Deep Gait Recognition: A Survey

Alireza Sepas-Moghaddam, *Member, IEEE* and Ali Etemad, *Senior Member, IEEE*

Abstract—Gait recognition is an appealing biometric modality which aims to identify individuals based on the way they walk. Deep learning has reshaped the research landscape in this area since 2015 through the ability to automatically learn discriminative representations. Gait recognition methods based on deep learning now dominate the state-of-the-art in the field and have fostered real-world applications. In this paper, we present a comprehensive overview of breakthroughs and recent developments in gait recognition with deep learning, and cover broad topics including datasets, test protocols, state-of-the-art solutions, challenges, and future research directions. We first review the commonly used gait datasets along with the principles designed for evaluating them. We then propose a novel taxonomy made up of four separate dimensions namely body representation, temporal representation, feature representation, and neural architecture, to help characterize and organize the research landscape and literature in this area. Following our proposed taxonomy, a comprehensive survey of gait recognition methods using deep learning is presented with discussions on their performances, characteristics, advantages, and limitations. We conclude this survey with a discussion on current challenges and mention a number of promising directions for future research in gait recognition.

Index Terms—Gait recognition, deep learning, gait datasets, body representations, temporal representation, feature representation

1 Introduction

Gait, defined as the way people walk, contains relevant cues about human subjects [1]. As a result, it has been widely used in different application areas such as affect analysis [2], [3], [4], sport science [5], [6], health [7], [8], [9], and user identification [10], [11], [12]. Gait information can be captured using a number of sensing modalities such as wearable sensors attached to the human body, for instance accelerometers, gyroscopes, and force and pressure sensors [13]. Non-wearable gait recognition systems predominantly use vision, and are therefore mostly known as vision-based gait recognition. These systems capture gait data using imaging sensors with no cooperation from the subjects and even from far away distances [14]. The focus of this paper is to survey vision-based gait recognition systems that have mainly relied on deep learning. We focus solely on vision-based gait recognition as a comprehensive review paper has recently been published, surveying wearable-based gait recognition approaches [13].

The performance of vision-based gait recognition systems, hereafter only referred to only as gait recognition, can be affected by *i)* variations in the appearance of the individual, such as carrying a handbag/backpack or wearing items of clothing such as a hat or a coat; *ii)* variations in the camera viewpoint; *iii)* occlusion factors, for instance where parts of the subject's body are partially covered by an object or by a part of the subject's own body in certain viewpoints (known as self-occlusion) [15], [16]; and *iv)* variations in the environment, such as complex backgrounds [17] and high or low levels of lighting [18], which generally make the segmentation and recognition processes more difficult.

• *The authors are with the Department of Electrical and Computer Engineering, and Ingenuity Labs Research Institute, Queen's University, Kingston, ON K7L 3N6, Canada. E-mail: {alireza.sepasmoghaddam, ali.etemad}@queensu.ca.*

Manuscript received 17 Feb. 2021; revised 16 Oct. 2021; accepted 13 Feb. 2022. Date of publication 15 Feb. 2022; date of current version 5 Dec. 2022. (Corresponding author: Alireza Sepas-Moghaddam.) Recommended for acceptance by L. Torresani. Digital Object Identifier no. 10.1109/TPAMI.2022.3151865

1.1 Unique Characteristics of Gait Recognition

Gait recognition systems pose challenges that are unique to this field, making it a problem that demands an independent treatment. From a 'biometrics' perspective, gait recognition has several unique characteristics that distinguishes it from other biometric modalities. For instance, in contrast to many other biometric systems such as face [19], ear [20], iris[21], [22], and fingerprint [23] recognition that require subjects to be quite close to acquisition systems, gait data can be captured from far away distances [14]. As a result, gait recognition videos may often be recorded with low spatial resolution, hence many details regarding the scene become challenging to detect by automated systems. Moreover, while most biometric recognition systems need the subjects' active cooperation towards acquisition, gait recognition data can be captured in a discrete manner [10]. As a result, the likelihood of recording gait patterns in an uncontrolled and non-obscured manner considerably increases. Interestingly, this very property makes gait difficult to forge by imposters, making it reliable for sensitive applications such as crime analysis [24].

What makes some of the challenges in gait recognition unique and distinct from general 'computer vision' problems is that most gait recognition methods learn representations from analysis of the *skeletons* or *silhouettes* of subjects. Meanwhile, other visual classification problems often heavily rely on derived features from *texture* in addition to shape and structure information. For example, despite the similarities of computer vision problems such as 'person re-identification' [25] and 'human activity recognition' [26] to gait recognition, gait data still pose challenges and properties that are unique to this field. Specifically, person re-identification methods identify subjects across multiple non-overlapping surveillance cameras, or possibly from the same camera but

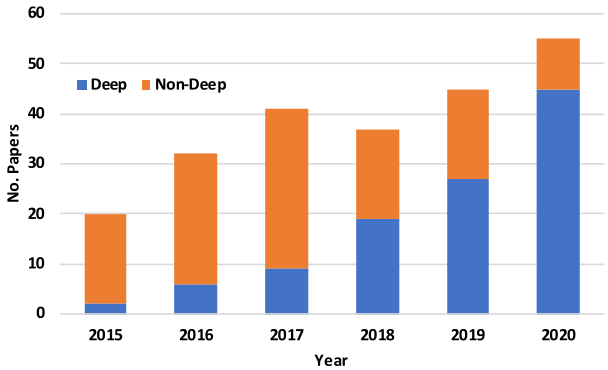


Fig. 1. The number of gait recognition papers published after 2015 using non-deep (orange) and deep (blue) gait recognition methods. These papers have been published in top-tier journals and conferences in the field. Journal publications include IEEE Transactions (19%) including *T-PAMI*, *T-IP*, *T-IFS*, *T-MM*, *T-CSVT*, and *T-Biom*, as well as other top journals (24%) such as *Pattern Recognition* and *Pattern Recognition Letter*. Conference publications include highly ranked computer vision and machine learning conferences (22%) including *CVPR*, *AAAI*, *ICCV*, *ECCV*, *ACCV*, *BMVC*, as well as other top relevant conferences (35%) such as *ICASSP*, *ICIP*, *ICPR*, *ICME*, *ACM Multimedia*, and *IJCB*. The figure shows clear opposing trends between the two approaches, indicating that, unsurprisingly, deep learning methods have become the dominant approach is recent years.

at different time instances. To this end, these methods aim to learn representations that capture appearance characteristics of individuals such as clothing and skin color tone, that are shared across multiple cameras [27]. On the contrary, gait recognition methods aim to learn suitable representations with which *walking patterns* can be disentangled from the visual appearance of the subjects and subsequently used for classification [10]. When comparing gait recognition to human activity recognition methods [28], the goal of the latter is to identify specific movements or actions of a subject from video clips, which can be considered as 'macro' motion patterns. Meanwhile, gait characteristics can be considered nuanced 'micro' patterns that sit on top of a specific activity class, namely *walking*. As a result, the detection of such subtle discriminative information are often more challenging than those dealt with for activity recognition. Furthermore, given the subtlety of gait patterns that make them unique to different subjects, they can often be highly influenced by the temporary personal state of the subject, for instance, fatigue [29], excitement and fear [30], and even injuries [31].

1.2 Motivation

In the past two decades, many gait recognition methods have been developed to tackle the above-mentioned problems. In recent years, there has been a clear trend in migrating from non-deep methods to deep learning-based solutions for gait recognition. To visualize this trend, we present Fig. 1, which illustrates the number of gait recognition papers published after 2015. It is observed that the majority of gait recognition methods in 2019 and 2020 have been designed based on deep neural networks. In Fig. 2, we illustrate the evolution of some of the most important gait recognition methods along with their associated accuracy on the CASIA-B [32] (perhaps the most popular dataset for gait recognition) when available. The first gait recognition method was proposed in 1997 [33], followed by the first shallow neural network for gait recognition in 2008 [34], consisting of only one input layer, one hidden layer, and one output layer. In 2015, the field witnessed significant breakthroughs, notably due to the popularization of deep neural networks [35], [36]. The method entitled GaitNet [37] was then proposed in 2016 based on a 6-layer convolutional neural network (CNN). In 2017, DBNGait [38] was proposed based on a deep belief network (DBN), and in [39] three different deep CNN architectures with different depths and architectures were fused for gait recognition. VGR-Net [40] was one of the important contributions in 2018, followed by the introduction of several significant methods in 2019, including PoseGait [41], DisentangledGait [42], and GaitSet [43], where the best recognition accuracy of 84.2% was achieved by GaitSet [43]. Remarkable advances have been made in 2020, notably by the appearance of several highly efficient methods, including PartialRNN [44], GaitPart [45], GLN [46], HMRGait [47], and 3DCNNGait [48]. The current state-of-the-art results on CASIA-B dataset [32] have been reported by 3DCNNGait [48] with a recognition accuracy of 90.4%.

Several survey papers [10], [11], [12], [13], [14], [49], [50], [51] have so far reviewed recent advances in gait recognition, where some of these papers, for instance [13], [50], [51], have focused on *non-vision-based* gait recognition methods. The most recent survey papers on *vision-based* gait recognition are [10], [11], [12], [14], [49], which only cover the papers published until mid 2018. Nonetheless, many important breakthroughs in gait recognition with deep learning have occurred since 2019, as observed in Figs. 1 and 2. Additionally, none of the surveys [10], [11], [12], [14], [49] have specifically focused on deep learning methods for gait recognition.

1.3 Contribution

This paper surveys the most recent advances in gait recognition until the end of January 2021, providing insights into both technical and performance aspects of the deep gait recognition methods in a systematic way. In this context, we first proposes a novel taxonomy with four dimensions, i.e.,

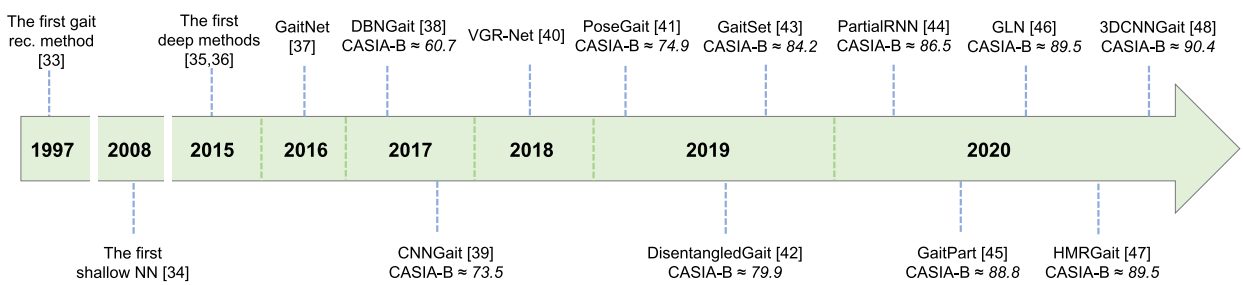


Fig. 2. The evolution of deep gait recognition methods.

body representation, temporal representation, feature representation, and neural architecture, to help characterize and organize the available methods. Following our proposed taxonomy, a comprehensive survey of all the available deep gait recognition methods is presented, along with discussions on their characteristics and performances. We have established certain search protocols to make sure other scholars can confidently use this survey in their future research.

Our key contributions are summarized as follows:

- We propose a novel taxonomy with four dimensions to characterize and organize the available deep gait recognition methods.
- We provide a taxonomy-guided review on the evolution of the deep gait recognition methods, where most of these methods have not been reviewed in previous surveys. This provides insights for new topic exploration and future algorithm design.
- We present comparisons between the state-of-the-art using the available results reported on large-scale public gait datasets, providing insight into the effectiveness of different deep gait recognition methods.
- We review 15 publicly available vision-based datasets for gait recognition, along with their associated test protocols.
- We discuss a number of open challenges and identify important future research directions that will be of benefit to researchers working on gait recognition.

1.4 Organization

The rest of this survey is structured as follows. We start with describing the systematic approach used to collect the papers and review the literature. Next, in Section 3, we review the available gait datasets along with their associated test protocols. We then use these datasets and protocols to report the existing performance results when reviewing the deep gait recognition methods. Section 4 presents our proposed taxonomy. Following, Section 5 surveys the state-of-the-art on deep gait recognition and discusses the evolutional trends of deep gait recognition over the past few years. Finally, Section 6 discusses some deep gait recognition challenges and identifies a number of future research areas.

2 Review Methodology

We employed a search protocol to ensure other scholars can confidently use this survey in their future research. To do so, we first discovered candidate papers through the Google Scholar [52] search engines and digital libraries, namely IEEE Xplore [53], ACM Digital Library [54], ScienceDirect [55], and CVF Open Access [56]. Our search terms included combinations of the following queries: "gait recognition", "gait identification", "gait biometric", "neural architecture", "deep learning", and "deep representations". We then filtered the search results, thus excluding papers that neither use deep learning methods for gait recognition nor demonstrate enough technical clarity/depth. To be more specific about the 'clarity/depth' criteria, we excluded the papers that *i*) use non-vision sensors for gait recognition; *ii*) do not propose a new solution; *iii*) use non-standard or private datasets for performance evaluation; *iv*) do not compare the performance of their solution to the state-of-the-art. In cases where other modalities were combined with vision-based sensors, only the technical solution focusing on the vision-based aspect was studied.

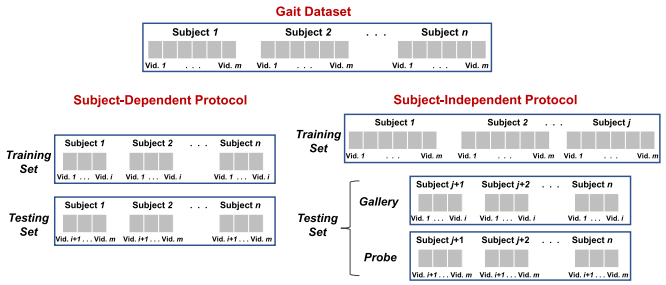


Fig. 3. An overview of test protocols is presented. These protocols can be categorized into subject-dependent or subject-independent according to whether test subjects appear in the training set or not.

Naturally, we imposed restrictions on the date of publications to only include search results after 2014, when deep neural networks were first used for biometric recognition [72], [73]. We then used the returned results in order to perform forward and backward searches, respectively identifying the other resources that have cited the returned articles and the references cited by the returned articles. We repeated this process with the new identified resources until we collected the most relevant papers to the best of our knowledge. We eventually ended up with a final set of publications that have used deep learning for gait recognition.

3 Test Protocols and Datasets

3.1 Protocols

Evaluation protocols for gait recognition solutions can generally be categorized into *subject-dependent* and *subject-independent*. As illustrated in Fig. 3, in the subject-dependent protocol, both the training and testing sets include samples from all the subjects. However, in the subject-independent protocol, the test subjects are disjoint from the training subjects. Here, the test data are further divided into gallery and probe sets, and the learned model on the disjoint training subjects is then used to extract features from gallery and probe sets. Finally, a classifier is used to compare the probe features with the gallery ones in order to identify the most similar gait patterns and label them as being from the same identity. Both subject-dependent and subject-independent protocols have been widely adopted for gait recognition. For example, in the TUM GAID [66] dataset, subject-dependent protocols have been often used, while in the CASIA-B [32] and OU-MVLP [68] large-scale datasets, subject-independent protocols are utilized. Gait recognition results in the literature have all been measured and presented using rank-1 recognition accuracy.

3.2 Datasets

In order to evaluate gait recognition systems, different datasets have been collected. These datasets cover various parameters related to acquisition viewpoints, environment conditions, and appearance of the subjects. Generally speaking, large datasets, both in terms of number and distribution

TABLE 1
Summary of Well-Known Gait Datasets Used in the Literature

Dataset	Year	Data Type	# of Subjects & Sequences	Environment	# of Views	Variations
CMU MoBo [57]	2001	RGB; Silhouette	25 / 600	Indoor	6	3 Walking Speeds; Carrying a Ball
SOTON [58]	2002	RGB; Silhouette	115 / 2,128	Indoor & Outdoor	2	Normal Walking on a Treadmill
CASIA-A [59]	2003	RGB	20 / 240	Outdoor	3	Normal Walking
USF HumanID [60]	2005	RGB	122 / 1,870	Outdoor	2	Outdoor Walking; Carrying a Briefcase; Time Interval
CASIA-B [32]	2006	RGB; Silhouette	124 / 13,680	Indoor	11	Normal Walking; Carrying a Bag; Wearing a Coat
CASIA-C [61]	2006	Infrared; Silhouette	153 / 1,530	Outdoor	1	3 Walking Speeds; Carrying a Bag
OU-ISIR Speed [62]	2010	Silhouette	34 / 306	Indoor	4	Nine walking speeds
OU-ISIR Clothing [63]	2010	Silhouette	68 / 2,746	Indoor	4	Up to 32 combinations of clothing
OU-ISIR MV [64]	2010	Silhouette	168 / 4,200	Indoor	25	24 azimuth views and 1 top view
OU-ISIR [65]	2012	Silhouette	4,007 / 31,368	Outdoor	4	Normal Walking
TUM GAID [66]	2012	RGB; Depth; Audio	305 / 3,737	Indoor	1	Normal Walking; Backpack; Wearing coating shoes
OU-ISIR LP Bag [67]	2017	Silhouette	62,528 / 187,584	Indoor	1	Seven different carried objects
OU-MVLP [68]	2018	Silhouette	10,307 / 259,013	Indoor	14	Normal Walking
CASIA-E [69], [70]	2020	Silhouette	1014 / Undisclosed	Indoor & Outdoor	15	3 Scenes; Normal Walk; Carrying a Bag; Wearing a Coat
OU-MVLP Pose [71]	2020	Skeleton	10,307 / 259,013	Indoor	14	Normal Walking

of samples and parameters, are often desired and preferred to allow for deep neural networks to be trained effectively. We present an overview of the main characteristics of well-known gait datasets in Table 1. These characteristics include the type and modality of data, number of subjects and sequences, number of viewpoints, and also the variations covered in each dataset. To show the chronological evolution of these datasets, we have sorted these datasets by the order of release date in Table 1. According to Table 1, CASIA B [32], CASIA-E [69], OU-ISIR MV [64], and OU-MVLP [68] cover the highest number of acquisition viewpoints while OU-ISIR [65], OU-ISIR LP Bag [67], and OU-MVLP [68] include the highest number of gait sequences. In the following we review the gait datasets available in Table 1 along with the associated test protocols.

CMU MoBo. The CMU Motion of Body (MoBo) [57] dataset is one of the first gait datasets in the literature, and consists of RGB and silhouette data from 25 different subjects who walk on a treadmill. This dataset covers three subsets including slow walking speed, fast walking speed, and walking when holding a ball. The test protocol proposed by the authors of this dataset suggests training with one subset and testing with the other subsets. All six combinations are often used to report the results.

SOTON. The SOTON dataset [58] contains data from 115 subjects. All the sequences have been recorded both indoor and outdoor, with a fixed camera, capturing the subjects walking along a straight path. The indoor gait data has been captured from the subjects when walking on a treadmill. Different papers have divided this dataset into training and testing sets differently, and there is no pre-defined test protocol presented with the dataset.

CASIA-A. CASIA-A [59] is a dataset that includes data from 20 subjects in outdoor environments. Participants have walked along a straight line, while three cameras positioned at 0°, 45°, and 90° have captured the gait videos with an average of 90 frames per a sequence. A cross-view test protocol is the most widely used protocol for this dataset, where solutions are trained with all the available views, excluding one which is then used for testing.

USF HumanID. The USF HumanID dataset [60] has been collected in the context of the HumanID gait challenge, and includes outdoor gait videos from 122 subjects who have walked in an elliptical path. This dataset covers challenging variations including carrying a briefcase, walking on different surfaces, wearing different shoes, and with acquisition times. The data has been captured from two viewing angles by left and right cameras. The evaluation study has been made available along with the dataset [60], which considers 12 different test protocols with respect to the above-mentioned variations.

CASIA-B. CASIA-B dataset [32] is the most widely used gait dataset, and contains multi-view gait data from 124 persons in the form of both RGB and silhouettes. Acquisition has been performed from 11 different viewing angles that range from 0° to 180° with 18° increments. The dataset considers three different walking conditions namely normal walking (NM), walking with a coat (CL), and walking with a bag (BG), respectively with 6, 2, and 2 gait sequences per person per view. The most frequently used test protocol for CASIA-B is a subject-independent protocol which uses the data from the first 74 subjects for training, and the remaining 50 subjects for testing. The test data is then split into a gallery set including the first four gait sequences from the NM gait data and the probe set consists of the rest of the sequences, namely the remaining 2 NM, 2 CL, and 2 BG sequences, per each subject per each view. The results have been mostly reported for all the viewing angles angles, excluding the probe sequences with angles identical to the references.

CASIA-C. The CASIA-C dataset [61] includes infrared and silhouette data from 153 different subjects, and the sequences have been captured under different variations at night. These variations include three different walking speeds namely slow walking (SW), normal walking (NW), and fast walking (FW), as well as carrying a bag (BW). There are 4 NW, 2 SW, 2 FW, and 2 BW sequences per each subject. As for the evaluation scheme, cross-speed walker identification tests have been considered.

OU-ISIR Speed. The OU-ISIR Speed dataset [62] provides silhouette data from 34 subjects. This dataset is suitable for evaluation of robustness of gait recognition methods with respect to walking speeds, as it includes nine different speeds, ranging from 2 km/h to 11 km/h, with 1 km/h interval. Cross-speed tests have been adopted for this dataset.

OU-ISIR Clothing. The OU-ISIR Clothing dataset [63] includes data from 68 subjects who wore up to 32 different

types of clothing. Gait sequences were collected in two indoor acquisition sessions on the same day. A subject-independent test protocol has been provided along with the dataset [65], which divides the data into pre-defined training, testing, and probe sets particularly with respect to the clothing conditions.

OU-ISIR MV. The OU-ISIR MV dataset [64] consists of gait silhouettes from 168 subjects with an age range of 4 to 75 years old, and almost equal number of male versus female participants. The data has been captured from a large range of view variations, including 24 azimuth views, along with 1 top view. A cross-view test protocols have been widely adopted for this dataset.

OU-ISIR. The OU-ISIR dataset [65] is a large-scale gait dataset, consisting of gait data from 4,007 subjects with almost equal gender distribution, and with ages ranging from 1 to 94 years old. The gait sequences have been captured in two different acquisition sessions in indoor halls using four cameras placed at 55°, 65°, 75°, and 85° degrees. As there are two sequences available for each subject, the test protocol uses the first sequences as gallery and the other one as probe samples.

TUM GAID. The TUM GAID [66] is a multi-modal gait dataset, including RGB, depth, and audio data from 305 subjects. For a selected set of 32 subjects, the dataset has been captured in two different outdoor acquisition sessions in winter and summer. 10 sequences have been captured from each subject, including normal walking (N), walking with a backpack (B), and walking with disposable shoe covers (S). The test protocol has been made available by the original authors, dividing the data into training, validation, and test sets. Recognition experiments are often then carried out with respect to the N, B, and S gait variations.

OU-ISIR LP Bag. The OU-ISIR LP Bag dataset [67] consists of gait videos from 62,528 subjects with carried objects, captured using one camera in constrained indoor environments. Three sequences have been obtained per each subject, one with a carried object and two without it. Following the test protocol proposed in [67], the training set contains data from 29,097 subjects with two sequences with and without the carried objects, and the test set includes the other 29,102 disjoint subjects. In order to divide the test set into probe and gallery sets, two approaches have been adopted, respectively under cooperative and uncooperative scenarios. For the cooperative scenario, the gallery set only contains sequences without carried objects, where the probe set includes sequences with seven different types of carried objects. In the uncooperative scenario, gallery and probe sets are randomly formed such that they both contain sequences with and without carried objects.

OU-MVLP. The OU-MVLP dataset [68] is the largest available gait dataset in term of number of gait sequences (259,013). The dataset provides videos of silhouettes and is acquired in two acquisition sessions per each subject. The gender distribution of subjects is almost equal with an age range of 2 to 87 years old. This dataset has been acquired from 14 different views, ranging from 0° to 90°, and 180° to 270°, where the angle change in each step is 15°. Pre-determined lists of 5153 and 5154 subjects have been designated and provided with the dataset as training and testing sets respectively. For testing, sequences from the first and second acquisition sessions respectively form gallery and probe sets. In most of the recent gait recognition papers, either all or four viewing angles, notably 0°, 30°, 60°, and 90°, are considered.

CASIA-E. CASIA-E dataset [69] consists of silhouettes from 1014 subjects with hundreds of sequences per subject, captured in three types of scenes with a simple static, a complex static, and a complex dynamic background. The data has been captured considering three different walking conditions, including normal walking (NM), walking with a coat (CL), and walking with a bag (BG). This dataset has been acquired from 15 different angles, including two vertical views with a height of 1.2 m and 3.5 m, as well as 13 horizontal views ranging from 0° to 180° with 15° increments. This dataset was recently used in the *TC4 Competition and Workshop on Human Identification at a Distance 2020* [73], where the training set included the entire data from the first 500 subjects, while 25 sequences from the last 514 subjects were used for validation. The remaining sequences were used for testing.

OU-MVLP Pose. The OU-MVLP Pose dataset [70] was built upon the OU-MVLP [68], extracting pose skeleton sequences from the RGB images available in OU-MVLP. Two subsets have been created using pre-trained versions of OpenPose [74] and AlphaPose [75] to extract human joint information. The test protocol is similar to the one proposed for OU-MVLP [68].

4 Proposed Taxonomy

In order to help illustrate an overall structure for the available gait recognition approaches, a few taxonomies have been proposed in the literature [18], [51], [76], which have organized the available solutions from different perspectives. The taxonomy proposed in [51] is based on the type of sensors, classifiers, and covariate factors such as occlusion types. The taxonomy in [76] categorizes gait recognition methods based on the type of features used. Finally, the one proposed in [18] considers user appearance, camera, light source, and environment-related factors. Nevertheless, despite the availability of these taxonomies, none focus on deep gait recognition methods that are most successful nowadays. We thus propose a new taxonomy in this paper to better illustrate the technological landscape of gait recognition methods with a particular focus on deep learning techniques. Fig. 4 presents our proposed taxonomy which considers four dimensions, namely body representation, temporal representation, feature representation, and neural architecture. The details of each of these dimensions are described in the following.

4.1 Body Representation

This dimension relates to the way the body is represented for recognition, which can be based on *silhouettes* or *skeletons*. Silhouette is the most frequently used body representation in the literature that can be easily computed by subtracting each image containing the subject from its background, followed by binarization. Gait silhouettes are proven to be effective and convenient for describing the

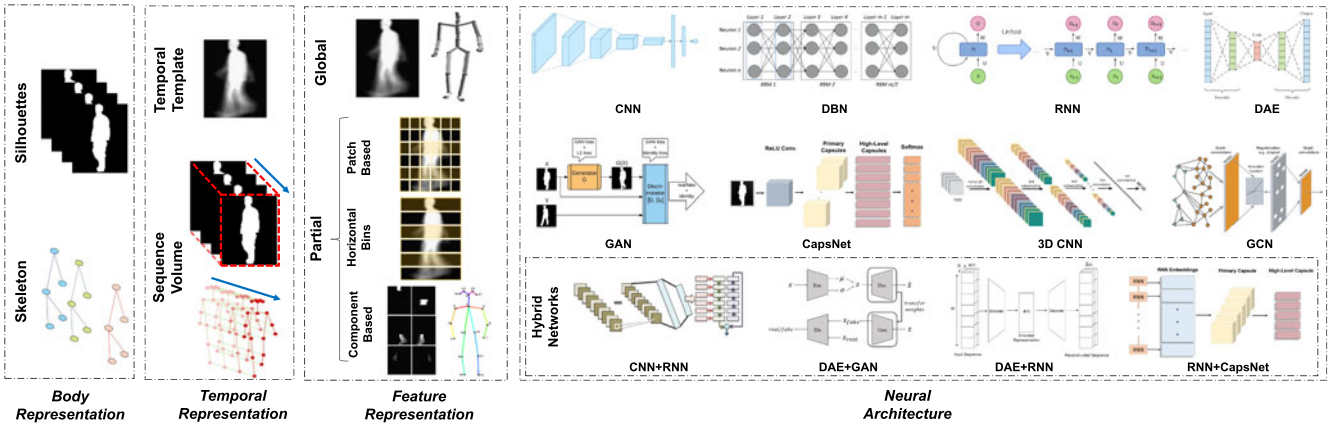


Fig. 4. Our taxonomy consisting of 4 dimensions: body representation, temporal representation, feature representation, and neural architecture.

body state in a single frame with low computational cost. This body representation forces recognition solutions to focus on 'gait' as opposed to clothing and other non-gait factors that could, from the perspective of a classifier, be used for identification. A sequence of silhouettes can represent useful gait features such as speed, cadence, leg angles, gait cycle time, step length, stride length, and the the ratio between swing and stance phases [77], [78]. It can also be processed to extract motion data, for example using optical flow calculation [41], [79], [80]. Nonetheless, gait silhouettes are more sensitive to changes in the appearance of the individuals, for instance via different clothing and carrying conditions.

Skeleton body representation can be captured using depth cameras [81] or alternatively be estimated using pose-estimation methods [82]. Static and dynamic features, for instance stride length, speed, distances, and angles between joints, can be obtained from skeleton joints [50]. Gait recognition methods based on this type of body representation are generally more robust against viewpoint changes due to the consideration of joint positions [83], as opposed to silhouette-based methods. Skeleton-based methods are also more robust against appearance changes [78] as the pose-estimation step generally learns to detect body joints over different clothing conditions, which is not the case for gait silhouettes. However, since these approaches rely heavily on accurate detection of body joints, they are generally more sensitive to occlusions [78]. Additionally, the use of pose-estimators imposes a computational overhead to these recognition systems [84].

4.2 Temporal Representation

This dimension deals with approaches used to represent the temporal information in gait sequences. Two types of representations, *templates* and *volumes*, have been commonly used in the literature. Following we describe these representations.

Templates aggregate temporal walking information over a sequence of silhouettes in a single map, for example by averaging the silhouettes over at least one gait cycle. This operation enables recognition solutions to be independent of the number of frames once template maps have been created. With respect to deep gait recognition architectures, gait silhouettes can be aggregated in the initial layer of a network (Fig. 5a), also known as *temporal templates*, where the aggregated map can then be processed by subsequent layers [85], [86], [87], [88], [89]. Gait silhouettes can alternatively be aggregated in an intermediate layer of the network after several convolution and pooling layers (Fig. 5b), also known as *convolutional template* [43], [44]. Examples of temporal templates include: (*i*) gait energy images (GEI) [85], which average gait silhouettes over one period/sequence (Fig. 5c); (*ii*) chrono gait images (CGI) [86], which extract the contour in each gait image to be then encoded using a multi-channel mapping function in the form a single map (Fig. 5d); (*iii*) frame-difference energy images (FDEI) [87], which preserve the kinetic information using clustering and denoising algorithms, notably when the silhouettes are incomplete (Fig. 5e); (*iv*) gait entropy images (GEnI) [88], computing entropy for each pixel in the gait frames to be then averaged in a single gait template (Fig. 5f); and (*v*) period energy images (PEI) [89], a generalization of GEI that preserves more spatial and temporal information by exploiting a multi-channel mapping function based on the

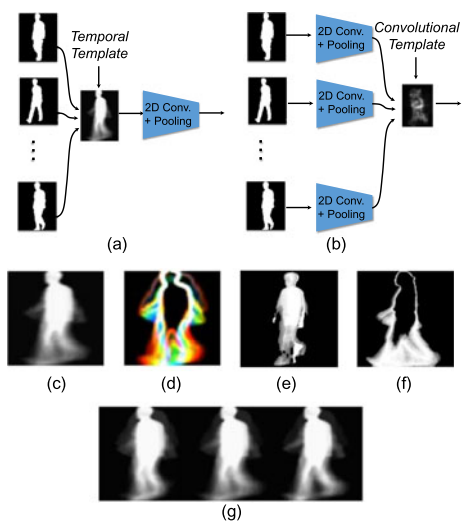


Fig. 5. Overview of temporal representations. Generating templates in: (a) the initial layer of a deep network; (b) an intermediate layer of the network after several convolution and pooling layers. Illustration of (c) GEI [85], (d) CGI [86], (e) FDEI [87], (f) GEnI [88], and (g) PEI [89] temporal gait templates.

TABLE 2
Overview of Recent CNN Architectures Adopted for Deep Gait Recognition

Method	Input Size	Total # of Layers	# of Conv. Layers	# of Pool. Layers	# of FC Layers
GEINet [37]	88×128	6	2	2	2
Ensem. CNNs [39]	128×128	7	3	2	2
MGANs [89]	64×64	8	4	1	3
EV-Gait [111]	128×128	9	6	0	2
Gait-joint [112]	64×64	16	12	2	2
Gait-Set [43]	64×64	9	6	2	1
Gait-RNNPart [44]	64×64	9	6	2	1
Gait-Part [45]	64×64	9	6	2	1
SMPL [47]	64×64	5	3	1	1
Caps-Gait [95]	64×64	9	6	2	1

amplitudes of frames (Fig. 5g). Examples of convolutional templates include set pooling [43] and gait convolutional energy maps (GCEM) [44], which average convolutional maps obtained by several convolution and pooling layers, over the whole sequence.

To preserve and learn from the order and relationship of frames in gait sequences, instead of aggregating them, sequence volume representations have be adopted (see Fig. 4, second box from the left). Then, to learn the temporal information, two different approaches have been adopted. In the first approach, the temporal dynamics over the sequences are learned using recurrent learning strategies, for example recurrent neural networks, where each frame is processed with respect to its relationships with the previous frames [47], [90], [91]. The second approach first creates 3D tensors from spatio-temporal information available in sequences, where the depth of the tensors represent the temporal information. These tensors are then learned, for example using 3D CNNs [48], [92], [93] or graph convolutional networks (GCNs) [94].

4.3 Feature Representation

This dimension encapsulates the region of support for representation learning, which can be either *global* or *partial*. The process of learning silhouettes or skeletons holistically is referred to as global representation learning. On the other hand, when learning partial representations, gait data is split into local regions, e.g., patches, body components, and vertical/horizontal bins (see Fig. 4, third box from the left). These local regions are then further processed, for example by recurrent neural networks [44], capsule networks [95], attention-based networks [96], or fully connected layers [43]. Methods based on global representations tend to be more sensitive to occlusions and appearance changes as well as missing key body parts [43], [97]. On the other hand, partial regions often maintain different contributions towards the final recognition performance, thus learning their importance can improve the overall performance of gait recognition methods [43], [45]. Additionally, the relations between these partial features can be learned, to preserve positional attributes such as scale, rotation, and location, which improve the robustness of gait recognition methods against orientation and view changes [44], [95].

4.4 Neural Architectures

Deep neural networks (DNNs) capture high-level abstractions using hierarchical architectures of multiple nonlinear transformations. Different neural architectures have been designed for gait recognition problems, whose descriptions are provided below.

4.4.1 Convolutional Neural Networks

Convolutional neural networks (CNNs) have been used the most for gait recognition. These models are generally used to learn an embedding where the body shape, represented as a silhouette or skeleton, is encoded in the feature space. Specifically, CNNs generally consist of different types of layers including convolutional, pooling, and fully connected layers. Convolutional layers convolve learned filters with the input image to create activation feature maps that capture features with varying levels of detail. The convolutional layers also include activation functions such as a ReLU [98] or a tanh [99] functions, to increase the non-linearity in the output. Pooling layers then reduce the spatial size of the feature maps by using nonlinear down-sampling strategies, such as average or maximum pooling, thus decreasing the complexity of the network. Fully connected layers are finally used to learn the resulting 2D feature maps into 1D vectors for further processing.

To better analyze CNNs adopted in the state-of-the-art gait recognition methods, we provide an overview of the most successful used architectures in Table 2. Note that for the methods combining CNNs with other types of networks, e.g., autoencoder, capsule, and Long Short-Term Memory (LSTM), we only present the architectures of the CNN components in the table. As can be seen, there is no need for state-of-the-art gait recognition models to exploit very deep CNN architectures. This is due to the fact that input gait data, either in the from of silhouettes or skeletons, do not present considerable complexity in terms of texture information. Hence, even fewer than 10 layers are shown to be sufficient for encoding gait frames. This is contrary to many other domains, such as face or activity recognition, where very deep networks such as ResNet [100] and inception [101] are used to learn highly discriminative features. In Table 2, we also present the size of CNN inputs, showing a trend toward a 64×64 resolution in the recent literature. Additionally, an analysis in [39] showed that the resolutions

of 64×64 and 128×128 lead to the best gait recognition results for several tested CNNs, where the input resolution of 128×128 works slightly better than 64×64. However, as a higher input resolution implies more convolutional and pooling layers, the input resolution of 64×64 has been most widely adopted to limit the computational complexity of the solutions.

4.4.2 Deep Belief Networks

A deep belief network (DBN) [102] is a probabilistic generative model, composed by staking restricted Boltzmann machines (RBMs) [103] to extract hierarchical representations from the training data. Each RBM is a two-layer generative stochastic model, including a visible and a hidden layer, with connections between the adjacent layers and without connections between the units within each layer. The weights and biases of the units define a probability distribution over the joint states of the visible and hidden units. DBNs have been used for gait recognition in [104] and [38]. In [104], fitting, body parameters, and shape features were extracted from silhouettes to be then learned by DBNs, thus extracting more discriminative features. In [38], gait has been first represented as motion and spatial components, and two separate DBNs were trained for each component. The extracted features were finally concatenated to represent the final feature.

4.4.3 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) have been widely applied to temporal or sequence learning problems, achieving competitive performances for different tasks [105], including gait recognition [42], [44], [91], [95], [106], [107], [108]. A layer of RNN is typically composed of several cells, each corresponding to one input element of the sequence, e.g., one frame of a gait video. RNNs can also stacks several layers to make the model deeper, where the output of the i^{th} cell in j^{th} layer feeds the i^{th} cell in the $(j+1)^{th}$ layer. Each cell is connected to its previous and subsequent cells, thus memorizing information from the previous time steps [105]. Among different RNN architectures, LSTM [109] and Gated Recurrent Units (GRU) [110] are the most widely used RNN architectures that have been used to learn the relationships available in a gait sequence using memory states and learnable gating functions. In an LSTM network [109], the cells have a common cell state, which keeps long-term dependencies along the entire LSTM cell chain, controlled by two gates, the so-called input and forget gates, thus allowing the network to decide when to forget the previous state or update the current state with new information. The output of each cell, the hidden state, is controlled by an output gate that allows the cell to compute its output given the updated cell state. GRU [110] is another form of RNN that does not use output activation functions as opposed to LSTM. This architecture also includes an update gate that allows the network to update the current state with respect to the new information. The output of the gate, also known as the reset gate, only maintains connections with the cell input.

There have been three different approaches to use RNNs for gait recognition systems. The first approach [91] (Fig. 6a) that has been mostly adopted for skeleton representations,

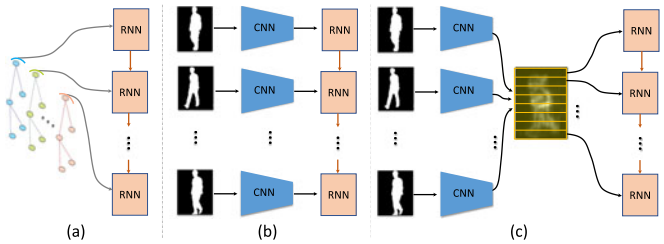


Fig. 6. Three different approaches for using RNNs in the context of deep gait recognition systems: (a) RNNs directly learn from the movement of joint positions; (b) RNNs are combined with CNNs; and (c) RNNs recurrently learn the relationships between partial representations in gait templates.

uses RNNs in order to learn from temporal relationships of joint positions. In the second approach [106], [107] (Fig. 6b), as will be discussed later in detail in Section 4.4.9, RNNs are combined with other types of the neural architectures, notably CNNs, for learning both spatial and temporal information. The last approach that has been recently adopted in [44], [95] (Fig. 6c) uses RNNs to recurrently learn the relationships between partial representations from a single gait template, for instance GCEM [44].

4.4.4 Deep AutoEncoders

Deep auto-encoder (DAE) is a type of network that aims to extract so called bottleneck features or latent space representations, using an encoder-decoder structure. The encoder transforms the input data into a feature representation and the decoder part transforms the representation back to the original input data. The encoder generally includes several fully connected and/or convolutional layers while the decoder consists of layers that perform the inverse operations. DAE networks are generally trained with the aim of minimizing the reconstruction error that measures the difference between the original input and the reconstructed version. Once a DAE is trained, the bottleneck features which are a latent/compressed representation of the knowledge of the original input, are extracted to be used for classification, i.e., gait recognition in our case. The method proposed in [113] uses a DAE network, first encoding the input temporal templates using four convolutional layers to extract feature. The decoder then reconstructs the input from the extracted features using four deconvolutional layers. In [114], an auto-encoder with 7 fully connected layers along with input and output layers was used to extract robust gait features. In [115], a DAE was used to disentangle the input temporal template into identity and covariate features. The backbone of the encoder was based on the Inception module in GoogLeNet [116], extracting multi-scale identity and covariate features. The decoder then took those features as input to reconstruct the temporal template using deconvolutional layers.

4.4.5 Generative Adversarial Networks

Generative Adversarial Networks (GANs) include a generator and a discriminator [117], where the generator aims to deceive the discriminator by synthesizing fake samples that resemble the real ones. In turn, the discriminator aims to distinguish between the fake and real samples. As a result of this minimax game between these two components, GANs can generate realistic synthesized samples. In the

context of gait recognition, GANs can be used to solve the problem of gait variations due to clothing, viewpoints, and carrying conditions. For instance, GANs can transform gait data from one view to another, or change the type of clothing worn by the subject, or even remove a backpack that was originally carried by the subject. Such disentanglement of identity from confounding factors often results in improvements in the performance of gait recognition systems [89], [118], [119], [120], [121], [122]. Nevertheless, one of the most important challenges toward manipulating gait data is the preservation of human identity features while modifying appearance characteristics in the representation space. To this end, two discriminators are often used [120]. The first discriminator is used to distinguish real versus fake samples in order to ensure that the generated images appear realistic. The second discriminator is exploited to ensure that identity information are preserved by taking a pair of source and target images as input and producing a scalar probability of whether the input pair belongs to the same person or not.

Different types of GANs have recently been adopted for gait recognition. Multi-task GAN (MGANs) [89] have been proposed for cross-view gait recognition, where a CNN is used to learn the temporal template as view-specific features in a latent space. Then, the features are transformed from one view to another using a view transform layer. The network is then trained with multi-task adversarial and pixel-wise losses. In another paper, Discriminant Gait GAN (DiGGAN) [120] considered the mechanisms of using two independent discriminators in order to transfer GEIs form a certain viewpoint to a different viewing angle while also preserving identity information. In [121] a Two-Stream GAN (TS-GAN) was proposed to learn both global and partial feature representations when transforming GEI temporal templates with different viewing angles to a GEI temporal templates with a standard view, i.e., 90°.

4.4.6 Capsule Networks

Capsule Networks (CapsNet) [123] have been proposed to address two important shortcomings in CNNs, namely the limits of scalar activations and poor information routing through pooling operations by respectively exploiting capsule activation values and routing-by-agreement algorithms. CapsNets are composed of capsules which are groups of neurons that explicitly encode the intrinsic viewpoint-invariant relationships available in different parts of the objects. In the context of gait representation learning, a CapsNet can model and understand the structural relationship between the various parts of the body, such as the relationships between legs and feet, upper body and lower body, and trunk and limbs, using a learnable pose matrix. A CapsNet can also be used to model internal hierarchical representations between multiple gait silhouettes or skeleton joint coordinates of a subject in a video. This is in contrast to the standard pooling layers often used in CNNs, which fail to preserve positional attributes in the human body, such as locations, scales, rotations, and relationships between the body parts. CapsNets generally include two blocks, primary and high-level group of capsules. The first block encodes spatial information with several layers including convolutional, reshaping, and squashing layers, followed by the second block that learns deeper part-whole relationships between hierarchical sub-parts. The concept of capsule network has been recently adopted for gait recognition [95], [124], [125]. The method proposed in [124] first learns the properties of GEI templates using a CNN. It then uses a CapsNet with dynamic routing to retain the relationship within each template with the aim of finding more robust features. Capsule networks have also been combined with other types of deep networks in [95] and [125], which we will review in Section 4.4.9.

4.4.7 3D Convolutional Neural Networks

3D Convolutional neural networks (3D CNNs) have been recently adopted for gait recognition to learn spatio-temporal dynamics over whole gait sequences [48], [93], [126]. 3D CNNs are able to extract features that are more robust to changes in camera viewpoints and the appearance of subjects. 3D CNNs take the stacked gait frames in the form of a 3D tensor as input, and then use multiple 3D convolution filters and pooling operations to extract the spatio-angular representations. The limitation of 3D CNNs for gait recognition is the lack of flexibility in processing variable length sequences. In [48], this shortcoming was addressed by exploiting multiple 3D CNNs to integrate temporal information at different scales. In [126], a 3D CNN network containing 13 3D convolution filters and pooling layers along with two fully connected layers was designed for gait recognition. The method in [93] is composed of several global and partial 3D convolutional layers, where the standard 3D pooling layer was modified to aggregate temporal information in local clips.

4.4.8 Graph Convolutional Networks

Graph convolutional networks (GCNs) have been recently developed to extend CNNs to a higher dimensional domain using arbitrarily structured graphs and graph convolution filters [127]. Given the inherent hierarchical and graph-like nature of the human body, GCNs can jointly model both the structural information of the human body and temporal relationships available between gait frames in order to learn discriminative and robust features with respect to camera viewpoints and subject appearances. Gait recognition methods based on GCNs consider gait sequence volumes as the spatio-temporal representations for gait recognition [94], [127]. In [94], gait features were extracted by forming a spatio-temporal graph from the available video sequences. The final features were then obtained using a joint relationship learning scheme by mapping the features onto a more discriminative subspace with respect to human body structure and walking pattern.

4.4.9 Hybrid Networks

A large number of hybrid deep networks that make use of two or more types of networks have been proposed to boost the performance of gait recognition systems. Among these, architectures with CNN+RNN, DAE+GAN, DAE+RNN, and RNN +CapsNet components are the most popular in the deep gait recognition literature (see Fig. 4). Following we provide the descriptions and examples of these four hybrid architectures.

CNN+RNN. Integration of CNNs with RNNs (notably LSTM and GRU) for learning the temporal relationships following spatial encoding is perhaps the most popular approach for spatio-temporal learning, which has also been used for gait recognition in the literature. In [128], a deep gait recognition system was proposed by combining eight different CNN architectures with LSTM to obtain spatio-temporal features from image sequences. The proposed method in [129] first divides gait silhouettes into 4 horizontal parts, where each part was fed to an individual CNN with 10 layers. An attention-based LSTM was then used to output frame-level attention scores for each sequence of CNN features. The CNN features were finally multiplied by their corresponding weights to selectively focus on the most important frames for gait recognition. In [44], convolutional maps from gait frames were first learned using an 8-layer CNN. The convolutional maps were then aggregated to form GCEM templates which were then split into horizontal bins. These partial features (horizontal bins) were finally learned by an attentive bi-directional GRU to exploit the relations between these parts of the embedding.

DAE+GAN. Recently, DAEs have been considered as the backbone of the generator and/or discriminator components in GANs for gait recognition [118], [119], [122], [130]. GaitGAN [118] and GaitGANv2 [119] used two discriminators with encoder-decoder structures, respectively for fake/real discrimination and identification. These two discriminators ensured that the generated gait images were realistic and that the generated images contained identity information. The Alpha-blending GAN (Ab-GAN) proposed in [122] exploits an encoder-decoder network as the generator to generate gait templates without carried objects. Cycle-consistent Attentive GAN (CA-GAN) was proposed in [130] and used an encoder-decoder structure for gait view synthesis. The proposed GAN contains two branches to simultaneously exploit both global and partial feature representations.

DAE+RNNs. The combination of DAEs and RNNs has recently been proposed for generating sequence-based disentangled features using an LSTM RNN [42], [108]. In this context, a deep encoder-decoder network with novel loss functions was first used to disentangle gait features, namely identity information from appearance and canonical features that mostly contain spurious information for gait recognition. A multi-layer LSTM was then used to capture temporal dynamics of the gait features to be finally aggregated for the recognition purpose [42], [108].

RNNs+CapsNets. Recurrently learned features obtained by RNNs can be treated as capsules [123], thus learning coupling weights between these capsules through dynamic routing. This encapsulated hierarchical part-whole relationships between the recurrently learned features that can make the hybrid network more robust against appearance and view changes. Additionally, the CapsNet can act as an attention mechanism, thus assigning more importance to the more relevant features. In [95], a CapsNet was used to treat the recurrently learned partial representations of a convolution template as capsules, thus learning the coupling weights between the partial features. This led to exploiting the relationships between the partial features while also preserving positional attributes. So, the model could generalize better to unseen gait viewpoints during testing. In [125], a capsule network with dynamic routing was used to exploit the spatial and structural relations between body parts. In this context, the recurrently learned features were first extracted using an LSTM network from a sequence of gait frames to feed the capsule network.

5 State-of-the-Art

In this section, we use our proposed taxonomy to survey deep gait recognition methods available in the literature. Table 3 summarizes the main characteristics of the existing deep solutions, sorted according to their publication dates. This table categorizes the available solutions based on the dimensions associated with the proposed taxonomy. This table also includes the loss functions that the methods have used for training as well as the datasets used for performance assessment. To better analyze the methods presented in Table 3, we also visualize them in Fig. 7a, where each entry represents a method categorized based on the three levels of our taxonomy as well as the publication year and month. We excluded the body representation dimension in this figure for better readability. The information about temporal representation, feature representation, and publication date are shown in the x, y, and z axes, respectively. Lastly, the color and marker symbol of each vertical line represent the neural architecture.

5.1 Analysis and Trends

Our analysis based on Table 3 and Fig. 7a allows us to reach some interesting conclusions about the recent evolution and trends in deep gait recognition technologies with respect to our proposed taxonomy. Following are the key ideas from the analysis.

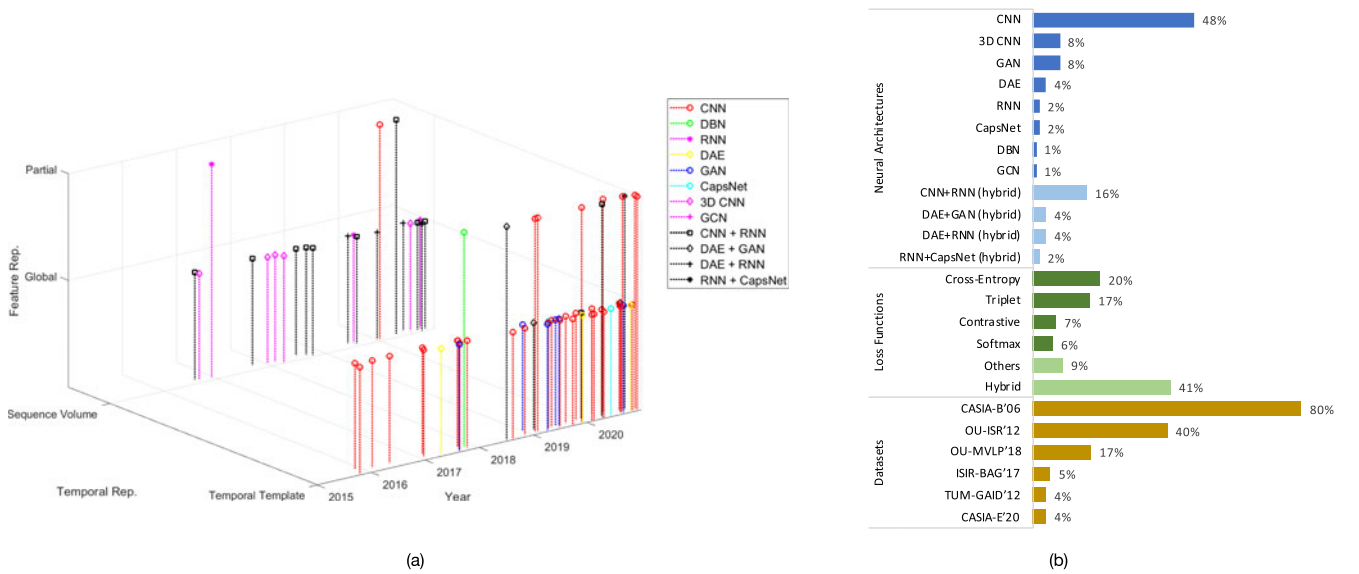
Body Representation. Silhouettes are the most widely adopted body representation for deep gait recognition, corresponding to over 81% of the literature. Skeletons have been considered less frequently compared to silhouettes, corresponding to only 13% of the available solutions. There have also been a few methods, i.e., approximately 5% of the available literature, that exploit both skeleton and silhouette representations, notably using disentangled representation learning or fusion strategies. Based on our analysis, high performing gait recognition methods such as [44], [45], [46], [47], [48], [95], [129], [157] have all adopted silhouette body representation. Nonetheless, due to recent advancements in effective pose-estimation techniques [74], [82], [84], [168] capable of extracting accurate and robust skeleton data from videos, we anticipate methods based on hybrid silhouettes-skeleton body representations to gain popularity in the near future.

Temporal Representation. Gait templates have been the most considered representation for capturing temporal gait information, corresponding to 70% of the proposed deep methods. Among different types of gait templates, GEI and set-pooling have been adopted the most. Around 30% of solutions adopt sequence volumes to preserve the order of available gait frames and to learn from their relationships. Given the frequent use of convolutional templates in some of the recent high-performing literature [43], [44], [45], [46], [95], we anticipate that these templates gain further popularity and surpass temporal templates in the future.

Feature Representation. Our analysis shows that over 87% of available methods are based on global feature representations,

TABLE 3
Classification of Deep Gait Recognition Methods Based on Our Proposed Taxonomy

Ref.	Year	Venue	Body Rep.	Temporal Rep.	Feat. Rep.	Neural Architecture	Loss Function	Dataset
[132]	2015	*T-MM*	Silhouettes	Tmp: GEI	Global	CNN	Cross-Entropy	CASIA-B
[35]	2015	*CISP*	Silhouettes	Tmp: GEI	Global	CNN	Undisclosed	CASIA-B
[36]	2016	*ICPR*	Skeleton	Sequence Volume	Partial	LSTM	Undisclosed	CASIA-B
[37]	2016	*ICB*	Silhouettes	Tmp: GEI	Global	CNN	Cross-Entropy	OU-ISIR
[127]	2016	*ICIP*	Silhouettes	Sequence Volume	Global	3D CNN	Undisclosed	CMU Mobo; USF HumanID
[133]	2016	*ICASSP*	Silhouettes	Tmp: GEI	Global	CNN	Contrastive	OU-ISIR
[92]	2016	*BMVC*	Skeleton	Sequence Volume	Global	CNN + LSTM	Cross-Entropy	CASIA-B; CASIA-A
[105]	2017	*Int. J. Biom.*	Silhouettes	Tmp: GEI	Partial	DBN	Undisclosed	CASIA-B
[134]	2017	*CVIU*	Silhouettes	Tmp: GEI	Global	CNN	Undisclosed	CASIA-B
[39]	2017	*IEEE T-PAMI*	Silhouettes	Tmp: GEI	Global	CNN	Cross-Entropy	CASIA-B; OU-ISIR
[135]	2017	*Applied Sci.*	Silhouettes	Tmp: Norm. AC	Global	CNN	Undisclosed	OU-ISIR
[136]	2017	*IEEE T-CSVT*	Silhouettes	Tmp: GEI	Global	CNN	Triplet Loss	OU-ISIR
[137]	2017	*BIOSIG*	Silhouettes	Tmp: GEI	Global	CNN	Undisclosed	TUM-GAID
[138]	2017	*MM*	Silhouettes	Tmp: GEI	Global	CNN	Triplet Loss	OU-ISIR
[91]	2017	*CCBR*	Skeleton	Sequence Volume	Global	CNN + LSTM	Undisclosed	CASIA-B
[119]	2017	*CVPRW*	Silhouettes	Tmp: GEI	Global	GAN	Cross-Entropy	CASIA-B
[115]	2017	*Neurocomp.*	Silhouettes	Tmp: GEI	Global	DAE	Euclidean	CASIA-B; SZU RGB-D
[93]	2018	*Elect. Imaging*	Silhouettes	Sequence Volume	Global	3D CNN	Undisclosed	CASIA-B
[129]	2018	*IEEE Access*	Silhouettes	Sequence Volume	Global	CNN + LSTM	Cross-Entropy	CASIA-C
[139]	2018	*Neuroinform.*	Silhouettes	Seq. Vol. + GEI	Global	3D CNN	Contrastive	OU-ISIR
[40]	2018	*DIC*	Skeleton	Tmp: GEI	Global	CNN	Cross-Entropy	CASIA-B
[140]	2018	*IEEE Access*	Silhouettes	Sequence Volume	Global	CNN + LSTM	Cross-Entropy	CASIA-B; OU-ISIR
[141]	2018	*ISBA*	Silhouettes	Sequence Volume	Global	3D CNN	Undisclosed	CASIA-B
[131]	2018	*ICME*	Silhouettes	Tmp: GEI	Part; Glob.	DAE + GAN	Cross-Entropy	CASIA-B
[142]	2018	*JVCIR*	Silhouettes	Tmp: GEI	Global	CNN	Cross-Entropy	CASIA-B; OU-ISIR
[143]	2018	*CCBR*	Skeleton	Sequence Volume	Global	CNN + LSTM	Undisclosed	CASIA-B
[144]	2019	*PRL*	Skel.; Silh.	Sequence Volume	Global	LSTM	Undisclosed	CASIA-B; TUM-GAID
[41]	2019	*IET Biom.*	Silhouettes	Tmp: Weight Avg.	Partial	CNN	Undisclosed	CASIA-B; TUM; OU-ISIR
[42]	2019	*CVPR*	Skel.; Silh.	Sequence Volume	Global	DAE + LSTM	Multiple Loss Functi ons	CASIA-B; FVG
[145]	2019	*J. Sys. Arch.*	Silhouettes	Tmp: GEI	Global	DAE + GAN	Multiple Loss Functions	CASIA-B; OU-ISIR
[113]	2019	*PR*	Silhouettes	Tmp: GEI	Global	CNN	Siamese	CASIA-B; SZU
[90]	2019	*IEEE T-IFS*	Silhouettes	Tmp: GEI	Global	GAN	Adversarial&Cross-Entropy	CASIA-B; OU-ISIR
[146]	2019	*PRL*	Silhouettes	Tmp: GEI	Global	CNN	Restrictive Triplet	CASIA-B; OU-ISIR
[147]	2019	*CVPR*	Silhouettes	Tmp: GEI	Global	CNN	Quintuplet	CASIA-B; OU-ISIR LP Bag
[122]	2019	*Neurocomp.*	Silhouettes	Tmp: GEI	Global	GAN	Pixel-wise and Entropy	CASIA-B; OU-ISIR
[148]	2019	*IJCNN*	Silhouettes	Tmp: GEI	Global	GAN	Multiple Loss Functions	CASIA-B
[114]	2019	*IEEE T-IFS*	Skeleton	Tmp: GEI	Global	DAE	Contrastive&Triplet Loss	OU-ISIR LP Bag; TUM-GAID
[80]	2019	*ICVIP*	Silhouettes	Tmp: Weight Avg.	Partial	CNN	View&Cross-Entropy	CASIA-B
[149]	2019	*IEEE T-MM*	Silhouettes	Sequence Volume	Global	CNN + LSTM	Contrastive	CASIA-B; OU-ISIR
[150]	2019	*IJCNN*	Silhouettes	Tmp: GEI	Global	GAN	Multiple Loss Functions	CASIA-B
[151]	2019	*NCAA*	Silhouettes	Tmp: GEI	Global	CNN	Undisclosed	CASIA-B; CASIA-A; OU-ISIR
[69]	2019	*PR*	Silhouettes	Tmp: Set pooling	Global	CNN	Center&Soft-Max	CASIA-B
[152]	2019	*NCAA*	Silhouettes	Tmp: GEI	Global	CNN	Undisclosed	CASIA-B; OU-ISIR
[153]	2019	*JVCI*	Silhouettes	Tmp: GEI	Global	CapsNet	Standard Capsule Loss	CASIA-B
[43]	2019	*AAAI*	Silhouettes	Tmp: Set Pooling	Partial	CNN	Batch All Triplet loss	CASIA-B; OU-MVLP
[154]	2020	*IEEE Access*	Skeleton	Sequence Volume	Global	DAE + LSTM	Mean Square Error	Walking Gait
[155]	2020	*PR*	Skeleton	Sequence Volume	Partial	CNN	Center&Soft-Max	CASIA-B; CASIA-E
[109]	2020	*IEEE T-PAMI*	Skel.; Silh.	Sequence Volume	Global	DAE + LSTM	Multiple Loss Functions	CASIA-B; FVG
[130]	2020	*IEEE T-IP*	Silhouettes	Sequence Volume	Partial	CNN + LSTM	Angle Center	CASIA-B; OU-MVLP; OU-LP
[123]	2020	*PR*	Silhouettes	Tmp: GEI	Global	GAN	Multiple Loss Functions	OULP-BAG; OU-ISIR LP Bag
[156]	2020	*MTAP*	Silhouettes	Tmp: MF-GEI	Global	CNN	Undisclosed	CASIA-B
[126]	2020	*KBS*	Silhouettes	Sequence Volume	Global	LSTM + Capsule	Capsule&Memory	CASIA-B; OU-MVLP
[157]	2020	*JINS*	Silhouettes	Tmp: GEI	Global	CNN + LSTM	Undisclosed	CASIA-B; OU-ISIR
[158]	2020	*IEEE T-CSVT*	Silhouettes	Tmp: GEI	Global	CNN	Contrastive&Triplet Loss	CASIA-B; OU-MVLP; OU-ISIR
[94]	2020	*arXiv*	Silhouettes	Sequence Volume	Global	3D CNN	Triplet Loss	CASIA-B; OU-MVLP
[159]	2020	*MTAP*	Silhouettes	Tmp: GEI	Partial	CNN	Undisclosed	CASIA-B; OU-ISIR
[95]	2020	*arXiv*	Skeleton	Sequence Volume	Global	GCN	Triplet Loss&ArcFace	CASIA-B
[160]	2020	*MTAP*	Silhouettes	Tmp: GEI	Global	CNN	Undisclosed	CASIA-B; OU-ISIR
[161]	2020	*JIPS*	Silhouettes	Tmp: GEI	Global	CNN	Soft-Max	CASIA-B; OU-ISIR
[162]	2020	*MTAP*	Silhouettes	Tmp: GEI	Global	CNN	Undisclosed	CASIA-B
[163]	2020	*J. SuperComp.*	Silhouettes	Tmp: GEI	Global	CNN	Undisclosed	CASIA-B; OU-ISIR; OU-MVLP
[153]	2020	*ITNEC*	Silhouettes	Tmp: GEI	Global	CapsNet	Standard Capsule Loss	CASIA-B; OU-ISIR
[45]	2020	*CVPR*	Silhouettes	Tmp: Hor. Pooling	Partial	CNN	Batch All Triplet loss	CASIA-B; OU-MVLP
[116]	2020	*CVPR*	Silhouettes	Tmp: GEI	Global	DAE	Contrastive&Triplet Loss	CASIA-B; OU-ISIR LP Bag
[71]	2020	*IEEE T-Biom*	Skeleton	Sequence Volume	Global	CNN + LSTM	Cross-Entropy&Center	OUMVLP-Pose
[164]	2020	*ACCVW*	Silhouettes	Tmp: Set Pooling	Partial	CNN	Batch All Triplet loss	CASIA-E
[165]	2020	*ACCVW*	Silhouettes	Tmp: Set Pooling	Partial	CNN	Batch All Triplet loss	CASIA-E
[96]	2020	*ICPR*	Silhouettes	Tmp: Set Pooling	Partial	CNN + GRU + Caps.	Triplet Loss&Cosine Prox.	CASIA-B; OU-MVLP
[44]	2020	*IEEE T-Biom.*	Silhouettes	Tmp: GCEM	Partial	CNN + GRU	Triplet Loss&Cross-Entropy	CASIA-B; OU-MVLP
[166]	2020	*IEEE Access*	Silhouettes	Tmp: Set Pooling	Global	CNN	Triplet Loss	CASIA-B
[167]	2019	*ICASSP*	Silhouettes	Tmp: Pooling	Global	CNN	Center-Ranked	CASIA-B; OU-MVLP
[168]	2020	*IJCB*	Silhouettes	Tmp: GEI	Global	DAE + GAN	Center&Soft-Max	CASIA-B; OU-ISIR
[47]	2020	*ACCV*	Skel.; Silh.	Sequence Volume	Global	CNN + LSTM	Multiple Loss Functions	CASIA-B; OU-MVLP
[48]	2020	*MM*	Silhouettes	Sequence Volume	Global	3D CNN	Multiple Triplet Losses	CASIA-B; OU-ISIR
[46]	2020	*ECCV*	Silhouettes	Tmp: Set Pooling	Global	CNN	Triplet Loss&Cross-Entropy	CASIA-B; OU-MVLP



(a) (b)

Fig. 7. (a) Visualization of deep gait recognition methods, according to three levels of our taxonomy and publication date; (b) The frequency of different neural architectures, loss functions, and gait datasets used in the literature.

where deep features are learned by considering gait information as a whole. Recently, many interesting and high performing methods [43], [44], [45], [95] have adopted partial representations by splitting the gait data into local regions. The performance of such techniques points to promising potential in partial representation learning for discriminating key gait features. Hence, we anticipate further research with convincing results in this area.

Neural Architectures. As presented in Fig. 7b, 2D CNNs are the most widely used DNN type for deep gait recognition with 48% of the published solutions utilizing only 2D CNN architectures for classification. 3D CNNs and GANs are the next popular categories, each corresponding to 8% of the literature. DAEs, RNNs, CapsNets, DBNs, and GCNs are less considered among DNNs, respectively corresponding to 4%, 2%, 2%, 1%, and 1% of the methods. Concerning hybrid methods which constitute 26% of the published solutions, CNN-RNN combinations are the most widely adapted approach with 16% share, while the combination of DAEs with GANs and RNNs corresponds to 8% of the methods, followed by RNN-CapsNet methods that make up 2% of the solutions. We expect that hybrid methods that make use of two or more types of DNN attract more attention in the near future and demonstrate robust performance in the field.

Loss Functions. Loss functions calculate a model's error during training and should ideally be designed to efficiently capture the properties of the problem for facilitating an effective training process [72]. Fig. 7b shows the usage frequency of different well-known loss functions that have been used by deep gait recognition literature. Among the single loss functions, cross-entropy [169] has been the most widely adopted with 20% of solutions having used it. This loss function takes the output probabilities of the predicated classes and makes the model output as close as possible to the ground-truth output. Triplet loss [170] is the next popular type with a usage frequency of 17%. This loss has been notably used by some of the most recent and state-of-the-art solutions [43], [44], [45], [48], [95], [115] and [46]. This loss function compares a baseline input, also known as *anchor*, to a *positive* sample with the same identity, and a *negative* sample with a different identity. The loss function then ensures that the dissimilarity between two feature vectors belonging to the same subject is lower than that between feature vectors belonging to two different subjects. Contrastive loss [171] corresponds to 7% of the recognition methods, and uses pairs of samples including anchor-neighbor or anchor-distant. If the pair of samples is anchor-neighbor, the loss function minimizes their distance; otherwise, it increases their distance. Th next popular loss function, corresponding to 6% of the recognition methods, is based on the softmax loss [172]. There have also been some other loss functions, such as arcface [173], center loss [174], and euclidean loss [175] with a combined usage frequency of 9% that have been less considered for gait recognition. Finally, there have been two classes of deep gait recognition methods that use multiple loss functions (usage frequency of 41%), including (*i*) methods such as [46], [94] that add together two or more loss functions to complement each other and compensate their weaknesses; and (*ii*) methods that have been designed based on networks with multiple components, such GANs with generators and discriminators [89], [121], [122], [147] and hybrid networks [42], [46], [47], [94], where different loss functions have been used to train different components. We expect that deep gait recognition methods based on multiple losses attract more attention and surpass other approaches in the near future.

Datasets. We tally the number of times each dataset has been used by the published literature and present the results in Fig. 7b. Fig. 7b does not include the datasets that have appeared less than 3 times in Table 3. In addition, a point to consider is that many of the literature use more than one dataset to perform the experiments. We observe that CASIA-B [32] is the most widely used dataset, appearing in 80% of the published literature, as it provides a large number of samples with variations in carrying and wearing conditions. OU-ISIR [65] was the largest gait dataset prior to 2018; we therefore found OU-ISIR to be the second most popular dataset having been used by 40% of the solutions.

TABLE 4
State-of-the-Art Results on CASIA-B Dataset [32]

Method			Performance			
Reference	Year	Venue	NM	BG	CL	Average
[131]	2015	*IEEE T-MM*	78.9	—	—	—
[104]	2017	*Int. J. Biom.*	90.8	45.9	45.3	60.7
[39]	2017	*IEEE T-PAMI*	94.1	72.4	54.0	73.5
[40]	2018	*DIC*	83.3	—	62.5	—
[69]	2019	*PR*	75.0	—	—	—
[89]	2019	*IEEE T-IFS*	79.8	—	—	—
[112]	2019	*PR*	89.9	—	—	—
[42]	2019	*CVPR*	93.9	82.6	63.2	79.9
[111]	2019	*CVPR*	89.9	—	—	—
[41]	2019	*IET Biom.*	94.5	78.6	51.6	74.9
[143]	2019	*PRL*	86.1	—	—	—
[43]	2019	*AAAI*	95.0	87.2	70.4	84.2
[108]	2020	*IEEE T-PAMI*	92.3	88.9	62.3	81.2
[129]	2020	*IEEE T-IP*	96.0	—	—	—
[157]	2020	*IEEE T-CSVT*	92.7	—	—	—
[165]	2020	*IEEE Access*	95.1	87.9	74.0	85.7
[95]	2020	*ICPR*	95.7	90.7	72.4	86.3
[44]	2020	*IEEE T-Biom.*	95.2	89.7	74.7	86.5
[45]	2020	*CVPR*	96.2	91.5	78.7	88.8
[115]	2020	*CVPR*	94.5	—	—	—
[46]	2020	ECCV	96.8	**94.0**	77.5	89.4
[47]	2020	*ACCV*	**97.9**	93.1	77.6	89.5
[48]	2020	*MM*	96.7	93.0	**81.5**	**90.4**

NM, BG, and CL are respectively normal walking, walking with a bag, and walking with a coat test protocols.

Since the introduction of OU-MVLP [68] in 2018, this dataset has been receiving considerable attention from the community and has been used by 18% of the methods in a span of only 2 years. The OU-ISIR LP Bag dataset [67] only consists of gait data with carried objects, so naturally it was only considered when designing solutions for specific applications such as those intended to be invariant to carrying conditions from a single viewpoint. As a result, this dataset was used for evaluation purposes by only 5% of methods. TUM GAID [66] has also been less considered by the community, corresponding to 5% of published literature. Finally, CASIA-E [69] which was developed in 2020 is the sixth most widely used, appearing in 4% of the literature. However, we anticipate that this dataset will become the standard benchmark dataset for gait recognition in the near future, due to the fact that it provides hundreds of video sequences per each subject with high variations in appearance and acquisition environments.

5.2 Performance Comparison

To shed more light on the performance of deep gait recognition methods, we summarize the performance of the methods tested on the three most popular gait datasets, namely CASIA-B [32], OU-ISIR [65], and OU-MVLP [68] datasets in Tables 4, 5, and 6 respectively. To perform a fair comparison, these tables only include methods that followed the standard test protocols designed for these datasets, as discussed in Section 3.2. The results show that the method proposed in [48] currently provides the best recognition results on CASIA-B (average performance result of 90.4%) and OU-ISIR (performance result of 99.9%). Concerning the OU-MVLP dataset, results show the superiority of the method proposed in [46] (performance result of 89.18%) over other methods. Apart from [48] and [46], there are several other methods including those proposed in [44], [45], [47], [95], [129], [157], whose performances are near the state-of-the-art for these datasets. Our analysis shows that some of these best performing methods, including [44], [47], [95], [129], make use of two or more types of neural architectures to boost the performance. Some other methods including [45], [46], [48], [157] use multiple loss functions to complement each other and compensate their weaknesses to boost performance. This analysis reveals the effectiveness of hybrid approaches, in term of either neural architectures as well as loss functions, for achieving strong performance in the area.

TABLE 5
State-of-the-Art Results on OU-ISIR [65] Dataset

Method			
Reference	Year	Venue	Performance
[132]	2016	*ICASSP*	90.71
[39]	2017	*IEEE T-PAMI*	92.77
[134]	2017	*Applied Sci.*	91.25
[138]	2018	*Neuroinform.*	88.56
[139]	2018	*IEEE Access*	95.67
[41]	2019	*IET Biom.*	97.40
[89]	2019	*IEEE T-IFS*	93.20
[145]	2019	*PRL*	94.62
[148]	2019	*IEEE T-MM*	97.26
[167]	2020	*IJCB*	94.17
[157]	2020	*IEEE T-CSVT*	98.93
[129]	2020	*IEEE T-IP*	99.27
[48]	2020	*MM*	**99.90**

6 Vulnerability to Adversarial Attacks

Traditional spoofing attacks to gait recognition systems were attempted by trained impostors, notably with similar body shapes and clothes, imitating the walking styles of target subjects [24], [176], [177]. However, these attacks are rather hard and limited as they require trained and qualified imitators. Different from the traditional spoofing attacks, *adversarial attacks* to gait recognition systems have been designed to imperceptibly fool recognition systems by synthesizing input gait videos with both good quantitative similarity and visual realism.

Despite the strong performance of deep learning solutions in computer vision, such solutions have been

TABLE 6
State-of-the-Art Results on OU-MVLP [68] Dataset

Method			
Reference	Year	Venue	Performance
[43]	2019	*AAAI*	83.40
[166]	2020	*ICASSP*	57.80
[157]	2020	*IEEE T-CSVT*	63.10
[129]	2020	*IEEE T-IP*	84.60
[95]	2020	*ICPR*	84.50
[44]	2020	*IEEE T-Biom.*	84.30
[45]	2020	*CVPR*	88.70
[46]	2020	*ECCV*	**89.18**

surprisingly vulnerable to adversarial attacks [178], [179]. These attacks introduce perturbations in visual content that can manipulate the predictions of deep models by resulting in embeddings capable of fooling the classifiers [180]. Since the introduction of the first adversarial attacks to deep neural networks in 2014 [181], these models have attracted significant attention from the research community in computer vision and pattern analysis. Among the adversarial attack approaches, GANs [117], [182], [183], [184] have been one of the most powerful methods for image and video manipulation.

The first attempt to investigates the vulnerability of gait recognition systems to adversarial attacks was done in [185] using a GAN for synthesizing gait images. In this context, the foreground of each frame of the source video is first segmented to be then fed to the generator along with the target background. The generator is composed of two parallel encoder-decoder networks, respectively dealing with foreground and background information. The corresponding feature representations from these two networks are fused at multiple scales. Static and dynamic silhouette-based losses have been designed in order to force the model to generate more realistic results for gait recognition. Additionally, triplet loss is used to preserve the similarity between the individuals in the source and generated videos. The performance of two state-of-the-art gait recognition systems, including CNNGait [39] and GaitSet [43], were evaluated using the generated gait samples. The results showed that the generated samples provide sufficient discriminative information to bypass gait recognition systems. To show how the sequence-based gait recognition is vulnerable to adversarial attacks, another study was done in [186]. In this study, a novel temporal sparse adversarial attack method was proposed based on a GAN to synthesize high-quality sequences of silhouette frames. To ensure imperceptibility of the proposed method, a few adversarial gait silhouettes are substituted or inserted in the sequence. Experimental results show that the state-of-the-art GaitSet [43] method has low robustness to this adversarial attack.

The presented results in [185], [186] suggest that adversarial attacks can surprisingly degrade the performance of deep gait recognition methods, thus posing a real threat to such recognition systems. This demonstrates the necessity of adopting efficient countermeasure techniques against adversarial attacks aimed towards deep gait recognition systems.

7 Challenges and Future Research Directions

Despite the great success of gait recognition using deep learning techniques, there still remains a large number of challenges that need to be addressed in the area. Here, we further point out a few promising future research directions and open problems in this domain. These directions may facilitate future research activities and foster real-world applications.

7.1 Disentanglement

Complex gait data arise from the interaction between many factors such as occlusion, camera view-points, appearance of individuals, sequence order, body part motion, or lighting sources present in the data [42], [108], [115]. These factors can interact in complex manners that can complicate the recognition task. There have recently been a growing number of methods in other research areas, such as face recognition [187], [188], action recognition [189], emotion recognition [190], and pose estimation [191], that focus on learning disentangled features by extracting representations that separate the various explanatory factors in the high-dimensional space of the data [192], [193], [194]. However, the majority of available deep gait recognition methods have not explored disentanglement approaches, and hence are not explicitly able to separate the underlying structure of gait data in the form of meaningful disjoint variables. Despite the recent progress in using disentanglement approaches in a few gait recognition methods [42], [108], [115], there is still room for improvement. To foster further progress in this area, the adaptation of novel generative models [195], [196] and loss functions [197] can be considered to learn more discriminative gait representations by explicitly disentangling identity and non-identity components.

7.2 Self-Supervised Learning

A considerable majority of available deep gait recognition methods follow the supervised learning paradigm, and thus require labeled data during training. Nevertheless, in real-world applications, labeled data may not always be readily available and labels are generally expensive and time-consuming to obtain. In order to utilize unlabeled gait data to learn more efficient and generalizable gait representations, self-supervised learning [198] can be exploited. In this context, general and rich high-level semantics can be captured without using any annotated labels. Self-supervised approaches can define various pretext tasks, such as body part motion or sequence order recognition for input sequences [199], [200], to be solved by a network. Through learning these pretext tasks, the network can then learn generic features. The network trained with the generated pre-text labels can then be fine-tuned with the actual labels in order to recognize the identity. Among self-supervised approaches, contrastive learning methods [201], including SimCLR [202], are promising approaches that learn representations by defining an *anchor* and a *positive* sample in the feature space, and then aim to make the anchor separable from the *negative* samples. One important challenge in using self-supervised learning in the context of gait recognition is to design effective pretext tasks to ensure the network can learn meaningful representations. Additionally, joint learning of several pretext tasks in a network [198], instead of a single pretext task, notably using several loss functions [203], can provide the network with more representative features [204], [205], [206]. We expect these challenges to gain increased popularity in the context of deep gait recognition in the near future.

7.3 Multi-Task Learning

Multi-task learning is generally performed to simultaneously learn multiple tasks using a shared model, thus learning more generalized and often reinforced representations [207], [208], [209]. In many cases, these approaches offer advantages such as increased convergence speed, improved learning by leveraging auxiliary information, and

reduced overfitting through shared representations [207], [208]. Despite the effectiveness of multi-task learning in a number of other domains [210], [211], most deep gait recognition solutions in the literature focus on the single task of identification. Thus, most existing works learn features that are sensitive to identity without considering interactions with other latent factors, such as affective states, gender, and age [35], [89], [212]. In this context, simultaneous learning of multiple tasks for gait recognition may present new design paradigms and optimization challenges, notably in terms of task identification and loss functions [213]. We expect these challenges to attract further attention in the near future and be tackled in the context of gait recognition with multi-task learning.

7.4 Data Synthesis and Domain Adaptation

Deep gait recognition methods require large amounts of data for effective training and reliable evaluation. This issue is evident in Fig. 7b where most of the deep gait recognition solutions [44], [45], [46], [47], [48], [95], [129], [157] used large-scale gait datasets for instance CASIA-B [32], OU-ISIR [65], and OU-MVLP [68]. In the context of deep gait recognition, data synthesis, for instance using GANs [214], [215], can be considered for creating large datasets or data augmentation [3], [216], [217]. Furthermore, developing synthesized datasets can also be advantageous in that subject privacy concerns could be alleviated with fake subject data. Similar approaches have been carried out for the more *privacy-sensitive* area of facial recognition [218], [219], where large datasets comprised only of fake data have been developed to be used in deep learning research [220], [221], [222]. In addition, such approaches can be used to increase the variance of existing datasets. For instance, large-scale gait datasets such as OU-ISIR [65] and OU-MVLP [68] only provide normal walking sequences with no variations in occlusion or carrying and clothing conditions. Thus, solutions trained on these datasets usually fail to generalize well when facing variations in appearance and environment during the testing phase. Here, domain adaptation [223], [224], [225], [226], [227] is a potential remedy for this problem that can modify existing datasets to include the desired variations, thus eliminating the necessity for collecting new data. Furthermore, gait synthesis can be performed for computer animation [228] and by game engines [229] to generate large-scale synthetic gait datasets. Hence, we anticipate that with advances in gait data synthesis and domain adaptation techniques, more complementary gait datasets will be constructed to enable the development of more robust solutions.

7.5 Cross-Dataset Evaluation

The practical value of gait recognition systems is strongly dependant in its ability to generalize to unseen data. To the best of our knowledge, cross-dataset gait recognition on well-known datasets such as CASIA-B [32], OU-ISIR dataset [65], and OU-MVLP [68], has not been performed in the literature as notable solutions available in the literature all use the same gait dataset for both training and testing. However, in many real applications such as deployed products, test or run-time data are often obtained in a variety of different conditions with respect to the training data. In order to examine the generalizability of gait recognition systems in real-world applications, cross-dataset evaluations should be adopted, for example using transfer learning techniques [230]. In this context, a solution trained on one dataset can be used to extract features from the test data (gallery and probe sets) of another dataset. The extracted features can then feed a classifier to perform gait recognition. Cross-dataset gait recognition can potentially be formulated as an out-of-distribution (OOD) testing problem, where the generalization ability of a deep model beyond the biases of the training set is evaluated [231]. We expect that OOD tests [232] become increasingly popular for evaluating the generalization ability of gait recognition methods.

7.6 Multi-View Recognition

A large number of gait datasets contain multi-view sequences, providing gait information captured from different view-points. Most of the current methods available in the literature only perform single-view gait recognition. These methods generally learn intra-view relationships and ignore inter-view information between multiple viewpoints. By casting the problem as multi-view, descriptors such as gate-level fusion LSTM [233], state-level fusion LSTM [233], spatio-temporal LSTM [234], multi-perspective LSTM [235], and multi-view LSTM [236], can be adopted to jointly learn both the intra-view and inter-view relationships. Another challenge in multi-view gait recognition is that most existing multi-view descriptors consider a well defined camera network topology with fixed camera positions. However, data collection in real-world environments is often uncontrollable, i.e., data might be captured from unpredictable viewing angles [237], [238] or even from moving cameras [239]. To this end, existing multi-view methods, which mostly rely on pre-trained descriptors, fail to bridge the domain gap between the training and run-time multi-view data. We expect that future research direction in this area will be shaped by proposing novel approaches, for example using clustering algorithms [240], combinatorial optimization [241], and self-supervised learning [242], for adopting generic gait descriptors for multi-view geometry.

7.7 Multi-Biometric Recognition

Some literature in the field have fused gait information with other biometric information such as face [19], [72] and ear [20], [243], [244], which can be obtained from high-quality gait videos. As we discussed earlier, gait recognition systems are generally challenged when facing variations in subject appearance and clothing, camera view-points, and body occlusions. On the other hand, additional sources of biometric information, notably face and ear, are less sensitive to some of these challenging factors. Instead, face and ear recognition systems can be negatively affected by some other factors such as low image quality, for instance blurred or low resolution images, varying lighting, or facial occlusions, which in turn have limited impact on the performance of gait recognition systems. Hence, various biometric modalities and gait can complement one another to compensate each others' weaknesses in the context of a multi-biometric system [245], [246]. Apart from the complementary (hard-)biometric traits, soft-biometric traits such as

age [247], height [248], [249], weight [250], gender [251], and particular body marks including tattoos [252] can also be included to boost overall performance. The combination of other soft- and hard- biometric traits with gait has mostly been done in the literature based on non-deep methods [253], [254], [255], [256], [257], [258], while multi-modal deep learning methods [259], [260], notably based on fusion [261], joint learning [233], and attention [262] networks, can also be adopted. Hence, we anticipate that research on deep multi-biometric recognition systems that include gait, gain popularity in the coming years.

8 Summary

We provided a survey of deep gait recognition methods that was driven by a novel taxonomy with four dimensions namely body representation, temporal representation, feature representation, and neural architectures. Following our taxonomy, we reviewed the most representative deep gait recognition methods and provided discussions on their characteristics, advantages, and limitations. We additionally reviewed the most commonly used datasets along with their evaluation protocols and corresponding performance results reported in the literature. We finally concluded this survey with a discussion on current challenges, pointing out a few promising future research directions in this domain. We expect that this survey provides insights into the technological landscape of gait recognition for guiding researchers in advancing future research.

Acknowledgments

The authors would like to thank the BMO Bank of Montreal and Mitacs for funding this research.

References

[1] F. Deligianni, Y. Guo, and G.-Z. Yang, "From emotions to mood disorders: A survey on gait analysis methodology," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 6, pp. 2302–2316, Nov. 2019.

[2] S. A. Etemad and A. Arya, "Expert-driven perceptual features for modeling style and affect in human motion," *IEEE Trans. Human-Mach. Syst.*, vol. 46, no. 4, pp. 534–545, Aug. 2016.

[3] A. Etemad and A. Arya, "Classification and translation of style and affect in human motion using RBF neural networks," *Neurocomputing*, vol. 129, no. 1, pp. 585–595, Apr. 2014.

[4] A. Etemad and A. Arya, "Correlation-optimized time warping for motion," *Vis. Comput.*, vol. 31, no. 1, pp. 1569–1586, Oct. 2015.

[5] J. M. Echterhoff, J. Haladjian, and B. Brügge, "Gait and jump classification in modern equestrian sports," in *Proc. ACM Int. Symp. Wearable Comput.*, 2018, pp. 88–91.

[6] H. Zhang, Y. Guo, and D. Zanotto, "Accurate ambulatory gait analysis in walking and running using machine learning models," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 1, pp. 191–202, Jan. 2020.

[7] T. T. Verlekar, P. Lobato Correia, and L. D. Soares, "Using transfer learning for classification of gait pathologies," in *Proc. Int. Conf. Bioinf. Biomed.*, 2019, pp. 1–6.

[8] A. Muro-de-la Herran, B. Garcia-Zapirain, and A. Mendez-Zorrilla, "Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications," *Sensors*, vol. 14, no. 2, pp. 3362–3394, Feb. 2014.

[9] D. Jarchi, J. Pope, T. K. M. Lee, L. Tamjidi, A. Mirzaei, and S. Sanei, "A review on accelerometry-based gait analysis and emerging clinical applications," *IEEE Rev. Biomed. Eng.*, vol. 11, no. 1, pp. 177–194, Feb. 2018.

[10] C. Wan, L. Wang, and V. V. Phoha, "A survey on gait recognition," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–35, Aug. 2018.

[11] I. Rida, N. Almaadeed, and S. Almaadeed, "Robust gait recognition: A comprehensive survey," *IET Biometrics*, vol. 8, no. 1, pp. 14–28, Jan. 2019.

[12] A. Nambiar, A. Bernardino, and J. C. Nascimento, "Gait-based person re-identification: A survey," *ACM Comput. Surv.*, vol. 52, no. 2, pp. 1–34, Apr. 2019.

[13] M. D. Marsico and A. Mecca, "A survey on gait recognition via wearable sensors," *ACM Comput. Surv.*, vol. 52, no. 4, pp. 1–39, Sep. 2019.

[14] J. P. Singh, S. Jain, S. Arora, and U. P. Singh, "Vision-based gait recognition: A survey," *IEEE Access*, vol. 6, pp. 70 497–70 527, Nov. 2018.

[15] 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, Aug. 2009.

[16] M. Z. Uddin, D. Muramatsu, N. Takemura, M. A. R. Ahad, and Y. Yagi, "Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion," *IPSJ Trans. Comput. Vis. Appl.*, vol. 11, no. 1, Nov. 2019, Art. no. 9.

[17] A. Ibrahim, W.-N. Mohd-Isa, and C. C. Ho, "Background subtraction on gait videos containing illumination variates," in *Proc. AIP Conf. Proc.*, 2018, Art. no. 020057.

[18] T. Verlekar, L. Soares, and P. Correia, "Gait recognition in the wild using shadow silhouettes," *Image Vis. Comput.*, vol. 76, no. 1, pp. 1–13, Aug. 2018.

[19] A. Sepas-Moghaddam, F. Pereira, and P. Correia, "Face recognition: A novel multi-level taxonomy based survey," *IET Biometrics*, vol. 9, no. 2, pp. 1–12, Mar. 2020.

[20] Ž. Emeršič, V. Štruc, and P. Peer, "Ear recognition: More than a survey," *Neurocomputing*, vol. 255, pp. 26–39, Sep. 2017.

[21] K. Nguyen, C. Fookes, R. Jillela, S. Sridharan, and A. Ross, "Long range iris recognition: A survey," *Pattern Recognit.*, vol. 72, no. 1, pp. 123–143, Dec. 2017.

[22] K. Nguyen, C. Fookes, R. Jillela, S. Sridharan, and A. Ross, "Long range iris recognition: A survey," *Pattern Recognit.*, vol. 72, pp. 123–143, Dec. 2017.

[23] K. Cao and A. K. Jain, "Automated latent fingerprint recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 788–800, Apr. 2019.

[24] A. Hadid, M. Ghahramani, V. Kellokumpu, M. Pietikäinen, J. Bustard, and M. Nixon, "Can gait biometrics be spoofed?," in *Proc. Int. Conf. Pattern Recognit.*, 2012, pp. 3280–3283.

[25] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 26, 2021, doi: 10.1109/TPAMI.2021.3054775.

[26] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–40, Jul. 2021.

[27] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: What features are important?," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 391–401.

[28] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019.

[29] J. L. Helbostad, S. Leirfall, R. Moe-Nilssen, and O. Sletvold, "Physical fatigue affects gait characteristics in older persons," *J. Gerontol. Ser. A: Biol. Sci. Med. Sci.*, vol. 62, no. 9, pp. 1010–1015, Sep. 2007.

[30] M. Karg, K. Kühnlenz, and M. Buss, "Recognition of affect based on gait patterns," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 1050–1061, Aug. 2010.

[31] I. Klöpfer-Krämer, A. Brand, H. Wackerle, J. Müßig, I. Kröger, and P. Augat, "Gait analysis–available platforms for outcome assessment," *Injury*, vol. 51, pp. S90–S96, May 2020.

[32] 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. Int. Conf. Pattern Recognit.*, 2006, pp. 441–444.

[33] D. Cunado, M. Nixon, and J. Carter, "Using gait as a biometric, via phase-weighted magnitude spectra," in *Proc. Int. Conf. Audio-Video-Based Biometric Pers. Authentication*, 1997, pp. 95–102.

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

[35] C. Yan, B. Zhang, and F. Coenen, "Multi-attributes gait identification by convolutional neural networks," in *Proc. Int. Congr. Image Signal Process.*, 2015, pp. 642–647.

[36] Y. Feng, Y. Li, and J. Luo, "Learning effective gait features using LSTM," in *Proc. Int. Conf. Pattern Recognit.*, 2016, pp. 325–330.

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

[38] B. M. Nair and K. D. Kendricks, "Deep network for analyzing gait patterns in low resolution video towards threat identification," *Electron. Imag.*, vol. 2016, no. 11, pp. 1–8, Feb. 2016.

[39] 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.

[40] 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. Digit. Image Comput., Techn. Appl.*, 2018, pp. 1–7.

[41] A. Sokolova and A. Konushin, "Pose-based deep gait recognition," *IET Biometrics*, vol. 8, no. 2, pp. 134–143, Feb. 2019.

[42] Z. Zhang *et al.*, "Gait recognition via disentangled representation learning," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, 2019, pp. 4705–4714.

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

[44] A. Sepas-Moghaddam and A. Etemad, "View-invariant gait recognition with attentive recurrent learning of partial representations," *IEEE Trans. Biometrics Behav. Identity Sci.*, vol. 3, no. 1, pp. 124–137, Jan. 2021.

[45] C. Fan *et al.*, "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, 2020, pp. 14213–14221.

[46] 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.

[47] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren, "End-to-end model-based gait recognition," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 1–17.

[48] B. Lin, S. Zhang, and F. Bao, "Gait recognition with multiple-temporal-scale 3D convolutional neural network," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 3054–3062.

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

[50] M. J. Nordin and A. Saadoon, "A survey of gait recognition based on skeleton model for human identification," *Res. J. Appl. Sci. Eng. Technol.*, vol. 12, no. 7, pp. 756–763, Apr. 2016.

[51] J. P. Singh, S. Jain, S. Arora, and U. P. Singh, "A survey of behavioral biometric gait recognition: Current success and future perspectives," *Arch. Comput. Methods Eng.*, vol. 28, no. 1, pp. 107–148, Nov. 2019.

[52] Google scholar, 2022. [Online]. Available: http://scholar.google.com/

[53] IEEE Xplore digital library, 2022. [Online]. Available: https://ieeexplore.ieee.org/

[54] ACM digital library, 2022. [Online]. Available: https://www.dl.acm.org/

[55] ScienceDirect digital library, 2022. [Online]. Available: https://www.sciencedirect.com/

[56] Computer vision foundation (CVF) open access, 2022. [Online]. Available: https://openaccess.thecvf.com/

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

[58] J. D. Shutler, M. G. Grant, M. S. Nixon, and J. N. Carter, "On a large sequence-based human gait database," in *Applications and Science in Soft Computing*. Berlin, Germany: Springer, 2004, pp. 339–346.

[59] 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.

[60] 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.

[61] D. Tan, K. Huang, S. Yu, and T. Tan, "Efficient night gait recognition based on template matching," in *Proc. Int. Conf. Pattern Recognit.*, 2006, pp. 1000–1003.

[62] A. Tsuji, Y. Makihara, and Y. Yagi, "Silhouette transformation based on walking speed for gait identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 717–722.

[63] M. A. Hossain, Y. Makihara, J. Wang, and Y. Yagi, "Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control," *Pattern Recognit.*, vol. 43, no. 6, pp. 2281–2291, 2010.

[64] Y. Makihara, H. Mannami, and Y. Yagi, "Gait analysis of gender and age using a large-scale multi-view gait database," in *Proc. Asian Conf. Comput. Vis.*, 2010, pp. 440–451.

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

[66] M. Hofmann, S. Bachmann, and G. Rigoll, "2.5D gait biometrics using the depth gradient histogram energy image," in *Proc. IEEE Int. Conf. Biometrics: Theory Appl. Syst.*, 2012, pp. 399–403.

[67] 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, pp. 1–11, May 2018.

[68] 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. 1, Feb. 2018, Art. no. 4.

[69] 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.

[70] 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.

[71] M. Wang and W. Deng, "Deep face recognition: A survey," 2019, *arXiv:1804.06655*.

[72] K. Sundararajan and D. L. Woodard, "Deep learning for biometrics: A survey," *ACM Comput. Surv.*, vol. 51, no. 3, pp. 1–34, May 2018.

[73] TC4 competition and workshop on human identification at a distance, 2020. [Online]. Available: http://hid2020.iapr-tc4.org/

[74] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1302–1310.

[75] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2353–2362.

[76] T. K. Lee, M. Belkhatir, and S. Sanei, "A comprehensive review of past and present vision-based techniques for gait recognition," *Multimedia Tools Appl.*, vol. 72, no. 3, pp. 2833–2869, Jul. 2014.

[77] M. Nieto-Hidalgo, F. J. Ferrández-Pastor, R. J. Valdivieso-Sarabia, J. Mora-Pascual, and J. M. García-Chamizo, "Vision based extraction of dynamic gait features focused on feet movement using RGB camera," in *Proc. Int. Conf. Ambient Intell. Health*, 2015, pp. 155–166.

[78] T. Verlekar, "Gait analysis in unconstrained environments," PhD dissertation, Instituto Superior Técnico, Univ. Lisbon, Lisbon, Portugal, Jun. 2019.

[79] A. Sokolova and A. Konushin, "View resistant gait recognition," in *Proc. Int. Conf. Video Image Process.*, 2019, pp. 7–12.

[80] F. M. Castro, M. J. Marin-Jimenez, N. Guil, S. Lopez-Tapia, and N. Perezde la Blanca, "Evaluation of CNN architectures for gait recognition based on optical flow maps," in *Proc. Int. Conf. Biometrics Special Int. Group*, 2017, pp. 1–5.

[81] E. Gianaria, N. Balossino, M. Grangetto, and M. Lucenteforte, "Gait characterization using dynamic skeleton acquisition," in *Proc. IEEE 15th Int. Workshop Multimedia Signal Process.*, 2013, pp. 440–445.

[82] R. Alp Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7297–7306.

[83] Y. Wang, J. Sun, J. Li, and D. Zhao, "Gait recognition based on 3D skeleton joints captured by kinect," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 3151–3155.

[84] D. Zhang and M. Shah, "Human pose estimation in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2012–2020.

[85] 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.
[86] C. Wang, J. Zhang, P. Xiaoru, and L. Wang, "Chrono-Gait image: A novel temporal template for gait recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2011, pp. 257–270.
[87] 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, Aug. 2009.
[88] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in *Proc. 3rd Int. Conf. Imag. Crime Detection Prevention*, 2009, pp. 1–6.
[89] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task GANs for view-specific feature learning in gait recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 1, pp. 102–113, Jan. 2019.
[90] 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. Chin. Conf. Biometric Recognit.*, 2017, pp. 474–483.
[91] D. Liu, M. Ye, X. Li, F. Zhang, and L. Lin, "Memory-based gait recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–12.
[92] W. Xing, Y. Li, and S. Zhang, "View-invariant gait recognition method by three-dimensional convolutional neural network," *J. Electron. Imag.*, vol. 27, no. 1, Jan. 2018, Art. no. 013010.
[93] B. Lin, S. Zhang, X. Yu, Z. Chu, and H. Zhang, "Learning effective representations from global and local features for cross-view gait recognition," 2020, *arXiv:2011.01461*.
[94] N. Li, X. Zhaoa, and C. Ma, "JointsGait:A model-based gait recognition method based on gait graph convolutional networks and joints relationship pyramid mapping," 2020, *arXiv:2005.08625*.
[95] A. Sepas-Moghaddam, S. Ghorbani, N. F. Troje, and A. Etemad, "Gait recognition using multi-scale partial representation transformation with capsules," in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 8045–8052.
[96] R. Imad, "Towards human body-part learning for model-free gait recognition," 2019, *arXiv:1904.01620*.
[97] Y. Fu *et al.*, "Horizontal pyramid matching for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8295–8302.
[98] A. F. Agarap, "Deep learning using rectified linear units (ReLU)," 2019, *arXiv:1803.08375*.
[99] B. Karlik and A. V. Olgac, "Performance analysis of various activation functions in generalized MLP architectures of neural networks," *Int. J. Artif. Intell. Expert Syst.*, vol. 1, no. 4, pp. 111–122, Dec. 2011.
[100] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[101] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.
[102] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, May 2006.
[103] G. E. Hinton *et al.*, "Learning and relearning in Boltzmann machines," *Parallel Distrib. Process., Explorations Microstruct. Cogn.*, vol. 1, no. 2, pp. 282–317, 1986.
[104] M. Benouis, M. Senouci, R. Tlemsani, and L. Mostefai, "Gait recognition based on model-based methods and deep belief networks," *Int. J. Biometrics*, vol. 8, no. 3, pp. 237–253, Mar. 2017.
[105] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," 2015, *arXiv:1506.00019*.
[106] Y. Feng, Y. Li, and J. Luo, "Learning effective gait features using LSTM," in *Proc. 23rd Int. Conf. Pattern Recognit.*, 2016, pp. 325–330.
[107] 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.
[108] 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.
[109] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[110] K. Cho, B. V. Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*.
[111] Y. Wang *et al.*, "EV-Gait: Event-based robust gait recognition using dynamic vision sensors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6351–6360.
[112] S. Yu *et al.*, "GaitNet: An end-to-end network for gait based human identification," *Pattern Recognit.*, vol. 96, no. 1, pp. 1–11, Dec. 2019.
[113] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Joint intensity transformer network for gait recognition robust against clothing and carrying status," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 12, pp. 3102–3115, Dec. 2019.
[114] 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.
[115] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait recognition via semi-supervised disentangled representation learning to identity and covariate features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13306–13316.
[116] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[117] I. J. Goodfellow *et al.*, "Generative adversarial networks," 2014, *arXiv:1406.2661*.
[118] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh, "GaitGAN: Invariant gait feature extraction using generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 532–539.
[119] S. Yu *et al.*, "GaitGANv2: Invariant gait feature extraction using generative adversarial networks," *Pattern Recognit.*, vol. 87, pp. 179–189, Mar. 2019.
[120] B. Hu, Y. Gao, Y. Guan, Y. Long, N. Lane, and T. Ploetz, "Robust cross-view gait identification with evidence: A discriminant gait GAN (DiGGAN) approach on 10000 people," 2018, *arXiv:1811.10493*.
[121] Y. Wang, C. Song, Y. Huang, Z. Wang, and L. Wang, "Learning view invariant gait features with two-stream GAN," *Neurocomputing*, vol. 339, pp. 245–254, Apr. 2019.
[122] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait recognition invariant to carried objects using alpha blending generative adversarial networks," *Pattern Recognit.*, vol. 105, pp. 1–12, Sep. 2020.
[123] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3859–3869.
[124] Z. Xu, W. Lu, Q. Zhang, Y. Yeung, and X. Chen, "Gait recognition based on capsule network," *J. Vis. Commun. Image Representation*, vol. 59, pp. 159–167, Feb. 2019.
[125] A. Zhao, J. Li, and M. Ahmed, "SpiderNet: A spiderweb graph neural network for multi-view gait recognition," *Knowl.-Based Syst.*, vol. 206, pp. 1–14, Oct. 2020.
[126] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 4165–4169.
[127] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5425–5434.
[128] G. Batchuluun, H. S. Yoon, J. K. Kang, and K. R. Park, "Gait-based human identification by combining shallow convolutional neural network-stacked long short-term memory and deep convolutional neural network," *IEEE Access*, vol. 6, pp. 63 164–63 186, 2018.
[129] Y. Zhang, Y. Huang, S. Yu, and L. Wang, "Cross-view gait recognition by discriminative feature learning," *IEEE Trans. Image Process.*, vol. 29, no. 1, pp. 1001–1015, Jul. 2019.
[130] S. Li, W. Liu, H. Ma, and S. Zhu, "Beyond view transformation: Cycle-consistent global and partial perception gan for view-invariant gait recognition," in *Proc. Int. Conf. Multimedia Expo*, 2018, pp. 1–6.
[131] 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.
[132] C. Zhang, W. Liu, H. Ma, and H. Fu, "Siamese neural network based gait recognition for human identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 2832–2836.
[133] M. Alotaibi and A. Mahmood, "Improved gait recognition based on specialized deep convolutional neural network," *Comput. Vis. Image Understanding*, vol. 164, pp. 103–110, Nov. 2017.

[134] C. Li, X. Min, S. Sun, W. Lin, and Z. Tang, "DeepGait: A learning deep convolutional representation for view-invariant gait recognition using joint Bayesian," *Appl. Sci.*, vol. 7, no. 3, Feb. 2017, Art. no. 210.
[135] 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.
[136] F. M. Castro, M. J. Marin-Jimenez, N. Guil, S. Lopez-Tapia, and N. Perezde la Blanca, "Evaluation of CNN architectures for gait recognition based on optical flow maps," in *Proc. Int. Conf. Biometrics Special Int. Group*, 2017, pp. 1–5.
[137] S. Tong, H. Ling, Y. Fu, and D. Wang, "Cross-view gait identification with embedded learning," in *Proc. Thematic Workshops ACM Multimedia*, 2017, pp. 385–392.
[138] W. Liu, C. Zhang, H. Ma, and S. Li, "Learning efficient spatial-temporal gait features with deep learning for human identification," *Neuroinformatics*, vol. 16, no. 3/4, pp. 457–471, Feb. 2018.
[139] S. Tong, Y. Fu, X. Yue, and H. Ling, "Multi-view gait recognition based on a spatial-temporal deep neural network," *IEEE Access*, vol. 6, pp. 57 583–57 596, 2018.
[140] D. Thapar, A. Nigam, D. Aggarwal, and P. Agarwal, "VGR-net: A view invariant gait recognition network," in *Proc. IEEE 4th Int. Conf. Identity Secur. Behav. Anal.*, 2018, pp. 1–8.
[141] H. Wu, J. Weng, X. Chen, and W. Lu, "Feedback weight convolutional neural network for gait recognition," *J. Vis. Commun. Image Representation*, vol. 55, pp. 424–432, Aug. 2018.
[142] W. An, R. Liao, S. Yu, Y. Huang, and P. C. Yuen, "Improving gait recognition with 3D pose estimation," in *Proc. Chin. Conf. Biometric Recognit.*, 2018, pp. 137–147.
[143] 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.
[144] S. Tong, Y. Fu, and H. Ling, "Gait recognition with cross-domain transfer networks," *J. Syst. Archit.*, vol. 93, pp. 40–47, Feb. 2019.
[145] S.-B. Tong, Y.-Z. Fu, and H.-F. Ling, "Cross-view gait recognition based on a restrictive triplet network," *Pattern Recognit. Lett.*, vol. 125, pp. 212–219, Jul. 2019.
[146] K. Zhang, W. Luo, L. Ma, W. Liu, and H. Li, "Learning joint gait representation via quintuplet loss minimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4695–4704.
[147] P. Zhang, Q. Wu, and J. Xu, "VT-GAN: View transformation GAN for gait recognition across views," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.
[148] S. Li, W. Liu, and H. Ma, "Attentive spatial–temporal summary networks for feature learning in irregular gait recognition," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2361–2375, Sep. 2019.
[149] P. Zhang, Q. Wu, and J. Xu, "VN-GAN: Identity-preserved variation normalizing GAN for gait recognition," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.
[150] X. Wang, J. Zhang, and W. Q. Yan, "Gait recognition using multichannel convolution neural networks," *Neural Comput. Appl.*, vol. 32, pp. 14275–14285, Oct. 2019.
[151] X. Wang and W. Q. Yan, "Cross-view gait recognition through ensemble learning," *Neural Comput. Appl.*, vol. 32, no. 11, pp. 7275–7287, May 2019.
[152] Y. Wu, J. Hou, Y. Su, C. Wu, M. Huang, and Z. Zhu, "Gait recognition based on feedback weight capsule network," in *Proc. IEEE 4th Inf. Technol. Netw. Electron. Autom. Control Conf.*, 2020, pp. 155–160.
[153] 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. 19 196–19 207, 2020.
[154] 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, pp. 1–11, Feb. 2020.
[155] X. Wang and J. Zhang, "Gait feature extraction and gait classification using two-branch CNN," *Multimedia Tools Appl.*, vol. 79, no. 3, pp. 2917–2930, Jan. 2020.
[156] 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, pp. 1–12, Jan. 2020.
[157] 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.
[158] X. Wang and W. Q. Yan, "Non-local gait feature extraction and human identification," *Multimedia Tools Appl.*, vol. 80, pp. 6065–6078, Oct. 2020.
[159] X. Wang and K. Yan, "Gait classification through CNN-based ensemble learning," *Multimedia Tools Appl.*, vol. 80, pp. 1–17, Sep. 2020.
[160] J. Wen, "Gait recognition based on GF-CNN and metric learning," *J. Inf. Process. Syst.*, vol. 16, no. 5, pp. 1105–1112, Oct. 2020.
[161] A. Mehmood *et al.*, "Prosperous human gait recognition: An end-to-end system based on pre-trained CNN features selection," *Multimedia Tools Appl.*, 2020. [Online]. Available: https://doi.org/10.1007/s11042-020-08928-0
[162] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and A. Bouridane, "Gait recognition for person re-identification," *J. Supercomput.*, vol. 77, pp. 3653–3672, 2021.
[163] J. Pan, H. Sun, Y. Wu, S. Yin, and S. Wang, "Optimization of GaitSet for gait recognition," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 1–8.
[164] P. Zhang, Z. Song, and X. Xing, "Multi-grid spatial and temporal feature fusion for human identification at a distance," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 1–5.
[165] G. Huang, Z. Lu, C.-M. Pun, and L. Cheng, "Flexible gait recognition based on flow regulation of local features between key frames," *IEEE Access*, vol. 8, pp. 75 381–75 392, 2020.
[166] J. Su, Y. Zhao, and X. Li, "Deep metric learning based on center-ranked loss for gait recognition," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 4077–4081.
[167] 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. Int. Joint Conf. Biometrics*, 2020, pp. 1–9.
[168] Z. Liu, J. Zhu, J. Bu, and C. Chen, "A survey of human pose estimation: The body parts parsing based methods," *J. Vis. Commun. Image Representation*, vol. 32, pp. 10–19, Oct. 2015.
[169] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Operations Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.
[170] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1–17.
[171] B. Ghojogh, M. Sikaroudi, S. Shafiei, H. R. Tizhoosh, F. Karray, and M. Crowley, "Fisher discriminant triplet and contrastive losses for training siamese networks," in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–7.
[172] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," 2017, *arXiv:1703.09507*.
[173] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4685–4694.
[174] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.
[175] G. De Soete and J. D. Carroll, "K-means clustering in a low-dimensional euclidean space," in *New Approaches in Classification and Data Analysis*. Berlin, Germany: Springer, 1994, pp. 212–219.
[176] B. Biggio, Z. Akhtar, G. Fumera, G. L. Marcialis, and F. Roli, "Security evaluation of biometric authentication systems under real spoofing attacks," *IET Biometrics*, vol. 1, no. 1, pp. 11–24, Mar. 2012.
[177] D. Gafurov, E. Snekkenes, and P. Bours, "Spoof attacks on gait authentication system," *IEEE Trans. Inf. Forensics Secur.*, vol. 2, no. 3, pp. 491–502, Sep. 2007.
[178] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.
[179] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Threat of adversarial attacks on deep learning in computer vision: Survey II," 2021, *arXiv:2108.00401*.
[180] N. Akhtar, M. Jalwana, M. Bennamoun, and A. S. Mian, "Attack to fool and explain deep networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 26, 2021, doi: 10.1109/TPAMI.2021.3083769.
[181] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–10.
[182] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1526–1535.

[183] Y. Wang, P. Bilinski, F. Bremond, and A. Dantcheva, "ImaGINator: Conditional spatio-temporal GAN for video generation," in *Proc. Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1149–1158.
[184] M. Treu, T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Fashion-guided adversarial attack on person segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 943–952.
[185] M. Jia, H. Yang, D. Huang, and Y. Wang, "Attacking gait recognition systems via silhouette guided GANs," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 638–646.
[186] Z. He, W. Wang, J. Dong, and T. Tan, "Temporal sparse adversarial attack on sequence-based gait recognition," 2020, *arXiv:2002.09674*.
[187] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker, "Reconstruction-based disentanglement for pose-invariant face recognition," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 1632–1641.
[188] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1283–1292.
[189] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 140–149.
[190] X. Liu, "Disentanglement for discriminative visual recognition," 2020, *arXiv:2006.07810*.
[191] J. Gu, Z. Wang, W. Ouyang, W. Zhang, J. Li, and L. Zhuo, "3D hand pose estimation with disentangled cross-modal latent space," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 380–389.
[192] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
[193] H. Li, S. Wang, R. Wan, and A. C. Kot, "GMFAD: Towards generalized visual recognition via multi-layer feature alignment and disentanglement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1289–1303, Mar. 2022.
[194] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," *J. Mach. Learn. Res.*, vol. 19, no. 1, pp. 1947–1980, Sep. 2018.
[195] E. Denton and V. Birodkar, "Unsupervised learning of disentangled representations from video," 2017, *arXiv:1705.10915*.
[196] Y. Khraimeche, G.-A. Bilodeau, D. Steele, and H. Mahadik, "Unsupervised disentanglement GAN for domain adaptive person re-identification," 2020, *arXiv:2007.15560*.
[197] K. Ridgeway and M. C. Mozer, "Learning deep disentangled embeddings with the F-statistic loss," 2018, *arXiv:1802.05312*.
[198] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.
[199] H. Rao *et al.*, "Self-supervised gait encoding with locality-aware attention for person re-identification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, Art. no. 125.
[200] H. Rao *et al.*, "A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification," 2020, *arXiv:2009.03671*.
[201] P. Khosla *et al.*, "Supervised contrastive learning," 2020, *arXiv:2004.11362*.
[202] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[203] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6706–6716.
[204] P. Sarkar and A. Etemad, "Self-supervised learning for ECG-based emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 3217–3221.
[205] P. Sarkar and A. Etemad, "Self-supervised ECG representation learning for emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Aug. 06, 2020, doi: 10.1109/TAFFC.2020.3014842.
[206] A. Saeed, T. Ozcelebi, and J. Lukkien, "Multi-task self-supervised learning for human activity detection," *ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 3, no. 2, pp. 1–30, 2019.
[207] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," 2020, *arXiv:2009.09796*.
[208] Y. Zhang and Q. Yang, "A survey on multi-task learning," 2017, *arXiv:1707.08114*.
[209] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1871–1880.
[210] N. Sarafianos, T. Giannakopoulos, C. Nikou, and I. A. Kakadiaris, "Curriculum learning for multi-task classification of visual attributes," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 2608–2615.
[211] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-Wild2, multi-task learning and ArcFace," 2019, *arXiv:1910.04855*.
[212] M. J. Marín-Jiménez, F. M. Castro, N. Guil, F. de la Torre, and R. Medina-Carnicer, "Deep multi-task learning for gait-based biometrics," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 106–110.
[213] S. Chennupati, G. Sistu, S. Yogamani, and S. A. Rawashdeh, "MultiNet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1200–1210.
[214] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4396–4405.
[215] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8107–8116.
[216] W. Xue, H. Ai, T. Sun, C. Song, Y. Huang, and L. Wang, "FrameGAN: Increasing the frame rate of gait videos with generative adversarial networks," *Neurocomputing*, vol. 380, pp. 95–104, Mar. 2020.
[217] X. Chen, X. Luo, J. Weng, W. Luo, H. Li, and Q. Tian, "Multi-view gait image generation for cross-view gait recognition," *IEEE Trans. Image Process.*, vol. 30, no. 1, pp. 3041–3055, Feb. 2021.
[218] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 2261–2268.
[219] B. Dolhansky *et al.*, "The DeepFake detection challenge dataset," 2020, *arXiv:2006.07397*.
[220] 100,000 faces generated by AI, 2022. [Online]. Available: https://generated.photos/
[221] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5780–5789.
[222] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, and J. Fierrez, "GANprintR: Improved fakes and evaluation of the state of the art in face manipulation detection," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 1038–1048, Aug. 2020.
[223] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Oct. 2018.
[224] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," 2017, *arXiv:1702.05374*.
[225] J. Chen, Y. Li, K. Ma, and Y. Zheng, "Generative adversarial networks for video-to-video domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 3462–3469.
[226] S. Huang, C. Lin, S. Chen, Y. Wu, P. Hsu, and S. Lai, "AugGAN: Cross domain adaptation with GAN-based data augmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 731–744.
[227] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.
[228] F. Multon, L. France, M.-P. Cani-Gascuel, and G. Debunne, "Computer animation of human walking: A survey," *J. Vis. Comput. Animation*, vol. 10, no. 1, pp. 39–54, Jun. 1999.
[229] H. Dou *et al.*, "VersatileGait: A large-scale synthetic gait dataset with fine-grained attributes and complicated scenarios," 2021, *arXiv:2101.01394*.
[230] J. Zhang, W. Li, and P. Ogunbona, "Transfer learning for cross-dataset recognition: A survey," 2017, *arXiv:1705.04396*.
[231] Y. Bengio *et al.*, "Deep learners benefit more from out-of-distribution examples," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, pp. 164–172.
[232] D. Teney, K. Kafle, R. Shrestha, E. Abbasnejad, C. Kanan, and A. Hengel, "On the value of out-of-distribution testing: An example of Goodhart's law," 2020, *arXiv:2005.09241*.

[233] A. Sepas-Moghaddam, A. Etemad, F. Pereira, and P. L. Correia, "Long short-term memory with gate and state level fusion for light field-based face recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, no. 1, pp. 1365–1379, Nov. 2021.
[234] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, Dec. 2018.
[235] A. Sepas-Moghaddam, F. Pereira, P. L. Correia, and A. Etemad, "Multi-perspective LSTM for joint visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16535–16543.
[236] S. S. Rajagopalan, L.-P. Morency, T. Baltrusaitis, and R. Goecke, "Extending long short-term memory for multi-view structured learning," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 338–353.
[237] B. H. D. Koh and W. L. Woo, "Multi-view temporal ensemble for classification of non-stationary signals," *IEEE Access*, vol. 7, pp. 32 482–32 491, 2019.
[238] M. C. Liem and D. M. Gavrila, "Joint multi-person detection and tracking from overlapping cameras," *Comput. Vis. Image Understanding*, vol. 128, pp. 36–50, Nov. 2014.
[239] B. Heo, K. Yun, and J. Y. Choi, "Appearance and motion based deep learning architecture for moving object detection in moving camera," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 1827–1831.
[240] S. A. Shah and V. Koltun, "Robust continuous clustering," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 37, pp. 9814–9819, Aug. 2017.
[241] H. Chen, P. Guo, P. Li, G. H. Lee, and G. Chirikjian, "Multi-person 3D pose estimation in crowded scenes based on multi-view geometry," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 541–557.
[242] M. Vo, E. Yumer, K. Sunkavalli, S. Hadap, Y. Sheikh, and S. G. Narasimhan, "Self-supervised multi-view person association and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2794–2808, Aug. 2021.
[243] S. Dodge, J. Mounsef, and L. Karam, "Unconstrained ear recognition using deep neural networks," *IET Biometrics*, vol. 7, no. 3, pp. 207–214, May 2018.
[244] A. Sepas-Moghaddam, F. Pereira, and P. L. Correia, "Ear recognition in a light field imaging framework: A new perspective," *IET Biometrics*, vol. 7, no. 3, pp. 224–231, May 2018.
[245] L. M. Dinca and G. P. Hancke, "The fall of one, the rise of many: A survey on multi-biometric fusion methods," *IEEE Access*, vol. 5, pp. 6247–6289, 2017.
[246] E. L. Oliveira, C. A. Lima, and S. M. Peres, "Fusion of face and gait for biometric recognition: Systematic literature review," in *Proc. Braz. Symp. Inf. Syst. Braz. Symp. Inf. Syst., Inf. Syst. Cloud Comput. Era*, 2016, pp. 108–115.
[247] A. Sakata, N. Takemura, and Y. Yagi, "Gait-based age estimation using multi-stage convolutional neural network," *IPSJ Trans. Comput. Vis. Appl.*, vol. 11, no. 1, pp. 1–10, Jun. 2019.
[248] C. BenAbdelkader, R. Cutler, and L. Davis, "View-invariant estimation of height and stride for gait recognition," in *Proc. Int. Workshop Biometric Authentication*, 2002, pp. 155–167.
[249] K. Koide and J. Miura, "Identification of a specific person using color, height, and gait features for a person following robot," *Robot. Auton. Syst.*, vol. 84, pp. 76–87, Oct. 2016.
[250] M. M. Samson, A. Crowe, P. De Vreede, J. A. Dessens, S. A. Duursma, and H. J. Verhaar, "Differences in gait parameters at a preferred walking speed in healthy subjects due to age, height and body weight," *Aging Clin. Exp. Res.*, vol. 13, no. 1, pp. 16–21, May 2001.
[251] A. Jain and V. Kanhangad, "Gender classification in smartphones using gait information," *Expert Syst. Appl.*, vol. 93, pp. 257–266, Mar. 2018.
[252] H. Han, J. Li, A. K. Jain, S. Shan, and X. Chen, "Tattoo image search at scale: Joint detection and compact representation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2333–2348, Oct. 2019.
[253] X. Xing, K. Wang, and Z. Lv, "Fusion of gait and facial features using coupled projections for people identification at a distance," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2349–2353, Dec. 2015.
[254] Y. Guan, X. Wei, C.-T. Li, G. L. Marcialis, F. Roli, and M. Tistarelli, "Combining gait and face for tackling the elapsed time challenges," in *Proc. IEEE 6th Int. Conf. Biometrics: Theory Appl. Syst.*, 2013, pp. 1–8.
[255] A. E. K. Ghalleb and N. E. B. Amara, "Remote person authentication in different scenarios based on gait and face in front view," in *Proc. 14th Int. Multi-Conf. Syst. Signals Devices*, 2017, pp. 486–491.
[256] A. El KissiGhalleb, R. BenSlamia, and N. E. BenAmara, "Contribution to the fusion of soft facial and body biometrics for remote people identification," in *Proc. Int. Conf. Adv. Technol. Signal Image Process.*, 2016, pp. 252–257.
[257] A. Kale, A. K. Roychowdhury, and R. Chellappa, "Fusion of gait and face for human identification," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2004, pp. V–901.
[258] Q. Yu, Y. Yin, G. Yang, Y. Ning, and Y. Li, "Face and gait recognition based on semi-supervised learning," in *Proc. Chin. Conf. Pattern Recognit.*, 2012, pp. 284–291.
[259] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Comput.*, vol. 32, no. 5, pp. 829–864, May 2020.
[260] V. Talreja, M. C. Valenti, and N. M. Nasrabadi, "Multibiometric secure system based on deep learning," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2017, pp. 298–302.
[261] J. Kim, J. Koh, Y. Kim, J. Choi, Y. Hwang, and J. W. Choi, "Robust deep multi-modal learning based on gated information fusion network," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 90–106.
[262] S. Liu *et al.*, "GIobalFusion: A global attentional deep learning framework for multisensor information fusion," *ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–27, Mar. 2020.

Alireza Sepas-Moghaddam (Member, IEEE) received the BSc and MSc (first class Hons.) degrees in computer engineering in 2007 and 2010, respectively. From 2011 to 2015, he was a lecturer with Shamsipour Technical University, Tehran, Iran. In 2015, he joined the Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal, where he completed his PhD degree (with distinction and Hons.) in electrical and computer engineering in 2019. From 2019 to 2021, he held a postdoctoral fellow position with Queen's University, Kingston, ON, Canada, where he worked on different research projects funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Mitacs, and also private sector. He is currently a senior computer vision data scientist with Socure, dealing with digital identity verification and fraud detection from visual signals. He has contributed more than 45 papers in notable conferences and journals in his area, including CVPR, ICCV, IEEE T-IP, IEEE T-IFS, IEEE T-CSVT, and IEEE T-Biom. His main research interests include theoretical and applied machine learning, notably deep learning, for biometrics, forensics, and affective computing. He was a program chair, publicity chair, and area chair for several conferences and workshops, including ICPR'22, AAAI-HCSSL'22, and EUVIP'22 and is a reviewer for multiple top-tier conferences and journals in the field.

Ali Etemad (Senior Member, IEEE) is currently an assistant professor, and also a mitchell professor of AI for Human Sensing & Understanding with the Department of Electrical and Computer Engineering, and Ingenuity Labs Research Institute, Queen's University, Canada. He leads the Ambient Intelligence and Interactive Machines (Aiim) Lab, where his main research interests include machine learning and deep learning focused on human-centered applications with wearables, smart devices, and smart environments. His work has appeared in top-tier venues, such as CVPR, AAAI, ICCV, ACM CHI, ICASSP, Interspeech, ICPR, FG, T-AFFC, T-IP, T-AI, T-ASLP, T-IFS, T-BIOM, T-HMS, T-NSRE, IEEE T-GRS, IEEE IoT J., and others. He is a co-inventor of nine patents and has given more than 20 invited talks at different venues. He is an associate editor for *IEEE Transactions on Artificial Intelligence*, and is a PC member/reviewer for many notable conferences and journals in the field. Heis the general chair of AAAI Workshop on Human-Centric Self-Supervised Learning (2022), publicity co-chair of European Workshop on Visual Information Processing (2022), and industry relations chair of Canadian Conference on AI (2019). His lab and research program have been funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada, Ontario Centers of Excellence (OCE), Canadian Foundation for Innovation (CFI), Mitacs, and other organizations,and also private sector.