Gait Embeddings

Martina Gassen

Abstract— **Differences in gait can be attributed to a multitude of entangled factors, such as physique and anatomy (interpersonal), mood (intrapersonal), and terrain (environmental). These factors are relevant in many aspects of robotics and Human Robot Interaction (HRI). Hence, this work aims to provide an overview of how gait can be incorporated into robotics, primarily in motion generation and secondarily in the classification of gaits. We focus on the use of gait embeddings to model the varying factors affecting gait. Specifically, we analyze the requirements emerging from the different applications and a breakdown of how current works aim to meet them. We additionally provide an overview of existing and available datasets relating to gait. Lastly, we discuss the state of the art, current limitations, and potential starting-off points for future work.**

I. Introduction

Gait describes the general manner of walking and refers to a wide variety of different walking patterns. These variations typically don't have a single cause. Instead, they are the result of a multitude of interrelated factors. Modeling these factors is relevant in many areas of robotics and Human Robot Interaction (HRI).

In human walking, changes in terrain can induce movement adaptations for safer and more efficient locomotion. Similarly, the speed at which areas are traversed also alters the movement. These adaptations can be significant enough to be considered different types of locomotion. Replicating this behavior can be desirable in the design of legged robots exposed to different and unseen environments. Ideally, gaits should adapt to input signals, and transitions between locomotion types should happen smoothly [56, 51].

Learning different types of locomotion is also relevant to the design of assistive devices, such as lower limb exoskeletons, used for patient rehabilitation [56]. The natural walking style of each patient is different due to their varying physique and other interpersonal differences. To ensure that assisted walking feels natural and comfortable to patients, assistive devices must take these differences into account when generating movements [52, 61, 26].

Another objective is the generation of human-like motion. This objective often requires some level of control over the generated motion. The classification of human gaits, which can be considered as the inverse of the motion generation task, can also be important in HRI to allow robots to adapt their behavior accordingly. A prominent example of this dichotomy is the generation and classification of emotive gaits. Conveying emotions can be helpful to increase the likeability of social and collaborative robots and to improve the experience of people interacting in HRI scenarios [42]. Moreover, emotions increase human likeness [13], which could potentially reduce vandalism directed towards robots [3]. Conversely, emotion recognition from human gaits [5, 6, 39] is an important problem for social robots during human-robot interactions.

Other generation and classification tasks can be derived from the various factors affecting gait. For example, gait changes caused by carrying heavy objects could be imitated to non-verbally communicate that a robot has reached its carrying capacity. The recognition of gait factors may also be relevant in fields outside of robotics and HRI, including but not limited to medical analysis (abnormal gait recognition [20] and estimation of anatomical conditions [37]), surveillance (detection of deceptive behaviour [40] and person identification (PID) [12, 22, 25, 35, 41]), and accident prevention (fatigue detection [38] and measuring distraction [4, 57]).

The isolation of factors is relevant in many classificational and generative cases [20], e.g. for PID, which is mainly concerned with isolating static interpersonal differences from other variable intrapersonal and environmental factors.

Another helpful aspect in the aforementioned applications is the proper modeling of spatio-temporal dependencies of human motion [6, 8, 5, 41, 25]. Although task-agnostic models have the potential to learn these dependencies with a sufficient amount of data, considering these in the design of models can increase performance.

Learning gait embeddings has been explored for many of the aforementioned applications. In generative tasks, gait embeddings benefit from being more descriptive than discrete class labels. They are also computationally more efficient than learning separate models for each class. In combination with Deep Generative Models (DGMs), they can be used to generate synthetic examples without the need for additional inputs. They furthermore allow transfer across different input modalities [20, 37, 52, 26]. Classification tasks use embedding to reduce the need for labeled data using self-supervision [6, 41, 35]. In PID tasks, gait embeddings can constitute templates for private and memory-efficient open-set identification [8, 25, 35]

This work aims to provide an extensive overview of the application of gait embeddings in current research. First, we introduce the different ways in which gaits can be captured and represented (Section II) and describe commonly used motion generation methods (Section III). In Section IV, we outline the requirements that arise from the different applications and discuss how different works attempt to meet them, specifically focusing on the use of gait embeddings. A collection of gait-related datasets is presented in Section V. Finally, in Section VI, we summarize the current progress

in the respective fields, discuss existing limitations, and highlight the resulting potential future work.

II. Gait Data

This section covers relevant methods for capturing gaits. First, we introduce general motion capture methods in Section II-A. We subsequently discuss auxiliary data used in works related to gait. These are used alongside classical motion representations in order to augment the overall representation and to capture a wider range of gait-relevant information.

A. Capturing Motion

Motion can be captured and represented in several ways. In the following, we present the most common representations, namely video, keypoint-based, and rotation-based representations.

Video is a fairly straightforward way of capturing motion. Compared to other data, video can be obtained cheaply and quickly without much preparation. However, the resulting representation is relatively dense and may also capture unwanted environmental information. It is also sensitive to changes in lighting and viewing angle among other things. Unnecessary information about changing backgrounds and illumination can be omitted by extracting silhouette images. Increasing the amount and variety of training data can also help mitigate the problems caused by the non-invariant nature of video data. This is achieved by recording more examples, data augmentation, or multi-view capturing. Aside from that, the application of video for robot motion generation is not trivial and usually requires extracting keypoints or joint angles for proper mapping.

In *keypoint*-based representations, the human body is represented by a set of specific points over time. These points typically coincide with relevant parts of the human body, such as joint locations or landmarks on the surface of the human body. Capturing the keypoints can be done either using Motion Capture (MoCap) systems, which typically require attaching markers to the subject's body, or indirectly using video and extracting the keypoints algorithmically using computer vision. The latter tends to be less accurate and, depending on the method used, may return only 2D coordinates, but can be applied anywhere and without much preparation.

Rotation-based methods similarly represent motion w.r.t. to joints, but instead of measuring their position, their rotation, relative to the previous joint, is captured. The angles can be captured using gyroscopic sensors at the corresponding joints. Like keypoint-based representations, they can also be estimated from video. They are usually represented as Euler angles, axis angles, or quaternions.

These representations can be used on their own, but oftentimes also in combination. Additionally, keypoint- and marker-based representations often also include first- and second-order dynamics, i.e., velocity and acceleration, or the forces and moments exerted.

Both joint-based representations have the computational advantage of being representationally sparse compared to video data. Although, some information is naturally omitted when representing motion this way. One consequence of this representation is that information about the volume of the associated limbs is not taken into account. Thus the body shape, the interaction of touching limbs, and the soft tissue motion cannot be captured. For this reason, some datasets additionally provide body meshes, to model the surface of the human body [29, 18].

B. Auxilliary Data

In addition to the kinematic data described in the previous section, many methods use auxiliary data to better represent gait.

Labeled gait events, such as foot strikes, can be used to aid the model in capturing more high-level data [56, 51, 22]. They can also be useful for segmenting gait sequences into gait cycles, allowing their duration to be normalized [56, 37, 52, 35]. The labeling of foot-based gait events is often done using insole foot pressure sensors for prediction [35, 56]. The readings from these sensors can also be used directly as auxiliary data [35].

Interpersonal differences such as anthropomorphic features also affect a person's gait. These can include height, weight, and age, as well as more specific parameters such as the length of body parts or muscle properties. Several methods use this information to learn individualized gait patterns [37, 61, 52, 26].

The acute physiological state of a person additionally also has an influence on their gait. The measurement of physiological processes can therefore be relevant for the adaptation of generated movements, e.g. for assistive devices. Commonly used sensors are heart rate sensors [59] and electromyographic (EMG) sensors that measure the electrical activity of individual muscles [31, 28, 49]. Furthermore, some medical works have examined the possibility of using neural activity [48].

III. Motion Generation Methods

In the following, we introduce key concepts in the field of DGMs, namely Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). We further elaborate on modifications for generating sequences and conditioning.

A. Deep Generative Models

DGMs, such as VAEs and GANs, can be used not only to predict motion from a given input condition but also to generate synthetic examples without the need for additional signals. Figure 1 provides a schematic overview of the different models used in motion motion generation. In the following, we also describe the most commonly used methods and briefly compare their respective strengths and weaknesses.

Variational Autoencoders (VAEs) [27] are a form of Autoencoders (AEs), that is, they consist of an encoder, which extracts a latent representation from the input, and

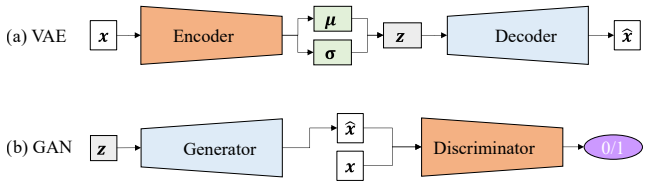

Fig. 1: An overview of different generative Models. Adapted from [60]

a decoder, which aims to reconstruct the input from the latent representation. In VAEs, the encoder produces a latent distribution $q_\phi(\mathbf{z}|\mathbf{x})$, which is conditioned to approximate a predefined prior $p_\theta(\mathbf{z})$ while being optimized to reconstruct the original input. Entirely new outputs can be synthesized by decoding random samples from the prior $p_\theta(\mathbf{z})$. The architecture of a VAE is depicted in Figure 1a.

Generative Adversarial Networks (GANs) [19] consist of two neural networks, namely a generator and a discriminator model. The Generator produces synthetic examples from random noise $\mathbf{z}$ sampled from some probability distribution $p_\mathbf{z}$. The generator implicitly learns the probability distribution p_g. This distribution cannot be evaluated directly, which is why we need the discriminator to encourage p_g to resemble the actual distribution of the data p_{data}. Concurrently, the discriminator tries to distinguish real from synthetic data, while the generator is trained to "fool" the discriminator, i.e. to maximize the probability of a generated image being classified as real by the discriminator. The general structure of a GAN is illustrated in Figure 1b.

Implicitly learning the distribution p_g allows it to be more complex than the distribution over the outputs of a VAE, resulting in higher quality examples. However, the latent space of VAEs tends to be more interpretable than that of GANs.

B. Generation of Temporal Data

The generation of any motion intrinsically requires the generation of sequential data. This can be achieved by generating an entire motion sequence in a single pass through the network (Figures 2a and 2e). This way, the model can always reference any previous or upcoming point in time. A major drawback of this approach is that it often requires all input data to be of the same length. This is not realistic for gait data, as the duration of recordings can vary depending. Therefore, the sequences must often be segmented into individual gait cycles. The duration of these cycles also varies, e.g. due to walking speed or anthropomorphic features such as leg length. Therefore, the duration of the gait cycles additionally needs to be normalized. Nevertheless, the resulting networks tend to be relatively large, due to the high-dimensional input and output, and consequently require a large amount of training data.

In contrast, other models predict each time step sequentially, allowing for sequences of arbitrary length. One way to implement this are autoregressive AEs (Figure 2b), where the readings from timestep t are used to predict the following timestep $t+1$, instead of reconstructing the readings from timestep t. Repeating this for all timesteps allows the reconstruction of the entire sequence. Alternatively, the temporal dependencies can be modeled with recurrent connections, e.g. using a Variational Recurrent Network (VRNN) [10] or an Recurrent GANs (RGANs) [14]. In this, the hidden state $\mathbf{h}_{t-1}$ from the previous timestep is passed to the model in the current timestep t. Depending on the architecture and design choices this can happen in various places in the model e.g. within the encoder and decoder / the generator and discriminator or within the latent space. Versions of the former are depicted in Figures 2c and 2c. The autoregressive approach as well as the use of simple recurrent connections model temporal dependencies in a Markovian way, in which the current timestep t depends only on the previous one $t-1$. Longer-term dependencies are typically captured using more complex recurrences, such as LSTMs [23], or by incorporating the Attention [2] mechanism.

Attention has become especially relevant with the rise of Transformers [50]. These models process sequences in parallel without recurrent units, allowing the model to efficiently learn long-term dependencies. The previously described drawbacks of this approach are addressed by parameter sharing and a high level of parallelization, which is possible through the use of the Attention mechanism. In this mechanism, each element in a sequence is represented as a pair of learned query (Q), key (K), and value (V) vectors. The output of the attention function is the weighted sum of the values of different timesteps w.r.t the weights assigned through the compatibility function of the queries and the keys. Like recurrent units, transformer blocks can be inserted almost anywhere within a model, enabling their usage in VAEs and GANs alike (see Figures 2d and 2g).

C. Conditioning

Another relevant aspect is the conditioning DGMs with a supplementary class label y. This modification is described by Conditional VAEs (CVAEs) [44] and Conditional GANs (CGANs) [34]. The class label conditions each part of the networks, i.e. the encoder and the decoder or the generator and the classifier, respectively. This allows for more control over the generated examples. A schematic representation of CVAEs and CGANs is depicted in Figure 3.

IV. Gait Embeddings

The generation and classification of gaits come with different requirements some general and some depending on the use case. The general requirements consist of learning spatio-temporal dependencies and isolating task-relevant information from interfering factors and noise. Task-specific requirements can be grouped as follows: The ability to condition gait generators on input signals, smooth transitions between locomotion types, and structured latent spaces for classification.

In the following subsections, we cover how these requirements have been addressed in different works. Furthermore,

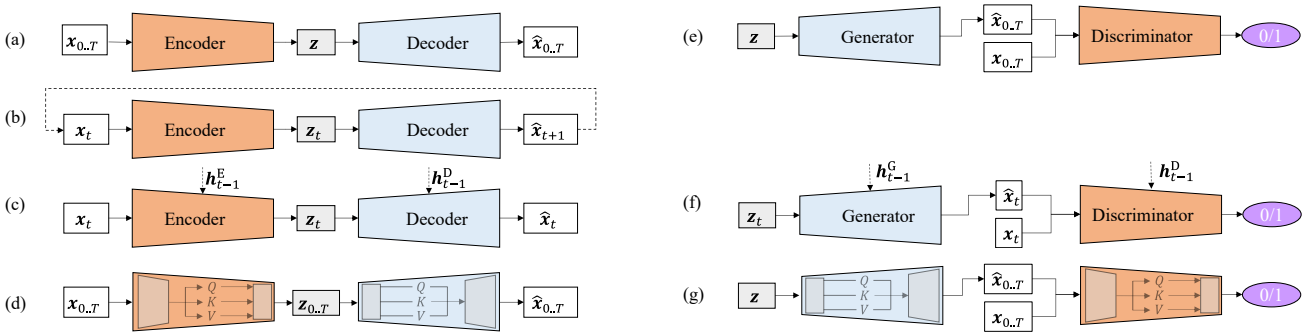

Fig. 2: Temporal DGMs. We show VAE variants (left), where the sampling step is omitted for better readability, and the corresponding GAN variants (right). Figures (a) and (e) depict the processing of the entire sequence in a single pass, Figure (b) depicts an autoregressive VAE, Figures (c) and (f) depict the use of recurrent units, and Figures (d) and (f) depict the Transformer approach.

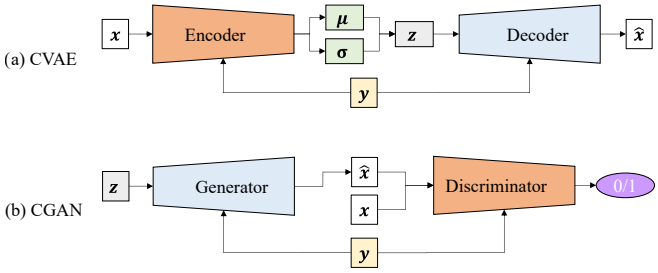

Fig. 3: Conditioning of DGMs

Table I compiles an overview of the publications covered in this work and the requirements they fulfill.

A. Learning Spatio-Temporal Dependencies

The motion of articulated objects is always constrained by their kinematic chain. This knowledge is not explicitly available to task-agnostic models and, therefore, typically needs to be learned implicitly. However, there are also approaches that aim to implement knowledge about limb dependencies directly into the model's structure, which can subsequently reduce the need for training data and increase accuracy.

Bhattacharya et al. [6], for instance, propose hierarchically pooling and unpooling the limb features, following the kinematic chain of human joints in the encoder and decoder, respectively.

GaitPT [8] similarly pools joints hierarchically but in a multi-step manner. This results in joint-level, limb-level, limb-group-level, and body-level representations that are concatenated to form the gait embedding.

By the same reasoning STEP [5] implements a Spatial Temporal Graph Convolutional Network (ST-GCN) [53] for encoding and a corresponding Spatial Temporal Graph Deconvolutional Network (ST-GDCN) for decoding. In this, a graph is constructed from skeleton sequences with inter-body connections reflecting the natural connections of the human body and inter-frame edges, connecting the same joint across neighboring frames. The resulting graph is used to perform graph convolution. This graph also induces temporal dependencies to the immediately connected frames.

Temporally adjacent poses usually exhibit high correlation. Modeling such relationships can increase the accuracy of a model and can help to avoid jerky outputs. Furthermore, human gaits are repetetive motions, if undisturbed. From a more global perspective, it can therefore be helpful to reference previous poses corresponding to the same phase of a gait cycle.

Self-supervised temporal attention is commonly used to learn dependencies on a wider scale [8, 41, 25, 54]. While this is efficient in learning long-term dependencies, the model is not guaranteed to learn local dependencies. Because of that, Rao et al. propose a locality-aware attention mechanism, in which the attention scores are conditioned to be higher in adjacent regions.

B. Feature Isolation

As discussed before, there are many factors from which differences in gait may arise. Therefore, it may be necessary to disentangle these factors to obtain more accurate results. Many works rely on a data-driven approach, where a dataset containing lots of variation implicitly guides the model towards learning task-relevant features.

Gu et al. [20], on the other hand, address this more explicitly, using a multi-encoder AE setup. They split the encoder into two branches, each embedding a different aspect of the gait, namely subject-specific and gait-pattern-specific features. These are disentangled using a cross-subject-reconstruction loss term and a triplet loss on both latent representations using the respective labels. The use of the first term relies on the availability of training data capturing the same person exhibiting multiple patterns separately. If there is a person exhibiting patterns A and B, the person specific pattern of that person is extracted from the recording exhibiting pattern A. The pattern-specific embedding of another person exhibiting pattern B is also extracted and transferred to the first person. Lastly, the newly comprised embedding is decoded and the reconstruction loss w.r.t the original recording of the first person exhibiting pattern B is computed.

Furthermore, by learning a model that produces an intended embedding using supervision, either concurrently or beforehand, we can more directly control the learned representation [6, 51, 61, 37]. This can be helpful when certain variables are known to affect gaits but they can't be directly measured or easily computed as hand-crafted features. Bidirectional GaitNet [37], for instance, generates gaits using labeled gait conditions and anatomical features at the skeletal and muscular levels. Accurate modeling of the latter requires invasive measurements or estimation trough physical exams conducted by experts. Because of this, the authors propose an automatic estimation. The model is trained in two steps. First, the decoder learns to predict gait sequences given a subset of sequences with muscle conditions manually annotated by an expert. Second, the decoder's parameters are frozen and the entire pipeline is trained with the objective. This way, the encoder implicitly learns to reproduce matching muscle conditions, even if the information is unknown.

Similarly, Bhattacharya et al. [6] learn to predict the embedding for a set of gait-based affective features that are supposed to capture a persons displayed emotions. They only constrain a subset of the embedding and allow the remainder to learn freely, so that hidden patterns that may not be obvious to the model designers can be learned aswell. Another advantage compared to handcrafted inputs is that such constrained embeddings can be conditioned using multiple objectives and therefore can be adapted to scenarios when needed, while still imposing structure on the model [51].

Self-supervised methods can also be conditioned more implicitly by extending or altering the inputs that should be reconstructed. Yu et al. [56] reconstruct the relative timing of gait events, with the goal of retaining relevant information and Rao et al. [41] reconstruct sequences in reverse order in order to capture more high-level information.

C. Conditioning on Input Signals

In gait generation, conditioning the generative model can be relevant if the end-result should be controllable. This is the case in the generation of movements associated to some class, e.g. the generation of a 'happy' as opposed to a 'sad' gait. Furthermore, some robots are required to adapt to input signals. This can be the adaptation to starting and ending poses, terrain, commands (e.g. speed), or anthropomorphic features of a patient.

Both STEP [5] and Bidirectional GaitNet [37] use a CVAE, where the conditional input, i.e. a class variable, is used to generate gaits, which enables us to sample new synthetic gaits.

This approach is naturally limited to previously labeled classes. The MoConVQ framework [54] aims to allow more open-ended conditioning using textual commands. To this end, they suggest two different approaches. The first integrates textual features into the temporal transformer of the network via an additional cross-attention layer. The second uses Large Language Models (LLMs) for obtaining the embeddings. In this, the LLM is provided with pairs of textual descriptions and a sequence of embeddings during training. It is subsequently prompted to produce embeddings given a textual description. The resulting embeddings are used to generate new synthetic motions.

Similarly, other methods don't sample the embedding directly, but instead obtain it algorithmically from the respective input signals [52, 26, 35, 56, 51, 61]. This can be achieved in a single step or in multiple steps. In single-step approaches, the control signals are used directly as inputs to a model that learns the embedding [51, 61, 20]. In multi-step approaches, an embedding is first learned, oftentimes in a self-supervised manner. A second model is then trained to predict the embedding. In both methods, the embedding is used to generate the output trajectory. Several works that implement gait generation methods for lower limb exoskeletons use the multi-step approach, using different models. IPGP [52] and Yu et al. [56] use an AE/VAE, to learn the embedding. Jisoo Hong et al. also utilize self supervision by using a Gaussian process dynamical model (GPDM). Secondly, both IPGP [52] and Jisoo Hong et al. [26] use Gaussian Process Regression (GPR) to predict the person-specific gait embedding in relation to anthropomorphic features. Other works use a Reinforcement Learning (RL) policy network to generate new embeddings [56, 54].

D. Smooth Transitions

Smooth transitions between different types of locomotion are important in motion generation to ensure safety and stability. This is especially important for assistive devices, as errors can increase the risk of falls and thus injury.

Wu et al. [51] learn a set of predefined gaits for quadruped robots, such as trotting and pacing. Each gait class is associated with a certain embedding tuple of leg phase offsets, frequency and proportional standing times. By encouraging the model to reproduce one of these tuples adversarially, alongside giving rewards for successfully treading terrains, they allow the robot to adapt the gait when needed and to transition between gaits. Yu et al. [56] capture transitions in terms of gait events. They construct a RL policy that aims to reduce gait event prediction errors in order to facilitate smooth transitions.

E. Structured Latent Space

The capability of classification methods depend on the underlying structure of the input, which is no different when working with embeddings. Because of this, the latent space needs to be structured. Gait embedding works oftentimes do this indirectly if the discriminator network is trained alongside the training network [6]. A more direct way to impose structure on the embeddings is to encourage class separation. This has been explored for person identification purposes using loss terms, e.g. using the triplet-loss [20, 35, 8] or a prototype-loss [35].

V. Gait Datasets

In the following section we provide a non-exhaustive overview of available gait datasets. We broadly summarize

Publication	Objective	ST-modelling	Feature Isolation	Conditioning	Transitions	Structure	Dataset	Model
Wu et al. [51]	Robot Control	✗	✓	✓	✓	✗	-	GAN + RL critic
MoConVQ [54]	Conditioned Motion Generation	✗	✗	✓	✗	✗	other	VQ-VAE
IPGP [52]	Gait Generation for Assistive Devices	✗	✗	✓	✗	✗	own	AE + GPR
Jisoo Hong et al. [26]	Gait Generation for Assistive Devices	✗	✗	✓	✗	✗	own	GPDM
Yu et al. [56]	Gait Generation for Assistive Devices	✗	(✓)	✓	✓	✗	own	VAE + RL critic
Bidirectional GaitNet [37]	Simulation of Anatomical Conditions	✗	✓	✓	✗	✗	own	CVAE
STEP [5]	Emotion Classification & Generation	✓	✗	✓	✗	✗	[5]	CVAE
Bhattacharya et al. [6]	Emotion Classification	(S)	✓	✗	✗	✗	[5]	AE
Gu et al. [20]	Gait Analysis	✗	✓	(✓)	✗	✗	own	AE
Rao et al. [41]	Person Identification	(T)	✗	✗	✗	✗	other	AE
Moon et al. [35]	Person Identification	✗	✗	✗	✗	✓	own	AE
GaitTAKE [25]	Person Identification	✓	✗	✗	✗	✓	[55, 46]	other
GaitPT [8]	Person Identification	✓	✗	✗	✗	✓	[55], other	other

TABLE I: Examined publications implementing gait embeddings. We report their main objective and whether they explicitly address the requirements outlined in Section IV. For spatio-temporal (ST) modelling we also denote works focusing on spatial modelling (S) or temporal modelling (T), only. We also report the model architecture and the used dataset.

how they differentiate or classify different gaits, how the gaits are represented, i.e. video, keypoints, or joint angles, and which auxiliary data is available. We further report the number of subjects and sequences. To facilitate a fair comparison, we report both the number of gaits and the number of total sequences, wherein the number of gaits refers to independent sequences, so additional views or additional synthetic data and data obtained trough data augmentation are not counted as extra sequences. For datasets that contain more than only gait-related data, only gait-related sequences are reported. Table II provides an overview over these attributes.

A. EmotionGait

The EmotionGait dataset [5] combines 2,177 real gait sequences with 1,000 synthetically generated examples, resulting in 3,177 sequences. Of the 2,177 real sequences, 342 explicitly depict emotion. The remaining sequences are taken from Habibie et al.'s Edinburgh Locomotion MOCAP Database [21] and have been manually labeled. Each sequence is labeled as one of 4 emotions, namely: angry, neutral, happy, and sad.

B. DeceptiveWalk

The DeceptiveWalk dataset [40] contains video footage and extracted poses (joint positions) of walks that are labeled as either being deceptive or natural. Each participant performs either a natural or a deceptive walk, which is randomly assigned. In the deceptive, setting the participants are instructed to conceal an object from potential onlookers (people unrelated to the experiment). Walking sequences are split into segments, that are interrupted by other tasks. Some segments had to be removed, if participants did not follow the instructions correctly, resulting in 589 separate sequences. The sequences are additionally labeled with regards to gestures performed by the participants while walking (e.g. hands in pockets or looking around).

C. DUO-GAIT

The DUO-GAIT Dataset [59] records walking under fatigued and dual-tasking conditions alone and in combination. Fatigue was induced via physical activity, and assessed by participants' perceived level of fatigue and changes in heart rate and blood lactate concentration. For dual-tasking, participants were asked to perform a cognitive task while walking. The dataset provides joint angle data measured using Inertial Measurement Units (IMUs)-sensors and participants' antropomorphic features (age, mass, height, leg length, and activity level).

D. CASIA Gait Database

The CASIA Gait Database consists out of 5 independent datasets [9, 55, 47, 58, 45] to date, each capturing identity while focusing on different variations.

CASIA-A [9] focuses on interpersonal differences and includes outdoor videos of three different walking directions (frontal, diagonal, and side view). Each of the 20 participants walked a total of 12 times (4 demonstrations per direction), resulting in 240 walking sequences in the form of silhouette videos.

CASIA-B [55] includes 11 views of each gait sequences. It also provides variations of clothing and carrying conditions. The 124 participants provided 10 demonstrations each: walking normally (6), with backpack (2), wearing a coat (2), resulting in 1240 independent sequences captured indoors. The dataset contains both plain videos and silhouette videos. *CASIA-B** [30], provides an updated version that used newer methods for extracting silhouette images for less noisy results. Furthermore, *CASIA-B-Pose* [15] provides 2D joint positions extracted from the original dataset.

CASIA-C [47] captures the gaits of 153 different subjects using an infrared camera. Each subject walked 10 times (= 1530 total sequences) with four different conditions: walking normally (4), walking with a bag, walking slowly, and walking quickly (twice each). The infrared videos and the extracted silhouette videos are provided.

Dataset	Captured Differences	Vid.	Kpts.	Ang.	Auxiliary Data	# Subj.	# Gaits	# Seq.
EmotionGait[5]	Emotion	✗	✓	✗	✗	-	2,177	3,177
DeceptiveWalk [40]	Deceptiveness	✓	✓	✗	(+)	162	589	589
DUO-GAIT [59]	Fatigue, Multi-tasking	✗	✗	✓	(AF), (+)	16	64	64
CASIA-A [9]	Identity	(S)	✗	✗	✗	20	240	240
CASIA-B [55, 30, 15]	Identity*	(S), (MV)	(2D)	✗	✗	124	1,240	13,640
CASIA-C [47]	Identity*	(IR)	✗	✗	✗	153	1,530	1,530
CASIA-D [58]	Identity	✓	✗	✗	(FP)	88	-	-
CASIA-E [45]	Identity*	(S), (IR), (MV)	✗	✗	(AF)	1,014	29,952	778,752
OU-ISIR Treadmill [32]	Identity*	(S), (MV)	✗	✗	✗	-	3,928	8,728
OUMVLP [46, 1, 29]	Identity	(S), (MV)	(2D)	✗	(BM)	10,307	41,228	288,596
TUM-IITKGP [24]	Identity*	(S)	✗	✗	✗	35	840	840
Fukuchi et al. [16]	Speed	✗	✗	(LB)	(AF), (FP)	42	462	462
Habibie et al. [21]	Speed	✗	✓	✗	✗	-	1, 835	1, 835
Macaluso et al. [31]	Perturbations*	✗	✗	(LB)	(FP), (EMG)	10	180	180
Reznick et al. [43]	Locomotion, Incline, Speed	✗	(LB)	(LB)	(FP)	10	-	-
MoVi [18]	Locomotion	(MV)	✓	✓	(BM)	90	-	-
Lencioni et al.. [28]	Locomotion	✗	(LB)	✓	(FP), (EMG), (+)	50	1,750	1,750
CMU MoCap [11]	Locomotion	✓	✓	✗	✗	144	-	-
100STYLE [33]	Locomotion, Style	✗	✓	✓	✗	1	810	810
Van Criekinge et al. [49]	Medical Conditions	✗	✗	✓	(AF), (FP), (EMG)	188	-	-

TABLE II: Gait related dataset, categorized by their captured differences. Entries marked with * denote that the additional differences are captured but are not the main focus of the dataset. These differences are described in the section regarding the specific dataset. The Vid. column denotes wheter videos are included in the dataset (✗ = no video, ✓ = plain video, (S) = silhouette videos, (MV) = multiple viewing angles, (IR) = infrared videos) The usage of keypoints and angles are denoted in the columns labelled Kpts. and Ang., respectively (✗ = no data, ✓ = full body, (2D) = 2D keypoints, (LB) = lower body). The availability of auxiliary data is further documented (✗ = no additional data, (FP) = foot pressure, (AF) = anthropomorphic features, (BM) = body meshes, (EMG) = electromyographic (EMG), (+) = other (documented in the respective dataset's section)). The columns labelled # Subj., # Gaits, and # Seq. denote the number of subjects, gaits and total sequences.

CASIA-D [58] contains videos of gaits captured indoors as well as corresponding cumulative foot pressure images, which represent the pressures exerted during a single gait cycle.

CASIA-E [45] provides a large-scale datasets of 1,014 different people captured from 26 different viewpoints (13 angles at 2 different heights). The data was captured in three different outdoor scenes under four different conditions: normal walking, carrying a bag, wearing a coat and pausing. The dataset comes with plain video and silhouette video, and also provides soft biometric features (age, gender, height, weight, and nationality). A subset of CASIA-E was captured using a thermal infrared camera in a single-view setting (540 videos).

E. OU-ISIR Gait Database

The OU-ISIR Gait Database is compiled form several independent datasets pertaining to identification using gait [36]. We cover the Treadmill Dataset [32] and the Multi-view Large Population (OUMVLP) Dataset and its variations [46, 1, 29].

The *Treadmill Dataset* [32] contains 3,928 independent gait sequences captured on a treadmill. The dataset itself is divided into four subsets A-D, dedicated to altering a single variable at once. All subset provide single-view silhouette and plain video recordings, except for the multi-view Subset C. This subset captures 200 participants from 25 different views under otherwise constant walking conditions. Subset A captures walking at 9 different speeds ranging from 2 km/h to 10 km/h in 1 km/h intervals (612 gaits from 34 participants). Clothing variations are captured by Subset B (2,746 gaits from 48 participants, up to 32 variations per participant). Lastly, Subset D focuses on capturing the stability of gaits, measuring the fluctuation between gait cycles and dividing the 185 participants into two subsets of people with low and high gait fluctuation.

The *OUMVLP* [46] provides gaits from 10,307 participants recorded with 7 cameras. The original version provides only silhouette videos, but later versions also provide 2D joint positions (*OUMVLP-Pose*) [1] and body meshes (*OUMVLP-Mesh*) [29] extracted from the base dataset's videos.

F. TUM-IITKGP

The TUM-IITKGP Dataset [24] provides 840 gait sequences from 35 different subjects in the form of silhouette videos. It captures interpersonal differences, 2 different walking styles, as well as different carrying conditions and introduces occlusion as a challenge. Each participant was filmed in 6 configurations: regular walking, hands in pocket, with a backpack, wearing a gown, dynamic occlusion (caused by other walking people), and static occlusion (caused by other standing people).

G. Fukuchi et al.

Fukuchi et al.'s dataset [16] contains gaits recorded both on a treadmill and overground at different speeds. The lower limb and pelvis keypoints of the 42 participants were captured in addition to external forces measured using force plates. The anthropomorphic attributes (age, height, and mass) of each participant are also reported.

H. Habibie et al.

Habibie et al.'s Edinburgh Locomotion MOCAP Database [21] models different gaits w.r.t to speed. The 1,835 sequences are represented using joint positions.

I. Macaluso et al.

The dataset by Macaluso et al. [31] captures the differences in gait in response to disturbances. The dataset is comprised of the control dataset and the perturbed dataset. In the former, ten able-bodied participants walked on a split-belt treadmill at three different inclines (-5°, 0°, and +5°) at a self-selected speed. In the latter, the same participants walked on the same treadmill with the same base speed. The gait was disturbed by randomly timed perturbation events, slowing down or accelerating the treadmill. Each participant completed 5 walking-sessions per incline, resulting in a total of 15 perturbed and 3 regular walking sequences per person. The dataset captures the lower limb joint angles, foot pressure, and EMG muscle activation of the lower body.

J. Reznick et al.

Reznick et al.'s dataset [43] focuses on different types of locomotion and the transitions between them, namely standing up, walking, running, and climbing stairs. Each locomotion type was captured under various conditions (e.g. inclines, speed, and acceleration). The actions were performed by ten able-bodies participants, equipped with markers on the lower body, for extracting joint position and angles, as well as estimating joint forces, moments, and powers. The action were performed on a treadmill equipped with force plates and a 4-Step adjustable stair set.

K. MoVi

The MoVi dataset [18] contains recordings of 20 pre-defined action, 5 of which relate to different locomotion types: namely walking, jogging, running in place, side gallop, and crawling. The dataset is multi-modal containing video from 4 different points of view, extracted body meshes, MoCap-data, and data from IMUs-sensors. The recordings of the different modalities, taken in different sessions, have been synchronized.

L. Lencioni et al.

The dataset by Lencioni et al. [28] contains multi-modal recordings of 4 different locomotion styles (toe-walking, heel-walking, stair-ascending, and stair-descending) at different velocities. The dataset contains full-body joint position, the joint angles of the lower body ground reaction forces and torques, center of pressure, lower limb joint mechanical moments and power, displacement of the body's center of mass as well as EMG signals of the main lower limb muscles.

M. CMU MoCap

The CMU Motion Capture Database (CMU MoCap) [11] covers a wide range of human motions, partially related to gait. It captures different types of locomotion (e.g. running, walking, and jumping), as well as gait differences caused by environmental disturbances (e.g. uneven terrain) and social interactions (e.g. walking with linked arms). Both video and joint position data is provided.

N. 100STYLE

Mason et al.'s 100STYLE dataset [33] contains different types of locomotion (e.g. forwards walking and sidestep running) in a total of 100 different styles ranging from natural variations such as walking with folded arms to theatrical ones such as walking like a zombie. All gaits were recorded once by a single actor and joint angles and positions were extracted from the resulting videos.

O. Van Criekinge et al.

Van Criekinge et al.'s dataset [49] captures the gaits of 138 able-bodied adults and 50 stroke survivors. It provides full-body joint angles, forces, moments and power, as well as the 3D center of mass, foot- pressure and EMG muscle activity of 14 back and lower limb muscles. They additionally report age, sex, body mass, height, and leg length of each participant.

VI. Discussion

Gait embeddings have been applied in numerous recent works regarding many different task and have generally yielded good results. Despite the different objectives, there are many shared techniques that can be found across task. Notably, there are many publications that utilize gait embeddings for classification purposes. Most of them are designed using some variation of an AE/VAE-architecture. Therefore, they could theoretically be adapted for gait generation and the efficacy of the different modeling choices that address general requirements could be tested in future work. Moreover, one could conceptualize a full pipeline in which human-gait is analyzed and a matching robots gait is generated accordingly. Such pipelines have been explored in HRI [17, 7] and could be relevant in the generation of interactive gaits, e.g. synchronizing walking rhythms while talking to a person or supporting a physically-impaired person.

Current works focusing on the generation of motion w.r.t. to class are limited to the classes known beforehand. In some cases new classes may become relevant overtime. Therefore, future work could explore a more open-ended approach inspired by few-shot learning and the methods used in open-set PID [20, 35, 8]

There also exist a wide range of gait-related datasets. While many gait factors are covered, identity is by far the most common one, not only in terms of dataset quantity but also in size. Other factors do not get as much coverage, which potentially slows research in these fields as researchers need to create their own datasets for experiments [52, 20, 37, 26]. Such datasets are often times not shared publicly which hinders comparability trough common benchmarks. We especially observe a large discrepancy between the number of publications and the availability of datasets dedicated to gait generation for assistive devices. The lack of datasets related to medical applications could be due to privacy concerns

and the reduced amount of potential participants. Similarly, the obtainment of datasets capturing emotive gaits can be challenging, because actively inducing negative emotions in participants is limited by ethical considerations. Because of this, some works rely on participants imitating the desired behaviour [40, 5, 20]. While this increases the availability of data, the quality depends heavily on the quality of imitation and there is a risk of missing crucial aspects of the genuine motions.

Another aspect is the data representation. Many datasets focus on video data and very few provide full-body 3D joint-data, which restricts their applicability in robotics. Therefore, there is still lots of potential future work that can be done in the creation of large scale and varied datasets capturing gaits on a joint-based basis.

REFERENCES

[1] W. An et al. "Performance Evaluation of Model-Based Gait on Multi-View Very Large Population Database With Pose Sequences". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 2.4 (2020), pp. 421–430.

[2] D. Bahdanau, K. Cho, and Y. Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2014.

[3] C. Bartneck and M. Keijsers. "The morality of abusing a robot". In: *Paladyn, Journal of Behavioral Robotics* 11.1 (2020), pp. 271–283.

[4] H. Bazzi and A. T. Cacace. "Altered gait parameters in distracted walking: a bio-evolutionary and prognostic health perspective on passive listening and active responding during cell phone use". In: *Frontiers in Integrative Neuroscience* 17 (2023). ISSN: 1662-5145.

[5] U. Bhattacharya et al. "STEP: Spatial Temporal Graph Convolutional Networks for Emotion Perception from Gaits". In: *CoRR* (2019).

[6] U. Bhattacharya et al. "Take an Emotion Walk: Perceiving Emotions from Gaits Using Hierarchical Attention Pooling and Affective Mapping". In: *CoRR* (2019).

[7] J. Bütepage et al. *Imitating by generating: deep generative models for imitation of interactive tasks*. 2019.

[8] A. Catruna, A. Cosma, and I. E. Radoi. *GaitPT: Skeletons Are All You Need For Gait Recognition*. Aug. 2023.

[9] *Center for Biometrics and Security Research*. URL: http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp (visited on 02/04/2024).

[10] J. Chung et al. "A Recurrent Latent Variable Model for Sequential Data". In: *CoRR* abs/1506.02216 (2015).

[11] *CMU Graphics Lab Motion Capture Database*. URL: http://mocap.cs.cmu.edu/ (visited on 02/28/2024).

[12] T. Connie, M. K. O. Goh, and A. B. J. Teoh. "A Grassmann graph embedding framework for Gait analysis". In: *Eurasip Journal on Advances in Signal Processing* 2014 (1 Feb. 2014), pp. 1–17.

[13] M. Destephe et al. "Emotional gait: Effects on humans' perception of humanoid robots". In: *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. 2014, pp. 261–266.

[14] C. Esteban, S. L. Hyland, and G. Rätsch. *Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs*. 2017.

[15] Y. Fu et al. *GPGait: Generalized Pose-based Gait Recognition*. 2023.

[16] C. A. Fukuchi, R. K. Fukuchi, and M. Duarte. "A public dataset of overground and treadmill walking kinematics and kinetics in healthy individuals". In: *PeerJ* 6 (4 2018). ISSN: 2167-8359.

[17] M. Gassen. "Learning a library of Physical Interactions for Social Robots". In: 2021.

[18] S. Ghorbani et al. "MoVi: A Large Multipurpose Motion and Video Dataset". In: *CoRR* abs/2003.01888 (2020).

[19] I. J. Goodfellow et al. *Generative Adversarial Networks*. 2014.

[20] X. Gu et al. "Cross-Subject and Cross-Modal Transfer for Generalized Abnormal Gait Pattern Recognition". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.2 (2021), pp. 546–560.

[21] I. Habibie et al. "A Recurrent Variational Autoencoder for Human Motion Synthesis". In: *Proceedings of the British Machine Vision Conference (BMVC)*. 2017.

[22] M. F. Hashmi et al. "GaitVision: Real-Time Extraction of Gait Parameters Using Residual Attention Network". In: *Complexity* 2021 (2021).

[23] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[24] M. Hofmann, S. Sural, and G. Rigoll. "Gait Recognition in the Presence of Occlusion: A New Dataset and Baseline Algorithms". In: 2011.

[25] H.-M. Hsu et al. *GaitTAKE: Gait Recognition by Temporal Attention and Keypoint-guided Embedding*. 2022.

[26] C. C. Jisoo Hong, S.-J. Kim, and F. C. Park. "Gaussian Process Trajectory Learning and Synthesis of Individualized Gait Motions". In: (2019).

[27] D. P. Kingma and M. Welling. *Auto-Encoding Variational Bayes*. 2013.

[28] T. Lencioni et al. "Human kinematic, kinetic and EMG data during different walking and stair ascending and descending tasks". In: *Scientific Data* 6 (Dec. 2019), p. 309.

[29] X. Li et al. "Multi-View Large Population Gait Database With Human Meshes and Its Performance Evaluation". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4.2 (2022), pp. 234–248.

[30] J. Liang et al. *GaitEdge: Beyond Plain End-to-end Gait Recognition for Better Practicality*. 2022.

[31] R. Macaluso et al. "Parameterizing Human Locomotion Across Quasi-Random Treadmill Perturbations and Inclines". In: *IEEE Transactions on Neural*

Systems and Rehabilitation Engineering 29 (2021), pp. 508–516.

[32] Y. Makihara et al. "The OU-ISIR Gait Database Comprising the Treadmill Dataset". In: *IPSJ Trans. on Computer Vision and Applications* 4 (Apr. 2012), pp. 53–62.

[33] I. Mason, S. Starke, and T. Komura. "Real-Time Style Modelling of Human Locomotion via Feature-Wise Transformations and Local Motion Phases". In: *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 5.1 (May 2022).

[34] M. Mirza and S. Osindero. *Conditional Generative Adversarial Nets*. 2014.

[35] J. Moon et al. "Explainable gait recognition with prototyping encoder–decoder". In: *PLoS ONE* 17 (3 Mar. 2022).

[36] *OU-ISIR Biometric Database*. URL: `http://www.am.sanken.osaka-u.ac.jp/BiometricDB/index.html` (visited on 02/28/2024).

[37] J. Park et al. "Bidirectional GaitNet: A Bidirectional Prediction Model of Human Gait and Anatomical Conditions". In: *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings*. SIGGRAPH '23. ACM, July 2023.

[38] H. R. Pavel et al. "Assessment of Cognitive Fatigue from Gait Cycle Analysis". In: *Technologies* 11.1 (2023). ISSN: 2227-7080.

[39] T. Randhavane et al. "Identifying Emotions from Walking using Affective and Deep Features". In: *CoRR* abs/1906.11884 (2019).

[40] T. Randhavane et al. "The Liar's Walk: Detecting Deception with Gait and Gesture". In: *CoRR* abs/1912.06874 (2019).

[41] H. Rao et al. "Self-Supervised Gait Encoding with Locality-Aware Attention for Person Re-Identification". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, July 2020.

[42] M. E. Reyes, I. V. Meza, and L. A. Pineda. "Robotics facial expression of anger in collaborative human–robot interaction". In: *International Journal of Advanced Robotic Systems* 16 (1 Jan. 2019).

[43] E. Reznick et al. "Lower-limb kinematics and kinetics during continuously varying human locomotion". In: *Scientific Data 2021 8:1* 8 (1 Oct. 2021), pp. 1–12. ISSN: 2052-4463.

[44] K. Sohn, H. Lee, and X. Yan. "Learning Structured Output Representation using Deep Conditional Generative Models". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015.

[45] C. Song et al. "CASIA-E: A Large Comprehensive Dataset for Gait Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.3 (2023), pp. 2801–2815.

[46] N. Takemura et al. "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition". In: *IPSJ Trans. on Computer Vision and Applications* 10.4 (2018), pp. 1–14.

[47] D. Tan et al. "Efficient Night Gait Recognition Based on Template Matching". In: *18th International Conference on Pattern Recognition (ICPR'06)*. Vol. 3. 2006, pp. 1000–1003.

[48] Y. Thenaisie et al. "Principles of gait encoding in the subthalamic nucleus of people with Parkinson's disease". In: *medRxiv* (2022).

[49] T. Van Criekinge et al. "A full-body motion capture gait dataset of 138 able-bodied adults across the life span and 50 stroke survivors". In: *Scientific Data 2023 10:1* 10 (1 Dec. 2023), pp. 1–8. ISSN: 2052-4463.

[50] A. Vaswani et al. *Attention Is All You Need*. 2017.

[51] J. Wu, Y. Xue, and C. Qi. *Learning Multiple Gaits within Latent Space for Quadruped Robots*. 2023. eprint: `2308.03014`.

[52] X. Wu et al. "Individualized Gait Pattern Generation for Sharing Lower Limb Exoskeleton Robot". In: *IEEE Transactions on Automation Science and Engineering* 15.4 (2018), pp. 1459–1470.

[53] S. Yan, Y. Xiong, and D. Lin. *Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition*. 2018.

[54] H. Yao et al. *MoConVQ: Unified Physics-Based Motion Control via Scalable Discrete Representations*. 2023.

[55] S. Yu, D. Tan, and T. Tan. "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition". In: *18th International Conference on Pattern Recognition (ICPR'06)*. Vol. 4. IEEE. 2006, pp. 441–444.

[56] Z. Yu et al. "Adaptive Gait Trajectory and Event Prediction of Lower Limb Exoskeletons for Various Terrains Using Reinforcement Learning". In: *Journal of Intelligent & Robotic Systems* (2023).

[57] M. H. Zaki and T. Sayed. "Exploring walking gait features for the automated recognition of distracted pedestrians". In: *IET Intelligent Transport Systems* 10.2 (2016), pp. 106–113.

[58] S. Zheng et al. "A cascade fusion scheme for gait and cumulative foot pressure image recognition". In: *Pattern Recognition* 45.10 (2012), pp. 3603–3610. ISSN: 0031-3203.

[59] L. Zhou et al. "DUO-GAIT: A gait dataset for walking under dual-task and fatigue conditions with inertial measurement units". In: *Scientific Data 2023 10:1* 10 (1 Aug. 2023), pp. 1–10.

[60] W. Zhu et al. *Human Motion Generation: A Survey*. 2023.

[61] C. Zou et al. "Learning gait models with varying walking speeds". In: *IEEE Robotics and Automation Letters* 6 (1 Jan. 2021), pp. 183–190.