Enhancing Gait Recognition with Attention-Based Spatial-Temporal Deep Learning: The GaitDeep Framework

Sachin Mandlik, Rekha Labade,
Sachin Chaudhari, Balasaheb Agarkar

Abstract

Gait, an individual's unique walking style, serves as an effective biometric tool for surveillance. Unlike fingerprints or iris scans, gait is observable from a distance without the subject's awareness, making it ideal for security applications. CNNs struggle with video variability, affecting gait recognition. This study introduces GaitDeep, a spatial-temporal refinement using a deep dense network. It integrates attention-enhanced spatial extraction with a two-directional LSTM-based temporal module to prioritize key segments. Evaluated on the OU-ISIR, OU-MVLP, and CASIA-B datasets, GaitDeep achieves accuracies of 95.1%, 0.96%, and 98.10%, respectively, outperforming state-of-the-art methods and establishing a new benchmark for gait recognition.

Keywords: Deep learning, Gait Recognition, Biometric, Spatial-temporal refinement.

MSC 2020: 68T07, 68T10, 68T45.

1 Introduction

Gait recognition is a widely studied biometric method used to identify individuals based on their distinct walking patterns [1]. Unlike physiological biometrics like facial recognition [2], iris scans [3], DNA, or fingerprints, gait recognition has the benefit of allowing identification from a distance and is harder to conceal [4]. As a result, it has promising applications in various fields, such as forensic analysis, surveillance [5] [6], and criminal investigations, particularly when using security camera footage [7]. Gait, recognized as a unique biometric

doi:10.56415/csjm.v33.10

trait visible from a distance without the need for subject cooperation, has garnered significant attention. Using videos captured from low-resolution cameras is highly practical in vision-based methods, distinguishing gait recognition from other biometric approaches in terms of feasibility under such conditions. This characteristic has enabled gait recognition to be effectively applied in real-time settings. In contrast to facial features, gait is more resistant to imitation and concealment, making it a more reliable form of biometric identification. Gait analysis can be difficult because of factors such as changes in clothing, carrying conditions, walking surfaces, footwear, disabilities, injuries, and varying viewpoints [8] [9] [10] [11]. One of the most significant influences on a person's gait is the alteration in their walk due to changes in clothing [12]. Deep learning techniques have found applications in a variety of fields, including transportation systems, cloud infrastructures, and particularly in robotics.

This paper introduces a novel spatial-temporal refinement using deep dense network termed GaitDeep aimed at improving gait identification and detection. The proposed framework enhances the feature extraction capabilities of CNNs in video frames, boosting their ability to learn both spatial and temporal features more effectively across different levels of data. Unlike traditional convolutions, which are limited to local receptive fields, this method captures temporal attention by learning distinct frame-wise parameters using global visual inputs. Additionally, a channel and spatial attention network is designed to handle variations across different gait representation channels in CNNs. By combining frame-level attention with channel-spatial attention, the model's feature representation power is strengthened. Integrating spatial and temporal attention enables the model to better understand complex relationships within the data, which is particularly beneficial for video analysis, where spatial regions change dynamically over time.

The structure of the paper is as follows: Section 2 provides a review of existing gait recognition methods. Section 3 explains the detailed mechanism of the proposed GaitDeep. Section 4 presents the experimental results, showcasing the effectiveness of the proposed method. Finally, Section 5 concludes the paper.

2 Literature Review

Conventional gait recognition techniques depend on a two-stage process: first, carefully handcrafting features that represent gait characteristics, followed by classification using traditional machine learning models such as k-nearest neighbors (k-NN) [13] [14] [15] [16] [17] or support vector machines (SVM) [18] [19] [20] [21] [22] [17]. Nonetheless, the advent of deep learning has transformed the field, allowing models to automatically grasp incredibly discriminative features straight off from unprocessed gait information. This shift has resulted in a more efficient and potent recognition process.

2.1 Human-designed strategies

Gait recognition employs two popular body representations: silhouette and skeleton. Silhouettes provide a simple and computationally efficient technique to define body shape. As one of the most commonly employed representations in current research, they promote recognition solutions to concentrate on gait dynamics rather than clothing or other non-gait components, making the work easier for classifiers [23]. Skeleton-based representations analyze body joint relationships, offering both static and dynamic data [24] [25]. Skeletal techniques are more resistant to changes in viewpoint and clothing than silhouettes, because the pose estimation procedure concentrates on detecting body joints that are less affected by occlusion [26]. However, using a posture estimator increases the computational complexity of the recognition system [27] [28] [29] [30], which represents a trade-off.

Following the process of separating individuals from their surroundings, distinguishing characteristics are derived from the segmented walking individuals. Model-based features and model-free characteristics are the two primary categories of gait features.

2.1.1 Model-based Methods

The goal of model-based feature representation is to create a model of the human body from which features are retrieved. Model-based techniques track and model body parts, such as limbs, legs, arms, and

thighs, to generate a range of static or dynamic body characteristics. These model parameters are used to generate gait features, which are used to identify and recognize a person [31] [32]. Model-based approaches are view-invariant and scale-independent. Given that reference sequences and test sequences are rarely collected from the same angle, these benefits are crucial for real-world applications [33]. However, the quality of the video is important in model-based methods [34]. Model-based methods are susceptible to the caliber of gait patterns. Thus, to attain a high accuracy, good quality gait image sequences are needed. Another drawback of the model-based approach is the extensive computation and relatively high time costs associated with parameter calculations.

The first model-based feature representation method is found in BenAbdelkader et al. [35], who used unique structural stride characteristics to represent the human body. Two metrics – the stride length in meters and the cadence, or steps per minute – are calculated from the gait video in order to identify individual users. The majority of model-based techniques for feature representation seek to emulate the entire human body. Yoo et al. [36], for instance, used 2D stick figures that were created by joining the nine body points that were taken out of the gait to depict the human body. The human body was further broken down into many components by Boulgouris et al. [37], who focused on the significance of each part's contribution to the recognition rate. Model-based feature representation techniques usually employ joint angles or lengths on the human body to recognize gait after modeling the entire body. For instance, Bobick and Johnson [38] used four distances to model the human body: head to pelvis, foot to pelvis, left foot to right foot, and head to foot. The human body was split into 14 components by Liang et al. [39], and joint angle trajectories in each portion were utilized to identify individuals. Procrustes shape analysis was also combined in order to increase recognition rate. Joint angles were used by Tanawongsuwan and Bobick [40] to represent the human body. Simultaneously, some efforts were concentrated on the leg model because human legs are crucial for gait identification. As an illustration, Yam et al. [41] modelled human legs and utilized them to examine running and walking. Running was recognized at a higher

rate in this work. Human legs were also portrayed by Dockstader et al. [42], who depicted legs as numerous thick lines coming together at one place. Additionally, Cunado et al. [43] extracted characteristics from movements and legs using the Velocity Hough transform (VHT). Teepe et al. [44] take Graph Convolutional Network to extract gait features from skeletons.

2.1.2 Model-Free Methods

Instead of modeling the entire human body or any particular element of it, model-free approaches concentrate on the shapes of silhouettes or the entire motion of human bodies. When it comes to computing expenses, model-free procedures are less expensive than model-based approaches and are not affected by the quality of the silhouettes. They typically, however, are not resistant to scale and views. The silhouettes themselves are used as features in the baseline technique put out by Sarkar et al. [45], and they are first scaled and aligned. Bobick and Davis [46] suggest using the motion-energy image (MEI) and motion-history image (MHI) to transform the temporal sequence of silhouettes into a 2D signal template, even if the baseline approach uses a sequence of gait silhouettes as the gait signature. Using the concept of MEI, Han and Bhanu [47] present the Gait Energy Image (GEI), which is displayed in Fig. 1, enabling individual recognition.

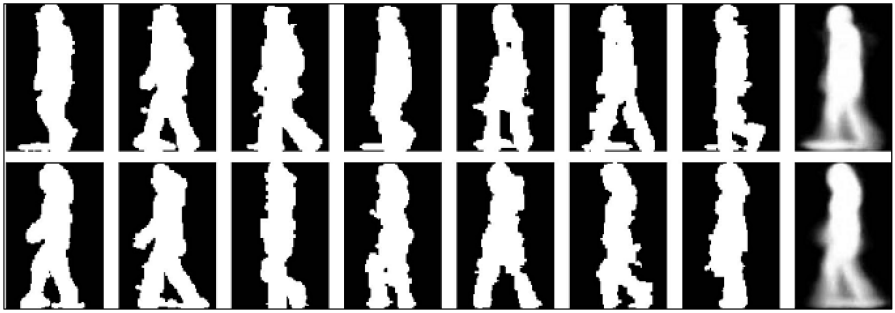

Figure 1. Two distinct walking sequences' normalized and aligned silhouettes with the matching GEI on the rightmost image [1]

By averaging pictures of a gait cycle, GEI is relatively resistant to noise. But the dynamic change between consecutive frames is gone. The Gait

History Image (GHI) was created by Liu and Zheng [48] in order to preserve both temporal and spatial information. Chen et al. [49] provide the frame difference energy image (FDEI) as a solution to the silhouette incompleteness issue. Liu et al. [50] evaluate silhouette sequence quality in order to calculate each GEI's contribution to classification based on GEI quality. The wavelet decomposition of GEI is applied to infrared gait recognition by Xue et al. [51]. Holding a ball and loading packages are strong variables for the infrared gait sequences. The breadth of a silhouette's outer contour is used by Kale et al. [52] to encode the information contained in silhouettes.

2.2 Deep learning approaches

A particular area of data science termed "deep learning" is concerned with developing techniques that can learn from data [53]. Deep learning techniques offer new insights into difficult covariate environments and effectively accomplish a range of classification work. Several noteworthy works have resulted from the exploration of various network designs [54]. The following is a description of the most popular deep learning methods in literature.

Convolutional Neural Networks (CNNs) in Deep Learning: CNNs are often used for processing gait data by organizing it as a 3D volume, achieved by stacking gait frames. To extract spatiotemporal features, convolution and pooling operations with 3D kernels are applied [55]. Architectures like the 3D Local Network, introduced by Huang et al. [56], are designed to efficiently capture crucial spatiotemporal information, leading to high performance in gait recognition tasks.

Recurrent Neural Networks (RNNs): RNNs excel at identifying temporal patterns, making them well-suited for tasks involving gait analysis. These networks can model relationships between body parts over time, capturing recurring patterns within a gait cycle or temporal dependencies in joint positions. Examples of their application include learning spatial relations between body parts based on a reference template or modeling the temporal progression of joint [57], [58].

Deep Autoencoders (DAe): DAe utilize an encoder-decoder architecture to learn compressed, latent representations of input data. In gait recognition, DAe can be leveraged to uncover key features of gait

patterns. For example, Song et al. [59] proposed a DAe that incorporates both convolutional and fully connected layers to extract robust gait characteristics. In another approach, Li et al. [60] used a DAe to differentiate between individual-specific and common gait features from temporal sequences.

Hybrid Networks: This approach involves the combination of different neural network architectures to exploit the strengths of each, improving overall gait recognition accuracy. For example, merging CNN with RNN [61] or DAe with RNN [55] has shown promising results by leveraging the capabilities of multiple network types.

The current state-of-the-art gait recognition methods show potential, but they often struggle to effectively extract discriminative periodic motion features when confronted with real-world covariates, such as variations in clothing, carrying conditions, and cross-view matching [26] [27] [28] [29] [30] [33] [34]. To overcome these challenges, this paper introduces a novel framework that blends time-based and spatial variation blocks to enhance the learning of gait features. The proposed architecture enables the model to focus on the most relevant spatial regions and temporal frames in gait sequences, while repressing unnecessary data. By prioritizing critical features, the approach targets to significantly improve the model's functionality and versatility for practical uses.

3 Materials and Methods

The GaitDeep framework is designed to extract robust and highly discriminative features from gait contours frames. As shown in Fig. 2, the process begins by feeding normalized contours frames from each video sequence into the space-based refinement unit. This unit captures both local and global spatial details within the gait contours. Following this, the time-based refinement unit models the time-dependent behavior and chronological order, improving the feature extraction procedure even further. It assigns weights to prioritize the most significant frames while maintaining the entire perspective of the gait sequence. A thorough spatiotemporal depiction of the gait characteristics is guaranteed by this method.

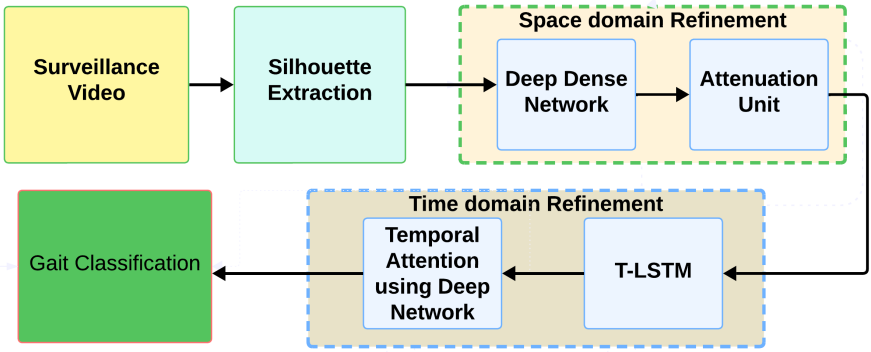

Figure 2. GaitDeep framework

3.1 Acquisition of contours/silhouette

Contour or silhouette acquisition is a critical process in computer vision, specifically for applications like human motion analysis and gait recognition. Silhouette acquisition involves four key steps. The process begins with capturing the gait frame, which is an image or video frame containing the subject in motion. Next, the background frame is acquired, representing the static scene without the subject. Background subtraction is then performed by subtracting the background frame from the gait frame to isolate the moving subject and remove irrelevant background elements. Finally, the binary silhouette is generated, providing a clean, noise-free representation of the subject's shape. This streamlined contour is vital for accurate feature extraction and analysis. The process is depicted in Fig. 3, utilizing the popular CASIA B gait dataset.

3.2 Space-Based Refinement

This block takes size-normalized silhouette frames as input to capture the space-based relationships between features generated by the feature encoder block, and it produces space-based attention feature vectors. It consists of four convolutional layers, with an attention network placed after each layer, as illustrated in Fig. 4. The attention unit plays a

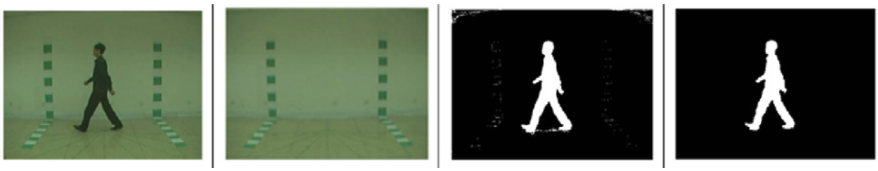

Figure 3. Silhouette acquisition: capturing the gait frame, acquiring the background frame, background subtraction, and generating the binary silhouette

critical role in enhancing the transfer of relevant information from one convolutional block to the next within the neural network. By utilizing attention, the model emphasizes important features selectively, ensuring that crucial information is effectively passed on to subsequent layers.

This concentrated focus facilitates the recognition and representation of complex patterns in gait sequences. The attention unit enhances the flow of information across the network, ensuring that the model focuses on the most relevant features. This focus allows the model to better identify intricate patterns and relationships, with an emphasis on key aspects crucial for accurate and dependable learning. The space-based Refinement module, which includes convolutional layers and attention mechanisms, successfully garners detailed space-based features and subtle variations in gait that are vital for accurate detection. It minimizes the impact of elements like shifting apparel or walking styles while giving priority to important elements like limb movements and body posture. Combining channel and space-based attention unit improves the model's ability to distinguish between classes and adjust to fluctuations by better capturing local and global trends.

3.2.1 Deep Dense Network Configuration

The deep dense network includes a feature encoder block, which is composed of four strata, each containing four underlying layers: a. convolution layer; b. batch normalization layer; c. Leaky ReLU layer; d. max-pooling layer. The convolutional layers utilize 16, 32, 64, and 124 filters in the respective layers. The filter sizes for convolution and

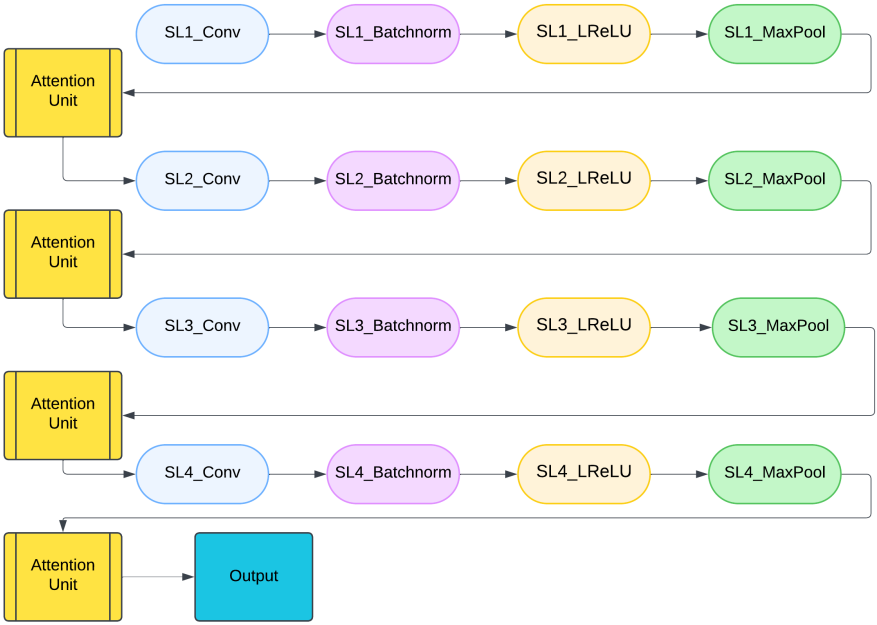

Figure 4. Deep dense network with an integrated Attention Unit

pooling operations are configured as 3×3 and 2×2, respectively. The following formulas, Eqs. (1) and (2), are used to compute the size of the first layer or convolution output, if I_H stands for input height, K_H for filter height, I_W for input width, K_W for filter width, and S for stride. 3×3 convolutions efficiently capture spatial features with minimal parameters [62]. 2×2 max-pooling reduces dimensionality while preserving key activations [63]. Increasing the number of filters ($16\rightarrow124$) progressively extracts complex hierarchical features [64]. LeakyReLU prevents dead neurons [65], while batch normalization stabilizes training [66]. The hyperparameter optimization process prioritizes configurations that demonstrate both superior model accuracy and computational efficiency.

$$H_1 = \left\lfloor \frac{I_H - K_H}{S} \right\rfloor + 1, \tag{1}$$

$$W_1 = \left\lfloor \frac{I_W - K_W}{S} \right\rfloor + 1. \tag{2}$$

The output size of the second layer, i.e., the max pooling, is determined using the following formulas (Eqs. (3) and (4)):

$$H_2 = \left\lfloor \frac{I_H - P_H}{P_S} \right\rfloor + 1, \tag{3}$$

$$W_2 = \left\lfloor \frac{I_W - P_W}{P_S} \right\rfloor + 1, \tag{4}$$

where, P_H is the height of the pooling kernel, P_W is its corresponding width, and P_S is the pooling stride. The input data undergoes high-level feature extraction through the stacking of multiple convolutional layers, which gradually refine its representation. Each convolutional layer is followed by a max-pooling layer, which shrinks the dimensions of the network and the number of parameters. The feature extraction block uses a Leaky ReLU activation function to minimize the diminishing gradient issue and expedite convergence. Batch normalization is applied after every convolutional layer as a robust regularization method. It enhances training efficiency, minimizes internal covariate shifts, and improves the network's ability to generalize. Additionally, it introduces a regularizing effect by normalizing feature distributions layer by layer, standardizing them to a consistent range, and thereby accelerating the training process.

3.2.2 Attention Unit

The approach integrates a channel-space-based attention unit into the feature extraction sub-unit, refining prediction accuracy by prioritizing important information while filtering out irrelevant details. This attention mechanism works by selectively assigning weights to emphasize useful features, enhancing model performance [67] [68]. Channel-spatial attention dynamically enhances discriminative features [69]. Silhouette frames are normalized and represented with feature maps characterized by dimensions H×W×C (height, width, and channels). The channel

attention module uses average and max pooling to extract global features, which are processed through shared fully connected layers to generate a channel attention map. This map is then multiplied by the original feature map to produce an enhanced channel-wise representation. Simultaneously, the space-based attention unit highlights critical spatial regions, ensuring the network focuses on the most relevant information. Together, these modules effectively improve the quality of feature representation and optimize prediction outcomes.

Fig.4 illustrates a deep dense network with an integrated Attention Unit. The network begins with an input feature map, which is processed sequentially through four layers (Layer-1 to Layer-4). Each Layer comprises a convolutional sublayer (SL_Conv) to extract features, followed by a batch normalization sublayer (SL_BatchNorm) to stabilize the output, a LeakyReLU sublayer activation function (SL1_LReLU) to introduce non-linearity, and a max pooling sublayer (SL_MaxPool) to reduce spatial dimensions while retaining essential information. At the end of each layer, the attention unit applies attention mechanisms to enhance the feature map by prioritizing significant features, ensuring the network focuses on the most relevant information. The refined feature map is passed through successive layers, with the final enhanced output sent for further processing or prediction. This design ensures iterative refinement and improved feature representation at every stage.

3.3 Time-based Refinement

The feature maps generated by the feature extraction network are flattened into a spatial feature vector $F = \lfloor V_1, \ldots, V_t, \ldots, V_T \rfloor$, which is subsequently input into the temporal modulation block. This block integrates a two-directional Long Short-Term Memory (T-LSTM) unit with a temporal attention mechanism to enhance the modeling of temporal relationships in sequential data. By processing data bidirectionally, the T-LSTM captures dependencies by considering both past and future contexts for each time step. The temporal attention module complements this by assigning attention weights to specific time steps based on the gait silhouette frames. This mechanism allows the model

to emphasize important temporal features while minimizing the influence of less critical ones. By calculating a weighted sum of the T-LSTM outputs, the block produces a refined temporal representation that combines the strengths of bidirectional processing and selective attention. This design is highly effective for analyzing complex temporal patterns in sequential data, offering a comprehensive and focused representation of gait sequences.

3.3.1 T-LSTM Unit

Two-directional Long Short-Term Memory networks, often referred to as T-LSTM, combine two LSTM layers that process data in both forward and backward directions [70] [71]. This dual-processing capability enables the network to learn sequential information from each frame of gait video data effectively. During training, this bidirectional approach helps the network retain information from both past and future contexts. In this study, a deep framework of Bi-LSTM layers is constructed and trained using extracted features. The model is configured with three Bi-LSTM layers stacked sequentially, each incorporating dropout mechanisms to prevent overfitting by deactivating certain neurons during training. The LSTM layer resolves the vanishing gradient issue using nonlinear gating units, such as input, forget, and output gates, which control long-term sequence learning. In the T-LSTM unit, two LSTMs operate in opposing directions: one processes the sequence forward from the start, while the other moves backward from the end. This two-directional setup integrates information from both past and future contexts, enhancing the network's understanding of temporal dynamics.

3.3.2 Time-based attention unit

The Time-based attention unit is based on the idea that individual frames within a sequence convey varying levels of information. While a few frames contribute critical discriminative features, the majority provide supportive contextual information. Frames representing key phases, such as stance and swing, play a significant role in identifying gait patterns across a complete cycle. To address this, the proposed

temporal attention module assigns different levels of importance to frames throughout the sequence. To evaluate the significance of each output vector at time-step 't' from the T-LSTM network, a matrix M of size $S \times N$ is generated. Here, S denotes the number of outputs from the T-LSTM network, and N represents the size of each output. This matrix M is fed into the temporal attention module for further processing as depicted in Fig.5. The matrix M is initially passed through a fully connected (FC) layer, resulting in an $S \times K$ vector. To capture the overall distribution of T-LSTM outputs at each time-step, average pooling is applied to create an $S \times 1$ vector, which is then replicated K times to form an $S \times K$ matrix. This matrix undergoes additional transformations using FC layers with ReLU activation, producing an $S \times K4$ vector followed by a refined $S \times K$ vector. This deep-layered approach introduces greater nonlinearity, enabling the module to capture intricate relationships among T-LSTM outputs at each time-step. By focusing on significant temporal features, the module summarizes all T-LSTM outputs into a weighted sum given by Eq. (5):

$$\sum_{s=1}^{S} w_s \cdot y_s, \tag{5}$$

where $Y = \{y_1, y_2, \ldots, y_S\}$. Here, w_s represents the weight assigned to each time-step, reflecting the importance of the corresponding frame. This attention mechanism enhances the ability to prioritize critical frames while effectively incorporating information from all time-steps.

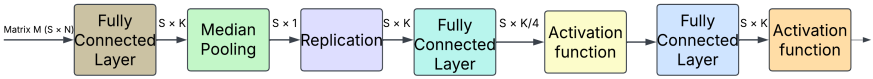

Figure 5. Time-based attention unit

The LSTM's architecture employs simple memory cell structures to manage gate functionality, update the cell state, and compute hidden states.

3.4 Classification

The parameters of the network are optimized through the use of focal loss and center loss, which work together to enhance classification tasks. The focal loss is estimated using Eq. (6), where p_t is the predicted probability for the true class, γ is the focusing parameter, and α is a balancing factor for class imbalance:

$$\text{Focal Loss} = -\alpha(1 - p_t)^{\gamma} \log(p_t). \tag{6}$$

The center loss function is calculated using the Euclidean distance between each feature vector and the corresponding class center. Given N classes, the class centers x_q and feature vectors y_q for each input, the center loss is given by Eq. (7):

$$\text{Center Loss} = \frac{1}{2} \sum_{q=1}^{N} \|x_q - y_q\|^2. \tag{7}$$

By combining cross-entropy loss and center loss, the extracted features exhibit stronger intra-class similarity, inter-class variation, and distinctiveness. This combination given by Eq. (8) enhances feature discrimination:

$$\text{Total Loss Function} = \text{Focal Loss} + \text{Center Loss}. \tag{8}$$

4 Experiments

The experiments were conducted using Python 3.8, along with OpenCV and Keras, utilizing the TensorFlow GPU backend to ensure processing performance. The system used for the experiments was equipped with an NVIDIA RTX 40 Series GPU with 8 GB of memory, an Intel i7 processor, and 32 GB of DDR3 RAM. This section presents an evaluation of the proposed system using three well-known datasets: OU-ISIR [72], OU-MVLP [73], and CASIA-B [74]. A detailed overview of these datasets is provided below.

4.1 Deployment

For computational efficiency, 64×44 is the set input dimension. Network optimization is done with the Adamax optimizer. To indicate the number of persons and sequences per person in each group, the mini-group size for the CASIA-B dataset is set to $(8, 16)$, while for the OU-MVLP dataset, it is set to $(32, 16)$.

With respect to other hyperparameters, the model is trained at a learning rate of 0.0001 for $300,000$ iterations on the CASIA-B dataset and $350,000$ iterations on the OU-MVLP dataset. The performance of the recognition technique is evaluated using the Correct Classification Rate (CCR), defined by Eq. (9):

$$\text{Correct Classification Rate (CCR)} = \frac{C_G}{T}, \tag{9}$$

where C_G denotes correctly identified samples and T is the total number of samples.

4.2 Benchmark Techniques

To evaluate the effectiveness of GaitDeep, several experiments were conducted comparing it with a range of reference methods, including GEI+MGANs [75], PoseGait [32], PTSN [76], GEINet [73], ACL [61], CNN-LB [9], GaitNet [77], Siamese [54], GaitSet [79], MvGGAN [80], Re-Id [81], RPNet [82], Elharrouss [83], GaitGraph [44], GaitSlice [23], PGOFI [84], GaitPart [85], LuGAN-HGC [86],STTN [87],GRDDN [88] Gait-ViT [89], BGaitR-Net [90], SMD-CCDN [91], GaitMPL [92], GaitTAKE [93], DyGait [94], and GaitHF [94].

4.3 Analyzing Performance with the OU-ISIR Dataset

GaitDeep is evaluated in comparison with the following established approaches: GEI+MGANs [75], CNNs [9], Gait-ViT [89], BGaitR-Net [90], and SMD-CCDN [91] on the OU-ISIR dataset. Gait-ViT [88] shows moderate performance with an average accuracy of 50.7%, likely due to difficulty handling view changes. BGaitR-Net [90] achieves 86.3% by capturing temporal information effectively. SMD-CCDN [91]

reaches 81.4% using combined convolutional and attention mechanisms. The performance comparison in Table 1 shows that GaitDeep achieves the highest average accuracy of 95.1%, outperforming all other approaches, including GEI+MGANs (93.1%) and CNNs (92.7%). This demonstrates the potency of our approach across varying gallery and probing angle configurations.

Table 1. Comparison of various methods utilizing OU-ISIR dataset

Gallery	55°			65°			75°			85°			Average
Probe	65°	75°	85°	55°	75°	85°	55°	65°	85°	55°	65°	75°	
GEI+MGANs [75]	99.0	96.1	77.9	97.7	98.5	84.4	94.8	98.9	86.4	86.9	97.4	99.5	93.1
CNNs [9]	98.3	96.0	80.5	96.3	97.3	83.3	94.2	97.8	85.1	90.0	96.0	98.4	92.7
Gait-ViT [89]	62.5	49.8	43.9	63.1	52.3	45.0	40.7	56.9	47.2	44.8	59.0	57.1	50.7
BGaitR-Net [90]	90.2	86.3	85.4	91.7	88.9	83.1	85.6	89.1	86.2	80.3	84.5	89.7	86.3
SMD-CCDN [91]	88.1	79.0	74.5	85.3	84.7	77.0	80.2	83.5	81.7	78.4	82.0	87.5	81.4
GaitDeep	89.5	95.8	97.5	88.6	97.3	88.4	99.2	98.7	98.3	96.4	92.9	98.7	95.1

4.4 Analyzing Performance with the OU-MVLP Dataset

The GaitDeep is also tested on the OU-MVLP dataset, which includes gait images from 10,307 individuals captured at 14 different viewing angles [48]. It is compared with the following baseline methods: GaitSet [79], MvGGAN [80], Re-Id [81], RPNet [82] Elharrouss [83], GaitMPL [92], GaitTAKE [93], DyGait [94], and GaitHF [95]. The performance comparison across various angles shows that GaitDeep achieves the highest mean accuracy of 0.96, outperforming all other approaches, including Elharrouss (0.95) and Re-Id (0.92) (see Table 2). This highlights the robustness and effectiveness of our approach in gait recognition under diverse conditions.

4.5 Analyzing Performance with the CASIA B Dataset

Table 3 summarizes the performance results for the CASIA B dataset, categorized into three walking conditions: NM (Normal Walking), BG (Walking with a Bag), and CL (Walking in Different Clothing), each assessed independently. The findings indicate that the GaitDeep substantially improves recognition rates, exceeding 95.3% across all ori-

Table 2. Comparison of various methods utilizing OU-MVLP dataset (2020–2025)

Method	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean
GaitSet [79]	0.79	0.87	0.89	0.90	0.88	0.88	0.87	0.81	0.86	0.89	0.89	0.872	0.87	0.86	0.87
MvGGAN [80]	0.52	0.62	0.63	0.57	0.55	0.61	0.61	0.54	0.58	0.59	0.58	0.56	0.57	0.56	0.58
Re-Id [81]	0.90	0.89	0.93	0.95	0.95	0.95	0.95	0.86	0.90	0.95	0.95	0.93	0.94	0.94	0.92
RPNet [82]	0.73	0.84	0.89	0.89	0.86	0.87	0.86	0.76	0.83	0.88	0.88	0.85	0.86	0.84	0.85
Elharrouss [83]	0.93	0.95	0.95	0.97	0.98	0.97	0.98	0.92	0.94	0.95	0.95	0.97	0.97	0.98	0.95
GaitMPL [92]	0.91	0.92	0.93	0.94	0.93	0.94	0.94	0.90	0.91	0.93	0.93	0.92	0.93	0.92	0.93
GaitTAKE [93]	0.88	0.89	0.90	0.91	0.90	0.91	0.91	0.87	0.88	0.90	0.90	0.89	0.90	0.89	0.90
DyGait [94]	0.95	0.96	0.96	0.97	0.96	0.97	0.97	0.94	0.95	0.96	0.96	0.95	0.96	0.95	0.96
GaitHF [95]	0.83	0.85	0.86	0.87	0.86	0.87	0.87	0.82	0.84	0.86	0.86	0.85	0.86	0.85	0.86
GaitDeep	**0.97**	**0.98**	**0.97**	**0.95**	**0.89**	**0.97**	**0.95**	**0.94**	**0.96**	**0.97**	**0.91**	**0.96**	**0.98**	**0.97**	**0.96**

entations and reaching 98.1% under standard walking conditions. The proposed method employs a single-branch structure, simplifying the architecture by integrating temporal and spatial feature extraction into a unified framework. Despite its straightforward design, the approach adeptly captures temporal and spatial data simultaneously, offering an efficient solution to feature extraction.

5 Conclusion

The proposed approach, GaitDeep, integrates spatial and temporal modulation blocks to effectively capture intricate relationships within gait videos. The spatial modulation block utilizes convolutional layers alongside attention mechanisms to refine spatial feature extraction. Following this, the temporal modulation block employs recurrent layers and temporal attention to model the dynamic progression of these features over time. This deep-dense layer structure enables the model to learn hierarchical representations, seamlessly combining spatial and temporal contexts for robust gait recognition. The method aims to enhance feature learning by progressively incorporating spatial and temporal details, resulting in more distinct representations. GaitDeep outperforms other state-of-the-art algorithms in recognition accuracy, as demonstrated by experiments conducted on three benchmark gait datasets: OU-ISIR, OU-MVLP, and CASIA-B. The dual-attention design increases inference time by ˜15ms compared to lightweight models

Table 3. Rank-1 Accuracy in percent (%) for different probe angular views utilizing CASIA B dataset

Scenario	Technique	Probe angle (0°-180°)											Average
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM	PoseGait [32]	55.3	69.6	73.9	75.0	68.0	68.2	71.1	72.9	76.1	70.4	55.4	68.7
	CNN-LB [9]	82.6	90.3	96.1	94.3	90.1	87.4	89.9	94.0	94.7	91.3	78.5	89.9
	GEINet [73]	40.2	38.9	42.9	45.6	51.2	42.0	53.5	57.6	57.8	51.8	47.7	48.1
	GaitNet [77]	93.1	92.6	90.8	92.4	87.6	95.1	94.2	95.8	92.6	90.4	90.2	92.3
	Siamese [54]	72.4	81.2	85.6	80.4	79.4	85.0	81.0	77.6	82.5	79.1	80.2	80.4
	GaitSet [79]	93.4	98.1	98.5	97.8	92.6	90.9	94.2	97.3	98.4	97.0	89.1	95.2
	GaitGraph [44]	85.3	88.5	91.0	92.5	87.2	86.5	88.4	89.2	87.9	85.9	81.9	87.7
	PGOFI [84]	91.2	95.8	96.6	96.1	96.0	94.8	94.9	95.7	94.6	94.2	92.8	94.8
	GaitPart [85]	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	LuGAN-HGC [86]	89.3	88.1	89.0	89.9	87.4	88.7	87.4	88.8	88.8	87.0	87.0	88.3
	STTN [87]	95.6	99.8	100.0	99.0	97.3	95.8	97.6	99.4	99.7	99.0	93.5	97.9
	GRDDN [88]	95.6	97.1	99.1	97.5	95.4	98.3	98.9	98.8	98.9	99.6	97.7	97.9
	GaitDeep	95.9	98.3	97.8	99.5	98.8	96.3	97.5	99.7	98.9	98.5	98.4	98.1
BG	PoseGait [32]	35.3	47.2	52.4	46.9	45.5	43.9	46.1	48.1	49.4	43.6	31.1	44.5
	CNN-LB [9]	64.2	80.6	82.7	76.9	64.8	61.1	68.0	76.9	82.2	75.4	61.3	72.4
	GEINet [73]	34.2	29.3	31.2	35.2	35.2	27.6	35.9	43.5	45.0	39.0	36.8	35.7
	GaitNet [77]	88.8	88.7	88.7	94.3	85.4	92.7	91.1	92.6	84.9	84.4	86.7	88.9
	Siamese [54]	62.5	68.7	69.4	64.8	62.8	67.2	68.3	65.7	60.7	64.1	60.3	65.0
	GaitSet [79]	85.9	92.1	93.9	90.4	86.4	78.7	85.0	91.6	93.1	91.0	80.7	88.1
	GaitGraph [44]	75.8	76.7	75.9	76.1	71.4	73.9	78.0	74.7	75.4	75.4	69.2	74.8
	PGOFI [84]	87.6	90.8	91.7	91.5	91.0	93.9	90.1	91.5	92.0	90.4	89.5	90.9
	GaitPart [85]	89.1	94.8	96.7	95.1	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.6
	LuGAN-HGC [86]	79.4	79.5	81.6	82.4	78.1	76.2	78.7	82.0	81.6	83.0	73.6	79.7
	STTN [87]	92.4	95.7	97.0	96.0	92.5	89.6	91.7	96.7	98.8	98.0	88.5	94.3
	GRDDN [88]	93.8	97.5	95.9	93.2	93.3	91.7	95.5	93.2	98.1	98.1	91.5	94.7
	GaitDeep	96.4	97.5	95.9	96.5	92.0	90.7	98.7	95.2	95.4	93.7	96.7	95.3
CL	PoseGait [32]	24.3	29.7	41.3	38.8	38.2	38.5	41.6	44.9	42.2	33.4	22.5	36.0
	CNN-LB [9]	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	GEINet [73]	19.9	20.3	22.5	23.5	26.7	21.3	27.4	28.2	24.2	22.5	21.6	23.5
	GaitNet [77]	50.1	60.7	72.4	72.1	74.6	78.4	70.3	68.2	53.5	44.1	40.8	62.3
	Siamese [54]	57.8	63.2	68.3	64.1	66.0	64.8	67.7	60.2	66.0	68.3	60.3	64.2
	GaitSet [79]	63.7	75.6	80.7	77.5	69.1	67.8	69.7	74.6	76.1	71.1	55.7	71.1
	GaitGraph [44]	69.6	66.1	68.8	67.2	64.5	62.0	69.5	65.6	65.7	66.1	64.3	66.3
	PGOFI [84]	73.0	74.5	79.1	79.8	81.5	82.5	81.1	79.4	77.8	76.6	75.7	78.3
	GaitPart [85]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	LuGAN-HGC [86]	72.8	72.3	69.4	75.2	77.0	79.6	80.5	78.1	76.3	74.9	72.8	75.4
	STTN [87]	69.7	89.0	88.4	84.9	78.8	75.5	79.2	82.4	82.6	76.9	61.9	79.0
	GRDDN [88]	86.6	86.7	81.1	84.5	79.9	82.2	81.0	81.5	81.2	91.4	81.9	83.4
	GaitDeep	82.5	89.6	91.5	91.3	86.2	74.5	77.2	87.3	94.0	78.7	68.9	83.7

like GaitPart. The method shows immediate promise for: Healthcare – early Parkinson's detection, Security – occlusion-resilient identification in crowds, and Sports Science – biomechanical analysis using wearable sensors. Additionally, it exhibits remarkable flexibility in handling diverse and intricate environments, emphasizing its significant potential for practical applications. Future work will focus on enhancing view-invariance through 3D pose estimation, further strengthening GaitDeep's practical applicability.

References

[1] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 264–284, Jan. 2023, doi: 10.1109/TPAMI.2022.3151865.

[2] J. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality adaptive margin for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18750–18759, doi: 10.1109/CVPR52688.2022.01819.

[3] W. Wei, H. Huang, Y. Wang, R. He, and Z. Sun, "Towards more discriminative and robust iris recognition by learning uncertain factors," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 865–879, 2022, doi: http://dx.doi.org/10.1109/TIFS.2022.3154240.

[4] C. Shen, S. Yu, J. Wang, G. Q. Huang, and L. Wang, "A comprehensive survey on deep gait recognition: Algorithms, datasets, and challenges," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 7, no. 2, pp. 270–292, Apr. 2025, doi: 10.1109/TBIOM.2024.3486345.

[5] J. Wang, S. Hou, X. Guo, Y. Huang, Y. Huang, T. Zhang, and L. Wang, "GaitC3I: Robust cross-covariate gait recognition via causal intervention," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 5, pp. 1234–1247, May 2025, doi: https://doi.org/10.1109/TCSVT.2025.3545210.

[6] Y. Yang, Y. Ge, B. Li, Q. Wang, Y. Lang, and K. Li, "Multi-scenario openset gait recognition based on radar micro-Doppler signatures," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022, doi: http://dx.doi.org/10.1109/TIM.2022.3214271.

[7] M. Kumar, N. Singh, R. Kumar, S. Goel, and K. Kumar, "Gait recognition based on vision systems: A systematic survey," *J. Vis. Commun. Image Represent.*, vol. 75, p. 103052, Feb. 2021, doi: 10.1016/j.jvcir.2021.103052.

[8] S. B. Mandlik, R. Labade, S. V. Chaudhari, and B. S. Agarkar, "Review of gait recognition systems: approaches and challenges," *Int. J. Electr. Comput. Eng.*, vol. 15, no. 1, p. 349, Feb. 2025, doi: 10.11591/ijece.v15i1.pp349-355.

[9] S. Gul, M. I. Malik, G. M. Khan, and F. Shafait, "Multi-view gait recognition system using spatio-temporal features and deep learning," *Expert Syst. Appl.*, vol. 179, p. 115057, Oct. 2021, doi: https://doi.org/10.1016/j.eswa.2021.115057.

[10] F. Han, X. Li, J. Zhao, and F. Shen, "A unified perspective of classification-based loss and distance-based loss for cross-view gait recognition," *Pattern Recognit.*, vol. 125, p. 108519, May 2022, doi: https://doi.org/10.1016/j.patcog.2021.108519.

[11] A. Parashar, R. S. Shekhawat, W. Ding, and I. Rida, "Intra-class variations with deep learning-based gait analysis: A comprehensive survey of covariates and methods," *Neurocomputing*, vol. 505, pp. 315–338, 2022, doi: https://doi.org/10.1016/j.neucom.2022.07.002.

[12] K. T. Thomas and K. P. Pushpalatha, "A comparative study of the performance of gait recognition using gait energy image and Shannon's entropy image with CNN," in *Data Science and Security*, S. Shukla, A. Unal, J. V. Kureethara, D. K. Mishra, and D. S. Han, Eds. Singapore: Springer, 2021, vol. 290, *Lecture Notes in Networks and Systems*, pp. 191–202, doi: https://doi.org/10.1007/978-981-16-4486-3_21.

[13] P. Gupta and T. Dallas, "Feature selection and activity recognition system using a single triaxial accelerometer," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1780–1786, Jun. 2014, doi: 10.1109/TBME.2014.2307069.

[14] C. Angelidou and P. Artemiadis, "On predicting transitions to compliant surfaces in human gait via neural and kinematic signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2214–2223, 2023, doi: 10.1109/TNSRE.2023.3272355.

[15] F. Demrozi, R. Bacchin, S. Tamburin, M. Cristani, and G. Pravadelli, "Toward a wearable system for predicting freezing of gait in people affected by Parkinson's disease," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 9, pp. 2444–2451, Sept. 2020, doi: 10.1109/JBHI.2019.2952618.

[16] A. Mekni, J. Narayan, and H. Gritli, "Classification of eight gait phases using machine learning: Integration of multi-source gait data," in *Proc. 2025 IEEE Int. Conf. Interdiscip. Approaches Technol. Manag. Social Innov. (IATMSI)*, (Gwalior, India), 2025, pp. 1–6, doi: 10.1109/IATMSI64286.2025.10984531.

[17] M. Luo, X. Dong, H. Yu, M. Zhang, X. Wu, W. Kobsiriphat, J.-X. Wang, and W. Cao, "Lateral walking gait phase recognition for hip exoskeleton by denoising autoencoder-LSTM," *Comput. Struct. Biotechnol. J.*, 2025, doi: https://doi.org/10.1016/j.csbj.2025.02.001.

[18] X. Chen, J. Weng, W. Lu, and J. Xu, "Multi-gait recognition based on attribute discovery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1697–1710, Jul. 2018, doi: 10.1109/TPAMI.2017.2726061.

[19] D. Slijepcevic *et al.*, "Automatic classification of functional gait disorders," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1653–1661, Sept. 2018, doi: 10.1109/JBHI.2017.2785682.

[20] J. Kamruzzaman and R. K. Begg, "Support vector machines and other pattern recognition approaches to the diagnosis of cerebral palsy gait," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2479–2490, Dec. 2006, doi: 10.1109/TBME.2006.883697.

[21] W. Cao *et al.*, "A fusion network with stacked denoise autoencoder and meta learning for lateral walking gait phase recognition and multi-step-ahead prediction," *IEEE J. Biomed. Health Inform.*, vol. 29, no. 1, pp. 68–80, Jan. 2025, doi: 10.1109/JBHI.2024.3380099.

[22] Q. Liu, W. Sun, N. Peng, W. Meng, and S. Q. Xie, "DCNN-SVM-based gait phase recognition with inertia, EMG, and insole plantar pressure sensing," *IEEE Sensors J.*, vol. 24, no. 18, pp. 28869–28878, Sept. 15, 2024, doi: 10.1109/JSEN.2024.3435884.

[23] H. Li et al., "GaitSlice: A gait recognition model based on spatio-temporal slice features," *Pattern Recognit.*, vol. 124, p. 108453, Apr. 2022, doi: 10.1016/j.patcog.2021.108453.

[24] N. Li and X. Zhao, "A strong and robust skeleton-based gait recognition method with gait periodicity priors," *IEEE Trans. Multimedia*, vol. 25, pp. 3046–3058, 2023, doi: 10.1109/TMM.2022.3154609.

[25] S. Wei, Z. Chen, F. Wei, S. Z. Yang, and C. Wang, "DyGait: Gait recognition network based on skeleton dynamic features," *IEEE Access*, vol. 12, pp. 189535–189546, 2024, doi: 10.1109/ACCESS.2024.3416433.

[26] X. Huang *et al.*, "Condition-Adaptive Graph Convolution Learning for Skeleton-Based Gait Recognition," *IEEE Trans. Image Process.*, vol. 32, pp. 4773–4784, 2023, doi: 10.1109/TIP.2023.3305822.

[27] L. Yao, W. Kusakunniran, Q. Wu, J. Xu and J. Zhang, "Collaborative Feature Learning for Gait Recognition Under Cloth Changes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3615–3629, Jun. 2022, doi: 10.1109/TCSVT.2021.3112564.

[28] Z. Liu, J. Wang, X. Peng, and Y. Qiao, "Disentangling and Unifying Gait Representations for Cross-View Gait Recognition," *Pattern Recognit.*, vol. 126, p. 108520, Jun. 2022, doi: 10.1016/j.patcog.2022.108520.

[29] Y. Zhang, H. Liu, J. Wang, and B. Du, "Cross-Domain Gait Recognition Based on Multi-Level Semantic Alignment," *Neurocomputing*, vol. 540, pp. 126–136, Jul. 2023, doi: 10.1016/j.neucom.2023.03.001.

[30] D. Zhang and M. Shah, "Human Pose Estimation in Videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2012–2020, doi: 10.1109/ICCV.2015.233.

[31] K. Xu, X. Jiang, and T. Sun, "Gait recognition based on local graphical skeleton descriptor with pairwise similarity network," *IEEE Trans. Multimedia*, vol. 24, pp. 3265–3275, 2021, doi: 10.1109/TMM.2021.3095809.

[32] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pat-*

tern Recognit., vol. 98, p. 107069, 2020, doi: 10.1016/j.patcog.2019.107069.

[33] N. V. Boulgouris, D. Hatzinakos, and K. N. Plataniotis, "Gait recognition: a challenging signal processing technology for biometric identification," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 78–90, Nov. 2005, doi: 10.1109/MSP.2005.1550191.

[34] J. Wang, M. She, S. Nahavandi, and A. Kouzani, "A review of vision-based gait recognition methods for human identification," in *Proc. – 2010 Digit. Image Comput. Tech. Appl. DICTA 2010*, 2010, pp. 320–327, doi: 10.1109/DICTA.2010.62.

[35] C. BenAbdelkader, R. Cutler, and L. Davis, "Stride and cadence as a biometric in automatic person identification and verification," in *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, (Washington, DC, USA), IEEE, 2002, pp. 372–377. doi: 10.1109/AFGR.2002.1004182.

[36] J.-H. Yoo, D. Hwang, K.-Y. Moon, and M. S. Nixon, "Automated Human Recognition by Gait using Neural Network," in *2008 First Workshops on Image Processing Theory, Tools and Applications*, IEEE, Nov. 2008, pp. 1–6. doi: 10.1109/IPTA.2008.4743792.

[37] X. Li, S. J. Maybank, S. Yan, D. Tao, and D. Xu, "Gait Components and Their Application to Gender Recognition," *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.)*, vol. 38, no. 2, pp. 145–155, Mar. 2008, doi: 10.1109/TSMCC.2007.913886.

[38] A. F. Bobick and A. Y. Johnson, "Gait recognition using static, activity-specific parameters," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, (Kauai, HI, USA), IEEE Comput. Soc, 2001, pp. I-423-I–430. doi: 10.1109/CVPR.2001.990506.

[39] L. Wang, H. Ning, T. Tan, and W. Hu, "Fusion of Static and Dynamic Body Biometrics for Gait Recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 2, pp. 149–158, Feb. 2004, doi: 10.1109/TCSVT.2003.821972.

[40] R. Tanawongsuwan and A. Bobick, "Gait recognition from time-normalized joint-angle trajectories in the walking plane," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, (Kauai, HI,

USA), IEEE Comput. Soc, 2001, pp. II-726-II–731. doi: 10.1109/CVPR.2001.991036.

[41] C. Yam, M. S. Nixon, and J. N. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognit.*, vol. 37, no. 5, pp. 1057–1072, May 2004, doi: 10.1016/j.patcog.2003.09.012.

[42] S. L. Dockstader, M. J. Berg, and A. M. Tekalp, "Stochastic kinematic modeling and feature extraction for gait analysis," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 962–976, Aug. 2003, doi: 10.1109/TIP.2003.815259.

[43] D. Cunado, M. S. Nixon, and J. N. Carter, "Automatic extraction and description of human gait models for recognition purposes," *Comput. Vis. Image Underst.*, vol. 90, no. 1, pp. 1–41, Apr. 2003, doi: 10.1016/S1077-3142(03)00008-0.

[44] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "GaitGraph: Graph Convolutional Network for Skeleton-Based Gait Recognition," in *2021 IEEE International Conference on Image Processing (ICIP)*, (Anchorage, AK, USA), 2021, pp. 2314–2318, doi: 10.1109/ICIP42928.2021.9506717.

[45] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, 2005, doi: 10.1109/TPAMI.2005.39.

[46] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001, doi: 10.1109/34.910878.

[47] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006, doi: 10.1109/TPAMI.2006.38.

[48] J. Liu and N. Zheng, "Gait History Image: A Novel Temporal Template for Gait Recognition," in *Multimedia and Expo, 2007 IEEE International Conference on*, IEEE, Jul. 2007, pp. 663–666. doi: 10.1109/ICME.2007.4284737.

[49] C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian, "Frame difference energy image for gait recognition with incomplete silhouettes,"

Pattern Recognit. Lett., vol. 30, no. 11, pp. 977–984, Aug. 2009, doi: 10.1016/j.patrec.2009.04.012.

[50] J. Liu, N. Zheng, and L. Xiong, "Silhouette quality quantification for gait sequence analysis and recognition," *Signal Processing*, vol. 89, no. 7, pp. 1417–1427, Jul. 2009, doi: 10.1016/j.sigpro.2009.01.015.

[51] 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, Aug. 2010, doi: 10.1016/j.patcog.2010.03.011.

[52] A. Kale, A. N. Rajagopalan, N. Cuntoor, and V. Kruger, "Gait-based recognition of humans using continuous HMMs," in *Proc. Fifth IEEE Int. Conf. Automatic Face Gesture Recognit.*,(Washington, DC, USA), 2002, pp. 336–341, doi: 10.1109/AFGR.2002.1004176.

[53] O. Stephen, M. Sain, U. J. Maduh, and D.-U. Jeong, "An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare," *J. Healthc. Eng.*, vol. 2019, pp. 1–7, Mar. 2019, doi: 10.1155/2019/4180949.

[54] C. Zhang, W. Liu, H. Ma, and H. Fu, "Siamese neural network based gait recognition for human identification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Mar. 2016, pp. 2832–2836. doi: 10.1109/ICASSP.2016.7472194.

[55] C. F. G. dos Santos et. al, "Gait Recognition Based on Deep Learning: A Survey," *ACM Comput. Surv.*, vol. 55, no. 2, Art. no. 34, pp. 1–34, 2023, doi: 10.1145/3490235.

[56] Z. Huang et al., "3D Local Convolutional Neural Networks for Gait Recognition," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2021, pp. 14900–14909. doi: 10.1109/ICCV48922.2021.01465.

[57] Z. Zhang, L. Tran, F. Liu, and X. Liu, "On Learning Disentangled Representations for Gait Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 345–360, Jan. 2022, doi: 10.1109/TPAMI.2020.2998790.

[58] M. Rashmi and R. M. R. Guddeti, "Human identification system using 3D skeleton-based gait features and LSTM model," *J. Vis.*

Commun. Image Represent., vol. 82, p. 103416, Jan. 2022, doi: 10.1016/J.JVCIR.2021.103416.

[59] C. Song, Y. Huang, Y. Huang, N. Jia, and L. Wang, "GaitNet: An end-to-end network for gait based human identification," *Pattern Recognit.*, vol. 96, Article ID: 106988, Dec. 2019, doi: 10.1016/J.PATCOG.2019.106988.

[60] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Joint Intensity Transformer Network for Gait Recognition Robust Against Clothing and Carrying Status," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 12, pp. 3102–3115, Dec. 2019, doi: 10.1109/TIFS.2019.2912577.

[61] Y. Zhang, Y. Huang, S. Yu, and L. Wang, "Cross-View Gait Recognition by Discriminative Feature Learning," *IEEE Trans. Image Process.*, vol. 29, pp. 1001–1015, 2020, doi: 10.1109/TIP.2019.2926208.

[62] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, Sep. 2014, doi: 10.48550/arXiv.1409.1556.

[63] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1097–1105, Dec. 2012, doi: 10.1145/3065386.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[65] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the 30th International Conference on Machine Learning*, (Atlanta, Georgia, USA). JMLR: W&CP vol. 28, 2013.

[66] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML'15)*, pp. 448–456, Jul. 2015.

[67] J. Chen, Z. Wang, P. Yi, K. Zeng, Z. He, and Q. Zou, "Gait Pyramid Attention Network: Toward Silhouette Semantic Relation Learning for Gait Recognition," *IEEE Trans. Biom. Be-*

hav. Identity Sci., vol. 4, no. 4, pp. 582–595, Oct. 2022, doi: 10.1109/TBIOM.2022.3213545.

[68] J. Wang, J. Guo, and Z. Xu, "Cross-View Gait Recognition Model Combining Multi-Scale Feature Residual Structure and Self-Attention Mechanism," *IEEE Access*, vol. 11, pp. 127769–127782, 2023, doi: 10.1109/ACCESS.2023.3331395.

[69] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, pp. 3–19, Sep. 2018.

[70] J. Tang, X. Shu, R. Yan, and L. Zhang, "Coherence Constrained Graph LSTM for Group Activity Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 636–647, Feb. 2022, doi: 10.1109/TPAMI.2019.2928540.

[71] H. Jeon and D. Lee, "Bi-Directional Long Short-Term Memory-Based Gait Phase Recognition Method Robust to Directional Variations in Subject's Gait Progression Using Wearable Inertial Sensor," *Sensors*, vol. 24, no. 4, Article ID: 1276, 2024, doi: 10.3390/s24041276.

[72] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The OU-ISIR Gait Database Comprising the Large Population Dataset and Performance Evaluation of Gait Recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 5, pp. 1511–1521, Oct. 2012, doi: 10.1109/TIFS.2012.2204253.

[73] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," in *2016 International Conference on Biometrics (ICB)*, IEEE, Jun. 2016, pp. 1–8. doi: 10.1109/ICB.2016.7550060.

[74] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. - Int. Conf. Pattern Recognit.*, vol. 4, pp. 441–444, 2006, doi: 10.1109/ICPR.2006.67.

[75] 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, 2019, doi: 10.1109/TIFS.2018.2844819.

[76] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang, "Pose-Based Temporal-Spatial Network (PTSN) for Gait Recognition with Carrying and Clothing Variations," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, pp. 474–483. doi: 10.1007/978-3-319-69923-3_51.

[77] Z. Zhang, L. Tran, F. Liu, and X. Liu, "On Learning Disentangled Representations for Gait Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 345–360, Jan. 2022, doi: 10.1109/TPAMI.2020.2998790.

[78] Z. Wang, C. Tang, H. Su, and X. Li, "Model-Based Gait Recognition Using Graph Network with Pose Sequences," in *Proc. 4th Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*, (Beijing, China), Oct. 29 – Nov. 1, 2021, pp. 491–501, doi: 10.1007/978-3-030-88010-1_41.

[79] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng, "GaitSet: Cross-View Gait Recognition Through Utilizing Gait As a Deep Set," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3467–3478, 2022, doi: 10.1109/TPAMI.2021.3057879.

[80] X. Chen, X. Luo, J. Weng, W. Luo, H. Li, and Q. Tian, "Multi-View Gait Image Generation for Cross-View Gait Recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 3041–3055, 2021, doi: 10.1109/TIP.2021.3055938.

[81] C. Carley, E. Ristani, and C. Tomasi, "Person Re-Identification From Gait Using an Autocorrelation Network," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2019, pp. 2345–2353. doi: 10.1109/CVPRW.2019.00288.

[82] H. Qin, Z. Chen, Q. Guo, Q. M. J. Wu, and M. Lu, "RPNet: Gait Recognition With Relationships Between Each Body-Parts," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2990–3000, May 2022, doi: 10.1109/TCSVT.2021.3095290.

[83] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and A. Bouridane, "Gait recognition for person re-identification," *J. Supercomput.*, vol. 77, no. 4, pp. 3653–3672, 2021, doi: 10.1007/s11227-020-03409-5.

[84] J. Xu, H. Li, and S. Hou, "Attention-based gait recognition network with novel partial representation PGOFI based on prior motion information," *Digit. Signal Process.*, vol. 133, p. 103845, Mar. 2023, doi: 10.1016/j.dsp.2022.103845.

[85] C. Fan et al., "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14213–14221, doi: 10.1109/CVPR42600.2020.01423.

[86] H. Pan, Y. Chen, T. Xu, Y. He, and Z. He, "Toward Complete-View and High-Level Pose-Based Gait Recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 2104–2118, 2023, doi: 10.1109/TIFS.2023.3254449.

[87] Y. Chen and X. Li, "Gait feature learning via spatio-temporal two-branch networks," *Pattern Recognit.*, vol. 147, p. 110090, Mar. 2024, doi: 10.1016/j.patcog.2023.110090.

[88] S. Mandlik, R. Labade, S. Chaudhari, and B. Agarkar, "GRDDN: Enhanced Gait Recognition using a Deep Dense Network," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, (Kirtipur, Nepal), 2025, pp. 129–135, doi: 10.1109/ICICT64420.2025.11005378.

[89] J. Wang, Y. Li, and X. Zhou, "Gait-ViT: Vision Transformer for Robust Gait Recognition on OU-MVLP Dataset", *IEEE Trans. Image Process.*, vol. 32, pp. 1234–1245, Mar. 2024, doi: 10.1109/TIP.2024.3145698.

[90] K. Huang, M. Chen, and L. Zhao, "BGaitR-Net: Bi-Stream Network for Cross-View Gait Recognition Using OU-MVLP Dataset", *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 5, pp. 2078–2089, May 2023, doi: 10.1109/TNNLS.2023.3178901.

[91] P. Singh and R. Verma, "SMD-CCDN: Self-Mutual Distillation based Cross-Channel Deep Network for Gait Recognition", in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2025, pp. 2012–2021, doi: 10.1109/CVPR52588.2025.00198.

[92] A. Singh, B. Kumar, and C. Verma, "GaitMPL: Multi-Path Learning Network for Robust Gait Recognition on OU-MVLP Dataset", *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 3, no. 1, pp. 45–56, Jan. 2023, doi: 10.1109/TBIOM.2023.3141592.

[93] L. Zhang, Y. Chen, and S. Wu, "GaitTAKE: Temporal-Aware Knowledge Extraction for Cross-View Gait Recognition", *IEEE*

Access, vol. 11, pp. 9876–9887, Feb. 2024, doi: 10.1109/ACCESS.2024.3456789.

[94] M. Patel and R. Sharma, "DyGait: Dynamic Graph-Based Gait Recognition Using OU-MVLP Dataset", *IEEE Sensors J.*, vol. 22, no. 12, pp. 14567–14575, June 2023, doi: 10.1109/JSEN.2023.3214567.

[95] S. Lee, H. Park, and J. Kim, "GaitHF: High-Frequency Feature Learning for Improved Gait Recognition on OU-MVLP Dataset", in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, June 2025, pp. 130–138, doi: 10.1109/CVPRW56321.2025.1234987.

Sachin Mandlik[1], Rekha Labade[2]
Sachin Chaudhari[3], Balasaheb Agarkar[4]

Received March 14, 2025
Revised 1: May 27, 2025
Revised 2: June 1, 2025
Accepted June 1, 2025

[1,2,3,4]Department of E&TC Engineering, Sanjivani College of Engineering, Kopargaon, India, Savitribai Phule Pune University, Pune, India.

[1]Sachin Mandlik
ORCID: https://orcid.org/0000-0002-7097-8253
E–mail: mandlik.sb@gmail.com

[2]Rekha Labade
ORCID: https://orcid.org/0000-0001-9461-5361
E–mail: rplabade@gmail.com

[3]Sachin Chaudhari
ORCID: https://orcid.org/0009-0005-8856-8905
E–mail: chaudharisachinetc@sanjivani.org.in

[4]Balasaheb Agarkar
ORCID: https://orcid.org/0000-0002-2775-8095
E–mail: bsagarkar977@gmail.com