Received: 25 November 2025 | Revised: 12 March 2026 | Accepted: 20 May 2026 | Published online: 3 June 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–15
DOI: 10.47852/bonviewAIA62028402

BON VIEW PUBLISHING

A Multi-Part Attention-Guided Spatial-Temporal GCN Framework for Gait-Based Person Recognition

Md. Khaliluzzaman[1,2] and **Kaushik Deb**[1,*]

[1] *Department of Computer Science and Engineering, Chittagong University of Engineering & Technology, Bangladesh*

[2] *Department of Computer Science and Engineering, International Islamic University Chittagong, Bangladesh*

Abstract: Gait recognition has appeared as an important biometric modality because of its nonintrusive nature and straightforward implementation, enabling identification without physical contact. In contrast to systems that rely on silhouette information and other visual attributes, skeleton-based approaches retrieve gait data independently of appearance indicators. However, most approaches in the field utilize manually extracted features and adjacency matrices based on joint physical connectivity. This dependence poses a significant challenge for acquiring semantically rich representations of joint interactions and fundamental motion patterns, which are crucial for real-world gait understanding. To focus on these issues, this paper introduces a skeleton-based multi-part attention-guided spatial-temporal graph convolutional network (ST-GCN) gait recognition approach, MAST-GCN, which enhances the modeling of spatial and temporal dependencies in skeletal data through a multi-part attention mechanism. Compared with ST-GCNs, which rely on rigid graph structures and struggle to capture long-range interactions essential for identifying subtle gait differences, our method divides the skeleton into distinct anatomical regions and applies a part-wise attention module. By integrating attention-weighted features through a hierarchical fusion process, the model effectively captures both detailed and broad gait patterns across multiple temporal scales. The framework's effectiveness has been verified on benchmark datasets such as the CASIA-B and OUMVLP-Pose, achieving rank-1 precisions of 95.8%, 91.8%, and 88.5% under normal walking (NM), carrying bag (BG), and wearing coat (CL) conditions, respectively, on the CASIA-B dataset and 93.0% on the OUMVLP-Pose dataset, showing superior performance. Our approach outperforms state-of-the-art methods, particularly highlighting the benefits of part-based, attention-driven feature extraction for robust, precise gait recognition.

Keywords: gait recognition, Spatial-Temporal Graph Convolutional Networks (ST-GCNs), multi-part attention-guided, CASIA-B, OUMVLP-Pose

1. Introduction

Vision-based gait identification identifies individuals by examining their unique walking patterns. The modality has some advantages over other biometric methods, such as fingerprint or facial recognition, including that it is nonintrusive and hard to disguise and can operate at a distance without requiring user interaction [1]. It is estimated that the global market for gait biometrics will reach about 58.62 million USD by 2028, with a mixed global growth rate of 10.6% [2]. Therefore, gait detection is particularly appropriate for human–robot interaction, access control, and intelligent surveillance systems [3, 4].

The gait recognition methods currently used can be roughly divided into appearance- and model-based approaches. In appearance-based techniques, silhouettes significantly capture an individual's body shape and size. However, these methods are susceptible to performance degradation involving covariate factors, for example, clothing or carried items. On the contrary, model-based methods for gait representation utilize a priori knowledge about human postures and motions and are thus inherently more robust to such variations [5]. Recent approaches in RGB-based pose estimation have also improved the reliability of these methods as they provide accurate skeletal representations [6].

Skeleton graphs show walking techniques with nodes representing joints and edges representing spatial and temporal interactions. GCNs have been widely used for modeling unstructured graph data with estimation of the statistical influences of joint connections [7]. Since gait is a complicated motion, it becomes crucial to recognize the unique patterns in the gait of subjects. The existing methods utilize only static filters, which restrict the expressive capacity and temporal accuracy required for effective modeling. Additionally, previous studies utilize only anatomically connected joints for graph generation. The model exhibits limited discriminative capacity.

***Corresponding author:** Kaushik Deb, Department of Computer Science and Engineering, Chittagong University of Engineering & Technology, Bangladesh. Email: debkaushik99@cuet.ac.bd

The skeleton graphs represent walking postures, and graph convolutional networks (GCNs) have been widely employed to model unstructured graph data by capturing the statistical influence of connected nodes [7]. Because gait is a complex movement, it is important to identify each subject's gait pattern. Moreover, in existing work, only anatomically connected joints are used to generate the graph. Therefore, the model's discriminative ability is limited.

GCNs have proven particularly effective for modeling the structured nature of human skeletons in motion. When extended to the spatial-temporal domain through spatial-temporal graph convolutional networks (ST-GCNs), these models learn both spatial structures and temporal evolution of joint movements. However, conventional ST-GCNs rely on a predefined graph structure that considers the adjacency of physical joints, which restricts the ability of these methods to model the cross-joint interactions at a distance, as shown in Figure 1. This becomes troublesome when trying to mimic sequences of coordinated activities, such as the combined use of arms and legs during walking, which must be careful to discriminate among more modest variances in gait. Failing to capture these long-range relationships typically makes models less useful in complex real-world situations.

To address these issues, GCN-based models have recently integrated adaptive graph learning and attention mechanisms to enhance flexibility. Based on these advances, we propose a multi-part attention (MPA)-based graph framework designed explicitly for skeleton-based gait recognition. The framework proposes an innovative technique that synergistically combines graph convolutional network (GCN), temporal convolutional network (TCN), and MPA to accurately extract, emphasize, and fuse key gait characteristics. In addition, this complementary framework can enhance the model's adaptability concerning multiple challenging conditions while focusing on the important joint interactions, thereby enabling the model to capture both local and long-range dependencies essential for gait analysis, as shown in Figure 2. This approach allows for exploring gait sequences' spatial and temporal correlations. At the same time, the part-wise attention mechanism learns to concentrate on the considerable discriminatory gait attributes in different body part levels, which we denote as multi-part attention-guided ST-GCN (MAST-GCN). The main structure of our framework comprises six GCN, TCN, and MPA modules customized to optimize the computational performance concerning the recognition task. We limit the units to six to keep the model light and efficient for real-time applications.

The contributions of this paper are presented below:

1) Proposed an efficient gait recognition framework, MAST-GCN, to identify and rank the most significant body parts that are physically apart and can handle real-world challenges, such as clothing variations and viewpoint changes.
2) Developed an MPA module that addresses the limitations of vanilla ST-GCN and dynamically captures the most relevant features across different body parts to improve the robustness of the gait recognition model.
3) The proposed MAST-GCN framework outperforms state-of-the-art approaches on gait identification tasks, as demonstrated by extensive evaluations on the CASIA-B and OUMVLP-Pose datasets. In addition, a detailed ablation study on the CASIA-B dataset shows that the MPA module captures subtle, discriminative information needed for practical gait analysis, thereby improving framework performance.

2. Related Works

This section summarizes current vision-based gait recognition technologies. These techniques are appearance-based or model-based, depending on the input vision modality.

2.1. Appearance-based gait recognition

Appearance-based gait recognition methods depict the human body's silhouette in images obtained through background subtraction or deep learning-based segmentation methods without keeping color or texture information. The disadvantage of this classification is its pure shape, which reduces the Kolmogorov complexity of the dataset and increases the convergence and processing efficiency of the recognition models [8]. One of the most prominent contributions to this line of work is from Asif et al. [9], who proposed a Gait Energy Image (GEI), which is an average of a time-normalized sequence of silhouettes. GEI simplifies the computation as compared to the processing of raw sequences of silhouettes. However, this static representation ignores the temporal dynamics that are essential for describing the movement characteristics of gait.

Figure 1
The processing example of ST-GCN: apply the ST-GCN on the skeleton joints, which is often prone to rely on static spatial partitioning strategies

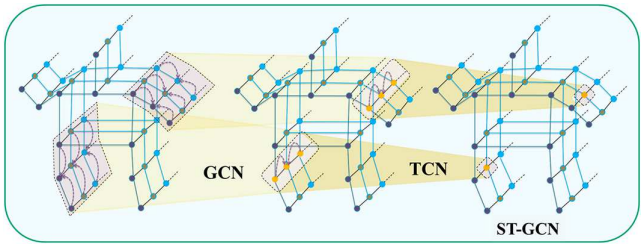
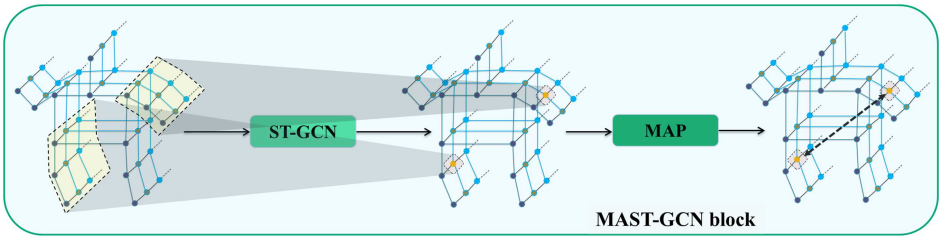


Figure 2
The processing example of MAST-GCN: apply the proposed multi-part attention process (MAP) on skeleton joints, which affects the correlation among the joint features that are physically apart



Many studies have attempted to use silhouette sequences as input to improve recognition performance beyond early innovations. The approaches can be roughly divided into set-based, part-based, 3D Convolutional Neural Network (CNN)-based, and disentanglement learning approaches. As an illustration, Zhao et al. [10] utilize unordered sequences of silhouettes and forward them through 2D CNNs to obtain both spatial and temporal features [3]. GaitGL [11] introduces two state networks to present the local and global information. Part-based methods, meanwhile, such as Chen's work [1] and micro-motion capture modules, are utilized to enhance localized motion characteristics [12], and they all learn differentiated features, for example, region-based features targeting different parts of the body. While recent part-based and set-based approaches take steps in that direction, they still tend to falter under similar variability, such as clothing changes or object carrying, leading to extreme silhouette changes.

To further analyze the effect of the covariates, researchers have proposed disentanglement methods that make use of encoder–decoder architectures to differentiate features of interest from others [13]. Additionally, Huang et al. [14] discuss the use of 3D CNNs. Jointly extracting spatial and temporal cues improves the discriminability of features but needs more computation. Appearance-based methods have achieved considerable success, but they still suffer from inaccurate silhouette segmentation in complex backgrounds and rely heavily on static body shape information.

2.2. Skeleton-based gait recognition

Skeleton-based gait recognition processes extract the joint angles and skeletal motions from gait sequences. The gait recognition approaches use the structural and dynamic features of the human body rather than its appearance, making them more resilient against variables like clothes and carrying conditions.

GCNs [15] illustrate a recent neural network structure that handles graph-based information to learn valuable spatial features. For the human skeleton investigation, spatial and temporal graph convolutions derive features spanning through spatial and temporal dimensions, where joints act as vertices while their connections act as edges. Early approaches, such as the one proposed by Monti et al. [16], used pose-based strategies with CNNs for feature extraction. Li et al. [13] first presented the GC process for gait recognition from the graph structure data.

Later refinements by Liao et al. [5, 17], Wei et al. [17], Shopon et al. [18], and Teepe et al. [7] combined 2D or 3D joint coordinates employing GCNs or CNNs but overlooked multi-part body features and disregarded joint, bone, and motion details. Lin et al. [19] introduced 3D-CNNs and Long Short-Term Memory (LSTMs) for feature extraction, sacrificing simpler architectures for improved performance, though with higher computational costs. Wang et al. [20] filled the gaps by combining motion, joint, and bone facts with graph structural features and increased the precision using the multi-order adjacency matrix. However, this method exhibited a bias toward nearer joints. Hasan et al. [21] suggested a hop-extraction method to balance attention across all joints, though it stayed suboptimal with regional joint bias. Zhang et al. [22] advanced the domain by merging spatial transformers (ST) with temporal convolutions (TC) for spatiotemporal feature extraction, leveraging multi-head self-attention tools. Zhu et al. [23] introduce GaitSkeleton, which improves skeleton-based gait recognition by explicitly modeling joint coordination associations. Ray et al. [24] present a multi-biometric feature extraction procedure for dragging the features from the different pose examination methods to improve gait recognition. Recent outcomes by Peng et al. [25], Li and Zhao [26], and Chen et al. [27] presented ST-GCN-based techniques, dragging spatial features via GCNs and temporal features via TCNs. Nonetheless, these techniques face limitations in static spatial partitioning, failing to account for dynamically disjointed joints and varying contextual relationships in human motion. To overcome the limitation of ST-GCN, Khaliluzzaman and Deb [28] proposed a fully connected skeleton operator with a spatial self-attention mechanism to adaptively capture the dependencies physically apart from joints. In addition, Khaliluzzaman et al. [29] proposed the attention-guided spatial-temporal GCN (AST-GCN) method, which incorporates frame and channel attention into an ST-GCN to overcome the limitations of the vanilla ST-GCN approach. In addition to the abovementioned research, several other substantial works [30–32]) also provide significant inspiration.

3. Method

The MAST-GCN method for gait recognition based on skeletal joint positions consists of several stages. Input per-frame gait video series data into a pose estimation algorithm to determine human stance joint locations. Joint positions are treated as vertices, and bone connections are treated as edges to construct a graph of gait sequences. An ST-GCN subsequently processes the sequence of graph structures. The processed spatial-temporal data are subsequently transmitted to the lower layers and integrated with an MPA module to determine and highlight individual body parts necessary for gait identification. Finally, the activation maps derived from gait sequences facilitate the classification of the corresponding sequence labels. Figure 3 presents a visual representation of the intended pipeline.

3.1. Preliminaries

The main aim of this work is to learn a mapping from a sequence of skeletal graphs to a gait feature representation that captures individuals' unique walking patterns. This problem can be formalized as follows: Let $G = (V, E)$ be a spatial-temporal graph, where $J = J_{t,j} | \ t = 1, \ldots, T, \ j = 1, \ldots, K$ represents the set of joints across T frames and E represents the edges connecting these joints. The adjacency matrix $A \in R^{JxJ}$ encodes the connectivity between joints, and the feature tensor $X \in R^{CxJxT}$ contains the joint coordinates and confidence scores. The edges in the space and time dimension can be formed as E_{sd} and E_{td}. The E_{sd} and E_{td} are defined as $E_{sd} = \left\{k_i^t, \ k_j^t\right\} | i, \ j = 1, \ldots, K, \ i \neq j; t = 1, \ldots, T\}$ and $E_{td} = \left\{k_i^t, \ k_j^{t+1}\right\} | i, \ j = 1, \ldots, K, \ i \neq j; t = 1, \ldots, T\}$. Our intent is to learn a process f that maps X to a gait feature representation $f_{MAST-GCN}$, which can be employed for recognition. The process f is parameterized by a neural network that includes GCNs, TCNs, and MPA features to grab both spatial and temporal dependencies in the gait series. The final gait attribute $f_{MAST-GCN}$ is then utilized to calculate similarities between gait arrangements for recognition purposes.

3.2. Joint estimation from video sequences

Pose estimation algorithms are employed in this study to retrieve skeletal data from video frames. In particular, we employ HRNet [33], a high-resolution network fine-tuned on COCO, to predict 2D coordinates of j joint is $v_j = (x_j, y_j)$ and

Figure 3
The intended pipeline of the multi-part attention-guided framework for skeleton-based gait recognition

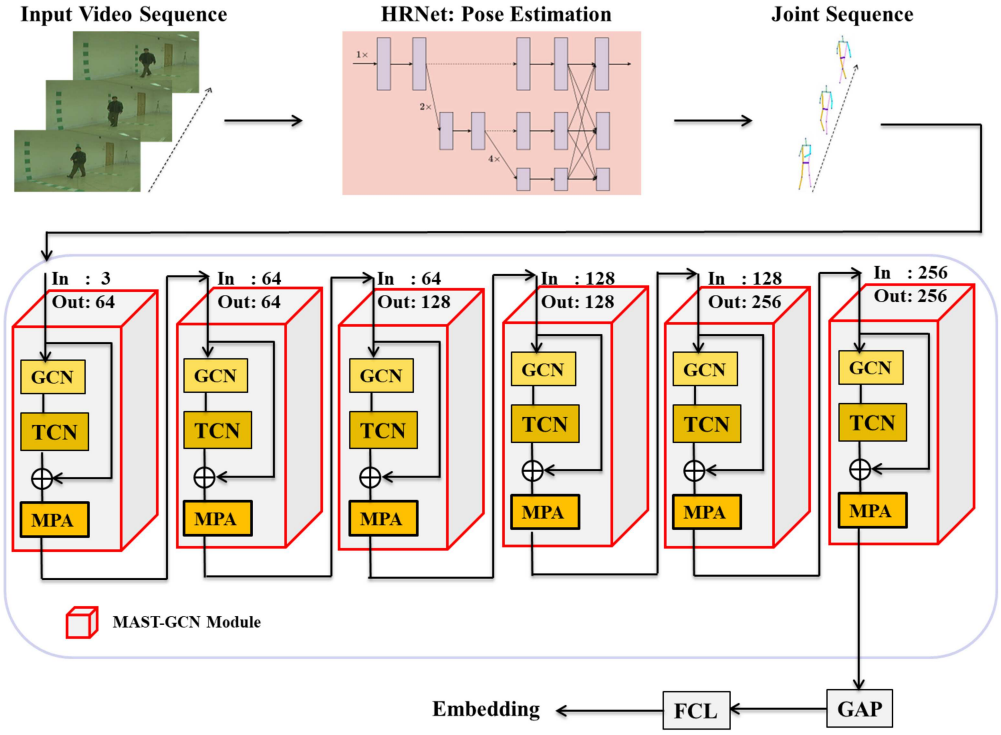
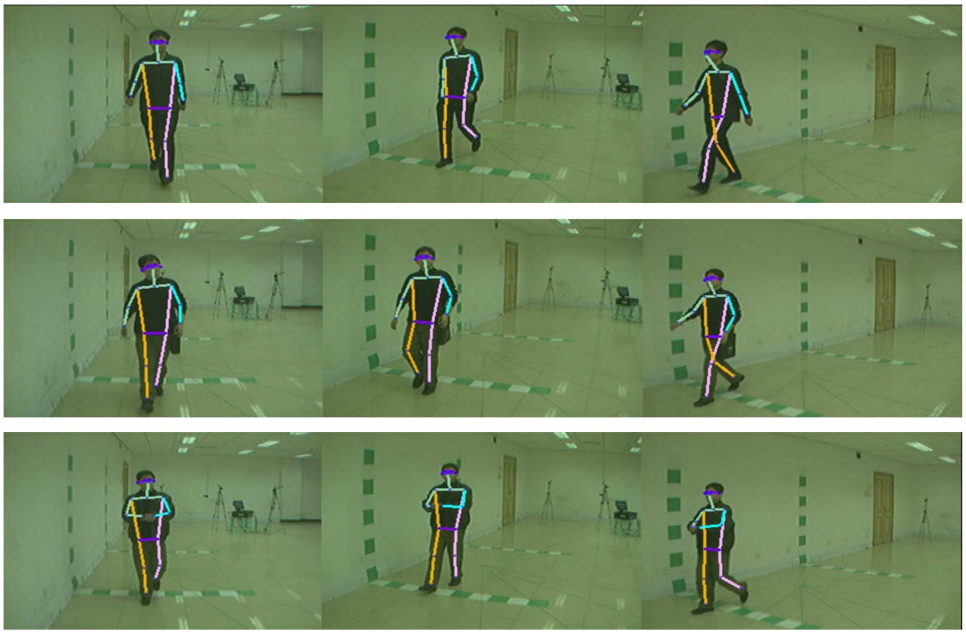


Figure 4
Skeletal sequences derived from the CASIA-B dataset utilizing the HRNet pose estimation framework: (a) normal walking, (b) walking with a bag, and (c) walking with a coat



confidence scores is θ_j for each body key point j. This step translates frames of the gait series into a skeleton graph. These skeleton sequences are further utilized as the input of our gait recognition model. For robustness intent, we pitch away samples with frames. Figure 4 delivers a skeletal sequence extracted based on the CASIA-B dataset using the HRNet pose estimation framework.

3.3. MAST-GCN: Multi-Part Attention-Guided Spatial-Temporal GCN

The ST-GCN model consists of 10 ST-GCN modules, which are categorized into three primary blocks. The fundamental differences among the blocks are the input and output channels. The initial block presents input to the 64 output channels. The

subsequent three consecutive units in the first block represent the 64 output channels. The following three modules present 128 output channels within the middle block. The concluding block comprises three modules, each with 256 output channels. Figure 5(a) illustrates the essential configuration of ST-GCN. The single ST-GCN module consists of a GCN and a TCN incorporated with a residual connection, as depicted in Figure 5(b). A trainable edge weight parameter quantifies the importance of node connections within individual ST-GCN modules. This analysis removed duplicative modules from the ST-GCN blocks to simplify their structure. Following the removal of duplicative blocks, the simplified ST-GCN model contains six ST-GCN modules in the three blocks, as displayed in Figure 6(a). The simplified ST-GCN model reduced the significant number of parameters.

Integrating the MPA process with ST-GCN shows meaningful benefits in gait recognition by improving the model's capability to concentrate on the considerable informative spatial and temporal features. MPA allows the model to prioritize key body parts in a gait sequence, extracting the most discriminative temporal patterns. Moreover, the attention module dynamically recalibrates the importance of different features of body parts, highlighting body parts that contribute most to gait recognition. Integrating the attention mechanism, ST-GCN can perform as a better model for the complex coordination between body parts, enhancing feature representation and robustness against variations such as clothing, carrying conditions, and viewing angles. The basic structure of the MAST-GCN is illustrated in Figure 6(b). The single module of MAST-GCN is depicted in Figure 6(c).

Current GCN-based methods utilize multi-scale graph convolutions employing diverse adjacency matrices to grab long-range connections from faraway neighbors. This process prioritizes tighter joints over those that are more remote. An m order adjacency matrix (A_m) is offered to handle the previous restrictions.

The adjacency matrix process is demonstrated as $A_m^{i,j} = 1$, for distance $(k_i,\ k_j) = m$ or $i = j$, where the remaining values are zero. Here, distance $(k_i,\ k_j)$ is the lowest distance within the joints k_i and k_j. The process of the GC in the spatial dimension is represented as Equation (1).

$$f_{gcn} = \sigma(\sum\nolimits_m \Delta_m^{-\frac{1}{2}}\ (A_m + \alpha)\,\Delta_m^{-\frac{1}{2}} f_{in} w_m) \tag{1}$$

Here, α denotes a learnable weight matrix that relieves nominal edge revisions within the skeleton graph. Δ_m conveys the normalized diagonal degree matrix, defined as $\Delta_m^{ij} = \sum_j A_m^{ij} w_m$, where w_m signifies the weight matrix containing various output channels' weight vectors and $\sigma\,(\cdot)$ signifies an activation function.

The output of f_{gcn} is conceded to the TC network. The TCN employed the 2D convolutional process with a fixed kernel $1 \times L$ on the TCN to capture temporally important properties from the f_{gcn} module outcomes. The procedure of TCN is assessed by Equation (2), where the output of $GCN\,(f_{gcn})$ is connected to the batch normalization ($bn1$), Rectified Linear Unit (ReLU), 2D Conv (W_c), batch normalization ($bn2$), and dropout operation (d_{out}).

$$f_{tcn}(x) = f_{dropout}(f_{bn2}(W_c\ ReLU\,(f_{bn1}(x))),\ d_{out}) \tag{2}$$

Actual gait recognition relies on extracting individual movement characteristics of vital body components, notably the arms and legs, reflecting important information for unique steps. Prioritizing modifications in joint positions in these regions allows a method to extract specific features critical for recognition

Figure 5
(a) ST-GCN model backbone and (b) ST-GCN unit

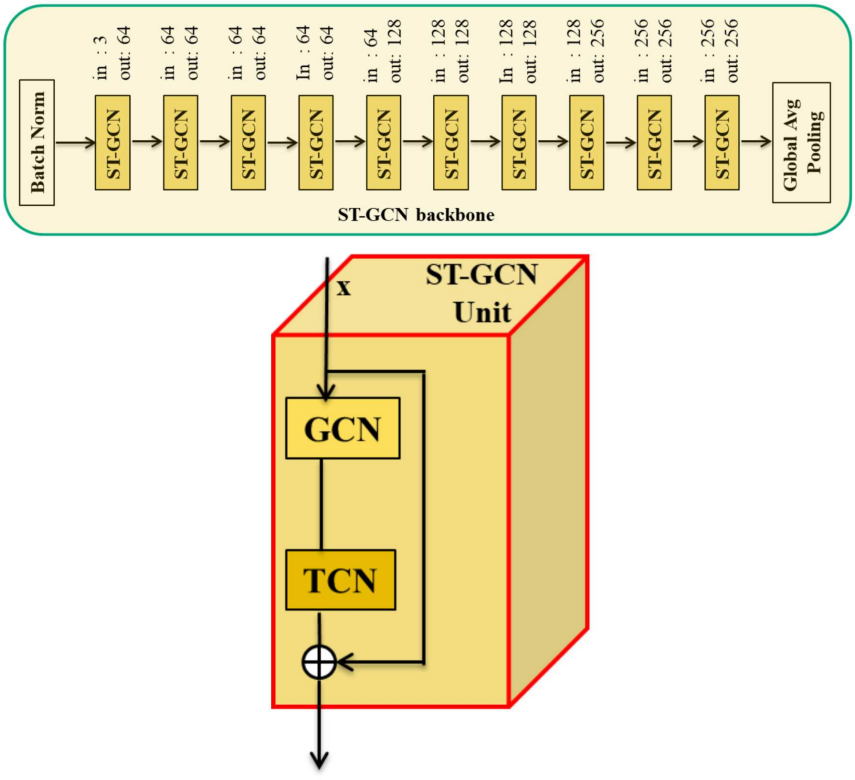
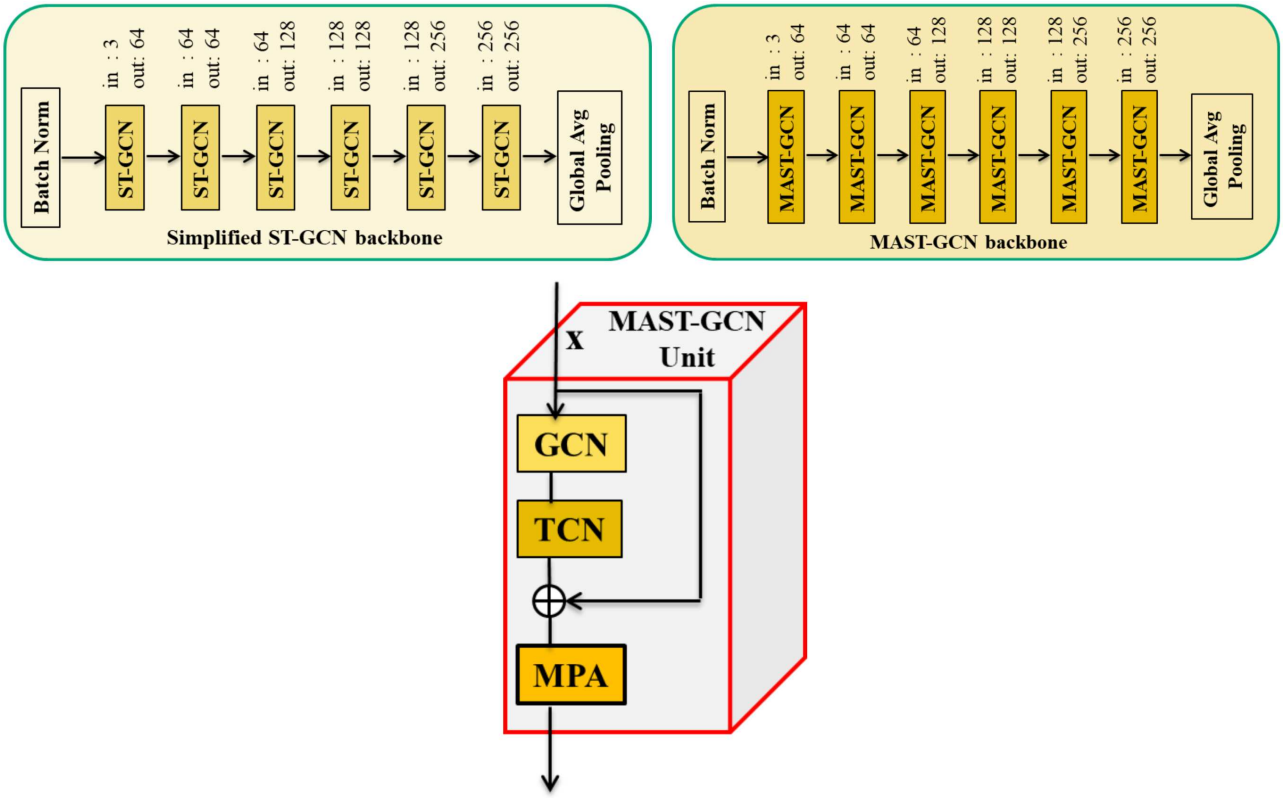


Figure 6
(a) Simplified ST-GCN model, (b) MAST-GCN model backbone, and (c) MAST-GCN module



while simultaneously decreasing the influence of noise from less dependable joints. Since invariant predictions of the body joints can be incorrect due to utilizing off-the-shelf pose estimation models trained on distinct datasets, including an attention mechanism is crucial. The attention mechanism assigns weighting to informative joints so that the more informative entities will have higher weight training, while the model focuses on not paying too much attention to overfit samples. Inspired by the split attention method, the proposed part-wise attention module divides the skeleton into five important parts and dynamically emphasizes vital joint relations. Typically, the five parts represent the left arm, the right arm, the left leg, the right leg, and the torso–head. This structure is consistent with the natural biomechanical structure of human locomotion. At the same time, it keeps the model simple and effective. These regions correspond to the main motion groups involved in walking. The legs generate propulsion, the arms maintain balance, and the torso–head stabilizes the body. This technique optimizes the extraction of both local and long-range dependencies that are essential for reliable gait recognition. Figure 7 illustrates that attention weights are obtained through a multilayer perceptron (MLP) comprising two layers that employ a softmax activation function.

The output of the MLP is normalized, ensuring that the weights of the attention sum to one across all components. The weighted attribute maps from each component are fused to form the significant feature representation. The joints are manually grouped into five anatomical regions from the input features to handle part-wise attention. The features of each segment are then concatenated and averaged along the temporal dimension. Finally, the new feature map is processed with fully connected layers (FCL), batch normalization, and ReLU activation functions.

Spatial attention operates by selectively re-weighting its video frames, while average pooling on a temporal dimension extracts global contextual feature maps across the entire sequence. This global pooling is also combined with batch normalization, which provides stability to the training process by ensuring similar weights across layers. Moreover, the ReLU activation function discards negative values and focuses only on the most significant parts of the input. The attention operates as a nonlinear operation, which enables the assignment of larger weights to informative features. It regularizes the network to capture the input's key properties and increase the discriminative capability. Then, five FCL, followed by part-level softmax, are used to generate attention scores for individual anatomical parts. The computed scores are used to scale the corresponding feature values, with the most significant contributions from each segment prioritized for improved representation.

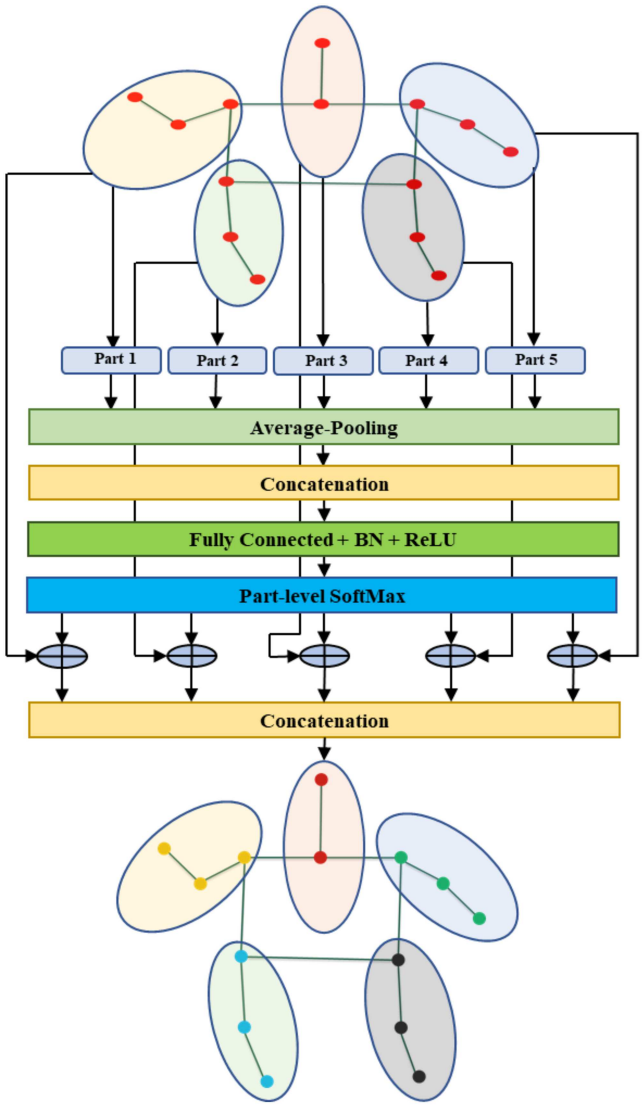
The feature values for each component, p_t, are computed using Equation (3).

$$f_{p_t}(x) = x(p_t) \bigotimes S_{part}(ReLU(T_{pool}(x)\, W) W_{pt}) \tag{3}$$

Here, x is the input feature, S_{part} is the softmax of the part level, and T_{pool} is the temporal average pooling. W and W_{pt} are parameters that can be learned, where W serves as a standard for every component of the body to facilitate reduction in dimensions and W_{pt} is calculated for individual body parts, indicating attention weights. The attribute vectors are combined to recreate the skeleton representation by Equation (4), denoted as f_{MPA}.

$$f_{MPA} = Concat_p \left(f_{pt} \middle| p_t = 1, \ldots, 5 \right) \tag{4}$$

Figure 7
Multi-part attention module estimates the weights for the five body parts through attention mechanism



The module will enhance the transparency and robustness of the models by determining the most important features driving the final forecast. Specific training on informative fields reduces training complexity. The focus is on allocating computing resources to the most important areas. This narrow focus optimizes resource utilization and shields the model from extraneous or distracting inputs. Thus, the network effectively removes the outside information and obtains relations among the activation maps that depend on the context, leading to more informed and accurate decisions. It is owing to the robustness to variation and clarity in reasoning that performance is better.

The process of MAST-GCN is processed according to Equation (5).

$$MSTA - GCN(x) = f_{MAST-GCN}(x) = f_{MPA}\left(f_{tcn}\left(f_{gcn}\left(x\right)\right)\right) \quad (5)$$

The attribute map built by the final MAST-GCN is directed to the global average pooling (GAP). The GAP significantly highlights important joints while eliminating less essential ones. The GAP attribute map is then processed via FCL of size 512. Ultimately, this results in the formation of the feature embedding vector, as illustrated in Figure 3.

4. Experimental Results

The proposed gait recognition model is evaluated for its efficacy by examining the publicly available CASIA-B [34] and OUMVLP-Pose [35] datasets. For that, initially, it presents an extensive summary of the datasets. The effectiveness of the proposed procedure is evaluated by comparing it with other advanced gait recognition methods. Finally, the ablation study examines the impact of several components of the method to support and assess the effectiveness of the suggested approach. The investigations were conducted in a PyTorch Lightning setting with a T4 NVIDIA GPU, 16 GB of Video Random Access Memory, and CUDA version 11.6, fine-tuned for artificial intelligence applications. The system also includes a 3.5 GHz CPU and 16 GB of RAM to meet the computing needs.

4.1. Dataset

CASIA-B dataset [34] has a large-scale multi-view gait series, one of the most widely operated standard datasets for gait recognition. CASIA-B retains 124 individuals (001–124), each carried in 11 distinct poses (0°, 18°, ..., 180°) and under three walking states: normal walking (NM), walking with a bag (BG), and walking while wearing a coat (CL). There are 110 series per individual, containing six series of NM, two series of BG, and two of CL per view angle. We track the standard experimental evaluation protocol, consisting of a training set (subjects 001–074) and a test set (subjects 075–124). The gallery consists of the first four NM series from the test set, while NM# 5–6, BG# 1–2, and CL# 1–2 are the probes. This design provides a complete test across diverse strategies, and thus, CASIA-B is a widely operated dataset to assess gait recognition applications.

OUMVLP-Pose dataset, introduced by An et al. [35], is a large-scale multi-viewpoint gait database containing 10,307 subjects. For each subject, 10 gait sequences were captured at 14 distinct viewing angles, spanning 0°–90° and 180°–270° in 15° increments. Pose annotations for these sequences were obtained by applying the state-of-the-art pose estimation algorithm OpenPose [36] to the original RGB video frames. The dataset was split into a training set of 5153 subjects and a test set of 5154 subjects. For evaluation purposes, the test set was further divided into separate gallery and probe subsets.

4.2. Augmentation

The model's generalization and robustness were enhanced by data augmentation. To skill up reverse walking and diversify the sample collection, we performed flipping on Video frames. Vertical-axis mirroring was applied to simulate an individual walking in the opposite direction to the recording device. Joint coordinate features were explicitly accounted for and mitigated with controlled Gaussian noise (sigma = 0.10) to remove common shortcomings of the upstream pose estimation pipeline. Significantly, the training subset was augmented only dynamically. The augmentation approach also protects against data leakage, which occurs when pre-split augmentation introduces structurally similar yet distinct examples of the same original work into the training and test sets [37], leading to overstated generalization accuracy.

4.3. Effect of confidence threshold

The baseline model performance was carefully examined across a range of potential thresholds. The process is performed to determine the appropriate confidence threshold (T_c) for filtering low-quality frames. The vanilla dataset ($T_c = 0$) was used, where no frames were deleted. The empirical study showed that at a 50% and 60% confidence level, 284 and 2910 frames were reduced, respectively. Additionally, 8152 frames were pruned at an index confidence threshold of 70%, which was much higher than the previous two confidence levels. These additional frames helped reduce the amount of sequential data required for effective gait representation, thereby improving the architecture's ability. Since the 284 and 2910 frames are considerable, the leading operating approach was to remove frames with average confidence below 60%.

4.4. Experimental settings

The well-known CASIA-B dataset is used to measure the rank-1 accuracy for the MAST-GCN method. The model was trained on the CASIA-B dataset with 200 epochs, batch size 128, Adam as an optimizer, and implementation in the PyTorch Lightning framework. The learning rate starts at 0.01 and drops by a factor of 10 every 20 epochs. Supervised Contrastive Loss [33] defined how far we wanted to be from what we were against. For the OUMVLP-Pose dataset, training was performed for 1.2×10^6 iterations with a batch size of 1024. The learning rate was decayed by a factor of 0.1 every 300 iterations. During testing, the similarity between each gallery and probe sequence was measured by computing the cosine similarity of their 128-dimensional feature vectors extracted from the model's fully connected (FC) layer.

4.5. MAST-GCN performance

The model is evaluated on the CASIA-B and OUMVLP-Pose datasets, and the results are presented in Tables 1 and 2, respectively. Results from the CASIA-B dataset are shown for walks under different conditions and for accuracy across 11 viewing angles. The model is trained with the 8140 gait sequences. To evaluate the proposed MAST-GCN model, the CASIA-B dataset is split into a probe and a gallery set. Among the evaluated 5300 gait sequences, 2200 sequences are used for the gallery set, and the remaining 3300 sequences are used for the probe set. The model is assessed using two factors: walking conditions and different view angles. The model is assessed across 11 view conditions, with each condition treated as an independent variable. For that, 11 unique experiments were conducted, and the accuracy for the different walking conditions is estimated and shown in Table 1. The mean and standard deviation (SD) are measured for the three different runs. The 95% confidence is also measured and presented in the table. The results presented in Table 1 demonstrate that the carrying bag and clothing conditions show moderate results compared to the normal walking condition. The reason is that both the carrying bag and the clothing condition are affected by different covariate factors and visual clues. Figure 8 illustrates the per-viewing angle and average accuracy under walking conditions.

The model is further evaluated on the OUMVLP-Pose dataset, which does not include any covariate factors other than the 14 different view angles. To measure the model's performance on the OUMVLP-Pose dataset, it is evaluated using 14 angles across three runs. The results for OUMVLP-Pose, with mean and SD, are shown in Table 2 for the 14 different viewing angles. The 95% confidence interval (CI) is presented in the table.

Table 1
The experimental result of gait recognition achieved by the proposed method for 11 different views and different walking conditions on the CASIA-B dataset with three runs

Gallery NM# 01-04	0–180 degree												
Probe	0	18	36	54	72	90	108	126	144	162	180	Mean ± SD	95% CI
NM# 5–6	95.3 ± 0.3	96.8 ± 0.4	96.6 ± 0.3	95.4 ± 1.1	95.8 ± 0.4	96.9 ± 1.4	94.5 ± 0.9	95.9 ± 0.3	97.1 ± 0.5	95.6 ± 0.4	94.6 ± 1.6	95.8 ± 1.12	[95.05, 96.55]
BG# 1–2	90.9 ± 0.1	92.6 ± 0.3	92.4 ± 0.1	91.6 ± 1.6	91.7 ± 0.3	92.6 ± 1.9	90.4 ± 1.2	92.5 ± 1.3	93.6 ± 0.3	91.7 ± 0.7	89.7 ± 3.1	91.8 ± 1.55	[90.76, 92.84]
CL# 1–2	85.8 ± 1.3	88.6 ± 1.1	88.6 ± 0.5	88.3 ± 0.8	89.2 ± 0.8	89.2 ± 1.3	89.1 ± 2.5	90.6 ± 1.9	90.8 ± 0.5	88.9 ± 0.8	85.2 ± 2.6	88.5 ± 2.05	[87.12, 89.88]
Mean	91.2 ± 0.6	93.2 ± 0.6	93.0 ± 0.3	92.3 ± 1.2	92.7 ± 0.5	91.9 ± 1.6	93.5 ± 1.5	94.2 ± 1.2	92.5 ± 0.4	92.6 ± 0.6	87.5 ± 2.4	92.5 ± 1.01	[91.82, 93.18]

Table 2
Average rank-1 performance on the OUMVLP-Pose dataset with three runs

Gallery # 01	0°–270°																
Probe #	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean ± SD	95% CI
00	89.9 ± 3.2	94.9 ± 0.6	93.5 ± 0.6	91.6 ± 2.0	92.5 ± 0.6	95.1 ± 2.4	90.5 ± 3.2	91.7 ± 1.4	94.9 ± 1.6	92.4 ± 1.2	92.5 ± 0.8	93.4 ± 0.5	96.2 ± 0.1	92.6 ± 0.9	93.0 ± 1.4	[92.82, 93.18]

4.6. Analysis of variance (ANOVA) for rank-1 accuracy

A two-way analysis of variance (ANOVA) [38, 39] was conducted to evaluate the influence of the walking condition and viewing angle on the gait recognition metric shown in Table 3. The analysis revealed that both primary factors had a highly statistically significant impact on the model's performance. The main effect of the walking condition was particularly pronounced, demonstrating an exceptionally large effect size, $F(2, 66) = 288.87$, $p < 0.001$, $\eta^2_p = 0.9$, confirming that variations in carrying or clothing status are the overwhelmingly dominant source of performance variability. In alignment with known challenges in appearance-based gait recognition, the viewing angle also presented a significant and substantial main effect, $F(10, 66) = 7.54$, $p < 0.001$, $\eta^2_p = 0.53$. Crucially, the interaction between walking condition and view angle was found to be statistically non-significant, $F(20, 66) = 1.15$, $p = 0.329$. This suggests that the performance degradation caused by a nonstandard walking condition does not reliably vary depending on the specific camera angle, indicating that these two performance challenges contribute independently to the overall recognition instability.

A one-way ANOVA was performed to determine the significance of the 14 distinct viewing angles on the gait recognition performance metric shown in Table 4. The analysis indicated a statistically significant effect for the view angle, $F(13, 28) = 2.51$, $p < 0.001$, confirming that performance means differ reliably across the various perspectives. Importantly, the practical significance of the factor was measured with eta squared (η^2), which was 0.62. This large effect size ($\eta^2 \geq 0.14$) implies that the viewing angle explains as much as 62% of the total variance in recognition performance. The statistical superiority obtained underscores the persistent difficulty of cross-view gait recognition. The process validates the conclusion that the camera's viewpoint remains the overwhelming determinant of performance stability, even when advanced techniques are used to improve view-invariance.

4.7. Performance of MAST-GCN

A model's complexity has an important bearing on its accuracy. The learnable parameters, floating-point operations (FLOPs), and inference time play an important role in the complexity and importance of the deep learning model. The complexity parameters of the proposed model are listed in Table 5. An attention module made of GCN and TCN has only a few more parameters than ST-GCN, owing to the attention mechanism.

4.8. Discussion

The presented model is a simplified ST-GCN approach. It is accomplished by extracting the repetitious ST-GCN modules from the fundamental ST-GCN model. The simplified ST-GCN module incorporates MPA with the ST-GCN module. The method specifics are presented in Section 3. The suggested model is assessed against top methods, including viewing angles and walking scenarios, normal, bag-carrying, and varying clothing conditions. Table 6 typically demonstrates the empirical superiority of our proposed approach against the present state-of-the-art methods.

As can be seen in Table 6, the proposed model outperforms other models under normal operating conditions with a mean accuracy of 95.8%, which is the highest accuracy among the

Figure 8
The accuracy of various angles and walking conditions: (a) accuracy across 11 viewing angles for three distinct walking conditions and (b) average accuracy for the three walking conditions

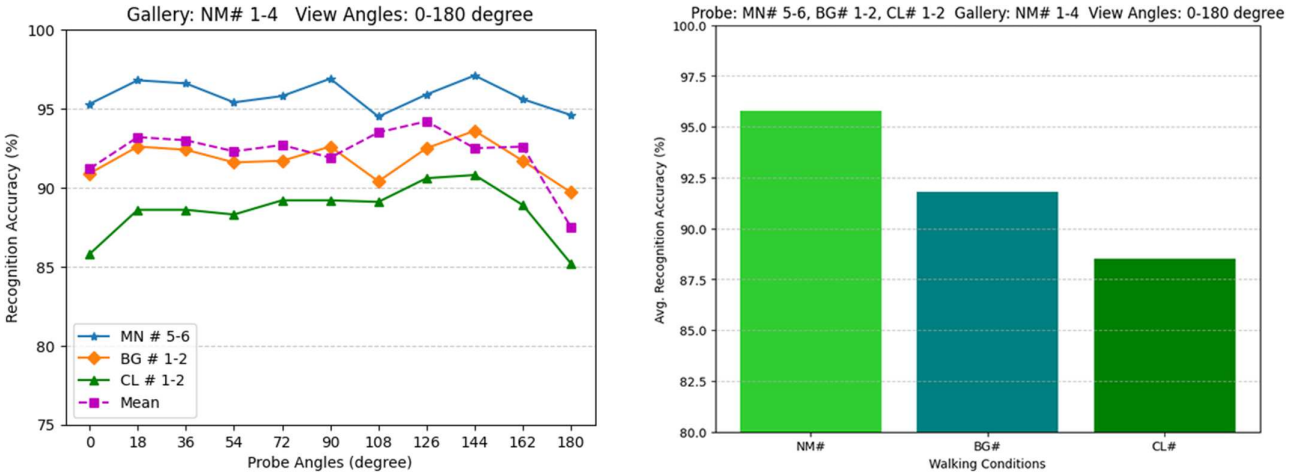


Table 3
Two-way ANOVA test results for rank-1 accuracy on the CASIA-B dataset

	Sum of squares (SE)	df	Mean squares	F	p	η^2_p
Walking	884.85	2	442.43	288.87	<0.001	0.9
View	115.46	10	11.55	7.54	<0.001	0.53
Walking × View	35.1	20	1.76	1.15	0.329	0.26
Error	101.08	66	1.53			

Table 4
One-way ANOVA test results for rank-1 accuracy on the OUMVLP-Pose dataset

	Sum of squares (SE)	df	Mean squares	F	p	η^2
View	126.61	13	9.74	3.51	<0.001	0.62
Error	77.68	28	2.77			

Table 5
Learnable parameters, FLOPs, and inference time in different network structures

Networks	#	MPA	FC + E	Params (M)	FLOPs (M)	InTime (s)
ST-GCN	10	x	√	2.8	277.7	0.08
EST-GCN	6	x	√	1.6	162.7	0.04
MAST-GCN	6	√	√	1.9	196.5	0.04

Note: # = number of blocks; **E** = embedding; **Params** = parameters; **InTime** = inference time.

selected models. The proposed model achieves 91.8% accuracy on the bag-carrying state (BG#). The proposed model improves the 10.5% relative accuracy against the best previous model. The proposed approach is shown to work well in clothing. Under the CL# case, the model achieved an accuracy of 88.5%, which is significantly better than the methods described in Table 6.

The proposed approach differs from many state-of-the-art methods using the skeleton-based module. Based on the attention module, the proposed method outperforms the MS-Gait under different walking conditions (NM#, BG#, CL#) with significant improvements. As shown in the table, the proposed model achieved the best performance on NM# walking conditions, as incorporating the MPA module with the ST-GCN operator considerably improved the representation of global body part dependencies in a regular walking sequence. In addition, the proposed MPA attention mechanism considers spatial and temporal dimensions to dynamically assign weights to the top two most relevant features. This process allows the model to condense on the most divergent features of gait, increasing resilience to nuances like carrying conditions and variations in perspective.

The OUMVLP-Pose dataset contains a large population of multi-view and multi-pose data, wherein the raw gait sequences are preprocessed. In the experiment, the OUMVLP-Pose dataset is used to compare the results with the baseline methods GaitGraph2 [33], Gait-D [36], and ResGait [37]. Table 7 presents the experimental results of MAST-GCN for the OUMVLP-Pose dataset compared to the other existing methods. The experiments show that the MAST-GCN outperforms Gait-D by around 2.0% in terms of the mean accuracy. From the results, it is shown that the MAST-GCN approach with part-based

Table 6
The results of the proposed framework are compared to the current state-of-the-art methods. The row indicates the variation in accuracy compared to state-of-the-art results. Bold represents the highest outcome

	Gallery NM#1–4	0–180 degree											
Probe	References	0	18	36	54	72	90	108	126	144	162	180	Mean
NM# 5–6	PoseGait [5]	55.3	69.6	73.9	75.0	68.0	68.2	71.1	72.9	76.1	70.4	55.4	68.7
	GaitGraph [7]	85.3	88.5	91.0	92.5	87.2	86.5	88.4	89.2	87.9	85.9	81.9	87.7
	MS-Gait [20]	89.4	91.7	91.6	90.2	90.6	90.6	90.4	90.9	90.4	88.5	85.6	90.0
	GaitGraph2 [33]	78.1	82.1	85.1	85.1	83.0	81.1	84.0	83.0	84.0	81.1	71.1	82.0
	GaitSkeleton [23]	87.8	90.6	92.0	93.3	90.1	91.5	90.2	89.6	91.0	90.8	85.0	90.2
	FLF [24]	93.7	93.8	95.8	95.8	91.4	92.3	91.7	93.5	94.3	93.3	91.0	93.3
	MAST-GCN (Ours)	95.3	96.8	96.6	95.4	95.8	96.9	94.5	95.9	97.1	95.6	94.6	95.8
BG# 1–2	PoseGait [5]	35.3	47.2	52.4	46.9	45.5	43.9	46.1	48.1	49.4	43.6	31.1	44.5
	GaitGraph [7]	75.8	76.7	75.9	76.1	71.4	73.9	78.0	74.7	75.4	75.4	69.2	74.8
	MS-Gait [20]	75.7	84.8	83.7	83.2	80.6	80.1	82.2	79.8	79.1	75.9	71.1	79.7
	GaitGraph2 [33]	69.9	75.9	78.1	79.3	71.4	71.7	74.3	76.2	73.2	73.4	61.7	73.2
	GaitSkeleton [23]	79.0	80.5	80.2	81.2	79.6	77.5	81.4	78.7	76.1	77.4	72.3	78.5
	FLF [24]	81.6	81.1	85.3	85.6	79.4	81.0	77.5	82.3	82.4	82.7	75.9	81.3
	MAST-GCN (Ours)	90.9	92.6	92.4	91.6	91.7	92.6	90.4	92.5	93.6	91.7	89.7	91.8
CL# 1–2	PoseGait [5]	24.3	29.7	41.3	38.8	38.2	38.5	41.6	44.9	42.2	33.4	22.5	35.9
	GaitGraph [7]	69.6	66.1	68.8	67.2	64.5	62.0	69.5	65.6	65.7	66.1	64.3	66.3
	MS-Gait [20]	75.1	79.7	80.5	84.7	84.0	82.4	79.8	80.4	78.3	78.0	70.9	79.4
	GaitGraph2 [33]	57.1	61.1	68.9	66.0	67.8	65.4	68.1	67.2	63.7	63.6	50.4	63.6
	GaitSkeleton [23]	71.2	72.9	68.2	71.2	70.0	69.9	77.7	72.8	74.7	76.8	71.8	72.4
	FLF [24]	72.3	72.0	73.8	77.9	61.3	67.1	73.7	74.5	76.6	75.3	72.1	72.1
	MAST-GCN (Ours)	85.8	88.6	88.6	88.3	89.2	89.2	89.1	90.6	90.8	88.9	85.2	88.5

Table 7
A comparative analysis was performed against existing state-of-the-art approaches on the OUMVLP-Pose dataset, using the average rank-1 recognition accuracy (%) as the primary evaluation metric

Gallery #01 Probe #00	Prove View (0–90) (180–270) degree													
Method	0	15	30	45	60	75	90	180	195	210	225	240	255	270	Mean
GaitGraph2 [33]	32.9	47.7	53.9	56.8	53.9	54.7	45.4	29.0	35.7	34.3	44.3	46.2	46.4	38.4	44.3
Gait-D [36]	84.3	92.6	90.6	92.1	90.5	91.3	92.1	87.6	90.4	92.6	91.3	92.2	94.5	92.3	91.0
ResGait [37]	39.6	49.3	56.2	58.1	57.3	59.6	47.7	35.5	40.2	43.3	47.2	54.9	55.3	46.2	49.3
MAST-GCN (Ours)	89.9	94.9	93.5	91.6	92.5	95.1	90.5	91.7	94.9	92.4	92.5	93.4	96.2	92.6	93.0
△	5.6	2.3	2.9	–0.5	2.0	3.8	1.6	4.1	4.5	–0.2	1.2	1.2	1.7	0.3	2.0

attention-driven feature extraction method is more accurate in enhancing the features' gait.

4.9. Cross-dataset evaluation

We present a summary comparison of the in-domain and cross-dataset rank-1 accuracies on the evaluated gait benchmarks (CASIA-B and OUMVLP-Pose) in Table 8. The in-domain results on CASIA-B are still high (NM: 95.8%, BG: 91.8%, and CL: 88.5%). This outcome is due to the regular motion patterns and controlled imaging setup. The method generalized in the work enables the model to learn robust spatial-temporal descriptors. The performance is significantly stable when the model is trained on OUMVLP-Pose and tested on CASIA-B (NM: 78.2%, BG: 71.9%, CL: 69.4%). This strength highlights the advantage of training on a large, multi-view dataset whose diversity equips the model with broadly transferable geometric and temporal priors. In contrast, transfer between CASIA-B→OUMVLP-Pose results in zero-shot accuracy of 65.4%, which is more moderate and reflects the common limitation of models trained from narrow, laboratory-limited sources when applied to a dataset with a large variation range of view angles, walking styles, and pose estimation noise. Together, these findings underline the asymmetric nature of cross-dataset generalization and emphasize the importance of scale and diversity in constructing transferable gait representations.

4.10. Ablation study

The proposed MAST-GCN method is built on the backbone ST-GCN method. It reduces the ST-GCN module by removing the extra blocks from the basic ST-GCN method. Then, the MPA module is fused with the ST-GCN model. First, we perform the ST-GCN module to justify the effect of the proposed method. Later, we apply the plain process. The proposed method is compared with ST-GCN and simplified ST-GCN. Investigational effects are reported in Table 9.

The proposed method surpasses ST-GCN and the simplified ST-GCN over all three walking scenarios (NM#, BG#, CL#) (Table 6). The MAST-GCN shows significant accuracy improvements over the plain ST-GCN and simplified ST-GCN under the normal, bag-carrying, and clothing conditions of 1.7%, 1.8%, 7.5%, and 0.8%, 0.8%, and 6.0%, respectively. This allows the MPA module to effectively learn multi-part body joint representations of high-capacity across joints in a skeleton. Since gait can be identified from different time frames, correlations between spatially distant body parts can significantly help identify unique movement styles like carrying a bag or wearing different clothes, which makes the approach efficient. In addition, both the ST-GCN basic network and the ST-GCN improved network have limitations in learning long-distance joint features.

4.11. Visualization

To interpret the discriminative patterns learned by our model, we employ the activation mapping technique to visualize joint-level activations across representative frames of input skeleton sequences [40]. As illustrated in Figure 9, the generated maps reveal the spatial emphasis of the model, where the gold and blue bounding boxes in the visualization represent the right and left sides of the body, respectively, and highlight joints and limbs of maximum displacement during gait cycles. The change in color from light green (low movement) to dark green (high

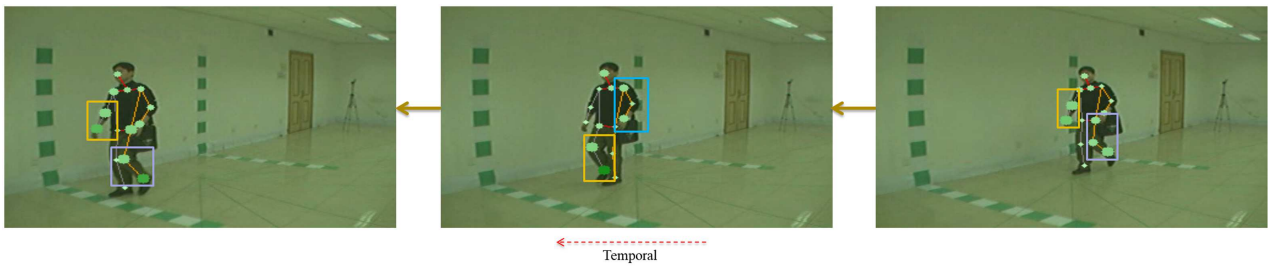
Table 8
Cross-datasets performance evaluation based on rank-1 accuracy (%)

	Test Dataset			
	CASIA-B			
Train Dataset	NM#	BG#	CL#	OUMVLP-Pose
CASIA-B	95.8	91.8	88.5	65.4
OUMVLP-Pose	78.2	71.9	69.4	93.0

Table 9
Experimental results of different framework components on the CASIA-B dataset

Framework components			Accuracy		
ST-GCN	Simplified ST-GCN	MAST-GCN	NM#	BG#	CL#
√	x	x	94.1	90.0	81.0
x	√	x	95.0	91.0	82.5
x	x	√	95.8	91.8	88.5

Figure 9
Activated joints of visualization



movement) and the scaling of the visual markers give a clear picture of the activity at the joint level. This pattern of asymmetric motion between the lateral elements of the body further supports the model's sensitivity to dynamic gait cues. The visual attention on those regions also provides evidence that MAST-GCN can model side-specific kinematic variations, which are important for robust and personalized gait recognition.

5. Conclusions

In this research paper, an MAST-GCN framework is proposed for gait-based person recognition, which employs the simplified ST-GCN. The simplified ST-GCN is formed by removing the replicated ST-GCN modules from the preliminary ST-GCN method. At the simplified ST-GCN module, an MPA process is incorporated to adaptively extract the reliance between the physically separated joints and overcome the limitations of conventional ST-GCN. The MPA module significantly enhances the capability to describe the global joint dependencies in further walking patterns and address the disturbances yielded by the bags and clothes during walking, improving the model's accuracy. The model shows excellent precision of 95.8%, 91.8%, and 88.5% at normal, carrying bag, and clothing conditions, respectively, on the CASIA-B dataset, while 93.0% on the OUMVLP-Pose dataset. However, the approach still struggles to capture features from the occluded joints. In the future, we will try to handle this situation and enhance the model's precision. Moreover, we will explore domain shift scenarios and adversarial conditions to enhance the understanding of the model's generalization capabilities in real-world scenarios.

Ethical Statement

The authors stated that this study was exempt from formal ethical approval because Chittagong University of Engineering and Technology does not require Institutional Review Board or ethics committee approval for secondary analysis of publicly available, de-identified benchmark datasets. This exemption is in line with the institution's Research Ethics Policy for the use of anonymized data for academic purposes.

The original CASIA-B and OUMVLP-Pose datasets consist of human subjects. However, this research utilized the dataset in a skeleton-only format. The data is in the form of spatiotemporal coordinates of joint movements, not raw video frames or facial images. Therefore, no personal characteristics or visual characteristics of the subjects could be obtained or processed with full compliance with the privacy protection standards and legal regulations for data subjects.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in CASIA-B at http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp and in OU-MVLP-Pose at http://www.am.sanken.osaka-u.ac.jp/BiometricDB/GaitMVLP.html.

Author Contribution Statement

Md. Khaliluzzaman: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing. **Kaushik Deb:** Validation, Resources, Writing – review & editing, Supervision, Project administration.

References

[1] Chen, J., Wang, Z., Yi, P., Zeng, K., He, Z., & Zou, Q. (2022). Gait pyramid attention network: Toward silhouette semantic relation learning for gait recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *4*(4), 582–595. https://doi.org/10.1109/TBIOM.2022.3213545

[2] Technavio. (2024). *Gait biometrics market analysis North America, Europe, APAC, South America, Middle East and Africa - US*. Canada, China, UK, Germany: size and forecast 2024–2028 https://www.technavio.com/report/biometrics-as-a-service-market-analysis

[3] Chao, H., Wang, K., He, Y., Zhang, J., & Feng, J. (2022). GaitSet: Cross-view gait recognition through utilizing gait as a deep set. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3467–3478. https://doi.org/10.1109/TPAMI.2021.3057879

[4] Khaliluzzaman, M., Uddin, A., Deb, K., & Hasan, M. J. (2023). Person recognition based on deep gait: A survey. *Sensors*, *23*(10), 4875. https://doi.org/10.3390/s23104875

[5] Liao, R., Yu, S., An, W., & Huang, Y. (2020). A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, *98*, 107069. https://doi.org/10.1016/j.patcog.2019.107069

[6] Fu, Y., Meng, S., Hou, S., Hu, X., & Huang, Y. (2023). GPGait: Generalized pose-based gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19595–19604. https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01795

[7] Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., & Rigoll, G. (2021). Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE International Conference on Image Processing*, 2314–2318. https://doi.org/10.1109/ICIP42928.2021.9506717

[8] Kabir, H., & Garg, N. (2023). Machine learning enabled orthogonal camera goniometry for accurate and robust contact angle measurements. *Scientific Reports*, *13*(1), 1497. https://doi.org/10.1038/s41598-023-28763-1

[9] Asif, M., Tiwana, M. I., Khan, U. S., Ahmad, M. W., Qureshi, W. S., & Iqbal, J. (2022). Human gait recognition subject to different covariate factors in a multi-view environment. *Results in Engineering*, *15*, 100556. https://doi.org/10.1016/j.rineng.2022.100556

[10] Zhao, L., Guo, L., Zhang, R., Xie, X., & Ye, X. (2022). mmGaitSet: Multimodal based gait recognition for countering carrying and clothing changes. *Applied Intelligence*, *52*(2), 2023–2036. https://doi.org/10.1007/s10489-021-02484-2

[11] Lin, B., Zhang, S., Wang, M., Li, L., & Yu, X. (2022). GaitGL: Learning discriminative global-local feature representations for gait recognition. arXiv. https://doi.org/10.48550/arXiv.2208.01380

[12] Xi, H., Xi, Y., Hu, C., & Yahyapour, R. (2025). FMCB-Gait: Fine-grained multimodal gait recognition with cues embedding and body parts distribution guidance. *Journal of King Saud University Computer and Information Sciences*, *37*(10), 311. https://doi.org/10.1007/s44443-025-00324-8

[13] Li, X., Makihara, Y., Xu, C., Yagi, Y., & Ren, M. (2020). Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13309–13319. https://doi.org/10.1109/CVPR42600.2020.01332

[14] Huang, X., Wang, X., Jin, Z., Yang, B., He, B., Feng, B., & Liu, W. (2023). Condition-adaptive graph convolution learning for skeleton-based gait recognition. *IEEE Transactions on Image Processing*, *32*, 4773–4784. https://doi.org/10.1109/TIP.2023.3305822

[15] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations*, 1–14.

[16] Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., & Bronstein, M. M. (2017). Geometric deep learning on graphs and manifolds using mixture model CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5115–5124. https://doi.org/10.1109/CVPR.2017.576

[17] Wei, S., Liu, W., Wei, F., Wang, C., & Xiong, N. N. (2024). Gaitdlf: Global and local fusion for skeleton-based gait recognition in the wild. *The Journal of Supercomputing*, *80*(12), 17606–17632. https://doi.org/10.1007/s11227-024-06089-7

[18] Shopon, M., Hsu, G. S. J., & Gavrilova, M. L. (2022). Multiview gait recognition on unconstrained path using graph convolutional neural network. *IEEE Access*, *10*, 54572–54588. https://doi.org/10.1109/ACCESS.2022.3176873

[19] Lin, B., Zhang, S., & Bao, F. (2020). Gait recognition with multiple-temporal-scale 3D convolutional neural network. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3054–3062. https://doi.org/10.1145/3394171.3413861

[20] Wang, L., Chen, J., Chen, Z., Liu, Y., & Yang, H. (2022). Multi-stream part-fused graph convolutional networks for skeleton-based gait recognition. *Connection Science*, *34*(1), 652–669. https://doi.org/10.1080/09540091.2022.2026294

[21] Hasan, M. B., Ahmed, T., Ahmed, S., & Kabir, M. H. (2023). GaitGCN++: Improving GCN-based gait recognition with part-wise attention and DropGraph. *Journal of King Saud University - Computer and Information Sciences*, *35*(7), 101641. https://doi.org/10.1016/j.jksuci.2023.101641

[22] Zhang, C., Chen, X.-P., Han, G.-Q., & Liu, X.-J. (2023). Spatial transformer network on skeleton-based gait recognition. *Expert Systems*, *40*(6), e13244. https://doi.org/10.1111/exsy.13244

[23] Zhu, D., Ji, L., Zhu, L., & Li, C. (2024). Gait coordination feature modeling and multi-scale gait representation for gait recognition. *International Journal of Machine Learning and Cybernetics*, *15*(9), 3791–3802. https://doi.org/10.1007/s13042-024-02120-8

[24] Ray, A., Uddin, M. Z., Hasan, K., Melody, Z. R., Sarker, P. K., & Ahad, M. A. R. (2024). Multi-biometric feature extraction from multiple pose estimation algorithms for cross-view gait recognition. *Sensors*, *24*(23), 7669. https://doi.org/10.3390/s24237669

[25] Peng, Y., Ma, K., Zhang, Y., & He, Z. (2024). Learning rich features for gait recognition by integrating skeletons

and silhouettes. *Multimedia Tools and Applications*, *83*(3), 7273–7294. https://doi.org/10.1007/s11042-023-15483-x

[26] Li, N., & Zhao, X. (2023). A multi-modal dataset for gait recognition under occlusion. *Applied Intelligence*, *53*(2), 1517–1534. https://doi.org/10.1007/s10489-022-03474-8

[27] Chen, G., Chen, X., Zheng, C., Wang, J., Liu, X., & Han, Y. (2024). Spatiotemporal smoothing aggregation enhanced multi-scale residual deep graph convolutional networks for skeleton-based gait recognition. *Applied Intelligence*, *54*(8), 6154–6174. https://doi.org/10.1007/s10489-024-05422-0

[28] Khaliluzzaman, M., & Deb, K. (2024). S2AT-GCN: A spatial self-attention temporal graph convolutional network for gait-based person recognition. In *2024 13th International Conference on Electrical and Computer Engineering*, 568–573. https://doi.org/10.1109/ICECE64886.2024.11024707

[29] Khaliluzzaman, M., Akhtar, M. N., & Deb, K. (2025). AST-GCN: An attention-guided spatial-temporal GCN framework for gait recognition. In *2025 2nd International Conference on Next-Generation Computing, IoT and Machine Learning*, 1–6. https://doi.org/10.1109/NCIM65934.2025.11160139

[30] Uddin, M. Z., Ray, A., Das, B., & Ahad, M. A. R. (2025). View-embedding GCN for skeleton-based cross-view gait recognition. *IEEE Transactions on Human-Machine Systems*, *55*(5), 674–685. https://doi.org/10.1109/THMS.2025.3595213

[31] Wang, A., Hou, Z., Lin, E., Li, X., Liang, J., & Zhou, X. (2025). GaitSTAGCN: Spatial-temporal attention graph convolutional networks for gait recognition. *Neurocomputing*, *654*, 131300. https://doi.org/10.1016/j.neucom.2025.131300

[32] Priyanka, D., & Mala, T. (2025). SFG-Net: Semantic relationship and hierarchical Fusion-based Graph Network for enhanced skeleton-based gait recognition. *Engineering Applications of Artificial Intelligence*, *148*, 110399. https://doi.org/10.1016/j.engappai.2025.110399

[33] Teepe, T., Gilg, J., Herzog, F., Hörmann, S., & Rigoll, G. (2022). Towards a deeper understanding of skeleton-based gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1569–1577. https://doi.org/10.1109/CVPRW56347.2022.00163

[34] Yu, S., Tan, D., & Tan, T. (2006). A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th international conference on pattern recognition*, *4*, 441–444. https://doi.org/10.1109/ICPR.2006.67

[35] An, W., Yu, S., Makihara, Y., Wu, X., Xu, C., Yu, Y., . . . , & Yagi, Y. (2020). Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *2*(4), 421–430. https://doi.org/10.1109/TBIOM.2020.3008862

[36] Gao, S., Yun, J., Zhao, Y., & Liu, L. (2022). Gait-D: Skeleton-based gait feature decomposition for gait recognition. *IET Computer Vision*, *16*(2), 111–125. https://doi.org/10.1049/cvi2.12070

[37] Gao, S., Tan, Z., Ning, J., Hou, B., & Li, L. (2023). ResGait: Gait feature refinement based on residual structure for gait recognition. *The Visual Computer*, *39*(8), 3455–3466. https://doi.org/10.1007/s00371-023-02973-0

[38] Catruna, A., Cosma, A., & Radoi, E. (2024). Gaitpt: Skeletons are all you need for gait recognition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–10. https://doi.org/10.1109/FG59268.2024.10581947

[39] Li, J., Wang, Z., Wang, C., & Su, W. (2024). GaitFormer: Leveraging dual-stream spatial–temporal Vision Transformer via a single low-cost RGB camera for clinical gait analysis. *Knowledge-Based Systems*, *295*, 111810. https://doi.org/10.1016/j.knosys.2024.111810

[40] Bao, T., Gao, J., Wang, J., Chen, Y., Xu, F., Qiao, G., & Li, F. (2023). A global bibliometric and visualized analysis of gait analysis and artificial intelligence research from 1992 to 2022. *Frontiers in Robotics and AI*, *10*, 1265543. https://doi.org/10.3389/frobt.2023.1265543

How to Cite: Khaliluzzaman, M., & Deb, K. (2026). A Multi-Part Attention-Guided Spatial-Temporal GCN Framework for Gait-Based Person Recognition. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62028402