changing views and clothes by averaging sequential skeletons instead of silhouettes to generate the GEI.

Though divided early fusion models have significantly boosted recognition accuracy by applying spatiotemporal feature fusion at the input level, more subsequent work [81], [129] indicated that the methods with spatiotemporal fusion at feature level capture more informative gait features spatial-temporally.

Divided late fusion model learn gait appearance before modeling and aggregating motion. This architecture advanced in fine-grained feature extraction on frame-level gait input rather than gait template, which helps to gain performance improvement and preserve spatial gait patterns in detail. Besides, the divided late fusion architecture fusing spatial-temporal features separately in two stages can help the community to investigate and understand human gait spatiotemporally better. With these advantages, many works have emerged for deep gait recognition.

Regarding spatial information modeling in the context of divided late fusion architecture, 2D convolution neural networks are typically the first choice to encode features from structured inputs, including RGB images, silhouettes, depths, and infrared images. A milestone of the 2D CNN-based method, *i.e.*, GaitSet [23], established a light and efficient network consisting of six convolutions along LeakyReLU activations and two max-pooling layers, which first achieved 70.3% cross-view recognition accuracy on CASIA-B dataset under clothes changing condition. The following works utilized such GaitSet-like backbone integrated with local [28], global-local [30], and multi-scale representation learning strategies [41], greatly boosting gait recognition performance on the laboratory datasets. However, these models are light and shadow, suffering a significant degradation when the models are employed from indoor to outdoor scenes. Fan et al. [35], [59] contributed to such performance degradation because shadow architectures are insufficient to learn discriminative features from unprecedented challenges within real-world scenes, such as complex backgrounds, harmful occlusion, unpredictable illumination, arbitrary viewpoints, and diverse clothing changes. Therefore, GaitBase [35] with residual learning deepened networks and improved performance on in-the-wild datasets significantly. DeepGaitV2-2D [59] further increased the depth of networks and demonstrated the recognition accuracy was positively relevant to the depth of models on in-the-wild datasets. In addition to convolutional neural networks as spatial feature encoders for structured inputs like silhouettes and RGB images, other networks such as graph networks [40] and multilayer perceptron networks [42] are used to encode gait features for different inputs, including skeletons and point clouds, respectively.

For temporal information modeling in the context of divided late fusion architecture, there are two main categories of temporal modeling methods: parameter-free and parameter-based. On the one hand, parameter-free methods use statistical functions to aggregate temporal information among sequences into sequence-level features. Temporal average pooling [24] is a common example of this approach, which calculates the average value of each feature over the entire temporal dimension of the sequence. Temporal max pooling [23] is another statistical function that summarizes temporal information with

superior recognition performance. However, they may not capture complex temporal dynamics and variations in the gait patterns as effectively as parameter-based methods. On the other hand, parameter-based methods use learnable parameters to capture temporal dependencies within the sequence data. These methods often incorporate recurrent neural networks, 1D convolutions [28], and attention mechanisms [125] to model the temporal dynamics of the gait sequence. While some works utilize learnable parameters to capture either long-term or short-term temporal features, recent work [32] suggests that multi-scale temporal feature learning with long-term, short-term, and frame-level features can better distinguish gait variants at different scales of temporal clues. Though these parameter-based methods help models learn informative gait dynamics, they come at the cost of increased computational complexity, which may not be practical for many real-world applications and even result in inferior performance compared to parameter-free methods [42], [43], [68].

Joint fusion models consider both dimensions simultaneously in the learning process, while divided fusion models learn space and time dimensions independently. Therefore these models can potentially capture more complex spatiotemporal patterns in gait sequences.

To learn spatiotemporal features from structural gait sequences, such as depths and images, 3D CNNs [148] are ideal processing units, as they understand sequential sequences as a 3D tensor with the shape of $C \times T \times H \times W$, where T represents the number of frames, H and W are the spatial dimensions, and C refers to the number of channels. The seminal work using the 3D convolutional network (C3D) for gait recognition is [19]. While inspiring, C3D is computationally intensive and hard to optimize on small-scale datasets. To address this, Lin et al. [30], [81] factorized a 3D kernel (*e.g.*, $3 \times 3 \times 3$) into two separate operations, a 2D convolution (*e.g.*, $1 \times 3 \times 3$) for spatial pattern modeling and a 1D convolution (*e.g.*, $3 \times 1 \times 1$) to construct temporal connections. Following this line, DeepGaitV2-P3D [59] compared 3D residual convolutions (R3D) and Pseudo 3D residual convolutions (P3D) [149] and found that two types of 3D convolutions achieve comparative performance, but P3D has fewer learnable parameters (11.1 vs. 27.5M) and lower computational cost (2.9 vs. 6.8GFLOPs). Additionally, DeepGaitV2-P3D greatly improved the recognition accuracy of DeepGaitV2-2D by up to 9.1% on various datasets with only a slight increase in parameters and computational cost.

In addition to 3D convolution, there are many alternatives for performing joint spatial-temporal modeling. Recently, temporal shift operation [124] further enhances temporal information modeling among neighboring frames. These methods are computationally efficient and do not require additional parameter tuning, making them popular in many gait recognition applications. Similarly, DyGait [57] and SiMo [150] explicitly model human dynamical motion with prior knowledge by facilitating the temporal difference operation between frames. Recently, many works have explored transformers [151] and spatial-temporal graph convolutional networks to model gait features from nonstructural inputs such as skeletons.

2) Multi-Modal Architecture: To construct discriminative deep features for gait recognition, alternative architectures explore the idea of utilizing multiple modalities. Unlike the aforementioned single-modality models that rely on architecture design to learn different semantic information, multi-modal methods [152], [153] explicitly utilize diverse semantic information from many input streams. In some literature, multi-modal architecture is defined as networks with multiple branches. In this paper, the term multi-modal models refers specifically to networks that utilize multiple modalities. Therefore a single modality having multiple pathways like GaitSet [23] is categorized into the single-modal model.

To capture motion features of human gait, considering optical flow appears to be a reasonable solution. The first attempt is introduced by the seminal paper on Two-Stream Networks (TSN) [154] for human action recognition, where TSN adds a second pathway to learn the temporal information in a video by training a CNN on the optical flow stream. Following this trend, Delgado-Escano et al. [155] extends to add a third path to capture structural information on the depth images. Furthermore, UGaitNet [156] inherits the multimodal recognition framework and further discusses a neglected problem within multimodal gait recognition that some modalities can be missing at test time. To tackle such a problem, UGaitNet integrated a gate mechanism and a merge operation to combine information from modality-specialized branches. In that way, when a modality is missing, the gate can disable the input of such modality to its corresponding branch in order not to penalize the performance.

To fully use body shape and structure information from the given modalities, the multi-modal methods integrating silhouettes and skeletons [126], [151], [157] are the most commonly used. Precisely, silhouettes maintain body shape information but are sensitive to the factors of changing human appearance, such as clothing and camera viewpoints. Skeletons are robust to the viewpoints but ignore the body shape details. GaitHybrid [158] was one of the first multi-stream networks that proposed a hybrid silhouette-skeleton body representation for gait recognition. Differently, GaitHybrid employed CNN on the skeleton heatmap to capture body structure information instead of unitizing GCN on the raw skeletons sequence with a shape of $C \times T \times N$, where N denotes the number of body key points, T represents the number of frames, and C refers to the number of channels of body keypoints. Typically, the silhouette-skeleton models as proposed in [126], [157] consist of a pose feature encoder and a silhouette feature encoder independently for two modalities. Instead of obtaining human structural information from human skeletons, recent methods target learning 3D body shapes from diverse modalities, including 3D human mesh [39], depth images [156], and 3D point clouds [61].

B. Generative Model

Unlike discriminative models that learn decision boundaries to make predictions, deep generative models learn the actual data distribution via reproducing samples as closely as possible from the given inputs [159], [160]. Specifically, generative

models summarize input data distribution and synthesize samples in the given data distribution. In the context of gait recognition, deep generative models can be used to eliminate the challenges of gait appearance changing caused by various real-world factors, including object carrying [161], clothes changing [27], and diverse viewpoints observing [160]. Generally, generative models for gait recognition can be divided into two main categories. One is the deep autoencoder network, constructing probability distribution of given data based on maximum likelihood estimation. The other one is the generative adversarial network, finding the Nash equilibrium of the minimax game between the data generator and sample discriminator. In the following parts, we provide a detailed introduction to these two models for generative gait models.

1) Deep Autoencoders: Deep autoencoders (DAEs) are a specific type of neural network architecture that aims to learn compact representations, also known as latent representations. The main objective of DAEs is to encode the input data into a lower-dimensional feature representation using an encoder. Then DAEs decode it back to reconstruct the original input using a decoder. The encoder typically consists of fully connected and/or convolutional layers, while the decoder performs the inverse operations. By minimizing the reconstruction error, which quantifies the dissimilarity between the input and reconstructed data, DAEs effectively learn latent features. These latent features serve as concise representations that capture the essential characteristics of the original data and can be extracted for various applications, including tasks such as generating new samples or performing gait recognition. Following this line, Yu et al. [87] proposed a uniform model named SPAE that adopted autoencoders to learn invariant gait features for gait recognition. With the stacked autoencoders, SPAE encodes the given GEIs from various viewpoints with carrying and clothing conditions from the input layers, and then constructs GEIs from the side view with the normal condition. With a similar spirit, Babaee et al. [162] proposed an incomplete to complete GEI converter, which is able to transform incomplete GEI or even a frame of silhouette to the corresponding complete GEI. Furthermore, recent methods integrated disentangled representation learning with encoder-decoder architecture. For instance, Li et al. [161] proposed ICDNet that consisted of an encoder to disentangle latent identity and covariant features from an input GEI and a decoder to perform two reconstructions. Zhang et al. [116] disentangled gait features in two components: pose and pose-irrelevant, while the improved version [27] further decoupled the pose-irrelevant features into discriminative canonical feature and appearance feature, improving gait recognition accuracy and providing interoperability. However, DAEs suffer from posterior collapse and imperfect reconstruction, resulting in uninformative latent space and blur sample synthesis.

2) Generative Adversarial Network: Generative Adversarial Networks (GANs) are representative generative models with a strong performance in data synthesis. These generative methods can effectively generate data with high quality and perverse identity consistency at the same time, which makes them very popular in the context of gait recognition for data generation. Generally, GANs involve two distinct networks, where a *generator* receives a latent code sampled from a prior distribution as input to generate data, and a *discriminator* distinguishes between real data and the synthesized data. Going through the zero-sum game between generator and discriminator, GANs can enjoy high-fidelity generation. Thus, GANs are gradually employed to solve many challenges of gait recognition, *i.e.*, differences in viewpoints, diverseness of clothing, and domain gap between different datasets. To perform gait recognition across two arbitrary views, GANs transform gaits into either a unified view or an identical view to the reference samples. For instance, Yu et al. [159] proposed a generative adversarial network called GaitGAN, which can transform the gait data from any view, clothing, and carrying conditions to the canonical side view of a walking gait in normal clothing without carrying objects. Nevertheless, it is a challenge to preserve human identity information when performing appearance transformation via generative models. To this end, GaitGAN adopts the identification discriminator to constrain the generated gait samples preserving identical identity information to the source gait samples, while GaitGANv2 further integrates a multi-loss strategy to enhance the preservation of human identity features further. A recent PSTN [166] had integrated a prior pairwise spatial transformer ahead of the recognition network, recusing spatial feature misalignment to enable recognition more robust on cross-view scenarios. Instead of focusing on the viewpoint transformation, Frame-GAN proposes to deal with the realistic challenge of low frame rate on gait recognition and then helps to improve gait recognition performance by increasing more gait frames. Toward gait recognition at low frame rate inputs, Xu et al. proposed a method named PA-GCRNet [164] that can even reconstruct a complete gait cycle of silhouettes.

The generative models have demonstrated the necessity to eliminate the challenges of various real-world factors. However, most generative methods heavily relied on supervised labels for training, lacking flexibility for practical usage. Recently GaitEditor [165] has adopted the GAN inversion technique to construct latent space with semantic separability, enabling multiple gait attribute manipulation. It shows the great potential of generative gait models with unsupervised learning toward generating a broad range of realistic attributes such as viewpoints, clothing, walking speed, gender, and age.

To clearly demonstrate our proposed taxonomy, we have selected several recent deep gait recognition methods and categorized them according to our taxonomy in Table II.

VI. Datasets and Evaluation

In this section, we first introduce the public gait datasets we can find in the literature to help researchers find their research of interest. Then we evaluate recent advances in the representative datasets to illustrate the development of deep gait models.

A. Datasets for Gait Recognition

We list all public gait datasets that we can find in the literature in Table III. It is concluded that large-scale gait datasets

TABLE II
CLASSIFICATION OF REPRESENTATIVE DEEP GAIT RECOGNITION METHODS BASED ON OUR PROPOSED TAXONOMY

	Method	Publication	Representation Learning						Model Architecture	
			Input-level			*Feature-level*				
Discriminative	GEINet [20]	ICB16	2D	Template	-	Global	Single-scale	Long-term	Single-modal	Divided early fusion
	GaitSet [23]	AAAI19	2D	Frames	Shuffled	Global & Local	Multi-scale	Long-term	Single-modal	Divided late fusion
	GaitPart [28]	CVPR20	2D	Frames	Ordered	Local	Single-scale	Long- & Short-term	Single-modal	Divided late fusion
	GaitGL [30]	ICCV21	2D	Frames	Ordered	Global & Local	Multi-scale	Long- & Short-term	Single-modal	Joint fusion
	SMPLGait [43]	CVPR22	3D	Frames	Shuffled	Global & Local	Multi-scale	Long-term	Multi-modal	Divided late fusion
	DyGait [57]	ICCV23	2D	Frames	Ordered	Global & Local	Multi-scale	Long- & Short-term	Single-modal	Joint fusion
	GPGait [41]	ICCV23	2D	Frames	Ordered	Global & Local	Multi-scale	Long-term	Single-modal	Divided late fusion
	GaitBase [35]	CVPR23	2D	Frames	Shuffled	Local	Single-scale	Long-term	Single-modal	Divided late fusion
	LidarGait [42]	CVPR23	3D	Frames	Shuffled	Local	Single-scale	Long-term	Single-modal	Divided late fusion
	MMGaitFormer [163]	CVPR23	2D	Frames	Ordered	Local	Multi-scale	Long- & Short-term	Multi-modal	Joint fusion
	BigGait [36]	CVPR24	2D	Frames	Shuffled	Local	Single-scale	Long-term	Single-modal	Divided late fusion
Generative	GaitGAN [159]	CVPRW17	2D	Template	-	Global	Single-scale	Long-term	Single-modal	GAN
	GaitNet [27]	CVPR19	2D	Frames	Ordered	Global	Single-scale	Long- & Short-term	Single-modal	DAE
	PA-GCRNet [164]	ECCV20	2D	Frames	Ordered	Global & Local	Single-scale	Long-term	Single-modal	GAN
	GaitEditor [165]	Arxiv23	2D	Frames	Ordered	Global & Local	Single-scale	Long- & Short-term	Single-modal	GAN

typically have limited attributes, while those with considerable attributes are small-scale. Large-scale gait datasets are essential to provide statistical evaluation for gait recognition. Especially in these years, the increasing data size and deep learning greatly advance performance. However, collecting a large-scale gait dataset demands significantly more time, storage, and cost compared to gathering a dataset of similar scale for face or fingerprint recognition.

To this end, one possible solution is that the research community can work together to collect data and train methods using federated learning or other privacy-protecting learning methods. The alternative is to create synthetic gait datasets using virtual human body models, which is also an exciting and promising solution. With concerns about data privacy and security, collecting a large gait dataset is much more challenging nowadays. Some laws, such as *European General Data Protection Regulation* (GDPR) [202] and the *Data Security Law of the P. R. of China* [203], put substantial restrictions to protect our privacy and improve data security in data collection and usage. It is a challenge and a new opportunity for the academic community to develop better methods to protect our privacy and improve the security of our society.

B. Performance Comparisons

Comparing deep gait models on diverse datasets is not easy since there are over 60 datasets, as reported in Table III, and each dataset might have diverse criteria and evaluation protocols. To present clearly, we select some representative datasets and report the state-of-the-art methods on these popular datasets, aiming to draw valuable observations from the performance comparison and analysis. To provide a comprehensive performance comparison on deep gait recognition, we evaluate deep gait recognition methods from three scenarios: (1) *cross-view scenarios*, focusing on evaluation on two popular datasets, CASIA-B [15] and OUMVLP [58]. (2) *in-the-wild scenarios*, reporting results on two popular outdoor datasets, Gait3D [43] and GREW [60]. (3) *Clothes-changing scenarios*, relating to three outdoor datasets with clothing change attributes, CASIA-E [133], CCPG [68], and FVG [27]. (4) *Point cloud-based scenarios*, focusing on gait recognition using 3D LiDAR point clouds and evaluating on two datasets, SUSTech1K [42] and LiCamGait [61].

1) Evaluation on Cross-View Gait Recognition: To explore cross-view gait recognition, the Institute of Automation, Chinese Academy of Sciences, released the *CASIA-B* [15] gait dataset in 2005, which is the first gait dataset containing over 100 subjects with walking sequences captured in dense viewpoints and diverse walking conditions, *e.g.*, walking normally (NM) or wearing a bag (BG) or a coat (CL). Alternatively, the Institute of Scientific and Industrial Research (*ISIR*), Osaka University (*OU*), also provided a cross-view dataset *OUMVLP* [58]. It is the largest multi-view gait dataset captured in the laboratory environment. OUMVLP contains 10307 subjects, and each participant contributes 28 sequences (7 cameras $\times$ 2 forward and backward $\times$ 2).

As listed in Table IV, many invaluable investigations have been proposed to tackle gait recognition toward cross-view scenarios. We can make three observations from the performance evaluation: (1) The deep models have progressively upgraded cross-view accuracy in both CASIA-B and OUMVLP datasets. The state-of-the-art methods [35], [57], [130] achieve over 90% rank-1 accuracy on two datasets under cross-view evaluation protocol. (2) The silhouette-based methods dominated the field of study, while methods taking other modalities (*i.e.*, RGB images and skeletons) are relatively lacking in the study. (3) These models behave differently on two datasets. In other words, some methods make comparable performance on CASIA-B, while there is a clear performance gap when implemented in OUMVLP. For example, GaitGCI [34] and DANet [130] achieve 94.5% and 94.6%, respectively, on CASIA-B. For OUMVLP, GaitGCI [34] outperforms DANet by 1.4% rank-1 recognition accuracy (92.1% v.s. 90.7%).

However, CASIA-B was introduced nearly two decades ago, its relevance has diminished in the face of more recent advancements and datasets. The evolution of gait recognition technology demands the use of datasets that present greater complexity and diversity, which CASIA-B no longer fully addresses. It suggests diverting gait recognition into more challenging settings like outdoor scenarios. (2) Recognizing pedestrians via other modalities is potential to investigate further. (3) For deep gait models, it is suggested to conduct statistically significant conclusions from the evaluation on a large-scale dataset.

2) Evaluation on In-the-Wild Gait Recognition: Although great achievements have been made in indoor laboratory

TABLE III
ALL PUBLIC GAIT DATASETS WE CAN FIND IN THE LITERATURE. THE RELATED INFORMATION OF EACH DATASET ARE ALSO LISTED

Institution	Dataset	Subjects	Sequences	Views	Variations	Environment	Available	Year
SUSTech,China	Scoliosis1K [167]	1,050	1,493	1	scoliosis	indoor	yes	2024
	SUSTech1K [42]	1,050	25,239	12	views, occlusion, clothing, illumination, 3D point cloud	outdoor	yes	2023
	GaitLU-1M [63]	1,035,309	1,035,309	1,379	Unlabelled	outdoor	yes	2023
CSU,China	CCGR [168]	970	1,580,617	33	53 diverse covariate	indoor	yes	2024
BNU,China	AerialGait [169]	533	82,454	10	aerial & ground views	outdoor	yes	2024
	DroneGait [170]	96	22,718	30	aerial views	outdoor	yes	2023
BJTU,China	CCPG [68]	200	16,566	10	clothing	outdoor	yes	2023
ZJU,China	VersatileGait [121]	10,000	1,320,000	44	age, gender, walking style	Unity3D	yes	2021
HDU,China	Gait3D [43]	4,000	253,309	39	views	in/outdoor	yes	2021
THU,China	GREW [60]	26,345	128,671	882	view, distractor, carrying, dressing, occlusion, surface, illumination, speed, shoes, trajectories	wild	yes	2021
SZU,China	ReSGait [171]	172	870	1	cl, carrying, trajectories	indoor	yes	2021
	RGB-D Gait [172]	99	792	2	views	indoor	yes	2013
OU-ISIR, Japan	OUMVLP Pose [95]	10,307	268,086	14	views	indoor	yes	2020
	OU-LP Bag [173]	62,528	177,973	1	carrying	indoor	yes	2018
	OUMVLP [58]	10,307	267,386	14	views	indoor	yes	2018
	OU-LP Age [174]	63,846	63,846	30	age	indoor	yes	2017
	Bag β [175]	2,070	4,140	1	carrying	indoor	yes	2017
	ST-1 [176]	179	-	1	speed	indoor	yes	2014
	ST-2 [176]	178	-	1	speed	indoor	yes	2014
	OU-LP c1v1 [177]	4,007	7,844	1	-	indoor	yes	2012
	OU-LP c1v2 [177]	4,016	7,860	1	-	indoor	yes	2012
	Speed [178]	34	612	1	speed	indoor	yes	2012
	clothing [178]	68	2,746	1	clothing	indoor	yes	2012
	view [178]	200	5,000	1	views	indoor	-	2012
	fluctuation [178]	185	370	1	fluctuation	indoor	yes	2012
UMA, Spain	MuPeG [179]	-	-	-	occlusion	indoor	yes	2020
Kyudai, Japan	PCG [180]	30	60	1	3D point cloud	-	yes	2020
MSU, US	FVG [116]	226	2,856	3	views, speed, carrying, cl	outdoor	yes	2019
IPVC, Portugal	GRIDDS [103], [27]	35	350	1	trajectories	indoor	yes	2019
ISR-Lisboa, Portugal	ks20	20	300	5	view	indoor	-	2017
GPJATK, Poland	GPJATK Dataset [181]	32	166	4	view, 3D data	indoor	-	2017
SDU, China	SDUGait [182]	52	1,040	2	trajectories views	indoor	yes	2016
MTASZTAKI, Hungary	SZTAKI-LGA [183]	28	11	1	3D point cloud	outdoor	yes	2016
WUST, Polan	BHV MoCap [184]	10	246	1	trajectories	-	yes	2015
IITs, Indian	Depth Gait [127]	29	464	2	view, occlusion, speed	-	yes	2015
PPGC-UFPel, Brasil	Kinect [185]	164	820	-	curve	indoor	yes	2015
KY, Japan	KY4D-B [186]	42	84	16	curve	indoor	yes	2014
	Shadow [187]	54	324	1	views, cl, bg	indoor	yes	2014
	KY4D-A [188]	42	168	16	views	indoor	yes	2010
A.V.A UCO, Spain	AVAMVG [189]	20	1,200	6	views, trajectories	indoor	yes	2013
WVU, US	WOSG [190]	155	-	8	views, Illumination	outdoor	-	2013
ITB, Indonesian	dataset [191]	212	-	1	-	indoor	-	2012
TUM, Germany	TUM-GAID [102]	305	3,370	1	time, carrying, shoes, occlusion	Indoor	yes	2012
	TUM-IITKGP [192]	35	1,645	1	times, appearance, bg, depth	indoor	no	2010
UAB, Spain	DGait [193]	55	605	1	trajectories	indoor	yes	2012
IIT, Italy	RGBD-ID [194]	79	316	1	trajectories, time, cl, speed	indoor	yes	2012
QUT, Australia	SAIVT-DGD [195]	35	700	1	speed, carrying, shoes		yes	2011
Soton, England	Multimodal [196], [69]	300	5,000	12	views	indoor	yes	2011
	Temporal [197]	25	2,280	12	views	indoor	yes	2011
	Small [3]	12	-	4	bg,cl,carrying, speed, footwear, views	indoor	yes	2002
	Large [16]	116	2,128	2	Terrain, direction, views	in/outdoor	yes	2002
	Early [3]	10	40	1	-	indoors	-	1997
TIT, Japan	TokyoTech DB [198]	30	1,902	-	speed	indoor	-	2010
CASIA, China	CASIA-D [199]	88	880	1	multi-modality	indoor	yes	2009
	CASIA-B [15]	124	13,640	11	views cl bg	indoor	yes	2005
	CASIA-C [17]	153	1,530	1	speed bg	outdoor	yes	2005
	CASIA-A [110]	20	240	1	walking direction	outdoor	yes	2001
BUAA, China	IRIP	60	4,800	8	gender view	indoor	-	2008
GT, US	GT Speed [106]	20	268	3	views time	in/outdoor	yes	2003
USF, US	USF [1], [137]	122	1,870	2	shoes, views, carrying, terrain, time, trajectories	outdoor	yes	2002
UMD, US	Dataset-1 [200]	25	100	4	views, long distance	outdoor	yes	2001
	Dataset-2 [200]	55	222	2	views, times	outdoor	yes	2001
	Dataset-3 [200]	12	-	1	views	outdoor	yes	2001
CMU, US	CMU-mobo [80]	25	600	6	speed, carrying, inclination	indoor	yes	2001
MIT, US	MITAI Gait [201]	24	194	1	times months	indoors	yes	2001
	Early [5]	5	26	1	-	indoor	-	1994
UCSD, US	UCSD [6]	6	42	1	-	outdoor	yes	1998
NTTBRL, Japan	NIT Gait [141]	7	70	1	same cl, shoes	-	-	1995

datasets, gait recognition would encounter more challenges in real-world environments. To delve into this open problem, Gait3D [43] and GREW [60] were constructed to promote gait recognition in real-world scenes. For Gait3D [43], it collected gait data from 39 cameras in a supermarket, which constructed a cross-camera gait recognition dataset with 4000 identities and 25,309 sequences in total. GREW [60] is the in-the-wild dataset with the most identities by far. It contains 26,345 subjects and 128,671 sequences from 882 cameras. Besides, it also provides 233,857 sequential distractors.

TABLE IV
CROSS-VIEW RECOGNITION PERFORMANCE ON TWO POPULAR DATASETS, CASIA-B [15] AND OUMVLP [58]

Input	Method	Publication	CASIAB [15]				OUMVLP [58]
			NM	BG	CL	Mean	Mean
Silhouette	GaitGAN [159]	CVPRW17	61.0	39.0	22.3	40.8	-
	ACL [111]	TIP19	96.0	-	-	-	89.0
	Song [78]	PR19	89.9	-	-	-	-
	Zhang *et al.* [129]	PR19	91.2	75.0	54.0	73.4	-
	GaitSet [23]	AAAI19	95.0	87.2	70.4	84.2	87.1
	EV-Gait [25]	CVPR19	94.1	-	-	-	-
	GaitPart [28]	CVPR20	96.2	91.5	78.7	88.8	88.7
	GLN [29]	ECCV20	96.9	94.0	77.5	89.5	89.2
	MT3D [81]	MM20	96.7	93.0	81.5	90.4	-
	Dresser [120]	TIP20	94.8	88.8	81.8	88.5	-
	PSTN [166]	TCSVT20	92.7	-	-	-	63.1
	PA-GCRNet [164]	ECCV20	-	-	-	74.7	-
	DV-GEIs [160]	IJCB20	83.4	-	-	-	-
	Yao *et al.* [204]	TMM2022	97.9	95.5	46.2	79.9	
	CSTL [32]	ICCV21	97.8	93.6	84.2	91.9	90.2
	3DLocal [31]	ICCV21	97.5	94.3	83.7	91.8	90.9
	GaitGL [30]	ICCV21	97.4	94.5	83.6	91.8	89.7
	SDML	ICASSP21	95.1	89.2	74.6	-	-
	SelfGait [205]	ICASSP21	93.5	90.1	81.3	88.3	89.9
	GaitMask [206]	BMCV21	96.8	93.2	84.2	91.4	90.2
	Koopman [207]	CVPR21	-	-	-	-	74.7
	Vi-GaitGL [119]	ICIP21	96.2	92.9	87.2	92.1	89.9
	CapsNet [208]	ICPR21	95.7	90.7	72.4	86.3	-
	SRN [209]	TBIOM21	97.1	94.0	81.8	91.0	89.9
	MvGGAN [210]	TIP21	97.1	91.9	75.6	88.2	58.4
	GaitMPL [211]	TIP22	97.5	94.5	88.0	93.3	89.6
	Lagrange [123]	CVPR22	96.9	93.5	86.5	92.3	90.0
	GaitSlice [212]	PR22	96.7	92.4	81.6	90.2	89.3
	MetaGait [125]	ECCV22	98.7	96.0	89.3	94.7	92.4
	STAR [213]	TBIOM22	97.3	93.9	84.0	91.7	89.7
	GaitStrip [214]	ACCV22	97.6	95.2	86.2	-	-
	GPAN [215]	TBIOM22	97.7	94.2	81.8	91.2	81.7
	ESNet [216]	TCSVT22	97.4	94.0	84.0	91.8	89.4
	GQAN [118]	TNNLS22	98.5	95.4	84.5	92.8	96.2
	GaitGCI [34]	CVPR23	98.4	96.6	88.5	94.5	92.1
	DANet [130]	CVPR23	98.0	95.9	89.9	94.6	90.7
	GaitBase [35]	CVPR23	97.6	94.0	77.4	89.7	90.8
	GaitCoTr [151]	ICASSP23	97.9	95.0	89.4	94.1	-
	Li *et al.* [150]	ICASSP23	98.0	96.1	88.1	94.1	90.5
	Wang *et al.* [157]	ICASSP23	97.7	93.8	92.7	94.7	91
Skeleton	PoseGait [37]	PR20	68.7	44.5	36.0	49.7	-
	GaitGraph [40]	ICIP21	87.7	74.8	66.3	76.3	20.4
	GaitGraph2 [217]	CVPRW22	82.0	73.2	63.6	72.9	67.1
	CNN-Pose [95]	TBIOM20	-	-	-	-	20.4
	GaitTAKE [218]	ICIP22	98.0	97.5	92.2	95.9	90.4
	GaitPoint [219]	ICIP22	93.0	92.0	81.1	88.7	-
	GaitMixer [220]	ICASSP23	94.9	85.6	84.5	88.3	-
	GPGait [41]	ICCV23	93.6	80.2	69.3	81.0	59.1
RGB	ModelGait [39]	ACCV20	97.9	93.1	77.6	89.5	95.8
	GaitNet [27]	CVPR20	92.3	88.9	62.3	81.2	-
	MvModelGait [221]	ICCVW21	98.1	93.4	80.7	90.7	89.7
	GaitEdge [62]	ECCV22	97.9	96.1	86.4	93.5	-
	Zhu *et al.* [222]	WACV23	97.5	94.8	84.1	92.1	89.8

TABLE V
GAIT RECOGNITION PERFORMANCE ON TWO POPULAR IN-THE-WILD DATASETS, GREW AND GAIT3D

Method	Venue	Gait3D [43]				GREW [60]			
		R@1	R@5	mAP	mINP	R@1	R@5	R@10	R@20
GEINet [20]	ICB16	5.4	14.2	5.1	3.14	6.8	13.4	17.0	21.0
GaitSet [23]	AAAI19	36.7	58.3	30	17.3	46.3	63.6	70.3	76.8
GaitPart [28]	CVPR20	28.2	47.6	21.6	12.4	44	60.7	67.3	73.5
GLN [29]	ECCV20	31.4	52.9	24.7	13.6	-	-	-	-
PoseGait [37]	PR20	0.2	1.1	0.5	0.3	0.2	1.1	2.2	4.3
GaitGraph [40]	ICIP21	6.3	16.2	5.2	2.4	1.3	3.5	5.1	7.5
CSTL [32]	ICCV21	11.7	19.2	5.6	2.6	-	-	-	-
GaitGL [30]	ICCV21	63.8	80.5	55.9	36.7	68	80.7	85	88.2
SMPLGait [43]	CVPR22	46.3	64.5	37.2	22.2	-	-	-	-
GaitGCI [34]	CVPR23	50.3	68.5	39.5	24.3	68.5	80.8	84.9	87.7
GaitBase [35]	CVPR23	64.6	-	-	-	60.1	-	-	-
DANet [130]	CVPR23	48	69.7	-	-	-	-	-	-
RealGait [115]	Arxiv23	-	-	-	-	54.1	71.5	77.6	81.7
DyGait [57]	ICCV23	66.3	80.8	56.4	37.3	71.4	83.2	86.8	89.5
MTSGait [124]	MM22	48.7	67.1	37.6	21.9	55.32	71.3	76.9	81.6
DeepGaitV2 [59]	Arxiv23	74.4	88	65.8		77.7	87.9	90.6	
GaitRef [223]	IJCB23	49	69.3	40.7	25.3	53	67.9	73	77.5
GaitCoTr [151]	ICASSP23	-	-	-	-	55.6	70.9	76.2	80.4
GPGait [41]	ICCV23	22.5	-	-	-	53.6	-	-	-
PAA [224]	ICCV23	38.9	59.1	-	-	38.7	62.1	-	-
HFSL [225]	ICCV23	61.3	76.3	55.5	34.8	62.7	76.6	81.3	85.2
VPNet [226]	CVPR24	75.4	87.1	-	-	80.0	89.4	-	-

From the analysis presented in Table V, several important observations can be made. Firstly, there is a significant drop in performance when deep gait methods are applied to outdoor datasets compared to indoor datasets [35]. Secondly, silhouette-based methods outperform those using estimated skeletons by a considerable margin [41]. Thirdly, deep models show better generalization performance on the GREW, despite GREW has more identities compared to Gait3D [59].

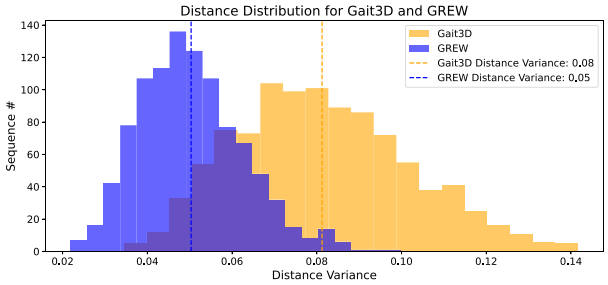


Fig. 7. Distance variance between each probe and gallery by randomly selecting 1,000 probes and comparing their distance variances with the gallery sequences. Our observations indicate that the GREW dataset exhibits less diversity in its sequences.

We consider the following reasons causing the abovementioned observations. Firstly, the performance downdegradation from indoor to outdoor datasets reflects that the challenges present in unconstrained environments, such as occlusion [192], diverse viewpoints [133], and poor resolution [60], having a detrimental effect on the performance of gait recognition models. Secondly, It raises significantly challenging to obtain ideal silhouettes and skeletons in the outdoor scenarios, highlighting the impact of realistic factors on the generation of robust gait representations. Thirdly, Gait3D suffers from spatial misalignment due to the collection of gait data from various vertical viewpoints and angles. Lastly but not at least, we observe that there is a noticed domain gap between GREW and Gait3D, where GREW has less viewpoint difference between its probe and gallery sequences. In another words, the probe and gallery samples within Gait3D typically has large viewpoint difference, while GREW has relatively limited viewpoint distribution as illustrated in Fig. 7. We conclude that the relatively limited diversity of viewpoints contributes to its better performance than Gait3D, because it is easier to concur the challenge of outdoor gait recognition with limited viewpoint changes.

3) Evaluation on Clothes-Changing Gait Recognition: FVG [27], CASIA-E [133], and CCPG [68] are three up-to-date benchmarks for clothes-changing gait recognition. Among them, FVG [27] is the smallest dataset consisting of 2,856 front-view walking sequences from 226 identities, and FVG also has a clothes-changing subset. CASIA-E [133] contains 1,014 people and 778,752 videos. Each participant involves varied appearances caused by changes in carrying and dressing. However, it only provides silhouettes and infrared thermal images. CCPG [68] provides 200 identities and over 16K sequences which are captured indoors and outdoors, while each identity has seven different cloth-changing statuses. Besides, both CCPG and FVG provide raw RGB images. As shown in Table VI, we observe that state-of-the-art methods can achieve over 70% recognition under clothes-changing settings in all three datasets, and deep models fed with RGB images as input can gain significant performance improvement. However, CCPG addressed that the performance improvement may come from utilizing gait-irrelevant information such as face and shoes. We suggest it is potential to study gait recognition from RGB images in an

TABLE VI
GAIT RECOGNITION PERFORMANCE ON THREE CLOTHES-CHANGING DATASETS, FVG [27], CASIA-E [133], AND CCPG [68]

FVG [27]							
		CL		ALL			
GEINet [20]	ICB16	6.5		13			
LBNet [21]	T-PAMI16	23.2		40.7			
GaitSetv1 [23]	CVPR19	56.8		81.2			
GaitSetv2 [24]	TPAMI21	70.4		91.9			
MMGaitFormer [163]	CVPR23	53.4		85.3			
CASIA-E [133]							
		NM		BG		CL	
GaitSet [23]	AAAI19	82.54		75.26		62.53	
GaitPart [28]	CVPR20	82.92		74.36		60.48	
GaitBase [35]	CVPR23	91.59		86.65		74.73	
CCPG [68]							
		RGB w/o face		RGB w/o face & foot		Silhouettes	
		R@1	mAP	R@1	mAP	R@1	mAP
AP3D [68]	ECCV20	86.7	60.1	55.1	27.3	-	-
BiCnet-TKS [68]	CVPR21	84.2	57.9	64.5	36.9	-	-
PSTA [68]	ICCV21	88.2	65.3	62.6	37.6	-	-
PiT [68]	TII22	85.1	60.1	57.1	30.8	-	-
GaitSet [23]	AAAI19	-	-	-	-	77.7	46.4
GaitPart [28]	CVPR20	-	-	-	-	77.8	45.5
GaitGL [30]	ICCV21	-	-	-	-	69.1	27

TABLE VII
GAIT RECOGNITION PERFORMANCE ON TWO LIDAR-BASED DATASETS, LICAMGAIT AND SUSTECH1K

LiCamGait [61]									
		2-8m				8-15m			
		NM	BG	CL	Mean	NM	BG	CL	Mean
PointNet [227]	CVPR17	32.3	29	44.8	35.4	29	19.4	34.4	27.6
PointMLP [228]	ICLR22	51.6	38.7	55.1	48.5	35.5	29	44.8	36.4
Han *et al.* [61]	Arxiv22	77.42	61.3	68.8	69.2	54.8	71	58.6	61.5
SUSTech1K [42]									
		NM	BG	CL	CR	UB	UF	OC	NT
PointNet [227]	CVPR17	43.6	37.3	25.7	28.8	19.9	30.1	44.3	27.4
PointNet++ [229]	NIPS17	55.9	52.2	41.6	49.6	47.8	45.9	54.2	52.5
SimpleView [230]	ICML21	72.3	68.8	57.2	63.3	49.2	62.5	79.7	66.5
PointGait [231]	IJCB23	68.6	65.2	48.1	56.8	35.6	55.0	68.8	61.7
LidarGait [42]	CVPR23	91.8	88.6	74.6	89	67.5	80.9	94.5	90.4

end-to-end manner, as RGB-based methods are efficient and contain more information.

4) Evaluation on 3D Gait Recognition: Recently, the integration of LiDAR sensors in gait recognition [42], [104] has gained attention to improve human perception in challenging lighting conditions and provide precise 3D information. Notably, datasets like SUSTech1K [42] and LiCamGait [61] have been introduced, offering a large-scale collection of sequences with various variations. Unlike silhouette-based methods that struggle with poor illumination, LiDAR-based gait recognition [42], [104] demonstrates promising performance in such conditions, as well as in scenarios with occlusion and other realistic challenges. As shown in Table VII, the 3D geometry information provided by LiDAR sensors allows for the learning of more informative and discriminative gait features, enhancing the practical applications of gait recognition.

The effectiveness of gait recognition systems depends on their ability to generalize to unseen data. Most studies have used the same dataset for both training and testing, which does not reflect real-world scenarios where test data often differs. To better improve generalizability, cross-domain evaluations are recommended. We recommend referring to BigGait [36] and GPGait [41] for further insights into this challenge.

VII. SECURITY AND PRIVACY OF GAIT RECOGNITION

The rapid development of gait recognition raises concerns that both the research community and society at large should address the potential effects. While the overview on biometrics by Jain et al. [232] provides a comprehensive description of security and privacy, it covers the broader field of biometrics and does not specifically focus on gait recognition. To bridge this gap, we will summarize the security and privacy challenges in biometrics in general and emphasize the specific concerns related to gait recognition.

A. Security

Like other biometrics systems, gait recognition should also be secure from various attacks. There are three kinds of attacks, according to the summary in [232], presentation attacks, adversarial attacks, and template attacks.

- *Presentation Attacks* refer to a type of attack where artificial objects are presented to the sensors of biometric systems. This form of attack is prevalent in face recognition systems, as highlighted in the FacePAD study [233], where face images displayed on devices or 3D silicon masks are used to deceive the system. In contrast, there have been limited studies on gait presentation attacks. The pioneering investigation on vision-based gait presentation attacks was conducted by Hadid et al. [234]. However, presentation attacks in gait recognition remain relatively unexplored, necessitating further research in this area.
- *Adversarial Attacks* are by digital synthetic data. Unlike presentation attacks, which rely on physical objects, adversarial attacks are based on generative models that can synthesize realistic data [159]. The advancement of generative models, such as stable diffusion models, has enabled the creation of high-quality and lifelike faces and human figures through extensive training on large datasets. While it is feasible to generate single frames that preserve identity with high fidelity, gait recognition systems require sequential data, necessitating both fidelity and temporal consistency. Further research is needed to explore these two critical aspects of fidelity and consistency in the context of gait recognition to enhance its robustness against adversarial attacks.
- *Template attacks* is to reconstruct images or videos from templates that are extracted by a biometrics system. Studies on face [235] have shown its feasibility. Unlike presentation and adversarial attacks, template attacks are manipulated at the feature level so that gait template attacks can borrow insights from other biometric features such as the face, fingerprint, or iris.

B. Privacy

Gait recognition systems, as their significant performance achieved in both indoor and outdoor scenarios, may pose greater privacy concerns compared to face recognition systems, which have already raised significant privacy issues globally. Gait can be captured from a greater distance than facial features. While it is natural to wear hats, sunglasses, or masks to protect one's face, concealing gait through completely different clothing is not as practical. In addition to identity, gait can also reveal information about gender and

Fig. 8. A shared video on the social media platform with many pedestrians [239].

health conditions, as demonstrated in studies [236]. These factors contribute to the increased privacy implications associated with gait recognition systems.

To protect data and privacy, various regulations and laws have been released worldwide. The European Parliament introduced the *General Data Protection Regulation (GDPR)* in 2016, and *Data Security Law of the P. R. of China* [203] went into effect in 2021. Similarly, Several states in the U.S. also passed similar laws, such as *California Consumer Privacy Act* [237] and *Biometric Information Privacy Act* of Illinois. In China, a national standard called *Information Security Technology—Personal Information Security Specification (GB/T 35273-2020)* [238] has been passed, providing detailed instructions for personal information security. Most laws and regulations concerning the privacy protection of biometric data share common principles. They typically outline the rights of individuals whose biometric data is collected, including the right 1) to know how the data will be used and stored, 2) to delete the data, and 3) to opt-out of the data usage. Meanwhile, it is worth noting that many laws primarily impose restrictions on the use of biometric data by private businesses, while allowing certain exceptions for government agencies for purposes such as public security. This aspect of the laws, which pertains to the usage of biometric data by governments, often lacks transparency and may raise concerns regarding the extent and potential misuse of such data for surveillance or other undisclosed purposes.

The privacy concerns on gait recognition may bring a crisis to video-sharing social platforms such as YouTube and TikTok. There are many kinds of videos, as in Fig. 8 online. If the walking pedestrians in the videos can be identified by their gait features, should we get a permit from them before posting? It is impossible to get permits from all pedestrians. Google Street has blurred all human faces on the street. Should the pedestrians in the videos be blurred also, or should those videos be deleted directly from the Internet? We leave these questions open for the community and society to discuss.

VIII. Challenges and Directions

Even though significant progress has been achieved in the past years, there are still many challenges in gait recognition. We think gait recognition can be improved from datasets, practicality, and trustworthiness. They are described in detail in the following part of this section.

A. Datasets

• *Learning from real scenarios.* Gait recognition using deep learning techniques has achieved promising results in indoor datasets captured in constrained scenarios. Although indoor laboratory datasets typically contain limited identities, these datasets help address many challenges of gait recognition, such as changes in clothes or viewpoints. Recently many outdoor datasets have been proposed to promote the study of gait recognition in unconstrained scenarios and extend the recognition research to realistic factors like poor illumination and low resolutions [42], [60]. In-the-wild datasets [42], [43], [60], [68] help develop a promising gait recognition system. However, collecting a large-scale gait dataset can be extremely difficult as gait is ambiguous to annotate over different viewpoints. The ambiguity of gait really prevents us from establishing a large-scale gait dataset with diverse viewpoints, clothing, and status for each pedestrian, and it is worth investigating potential solutions to extend gait recognition in real scenarios.

• *Learning from synthetic data.* Since it is challenging to collect gait data in real scenarios, one possible solution is to use synthetic methods to generate gait data. Synthetic data is prevalent in many studies, including face spoofing, self-driving [240], and 3D object classification. VersatileGait [121], a synthetic dataset for gait recognition, introduces an affordable and easy way to access diverse attributes and large-scale identities. However, defining an identity may be a problem. There are many parameters to define a 3D body and its motion. If we consider two persons to be two different identities, how to determine the threshold of their similarity?

• *Learning from unlabelled data.* Collecting a large gait dataset in real scenarios is difficult and potentially causes privacy concerns. Another possible solution is to use unlabelled data. We can collect many gait data from videos online and other sources [63]. Gait data can be automatically detected and segmented from videos. Although it is difficult to label the identities in those videos, deep models still have the potential to learn informative representations from unlabelled data via self-supervised learning methods. We value semi- and self-supervised learning methods to promote the study of gait recognition in the future greatly.

B. Gait Recognition Toward Reality

• *Gait-Changing Gait Recognition.* Several common factors can significantly impact gait and result in performance degradation in gait recognition systems. Camera viewpoints, variations in pedestrian dressing, and occlusion are examples of factors that can lead to significant differences in human appearance, making it challenging to match samples from the same subject successfully. To tackle this challenge, researchers have explored potential solutions such as eliminating gait-irrelevant information through the disentanglement method [27] or modeling gait motion information to be robust to appearance changes [125], [130]. However, it is important to note that gait motion itself can also vary due to factors

such as aging and disease. This aspect is often overlooked but presents valuable opportunities for further exploration.

• *End-to-end Gait Recognition.* The current focus of gait recognition has predominantly been on the recognition process itself, ignoring the importance of upstream tasks that provide gait representations. These upstream tasks include human foreground segmentation and pose estimation methods, which are crucial for obtaining silhouettes and skeletons. However, recent research has highlighted that using off-the-shelf segmentation methods can negatively impact representation learning. As a result, there is a growing trend toward learning deep gait representations in an end-to-end manner. This approach involves applying deep models directly to raw input data, such as in the works of Liang et al. [62], Song et al. [78], and Zhang et al. [27]. Besides, BigGait [36] leverages task-agnostic Large Vision Models (LVMs) to derive implicit gait representations efficiently. BigGait outperforms previous methods across various datasets, offering a promising direction for next-generation gait recognition. The objective is to fully leverage the whole-body information from the raw input, simplifying the overall pipeline of human identification.

• *Heterogeneous Gait Recognition.* Most gait recognition methods primarily rely on camera-based modalities, which restricts their effectiveness in poor illumination conditions, particularly during nighttime. However, recent advancements in the field have introduced alternative modalities, including infrared cameras and LiDAR sensors, to overcome the limitations of camera-based approaches. These new modalities not only offer a promising solution for low-light gait recognition but also present new challenges, such as multimodal fusion and cross-modal retrieval [241], which require further investigation in the coming years.

• *Efficient Gait Recognition.* The existing gait recognition methods have shown satisfactory performance on large-scale datasets [60]. However, they still fall short of meeting the requirements of practical surveillance systems, which demand real-time feature extraction and retrieval. The importance of efficiency in real-world applications is often overlooked [117]. In other fields of study, advancements have been made in lightweight models that offer comparable performance with reduced computational costs. Additionally, the existing literature fails to address the challenges posed by real-world applications that may involve galleries with millions or even billions of samples. The process of comparing a probe feature with such a large gallery can be extremely time-consuming. Therefore, there is a need to explore efficient gait recognition techniques that can meet the demands of practical applications.

• *Transfer learning for Gait Recognition.* Transfer learning in the context of gait recognition holds significant value and potential for various applications. It offers the advantage of training a model on a large dataset in one domain and then applying the learned knowledge to a different domain for evaluation. Two examples illustrate the benefits of transfer learning. Firstly, by training a deep model on a source domain like GREW [60], we can leverage the learned knowledge to achieve promising results on a targeted domain such as CASIA-B [15]. This transfer of knowledge helps in improving performance and generalization [205]. Additionally, transfer learning enables the utilization of learned knowledge from human gait recognition in animal recognition scenarios. By leveraging the insights gained from human gait recognition, we can potentially apply similar techniques and models to recognize and analyze the gaits of animals. Overall, transfer learning plays a vital role in gait recognition by enabling knowledge transfer across domains, leading to improved performance and the exploration of new application scenarios.

C. Trustworthy Gait Recognition

• *Gait Privacy Protection.* In addition to identifying individuals, gait recognition also raises privacy concerns regarding sensitive information such as race, gender, age, dress, and other attributes, similar to face recognition [242]. It is crucial to address these privacy concerns and investigate potential biases associated with these factors in gait recognition systems. Future research should focus on understanding and mitigating these privacy issues to ensure fairness, transparency, and responsible deployment of gait recognition technology.

• *Gait Encryption for Security Protection* As discussed in Section VII-B, the growing accuracy of gait recognition poses challenges for online video-sharing platforms. Blurring or masking all pedestrians in videos can significantly degrade the visual quality, discouraging users from sharing such content. To address this issue, one potential solution could be to encrypt gait videos using modifications that render them unrecognizable by gait recognition methods while preserving the visual appeal. However, research on this specific topic is currently lacking, despite the existence of encryption methods for face recognition [243]. Further exploration and investigation are needed to develop effective and practical encryption techniques for protecting privacy in gait recognition videos.

IX. Conclusion

This survey provided a comprehensive overview of deep gait recognition, covering its fundamental concepts, challenges, advancements, and potential future directions. The survey highlighted the various approaches and techniques, including traditional and deep learning-based approaches. It also discussed the importance of datasets, challenges related to privacy and security, and the emergence of alternative modalities. Furthermore, the survey reviewed the available deep gait recognition methods from feature learning and architecture perspectives following our proposed taxonomy, which provides a clear and thorough understanding of the big picture of the field. Overall, we hope this survey can serve as a valuable resource for researchers and practitioners in the field of gait recognition, shedding light on its advancements and potential for further development.

References

[1] 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.

[2] M. S. Nixon, J. N. Carter, D. Cunado, P. S. Huang, and S. Stevenage,"Automatic gait recognition," in *Biometrics*. Boston, MA, USA: Springer, 1996, pp. 231–249.

[3] M. S. Nixon, T. Tan, and R. Chellappa, *Human Identification Based on Gait*, vol. 4. New York, NY, USA: Springer, 2010.

[4] T. Singhal, "A review of coronavirus disease-2019 (COVID-19)," *Indian J. Pediatr.*, vol. 87, no. 4, pp. 281–286, 2020.

[5] Niyogi and Adelson, "Analyzing and recognizing walking figures in XYT," in *Proc. CVPR*, 1994, pp. 469–474.

[6] J. Little and J. Boyd, "Recognizing people by their gait: The shape of motion," *Videre, J. Comput. Vis. Res.*, vol. 1, no. 2, pp. 1–32, 1998.

[7] P. Phillips, "Human identification technical challenges," in *Proc. ICIP*, vol. 1, 2002, p. 1.

[8] P. J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. Bowyer, "Baseline results for the challenge problem of HumanID using gait analysis," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2002, pp. 137–142.

[9] 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.

[10] C. Wang, J. Zhang, J. Pu, X. Yuan, and L. Wang, "Chrono-gait image: A novel temporal template for gait recognition," in *Proc. ECCV*, 2010, pp. 257–270.

[11] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[12] 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.

[13] 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.

[14] D. K. Wagg and M. S. Nixon, "Automated markerless extraction of walking people using deformable contour models," *Comput. Animat. Virtual Worlds*, vol. 15, nos. 3–4, pp. 399–406, 2004.

[15] 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. ICPR*, vol. 4, 2006, pp. 441–444.

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

[17] D. Tan, K. Huang, S. Yu, and T. Tan, "Efficient night gait recognition based on template matching," in *Proc. ICPR*, vol. 3, 2006, pp. 1000–1003.

[18] A. Kale, A. Rajagopalan, N. Cuntoor, and V. Kruger, "Gait-based recognition of humans using continuous HMMs," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2002, pp. 336–341.

[19] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3D convolutional neural networks," in *Proc. ICIP*, 2016, pp. 4165–4169.

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

[21] 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.

[22] 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.

[23] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in *Proc. AAAI*, 2019, pp. 8126–8133.

[24] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng, "GaitSet: Cross-view gait recognition through utilizing gait as a deep set," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3467–3478, Jul. 2022.

[25] Y. Wang et al., "EV-Gait: Event-based robust gait recognition using dynamic vision sensors," in *Proc. CVPR*, 2019, pp. 6358–6367.

[26] K. Zhang, W. Luo, L. Ma, W. Liu, and H. Li, "Learning joint gait representation via quintuplet loss minimization," in *Proc. CVPR*, 2019, pp. 4700–4709.

[27] Z. Zhang, L. Tran, F. Liu, and X. Liu, "On learning disentangled representations for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 345–360, Jan. 2022.

[28] C. Fan et al., "GaitPart: Temporal part-based model for gait recognition," in *Proc. CVPR*, 2020, pp. 14213–14221.

[29] S. Hou, C. Cao, X. Liu, and Y. Huang, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 382–398.

[30] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proc. ICCV*, 2021, pp. 14648–14656.

[31] Z. Huang et al., "3D local convolutional neural networks for gait recognition," in *Proc. ICCV*, 2021, pp. 14920–14929.

[32] X. Huang et al., "Context-sensitive temporal feature learning for gait recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 12909–12918.

[33] C. Shen, B. Lin, S. Zhang, X. Yu, G. Q. Huang, and S. Yu, "Gait recognition with mask-based regularization," in *Proc. IJCB*, 2023, pp. 1–10.

[34] H. Dou, P. Zhang, W. Su, Y. Yu, Y. Lin, and X. Li, "GaitGCI: Generative counterfactual intervention for gait recognition," in *Proc. CVPR*, 2023, pp. 5578–5588.

[35] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang, and S. Yu, "OpenGait: Revisiting gait recognition towards better practicality," in *Proc. CVPR*, 2023, pp. 9707–9716.

[36] D. Ye, C. Fan, J. Ma, X. Liu, and S. Yu, "BigGait: Learning gait representation you want by large vision models," in *Proc. CVPR*, 2024, pp. 200–210.

[37] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107069.

[38] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang, "Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations," in *Proc. CCBR*, 2017, pp. 474–483.

[39] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren, "End-to-end model-based gait recognition," in *Proc. ACCV*, 2020, pp. 3–20.

[40] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "GaitGraph: Graph convolutional network for skeleton-based gait recognition," in *Proc. ICIP*, 2021, pp. 2314–2318.

[41] Y. Fu, S. Meng, S. Hou, X. Hu, and Y. Huang, "GPGait: Generalized pose-based gait recognition," in *Proc. ICCV*, 2023, pp. 19595–19604.

[42] C. Shen, C. Fan, W. Wu, R. Wang, G. Q. Huang, and S. Yu, "LidarGait: Benchmarking 3D gait recognition with point clouds," in *Proc. CVPR*, 2023, pp. 1054–1063.

[43] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei, "Gait recognition in the wild with dense 3D representations and a benchmark," in *Proc. CVPR*, 2022, pp. 20228–20237.

[44] Y. Feng, Y. Li, and J. Luo, "Learning effective gait features using LSTM," in *Proc. ICPR*, 2016, pp. 325–330.

[45] J. Zheng, X. Liu, S. Wang, L. Wang, C. Yan, and W. Liu, "Parsing is all you need for accurate gait recognition in the wild," in *Proc. ACM Multimedia*, 2023, pp. 116–124.

[46] F. M. Castro, R. Delgado-Escaño, R. Hernández-García, M. J. Marín-Jiménez, and N. Guil, "AttenGait: Gait recognition with attention and rich modalities," *Pattern Recognit.*, vol. 148, Apr. 2024, Art. no. 110171.

[47] G. Li, L. Guo, R. Zhang, J. Qian, and S. Gao, "Transgait: Multimodal-based gait recognition with set transformer," *Appl. Intell.*, vol. 53, no. 2, pp. 1535–1547, 2023.

[48] J. Chen, Z. Wang, C. Zheng, K. Zeng, Q. Zou, and L. Cui, "GaitAMR: Cross-view gait recognition via aggregated multi-feature representation," *Inf. Sci.*, vol. 636, Jul. 2023, Art. no. 118920.

[49] Z. Wang et al., "QAGait: Revisit gait recognition from a quality perspective," in *Proc. AAAI*, vol. 38, 2024, pp. 5785–5793.

[50] H. Pan, Y. Chen, T. Xu, Y. He, and Z. He, "Toward complete-view and high-level pose-based gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 2104–2118, 2023.

[51] A. Catruna, A. Cosma, and E. Radoi, "GaitPT: Skeletons are all you need for gait recognition," in *Proc. IEEE 18th Int. Conf. Autom. Face Gesture Recognit.*, 2024, pp. 1–10.

[52] A. Gupta and R. Chellappa, "You can run but not hide: Improving gait recognition with intrinsic occlusion type awareness," in *Proc. WACV*, 2024, pp. 5893–5902.

[53] Z. Wang et al., "LandmarkGait: Intrinsic human parsing for gait recognition," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 2305–2314.

[54] H. Dou, P. Zhang, Y. Zhao, L. Jin, and X. Li, "CLASH: Complementary learning with neural architecture search for gait recognition," *IEEE Trans. Image Process.*, early access, Feb. 16, 2024, doi: 10.1109/TIP.2024.3360870.

[55] G. Habib, N. Barzilay, O. Shimshi, R. Ben-Ari, and N. Darshan, "Watch where you head: A view-biased domain gap in gait recognition and unsupervised adaptation," in *Proc. WACV*, 2024, pp. 6109–6119.

[56] Z. Lv, X. Xing, K. Wang, and D. Guan, "Class energy image analysis for video sensor-based gait recognition: A review," *Sensors*, vol. 15, no. 1, pp. 932–964, 2015.

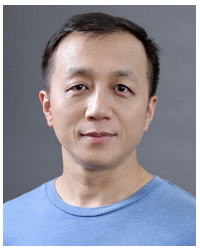
[57] M. Wang et al., "DyGait: Exploiting dynamic representations for high-performance gait recognition," in *Proc. ICCV*, 2023, pp. 1–10.
[58] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 4, pp. 1–14, 2018.
[59] C. Fan, S. Hou, Y. Huang, and S. Yu, "Exploring deep models for practical gait recognition," 2023, *arXiv:2303.03301*.
[60] Z. Zhu et al., "Gait recognition in the wild: A benchmark," in *Proc. ICCV*, 2021, pp. 14789–14799.
[61] X. Han, P. Cong, L. Xu, J. Wang, J. Yu, and Y. Ma, "LiCamGait: Gait recognition in the wild by using LiDAR and camera multi-modal visual sensors," 2022, *arXiv.2211.12371*.
[62] J. Liang, C. Fan, S. Hou, C. Shen, Y. Huang, and S. Yu, "Gaitedge: Beyond plain end-to-end gait recognition for better practicality," in *Proc. ECCV*, 2022, pp. 375–390.
[63] C. Fan, S. Hou, J. Wang, Y. Huang, and S. Yu, "Learning gait representation from massive unlabelled walking videos: A benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14920–14937, Dec. 2023.
[64] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 264–284, Jan. 2023.
[65] J. Wang et al., "Free lunch for gait recognition: A novel relation descriptor," in *Proc. ECCV*, 2023, pp. 39–56.
[66] J. Wang et al., "Causal intervention for sparse-view gait recognition," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 77–85.
[67] C. F. G. dos Santos et al., "Gait recognition based on deep learning: A survey," *ACM Comput. Surv.*, vol. 55, no. 2, pp. 1–34, 2022.
[68] W. Li et al., "An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions," in *Proc. CVPR*, 2023, pp. 13824–13833.
[69] Y. Makihara, M. S. Nixon, and Y. Yagi, "Gait recognition: Databases, representations, and applications," in *Computer Vision: A Reference Guide*. Cham, Switzerland: Springer, 2020, pp. 1–13.
[70] I. Rida, N. Almaadeed, and S. Almaadeed, "Robust gait recognition: A comprehensive survey," *IET Biom.*, vol. 8, no. 1, pp. 14–28, 2018.
[71] R. Liao, Z. Li, S. S. Bhattacharyya, and G. York, "*PoseMapGait*: A model-based gait recognition method with pose estimation maps and graph convolutional networks," *Neurocomputing*, vol. 501, pp. 514–528, Aug. 2022.
[72] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*. New York, NY, USA: Springer, 2007.
[73] S. Hou, C. Fan, C. Cao, X. Liu, and Y. Huang, "A comprehensive study on the evaluation of silhouette-based gait recognition," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 5, no. 2, pp. 196–208, Apr. 2023.
[74] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in *Proc. ICLR*, 2019, pp. 1–22.
[75] M. Piccardi, "Background subtraction techniques: A review," in *Proc. SMC*, vol. 4, 2004, pp. 3099–3104.
[76] G. V. Veres, M. S. Nixon, and J. N. Carter, "Model-based approaches for predicting gait changes over time," in *Proc. ISSNIP*, 2005, pp. 325–330.
[77] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. CVPR*, 2019, pp. 5693–5703.
[78] C. Song, Y. Huang, Y. Huang, N. Jia, and L. Wang, "GaitNet: An end-to-end network for gait based human identification," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106988.
[79] Z. Liu, L. Malave, and S. Sarkar, "Studies on silhouette quality and gait recognition," in *Proc. CVPR*, vol. 2, 2004, pp. 1–8.
[80] R. T. Collins, R. Gross, and J. Shi, "Silhouette-based human identification from body shape and gait," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2002, pp. 366–371.
[81] B. Lin, S. Zhang, and F. Bao, "Gait recognition with multiple-temporal-scale 3D convolutional neural network," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3054–3062.
[82] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999.
[83] R. Begg and J. Kamruzzaman, "A machine learning approach for automated recognition of movement patterns using basic, kinetic and kinematic gait data," *J. Biomech.*, vol. 38, no. 3, pp. 401–408, 2005.
[84] H. Lu, K. Plataniotis, and A. Venetsanopoulos, "Boosting LDA with regularization on MPCA features for gait recognition," in *Proc. Biom. Symp.*, 2007, pp. 1–6.
[85] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in *Proc. 3rd Int. Conf. Imag. Crime Detect. Prevent.*, 2009, pp. 1–6.
[86] Q. Ma, S. Wang, D. Nie, and J. Qiu, "Recognizing humans based on gait moment image," in *Proc. SNPD*, vol. 2, 2007, pp. 606–610.
[87] 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, May 2017.
[88] F. M. Castro, M. J. Marín-Jiménez, N. Guil, S. López-Tapia, and N. P. de la Blanca, "Evaluation of CNN architectures for gait recognition based on optical flow maps," in *Proc. BIOSIG*, 2017, pp. 1–5.
[89] A. Dosovitskiy et al., "FlowNet: Learning optical flow with convolutional networks," in *Proc. ICCV*, 2015, pp. 2758–2766.
[90] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. CVPR*, 2017, pp. 1647–1655.
[91] T. Hui, X. Tang, and C. C. Loy, "LiteFlowNet: A lightweight convolutional neural network for optical flow estimation," in *Proc. CVPR*, 2018, pp. 8981–8989.
[92] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. CVPR*, 2014, pp. 1653–1660.
[93] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
[94] A. Cosma and E. Rădoi, "GaitMorph: Transforming gait by optimally transporting discrete codes," in *Proc. IJCB*, 2023, pp. 1–11.
[95] W. An et al., "Performance evaluation of model-based gait on multi-view very large population database with pose sequences," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 4, pp. 421–430, Oct. 2020.
[96] C. Fan, J. Ma, D. Jin, C. Shen, and S. Yu, "SkeletonGait: Gait recognition using skeleton maps," in *Proc. AAAI*, 2024, pp. 1662–1669.
[97] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2Mesh: Generating 3D mesh models from single RGB images," in *Proc. ECCV*, 2018, pp. 52–67.
[98] Y. Dong et al., "HybridGait: A benchmark for spatial-temporal cloth-changing gait recognition with hybrid explorations," in *Proc. AAAI*, 2024, pp. 1600–1608.
[99] N. Kolotouros, G. Pavlakos, D. Jayaraman, and K. Daniilidis, "Probabilistic modeling for human mesh recovery," in *Proc. ICCV*, 2021, pp. 11605–11614.
[100] G. Georgakis, R. Li, S. Karanam, T. Chen, J. Košecká, and Z. Wu, "Hierarchical kinematic human mesh recovery," in *Proc. ECCV*, 2020, pp. 768–784.
[101] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, Feb. 2012.
[102] 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 Represent.*, vol. 25, no. 1, pp. 195–206, 2014.
[103] J. F. Nunes, P. M. Moreira, and J. M. R. Tavares, "Benchmark RGB-D gait datasets: A systematic review," in *Proc. VipIMAGE*, 2019, pp. 366–372.
[104] J. Ahn, K. Nakashima, K. Yoshino, Y. Iwashita, and R. Kurazume, "2V-Gait: Gait recognition using 3D LiDAR robust to changes in walking direction and measurement distance," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, 2022, pp. 602–607.
[105] A. F. Bobick and A. Y. Johnson, "Gait recognition using static, activity-specific parameters," in *Proc. CVPR*, 2001, pp. I–I.
[106] A. Y. Johnson and A. F. Bobick, "A multi-view method for gait recognition using static body parameters," in *Proc. AVBPA*, 2001, pp. 301–311.
[107] N. V. Boulgouris and Z. X. Chi, "Human gait recognition based on matching of body components," *Pattern Recognit.*, vol. 40, no. 6, pp. 1763–1770, 2007.
[108] D. Cunado, M. S. Nixon, and J. N. Carter, "Using gait as a biometric, via phase-weighted magnitude spectra," in *Proc. AVBPA*, 1997, pp. 93–102.
[109] D. Cunado, J. M. Nash, M. S. Nixon, and J. N. Carter, "Gait extraction and description by evidence-gathering," in *Proc. AVBPA*, 1999.
[110] 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.

[111] Y. Zhang, Y. Huang, S. Yu, and L. Wang, "Cross-view gait recognition by discriminative feature learning," *IEEE Trans. Image. Process*, vol. 29, pp. 1001–1015, 2019.
[112] M. H. Khan, F. Li, M. S. Farid, and M. Grzegorzek, "Gait recognition using motion trajectory analysis," in *Proc. CORES*, 2017, pp. 73–82.
[113] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," in *Shape, Contour Grouping Computer Visual*. Berlin, Germany: Springer, 1999, pp. 319–345.
[114] 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.
[115] S. Zhang, Y. Wang, T. Chai, A. Li, and A. K. Jain, "RealGait: Gait recognition for person re-identification," 2022, *arXiv:2201.04806*.
[116] Z. Zhang et al., "Gait recognition via disentangled representation learning," in *Proc. CVPR*, 2019, pp. 4710–4719.
[117] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. CVPR*, 2016, pp. 779–788.
[118] S. Hou, X. Liu, C. Cao, and Y. Huang, "Gait quality aware network: Toward the interpretability of silhouette-based gait recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8978–8988, Nov. 2023.
[119] T. Chai, X. Mei, A. Li, and Y. Wang, "Silhouette-based view-embeddings for gait recognition under multiple views," in *Proc. ICIP*, 2021, pp. 2319–2323.
[120] H. Wu, J. Tian, Y. Fu, B. Li, and X. Li, "Condition-aware comparison scheme for gait recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 2734–2744, 2020.
[121] P. Zhang et al., "A large-scale synthetic gait dataset towards in-the-wild simulation and comparison study," *Trans. Multimedia Comput. Commun. Appli.*, vol. 19, no. 1, pp. 1–23, 2023.
[122] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *J. Mach. Learn. Res.*, vol. 11, no. 36, pp. 1109–1135, 2010.
[123] T. Chai, A. Li, S. Zhang, Z. Li, and Y. Wang, "Lagrange motion analysis and view embeddings for improved gait recognition," in *Proc. CVPR*, 2022, pp. 20249–20258.
[124] J. Zheng et al., "Gait recognition in the wild with multi-hop temporal switch," in *Proc. ACM Multimedia*, 2022, pp. 6136–6145.
[125] H. Dou, P. Zhang, W. Su, Y. Yu, and X. Li, "Metagait: Learning to learn an omni sample adaptive representation for gait recognition," in *Proc. ECCV*, 2022, pp. 357–374.
[126] Y. Peng, K. Ma, Y. Zhang, and Z. He, "Learning rich features for gait recognition by integrating skeletons and silhouettes," *Multimedia Tools Appl.*, vol. 83, no. 3, pp. 7273–7294, 2024.
[127] P. Chattopadhyay, S. Sural, and J. Mukherjee, "Frontal gait recognition from occluded scenes," *Pattern Recognit. Lett.*, vol. 63, pp. 9–15, Oct. 2015.
[128] C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian, "Frame difference energy image for gait recognition with incomplete silhouettes," *Pattern Recognit. Lett.*, vol. 30, no. 11, pp. 977–984, 2009.
[129] 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, Sep. 2019.
[130] K. Ma, Y. Fu, D. Zheng, C. Cao, X. Hu, and Y. Huang, "Dynamic aggregated network for gait recognition," in *Proc. CVPR*, 2023, pp. 22076–22085.
[131] M. A. Goodale and A. D. Milner, "Separate visual pathways for perception and action," *Trends Neurosci.*, vol. 15, no. 1, pp. 20–25, 1992.
[132] K. Jun, D.-W. Lee, K. Lee, S. Lee, and M. S. Kim, "Feature extraction using an RNN autoencoder for skeleton-based abnormal gait recognition," *IEEE Access*, vol. 8, pp. 19196–19207, 2020.
[133] C. Song, Y. Huang, W. Wang, and L. Wang, "CASIA-E: A large comprehensive dataset for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2801–2815, Mar. 2023.
[134] E. Hossain and G. Chetty, "Multimodal feature learning for gait biometric based human identity recognition," in *Proc. Int. Conf. Neural Inf. Process.*, 2013, pp. 721–728.
[135] W. An, R. Liao, S. Yu, Y. Huang, and P. C. Yuen, "Improving gait recognition with 3D pose estimation," in *Proc. CCBR*, 2018, pp. 137–147.
[136] F. Battistone and A. Petrosino, "TGLSTM: A time based graph deep learning approach to gait recognition," *Pattern Recognit. Lett.*, vol. 126, pp. 132–138, Sep. 2019.
[137] Z. Liu, L. Malave, A. Osuntogun, P. Sudhakar, and S. Sarkar, "Toward understanding the limits of gait recognition," in *Proc. SPIE Biom. Technol. Human Identif.*, 2004, pp. 195–205.
[138] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. ICCV*, 2017, pp. 3980–3989.
[139] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. CVPR*, 2018, pp. 2285–2294.
[140] A. Nandy, R. Chakraborty, and P. Chakraborty, "Cloth invariant gait recognition using pooled segmented statistical features," *Neurocomputing*, vol. 191, pp. 117–140, May 2016.
[141] H. Murase and R. Sakai, "Moving object recognition in eigenspace representation: Gait analysis and lip reading," *Pattern Recognit. Lett.*, vol. 17, no. 2, pp. 155–162, 1996.
[142] Y. Fu et al., "Horizontal pyramid matching for person re-identification," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, pp. 8295–8302.
[143] C. Mao, Y. Li, Y. Zhang, Z. Zhang, and X. Li, "Multi-channel pyramid person matching network for person re-identification," in *Proc. AAAI*, 2018, pp. 7243–7250.
[144] D. Liu, M. Ye, X. Li, F. Zhang, and L. Lin, "Memory-based gait recognition," in *Proc. BMVC*, 2016, pp. 1–12.
[145] Y. Sun, H. Long, X. Feng, and M. Nixon, "GaitASMS: Gait recognition by adaptive structured spatial representation and multi-scale temporal aggregation," *Neural Comput. Appl.*, vol. 36, no. 13, pp. 7057–7069, 2024.
[146] X. Wang and W. Q. Yan, "Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory," *Int. J. Neural Syst.*, vol. 30, no. 1, 2020, Art. no. 1950027.
[147] L. Yao, W. Kusakunniran, Q. Wu, J. Zhang, and Z. Tang, "Robust CNN-based gait verification and identification using skeleton gait energy image," in *Proc. DICTA*, 2018, pp. 1–7.
[148] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," in *Proc. ICML*, 2010, pp. 495–502.
[149] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. ICCV*, 2017, pp. 5533–5541.
[150] J. Li, J. Gao, Y. Zhang, H. Shan, and J. Zhang, "Motion matters: A novel motion modeling for cross-view gait feature learning," in *Proc. ICASSP*, 2023, pp. 1–5.
[151] J. Li, Y. Zhang, H. Shan, and J. Zhang, "Gaitcotr: Improved spatial-temporal representation for gait recognition with a hybrid convolution-transformer framework," in *Proc. ICASSP*, 2023, pp. 1–5.
[152] S. Zou, J. Xiong, C. Fan, C. Shen, S. Yu, and J. Tang, "A multi-stage adaptive feature fusion neural network for multimodal gait recognition," *IEEE Trans. Biometrics, Behav., Identity Sci.*, early access, Apr. 3, 2024, doi: 10.1109/TBIOM.2024.3384704.
[153] Y. Sun, X. Feng, L. Ma, L. Hu, and M. Nixon, "TriGait: Aligning and fusing skeleton and silhouette gait data via a tri-branch network," in *Proc. IJCB*, 2023, pp. 1–9.
[154] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 568–576.
[155] R. Delgado-Escano, F. M. Castro, N. Guil, V. Kalogeiton, and M. J. Marin-Jimenez, "Multimodal gait recognition under missing modalities," in *Proc. ICIP*, 2021, pp. 3003–3007.
[156] M. J. Marín-Jiménez, F. M. Castro, R. Delgado-Escaño, V. Kalogeiton, and N. Guil, "UGaitNet: Multimodal gait recognition with missing input modalities," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 5452–5462, 2021.
[157] L. Wang, R. Han, and W. Feng, "Combining the silhouette and skeleton data for gait recognition," in *Proc. ICASSP*, 2023, pp. 1–5.
[158] N. Cai et al., "Hybrid silhouette-skeleton body representation for gait recognition," in *Proc. IHMSC*, 2021, pp. 216–220.
[159] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh, "GaitGAN: Invariant gait feature extraction using generative adversarial networks," in *Proc. CVPR Workshops*, 2017, pp. 30–37.
[160] R. Liao, W. An, S. Yu, Z. Li, and Y. Huang, "Dense-view GEIs set: View space covering for gait recognition based on dense-view GAN," in *Proc. IJCB*, 2020, pp. 1–9.
[161] 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. CVPR*, 2020, pp. 13306–13316.
[162] M. Babaee, L. Li, and G. Rigoll, "Person identification from partial gait cycle using fully convolutional neural networks," *Neurocomputing*, vol. 338, pp. 116–125, Apr. 2019.
[163] Y. Cui and Y. Kang, "Multi-modal gait recognition via effective spatial-temporal feature fusion," in *Proc. CVPR*, 2023, pp. 17949–17957.

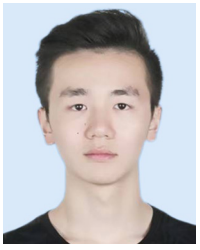
[164] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Gait recognition from a single image using a phase-aware gait cycle reconstruction network," in *Proc. ECCV*, 2020, pp. 386–403.
[165] D. Ye, J. Ma, C. Fan, and S. Yu, "GaitEditer: Attribute editing for gait representation learning," 2023, *arXiv:2303.05076*.
[166] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Cross-view gait recognition using pairwise spatial transformer networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 260–274, Jan. 2021.
[167] Z. Zhou, J. Liang, Z. Peng, C. Fan, F. An, and S. Yu, "Gait patterns as biomarkers: A video-based approach for classifying scoliosis," in *Proc. MICCAI*, 2024, pp. 284–294.
[168] S. Zou, C. Fan, J. Xiong, C. Shen, S. Yu, and J. Tang, "Cross-covariate gait recognition: A benchmark," in *Proc. AAAI*, 2024, pp. 7855–7863.
[169] A. Li, S. Hou, C. Wang, Q. Cai, and Y. Huang, "AerialGait: Bridging aerial and ground views for gait recognition," in *Proc. ACM MM*, 2024, pp. 1139–1147.
[170] A. Li, S. Hou, Q. Cai, Y. Fu, and Y. Huang, "Gait recognition with drones: A benchmark," *IEEE Trans. Multimedia*, vol. 26, pp. 3530–3540, 2024.
[171] Z. Mu, F. M. Castro, M. J. Marín-Jiménez, N. Guil, Y.-R. Li, and S. Yu, "ReSGait: The real-scene gait dataset," in *Proc. IJCB*, 2021, pp. 1–8.
[172] S. Yu, Q. Wang, and Y. Huang, "A large RGB-D gait dataset and the baseline algorithm," in *Proc. CCBR*, 2013, pp. 417–424.
[173] Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi, "Joint intensity and spatial metric learning for robust gait recognition," in *Proc. CVPR*, 2017, pp. 6786–6796.
[174] 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.
[175] M. Z. Uddin et al., "The OU-ISIR large population gait database with real-life carried object and its performance evaluation," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 1, p. 5, 2018.
[176] A. Mansur, Y. Makihara, R. Aqmar, and Y. Yagi, "Gait recognition under speed transition," in *Proc. CVPR*, 2014, pp. 2521–2528.
[177] 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 Security*, vol. 7, pp. 1511–1521, 2012.
[178] Y. Makihara et al., "The OU-ISIR gait database comprising the treadmill dataset," *IPSJ Trans. Comput. Vis. Appl.*, vol. 4, pp. 53–62, Apr. 2012.
[179] R. Delgado-Escano, F. M. Castro, J. R. Cózar, M. J. Marin-Jimenez, and N. Guil, "MuPeG—The multiple person gait framework," *Sensors*, vol. 20, no. 5, p. 1358, 2020.
[180] H. Yamada, J. Ahn, O. M. Mozos, Y. Iwashita, and R. Kurazume, "Gait-based person identification using 3D LiDAR and long short-term memory deep networks," *Adv. Robot.*, vol. 34, no. 18, pp. 1201–1211, 2020.
[181] B. Kwolek, A. Michalczuk, T. Krzeszowski, A. Switonski, H. Josinski, and K. Wojciechowski, "Calibrated and synchronized multi-view video and motion capture dataset for evaluation of gait recognition," *Multimedia Tools Appl.*, vol. 78, no. 22, pp. 32437–32465, 2019.
[182] Y. Wang, J. Sun, J. Li, and D. Zhao, "Gait recognition based on 3D skeleton joints captured by Kinect," in *Proc. ICIP*, 2016, pp. 3151–3155.
[183] C. Benedek, B. Gálai, B. Nagy, and Z. Jankó, "Lidar-based gait analysis and activity recognition in a 4D surveillance system," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 1, pp. 101–113, Jan. 2018.
[184] K. Galińska, P. Luboch, K. Kluwak, and M. Biegański, "A database of elementary human movements collected with RGB-D type camera," in *Proc. CogInfoCom*, 2015, pp. 575–580.
[185] V. O. Andersson and R. M. de Araújo, "Person identification using anthropometric and gait data from Kinect sensor," in *Proc. AAAI*, 2015, pp. 425–431.
[186] Y. Iwashita, K. Ogawara, and R. Kurazume, "Identification of people walking along curved trajectories," *Pattern Recognit. Lett.*, vol. 48, pp. 60–69, Oct. 2014.
[187] Y. Iwashita, R. Kurazume, and A. Stoica, "Gait identification using invisible shadows: Robustness to appearance changes," in *Proc. 5th Int. Conf. Emerg. Security Technol.*, 2014, pp. 34–39.
[188] Y. Iwashita, R. Baba, K. Ogawara, and R. Kurazume, "Person identification from spatio-temporal 3D gait," in *Proc. Int. Conf. Emerg. Security Technol.*, 2010, pp. 30–35.
[189] D. López-Fernández, F. J. Madrid-Cuevas, Á. Carmona-Poyato, M. J. Marín-Jiménez, and R. Muñoz-Salinas, "The AVA multi-view dataset for gait recognition," in *Proc. AMMDS*, 2014, pp. 26–39.
[190] B. DeCann, A. Ross, and J. Dawson, "Investigating gait recognition in the short-wave infrared (SWIR) spectrum: Dataset and challenges," in *Proc. SPIE*, 2013, Art. no. 87120J.
[191] A. I. Mahyuddin, S. Mihradi, T. Dirgantara, M. Moeliono, and T. Prabowo, "Development of Indonesian gait database using 2D optical motion analyzer system," *ASEAN Eng. J.*, vol. 2, no. 2, pp. 62–72, 2012.
[192] M. Hofmann, S. Sural, and G. Rigoll, "Gait recognition in the presence of occlusion: A new dataset and baseline algorithms," in *Proc. WSCG*, 2011, pp. 1–6.
[193] R. Borràs, À. Lapedriza, and L. Igual, "Depth information in human gait analysis: An experimental study on gender recognition," in *Proc. ICIAR*, 2012, pp. 98–105.
[194] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with RGB-D sensors," in *Proc. ECCV*, 2012, pp. 433–442.
[195] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes, "Gait energy volumes and frontal gait recognition using depth images," in *Proc. IJCB*, 2011, pp. 1–6.
[196] S. Samangooei, J. Bustard, M. S. Nixon, and J. N. Carter, "On acquisition and analysis of a dataset comprising of gait, ear and semantic data," in *Multibiometrics Human Identification*. Cambridge, U.K.: Cambridge Univ., 2011, pp. 277–301.
[197] D. S. Matovski, M. S. Nixon, S. Mahmoodi, and J. N. Carter, "The effect of time on the performance of gait biometrics," in *Proc. BTAS*, 2010, pp. 1–6.
[198] M. R. Aqmar, K. Shinoda, and S. Furui, "Robust gait recognition against speed variation," in *Proc. ICPR*, 2010, pp. 2190–2193.
[199] S. Zheng, K. Huang, T. Tan, and D. Tao, "A cascade fusion scheme for gait and cumulative foot pressure image recognition," *Pattern Recgonit.*, vol. 45, no. 10, pp. 3603–3610, 2012.
[200] "Human identification at a distance, UMD database." 2001. [Online]. Available: http://www.umiacs.umd.edu/labs/pirl/hid/data.html
[201] L. Lee and W. E. L. Grimson, "Gait analysis for recognition and classification," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2002, pp. 155–162.
[202] (Eur. Parliament Council, Strasbourg, France). *General Data Protection Regulation (GDPR)*, 2016. [Online]. Available: https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en
[203] "The data security law of the people's republic of China," presented at the 29th Meeting Stand. Comm. 13th Nat. People's Congr., 2021. [Online]. Available: https://www.gov.cn/xinwen/2021-06/11/content_5616919.htm
[204] L. Yao, W. Kusakunniran, P. Zhang, Q. Wu, and J. Zhang, "Improving disentangled representation learning for gait recognition using group supervision," *IEEE Trans. Multimedia*, vol. 25, pp. 4187–4198, Oct. 2023.
[205] Y. Liu, Y. Zeng, J. Pu, H. Shan, P. He, and J. Zhang, "SelfGait: A spatiotemporal representation learning method for self-supervised gait recognition," in *Proc. ICASSP*, 2021, pp. 2570–2574.
[206] B. Lin, Y. Liu, and S. Zhang, "Gaitmask: Mask-based model for gait recognition," in *Proc. BMVC*, 2021, pp. 1–12.
[207] S. Zhang, Y. Wang, and A. Li, "Cross-view gait recognition with deep universal linear embeddings," in *Proc. CVPR*, 2021, pp. 9095–9104.
[208] A. Sepas-Moghaddam, S. Ghorbani, N. F. Troje, and A. Etemad, "Gait recognition using multi-scale partial representation transformation with capsules," in *Proc. ICPR*, 2021, pp. 8045–8052.
[209] S. Hou, X. Liu, C. Cao, and Y. Huang, "Set residual network for silhouette-based gait recognition," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 3, pp. 384–393, Jul. 2021.
[210] X. Chen, X. Luo, J. Weng, W. Luo, H. Li, and Q. Tian, "Multi-view gait image generation for cross-view gait recognition," *Trans. Image Process.*, vol. 30, pp. 3041–3055, 2021.
[211] H. Dou, P. Zhang, Y. Zhao, L. Dong, Z. Qin, and X. Li, "GaitMPL: Gait recognition with memory-augmented progressive learning," *IEEE Trans. Image Process.*, vol. 33, pp. 1464–1475, 2024.
[212] H. Li et al., "GaitSlice: A gait recognition model based on spatio-temporal slice features," *Pattern Recognit.*, vol. 124, Apr. 2022, Art. no. 108453.
[213] X. Huang, X. Wang, B. He, S. He, W. Liu, and B. Feng, "Star: Spatio-temporal augmented relation network for gait recognition," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 5, no. 1, pp. 115–125, Jan. 2023.
[214] M. Wang et al., "GaitStrip: Gait recognition via effective strip-based feature representations and multi-level framework," in *Proc. ACCV*, 2022, pp. 536–551.

[215] J. Chen, Z. Wang, P. Yi, K. Zeng, Z. He, and Q. Zou, "Gait pyramid attention network: Toward silhouette semantic relation learning for gait recognition," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 4, no. 4, pp. 582–595, Oct. 2022.
[216] T. Huang, X. Ben, C. Gong, B. Zhang, R. Yan, and Q. Wu, "Enhanced spatial-temporal salience for cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6967–6980, Oct. 2022.
[217] T. Teepe, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Towards a deeper understanding of skeleton-based gait recognition," in *Proc. CVPR Workshops*, 2022, pp. 1569–1577.
[218] H.-M. Hsu, Y. Wang, C.-Y. Yang, J.-N. Hwang, H. L. U. Thuc, and K.-J. Kim, "GAITTAKE: Gait recognition by temporal attention and keypoint-guided embedding," in *Proc. ICIP*, 2022, pp. 2546–2550.
[219] J. Chen, H. Ren, F. S. Chen, S. Velipasalar, and V. V. Phoha, "Gaitpoint: A gait recognition network based on point cloud analysis," in *Proc. ICIP*, 2022, pp. 1916–1920.
[220] E. Pinyoanuntapong, A. Ali, P. Wang, M. Lee, and C. Chen, "GaitMixer: Skeleton-based gait representation learning via wide-spectrum multi-axial mixer," in *Proc. ICASSP*, 2023, pp. 1–5.
[221] X. Li, Y. Makihara, C. Xu, and Y. Yagi, "End-to-end model-based gait recognition using synchronized multi-view pose constraint," in *Proc. ICCVW*, 2021, pp. 4106–4115.
[222] H. Zhu, Z. Zheng, and R. Nevatia, "Gait recognition using 3-D human body shape inference," in *Proc. WACV*, 2023, pp. 909–918.
[223] H. Zhu, W. Zheng, Z. Zheng, and R. Nevatia, "GaitRef: Gait recognition with refined sequential skeletons," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, 2023, pp. 1–10.
[224] H. Guo and Q. Ji, "Physics-augmented autoencoder for 3D skeleton-based gait recognition," in *Proc. ICCV*, 2023, pp. 19627–19638.
[225] L. Wang, B. Liu, F. Liang, and B. Wang, "Hierarchical spatio-temporal representation learning for gait recognition," in *Proc. ICCV*, 2023, pp. 19639–19649.
[226] K. Ma, Y. Fu, C. Cao, S. Hou, Y. Huang, and D. Zheng, "Learning visual prompt for gait recognition," in *Proc. CVPR*, 2024, pp. 593–603.
[227] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. CVPR*, 2017, pp. 652–660.
[228] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu, "Rethinking network design and local geometry in point cloud: A simple residual MLP framework," 2022, *arXiv:2202.07123*.
[229] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5105–5114.
[230] A. Goyal, H. Law, B. Liu, A. Newell, and J. Deng, "Revisiting point cloud shape classification with a simple and effective baseline," in *Proc. ICML*, 2021, pp. 3809–3820.
[231] R. Wang, C. Shen, C. Fan, G. Q. Huang, and S. Yu, "PointGait: Boosting end-to-end 3D gait recognition with point clouds via spatiotemporal modeling," in *Proc. IJCB*, 2023, pp. 1–10.
[232] A. K. Jain, D. Deb, and J. J. Engelsma, "Biometrics: Trust, but verify," *IEEE Trans. Biom., Behav., Identity Sci.*, vol. 4, no. 3, pp. 303–323, Jul. 2022.
[233] F. Abdullakutty, E. Elyan, and P. Johnston, "A review of state-of-the-art in face presentation attack detection: From early development to advanced deep learning and multi-modal fusion methods," *Inf. Fusion*, vol. 75, pp. 55–69, Nov. 2021.
[234] A. Hadid, M. Ghahramani, V. Kellokumpu, M. Pietikäinen, J. Bustard, and M. Nixon, "Can gait biometrics be spoofed?" in *Proc. PICPR*, 2012, pp. 3280–3283.
[235] G. Mai, K. Cao, P. C. Yuen, and A. K. Jain, "On the reconstruction of face images from deep face templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1188–1202, May 2019.
[236] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu, "A study on gait-based gender classification," *IEEE Trans. Image Process.*, vol. 18, pp. 1905–1910, 2009.
[237] *California Consumer Privacy Act (CCPA)*, California State Legislature, Sacramento, CA, USA, 2018.
[238] "Information security technology—Personal information security specification," Nat. Stand. People's Republ. China, Standard. Admin. China, Beijing, China, document GB/T 35273-2020, 2020.
[239] *Walking Times Square Midtown Manhattan New York City 2020*. (Sep. 8, 2020). [Online Video]. Available: https://www.youtube.com/watch?app=desktop&v=Ph77Bt51WVk
[240] Y. Chen et al., "GeoSim: Realistic video simulation via geometry-aware composition for self-driving," in *Proc. CVPR*, 2021, pp. 7230–7240.
[241] R. Wang, C. Shen, M. J. Marin-Jimenez, G. Q. Huang, and S. Yu, "Cross-modality gait recognition: Bridging LiDAR and camera modalities for human identification," 2024, *arXiv:2404.04120*.
[242] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information Maximization adaptation network," in *Proc. ICCV*, 2019, pp. 692–702.
[243] X. Yang et al., "Towards face encryption by generating adversarial identity masks," in *Proc. ICCV*, 2021, pp. 3897–3907.

Chuanfu Shen received the B.E. degree from the Southern University of Science and Technology, Shenzhen, China, in 2019. He is currently pursuing the Ph.D. degree jointly enrolled in computer science and engineering with the Southern University of Science and Technology and in industrial and manufacturing systems engineering with The University of Hong Kong. His research interests include human retrieval, gait recognition, and 3-D computer vision.



Shiqi Yu (Member, IEEE) received the B.E. degree in computer science and engineering from the Chu Kochen Honors College, Zhejiang University in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences in 2007. He is currently an Associate Professor with the Department of Computer Science and Engineering, Southern University of Science and Technology, China. He worked as an Assistant Professor and an Associate Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences from 2007 to 2010, and an Associate Professor with Shenzhen University from 2010 to 2019. His research interests include gait recognition, face detection, and computer vision.



Jilong Wang received the B.S. degree from the University of Science and Technology Beijing in 2019. He is currently pursuing the Ph.D. degree with the University of Science and Technology of China and studies with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests are in artificial intelligence, machine learning, and computer vision.

George Q. Huang (Fellow, IEEE) received the B.Eng. degree from Southeast University, and the Ph.D. degree from Cardiff University. He is a Chair Professor of Smart Manufacturing with the Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University. He has conducted research projects in the field of physical Internet (Internet of Things) for manufacturing and logistics with substantial government and industrial grants. He has published extensively, including over 200 refereed journal papers in addition to over 200 conference papers and ten monographs, edited reference books, and conference proceedings. His research works have been widely cited in the relevant field. He serves as an associate editor and an editorial member for several international journals. He is a Chartered Engineer, a Fellow of ASME, HKIE, IET, and CILT, and a member of IIE.

Liang Wang (Fellow, IEEE) received the B.Eng. and M.Eng. degrees from Anhui University in 1997 and 2000, respectively, and the Ph.D. 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, U.K., and Monash University, Australia, a Research Fellow with the University of Melbourne, Australia, and a Lecturer with the University of Bath, U.K. He is currently a Full Professor of the Hundred Talents Program with the New Laboratory of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He is an IAPR Fellow.