

# Associated Spatio-Temporal Capsule Network for Gait Recognition

Aite Zhao , Junyu Dong , Jianbo Li , Lin Qi, and Huiyu Zhou 

**Abstract**—It is a challenging task to identify a person based on her/his gait patterns. State-of-the-art approaches rely on the analysis of temporal or spatial characteristics of gait, and gait recognition is usually performed on single modality data (such as images, skeleton joint coordinates, or force signals). Evidence has shown that using multi-modality data is more conducive to gait research. Therefore, we here establish an automated learning system, with an associated spatio-temporal capsule network (ASTCapsNet) trained on multi-sensor datasets, to analyze multimodal information for gait recognition. Specifically, we first design a low-level feature extractor and a high-level feature extractor for spatio-temporal feature extraction of gait with a novel recurrent memory unit and a relationship layer. Subsequently, a Bayesian model is employed for the decision-making of class labels. Extensive experiments on several public datasets (normal and abnormal gait) validate the effectiveness of the proposed ASTCapsNet, compared against several state-of-the-art methods.

**Index Terms**—Associated capsules, capsule network, gait recognition, multi-sensor, spatio-temporal.

## I. INTRODUCTION

NOWADAYS, there are a number of technologies for identifying a person, and one of them is face recognition, which has been applied in a wide range of applications [1], [2]. Additionally, fingerprint recognition, palmprint and action recognition have attracted large attention from the research community [3]–[5]. Gait recognition is a biometric technology that recognizes people at a longer distance by acquiring walking

Manuscript received June 24, 2020; revised October 22, 2020 and January 9, 2021; accepted February 5, 2021. Date of publication February 19, 2021; date of current version February 8, 2022. This work was supported in part by National Key Research and Development Plan Key Special Projects under Grant 2018YFB2100303, in part by Shandong Province Colleges and Universities Youth Innovation Technology Plan Innovation Team Project under Grant 2020KJN011, in part by Shandong Provincial Natural Science Foundation under Grant ZR2020MF060, in part by Program for Innovative Postdoctoral Talents in Shandong Province under Grant 40618030001, in part by National Natural Science Foundation of China under Grant 61802216, and in part by Postdoctoral Science Foundation of China under Grant 2018M642613. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Guo-Jun Qi. (*Corresponding author: Jianbo Li.*)

Aite Zhao and Jianbo Li are with the College of Computer Science and Technology, Qingdao University, Qingdao 266071, Shandong, China (e-mail: zhaoaite@gmail.com; lijianbo@qdu.edu.cn).

Junyu Dong and Lin Qi are with the College of Information Science and Engineering, Ocean University of China, Qindao 266100, China (e-mail: dongjunyu@ouc.edu.cn; qilin@ouc.edu.cn).

Huiyu Zhou is with the School of Informatics, University of Leicester, Leicester LE1 7RH, Leicestershire, U.K. (e-mail: hz143@leicester.ac.uk).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TMM.2021.3060280>.

Digital Object Identifier 10.1109/TMM.2021.3060280

data [6], [7], which aims to analyze the overall human activity rather than part of the human body, and gait is not easily imitated and hence more reliable for the recognition task.

Human beings have a unique visual system, which can directly identify people based on human gait beyond a certain distance. The input data of a gait recognition model is generally a sequence of walking videos or images, and its data collection is similar to face recognition, which is noninvasive. However, due to a large amount of data from image sequences, the computational complexity of gait recognition is relatively high, and the processing is not straightforward. A gait recognition system extracts key features from the images of the joint movement of a walking person. So far, there is no commercialized gait-based identity authentication system.

In previous research, many types of sensors have been used for gait data collection and recognition - image, depth and inertial sensors, which acquire RGB images/videos, motion distance and acceleration/orientation. Among them, captured images have the advantages of convenience, non-contact, and easy understanding. Besides, skeleton data well preserve the integrity of gait movement whilst minimizing the interference of color and shape of the human body in the image. Some existing work can directly extract 3D skeleton nodes from the images of the human body and this character makes up for the shortcomings of 2D RGB images [8], [9].

On the one hand, traditional gait recognition algorithms have been well studied including Hough transform, contour extraction, and three-dimensional wavelet moments. Model-based algorithms are well investigated in multiple data sources. Besides, deep learning methods, such as convolutional neural networks (CNN) and restricted Boltzmann machine (RBM) have been employed and achieved promising results in gait recognition or other relevant fields [10]–[12]. Moreover, there are several studies on image segmentation and image classification, which can be applied to the field of gait recognition to solve the problem of multimodal gait data classification [13], [14]. However, these approaches are not designed to deal with sequential data.

On the other hand, although some feature extraction algorithms take temporal changes into account, they lack the analysis of spatial information. These methods use the relationship between time series to represent the dynamics in the gait cycle, but do not analyze each frame in detail. Specifically, there are several state-of-the-art methods for sequential information acquisition [8], [15]–[19], such as original, bi-directional, and attention-based LSTM. However, these methods come from the standard LSTM in order to enhance its performance,

but the spatial mapping of the input data has not been fully addressed.

Apart from temporal and spatial characteristics, we also consider the structural relationship between the various parts of the human body when analyzing gait, such as the relationship between feet and legs, head and body, and limbs and trunk. Unlike CNN, capsule network (CapsNet) [20] does not bypass the position of the entity (body parts) within the region (human body). For low-level capsules (body parts), location information is “place-coded” whereas capsule is active. Using an iterative routing process, each active capsule will choose a capsule in the layer to be its parent in the tree. For the task of gait recognition, this iterative process will be solving the problem of locating the body parts.

In order to extract temporal, spatial, and structural features from multimodal data, we design a capsule network to synchronously analyze gait patterns and changes. It mainly includes two tasks, feature extraction and classification. We extract and classify features, and evaluate the effectiveness of feature extraction in classification. Inspired by the learning process of CapsNet [20] and the gated recurrent unit (GRU) [15], an associated spatio-temporal capsule network (ASTCapsNet) is here proposed for gait recognition, including three modules: a low-level feature extractor, a high-level feature extractor, and a decision-making layer.

The low-level feature extractor aims to extract spatio-temporal features of the input data, including a novel expandable memory module and a convolution layer; the high-level feature extractor mainly performs matrix operation on the feature map in the form of capsules, including the primary capsule layer, the relationship layer and the digital capsule layer with the dynamic routing algorithm.

In the decision-making layer, we design four softmax classifiers to analyze the output of all layers, and the parameters of each layer of the network are optimised until the redefined joint loss converges to the minimum. Moreover, the standard Bayesian model is employed to fit the results of the classifiers and regain as a new feature input. Furthermore, we also propose a relationship layer between the capsules to calculate the relative relationship between the local features of gait.

The main contributions of this paper are summarized as follows:

- 1) An associated spatio-temporal capsule network (ASTCapsNet) is proposed for spatio-temporal and structural feature extraction and classification using multimodal gait data.
- 2) Two feature extractors are designed for low- and high-level gait feature analysis and extraction. The low-level feature extractor contains a novel memory module and a convolution layer to extract spatio-temporal features. In the high-level feature extractor, we propose a relationship layer to calculate the relationship matrix between two capsules to ensure the structural integrity of the input.
- 3) We develop a recurrent memory unit to enable the model to converge faster than the original GRU unit.
- 4) The decision-making layer is used to construct the evaluation mechanism of ASTCapsNet, through four softmax classifiers and a Bayesian model to select optimal features.

- 5) Compared against the other state-of-the-art technologies, the proposed method achieves superior results on five challenging datasets, including both normal and abnormal gait data, for multimodal gait recognition.

This paper is organized as follows. Section II describes existing studies on gait recognition. Section III introduces the proposed ASTCapsNet model, and Section IV shows experimental results. We discuss the performance of ASTCapsNet, and point out its weaknesses in Section V. We conclude the whole paper in Section VI.

## II. RELATED WORK

Significant efforts have been made to progress the state-of-the-art of gait recognition. From handcrafted feature recognition to model-based recognition, gait recognition methods have undergone significant progress. It must be pointed out that gait data collection process has become diverse.

### A. Multi-Sensor Based Methods

There are several ways to obtain gait data collected by sensors. The types and output of the sensors are different and can be roughly classified into the following types: cameras, infrared, depth, force sensors, accelerators and gyroscopes. Camera acquisition is convenient and non-invasive, which can capture people’s gait form at a certain distance. Chattopadhyay *et al.* combined front and back view features from RGB-D images. The data acquisition was conducted at the airport security checkpoint with two depth cameras installed on top of the metal detector door outside the yellow line [21]. Xue *et al.* used an infrared thermal imaging camera to collect gait images to establish an infrared thermal gait database, which can detect the human body and remove noise from complex background [22].

There is also the use of gyroscopes and accelerators to develop wearable measurement modules for measuring inertial signals generated while walking and to assess gait, as a basis for new hemiplegia diagnostic techniques using wearable devices [23].

Besides, some studies were based on foot force datasets. Force sensors were installed on the sole to capture the foot movement data of a patient in order to monitor the progress of neurodegenerative diseases [19].

### B. Hand-Crafted Features Based Approaches

Based on the above-mentioned data sources, early gait recognition techniques often use traditional computer vision methods for image understanding and produce hand-crafted features from raw data. Gianaria *et al.* [24] defined a dataset of physical and behavioral features to identify relevant parameters for gait description. Andersson *et al.* [25] acquired gait attributes in each gait cycle, such as leg angle, step length, stride length, and anthropometric attributes such as mean and standard deviation. However, these features are in a low-dimensional space and are mainly based on human experience. Therefore, it is difficult to obtain descriptive features in a high-dimensional space.

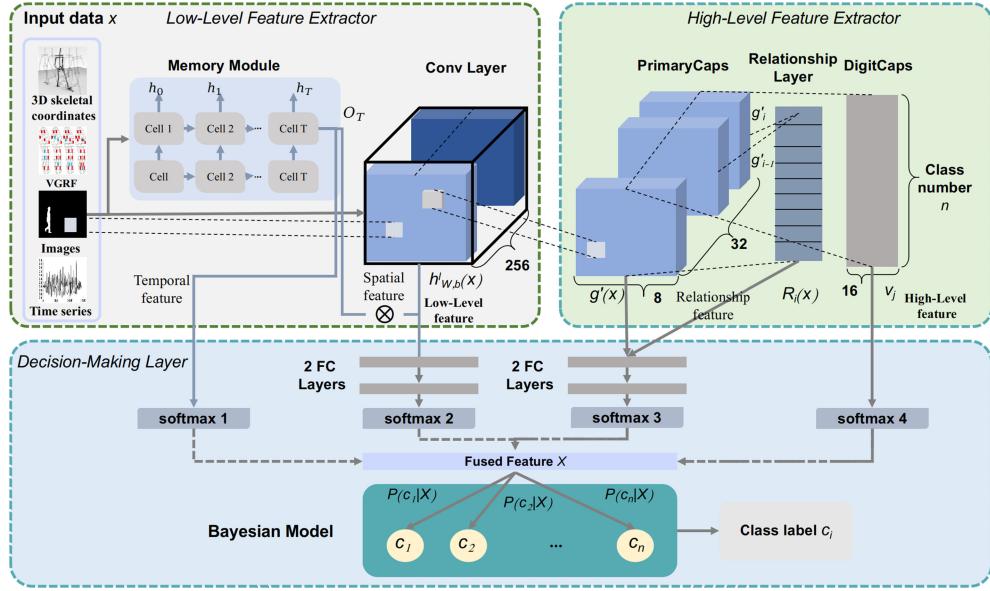


Fig. 1. The structure of ASTCapsNet. It is divided into three modules: A low-level feature extractor for spatio-temporal feature modeling, including a memory module and a convolution layer; a high-level feature extractor for learning structural and relationship features, including a primary capsule layer, a relationship layer, and a digit capsule layer; and a decision-making layer for classification, including four softmax classifiers and a Bayesian model. Four softmax classifiers behind the first two modules are utilized to evaluate the system performance. The Bayesian model utilizes the outputs of the previous two feature extractors to determine the final class labels.

### C. Model-Based Approaches

With remarkable advantages that deep learning plays in various fields, gait recognition technologies have also employed deep learning algorithms to model the dynamic changes of gait. When observing human walking, people can describe the structure of the human body and detect motion patterns of human limbs. The structure of the human body can be interpreted based on their prior knowledge. Model-based gait recognition methods play an active role in recent development.

Feng *et al.* proposed a new feature learning method to learn temporal information in a gait sequence for cross-view gait recognition. Heat maps extracted by a CNN-based pose estimation method were used to describe the gait information in every frame, and the standard LSTM was adopted to model a gait sequence [26].

Wu *et al.* utilized deep CNNs for human identification via similarity learning, which can recognize the most discriminative changes of gait patterns [10]. In order to extract gait spatio-temporal features, CNN and LSTM have also been simultaneously applied to establishing models for gait recognition [27]. Although these methods explain gait patterns well, the characteristics of gait are not only the stacking of features, but also include structural and positional information, such as the positional relationship between arms and legs, and so on.

The activities of the neurons within CapsNet represent the various properties of a particular entity that is present in the data. These properties can include many different types of instantiation parameters such as pose, deformation, velocity, and texture, which can improve gait recognition. Even though Capsule Network has been applied [28], [29], gait recognition is seldomly involved.

In this paper, we propose an associated spatio-temporal capsule network (ASTCapsNet) to process the input gait data and learn more powerful multi-level structured representations at each class of the streaming gait data. The newly proposed low- and high-level feature extractors enhance the capability of our framework for gait recognition in different gait streams. Moreover, we introduce softmax classifiers and the Bayesian model to evaluate the importance of the features generated by each module, which increases the accuracy of classification. Furthermore, we extensively evaluate the proposed gait recognition framework over five datasets, including the KinectUNIITO, SDUgait, CASIA-A datasets for people identification, and NDDs, PD datasets for disease diagnosis.

### III. ASSOCIATED SPATIO-TEMPORAL CAPSULE NETWORK

We introduce the proposed architecture, associated spatio-temporal capsule network (ASTCapsNet), for gait recognition in this section. The overall schema of this method is illustrated in Fig. 1. Here, we take the input gait images as an example to illustrate that other sensor data also follows this step. Firstly, the image is fed into the memory module and the convolution layer at the same time, and the generated spatio-temporal features are used as the input of softmax 1 and 2. Then, after the high-level feature extractor receives the low-level spatio-temporal features, the capsule layer and the relationship layer are used to extract the high-level structure and relationship features as the input of softmax 3 and 4, and the decision-making module utilizes the fused feature for the final decision.

In the proposed network, the inputs of ASTCapsNet are continuous sensor output data at each time step including images, force-sensitive data, 3 d skeleton joint data, and time series.

The structure of the gated recurrent unit (GRU) is performed in the time-domain to model the gait dynamics over data frames. To effectively extract high-level features from low-level features and determine the relationship between them, a dynamic routing algorithm is used for weighing operations. To better deal with different levels of feature extracted by the network, a Bayesian modeling mechanism is also introduced to process the output of each layer for our network.

On the one hand, for extracting spatial information from gait, CNN (convolutional neural network) is one of the most popular methods which can implement feature extraction by accumulating adjustments layer by layer. However, in this process, there is an important information loss - the relative position and relationship between features [20]. Precisely it is since the pooling operation in CNN can only provide rough position information, allowing the model to ignore small spatial changes, and cannot accurately learn the position correlation of different objects, and to this end, the capsule network (CapsNet) emerged.

On the other hand, for capturing the temporal information, gated recurrent unit (GRU) is a recurrent neural network that can be expanded into several cells (a chainlike model), which can extract information at each time step, and the connection between hidden units is iterative.

#### A. Low-Level Feature Extractor

Capsule network (CapsNet) [20] has verified their superior strength in extracting a spatial feature from different data types [28]. Inspired by the success of these approaches in the analysis of different data, we use a convolutional mechanism and capsule structure in CapsNet to extract spatial features. Besides, by adopting the advantages of processing temporal sequential data of GRU, we design a temporal mask added in each layer to represent the unique spatial and temporal features. Specifically, a hierarchical architecture with a spatio-temporal feature extraction layer is utilized in our model to learn a comprehensive representation.

To introduce the ASTCapsNet recurrent neural networks, we begin with the definition of several notations. Suppose we are given an input sequence  $D = \{x_i \in \mathbb{R}^N, i = 1, 2, \dots, M\}$ , where  $M$  is the sample number and  $N$  is the feature dimension in each sample, with corresponding class label sequences  $L = \{y_i \in \mathbb{R}^1, i = 1, 2, \dots, M\}$ . For the task of gait recognition, our aim is to address three problems: (1) Successfully obtaining spatio-temporal feature. (2) Acquiring high-level features, and (3) seeking the final class label.

1) *Convolution Layer*: The first building block of our network model is the convolution layer. The purpose of this layer is to use several sliding convolution kernels with a step size of 1 to filter the input data and extract multiple feature maps.

Firstly, we utilize  $L_2$  normalization to scale the original input  $D = \{x_i \in \mathbb{R}^N, i = 1, 2, \dots, M\}$  to  $[0,1]$  with the output of  $D'$ . And then, we feed  $D' = \{x'_i \in \mathbb{R}^N, i = 1, 2, \dots, M\}$  into memory and convolution modules, respectively. For the pre-processing of  $D'$ , we reshape each training sample in a format of  $K \times T = N$ , where  $K$  and  $T$  are the numbers of rows and columns of each sample so that convolution operation can

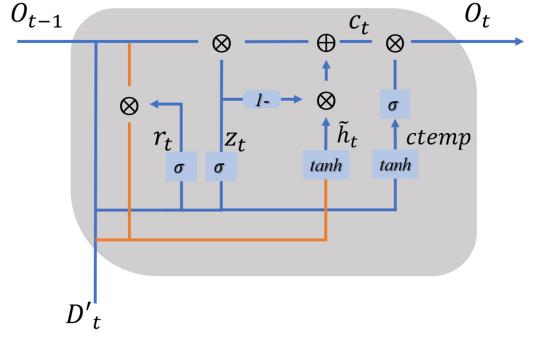


Fig. 2. The proposed recurrent memory unit. The internal structure and operations are shown in this figure, and  $ctemp$  is the new added path in the original GRU cell.

be carried out in the form of matrix no matter what the data type is. The sizes of the convolutional kernels must be smaller than  $K$  and  $T$  respectively. The feature map is demonstrated as follows:

$$h_{W,b}^l(x') = f\left(\sum_{i=1}^M W_i^{l-1} * x_i'^{(l-1)} + b^{l-1}\right). \quad (1)$$

The  $*$  operation above is essentially a process in which the convolution kernel (here the shared weight is the convolution kernel)  $W_i$  performs convolution operations on all the related feature maps at the  $(l-1)^{th}$  layer, then sums them up and adds a bias parameter  $b$  to obtain the final excitation value  $h_{W,b}^l(x')$  at layer  $l$  by taking sigmoid.

2) *The Proposed Recurrent Memory Unit*: We feed the original data into a two-layer memory module to extract temporal information and obtain the corresponding feature matrix. A variant of long short-term memory (LSTM), called the gated recurrent unit (GRU), is proposed to learn long-term dependency information from the input dynamic data, merging the input and forgot gates into an update gate and fusing node and hidden states in each node of the original LSTM, which was first proposed in [30]. We improve the performance of the original GRU cell by adding a temporary path and modifying the state settings. Fig. 2 shows the internal structure of the proposed recurrent memory unit and illustrates the cooperation of all the gates.

Because of slow convergence of the original GRU in gait data classification, we improve the gating structure in its internal deployment nodes by adding a new path and a temporary state to extract representative features. The improved GRU cell is shown in Fig. 2. The new cell of GRU is illustrated in Eqs. (2)–(7)

$$z_t = \sigma(W_z \cdot [O_{t-1}, x'_t]), \quad (2)$$

$$r_t = \sigma(W_r \cdot [O_{t-1}, x'_t]), \quad (3)$$

$$\tilde{h}_t = \tanh(W \cdot [r_t \odot O_{t-1}, x'_t]), \quad (4)$$

where  $x'_t$  and  $O_{t-1}$  are the input and previous hidden states respectively in each memory unit,  $\sigma$  and  $\tanh$  are the logistic functions.  $z_t$  indicates the output of update gate to the network at time step  $t \in \{1, 2, 3, \dots, T\}$ ,  $r_t$  denotes the reset gate. The update gate  $z_t$  determines whether or not the hidden state is to be updated with a new hidden state  $\tilde{h}_t$ . The reset gate  $r_t$  decides

whether or not the previous hidden state is ignored.

$$ctemp = \tanh(W_{ctemp} \cdot [O_{t-1}, x'_t]), \quad (5)$$

$$c_t = (1 - z_t) \odot \tilde{h}_t + z_t \odot O_{t-1}, \quad (6)$$

$$O_t = c_t \odot \sigma(ctemp), \quad (7)$$

where  $ctemp$  represents the temporary state to select the information of  $x'_t$  and  $O_{t-1}$  while  $c_t$  denotes the final state of the original GRU. It means that we take the advantage of the information selector to extract feature from  $c_t$ .  $O_t$  indicates actual activation of the proposed node at time step  $t \in \{1, 2, 3, \dots, T\}$ . Experiments show that it can converge faster than the original GRU.

Once having obtained the temporal information from the improved GRU, we use a stacking method to mask the output of each convolution layer with temporal features, forming a three-dimensional cube containing both temporal and spatial features. The representation is shown as follows:

$$g_{st}(x) = tile(O_t) + h_{W,b}^l(x'), \quad (8)$$

where  $g_{st}(x)$  is the function to obtain spatio-temporal features,  $tile(O_t)$  is the function to copy the input into multi-dimensional data, which is consistent with the output of  $h_{W,b}^l(x')$ , and then add it to  $h_{W,b}^l(x')$  to form a tensor  $g$ . We use a 3D convolutional kernels/capsules with a step size of 1 to get the output of the primary capsule layer  $g' = W_{conv}g$ .

### B. High-Level Feature Extractor

After having extracted temporal and spatial features (low-level features), we use capsules to process these features to obtain structural and relationship information (high-level features) of gaits, proposed to solve the problem of information loss on CNN. Different capsules represent different target positions, revealing the relative relationship between different spatial and temporal feature points. Then, according to the relationship between capsules, more important features are selected, and higher-level features are extracted by a dynamic routing algorithm.

In CapsNet, detailed posture information (such as exact locations, rotation, thickness, tilt and size of objects) will be saved in the network before we can restore it. Slight changes in the input will bring small changes to the output, and information will be saved. This allows CapsNet to use a simple and unified architecture for different visual tasks. Therefore, we design a capsule layer as the first layer in the high-level feature extractor.

1) *Capsule Layer*: The calculation process of the primary capsule layer is a stack of 8 parallel convolutional layers, each of which has a 3D convolutional kernel/capsule with a step size of 1.

The operation of this layer is similar to the convolution that is divided into different capsules. Since the output of the capsule is a vector, a powerful dynamic routing mechanism can be used to ensure that the output is sent to the appropriate parent in the previous layer. Therefore, the structure and direction of the feature are better reflected.

The input vector of the primary capsule layer is equivalent to the scalar input of traditional neural network neurons, and the calculation of the vector is equivalent to the mode of propagation and connection between the primary capsule and digit capsule.

The calculation of input vectors can be divided into two stages, i.e. linear combination and dynamic routing. The linear combination process can be expressed by the following formula:

$$\hat{g}'_{j|i} = W_{ij}g'_i, \quad (9)$$

where  $\hat{g}'_{j|i}$  is a linear combination of  $g'_i$ , which can be seen as the output of the first layer of neurons in the fully-connected network to a certain neuron in the next layer with different strength connections.  $\hat{g}'_{j|i}$  represents the output vector of the first capsule of the previous layer and the corresponding weight vector ( $W_{ij}$  represents the vector rather than the element).  $\hat{g}'_{j|i}$  can also be understood as the strength of connecting to  $j^{th}$  Capsule of the next layer when the former layer is the Capsule  $i$ .

After determining  $\hat{g}'_{j|i}$ , we need to use the dynamic routing algorithm to allocate the second stage to calculate the output  $s_j$ :

$$s_j = \sum_i c_{ij} \hat{g}'_{j|i}, \quad (10)$$

The parameter  $c_{ij}$  can be updated iteratively. The input  $s_j$  of the next capsule can be obtained by Routing, and then squashing non-linear function can activate the input  $s_j$ .

By adding the routing mechanism to the capsule, a set of coupling coefficients  $c_{ij}$  can be found, which are updated and determined iteratively by dynamic routing process:

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}. \quad (11)$$

which makes the prediction vector  $\hat{g}'_{j|i}$  consistent with the output vector.  $b_{ij}$  depends on the location and type of two capsules, and  $c_{ij}$  can be iteratively updated with the consistency of the measurement. The dissemination and distribution of the whole hierarchy can be divided into two parts. The first part is the linear combination between  $g'_i$  and  $\hat{g}'_{j|i}$  in the following figure, and the second part is the routing process between  $\hat{g}'_{j|i}$  and  $s_j$ . The information transmission and calculation between the two capsule layers are shown in Fig. 3.

Assuming there are two capsule units ( $g'_i$ ) in the primary capsule layer, and four capsules are passed from this level to the next level  $v_j$ .  $g'_1$  and  $g'_2$  are the tensors obtained from the spatio-temporal feature extraction layer, i.e. capsule units containing a group of neurons, which are multiplied by different weights  $W_{ij}$  to form  $\hat{g}'_{j|i}$ , respectively. The predictive vector is then multiplied by the corresponding coupling coefficient  $c_{ij}$  and passed to a specific digit capsule unit. The input  $s_j$  of different capsule units is the weighted sum of all possible incoming prediction tensors. Then we collect different input tensors  $s_j$ . By putting the input tensors into the squashing non-linear function, we have the output tensors  $v_j$  of the latter capsule unit. We use the product of the output vector  $v_j$  and the corresponding prediction vector  $\hat{g}'_{j|i}$  to update the coupling coefficient  $c_{ij}$ . Such an iterative update does not need to apply back-propagation.

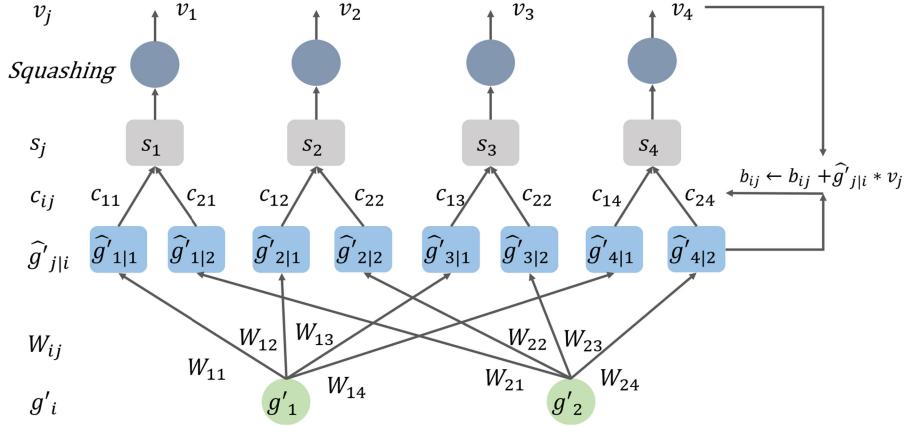


Fig. 3. The propagation process between the two capsule layers. There are three steps in this process: 3D convolution ( $\hat{g}'_{j|i}$ ), dynamic routing ( $s_j$ ), and squashing ( $v_j$ ).

The activation of neurons in a capsule indicates the properties of specific entities in data. These properties include different parameters, such as gait data position, size, direction, deformation, speed, reflectivity, color, and texture. The length of the input and output tensors represents the probability of the occurrence of a feature, so its value must be between 0 and 1.

In order to achieve this compression and implement the activation function at the capsule level, a non-linear function called “squashing” ensures that the length of short tensors can be reduced to almost zero, while the length of long tensors can be reduced to near but no more than 1. The following is the expression of the non-linear function:

$$v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|}, \quad (12)$$

where  $v_j$  is the output vector of capsule  $j$ ,  $s_j$  is the weighted sum of all the tensors of the capsule output from the previous layer to capsule  $j$  of the current layer, and  $s_j$  is the input vector of capsule  $j$ . The non-linear function can be divided into two parts. The first part  $\frac{\|s_j\|^2}{1 + \|s_j\|^2}$  is the scalar of the input vector  $s_j$ , and the second part  $\frac{s_j}{\|s_j\|}$  is the unit vector of the input vector  $s_j$ . The non-linear function not only retains the direction of the input vector but also compresses the length of the input vector into the interval [0,1]. The non-linear function can be regarded as compression and reallocation of the length of the vector, so it can also be regarded as a way of activating the output vector after the input vector.

Obtaining the output high-level feature of the digit capsule layer, we use the *softmax* function for classification. In order to comprehensively consider the effectiveness of feature extraction of the whole network, we classify the output features of each level and construct a Bayesian model to determine the final classification results using those predictions.

2) *Relationship Layer*: In terms of the primary capsule layer, all the capsules are independent of each other. The extracted features can not only represent the local features (arms, legs, or head) of gait, but also interpret the relationship between local and global features (the whole body). However, the relationship

between capsules has been neglected, so we design a capsule relationship layer to preserve the relationship between capsules, that is, the relationship between local features.

In order to preserve the relative position, direction, and state between local features, we use the transfer matrix of the former capsule and the latter capsule to represent their relationship and send them to the *softmax* classifier. The specific calculation is as follows:

$$R_i = g'_i * (g'_{i+1})^{-1}, \quad i < 32. \quad (13)$$

Assuming  $R_i$  is the relationship matrix of the capsules in the primary capsule layer.  $g'_i$  and  $g'_{i+1}$  are adjacent capsules. We use the inverse of  $(g'_{i+1})$  to find the transfer matrix  $R_i$ .

### C. Decision-Making Layer

According to the output of different feature extraction layers, we add four classifiers to evaluate the features of the four layers and output specific class labels. For the classification of the extracted features, the Bayesian model is built for the output of the four classifiers to select more accurate classification results.

1) *Classification Layers*: Firstly, we embed a softmax multi-classifier on the GRU feature extractor to evaluate whether or not temporal features are beneficial to classification. Furthermore, to make spatial and temporal features more distinct and easy to classify, we add two fully-connected layers to the convolutional layer, followed by a softmax function to classify the result, and integrate the feature cubes to obtain high-level information.

2) *Bayesian Modeling Process*: After classification, we have four label vectors. We merge these four results into a label matrix, regard the four label values predicted as features, and use the Bayesian model to maximize the posterior probability of class estimation. Assuming that there are  $n$  classes, using  $c_1, c_2$ , and  $c_n$  to denote,  $X = x_1, x_2, \dots, x_n$  represents each sample in a training batch with a certain label  $c_i, i \in [1, n]$ .

$$\max P(c_i|X) = \max \frac{P(X|c_i)P(c_i)}{P(X)}. \quad (14)$$

Because  $P(X)$  is constant for all the classes, the process of maximizing a posteriori probability  $P(c_i|X)$  can be transformed into maximizing probability  $P(X|c_i)P(c_i)$ . It is usually assumed that the attributes are independent of each other, and prior probability  $P(x_1|c_i), P(x_2|c_i), \dots, P(x_n|c_i)$  can be obtained from the training dataset. According to this method, the probability  $P(X|c_i)P(c_i)$  that  $X$  belongs to each category of  $c_i$  can be calculated for a sample  $X$  of an unknown category, and then the category with the greatest probability can be selected as its class label.

#### D. Loss Function

Combining standard categorical cross-entropy loss of the first three outputs of ASTCapsNet, and the margin loss  $l_{dc}$  of the last output of the digit capsule layer, the final joint loss function is formed as:

$$\begin{aligned} l_{dc} = & T_c \max(0, m^+ - \|v_c\|)^2 \\ & + \lambda(1 - T_c) \max(0, \|v_c\| - m^-)^2, \end{aligned} \quad (15)$$

where  $c$  is the class label of a sample,  $T_c$  is the indicate function (1 if  $c$  exists, else 0),  $m^-$  is the top margin of  $\|v_c\|$ ,  $m^+$  is the bottom margin of  $\|v_c\|$ .

$$\begin{aligned} \mathcal{L}(y, y') = & l_{tp} + l_{st} + l_{pc} + l_{dc} \\ = & - \sum_j y'_{tpj} \log \left( \frac{e^{y_{tpj}}}{\sum_{i=1}^n e^{y_{tpi}}} \right) - \sum_j y'_{stj} \log \left( \frac{e^{y_{stj}}}{\sum_{i=1}^n e^{y_{sti}}} \right) \\ & - \sum_j y'_{pcj} \log \left( \frac{e^{y_{pcj}}}{\sum_{i=1}^n e^{y_{pci}}} \right) + \sum_c l_{dc}, \end{aligned} \quad (16)$$

where  $l_{tp}$  and  $l_{st}$  represent the classification loss of temporal and spatio-temporal data,  $l_{pc}$  denotes the classification loss in the primary capsule layer.  $l_{dc}$  is the margin loss in the digit capsule layer.  $\mathcal{L}(y, y')$  is the joint loss of the four classification results. In the optimization process, we use Adam optimizer to consider the first and second moments of the gradients in order to reduce the loss quickly.  $y'_{tpj}, y'_{stj}, y'_{pcj}$  denote the  $j^{th}$  true label of a training batch while  $y_{tpj}, y_{stj}, y_{pcj}$  represents the predicted label.

## IV. EXPERIMENT

The experiment was conducted on five different datasets, and the results of the experiment were evaluated against several state-of-the-art methods. The following is a list of all the comparison methods:

#### Classifiers:

- 1) DT (decision tree) refers to a decision support tool using a tree-like model of decisions and their possible consequences.
- 2) GBDT (gradient boosting decision tree) is an iterative decision tree algorithm, which is composed of multiple decision trees. The results of all the trees are accumulated to determine the final label.

3) LR (logistic regression) is a kind of nonlinear regression models and a machine learning method for probability estimation.

4) RF (random forest) is a classifier with multiple decision trees.

5) KNN (k-nearest neighbor) is a classifier that finds  $k$  nearest instance and votes to determine the class name of the new instances.

#### Deep Models:

- 1) GRU (gated recurrent unit) is a gating mechanism in recurrent neural networks, which has fewer parameters than LSTM.
- 2) LSTM (long-short term memory) is a model with expandable nodes that are suitable for temporal data.
- 3) BiLSTM (bidirectional long-short term memory) is composed of a forward LSTM and a backward LSTM.
- 4) CapsNet (capsule network) is a deep neural network for dynamic analysis of data structure and spatial information.
- 5) Attention LSTM is an LSTM model with an attention mechanism.
- 6) CNN (convolutional neural network) is a network for learning spatial features.

Additionally, we also use several gait recognition methods from the literature because they use the dataset mentioned in this paper.

#### Other Studies:

- 1) TL-LSTM [18] is a two-layer LSTM for gait recognition.
- 2) HCF+SVM [31] is a method using hand-crafted features and SVM for classification.
- 3) GCP [32] is a gait cycle partitioning method.
- 4) WT+GA [33] is the combination of wavelet transform and a genetic algorithm.
- 5) DT+RF [34] is the combination of distance transform and random forests.
- 6) SD [35] (static and dynamic feature extraction method) is a human walking model including the static and dynamic gait features.
- 7) FLM [36] (frame-level matching) is a frame-level matching method for gait recognition.
- 8) RBF [37] (radial basis function) is a neural network for gait analysis and recognition.
- 9) DCLSTM [19] (dual-channel LSTM) is a temporal LSTM model for multimodal gait recognition.
- 10) Q-BTDNN [38] (Q-backpropagated time-delay neural network) is presented to identify the gait disturbances in Parkinson's disease.
- 11) 2D-CNN+LSTM [39] is the combination of 2D-CNN and LSTM for gait spatio-temporal information extraction.
- 12) SFM [16] (state-frequency memory) is a recurrent network that allows separating dynamic patterns across different frequency components and their impacts on modeling the temporal contexts of the sequences.
- 13) CapProNet [29] (capsule projection network) can learn an orthogonal projection matrix for each capsule subspace to classify different objects.

### A. Experimental Setup

First, we introduce the datasets used in our experiments and then describe the training details.

**1) Datasets:** We perform experiments on three gait datasets (normal people), including CASIA Gait dataset [40], UNITO dataset [24], and SDUgait dataset [41], and also verify their extensibility on two gait datasets (neurodegenerative patients) - NDDs dataset [42] and PD dataset [43]. These datasets are collected using different sensors, such as cameras, infrared sensors, and force sensors. The data types include human skeletons, gait contour maps, and so on. We will introduce these datasets in detail.

**UNITO Dataset:** This includes skeleton data of 20 subjects, acquired with Kinect For Windows SDK 1.x. The subjects were asked to walk naturally along a corridor, towards the camera (FRONT view) and away from the camera (REAR view), for 10 times, with a total of 20 gait sequences for each subject (10 FRONT and 10 REAR). The order of the 20 joints is the same as the Kinect Skeleton Map. For each joint, 3D coordinates of subjects are recorded.

**CASIA Gait Dataset:** In this dataset, we choose Dataset A (former NLPR Gait Dataset), which includes 20 persons. Each person has 12 preprocessed gait contour sequences, 4 sequences for each of the three directions, i.e. parallel, 45 degrees, and 90 degrees to the image plane. The length of the sequences is not identical for the variation of the walker's speed, but it must range from 37 to 127.

**SDUgait dataset:** The dataset includes 52 subjects, 28 males and 24 females with an average age of 22. Each subject has 20 sequences with at least 6 fixed walking directions and 2 arbitrary directions, and a total of 1040 sequences; The dataset is based on the second-generation Kinect (Kinect V2). The Kinect V2 does not only have a broader view angle but also can produce higher-resolution of depth images with 25 body joints.

**NDDs dataset:** The dataset contains the gait signals of 48 patients with NDDs (ALS, HD, and PD) and 16 healthy controls (CO). Participants were walking at their usual pace along a 77-m-long hallway for 5 minutes including the recorded signal of stride, swing, and stand times for each leg and double support signals for both legs. An expert physician labeled patients' states from 0 to 13 (0 equal to the most severe state and 13 for a healthy one). To measure time intervals, a 12-bit onboard analog-to-digital converter samples the output of foot switches at 300 Hz [44].

**PD Dataset:** In this study, we have utilized gait signals from PhysioNet. The dataset consists of three PD gait sub-datasets, which are contributed by three researchers (Ga [45], Ju [46] and Si [47]). We only select the subset of Ga [45] as an example of PD severity classification.

The dataset includes gait information from 93 patients with the idiopathic PD and 73 healthy controls (average age 66.3, 55% male). Every participant was asked to walk at their usual, self-selected pace for about two minutes while wearing a pair of shoes with 8 force sensors located under each foot. The sensors measure vertical ground reaction force (VGRF, in Newton) as a function of time at 100 samples per second. The dataset

also contains the specific situation of each participant including gender, age, height, weight, and severity level of PD. The PD severity level is graded according to two scales (H&Y, UPDRS).

**2) Training Details:** We employ different training settings on these datasets. For the UNITO Dataset, we use 5\*5 and 3\*3 convolutional kernels in the convolutional layer and primary capsule layer, 7-time steps in LSTM with the batch size of 128, and input dimension of 9 (63D, 21 joints, 3D coordinates), the dropout in LSTM is set to 0.5 to avoid over-fitting.

For CASIA Gait Dataset, we apply all of three directions of the gait contour images and convert the image into text sequences. After resizing the original image to 50 by 50 pixels, the gait data is fed into ASTCapsNet with 9\*9 convolutional kernels in the convolutional layer and the primary capsule layer. The time step and input vector are set to 50 and 50 to match the image size. The learning rate is 0.001 in LSTM and other components. The input and time steps in LSTM are 50 and 50 respectively.

For the SDUgait dataset, we only use the skeleton data of the gait, which includes the 3D positions of 21 joints. Differently, the time step in LSTM is set to 7, and the input is 9D (63D, 3\*21), that is, the dimension of the input in LSTM is 63. 5\*5 and 3\*3 convolutional kernels can be used in the convolutional layer and the primary capsule layer respectively.

For the NDDs dataset, the input matrix of a training sample is 12\*10 that the input is 12D and the time step is 10. The size of the convolutional kernels is set to 5\*5 based on the scale of the input data. The hidden output of LSTM is 128D to represent the temporal feature of the gait.

For the PD Dataset, the size of a training sample is 19\*100, which indicates that the input is 19D and the time step is 100. The size of the convolutional kernels is set to 9\*9 based on the scale of the input data. The hidden output of LSTM is 128D to represent the temporal feature of gait.

The parameters of other layers are set according to different datasets and fine-tuned to achieve good performance. The model is implemented with Tensorflow.

### B. Experiments on Normal Gait Datasets

In this subsection, we evaluate the ASTCapsNet model on normal gait datasets for identity recognition: UNITO, CASIA and SDUgait datasets.

**1) Experiments on the UNITO Skeleton Dataset:** This experiment aims to verify that our proposed model can perform well in the three-dimensional skeletal dataset, which outperforms the other methods on this dataset.

Firstly, in this dataset, we have 20 subjects for classification, and use 80% of the data for training and 20% for testing. The performance of the proposed model is shown in Fig. 6. It is obvious that the gait recognition rate of the model is higher than 99.0% for each of 20 subjects, and we finally achieve the test result of 99.65%.

Furthermore, we compare our model with several advanced algorithms for this dataset in recent years, which are demonstrated in Table I. In the light of the results of five traditional classifiers, i.e., decision tree (DT), gradient boosting decision tree

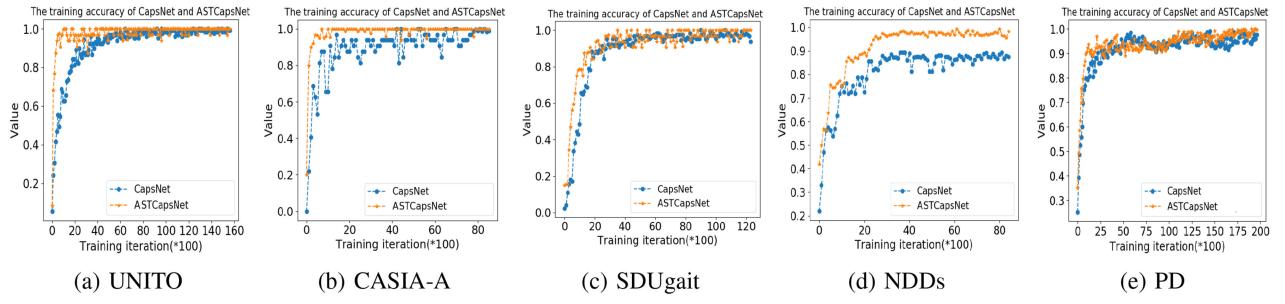


Fig. 4. The training accuracy of the original CapsNet and ASTCapsNet on the five datasets.

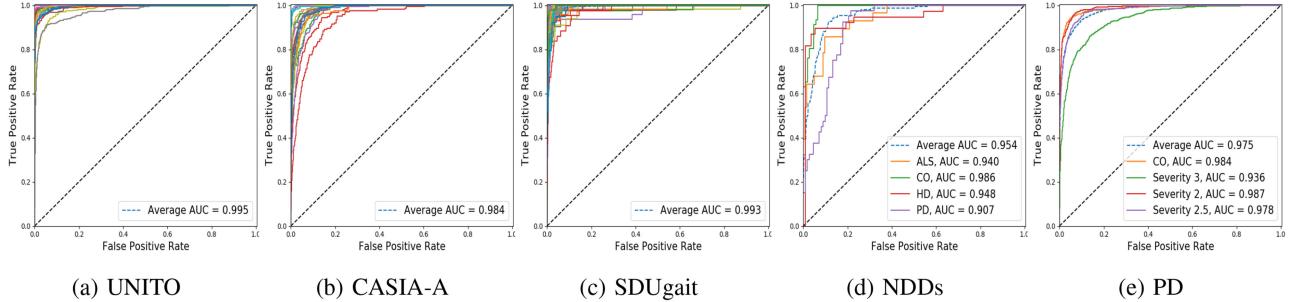


Fig. 5. ROC curves of ASTCapsNet on the five datasets. The legend shows the AUC values of different classes in the dataset, and there are too many classes of UNITO, CASIA-A, and SDUGait datasets to be fully represented by the legend, so we only give average AUC values. Each line represents the AUC value of the corresponding class. ASTCapsNet has a high recognition rate of each class and the curve is concentrated in the upper left corner. Besides, the micro-average AUC values are compared, which shows the strong learning ability of ASTCapsNet.

TABLE I  
RESULTS ON THE UNITO SKELETON DATASET

Classifiers	Acc	Deep Models	Acc	Others	Acc
DT	87.29%	GRU	97.95%	TL-LSTM [18]	97.33%
GBDT	90.82%	LSTM	95.44%	HCF+SVM [31]	97.00%
LR	67.51%	BiLSTM	96.45%	SFM [16]	96.97%
RF	96.47%	CapsNet	97.33%	CapProNet [29]	98.69%
KNN	90.00%	Attention LSTM	97.92%		
ASTCapsNet	<b>99.65%</b>	CNN	95.84%		

(GBDT), logic regression (LR), random forest (RF), k-nearest neighbor (KNN), and temporal models, i.e., gated recurrent unit (GRU), long-short term memory (LSTM), attention LSTM and bi-directional long short-term memory (BiLSTM), we can see that the average result of the temporal model is higher than that of the traditional classifiers, and the performance of RF and attention LSTM is the best among the models.

For other studies, Gianaria *et al.* [31] uses handcrafted features as input to the SVM classifier and achieves the result of 97%, which needs to calculate the relative distance between joints (FoRD) and the sway of joints (FoS). Li *et al.* [18] apply a two-layer LSTM in this dataset and obtain the result of 97.33% which considers the temporal changes of gait. The performance of SFM [16] exceeds LSTM and BiLSTM, indicating the advantages of using state-frequency components. Although the original CapProNet [29] can perform better than these methods by applying capsule subspace, our model achieves the highest accuracy among them.

Additionally, we also show the training accuracy of the original CapsNet and ASTCapsNet in Fig. 4. We can see that ASTCapsNet converges faster than the original CapsNet, and the final training accuracy is closer to 100%. Fig. 5 shows the receiver operating characteristic curve (ROC) for the evaluation of the proposed model. The curve is close to the upper left corner, and the average area under curve (AUC) value approaches to 1, indicating ASTCapsNet achieves satisfactory performance.

2) *Experiments on the CASIA Gait Dataset:* This experiment aims to verify that our proposed model can perform well in the 2D silhouette image dataset, which also outperforms the other methods for this dataset.

In this dataset that includes 20 subjects for identification, we use 80% of the data for training and 20% for testing. The performance of the proposed model is shown in Fig. 7. The correct number and accuracy of the samples on the diagonal line show that the model can accurately determine the identity via gait.

Beyond that, we also compare the experimental results in the past five years, and the model presented in this paper also stands out, which are illustrated in Table II. Because of the temporality of the gait data, the temporal model is more suitable for learning, which achieves better results. Hong *et al.* [32] use a gait cycle partitioning method for gait representation. Distance transform and random forest are also compared for this task [34]. [33] propose to use wavelet transform and a genetic algorithm for identification, and the result is also shown in Table II. In general, the temporal models are better than the other models, where capsule networks, i.e., CapsNet and CapProNet can effectively analyze the structural and spatial information of gait.

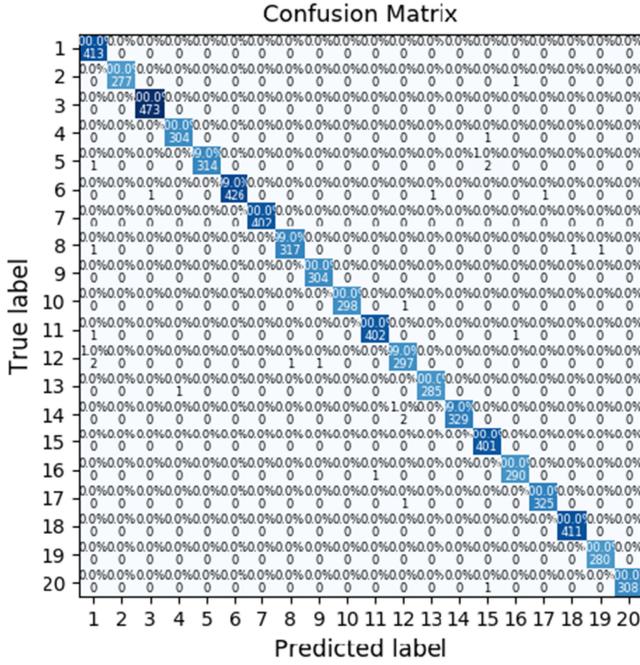


Fig. 6. The confusion matrix of classification results on the UNITO Skeleton Dataset.

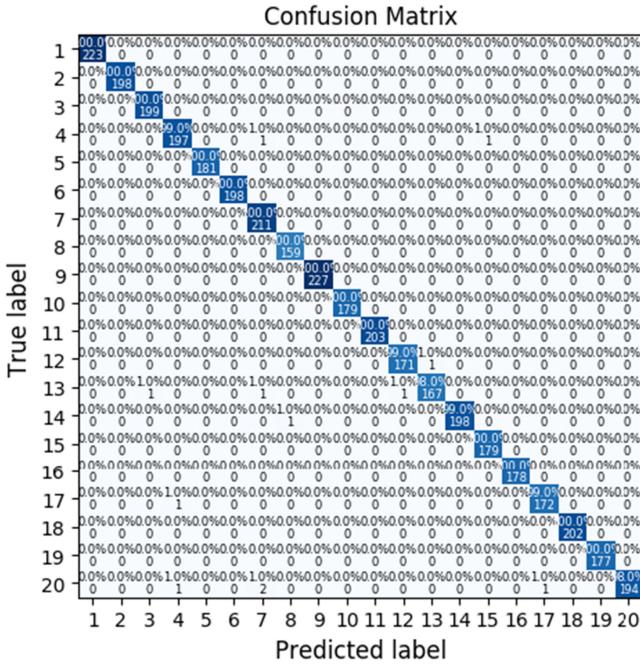


Fig. 7. The confusion matrix of classification results on the CASIA-A Gait Dataset.

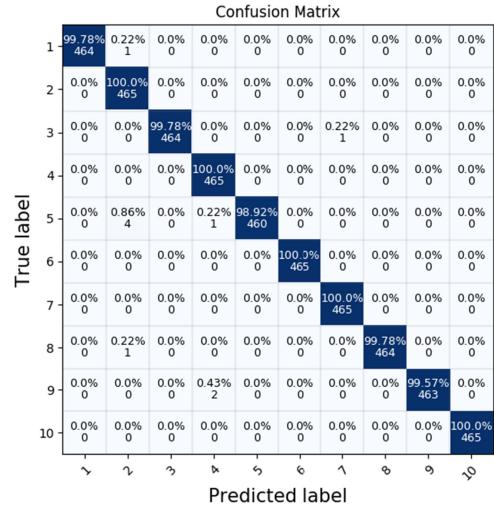


Fig. 8. The confusion matrix of the classification results on the SDUGait Dataset. Since the recognition results of 52 subjects cannot be completely shown due to the page size, we select the subjects labeled 1–10 to show the recognition results which is 99.13%.

TABLE III  
RESULTS ON THE SDUGAIT GAIT DATASET

Classifiers	Acc	Deep Models	Acc	Others	Acc
DT	80.33%	GRU	95.30%	SD [35]	94.23%
GBDT	85.02%	LSTM	94.80%	FLM [36]	98.08%
LR	63.08%	BiLSTM	96.38%	SFM [16]	95.02%
RF	94.26%	CapsNet	97.00%	CapProNet [29]	97.64%
KNN	85.28%	Attention LSTM	96.33%	CNN	92.85%
ASTCapsNet	<b>99.13%</b>				

Fig. 4 shows the training accuracy of the original CapsNet and ASTCapsNet on the CASIA-A Dataset. After the 5<sup>th</sup> training iteration, ASTCapsNet achieves higher accuracy than the original CapsNet and quickly converges to nearly 100%. The average AUC value of the CASIA-A dataset is 98.4%, the true positive rate (TPR) approaches 1 when the false positive rate (FPR) is 0.3, which verifies the effectiveness of the ASTCapsNet (Fig. 5).

3) *Experiments on the SDUGait Dataset:* This experiment aims to verify that our proposed model can perform well in the 3D skeletal dataset, which also outperforms the other methods applied to this dataset.

Firstly, in this dataset that contains 52 subjects for classification, we use 80% of the data for training and 20% for testing. The performance of the proposed model is shown in Fig. 8. There are so many categories that the recognition results of each category are not shown and the outline of the color confusion matrix is displayed. We can also see that the recognition error rate of the model for different gaits is very small.

Moreover, we compare this result with some advanced algorithms for this dataset, which are shown in Table III. The average result of the temporal model is much higher than that of the traditional classifier, and RF and BiLSTM perform the best. Wang *et al.* [35] obtains the classification result of 94.23% on this dataset by using the static and dynamic feature extraction methods, being view-invariant for gait recognition. Although we only use the 3D coordinates of the skeletal joints in the SDUGait

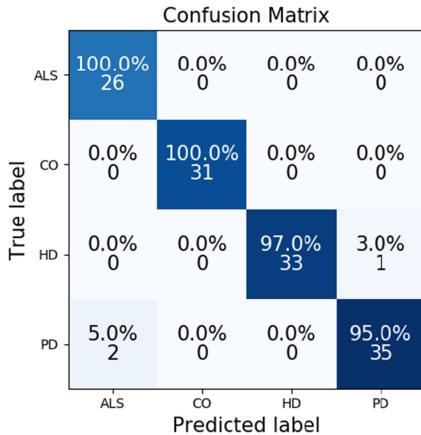


Fig. 9. The confusion matrix of classification results on the NDDs Dataset.

TABLE IV  
RESULTS ON THE NDDS DATASET

Classifiers	Acc	Deep Models	Acc	Others	Acc
DT	49.34%	GRU	92.76%	RBF [37]	93.75%
GBDT	67.11%	LSTM	92.11%	DCLSTM [19]	95.67%
LR	51.32%	BiLSTM	93.82%	SFM [16]	93.59%
RF	65.13%	CapsNet	88.86%	CapProNet [29]	92.11%
KNN	73.03%	Attention LSTM	93.41%		
ASTCapsNet	<b>95.78%</b>	CNN	85.58%		

dataset, the proposed model can also well learn the gait patterns. Besides, Choi *et al.* [36] propose a robust frame-level matching method for gait recognition that minimizes the influence of noisy patterns as well as secures the frame-level discriminative power and achieves accuracy of 98.08%, little lower than ASTCapsNet. The SFM [16] here does not show outstanding advantages due to the sparsity of gait data and unobvious frequency variation. The result of CapProNet [29] is still slightly higher than that of CapsNet.

Furthermore, we also show the training accuracy of the original CapsNet and ASTCapsNet in Fig. 4. The curves of the training accuracy on this dataset are similar. In the 20<sup>th</sup> training iteration, ASTCapsNet begins to be slightly higher than CapsNet until the end of the training, indicating the stability of the model. Fig. 5 shows the ROC curve of the proposed ASTCapsNet. We can see that except for one subject with a high detection error rate, the AUC of the other subjects is more than 97%, reflecting the high performance of the model.

### C. Experiments on Abnormal Gait Datasets

In this subsection, we evaluate the ASTCapsNet model on the abnormal gait datasets for disease classification and diagnosis: NDDs and PD datasets.

1) *Experiments on the NDDs Dataset:* In this Dataset, our model is used to identify 3 diseases and 1 healthy control group (ALS, HD, PD, and CO). The performance is explained in Fig. 9. CO and ALS obtain the result of 100% while two samples of PD are misclassified into ALS.

Moreover, there are several comparative experiments implemented on this dataset, as illustrated in Table IV. The five classifiers can not provide a good reference because of the small

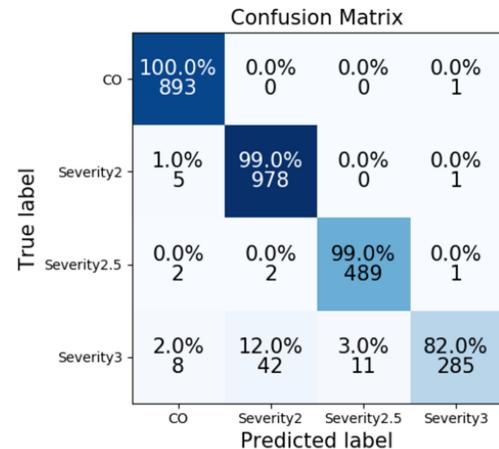


Fig. 10. The confusion matrix of classification results on the PD Dataset.

amount of data, and the time series model can capture the dynamic changes of the gait data by more than 90%. It can be seen that the effect of CapsNet on this dataset is under 90%, showing a slightly weaker classification ability than CapProNet [29]. The effect of the two-channel LSTM using two different sensor data is only inferior to that of ASTCapsNet.

Apart from this observation, we also show the training accuracy of the original CapsNet and ASTCapsNet in Fig. 4. The proposed model is approximately 10% higher than the original CapsNet because the temporal LSTM and Bayesian model are added to compensate for the loss of the features caused by CapsNet. Moreover, the small amount of data does not allow the original CapsNet to learn well. As shown in Fig. 5, the ROC curves of CO tend to be 1, while the curves of ALS, PD, and HD slightly deviate, indicating that the model is more sensitive to the classification of CO due to the big difference between healthy people and patients.

2) *Experiments on the PD Dataset:* In the PD Dataset, we only use the Ga [45] subset as an example to classify the severity of Parkinson's disease based on the gait performance. There are three categories of severity according to the criteria of H & Y scale (2 = Bilateral involvement without impairment of balance, 2.5 = Mild bilateral disease with recovery on the pull test, 3 = Mild to moderate bilateral disease; some postural instability; physically independent), and the four categories are compared with the healthy control group.

After training ASTCapsNet, the performance is shown in Fig. 10. The model is 100% for healthy people and 82% for patients with severity 3. We witness that the model does not distinctly extract patient features for severity 3, which leads to misclassification.

Compared with other experiments on this dataset, our proposed model achieves higher classification accuracy. The details are illustrated in Table V. The results of RF and KNN are better, and GRU also shows advantages in temporal models. Q-BTDNN (Q-backpropagated time-delay neural network) is designed for severity classification which can diagnose the gait disturbances in Parkinson's disease. Another study [39] uses both 2D-CNN and LSTM to obtain a promising result of 96.00%, while

TABLE V  
RESULTS ON THE PD DATASET

Classifiers	Acc	Deep Models	Acc	Others	Acc
DT	77.90%	GRU	94.89%	Q-BTDNN [38]	93.10%
GBDT	84.48%	LSTM	93.95%	2D-CNN+LSTM [39]	96.00%
LR	57.23%	BiLSTM	94.78%	SFM [16]	93.63%
RF	91.84%	CapsNet	95.30%	CapProNet [29]	96.58%
KNN	92.02%	Attention LSTM	93.53%		
ASTCapsNet	<b>97.31%</b>	CNN	89.76%		

CapProNet [29] performs better. ASTCapsNet still outperforms the original CapsNet and CapProNet [29].

Additionally, we also show the training accuracy of the original CapsNet and ASTCapsNet in Fig. 4. We can see that after the 125<sup>th</sup> training iteration, ASTCapsNet begins to achieve slightly higher training accuracy than CapsNet until the end of the iteration. The AUC value of the four-class classification for Parkinson’s disease is 97.5%, a high level of recognition (5). The AUC value of severity 3 is low, which means that the boundary between severity 3 and other severity degrees is not clear, and the difference is minor.

#### D. Performance of the Whole Model

In this section, we evaluate and compare the performance of the whole model and its internal components to verify the model’s ability to extract and classify features.

1) *Performance of Split Components:* First, we provide the recognition results of each module in Table VI. The performance of six individual components is demonstrated on the five datasets respectively. As shown in Fig. 1, the low-level feature extractor includes the memory and convolution modules for spatio-temporal feature extraction. The high-level feature extractor includes the capsule and capsule relationship layers for relationship and structure feature extraction. “No relationship layer” is the output of ASTCapsNet after we have removed the relationship layer, and “no memory module” is the output of ASTCapsNet after removing the memory module. “Bayesian” model is the original Bayesian model, and “Bayesian model (last layer)” is the total output of ASTCapsNet.

It can be seen that the average output of the high-level feature extractor is better than that of the low-level feature extractor on these five datasets, which reflects the superiority of the module integration scheme. The performance of the “no relationship layer” and the “no memory module” is similar, depending on the fitting degree of the model and data. However, the data does not conform to the Gaussian distribution, leading to the poor outcome of using the Bayesian model directly, and the output of the softmax layer conforms to the Gaussian distribution so that the Bayesian model can be placed in the last layer to retain higher accuracy. Briefly, in the comparison of all the individual modules, the high-level feature extractor makes the most, followed by the memory module, and then the relationship layer.

2) *Separability of the Extracted Features:* We discuss the separability of the features extracted by ASTCapsNet, and take the outputs of *softmax 2* and *softmax 4* in the low- and high-level feature extractors as an example. All the features are

processed by t-SNE (t-distributed stochastic neighbor embedding) for dimensionality reduction. As shown in Fig. 11, the first row represents the performance of the low-level extractor on the five datasets after dimensionality reduction, and the second row represents the performance of the high-level extractor after dimensionality reduction. We can clearly observe that the high-level features are more discriminative than the low-level features, and can represent the critical information of each class. The overlapping probability of low-level features is slightly larger than that of the high-level features. The high-level features have clear boundaries with the other classes after dimensionality reduction and have a high degree of aggregation with the samples in the same class, demonstrating the effectiveness of our proposed method.

## V. DISCUSSION

As shown in the experiment, ASTCapsNet can effectively learn the human gait pattern including healthy people and NDDs patients. By experimenting on five gait datasets, the highest accuracy of our model are 99.65%, 99.69%, 99.13%, 95.78%, 97.31% respectively, superior to the other gait recognition methods on the classification of these datasets. To evaluate the performance of each component in ASTCapsNet, we compare the classification results of each softmax layer with the output of the Bayesian model, shown in Table VII. We can find that the performance of softmax 4 surpasses the other layers because this layer contains rich features calculated from the previous layers. Softmax 3 also performs well, reflecting the validity of the relationship layer. Furthermore, the model parameters have been found to affect the performance of classification, and the optimal setting is reported in Section IV.

In the field of gait recognition, there is still a lack of public datasets that have both large view variations and diversity of data. This factor affects the possibility to train a unified and reliable model for practical scenarios in medical diagnosis. For example, if the data of one modality is missing in the training, the final classification system may not be able to effectively handle this modality in testing. Moreover, it is expensive to obtain the labeled gait sequences. In this section, we summarize several possible ways to improve the practicality and effectiveness of our proposed model as follows.

**A. Cross-view gait recognition.** In our ASTCapsNet, we only recognize the gait from different sources, without considering the relationship between multi-modal gait data. [48] and [49] have undertaken research and discussion on cross-view and multi-modal gait recognition. We will use canonical correlation analysis (CCA), kernel CCA (KCCA), and other data fusion methods to analyze the correlation between different sensors or views.

**B. Unsupervised or Semi-Supervised Learning.** The model discussed in this paper only refers to the encoder, which does not have the function of generating new samples and cannot deal with unlabeled gait data. How to utilize unlabeled samples for unsupervised or semi-supervised learning is still an open problem in this study. [48] reports multi-task generative adversarial

TABLE VI  
PERFORMANCE OF INDIVIDUAL COMPONENTS IN ASTCAPSNET

Dataset	Low-Level Feature Extractor	High-Level Feature Extractor	No Relationship Layer	No Memory Module	Bayesian Model (Raw Data)	Bayesian Model (Last Layer)
UNITO	95.87%	99.56%	97.41%	97.10%	77.54%	<b>99.65%</b>
CASIA-A	97.32%	99.23%	97.62%	95.73%	73.29%	<b>99.69%</b>
SDUgait	95.33%	98.56%	96.50%	97.94%	78.56%	<b>99.13%</b>
NDDs	93.64%	95.33%	94.14%	90.39%	60.85%	<b>95.78%</b>
PD	94.37%	96.38%	95.72%	95.87%	69.78%	<b>97.31%</b>

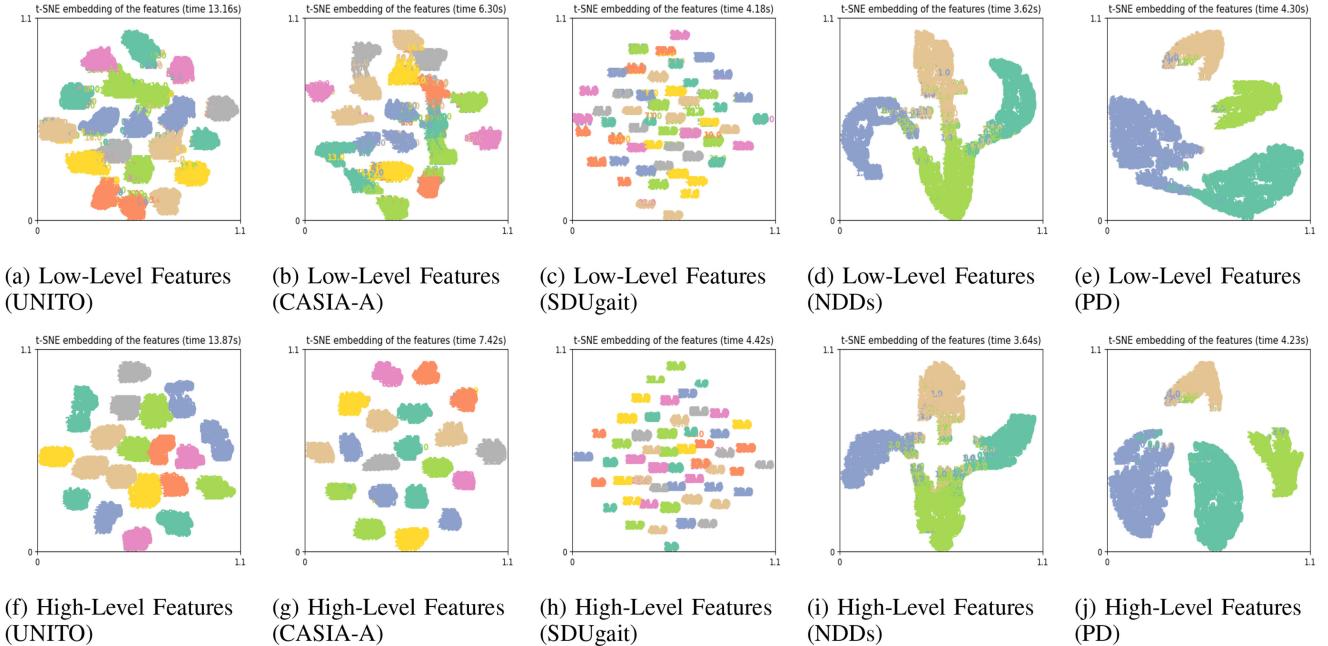


Fig. 11. Dimensionality reduction comparison between the extracted features and original data using t-SNE.

TABLE VII  
PERFORMANCE OF DIFFERENT LAYERS IN ASTCAPSNET

Dataset	Softmax 1	Softmax 2	Softmax 3	Softmax 4	Total
UNITO	95.99%	95.87%	99.12%	99.56%	<b>99.65%</b>
CASIA Gait	96.65%	97.32%	98.22%	99.23%	<b>99.69%</b>
SDUgait	95.30%	95.33%	98.77%	98.56%	<b>99.13%</b>
NDDs	93.86%	93.64%	94.98%	95.33%	<b>95.78%</b>
PD	94.02%	94.37%	96.43%	96.38%	<b>97.31%</b>

networks (MGANs) for learning view-specific feature representations, which provides a possible way to utilize unlabeled data in our model. We will explore to enhance our method by exploiting unlabeled information in future work.

## VI. CONCLUSION

In this paper, we have proposed an end-to-end learning model, ASTCapsNet, for gait recognition in multimodal gait sequences. The convolutional layer and a novel recurrent memory unit were introduced to model the dynamics and dependency in the spatio-temporal dimension for low-level feature extraction. Furthermore, a relationship layer was designed for high-level feature extraction, calculated between different capsules. Finally, a Bayesian model was also proposed for ASTCapsNet, with which our network can assess the quality of the features extracted from

each layer using softmax. Our proposed method yields superior performance on the benchmark datasets for handling the gait recognition problem. This network will be extended to address the task of multi-view gait detection and recognition for intelligently streaming scalable video sequences, which requires analyzing gait changes in the video sequences and predicts the class of each person in real-time.

## REFERENCES

- [1] T. Luan, 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.
- [2] J. Sun, W. Yang, J.-H. Xue, and Q. Liao, "An equalized margin loss for face recognition," *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 2833–2843, Nov. 2020.
- [3] 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.
- [4] W. Jia *et al.*, "Palprint recognition based on complete direction representation," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4483–4498, Sep. 2017.
- [5] T. Liu, X. Dong, Y. Wang, X. Dai, Q. You, and J. Luo, "Double-layer conditional random fields model for human action recognition," *Signal Process. Image Commun.*, vol. 80, 2020, Art. no. 115672.
- [6] C. Bregler, "Learning and recognizing human dynamics in video sequences," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 1997, pp. 568–574.

- [7] M. Ye, C. Yang, V. Stankovic, L. Stankovic, and S. Cheng, "Distinct feature extraction for video-based gait phase classification," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1113–1125, May 2020.
- [8] J. Liu, A. Shahroudy, D. Xu, A. K. Chichung, 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.
- [9] 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.
- [10] 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.
- [11] E. Hossain and G. Chetty, "Multimodal feature learning for gait biometric based human identity recognition," *ICONIP 2013: Neural Inf. Process.*, vol. 8227, pp. 721–728, 2013.
- [12] Qi and Guo-Jun, "Hierarchically gated deep networks for semantic segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 2267–2275.
- [13] X. Shu, J. Tang, G. J. Qi, Z. Li, Y. G. Jiang, and S. Yan, "Image classification with tailored fine-grained dictionaries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 2, pp. 454–467, Feb. 2018.
- [14] Y. Zhao *et al.*, "An adversarial approach to hard triplet generation," *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 501–517.
- [15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [16] H. Hu and G.-J. Qi, "State-frequency memory recurrent neural networks," in *Proc. 34th Int. Conf. Mach. Learn., ser. ICML2017*, 2017, pp. 1568–1577.
- [17] X. Shu, J. Tang, G.-J. Qi, W. Liu, and J. Yang, "Hierarchical long short-term concurrent memory for human interaction recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1110–1118, Mar. 2021.
- [18] J. Li, L. Qi, A. Zhao, X. Chen, and J. Dong, "Dynamic Long Short-Term Memory Network for Skeleton-Based Gait Recognition," in *Proc. IEEE SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI*, Aug 2017, pp. 1–6.
- [19] A. Zhao, L. Qi, J. Dong, and H. Yu, "Dual channel lstm based multi-feature extraction in gait for diagnosis of neurodegenerative diseases," *Knowl.-Based Syst.*, vol. 145, pp. 91–97, 2018.
- [20] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. 31st Conf. Neural Inf. Process. Syst.* 2017, pp. 3859–3869.
- [21] P. Chattopadhyay, S. Sural, and J. Mukherjee, "Frontal gait recognition from incomplete sequences using rgb-d camera," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 11, pp. 1843–1856, Nov. 2014.
- [22] Z. Xue, D. Ming, W. Song, B. Wan, and S. Jin, "Infrared gait recognition based on wavelet transform and support vector machine," *Pattern Recognit.*, vol. 43, no. 8, pp. 2904–2910, 2010.
- [23] S. Park, J. S. Lee, J. Kwak, and H. Shin, "Design of the wearable device for hemiplegic gait detection using an accelerometer and a gyroscope," in *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, pp. 1409–1412, 2017.
- [24] E. Gianaria, M. Grangetto, M. Lucenteforte, and N. Balossino, *Human Classification Using Gait Features*. Berlin, Germany: Springer, 2014.
- [25] V. O. Andersson and R. M. de Araújo, "Person identification using anthropometric and gait data from kinect sensor," in *AAAI*, 2015, pp. 425–431.
- [26] Y. Feng, Y. Li, and J. Luo, "Learning effective gait features using lstm," in *Proc. Int. Conf. Pattern Recognit.*, 2017, pp. 325–330.
- [27] M. Alotaibi and A. Mahmood, "Improved gait recognition based on specialized deep convolutional neural network," *Comput. Vis. Image Understanding*, vol. 164, pp. 103–110, 2017.
- [28] Y. Zhao, T. Birdal, H. Deng, and F. Tombari, "3 d point capsule networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1009–1018.
- [29] L. Zhang, M. Edraki, and G.-J. Qi, "Capronet: Deep feature learning via orthogonal projections onto capsule subspaces," in *Proc. Adv. Neural Inf. Process. Syst.*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 5814–5823.
- [30] J. Chung, C. Gulcehre, K. H. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, arXiv:1412.3555.
- [31] E. Gianaria and M. Grangetto, "Robust gait identification using kinect dynamic skeleton data," *Multimedia Tools Appl.*, vol. 78, no. 10, pp. 13 925–13 948, 2019.
- [32] S. Hong and E. Kim, "A new automatic gait cycle partitioning method and its application to human identification," *Int. J. Fuzzy Log. Intell. Syst.*, vol. 17, no. 2, pp. 51–57, 2017.
- [33] Y. Shi, X. Q. Wang, and Y. C. Zhang, "A recognition method of gait by wavelet transform and genetic algorithm," *Appl. Mechanics & Mater.*, vol. 701–702, pp. 274–278, 2015.
- [34] V. Ho, H. Huynh, and V. Ngo, "Direction-free person identification with distance transform and random forest," in *Proc. IEEE-RIVF Int. Conf. Comput. Commun. Technol.*, Mar. 2019, pp. 1–6.
- [35] Y. Wang, J. Sun, J. Li, and D. Zhao, "Gait recognition based on 3 d skeleton joints captured by kinect," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2016, pp. 3151–3155.
- [36] S. Choi, J. Kim, W. Kim, and C. Kim, "Skeleton-based gait recognition via robust frame-level matching," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 10, pp. 2577–2592, Oct. 2019.
- [37] W. Zeng and C. Wang, "Classification of neurodegenerative diseases using gait dynamics via deterministic learning," *Inf. Sci.*, vol. 317, pp. 246–258, 2015.
- [38] Y. N. Jane, H. K. Nehemiah, and K. Arputharaj, "A q-backpropagated time delay neural network for diagnosing severity of gait disturbances in parkinson's disease," *J. Biomed. Informat.*, vol. 60, pp. 169–176, 2016.
- [39] A. S. Alharthi and K. B. Ozanyan, "Deep learning for ground reaction force data analysis: Application to wide-area floor sensing," in *Proc. IEEE 28th Int. Symp. Ind. Electron.*, Jun. 2019, pp. 1401–1406.
- [40] C.-A. database, "Casia-a database," [Online]. Available: <http://www.cbsr.ia.ac.cn/china/Gait%20Databases%20CH.asp>.
- [41] SDU, "Sdu gait database," 2021. [Online]. Available: <http://mla.sdu.edu.cn/info/1006/1195.htm>.
- [42] physionet ndd, "physionet database," 2021. [Online]. Available: <https://physionet.org/physiobank/database/gaitndd/>.
- [43] physionet pd, "physionet database," 2021. [Online]. Available: <http://physionet.org/pn3/gaitpd/>.
- [44] J. M. Hausdorff, M. E. Cudkowicz, R. Firtion, J. Y. Wei, and A. L. Goldberger, "Gait variability and basal ganglia disorders: Stride-to-stride variations of gait cycle timing in Parkinson's disease and huntington's disease," *Movement Disord.*, vol. 13, no. 3, pp. 428–437, 1998.
- [45] G. Yogev *et al.*, "Dual tasking, gait rhythmicity, and parkinson's disease: Which aspects of gait are attention demanding?" *Eur. J. Neurosci.*, vol. 22, no. 5, pp. 1248–1256, 2005.
- [46] J. M. Hausdorff *et al.*, "Rhythmic auditory stimulation modulates gait variability in parkinson's disease," *Eur. J. Neurosci.*, vol. 26, no. 8, pp. 2369–2375, 2007.
- [47] S. Frenkel-Toledo *et al.*, "Treadmill walking as an external pacemaker to improve gait rhythm and stability in parkinson's disease," *Movement Disord.*, vol. 20, no. 9, pp. 1109–1114, 2005.
- [48] 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.
- [49] P. Kumar *et al.*, "Multimodal gait recognition with inertial sensor data and video using evolutionary algorithm," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 5, pp. 956–965, May 2019.



**Aite Zhao** received the bachelor's degree in software engineering from the Qingdao University of Technology, Qingdao, China, in 2013. She is currently working toward the Ph.D. degree with the College of Information Science and Engineering, Ocean University of China, Qingdao, China. She is currently a Visiting Ph.D. Researcher with the Department of Informatics, University of Leicester, Leicester, U.K. Her research interests include computer vision, pattern recognition, machine learning, data analysis, and robotics.



**Junyu Dong** received the B.Sc. and M.Sc. degrees from the Department of Applied Mathematics, Ocean University of China, Qingdao, China, in 1993 and 1999 respectively, and the Ph.D. degree from Heriot-Watt University, Edinburgh, U.K., in 2003. He is currently a Professor and the Vice Dean with the College of Information Science and Engineering, Ocean University of China. He has authored or coauthored more than 100 journal and conference papers. His research interests include computer vision, underwater image processing, and machine learning, with more than 10 research projects supported by NSFC, MOST, and other funding agencies.



**Jianbo Li** received the bachelor's and master's degrees from Computer Science Department, Qingdao University, Qingdao, China, in 2002 and 2005, respectively, and the Ph.D. degree from Computer Science Department, University of Science and Technology of China, Hefei, China, in 2009. He is currently a Professor and the Dean with the College of Computer Science and Technology, Qingdao University. His research interests include urban computing, mobile social networks, and machine learning.



**Lin Qi** received the B.Sc. and M.Sc. degrees from Ocean University of China, Qingdao, China, in 2005 and 2008 respectively, and the Ph.D. degree in computer science from Heriot-Watt University, Edinburgh, U.K., in 2012. He is currently an Associate Professor with the Department of Computer Science and Technology, Ocean University of China. His research interests include computer vision and visual perception.



**Huiyu Zhou** received the B.E. degree in radio technology from Huazhong University of Science and Technology, Wuhan, China and the M.Sc. degree in biomedical engineering from University of Dundee, Dundee, U.K. He is currently a Professor and heads the Applied Algorithms and AI (AAAI) Theme with University of Leicester, Leicester, U.K. From 2012 to 2017, he was a Lecturer with the School of Electronics, Electrical Engineering and Computer Science, Queen's University, Belfast, U.K. He has published widely in the field. He was the recipient of the Doctor of Philosophy Degree in Computer Vision from Heriot-Watt University, Edinburgh, U.K., the CVIU 2012 Most Cited Paper Award, the MIUA 2020 Best Paper Award, the ICPRAM 2016 Best Paper Award in the Area of Applications, and was shortlisted for the ICPRAM 2017 Best Student Paper Award and the MBEC 2006 Nightingale Prize.