GaitCycFormer: Leveraging Gait Cycles and Transformers for Gait Emotion Recognition

Qingyang Zeng[1, 2], **Lin Shang**[1, 2]

[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]Department of Computer Science and Technology, Nanjing University
qy_zeng@smail.nju.edu.cn, shanglin@nju.edu.cn

Abstract

Gait Emotion Recognition (GER) is an emerging task within Human Emotion Recognition. Skeleton-based GER requires discriminative spatial and temporal features. However, current methods primarily focus on capturing spatial topology information but fail to effectively learn temporal features from long-distance frames. Moreover, these methods are mostly sensitive to the order of sampled sequences, resulting in significant accuracy drops when sequences are randomly sampled. In order to obtain a more robust and comprehensive spatial-temporal representation of gait, we introduce the Graph-Transformer architecture into GER for the first time, proposing a novel framework named GaitCycFormer. Specifically, we designed a Cycle Position Encoding (CPE) based on the gait cycle, which explicitly segments any gait sequence into more manageable periodic units, to enhance temporal feature modeling. Additionally, we incorporate a bi-level Transformer, consisting of an Intra-cycle Transformer and an Inter-cycle Transformer to capture local and global temporal information within each gait cycle and between gait cycles respectively. Experiments demonstrate that our GaitCycFormer achieves state-of-the-art performance on popular datasets, and proves to be more reliable and robust.

Introduction

Human emotion recognition is a crucial aspect of affective computing, influencing various fields such as human-computer interaction (Narayanan et al. 2023), video surveillance (Arunnehru and Kalaiselvi Geetha 2017) and healthcare (Lu et al. 2022). Gait analysis has emerged as a promising alternative for emotion recognition (Li et al. 2018; Randhavane et al. 2019; Sun, Su, and Fan 2022). which does not require close interaction or active cooperation from the subjects. Previous studies have demonstrated that distinct gait behaviors associated with different emotional states (Kleinsmith and Bianchi-Berthouze 2013). Recent advancements in human pose estimation and depth sensor technologies (Martinez et al. 2017; Song et al. 2021) have further facilitated the use of 3D skeleton data, enabling more robust and accurate gait analysis for emotion recognition.

This paper focuses on developing and evaluating methods for Gait Emotion Recognition (GER) using 3D skeleton data. Given the natural structure of skeleton data as graphs in non-Euclidean geometric spaces, graph-based methods have gained significant attention recently (Lima et al. 2024; Yin et al. 2024; Zhang et al. 2024; Chen and Sun 2023; Zhuang et al. 2020). In these studies, the initial gait sequence length is typically set to 240 frames according to the setting of the EGait dataset (Bhattacharya et al. 2020b), utilizing the Temporal Convolution Network (TCN) module to learn temporal features. Some studies downsample the sequence interval to 48 frames (Yin et al. 2024) or 75 frames (Narayanan et al. 2020). However, these single sampling and convolution strategies often cause models to rely on a specific sampling order, hindering their ability to learn deep temporal features and limiting their generalization capacity, especially in random sampling settings or real-world environments where the sequence order may vary.

To address these limitations, recent methods (Chen and Sun 2023; Zhang et al. 2024) have advanced temporal feature extraction. STA-GCN (Chen and Sun 2023) employed a multi-scale temporal convolution module, which was widely used in skeleton-based action recognition (Chen et al. 2021). However, this direct module migration did not fully consider the inherent temporal characteristics of gait. As illustrated in Fig. 1, gait is cyclic and symmetrical. Each gait cycle reflects a local continuous motion process, and frames at the same position in different gait cycles reflect similar postures, although walking directions may differ due to changes in the

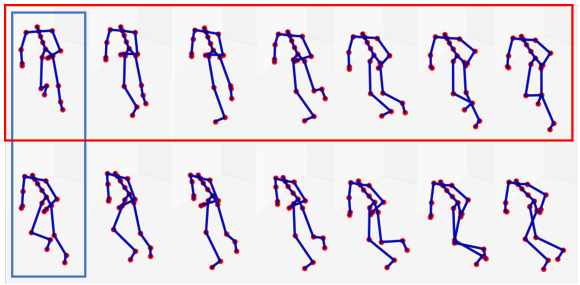

Figure 1: Gait Cycle: The interval between adjacency frames where the same foot touches the ground. The example in the figure is a part of the gait sequence of downsampled 48-frame sequence from subject '1684' of EGait. Gait frames in each row represents a gait cycle and gait frames in each colomn mean the same position in different gait cycles.

photographer's position.

The gait cycle serves as a unique temporal feature for gait analysis. It is usually employed as a crucial handcrafted feature (Randhavane et al. 2019; Bhattacharya et al. 2020a). We have observed that while gait cycles associated with different emotions exhibit subtle variations, they hold potential for modeling and simplifying long-distance time series. TT-GCN (Zhang et al. 2024) incorporated the gait cycle using causal-TCN and dilated convolution, with an expanded receptive field to enhance correlations between gait cycles. However, it did not thoroughly explore the gait cycle or explicitly address how to effectively capture and utilize temporal information within and between gait cycles.

Transformers demonstrate superior suitability compared to Temporal Convolutional Networks (TCNs) when examining the continuity of sequential frames within a gait cycle and the consistency of frames exhibiting similar positional attributes across cycles. The positional encodings inherent in Transformers facilitate the establishment of temporal continuity, particularly when aligning cycle and positional encodings, thereby enhancing attention mechanisms within a given cycle. By leveraging a comprehensive attention matrix, the set of frames at identical positions across cycles can effectively capture subtle variations in gait, especially in response to changes in direction and angle. These capabilities empower Transformers to dynamically adapt to nuanced changes and patterns both within and between gait cycles, a task that poses significant challenges for TCNs.

In this paper, we introduce a novel framework for Gait Emotion Recognition, termed as **Gait Cyc**le Trans**former** (**GaitCycFormer**), which is focused on the gait cycle to effectively capture local and global temporal information within and between gait cycles. We begin by employing a 3-layer Spatial-Temporal Graph Convolutional Network (ST-GCN) as the backbone to extract foundational spatial-temporal features. To enhance these features, we propose Graph Temporal Transformer (GTT) Blocks, which integrate a Graph-Transformer architecture. This design preserves essential spatial characteristics through the GCN while simultaneously leveraging the advanced temporal modeling capabilities of the Transformer's multi-head attention mechanism.

A key innovation is the incorporation of Cycle Position Encoding (CPE), which is derived from carefully extracted gait cycles across various sampled sequences. We optimize this extraction method to ensure consistent representation of gait cycles, accommodating different sampling sequences. Notably, our improved approach guarantees that when sequences are downsampled, the calculated gait cycles exhibit uniform reduction patterns. This consistency is crucial for accurately capturing the cyclical nature of gait, as it allows for reliable encoding of each frame's position within the gait cycle. By enriching the Transformer's temporal modeling capabilities, our CPE facilitates a more profound understanding of gait dynamics.

To further refine our approach, we utilize two distinct Transformer modules: the Intra-cycle Transformer and the Inter-cycle Transformer. The Intra-cycle Transformer effectively captures local temporal information within individual gait cycles, allowing for detailed analysis of joint movements and their temporal dependencies. Conversely, the Inter-cycle Transformer addresses the global temporal information across multiple gait cycles, facilitating the recognition of patterns that extend beyond single cycles.

By combining these modules, we achieve a robust understanding of both local and global temporal features, significantly enhancing performance in GER. Finally, we concatenate shallow and deep features to further improve overall model efficiency, demonstrating the strengths of our framework in addressing the complexities of gait analysis.

The main contributions of the proposed method are summarized as follows:

- We propose a novel and more accurate method for calculating gait cycles. This method improves the segmentation and sampling of gait sequences, and further leads to design CPE that models the fine-grained temporal features within a gait cycle and the global attention between cycles.
- We introduce a Graph-Transformer hybrid architecture that leverages the Transformer's self-attention mechanism for temporal feature fusion. The Graph-Transformer framework is firstly used to obtain both spatial and temporal dependencies for GER taking the advantage of graph-based methods and Transformers.
- GaitCycFormer achieves state-of-the-art performance on public datasets. Additionally, our model proves to be adaptable to handle variations in sampling sequences, making it more reliable and robust.

Related Work

Gait Emotion Recognition

GER research focuses on three key areas: feature design, deep learning models, and learning paradigms. Initially, GER relied on traditional machine learning with handcrafted features (Crenn et al. 2016). Although recent approaches have shifted towards deep learning to derive gait representations, some studies still integrate hand-crafted and deep learning features to improve accuracy, as seen in (Bhattacharya et al. 2020a) with ST-GCN features combined with manually calculated 29-dimensional emotion features. Hand-crafted features, however, are limited due to complex modeling and lack of universality across different data samples. whereas the gait cycle offers a universal and lightweight temporal feature applicable to various skeleton data.

Deep learning models are primarily sequence-based, image-based or graph-based approaches (Hu et al. 2022). Graph-based methods have emerged as the most popular due to their ability to effectively capture complex spatial-temporal relationships. They utilize spatial-temporal graph convolutional networks to model joint interactions. (Bhattacharya et al. 2020a) applied ST-GCN to enhance the spatial and temporal joint relationships. Similarly, (Zhuang et al. 2020) adopted a globally connected graph approach for spatial feature extraction, while (Chen and Sun 2023) emphasized adaptive graph aggregation to capture both implicit and explicit connections. The increasing popularity of

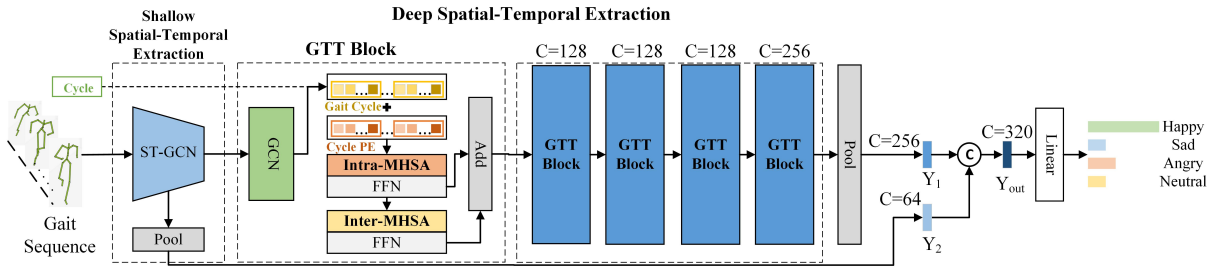

Figure 2: The pipeline of our GaitCycFormer. The ST-GCN extracts the shallow features from the initial 3D skeleton data. The GTT Block, consisting of a GCN, and a bi-level Transformer, generates basic spatial features and deep temporal features respectively. The Intra-cycle Transformer and Inter-cycle Transformer in the bi-level Transformer captures local and global temporal information at both fine and coarse levels.

graph-based methods highlights their effectiveness in representing the intricate dynamics of gait data.

Lastly, learning paradigms in GER can be categorized into fully supervised and self-supervised. Self-supervised learning has emerged as a new paradigm for GER and we will discuss their performance in the experiment session.

Graph Transformer for 3D Skeletons

Compared to GCN-based methods, Transformer can rapidly capture global spatial and temporal information. When Pure Transformers are introduced into 3D skeleton data, GCNs are used as supplements to address the limitations of Transformers, resulting in what is called a Graph Transformer.

Two main approaches exist: Transformer-style GCNs, which construct adjacency matrices via self-attention, and GCNs integrated into Transformers to bolster local feature extraction. Both utilize Positional Embedding and Multi-Head Attention, constructing fully connected matrices for spatial topology and temporal dimensions.

We focus on hybrid architectures that combine GCN and Transformer for spatial-temporal feature extraction, emphasizing the Transformer's temporal modeling capability. (Plizzari, Cannici, and Matteucci 2021) introduced the ST-TR network with a Temporal Self-Attention module (TSA) for long-term dependencies in skeleton activity recognition. (Li et al. 2020) developed TE-GCN to capture correlations between non-adjacent temporal distances using temporal self-attention. However, fully connected attention matrices often result in high computational complexity. To mitigate this, (Do and Kim 2024) proposed a skeletal-temporal Transformer that partitions frames by temporal relationships, reducing computational demands. Meanwhile, our proposed method also improve the model efficiency through selectively retaining temporal attention.

Method

Overview of Pipeline

The overall architecture of our proposed GaitCycFormer is depicted in Fig. 2. A skeleton sequence, corresponding to a single emotion, is sampled to maintain a consistent frame count of T, resulting in $\mathbf{X} \in \mathbb{R}^{T\times V\times C}$ where V represents the number of joints per frame, C is the dimensionality of data representing a single joint ($C = 3$ for 3D skeletons). The basic feature maps denoted as $\mathbf{S}_F = \{F_i \in \mathbb{R}^{T\times V\times C_1} | i = 1, 2..., T\}$, are extracted using the ST-GCN backbone (Yan, Xiong, and Lin 2018). These extracted feature maps are then fed into two branches.

The first branch performs pooling on the feature maps to obtains the shallow spatial-temporal feature representation Y_2. The second branch, the temporal Transformer stream, consists of five GTT Blocks, each comprising a GCN layer, an Intra-cycle Transformer and an Inter-cycle Transformer. The feature obtained from the temporal Transformer stream undergo pooling to produce the deep spatial-temporal feature represent Y_1. Finally, we concatenate Y_1 and Y_2 to form the final feature representation Y_{out} for the gait sequence. This representation is passed through a linear layer to produce the outcome $\hat{y} \in \mathbb{R}^{N_c}$ where N_c represents the number of emotion classes. We utilize the standard cross-entropy loss to train the proposed network.

Spatial-Temporal Backbone

We utilitze the ST-GCN backbone to extract fundamental shallow spatial-temporal features. Given the spatial-temporal data defined above, spatial-temporal graph convolution is applied across multiple layers using a predefined graph structure. The approach adopted by ST-GCN is similar to the method (Kipf and Welling 2017), where the skeleton graph is represented by the adjacency matrix A and an identity matrix I. The convolutional operation can be described in a vectorized form as Eq. (1):

$$f_{out} = \sum_{v}^{K_v} A_k \odot M_k \left(f_{in} W_k\right) \tag{1}$$

Here, K_v represents the kernel size of the spatial dimension, set to 3 in accordance with the spatial configuration partitioning strategy.

The adjacency matrix $A_k = \hat{D}^{-\frac{1}{2}}A\hat{D}^{-\frac{1}{2}}$ (with $A_k \in \mathbb{R}^{V\times V}$) is defined based on the partitioning strategy explained, where $\hat{D}^{ii} = \sum_j A^{ij} + I^{ij}$. $W \in \mathbb{R}^{C_{out}\times C_{in}\times 1\times 1}$ is learnable for the 1×1 convolution. $M \in \mathbb{R}^{V\times V}$ is also learnable and indicates the importance of each vertex. The symbol $\odot$ denotes the element-wise product between two matrices. For the temporal convolution, a $K \times 1$ convolutional layer, with K set to 9, is applied to learn representational features from the adjacent frames.

Gait Cycle Analysis

Motivation. As a sequence-based recognition task, the temporal relationships between gait frames contain unique biological information that is critical for accurate recognition. In the study of the temporal dimension of gait sequences, the gait cycle servers as a natural and fundamental time unit, enabling the decomposition of long sequences into manageable segments. The gait cycle is defined as the period that begins when one foot touches the ground and ends when the same foot touches the ground again. Typically, a normal gait cycle spans approximately 20-30 frames (Li and Zhao 2022). However, the gait cycles calculated by existing methods (Randhavane et al. 2019; Bhattacharya et al. 2020a; Zhang et al. 2024) have proven to be insufficiently accurate, resulting in cycles shorter than 15 frames. Therefore, we propose a novel method based on window to calculate the gait cycle.

Window-based Calculation. Gait cycles were calculated by detecting the frame interval between the key frames where the foot joint on the same side touches the ground in previous works (Randhavane et al. 2019; Bhattacharya et al. 2020a; Zhang et al. 2024). If the vertical coordinate of the foot joint remains unchanged across 3 consecutive frames or if the vertical axis coordinate of the middle frame is the lowest, it is determined to be a key frame of the gait cycle. The interval between two key frames defines as a gait cycle, and the average of all cycles in a sequence provides the average gait cycle. However, downsampling a sequence reduces the frame count in each gait cycle, ideally proportionate to the sequence length reduction. Unfortunately, we observed that the above method results in inaccurate gait cycle calculations. Specifically, the gait cycles calculated using the initial sequence were shorter than expected, and the rate of decrease in the gait cycle length during downsampling was also inaccurate.

In this work, we propose a window-based method for calculating gait cycles, where the window size adjusts dynamically based on the total gait sequence length and the number of sampled frames. By calculating the ratio coefficient between the total number of frames in the gait sequence T_t (with T_t = 240 in our case) and the sampled frames T_s, we determine a window size that adapts to the length of the current sequence length, ensuring the window contains at least three frames. The window size is denoted as Eq. (2):

$$W = max(\frac{T_s}{T_t} \times a, 3) \tag{2}$$

where a is a hyperparameter set according to the frame rate (with a = 15 in our case). As we traversal the gait sequence, we identify key frames by finding the lowest vertical coordinate of the foot joint within each window. To optimize the detection and avoid misdetection when foot joints share the same coordinates, we set a minimum interval threshold of 3 to ensure adequate spacing between detection frames. Our method adapts to gait sequences of varying lengths and reduces misdetection, thereby improving the accuracy of gait cycle calculations.

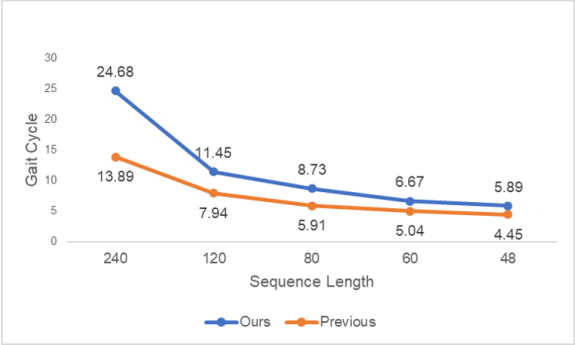

Figure 3: The gait cycles corresponding to different sampling sequences calculated by our method and previous method (Bhattacharya et al. 2020a). As the sequences are downsampled, our gait cycles follow an ideal downward trend.

The gait cycles corresponding to different sampling sequences calculated by our method and previous method (Bhattacharya et al. 2020a) are shown in the Fig. 3.

Cycle Position Encoding

Positional encoding assigns each position a unique embedding. To enhance the modeling of temporal information for gait sequences in the Transformer layer, we introduce Cycle Position Encoding (CPE), which aligns with the accurately determined gait cycle length.

The CPE is denoted as $P_S = \{p_i \mid i = 1, ...S\}$, where S is the cycle size of the position encoding, corresponding to the gait cycle length of the current input sequence. We repeat position encoding embeddings to match the length of the input sequence, ensuring the positional information is consistent throughout the sequence. This process simulates the cyclic nature of the gait cycle, with the cycle size being calculated based on the number of sampled sequence frames, ensuring interpretability. Moreover, the continuity of different position encodings within the same cycle, along with the repeatability of the same position encoding across different cycles, provide extensive learnable information for the subsequent attention mechanism. Sine and cosine functions with varying frequencies are employed as the encoding functions in Eq. (3) and Eq. (4):

$$PE(p, 2i) = sin(\frac{p \ mod \ L}{L} \cdot 10000^{-\frac{2i}{C_{in}}}) \tag{3}$$

$$PE(p, 2i+1) = cos(\frac{p \ mod \ L}{L} \cdot 10000^{-\frac{2i}{C_{in}}}) \tag{4}$$

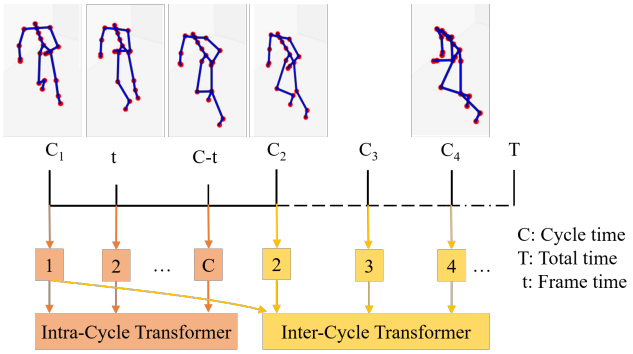

Figure 4: Frames of interest for Intra-cycle Transformer and Inter-cycle Transformer.

where p and i denote the position of the joint and the dimension of the position encoding vector, and L represents the length of the gait cycle.

GTT Block

Motivation. Tradition Transformers utilize a global attention mechanism to capture information between any pair of frames, which can lead to excessive attention to long-distance frames and introduce irrelevant global information. This phenomenon, termed the 'over-globalizing problem' by (Xing et al. 2024), can be problematic. Inspired by (Xing et al. 2024) and considering the periodicity of gait, we propose a bi-level Transformer module consisting of an Intra-cycle Transformer layer and an Inter-cycle Transformer layer to effectively decouple the temporal information, as shown in Fig. 4. This design ensures that local information is captured within each gait cycle, followed by the calculation of global attention between different cycles, thereby alleviating the over-globalizing problem.

Intra-cycle Transformer. After dividing the gait sequence into periodic units, local information is typically contained within each gait cycle, so we utilize an Intra-cycle Transformer. We start by dividing the gait sequence into gait cycles, denoted as $T_{n,s} = \{t_{i,j} \mid i = 1, ...n, j = 1...s\}$, where n and s represent the number of gait cycles and the length of gait cycle respectively. The node features in gait cycle T_n are represented as $\mathbf{X}_n \in \mathbb{R}^{(N\times V)\times C_{in}\times s}$, where N represents the batch size. Notably, the joint dimension is incorporated within the batch to process each joint along the time axis separately. Concurrently, we generate the CPE, denoted as $P_s = \{p_i \mid i = 1, ...s\}$ that matches the gait cycle length. The CPE is reshaped to facilitate element-wise addition with $\mathbf{X}_n$ as the sequence embedding $\mathbf{H}_n$ and fed into the Intra-cycle Transformer and learn the updated representations $\hat{\mathbf{H}}_n$, as illustrated in Eq. (5):

$$\begin{aligned} \hat{\mathbf{H}}_n &= FFN\left(Softmax\left(\frac{Q_n K_n^T}{\sqrt{s}}\right)V_n\right) \\ Q_n &= \mathbf{H}_n W_Q, K_n = \mathbf{H}_n W_K, V_n = \mathbf{H}_n W_V \end{aligned} \tag{5}$$

where W_Q, W_K and $W_V \in \mathbb{R}^{s\times s}$ are learnable weights, and FFN represents a Feed-Forward Neural Network.

Inter-cycle Transformer. Subsequently, we reshape the representations $\hat{\mathbf{H}}$ from $(N * V * n, C, 1, s)$ to $(N * V * s, C, 1, n)$, which means the time dimension of interest changes from a sequence of frames with different position encodings within a cycle to a collection of frames in all cycles with the same position coding. Following this, the reshaped $\hat{\mathbf{H}}_s$ are fed into the Inter-cycle Transformer, as illustrated in Eq. (6):

$$\begin{aligned} \hat{\mathbf{H}}_s^{'} &= FFN\left(Softmax\left(\frac{Q_s^{'} K_s^{'T}}{\sqrt{n}}\right)V_s^{'}\right) \\ Q_s^{'} &= \hat{\mathbf{H}}_s W_{Q'}, K_s^{'} = \hat{\mathbf{H}}_s W_{K'}, V_s^{'} = \hat{\mathbf{H}}_s W_{V'} \end{aligned} \tag{6}$$

where $W_{Q'}, W_{K'}$ and $W_{V'} \in \mathbb{R}^{n\times n}$ are learnable weights.

The sequence of gait frames we focus on consist of frames taken from the same position within each gait cycle, all sharing identical position encoding. These frames exhibit the most similarity in terms of walking mode. However, as illustrated in Fig. 1, the angle between the monocular camera and the subject can cause the captured gait to frequently change direction. Consequently, although the gaits at the same position in different cycles are similar in posture, their directions vary. By associating these gait sequences across different directions, the model can intuitively learn the states of the gaits in various orientations for the same emotional state.

Moreover, the inherent periodicity of the input sequences allows the model to comprehend the complete cycle information of the gait sequence. This cyclical nature ensures that the model captures global periodic attention, enabling a more comprehensive understanding of gait emotions. The periodic input not only highlights the subtle directional changes but also provides the model with a holistic view of the entire gait sequence, encompassing all cycles and offering a thorough grasp of the temporal structure and emotional cues embedded within the gait.

Finally, the output attention $\mathbf{A} \in \mathbb{R}^{T\times V\times C}$ is obtained by summing $\hat{\mathbf{H}}_n$ and $\hat{\mathbf{H}}^{'}s$, followed by a residual connection(He et al. 2016).

To summarize, the GTT Block consists of a GCN layer and a bi-level Transformer module, as illustrated in Fig. 2. This robust framework enables the extraction of both spatial and temporal features effectively. Retaining the GCN layer in each GTT Block allows for the exploration of complex temporal features while preserving the essential spatial features. The Intra-cycle Transformer focuses on learning detailed temporal dependencies within each gait cycle, ensuring that fine-grained features are captured. Concurrently, the Inter-cycle Transformer identifies the relationships between different gait cycles, thus providing a comprehensive understanding of the temporal structure of the gait sequence. Furthermore, compared to the traditional Transformer, which employs a fully connected relationship matrix, our configuration significantly reduces computational complexity and enhances model efficiency since we focus on Intra-cycle attention and Inter-cycle attention.

Dataset	Paradigm	Method		Accuracy
EGait	Self-supervised	CNN	CAGE(Lu, Hu, and Hu 2023)	79.6%
		Graph	SSAL(Song et al. 2024)	81.2%
	Supervised	CNN	Proxemo(Narayanan et al. 2020)	82.4%
		Transformer	TNTC(Hu et al. 2022)	83.2%
		Graph	STEP(Bhattacharya et al. 2020a)	78.2%
			TT-GCN(Zhang et al. 2024)	80.1%
			G-GCSN(Zhuang et al. 2020)	81.5%
			MSA-GCN(Yin et al. 2024)	83.5%
			STA-GCN(Chen and Sun 2023)	85.8%
		Graph-Transformer	Ours	**86.3%**
EWalk	Supervised	Graph	STEP(Bhattacharya et al. 2020a)	76.5%
		Graph	STA-GCN(Chen and Sun 2023)	82.8%
		Graph-Transformer	Ours	**83.4%**

Table 1: Comparison with state-of-the-art methods on public datasets.

Experiment

Datasets

EGait. The EGait dataset contains 1,835 gait samples spanning for four emotions: happy, sad, angry, neutral. Each sample is meticulously annotated by 10 domain experts. All the samples are represented as skeleton data, each containing $T = 240$ frames. For further details about EGait dataset, please refer to (Bhattacharya et al. 2020b). In our experiments, we maintain a fixed 8:1:1 split for training, validating and testing sets. This split is consistently applied across all experiments to ensure reliable and reproducible results.

EWalk. The EWalk dataset, collected by (Bhattacharya et al. 2020a), includes 342 gait samples from 90 participants. Participants were instructed to imagine different emotions (happy, sad, angry, neutral) while walking. Since the gait sequences in this dataset vary in length, so we standardize each sequence to 240 frames by looping them.

Implementation Details

Hyper-parameters. (**1**) The shape of the skeleton sequence fed into the model is $T \times V \times C$. $T = 240$ is the number of frames in the sequence, $V = 16$ is the number of joints of the skeleton, and $C = 3$ is the dimension of the 3D coordinates of the joints. (**2**) Shallow feature extractions from the 3-layer ST-GCN has an output channel of 64. The output channels of the five GTT Blocks in the temporal Transformer stream are set to 128, 128, 128, 256 and 256. The combined output channel of the shallow feature and deep feature is 320. (**3**) In the TCN layer, the kernel size K is set to 9. In the Transformer layer, the head number of each self-attention block is set to 8.

Training details. All experiments are conducted using the Pytorch framework, running on one NVIDIA 1080 Ti GPU and two NVIDIA 2080 Ti GPUs. SGD is used as the optimizer with a weight decay of 1×10^{-3} and an initial learning rate of 1×10^{-1}. The model is trained for 100 epochs with a batch size of 16, and validation is performed every 10 epochs. During testing, the entire sequence is input into the network for gait representations. We take the top-1 accuracy as evaluation criteria.

Comparison with State-of-the-Art Methods

We compare our method with state-of-the-art methods recently reported on the EGait and EWalk datasets, which are based on different paradigms and model architectures. Most of the works report their accuracy on the EGait dataset, as it is a relatively larger and more standardized dataset. It is worth noting that self-supervised learning has emerged in the latest methods, which adopts a contrastive learning paradigm to learn emotional representations from unlabeled gait sequences (Lu, Hu, and Hu 2023; Song et al. 2024). Despite its potential, self-supervised learning still lags behind fully supervised methods in terms of performance. As shown in Tab. 1, our proposed GaitCycFormer outperfoms all prior state-of-the-art methods on both datasets, achieving the hightest performance metrics.

Abalation Study

In this section, we evaluate the effectivess of different components of our method. Several abalation studies with various settings are conducted on EGait.

Effectiveness of GTT Block. To validate the effectiveness of the proposed GTT blocks, we establish a baseline by removing the GTT blocks, relying solely on short-term feature representations for recognition. We then conducted experiments to compare the performance of the proposed Transformer with various settings. Tab. 2 shows the comparison. (**1**) It is evident that all experiments utilizting Transformer to extract deep spatial-temporal features outperform the base-

Method	Accuracy
Baseline	79.8%
Baseline + GTT w/o Inter-Tr	83.6%
Baseline + GTT W/o Intra-Tr	83.2%
Baseline + GTT (Our GaitCycFormer)	86.3%

Table 2: Effectiveness of GTT Block. **Intra-Tr**: Intra-cyle Transformer. **Inter-Tr**: Inter-cycle Transformer.

Method	w/o PE	w/ PE	w/ CPE
Vanilla Transformer	82.6%	83.2%	84.6%
Intra-Inter Transformer	84.5%	85.3%	86.3%

Table 3: Importance of Position Encoding. **Vanilla**: Vanilla Transformer(Vaswani et al. 2017). **Intra-Inter**: Our Intra-cycle Transformer and Inter-cycle Transformer. **w/o PE**: remove the Position Encoding in Transformer. **w/ PE**: Position Encoding by vanilla Position Encoding(Vaswani et al. 2017). **w/ CPE**: Position Encoding by our proposed Cycle Position Encoding.

line, demonstrating its advantage in capturing complex temporal dependencies in gain sequences. (**2**) Removing either the Intra-cycle Transformer or the Inter-cycle Transformer from the model results in decreased performance, highlighting the crucial role both components play in effective temporal information extraction. The Intra-cycle Transformer captures local temporal dependencies within a single gait cycle, while the Inter-cycle Transformer captures global dependencies across multiple gait cycles. Both are essential for robust gait emotion recognition.

Importance of Position Encoding. In Tab. 3, we show the effectiveness of our position encoding in the Transformer module. We use both vanilla Transformer and our Intra-Inter Transformer as architectures, considering that vanilla Transformer with vanilla Position Encoding (PE) may benefits in the whole gait sequence. (**1**) When no position encoding is performed on the input sequences, the accuracy decreased by **2.0%** in Vanilla Transformer and **1.8%** in Intra-Inter Transformer, respectively. (**2**) When using the vanilla PE, the accuracy reduced by **1.4%** in Vanilla Transformer and **1.0%** in Intra-Inter Transformer, respectively. (**3**) When using the CPE, the accuracy in vanilla Transformer is **1.7%** less than Intra-Inter Transformer. Considering that the fully attention connection in vanilla Transformer takes too many training parameters may lead to poor model performance because of overfitting, the Intra-Inter Transformer focused on the gait cycle capture more effective local and global temporal features. The result also demonstrate that our proposed CPE model the temporal information of gait sequences more effectively.

Frames	Accuracy	
	TCN	Ours
240	79.8%	86.3%
120	78.8%	84.6%
120 (r)	74.5%	84.8%
48	78.2%	83.7%
48 (r)	71.8%	83.6%

Table 4: Robustness of Sampled Sequence. r represents the random sampling setting, in other words, the sampling order intervals of the sequence are not strictly the same.

Robustness of Sampled Sequence

We compare the performance of TCNs and our proposed temporal module under different sampling settings, as shown in Tab. 4. When the number of sampled sequences decreases, the recognition accuracy of TCNs does not significantly decline, indicating that the TCN module does not effectively enhance recognition across long-distance temporal dimensions of the gait sequences. However, under random sampling settings, the recognition accuracy of the TCN modules decreases **4.3%** in 120 frames and **6.4%** in 48 frames respectively, demonstrating that the TCN's performance depends heavily on the sequence order of sampling and lacks generalization capacity.

In contrast, under the random sampling setting, the accuracy of our model is less than **0.1%** compared to normal sampling in 48 frames, and even exceeds the accuracy of the original normal sampling by **0.2%** in 120 frames. This robustness is due to our method of calculating cycles, which is not significantly affected by random sampling. Thus, the Cycle Position Encoding and the Intra-Inter Transformer components continue to function effectively, ensuring stable performance. This highlights the superior ability of our model to handle variations in sampling sequences, making it more reliable for gait emotion recognition tasks. By maintaining high accuracy regardless of sampling strategy, our model proves to be adaptable and resilient, offering consistent performance in diverse scenarios.

Conclusion

In this paper, we propose a novel Graph Transformer framework for gait emotion recognition task, named GaitCycFormer. The framework is specifically designed to focus on the gait cycle, enabling it to effectively capture both local and global spatial-temporal information. We propose a window-based method for accurately calculating the gait cycle, which adapts to various lengths of sampled gait sequences. Additionally, We present Cycle Position Encoding, along with a bi-level Transformer, aiming at modeling temporal information both within and between gait cycles. Experiments demonstrate that GaitCycFormer significantly outperforms existing methods, showcasing superior accuracy and robustness, and it can serves as a valuable reference for other skeleton-based gait tasks.

References

Arunnehru, J.; and Kalaiselvi Geetha, M. 2017. Automatic human emotion recognition in surveillance video. *Intelligent techniques in signal processing for multimedia security*, 321–342.

Bhattacharya, U.; Mittal, T.; Chandra, R.; Randhavane, T.; Bera, A.; and Manocha, D. 2020a. STEP: Spatial Temporal Graph Convolutional Networks for Emotion Perception from Gaits. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 1342–1350. AAAI Press.

Bhattacharya, U.; Roncal, C.; Mittal, T.; Chandra, R.; Kapsaskis, K.; Gray, K.; Bera, A.; and Manocha, D. 2020b. Take an Emotion Walk: Perceiving Emotions from Gaits Using Hierarchical Attention Pooling and Affective Mapping. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X*, volume 12355 of *Lecture Notes in Computer Science*, 145–163. Springer.

Chen, C.; and Sun, X. 2023. STA-GCN:Spatial Temporal Adaptive Graph Convolutional Network for Gait Emotion Recognition. In *IEEE International Conference on Multimedia and Expo, ICME 2023, Brisbane, Australia, July 10-14, 2023*, 1385–1390. IEEE.

Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; and Hu, W. 2021. Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 13339–13348. IEEE.

Crenn, A.; Khan, R. A.; Meyer, A.; and Bouakaz, S. 2016. Body expression recognition from animated 3D skeleton. In *International Conference on 3D Imaging, IC3D 2016, Liège, Belgium, December 13-14, 2016*, 1–7. IEEE.

Do, J.; and Kim, M. 2024. SkateFormer: Skeletal-Temporal Transformer for Human Action Recognition. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLI*, volume 15099 of *Lecture Notes in Computer Science*, 401–420. Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society.

Hu, C.; Sheng, W.; Dong, B.; and Li, X. 2022. TNTC: Two-Stream Network with Transformer-Based Complementarity for Gait-Based Emotion Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, 3229–3233. IEEE.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Kleinsmith, A.; and Bianchi-Berthouze, N. 2013. Affective Body Expression Perception and Recognition: A Survey. *IEEE Transactions on Affective Computing*, 15–33.

Li, B.; Zhu, C.; Li, S.; and Zhu, T. 2018. Identifying Emotions from Non-Contact Gaits Information Based on Microsoft Kinects. *IEEE Transactions on Affective Computing*, 585–591.

Li, J.; Xie, X.; Zhao, Z.; Cao, Y.; Pan, Q.; and Shi, G. 2020. Temporal Graph Modeling for Skeleton-based Action Recognition. arXiv:2012.08804.

Li, N.; and Zhao, X. 2022. A strong and robust skeleton-based gait recognition method with gait periodicity priors. *IEEE Transactions on Multimedia*, 25: 3046–3058.

Lima, M. L.; de Lima Costa, W.; Martínez, E. T.; and Teichrieb, V. 2024. ST-Gait++: Leveraging spatio-temporal convolutions for gait-based emotion recognition on videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, 302–310. IEEE.

Lu, H.; Hu, X.; and Hu, B. 2023. See Your Emotion from Gait Using Unlabeled Skeleton Data. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 1826–1834. AAAI Press.

Lu, H.; Xu, S.; Hu, X.; Ngai, E.; Guo, Y.; Wang, W.; and Hu, B. 2022. Postgraduate Student Depression Assessment by Multimedia Gait Analysis. *IEEE MultiMedia*, 56–65.

Martinez, J.; Hossain, R.; Romero, J.; and Little, J. J. 2017. A Simple Yet Effective Baseline for 3d Human Pose Estimation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2659–2668. IEEE Computer Society.

Narayanan, V.; Manoghar, B. M.; Dorbala, V. S.; Manocha, D.; and Bera, A. 2020. ProxEmo: Gait-based Emotion Learning and Multi-view Proxemic Fusion for Socially-Aware Robot Navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2020, Las Vegas, NV, USA, October 24, 2020 - January 24, 2021*, 8200–8207. IEEE.

Narayanan, V.; Manoghar, B. M.; RV, R. P.; and Bera, A. 2023. EWareNet: Emotion-Aware Pedestrian Intent Prediction and Adaptive Spatial Profile Fusion for Social Robot Navigation. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, 7569–7575. IEEE.

Plizzari, C.; Cannici, M.; and Matteucci, M. 2021. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208: 103219.

Randhavane, T.; Bhattacharya, U.; Kapsaskis, K.; Gray, K.; Bera, A.; and Manocha, D. 2019. Identifying emotions from walking using affective and deep features. arxiv:1906.11884.

Song, C.; Lu, L.; Ke, Z.; Gao, L.; and Ding, S. 2024. Self-supervised Gait-based Emotion Representation Learning from Selective Strongly Augmented Skeleton Sequences. arxiv:2405.04900.

Song, L.; Yu, G.; Yuan, J.; and Liu, Z. 2021. Human pose estimation and its application to action recognition: A survey. *Journal of Visual Communication and Image Representation*, 76: 103055.

Sun, X.; Su, K.; and Fan, C. 2022. VFL—A deep learning-based framework for classifying walking gaits into emotions. *Neurocomputing*, 473: 1–13.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xing, Y.; Wang, X.; Li, Y.; Huang, H.; and Shi, C. 2024. Less is More: on the Over-Globalizing Problem in Graph Transformers. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In McIlraith, S. A.; and Weinberger, K. Q., eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 7444–7452. AAAI Press.

Yin, Y.; Jing, L.; Huang, F.; Yang, G.; and Wang, Z. 2024. MSA-GCN: Multiscale Adaptive Graph Convolution Network for gait emotion recognition. *Pattern Recognition*, 147: 110117.

Zhang, T.; Chen, Y.; Li, S.; Hu, X.; and Chen, C. L. P. 2024. TT-GCN: Temporal-Tightly Graph Convolutional Network for Emotion Recognition From Gaits. *IEEE Transactions on Computational Social Systems*, 11(3): 4300–4314.

Zhuang, Y.; Lin, L.; Tong, R.; Liu, J.; Iwamoto, Y.; and Chen, Y. 2020. G-GCSN: Global Graph Convolution Shrinkage Network for Emotion Perception from Gait. In Sato, I.; and Han, B., eds., *Computer Vision - ACCV 2020 Workshops - 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - December 4, 2020, Revised Selected Papers*, volume 12628 of *Lecture Notes in Computer Science*, 46–57. Springer.