

Article

Exploring Gait Recognition in Wild Nighttime Scenes

Haotian Li ¹, Wenjuan Gong ^{1,*}, Yutong Li ¹, Yikai Wu ¹, Kechen Li ¹ and Jordi González ²¹ Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China; hotelslee@foxmail.com (H.L.)² Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain;

* Correspondence: wenjuangong@upc.edu.cn

Abstract: Currently, gait recognition research is gradually expanding from ideal indoor environments to real-world outdoor scenarios. However, recognition scenarios in practical applications are often more complex than those considered in existing studies. For instance, real-world scenarios present multiple influencing factors, such as viewpoint variations and diverse carried items. Notably, many gait recognition tasks occur under low-light conditions at night. At present, research on gait recognition in nocturnal environments is relatively limited, and effective methods for nighttime gait recognition are lacking. To address this gap, this study extends gait recognition research to outdoor nighttime environments and introduces the first wild gait dataset encompassing both daytime and nighttime data, named Gait Recognition of Day and Night (GaitDN). Furthermore, to tackle the challenges posed by low-light conditions and other influencing factors in outdoor nighttime gait recognition, we propose a novel pose-based gait recognition framework called GaitSAT. This framework models the intrinsic correlations of human joints by integrating self-attention and graph convolution modules. We conduct a comprehensive evaluation of the proposed method and existing approaches using both the GaitDN dataset and other available datasets. The proposed GaitSAT achieves state-of-the-art performance on the OUMVLP, GREW, Gait3D, and GaitDN datasets, with Rank-1 accuracies of 60.77%, 57.37%, 22.90%, and 86.24%, respectively. Experimental results demonstrate that GaitSAT achieves higher accuracy and superior generalization capabilities compared to state-of-the-art pose-based methods.



Academic Editor: Pedro Couto

Received: 21 November 2024

Revised: 13 December 2024

Accepted: 14 December 2024

Published: 2 January 2025

Citation: Li, H.; Gong, W.; Li, Y.; Wu, Y.; Li, K.; González, J. Exploring Gait Recognition in Wild Nighttime Scenes. *Appl. Sci.* **2025**, *15*, 350. <https://doi.org/10.3390/app15010350>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: gait recognition; gait recognition dataset; graph convolution networks; temporal; convolutional network; deep learning

1. Introduction

Gait recognition aims to identify individuals based on their walking patterns, representing a crucial task in human identity recognition. Compared to widely applied identification methods such as facial recognition and fingerprinting, gait recognition offers advantages including enhanced covertness, long-distance recognition capability, and the ability to operate without subject cooperation. Over the past few decades, gait recognition algorithms have demonstrated significant progress, showing substantial potential in emerging industries such as smart communities and intelligent security systems.

Gait recognition in practical applications faces challenges from various complex and dynamic factors. To address these challenges, it is crucial to understand the impact of different influencing factors and develop datasets that reflect real-world scenarios. In early investigations, researchers identified several factors that potentially influence gait recognition accuracy and intentionally collected gait data encompassing various influencing

factors to conduct comprehensive studies. These factors include, but are not limited to, viewpoint variations, subjects' attire, and carried items. The development and analysis of diverse datasets play a vital role in advancing gait recognition technology and bridging the gap between laboratory conditions and real-world applications. Several datasets have been created to study these influencing factors. For instance, the CASIA-B [1] dataset incorporates gait sequences under three different walking conditions (normal, wearing a coat, and carrying a bag) from multiple viewpoints. However, early datasets such as CASIA-B and OU-MVLP [2] were collected in indoor laboratory environments, as shown in Figure 1a,b. These idealized scenarios fail to fully reflect the complexities encountered in real-world applications. Consequently, methods based on such indoor, idealized data often exhibit suboptimal performance when applied to outdoor, real-world scenarios [3].



Figure 1. Examples of existing mainstream datasets. (a) CASIA-B. (b) OU-MVLP. (c) GREW. (d) Gait3D.

In recent years, researchers have increasingly focused on gait recognition in the wild [3–5], attempting to investigate gait recognition in more realistic environments. This shift toward real-world scenarios has led to the development of more comprehensive and challenging datasets. For instance, the GREW dataset provides a large-scale benchmark for gait recognition in the wild, with the addition of a distractor set to enhance its applicability to real-world scenarios. The Gait3D dataset, collected in a large supermarket, offers various human representations. Examples of pedestrians from GREW and Gait3D are illustrated in Figure 1c,d, respectively. However, researchers have paid limited attention to variations in lighting conditions in outdoor environments, particularly regarding gait recognition in low-light nighttime scenarios. Under such conditions, the clarity of video captured by cameras significantly deteriorates, posing substantial challenges to existing gait recognition methods. Consequently, improving the accuracy of gait recognition in outdoor low-light conditions has emerged as a pressing issue that demands resolution.

Nevertheless, there is currently a lack of suitable datasets that provide gait recognition data in real-world nighttime scenarios. To further explore nighttime gait recognition, we constructed a dataset named Gait Recognition of Day and Night (GaitDN) using videos captured with outdoor public area cameras. GaitDN was collected in outdoor public areas near residential zones and comprises 1009 subjects with over 3000 gait sequences captured in unconstrained wild environments. To facilitate research and comparative analysis, GaitDN also includes gait data collected from the same area during daytime with ample illumination, as illustrated in Figure 2.

Through the analysis of the collected data, we observed a significant decrease in video clarity under low-light nighttime conditions, as illustrated in Figure 3. This results in blurred boundaries between pedestrians and backgrounds, making it difficult to extract clear human silhouettes from the original images. While silhouette-based methods have

been widely adopted and have shown promising results in indoor conditions [6,7], they face critical challenges in low-light outdoor environments. Imperfect silhouette extraction due to blurred boundaries leads to a loss of fine-grained details, particularly in low-light conditions. This loss of detail affects critical gait features such as the precise silhouettes of limb movements and subtle body posture changes during the gait cycle. Consequently, the degradation of these key characteristics significantly compromises the effectiveness of silhouette-based methods, which rely heavily on accurate and detailed body silhouettes for recognition. The aforementioned limitations highlight the need for alternative approaches that can maintain robust performance across various lighting conditions. To address this challenge, we propose a pose-based gait recognition framework, GaitSAT. Utilizing human pose information, which can be more reliably extracted in low-light conditions compared to silhouettes, GaitSAT maintains consistent performance across diverse lighting scenarios, solving the problem of performance degradation in challenging nighttime environments that plague silhouette-based methods. By focusing on body joints and their relationships, GaitSAT aims to capture essential gait characteristics even when detailed silhouettes are not discernible. The algorithm combines graph convolutional neural networks with self-attention mechanisms to effectively learn and analyze structural changes in the human body during walking, which can maximally mitigate the effects of lighting variations.

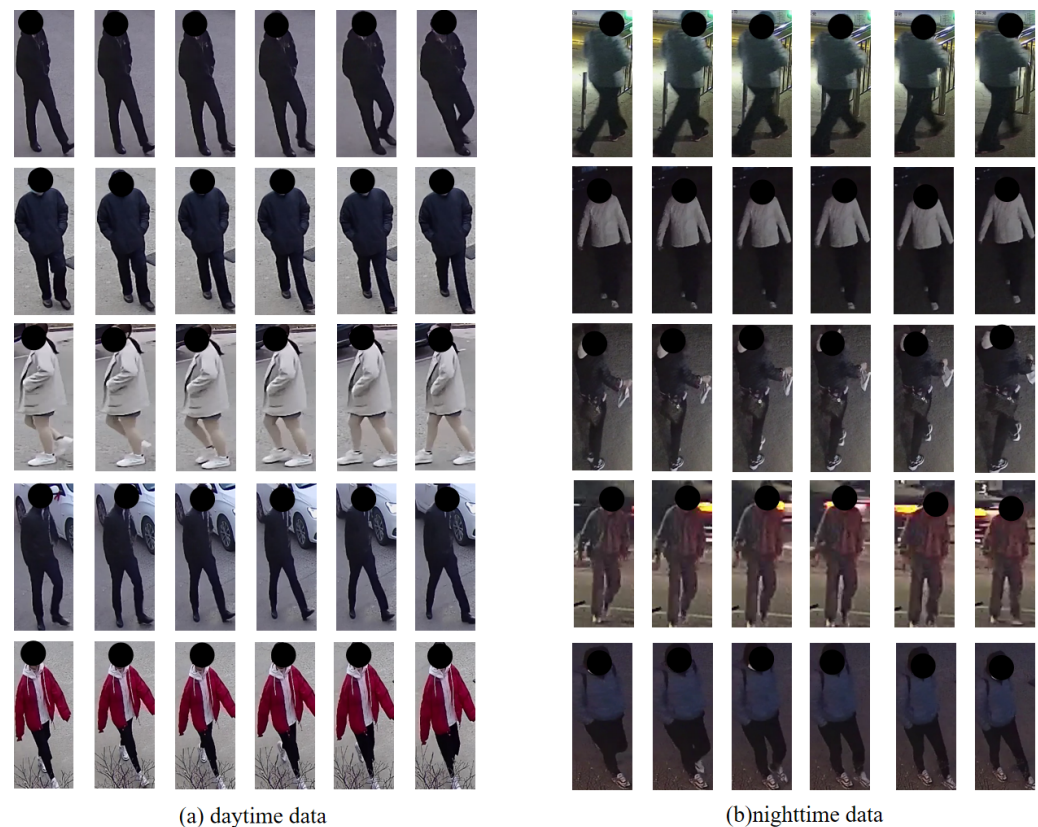


Figure 2. Examples of GaitDN dataset (with key information obscured). (a) Daytime data. (b) Night-time data.

Furthermore, considering that human walking is a process with periodic patterns, extracting effective temporal features is crucial. Many existing methods, such as those using Gait Energy Images (GEIs) or temporal convolutional networks [8], have inherent mechanisms to handle variations in walking speed to some extent. For instance, GEI-based [9,10] approaches implicitly average out speed variations over a gait cycle, while temporal convolutions can capture patterns across different time scales. However, in complex outdoor scenarios, pedestrians exhibit a wide range of walking speeds, which

poses significant challenges to existing methods. Although current temporal convolution approaches [8,11] have considered the importance of temporal features, they lack the adaptability required for more complex cases of speed variation, potentially leading to a loss in critical temporal information. To enhance the model's robustness and accuracy in these scenarios, we introduced a multi-scale temporal rhythm perception module. By integrating temporal information from different time steps, the module enhances the model's ability to handle diverse speed variations and walking patterns, improving recognition accuracy in complex outdoor environments. This adaptation significantly boosts the model's overall performance, providing a more comprehensive and effective approach to gait recognition in real-world settings.

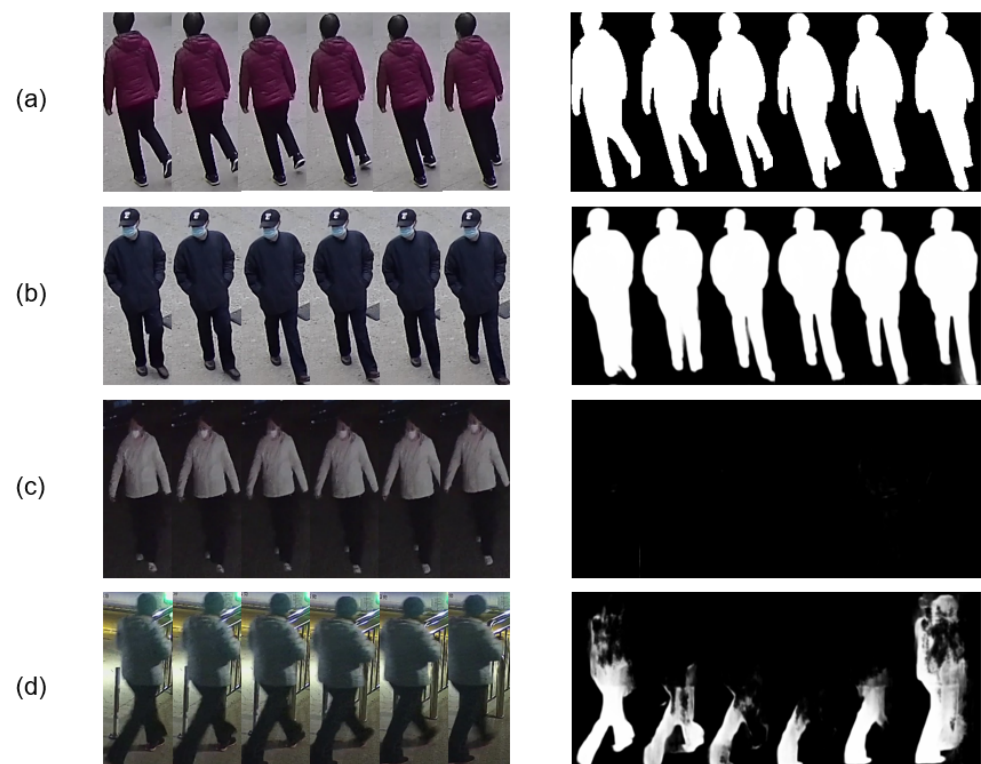


Figure 3. Collected daytime and nighttime data with corresponding human silhouettes. (a,b) show data collected during the daytime, where clear human silhouettes can be successfully extracted. (c,d) show data collected during the nighttime, where the silhouettes are either unidentifiable or incomplete.

In summary, the contributions of this paper are as follows:

- To address the limitations in existing research on nighttime gait recognition, we constructed the first outdoor gait dataset that encompasses both daytime and nighttime data, named Gait Recognition of Day and Night (GaitDN). This dataset provides data support for gait recognition research and applications, particularly in low-light nighttime scenarios.
- A novel gait recognition framework GaitSAT is introduced, which is particularly suited for processing gait recognition tasks in complex nighttime scenarios. GaitSAT not only enhances the model's adaptability to low-light outdoor scenes but also provides a flexible and generalizable framework for future research.
- Experimental results demonstrate that GaitSAT achieves state-of-the-art recognition accuracy and generalization performance among existing skeleton-based gait recognition methods in nighttime outdoor scenarios. These findings substantiate the effective-

ness of our proposed method in practical applications and showcase the potential of GaitSAT in addressing more complex recognition scenarios.

2. Related Works

2.1. Gait Recognition Datasets

Gait recognition datasets have evolved significantly, reflecting the field's progression from controlled laboratory scenarios to more complex, real-world scenarios. This evolution can be broadly categorized into three stages: early controlled environments, transition to outdoor scenarios, and recent efforts to capture real-world complexity.

In the initial stage, researchers focused on isolating specific factors affecting gait recognition. Datasets like the CASIA series [1,12] and OU-ISIR series [2,13,14] were instrumental in this phase. The CASIA-B dataset, for instance, investigates walking states and viewpoint variations, offering three distinct conditions (normal walking, carrying a bag, and wearing a coat) from 11 viewpoints. Similarly, the OU-ISIR Speed [13] dataset explores the impact of walking speed, while the OU-ISIR LP [2] Bag dataset addresses the challenge of recognizing individuals carrying objects. These datasets, collected in controlled indoor environments, allowed researchers to systematically study individual factors influencing gait recognition.

However, the controlled nature of these datasets created a significant gap between research and real-world applications. Recognizing this limitation, the field began transitioning toward more realistic scenarios. This shift is exemplified by datasets like GREW and Gait3D. The GREW dataset provides a large-scale benchmark for gait recognition in the wild, encompassing 26,345 pedestrians and 128,671 gait sequences, with the addition of a distractor set to enhance its applicability to real-world scenarios. The Gait3D dataset, collected in a large supermarket, comprises 4000 pedestrians and 25,309 gait sequences, offering various human representations including silhouettes, 2D/3D skeletons, 3D meshes, and 3D SMPL data. Recent efforts have focused on capturing even more nuanced aspects of real-world gait recognition. The CCPG [5] dataset, for example, addresses the challenge of clothing changes, offering refined conditions including overall fit changes, top and bottom changes, and carrying bags. This progression toward more complex and varied datasets reflects the field's growing emphasis on developing robust recognition algorithms capable of performing in diverse, real-world conditions.

Despite these advancements, a significant gap remains in the realm of nighttime gait recognition [15,16]. The challenges posed by low-light conditions in outdoor environments have not been adequately addressed by existing datasets. The CASIA Infrared [12] dataset, while pioneering in nighttime gait recognition, lacks the viewpoint variations and environmental complexities characteristic of real-world scenarios. This gap in available data highlights the need for a comprehensive dataset that captures the unique challenges of nighttime gait recognition in outdoor settings. Such a dataset would need to account for factors like varying lighting conditions, complex backgrounds, and the potential for irregular walking patterns exacerbated by low visibility. Our research aims to address this crucial need by developing a gait dataset that richly captures these nighttime variations.

2.2. Gait Recognition Methods

Based on the distinct features utilized by the models, gait recognition methods can be broadly categorized into two types: appearance-based methods and pose-based methods. Appearance-based methods typically utilize image sequences containing human silhouette information as input, primarily including binary black and white silhouettes obtained through background subtraction algorithms [17] and Gait Energy Images (GEIs) derived from averaging gait silhouettes over a complete cycle [9]. Due to the outstanding performance of deep learning [18] in computer vision tasks [19], and the proven effectiveness

of deep convolutional neural networks (CNNs) in extracting image features, CNNs dominate appearance-based methods. For Gait Energy Images, methods such as GEINet [9,10] construct models based on CNNs to extract effective gait information from GEIs, significantly outperforming previous approaches. Recent methods [3,20–24] are mostly based on human silhouettes, exploring the use of different CNNs or multi-scale structures to learn discriminative features directly from silhouette sequences. For instance, GaitSet [24] treats gait sequences as deep sets and uses CNNs to independently extract frame-level features from each silhouette. GaitPart [20] introduces focus convolution, enabling top convolutional kernels to attend to more local details within specific parts of input frames, enhancing the fine-grained learning of part-level spatial features, modeling of local micro-motion features, and global understanding of the entire gait sequence. CSTL [21] fuses multi-scale temporal features from local and global perspectives, treating cross-scale contextual information as guidance for temporal aggregation, thereby providing high-quality spatial cues. GaitGL [22] establishes global and local feature extractors and constructs a local temporal aggregation module to replace traditional spatial pooling layers, thereby preserving more spatial information. MTSGait [3] is a 2D convolution-based multi-hop temporal transform method, ensuring that the model can learn both spatial and multi-scale temporal information simultaneously while avoiding issues of excessive model parameters and training difficulties. GLN [23] utilizes the inherent feature pyramid in deep CNNs to enhance gait representation, fusing silhouette-level and set-level features extracted at different stages in a top-down manner with lateral connections to learn discriminative and compact representations from gait silhouettes. Although these methods have significantly improved performance on indoor datasets, they are susceptible to factors such as clothing, camera angle, walking speed variations, and lighting changes, resulting in poor robustness and transferability.

Pose-based methods utilize human skeleton sequences extracted from raw images as input, learning intrinsic correlations between body skeletons to extract gait features, often demonstrating stronger robustness. Posegait [25] uses 2D skeletons extracted from RGB images to estimate 3D pose information [26], combining a CNN with LSTM to design a dedicated gait feature extractor for 3D pose information. Given the success of graph convolutional neural networks in action recognition, graph convolution-related methods have gradually been introduced into gait recognition. GaitGraph [27] uses HR-Net [28] to extract 2D human skeleton points and applies a ResGCN module architecture network to learn gait features. GPGait [29] proposes a series of human-centric preprocessing operations to overcome environmental covariance influences, establish unified human pose representations, and achieve efficient graph partitioning and local–global feature relationship extraction. Furthermore, with the popularity of Transformers and their ability to capture long-range dependencies effectively, Ali et al. proposed a novel heterogeneous spatiotemporal axial mixer, GaitMixer [30], that can efficiently learn discriminative gait representations by capturing high-frequency and low-frequency features. Although these methods have shown improvements in robustness and generalization performance, experiments conducted on our newly collected GaitDN dataset, which includes nighttime data, indicate that their effectiveness in outdoor nighttime environments is not satisfactory. For such complex nighttime scenarios in outdoor low-light environments, there is still a lack of effective gait recognition methods.

3. GaitSAT Framework

In this section, we comprehensively illustrate the proposed method, GaitSAT. GaitSAT initially employs a self-attention module to extract spatial features, aiming to thoroughly learn the intrinsic relationships between various joints of the human body in each frame. Subsequently, to fully capture the walking patterns of pedestrians in outdoor environments along the temporal dimension, we input the extracted spatial vectors into a multi-scale temporal rhythm perception module. This module learns the rich periodic features contained within the gait sequence. In the following subsections, we will provide detailed descriptions of the aforementioned modules.

3.1. Spatial Feature Extraction

The human skeleton can be represented as a graph $G(V, E)$, where joints serve as vertices $V = \{v_i | i = 1, \dots, N\}$ containing N vertices, and the bones as edges E . The edges can be represented by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, where $\mathbf{A}_{i,j} = 1$ if joint i and j are physically connected; otherwise, $\mathbf{A}_{i,j} = 0$. A skeleton sequence is represented as a joint feature tensor $\mathbf{S} \in \mathbb{R}^{T \times N \times C}$, where T is the number of frames in the sequence, and C is the feature dimension.

For each individual skeleton frame, our objective is to effectively capture the intrinsic relationships between different joints. Considering the advantages of self-attention mechanisms in capturing global information, we adopt a Self-Attention-based Graph Convolutional module (SA-GC) [31] to process 2D skeleton features, as illustrated in Figure 4a.

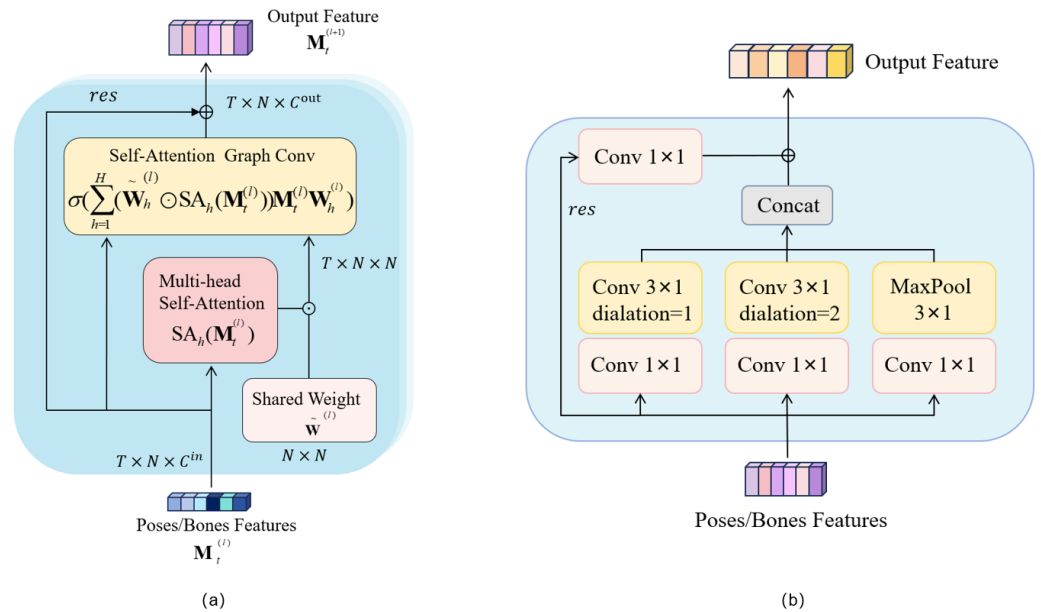


Figure 4. Detailed structure of (a) Self-Attention Graph Convolution (SA-GC) and (b) Multi-Scale Temporal Convolutional Network (MS-TCN).

The Self-Attention Graph Convolutional (SA-GC) module utilizes the self-attention of joint features to infer their intrinsic relationships. The SA-GC derives a positive, bounded weight, termed as the self-attention map, to represent the strength of relationships between all joints. Specifically, we employ learnable matrices $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{D \times D'}$ to linearly project the joint representation $\mathbf{S}_t$ into D' dimensional queries and keys to obtain the self-attention map, which can be represented as follows:

$$\text{SA}(\mathbf{S}_t) = \text{softmax}\left(\frac{\mathbf{S}_t \mathbf{W}_K (\mathbf{S}_t \mathbf{W}_Q)^T}{\sqrt{D'}}\right) \quad (1)$$

where t is the time index.

In addition to the self-attention map, we incorporate a learned weight $\tilde{\mathbf{W}}_h$, which is shared across time and instances. The shared weight have H heads, allowing the model to attend to different feature subspaces and capture the subtle connections between different joints.

For each head $1 \leq h \leq H$, the shared weights are combined with the self-attention map to obtain the intrinsic correlations:

$$\tilde{\mathbf{W}}_h \odot \text{SA}_h(\mathbf{S}_t) \in \mathbb{R}^{T \times N \times N} \quad (2)$$

where $\odot$ denotes the element-wise product. The overall update rule for joint representations can be expressed as

$$\mathbf{s}_t^{(l+1)} = \sigma \left(\sum_{h=1}^H (\tilde{\mathbf{W}}_h \odot \text{SA}_h(\mathbf{s}_t^{(l)})) \mathbf{s}_t^{(l)} \mathbf{w}_h^{(l)} \right) \quad (3)$$

where $\mathbf{W}^{(l)} \in \mathbb{R}^{D^{(l)} \times D^{(l+1)}}$ is the learnable weight of the l layer, and $\sigma(\bullet)$ represents the nonlinear activation function.

3.2. Multi-Scale Temporal Rhythm Perception Module

Human gait, as a process with periodic patterns, inherently encompasses regular changes across multiple temporal scales. These rhythmic variations, while generally consistent for an individual, can differ significantly among people due to variations in walking speed. The diversity in walking speeds leads to distinctive temporal patterns in gait cycles. In real-world scenarios, pedestrian walking speeds vary considerably, as illustrated in Figure 5, further emphasizing the need to account for diverse temporal patterns. Existing temporal convolution approaches with a fixed time-step [8,11], when processing these variations, often extract features only at specific time intervals, limiting the ability to accommodate the diverse walking speeds and patterns exhibited in real-world scenarios. Specifically, convolution kernels with fixed strides can only slide over a particular time stride, leading to the loss of critical temporal information and comprehensive gait patterns.

To address this issue, we propose a multi-scale temporal rhythm perception module, as illustrated in Figure 6. This module comprises three parallel branches designed to capture gait patterns at different temporal granularities: the fine temporal branch (temporal stride of 1), the middle-scale temporal branch (temporal stride of 2), and the coarse temporal branch (temporal stride of 3), which significantly outperforms fixed-stride models in several aspects: (1) The multi-scale structure enables comprehensive feature extraction across different temporal resolutions, preventing the loss of critical temporal information that often occurs with fixed-stride approaches. (2) The parallel processing of different temporal scales allows the model to adaptively capture both rapid, subtle movements and slower, more sustained gait patterns simultaneously. (3) This architecture inherently enhances the model's robustness to variations in walking speed and irregular gait patterns, as it can effectively process temporal information at multiple scales concurrently. (4) The combination of different temporal scales creates a more complete representation of gait dynamics, leading to more reliable and accurate recognition results in challenging real-world scenarios. In each branch, we introduce Self-Attention Temporal Fusion Units (SAT Units) as the basic modules. Each SAT Unit consists of an SA-GC block and a Multi-Scale Temporal Convolutional Network (MS-TCN) [32] block connected in series. The MS-TCN includes three convolution branches with different combinations of kernel sizes and dilation rates, each used to learn temporal features of varying granularity. Additionally, the MS-TCN employs 1×1 convolutions as residual modules, as shown in Figure 4b. We construct

each branch's structure by stacking multiple SAT Units with the same time stride, where the number of SAT Units can be dynamically adjusted based on the characteristics of different datasets. For instance, for the Gait3D dataset, we stack three SAT Units in each branch. Our approach possesses inherent dynamic adaptability, allowing for the flexible adjustment of the number of SAT Units according to task requirements. This dynamic structure is not only simple but it also enables the model to adaptively handle gait recognition tasks of varying complexity. While maintaining a consistent core architecture, it achieves the efficient processing of datasets of various scales and characteristics, significantly enhancing its flexibility and efficiency in handling diverse gait data.



Figure 5. Examples of pedestrians with varying walking speeds in GaitDN. (a–c) illustrate the process of a pedestrian taking a step, from the moment the left foot leaves the ground to the moment the right foot leaves the ground in the subsequent step. (a) The pedestrian takes one step in 12 frames. (b) The pedestrian takes one step in 16 frames. (c) The pedestrian takes one step in 18 frames.

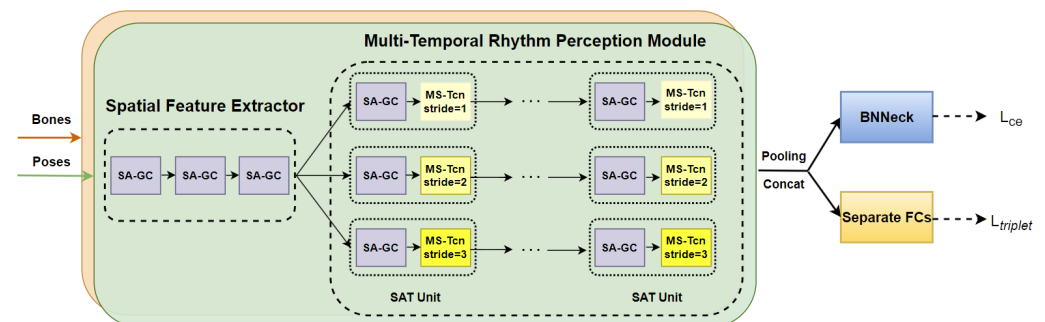


Figure 6. The framework of GaitSAT.

3.3. Network Structure

To capture the multifaceted features of human gait, our network integrates both 2D pose and bone data. This dual-input approach ensures a comprehensive representation of gait characteristics. The backbone network consists of two parallel streams, each dedicated to processing one of the two input types, allowing for specialized feature extraction from distinct gait representations. In each stream, we employ a series of SA-GC blocks to model the intrinsic relationships between joints and bones within individual frames of a gait sequence, effectively learning spatial features. However, the intrinsic connections within single skeleton frames cannot reflect the periodic patterns of gait. It is essential to capture the temporal dynamics of gait patterns. Consequently, we employ the proposed multi-scale temporal rhythm perception module to learn gait features at different time strides,

enriching the temporal dimension of gait features. Finally, we perform pooling operations on the features learned at different time strides to obtain a robust and comprehensive final embedding that encapsulates both spatial and temporal gait characteristics. The network architecture of GaitSAT is illustrated in Figure 6. Suppose the output of the multi-scale temporal rhythm perception module is denoted as $f_S \in \mathbb{R}^{T \times N \times C^{out}}$, where C^{out} represents the number of output channels, T represents the number of frames in the sequence, and N represents the number of key points. The spatial feature mapping can be represented as

$$f_{np}^v = \text{ap}_v(f_S) + \text{mp}_v(f_S) \quad (4)$$

where $\text{ap}(\bullet)$ refers to average pooling, $\text{mp}(\bullet)$ refers to max pooling along the N dimension, and v denotes the output of the v -th branch of the multi-scale temporal rhythm perception module. The average pooling operation captures the overall distribution of joint features, while max pooling identifies the most salient features across the joint dimension. By combining both average and max pooling along the joint dimension, we aim to create a comprehensive representation that captures both the general patterns and the distinctive features of an individual's gait across different joints and bones. Next, set pooling is employed to aggregate the temporal information:

$$f_{tp}^v = \text{sp}_v(f_{np}^v) \quad (5)$$

where $\text{sp}(\bullet)$ refers to max pooling along the T dimension. Finally, the outputs from both branches (including the joint branch and the bone branch) are concatenated together to form the final recognition representation:

$$\begin{aligned} F_{branch} &= \text{cat}(f_{tp}^1, \dots, f_{tp}^v) \\ F_{final} &= \text{cat}(F_{joint}, F_{bone}) \end{aligned} \quad (6)$$

This concatenation step integrates the complementary information learned from both joint and bone data, creating a comprehensive feature vector that encapsulates the full range of gait characteristics observed from different anatomical perspectives.

4. The GaitDN Dataset

To facilitate research on gait recognition in nighttime scenarios, which are prevalent in practical applications, we compiled a compact gait dataset named Gait Recognition of Day and Night (GaitDN). This dataset represents the first wild gait dataset that concurrently encompasses both daytime and nighttime data. The GaitDN dataset comprises 1009 subjects, over 3000 sequences, and more than 100,000 pedestrian images. Specifically, it includes 762 subjects collected during daytime and 247 subjects collected at night. GaitDN was collected in public areas where individuals typically carry various items and walk in irregular patterns and speeds. This characteristic not only incorporates variations in lighting conditions but also includes diverse interference factors, more closely approximating real-world gait recognition scenarios. The dataset thus provides a more authentic representation of the challenges encountered in practical applications. A comparison between GaitDN and other mainstream datasets is shown in Table 1.

Table 1. Comparison of publicly available datasets for gait recognition. Wild, Daytime Data, and Nighttime Data indicate whether the dataset was captured in the wild, has daytime data, and has nighttime data, respectively.

Dataset	Year	Subject	Seq	Cam	Data Type	Wild	Daytime Data	Nighttime Data
CASIA-A	2003	20	240	3	RGB, Silh.	×	×	×
CASIA Infrared	2006	153	1530	1	Infrared, Silh.	×	×	✓
CASIA-B	2006	124	13,640	11	RGB, Silh.	×	×	×
OU-ISIR Speed	2010	34	306	1	Silh.	×	×	×
OU-ISIR-LP	2012	4007	31,368	2	Silh.	×	×	×
OU-LP Bag	2018	62,528	187,584	1	Silh.	×	×	×
OU-MVLP	2018	10,307	288,596	14	Silh.	×	✓	×
OU-MVLP Pose	2020	10,307	288,596	14	2D Pose	×	✓	×
GREW	2021	26,345	128,671	882	Silh., 2D/3D Pose, Flow	✓	✓	×
Gait3D	2022	4000	25,309	39	Silh., 2D/3D Pose, 3D Mesh&SMPL	✓	✓	×
GaitDN	-	1009	3300	3	2D Pose	✓	✓	✓

The construction of the GaitDN dataset primarily involves four steps: video collection, pedestrian detection, pedestrian clustering, and feature extraction. To collect gait data from real-world outdoor scenarios, we utilized three cameras installed near public area entrances to gather over 100 h of video data with a resolution of 1920×1080 and a frame rate of 25 fps. For each frame of the original video, we employed a YOLOv3-based [33] pedestrian detection model to segment and extract pedestrian images, chosen for its speed and accuracy. This process yielded 3300 pedestrian gait sequences, each containing 25 to 60 frames. To cluster sequences of the same individual, we extracted pedestrian re-identification features using the open source FastReID [34] framework and applied unsupervised clustering methods. However, for nighttime data with reduced image quality, these methods proved inadequate for accurate identity classification. Consequently, we employed human annotators to assist in classifying this portion of the nighttime data. For the clustered pedestrian images, we used a pre-trained HRNet [28] model to estimate 2D skeleton data for each image. HRNet is currently the most widely adopted human pose extraction method, offering high accuracy and robust performance in challenging conditions.

To ensure that the collected data closely reflect real-world application scenarios, we intentionally did not control any variables other than clearly distinguishing between daytime and nighttime data, as we aimed to gather data that are sufficiently complex to encompass a wide range of variations. Consequently, GaitDN not only incorporates lighting variations but also includes multiple potential factors that may impact recognition accuracy, comprising cluttered backgrounds, environmental occlusions of pedestrians, individuals carrying bags, and pedestrians wearing heavy coats, as illustrated in Figure 2. Some extreme cases of pedestrian gait sequences are shown in Figure 7. For pedestrian a, the two gait sequences exhibit significant variations in lighting direction and color, with clothing colors blending closely with the background, which poses a challenge for feature extraction algorithms. For pedestrian b, the sequences show not only changes in lighting color and intensity but also significant differences in clarity and camera angles. For pedestrian c, the background includes intense lighting variations and moving vehicles, leading to dynamic background changes. These examples highlight the complexity of the data in GaitDN, effectively simulating the challenging scenarios of real-world nighttime environments. To protect privacy, the faces of subjects in the samples were pixelated. The diversity of influencing factors enables GaitDN to better simulate various complex situations that may be encountered in practical applications, providing researchers with a more comprehensive and authentic testing platform. This not only aids in enhancing model performance in complex environments but also provides a crucial data foundation for developing more robust and adaptive gait recognition algorithms.



Figure 7. Example of pedestrian sequence in extreme conditions (with key information obscured). For pedestrian (a), the two sequences demonstrate notable variations in lighting direction and color, where the clothing blends closely with the background, making feature extraction more challenging. For pedestrian (b), the sequences reveal changes in lighting color and intensity, as well as significant differences in clarity and camera angles. For pedestrian (c), the background includes strong lighting variations and dynamic changes caused by moving vehicles, posing additional challenges for accurate gait recognition.

5. Results and Discussion

We conducted experiments on indoor gait datasets (CASIA-B and OU-MVLP), outdoor gait datasets (GREW and Gait3D), and our proposed GaitDN dataset, which includes nighttime data. We compared several current state-of-the-art pose-based methods with our GaitSAT approach. The experimental results demonstrate that GaitSAT exhibits superior recognition accuracy and generalization performance in nighttime scenarios while maintaining high-level performance on daytime data.

5.1. Experimental Settings

GaitSAT, as a flexible framework, can be dynamically adjusted according to the scale of different datasets. In the spatial feature extraction module, to thoroughly extract spatial dimension features, we employ three cascaded SA-GC blocks, each with a different number of output channels, aiming to comprehensively learn the intrinsic connections and complex interactions between various body parts. The multi-scale temporal rhythm perception module consists of three branches, each composed of serially connected SAT Units. The number of SAT Units was dynamically adjusted based on the dataset. Specifically, for the smaller-scale CASIA-B dataset, the number of SAT Units was 1; for Gait3D, the number of SAT Units was 2; for OU-MVLP and GREW, the number of SAT Units was 3. The output channel details for each basic block are shown in Table 2.

During the training phase, to enhance the model's robustness and generalization capabilities, we implemented various data augmentation strategies. First, we added Gaussian noise to each keypoint with a probability of 0.3, simulating potential keypoint detection errors in real-world scenarios. Second, we applied an inverse operator to the entire skeleton with a probability of 0.01, resulting in a left-right flip of all skeleton points. Lastly, we randomly selected skeleton sequences with a fixed length of 60 to ensure input data consistency. For optimization, we employed the Adam optimizer in conjunction with the OneCycle learning rate scheduling strategy. The initial, maximum, and final learning

rates were set to 1×10^{-5} , 1×10^{-3} , and 1×10^{-8} , respectively. This configuration allows the model to converge rapidly in the early stages of training and fine-tune parameters in the later stages, thereby achieving optimal performance. We adjusted the batch size and number of training epochs according to the scale of each dataset, as shown in Table 2.

Table 2. Training parameters and network parameters for different datasets.

Dataset	Batchsize	Iterations	Spatial Feature Extractor	Multi-Scale Temporal Rhythm Perception Module	
			Output Channels	Number of SAT Units	Output Channels
GaitDN	(32, 4)	30 k	(64, 64, 128)	2	(128, 256)
CASIA-B	(4, 32)	40 k	(64, 64, 128)	1	(256)
OUMVLP	(32, 16)	150 k	(64, 128, 128)	3	(128, 256, 256)
GREW	(32, 8)	200 k	(64, 128, 128)	3	(128, 256, 256)
Gait3D	(32, 4)	60 k	(64, 64, 128)	2	(128, 256)

5.2. Ablation Study

5.2.1. Analysis of Input Feature Types

We compared the performance differences between models trained using joint point data or skeleton data alone, and models trained using both features simultaneously. The experimental results are shown in Table 3. Using either joint point or skeleton features independently failed to achieve optimal results. However, when combining both joint point and skeleton features, GaitSAT achieved the best performance across multiple datasets. Notably, on GaitDN, compared to using joint point and skeleton features separately, the combined use of both features increased the recognition accuracy by 14.11% and 5.29%, respectively. This result indicates that integrating joint point and skeleton features captures richer gait information, thereby enhancing the model recognition capability. The superior performance of integrating joint and skeleton features emerges from their intrinsically complementary characteristics. Specifically, joint features excel at capturing local motion patterns and precise positioning information, while skeleton features provide comprehensive global structural insights and body configuration. When strategically integrated, these features synergistically construct a more robust and nuanced representation, demonstrating exceptional adaptability, particularly in challenging environmental conditions such as low-light or nighttime scenarios, where individual feature types could independently exhibit limitations. Furthermore, we observed that compared to the CASIA-B indoor dataset, GaitSAT achieved more significant performance improvements when using both features in the outdoor datasets Gait3D and GaitDN. This finding emphasizes the importance and necessity of considering multiple features when addressing complex gait recognition tasks in outdoor scenarios. This not only validates the potential of our method in practical applications but also provides important insights for future gait recognition research in more complex and variable environments.

Table 3. Analysis of input feature types, in which we keep the network architecture consistent.

Setting		CASIA-B	Gait3D	GaitDN
Joints	Bones			
✓		73.56	19.30	72.13
	✓	72.44	17.80	80.95
✓	✓	76.32	22.90	86.24

5.2.2. Analysis of the Spatial Feature Extraction Module

To validate the effectiveness of the SA-GC module in spatial feature extraction, we conducted comparative experiments with several advanced spatial feature extraction modules [27,29,35,36] while maintaining the overall structure and parameters of the network. The experimental results are shown in Table 4. Compared with other modules, our SA-GC module demonstrated superior performance, which confirms the efficacy of our chosen module. The improved performance can be attributed to the SA-GC module's unique ability to capture the intrinsic topological structure of joints in the skeleton, going beyond mere physical connections. By employing self-attention mechanisms [37], the module infers complex, potentially asymmetric relationships between joints, reflecting their actual interactions in gait patterns. This approach allows the SA-GC to adaptively learn and represent rich spatial features of the human pose, leading to more discriminative gait representations.

Table 4. Analysis of the spatial feature extraction module, in which we keep the network architecture consistent.

Block	CASIA-B	Gait3D	GaitDN
ResGCN [27]	63.74	15.33	65.83
ST-GCN [35]	73.60	19.89	74.33
Shift-GCN [36]	71.26	18.44	74.49
PAGCN [29]	74.10	22.56	84.77
SA-GC	76.32	22.90	86.24

5.2.3. Analysis of Multi-Scale Temporal Rhythm Perception Module

To verify the effectiveness of the proposed multi-scale temporal rhythm perception module, we designed ablation experiments for validation. The comparative network structure remained unchanged, but each branch adopted the same temporal stride. The experimental results are shown in Table 5. When we employed the multi-scale temporal rhythm perception module, setting the temporal strides of the three branches to 1, 2, and 3, respectively, optimal results were achieved across all three datasets. In contrast, when all branches used the same temporal stride, recognition accuracy declined to varying degrees.

This experimental result further validates the efficacy of the multi-scale temporal rhythm perception module. In setting different temporal strides, the model can better capture gait feature variations across different time scales, thereby enhancing the overall recognition performance. When all branches use the same temporal stride, the model's temporal perception ability is limited, leading to decreased recognition accuracy. This indicates that the multi-scale temporal rhythm perception module can effectively improve the model's ability to capture gait information, demonstrating significant advantages in processing complex temporal data.

Furthermore, to validate the applicability of our proposed method to other temporal aggregation approaches, we substituted the MS-TCN with two different temporal feature extraction modules [8,11]. The experimental results show that our MS-TCN module achieves optimal performance under various temporal stride settings, further confirming the effectiveness of our chosen module.

Table 5. Analysis of different temporal aggregation methods using proposed multi-scale temporal rhythm perception module.

Block	Setting	CASIA-B	Gait3D	GaitDN
	Stride			
TCN [8]	All stride = 1	58.32	13.65	63.67
	All stride = 2	56.33	13.98	62.69
	All stride = 3	55.97	12.11	58.98
	stride = 1, 2, 3	59.86	14.92	54.10
Temporal Bottleneck Block [11]	All stride = 1	69.20	17.53	77.38
	All stride = 2	67.56	17.04	76.52
	All stride = 3	63.05	15.54	74.81
	stride = 1, 2, 3	71.97	18.67	78.96
MS-TCN	All stride = 1	73.74	22.60	85.19
	All stride = 2	67.56	17.04	76.52
	All stride = 3	72.41	20.90	73.77
	stride = 1, 2, 3	76.32	22.90	86.24

5.3. Performance Comparison with State-of-the-Art Methods

Our method uses human poses as input and belongs to the pose-based approach category. Therefore, we selected three mainstream pose-based gait recognition methods for comparison with our proposed method. GaitGraph, designed based on gait data from ideal environments, does not perform well on outdoor gait datasets. While this method achieves higher accuracy on indoor datasets compared to other pose-based methods, its performance on outdoor datasets does not reach the same level as on indoor datasets. GP-Gait demonstrates robust generalization capabilities and performs excellently on outdoor datasets. The selected methods are briefly introduced as follows:

1. GaitGraph is a representative pose-based method that first introduced 2D human poses into gait recognition. It employs graph convolutional networks [38] for spatiotemporal modeling and uses supervised contrastive loss for training.
2. GaitTR pioneers the incorporation of spatial Transformers in gait recognition, combining them with temporal convolutional networks to extract gait features from skeletons.
3. GPGait proposes a series of human-oriented methods for preprocessing human poses, aiming to obtain unified and rich representations. It also utilizes mask design to achieve effective graph partitioning and extraction of local–global feature relationships.

The experimental results of existing pose-based methods and our GaitSAT on various datasets are presented in Table 6. GaitSAT demonstrated outstanding performance on outdoor datasets, achieving Rank-1 accuracies of 22.90% and 57.37% on the outdoor datasets Gait3D and GREW, respectively, representing the current state-of-the-art performance. For our proposed outdoor dataset GaitDN, which includes nighttime data, GaitSAT achieved a Rank-1 accuracy of 86.24%, also the best performance to date. These results indicate that GaitSAT is not only suitable for outdoor scenarios but also performs well in low-light nighttime conditions, maintaining excellent performance when facing complex recognition tasks. For indoor datasets, GaitSAT's performance on the smaller-scale CASIA-B was not superior to existing methods but still demonstrated stable performance. Moreover, on OU-MVLP, currently the largest indoor dataset, GaitSAT achieved state-of-the-art performance with a

Rank-1 accuracy of 60.77%. GaitSAT’s excellent performance across multiple large-scale datasets proves its superiority in processing large-scale data. These results demonstrate that GaitSAT possesses outstanding adaptability in various environments, particularly excelling in outdoor and complex scenarios, showing broad potential for practical applications.

Through a comprehensive analysis of computational resources, we thoroughly evaluated the scalability of GaitSAT. The experimental results in Table 7 show that while GaitSAT has more parameters (2.961 M) compared to GaitGraph (0.320 M) and GaitTR (0.512 M), this increase is justified by its significant performance improvements. Specifically, GaitSAT demonstrates substantial performance advantages on the OUMVLP dataset, achieving 56.53% and 21% improvements in Rank-1 accuracy compared to GaitGraph and GaitTR, respectively. Notably, when compared to the current state-of-the-art GPGait (3.655 M), GaitSAT achieves superior performance with fewer parameters and inference time (0.0273 vs. 0.0428 sequences/second*GPU). These experimental results demonstrate that, despite its multimodule and cascade structure, GaitSAT’s excellent computational efficiency and resource utilization make it feasible for large-scale practical deployment.

Table 6. Comparison of Rank-1 accuracy (%) of GaitSAT and recent state-of-the-art pose-based gait recognition methods on GaitDN and four popular datasets.

Method	CASIA-B				OUMVLP	Gait3D	GREW	GaitDN
	NM	BG	CL	Mean				
GaitGraph	86.37	76.5	65.24	76.04	4.24	8.60	10.18	69.38
GaitTR	94.72	89.29	86.65	90.22	39.77	7.20	48.58	60.66
GPGait	93.60	80.15	69.29	81.01	59.11	22.40	57.04	85.71
GaitSAT	90.37	73.84	64.04	76.32	60.77	22.90	57.37	86.24

Table 7. Comparison with SOTA methods on CASIA-B datasets in parameter number (million) and inference time (second/(sequences* GPU)).

Method	Param.	Inference Time
GaitGraph	0.320	0.0029
GaitTR	0.512	0.0051
GPGait	3.655	0.0428
GaitSAT	2.961	0.0273

5.4. Analysis of Generalization Performance

To further investigate nighttime gait recognition and explore the model’s generalization performance, we conducted more extensive cross-domain experiments. These experiments involved training on a source dataset and subsequently testing on a target dataset. We selected current mainstream public datasets as source datasets for training and tested separately on daytime data (GaitDN_Day) and nighttime data (GaitDN_Night) from the GaitDN dataset. This approach was chosen to assess the model’s adaptability to complex scenarios, particularly nighttime scenarios, that were not included in the training process. The experimental results are presented in Table 8:

1. When the target dataset is the daytime data from GaitDN, GaitSAT achieved the current best performance when using GREW and Gait3D as source datasets. The Rank-1 accuracies for GREW→GaitDN_Day (where “→” indicates the cross-domain direction from source to target dataset) and Gait3D → GaitDN_Day were 88.89% and

86.78%, respectively. However, when CASIA-B, OUMVLP, and GaitDN were used as source datasets, the cross-domain performance was slightly inferior to GPGait.

- When the target dataset is the nighttime data from GaitDN, GaitSAT demonstrated the strongest cross-domain performance across all source datasets. Notably, the Rank-1 accuracies for GREW → GaitDN_Night and Gait3D → GaitDN_Night reached 71.67% and 70.03%, respectively. These results comprehensively showcase GaitSAT's excellent transferability on nighttime data. Regardless of the source dataset, the trained model maintains outstanding performance on nighttime data, indicating GaitSAT's robust domain adaptation capability, particularly suitable for nighttime gait recognition.

Table 8. Cross-Domain performance comparison between GaitSAT and state-of-the-art methods on daytime and nighttime data from GaitDN. The source datasets are mainstream datasets, and the target datasets are the daytime and nighttime data from GaitDN, respectively.

Source Dataset	Method	Target DataSet	
		GaitDN_Day	GaitDN_Night
CASIA-B	GaitGraph	21.56	6.52
	GaitTR	21.16	18.33
	GPGait	81.48	61.67
	GaitSAT	78.84	61.96
OUMVLP	GaitGraph	40.52	26.08
	GaitTR	55.55	40.10
	GPGait	80.95	56.37
	GaitSAT	77.78	56.67
GREW	GaitGraph	84.97	69.56
	GaitTR	64.55	53.33
	GPGait	87.83	71.56
	GaitSAT	88.89	71.67
Gait3D	GaitGraph	65.39	43.48
	GaitTR	40.74	28.33
	GPGait	85.56	69.89
	GaitSAT	86.78	70.00
GaitDN	GaitGraph	72.19	65.36
	GaitTR	62.93	57.44
	GPGait	87.83	66.07
	GaitSAT	86.17	68.01

6. Conclusions

In this work, to explore more effective gait recognition in the wild, we collected GaitDN, the first gait dataset explicitly distinguishing between daytime and nighttime data, and proposed GaitSAT, a method more adaptable to outdoor scenarios. GaitSAT employs attention-based graph convolution to model the intrinsic connections between human skeleton joints and uses a multi-scale temporal rhythm perception module to aggregate temporal features, ultimately obtaining rich human gait representations. GaitSAT not only outperforms baseline models on public datasets but also further validates its excellent cross-domain performance and robustness in cross-domain tasks. Additionally,

ablation experiments demonstrate the crucial role of each GaitSAT component in enhancing model performance.

Future work could explore several potential directions. Firstly, improving the quality of original images is crucial for gait recognition in complex scenarios. Besides upgrading camera equipment, effective data preprocessing methods could be explored to enhance image clarity by adjusting parameters such as contrast and brightness while minimizing interference with original human information, thus providing higher-quality input for silhouette-based methods. Secondly, using multi-modal approaches [39–42] to simultaneously learn gait representations from silhouette and skeleton information is a promising research direction. How to effectively combine silhouette and skeleton information merits further exploration by researchers. Lastly, we are committed to building a comprehensive gait recognition platform. In future work, we plan to introduce additional factors that may influence gait recognition, such as terrain and weather conditions. Our ultimate goal is to develop gait recognition models with greater robustness and scalability, thereby advancing the practical applications of gait recognition.

Furthermore, the relevant data in this paper contain privacy information of the subjects, and we will make every effort to protect their privacy. Firstly, all collected data are for research purposes only, and any original videos or RGB images related to the subjects will not be made public. The samples shown in this paper had their faces pixelated. Secondly, processed data, such as 2D skeleton data, will only be released after strict approval from the relevant management level, and researchers will need to sign a rigorous application form before using the data for academic research.

Author Contributions: Conceptualization, H.L.; methodology, H.L.; software, H.L.; validation, H.L.; formal analysis, W.G.; data curation, Y.W., Y.L. and K.L.; writing—original draft preparation, H.L.; writing—review and editing, W.G.; visualization, W.G.; supervision, J.G.; project administration, J.G.; funding acquisition, J.G. All authors have read and agreed to the published version of the manuscript.

Funding: The present study was partly supported by the Natural Science Foundation of Shandong Province under Grant ZR2023MF041, the National Natural Science Foundation of China 532 under Grant (62072469), and Shandong Data Open Innovative Application Laboratory.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used for the proposed method are the publicly available CASIA-B dataset (<http://www.cbsr.ia.ac.cn/china/Gait%20Databases%20CH.asp>, accessed on 20 November 2024), OU-MVLP dataset (<http://www.am.sanken.osaka-u.ac.jp/BiometricDB/GaitLPPose.html>, accessed on 20 November 2024), Gait3D dataset (<https://gait3d.github.io>, accessed on 20 November 2024), and GREW dataset (<https://github.com/GREW-Benchmark/GREW-Benchmark>, accessed on 20 November 2024).

Acknowledgments: Wenjuan Gong acknowledges the support of the Natural Science Foundation of Shandong Province under Grant ZR2023MF041, the National Natural Science Foundation of China under Grant (62072469), and Shandong Data Open Innovative Application Laboratory. Jordi González acknowledges the support of the Spanish Ministry of Economy and Competitiveness (MINECO) and the European Regional Development Fund (ERDF) under Project No. PID2020-120611RBI00/AEI/10.13039/501100011033.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SA-GC Self-Attention Graph Convolution;
MS-TCN Multi-Scale Temporal Convolutional Network.

References

1. Yu, S.; Tan, D.; Tan, T. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, 20–24 August 2006; Volume 4, pp. 441–444.
2. Takemura, N.; Makihara, Y.; Muramatsu, D.; Echigo, T.; Yagi, Y. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Trans. Comput. Vis. Appl.* **2018**, *10*, 4. [\[CrossRef\]](#)
3. Zheng, J.; Liu, X.; Liu, W.; He, L.; Yan, C.; Mei, T. Gait recognition in the wild with dense 3d representations and a benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20228–20237.
4. Zhu, Z.; Guo, X.; Yang, T.; Huang, J.; Deng, J.; Huang, G.; Du, D.; Lu, J.; Zhou, J. Gait recognition in the wild: A benchmark. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 14789–14799.
5. Li, W.; Hou, S.; Zhang, C.; Cao, C.; Liu, X.; Huang, Y.; Zhao, Y. An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 13824–13833.
6. Sepas-Moghaddam, A.; Etemad, A. Deep gait recognition: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 264–284. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Kukreja, V.; Kumar, D.; Kaur, A. Deep learning in Human Gait Recognition: An Overview. In Proceedings of the 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 4–5 March 2021; pp. 9–13.
8. Lea, C.; Flynn, M.D.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal convolutional networks for action segmentation and detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 156–165.
9. Shiraga, K.; Makihara, Y.; Muramatsu, D.; Echigo, T.; Yagi, Y. Geinet: View-invariant gait recognition using a convolutional neural network. In Proceedings of the 2016 International Conference on Biometrics (ICB), Halmstad, Sweden, 13–16 June 2016; pp. 1–8.
10. Wu, Z.; Huang, Y.; Wang, L.; Wang, X.; Tan, T. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 209–226. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Song, Y.F.; Zhang, Z.; Shan, C.; Wang, L. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1625–1633.
12. Tan, D.; Huang, K.; Yu, S.; Tan, T. Efficient night gait recognition based on template matching. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; Volume 3, pp. 1000–1003.
13. Iwama, H.; Okumura, M.; Makihara, Y.; Yagi, Y. The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 1511–1521. [\[CrossRef\]](#)
14. An, W.; Yu, S.; Makihara, Y.; Wu, X.; Xu, C.; Yu, Y.; Liao, R.; Yagi, Y. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE Trans. Biom. Behav. Identity Sci.* **2020**, *2*, 421–430. [\[CrossRef\]](#)
15. Tan, D.; Huang, K.; Yu, S.; Tan, T. Recognizing night walkers based on one pseudoshape representation of gait. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
16. DeCann, B.; Ross, A. Gait curves for human recognition, backpack detection, and silhouette correction in a nighttime environment. In Proceedings of the Biometric Technology for Human Identification VII, Orlando, FL, USA, 5–6 April 2010; Volume 7667, pp. 248–260.
17. Liu, Z.; Malave, L.; Sarkar, S. Studies on silhouette quality and gait recognition. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, Washington, DC, USA, 27 June–2 July 2004; Volume 2, p. II.

18. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 2872–2893. [[CrossRef](#)] [[PubMed](#)]
19. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [[CrossRef](#)] [[PubMed](#)]
20. Fan, C.; Peng, Y.; Cao, C.; Liu, X.; Hou, S.; Chi, J.; Huang, Y.; Li, Q.; He, Z. Gaitpart: Temporal part-based model for gait recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14225–14233.
21. Huang, X.; Zhu, D.; Wang, H.; Wang, X.; Yang, B.; He, B.; Liu, W.; Feng, B. Context-sensitive temporal feature learning for gait recognition. In Proceedings of the IEEE/CVF international conference on computer vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12909–12918.
22. Lin, B.; Zhang, S.; Wang, M.; Li, L.; Yu, X. Gaitgl: Learning discriminative global-local feature representations for gait recognition. *arXiv* **2022**, arXiv:2208.01380.
23. Hou, S.; Cao, C.; Liu, X.; Huang, Y. Gait lateral network: Learning discriminative and compact representations for gait recognition. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 382–398.
24. Chao, H.; He, Y.; Zhang, J.; Feng, J. Gaitset: Regarding gait as a set for cross-view gait recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8126–8133.
25. Liao, R.; Yu, S.; An, W.; Huang, Y. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognit.* **2020**, *98*, 107069. [[CrossRef](#)]
26. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
27. Teepe, T.; Khan, A.; Gilg, J.; Herzog, F.; Hörmann, S.; Rigoll, G. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 2314–2318.
28. Seong, S.; Choi, J. Semantic segmentation of urban buildings using a high-resolution network (HRNet) with channel and spatial attention gates. *Remote Sens.* **2021**, *13*, 3087. [[CrossRef](#)]
29. Fu, Y.; Meng, S.; Hou, S.; Hu, X.; Huang, Y. Gpgait: Generalized pose-based gait recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 19595–19604.
30. Pinyoanuntapong, E.; Ali, A.; Wang, P.; Lee, M.; Chen, C. Gaitmixer: Skeleton-based gait representation learning via wide-spectrum multi-axial mixer. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–9 June 2023; pp. 1–5.
31. Chi, H.g.; Ha, M.H.; Chi, S.; Lee, S.W.; Huang, Q.; Ramani, K. Infogcn: Representation learning for human skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20186–20196.
32. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and unifying graph convolutions for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 143–152.
33. Farhadi, A.; Redmon, J. Yolov3: An incremental improvement. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; Springer: Berlin/Heidelberg, Germany, 2018; Volume 1804, pp. 1–6.
34. He, L.; Liao, X.; Liu, W.; Liu, X.; Cheng, P.; Mei, T. Fastreid: A pytorch toolbox for general instance re-identification. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 9664–9667.
35. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
36. Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; Lu, H. Skeleton-based action recognition with shift graph convolutional network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 183–192.
37. Vaswani, A. Attention is all you need. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2017.
38. Zhang, S.; Tong, H.; Xu, J.; Maciejewski, R. Graph convolutional networks: A comprehensive review. *Comput. Soc. Netw.* **2019**, *6*, 1–23. [[CrossRef](#)] [[PubMed](#)]
39. Zhu, H.; Zheng, Z.; Nevatia, R. Gait recognition using 3-d human body shape inference. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 909–918.
40. Zou, S.; Xiong, J.; Fan, C.; Shen, C.; Yu, S.; Tang, J. A multi-stage adaptive feature fusion neural network for multimodal gait recognition. *IEEE Trans. Biom. Behav. Identity Sci.* **2024**, *6*, 539–549. [[CrossRef](#)]

41. Marín-Jiménez, M.J.; Castro, F.M.; Delgado-Escañó, R.; Kalogiton, V.; Guil, N. UGaitNet: Multimodal gait recognition with missing input modalities. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 5452–5462. [[CrossRef](#)]
42. Li, G.; Guo, L.; Zhang, R.; Qian, J.; Gao, S. Transgait: Multimodal-based gait recognition with set transformer. *Appl. Intell.* **2023**, *53*, 1535–1547. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.