FSGait: Fine-Grained Self-Supervised Gait Abnormality Detection

Bingzhi Duan, Xiaoyue Wan, and Xu Zhao*

Shanghai Jiao Tong University, Shanghai, China
{duanbingzhi,sherrywaan,zhaoxu}@sjtu.edu.cn

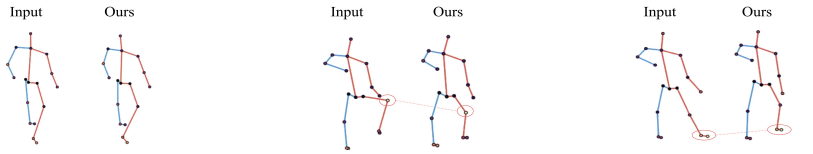
Abstract. *Gait Abnormality Detection* (GAD) plays an important role in diagnosing diseases associated with abnormal gait patterns. However, existing works are limited in generalization capability and granularity, which indicates they detect only the types of abnormalities seen during training and sequence-level gait anomalies. The reason is the restricted variety of abnormalities in datasets and the reliance on supervised learning algorithms. Therefore, we propose a *Fine-grained Self-supervised Gait Abnormality Detection* method (FSGait). We divide gait abnormality into two sub-problems: postural anomaly and temporal anomaly, which are solved by two designed modules, *Gait Reconstruction Module* (GRM) and *Gait Prediction Module* (GPM). The two modules are trained in self-supervised way on normal gait data. In this way, they capture normal gait patterns to distinguish abnormalities, thereby enhancing the generalization capability. For fine-grained detection, three-level (*Sequence, Frame and Joint*) abnormal detections are achieved with the intermediate results of these two modules. FSGait has a high degree of granularity and holds significant potential for aiding medical diagnosis and automating disease detection. Experiments on two datasets show that FSGait achieves state-of-the-art performance in frame-level GAD, while maintaining high sequence-level GAD accuracy. The joint-level detection results are presented with visualization.

Keywords: Gait Abnormality Detection · Fine-Grained · Generalization Capability · Reconstruction · Prediction

1 Introduction

Gait has a multitude of applications, such as gait recognition [5,17,28,30,36] and Gait Abnormality Detection (GAD) [13, 15, 20, 25, 31]. GAD focuses on identifying abnormalities within a gait sequence. Presently, GAD is widely utilized in medical field, as numerous diseases exhibit characteristic pathological gait patterns. For example, gait abnormalities in Parkinson [7, 9, 16] are characterized by disturbances in gait posture, as well as *Freezing Of Gait* (FOG).

GAD methods are categorized into wearable device-based approaches [20, 31] and vision-based approaches [9, 13, 15, 25]. Compared with wearable device-based methods, vision-based methods have higher flexibility in detection, lower detection cost, and a wider range of application scenarios.



(a) Normal Gait **(b)** Steppage Gait **(c)** Stiff-legged Gait

Fig. 1: Difference between normal and abnormal gait detection. Left pose denotes the input, right pose denotes our reconstructed (predicted) pose. Parts in red circle are the abnormal parts detected by comparing the left and the right.

The majority of existing vision-based studies employ supervised learning algorithms. However, two challenges persist. Firstly, the datasets encompass a limited variety of abnormal gait sequences. This leads to a lack of generalization capability, which means that these studies fail in accurately detecting unseen anomalies. Secondly, there currently exists few abnormal gait dataset labeled at the frame level precisely. Consequently, most existing methods are confined to sequence-level. As far as we know, few work [22] manage to achieve frame-level detection, and the accuracy tends to be low. This leads to a lack of granularity.

To be practically applicable in the medical field, such as detecting freezing gait in Parkinson's disease patients and pinpointing the affected body parts, a more refined approach is essential. It requires the capability to identify abnormalities within a gait sequence, determine the specific abnormal frame, and localize the abnormal joints accurately.

Therefore, we propose a fine-grained and self-supervised GAD method. The self-supervised training technique significantly improves generalization ability, while the reconstruction and prediction mechanisms enhance granularity. For abnormal gait, we believe that two different cases exist: postural anomaly and temporal anomaly. Postural anomaly refers to the abnormal posture in gait that is different from the normal walking posture, while temporal anomaly refers to the gait abnormality that is mainly reflected in temporal, such as FOG. We posit that human recognition of abnormal motion stems from comparisons with large amounts of normal data in daily life. For the two abnormal types, we simulate human's ability, designing two self-supervised module, Gait Reconstruction Module (GRM) with memory bank and Gait Prediction Module (GPM) based on Transformer, to learn normal gait patterns. Based on this, our approach treats abnormality as opposite to normality, and assess whether a frame is abnormal by comparing it with its closest normal gait frame which stems from GRM and GPM, as shown in Fig. 1. It enables the identification of abnormal gait at three levels, namely sequence, frame and joint levels, thereby enhancing granularity.

In this paper, our contributions are summarized as follows.

– To overcome the limitation of generalization ability and granularity in supervised GAD , we propose a Fine-grained Self-supervised GAD method (FSGait) which can achieve three-level abnormal gait detection.

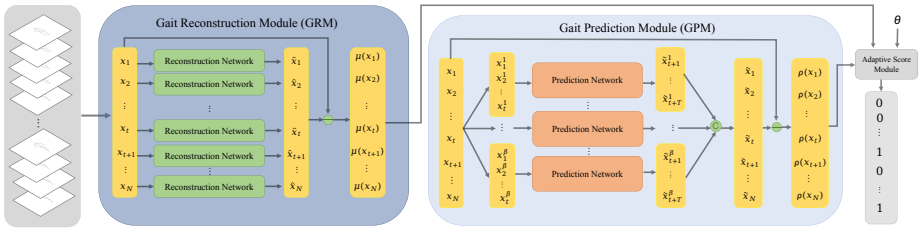
– We divide gait abnormalities into postural and temporal anomaly. To capture these, we design GRM and GPM respectively. And synthesize the results of two modules by an adaptive scoring module.
– We label a portion of a pathological gait dataset [14] at the frame level and select a subset of CMU MOCAP dataset [1] named AGD-CMU for test. FSGait achieve state-of-the-art performance for frame-level GAD and is the first work to explore joint-level GAD.

2 RELATED WORK

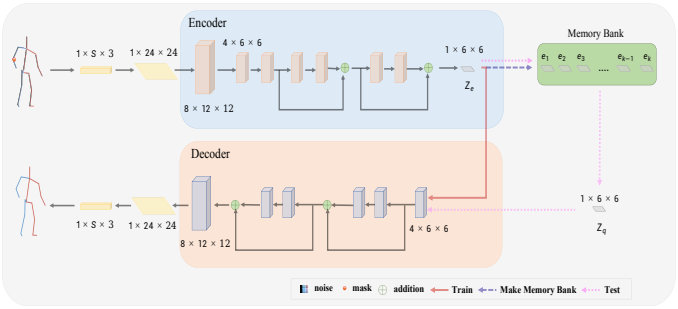
Abnormal Detection Granularity. Currently, GAD researches are mainly confined to sequence level. This implies that these studies are only capable of determining whether a gait sequence is abnormal or assess the severity of abnormal gait sequence [9]. For example, in [4, 21], researchers extract physical parameters to classify the gait sequence as either normal or abnormal. Since 2016, some neural network-based methods [8, 11, 15, 25] have emerged. Researchers utilize models that are more sensitive to timing, such as RNN [13], LSTM [15, 25] and Transformer [9], to perform this task. More recently, sensor-based approaches increase, such as IMU based [31], mobile phone sensor-based [3], and GRF-based [12] methods. These studies have continuously improved the accuracy of abnormal gait sequence detection, with the normal and abnormal binary classification of [15] achieving an accuracy of 98%. However, none of these works has refined the detection level to frame and joint. As far as we know, only the work [22] implements binary classification task of frames which is not precise. As far as we know, there is no joint-level detection method at present. This can be attributed to one main factor: the absence of datasets labeled at the frame level and joint level. In this paper, we introduce a three-level anomaly detection method. By progressing from the sequence level to the frame level, and finally to the joint level, we are able to achieve fine-grained GAD.

Abnormal Detection Generalization Capability. Supervised learning is limited to learn abnormal types with labels. For instance, [15] identified five types of abnormal gait, specifically Steppage Gait, Lurching Gait, Antalgic Gait, Stiff-legged Gait, and Trendelenburg Gait. In contrast, [22] recognized a different set of abnormal gait types, including heel abnormality gait, Parkinson gait, stroke gait, left/right leg half step gait, and full body abnormal gait. [31] identified FOG as an abnormal gait type, while [12] identified four types of abnormal gait, which include Hip disorder gait, Knee disorder gait, Ankle disorder gait, and Calcaneus disorder gait. Despite all this work, there remains an urgent need for a universal method capable of identifying a wide range of gait abnormalities. In response to this need, this paper proposes a self-supervised GAD method aimed at enhancing the generalization capability of GAD.

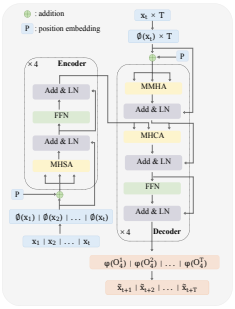
Data Modality. In terms of data modalities, particularly when concentrating on vision-based methodologies, a variety of data types have been utilized in the field of GAD. These include RGB [24, 35], silhouette [2], 2D pose [13, 32], 3D pose [22], and point cloud [23]. However, 3D pose data, which can be derived



(a) Pipeline.



(b) Reconstruction Network.



(c) Prediction Network.

Fig. 2: The overall framework of our proposed FSGait. The reconstruction network and prediction network in pipeline are shown in (b) and (c).

from video, IMU data, and other data types, exhibits superior expansibility. We refer to several articles on gait recognition [19, 33]. For the representation of gait sequence, compared with 2D pose and silhouette, 3D pose or skeleton data contain more spatio-temporal information and can better reflect some subtle anomalies. In addition, 3D pose consumes little computing resources and storage space, which facilitates the production of memory bank in our method.

3 METHOD

3.1 Framework Architecture

The proposed framework is shown in Fig. 2. We input a gait sequence denoted as $X = \{\boldsymbol{x_i}\}_{i=1}^{N}, \boldsymbol{x_i} \in R^{S\times 3}$, where N represents the number of frames and each frame contains S joints. We input X into GRM and GPM simultaneously. The reconstruction network is responsible for rebuilding each frame, the output is represented as $\widehat{\boldsymbol{X}} = \{\widehat{\boldsymbol{x}}_{\boldsymbol{i}}\}_{i=1}^{N}, \widehat{\boldsymbol{x}}_{\boldsymbol{i}} \in R^{S\times 3}$, following which we compute the *Mean Per Joint Position Error* (MPJPE), denoted as $\mu(\widehat{X}) = \{\mu(\widehat{\boldsymbol{x}}_{\boldsymbol{i}})\}_{i=1}^{N}, \mu(\widehat{\boldsymbol{x}}_{\boldsymbol{i}}) \in R$, between $\widehat{\boldsymbol{X}}$ and $\boldsymbol{X}$. The compute method of $\mu(\widehat{\boldsymbol{x}}_{\boldsymbol{i}})$ is illustrated in Eq. (1).

$$MPJPE = \frac{1}{S} \cdot \sum_{j=1}^{S} ||x_i^j - \widehat{x}_i^j|| \tag{1}$$

When X is input into GPM to detect the temporal anomaly, we initially divide X into β sliding windows. The length of each sliding window is $t+T$. Then we use the pre-t frames of each sliding window to predict the later $t+1 \sim t+T$ frames, and concatenate the prediction results to obtain the $\widetilde{X} = \{\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}\}_{i=t+1}^{N}$, $\widetilde{\boldsymbol{x}}_{\boldsymbol{i}} \in R^{S\times 3}$. After that, the $MPJPE$ ($\rho(\widetilde{X}) = \{\rho(\widetilde{\boldsymbol{x}}_{\boldsymbol{i}})\}_{i=t+1}^{N}$, $\rho(\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}) \in R$) between $\widetilde{X}$ and X is calculated. $\rho(\widetilde{\boldsymbol{x}}_{\boldsymbol{1}}) \sim \rho(\widetilde{\boldsymbol{x}}_{\boldsymbol{t}})$ is zero in default. Finally, we acquire the $\rho(\widetilde{X}) = \{\boldsymbol{\rho(\widetilde{x}_i)}\}_{i=1}^{N}$.

$\mu(\widehat{X})$ and $\rho(\widetilde{X})$, which contain spatial and temporal information, are fed into the scoring module for information fusion. Subsequently, we generate the classification outcome that determines whether each frame is normal or abnormal. On this basis, we determine whether a sequence is abnormal by identifying the presence of anomalous frames within the sequence.

Regarding the joint, we employ the Euclidean distance of each joint pair between $\boldsymbol{x}_{\boldsymbol{i}}$ and $\widehat{\boldsymbol{x}}_{\boldsymbol{i}}/\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}$, as a measure of the joint's abnormal weight, as illustrated in Fig. 1. This method allows us to pinpoint the joint with the most abnormal, which is instrumental in identifying the patient's abnormal region.

3.2 Gait Reconstruction Module

A multitude of anomalies can result in abnormal walking postures. For instance, cerebral palsy-induced muscle rigidity can lead to scissors gait. We believe that humans judge such anomalies by comparing them to numerous normal gait posture in their memory. Therefore, we propose a *Gait Reconstruction Module* (GRM) with memory bank underpinned by a 2D Convolutional Neural Network (CNN). The reasons for model selection are detailed in the experiment section.

Gait Reconstruction Network. The network is a pose reconstruction network trained by numerous normal gait poses. Refer to article [6,29], adding noise and mask to data can increase the robustness and accuracy of the model. For our task, mask helps extract the spatial relationship between the joints and the global information of pose, so we randomly mask a certain joint in the input, and add random noise to each joint.

As illustrated in Fig. 2 (b), we input a frame of gait pose denoted as $\boldsymbol{x}_{\boldsymbol{i}}$. For the pose composed of multiple joints, we believe that in the process of gait, there is not only a relationship between each joint, but also a certain relationship between the coordinate of each dimension of joints. Therefore, before entering the reconstruction network, we flat the input as R^{S*3} and embed it to R^{24*24} through the full connection layer to capture the association between joints and the global information of pose. After that, we reshape it to $\gamma(\boldsymbol{x}_{\boldsymbol{i}})$, $\gamma(\boldsymbol{x}_{\boldsymbol{i}}) \in R^{1\times 24\times 24}$. Following pose embedding, we employ 2D CNN to extract the spatial features. The parameters are presented in Supplementary Materials. After encoder, we obtain $\boldsymbol{Z}_{\boldsymbol{e}}$, $\boldsymbol{Z}_{\boldsymbol{e}} \in R^{1\times 6\times 6}$ vector. During training, we input $\boldsymbol{Z}_{\boldsymbol{e}}$ into decoder and we

restore $\boldsymbol{Z_e}$ to $R^{1\times24\times24}$, and then to the reconstructed pose $\widehat{\boldsymbol{x}}_{\boldsymbol{i}} \in R^{S\times3}$. During the training phase, the data utilized are all data during normal walking, with no involvement of abnormal conditions.

Memory Bank Establishment. After training, the $MPJPE$ between $\widehat{\boldsymbol{x}}_{\boldsymbol{i}}$ and $\boldsymbol{x_i}$ can reach about 4.2 mm. Then we freeze the parameters of encoder to establish a memory bank. We input a normal pose from the training set and put it's $\boldsymbol{Z_e}$, $\boldsymbol{Z_e} \in R^{1\times6\times6}$ into the memory bank denoted as $M = \{\boldsymbol{e_i}\}_{i=1}^{K}$, $\boldsymbol{e_i} \in R^{1\times6\times6}$. Instead of storing pose directly, we store $\boldsymbol{Z_e}$, which consumes less storage space, and the query time is greatly reduced when using memory bank. Moreover, $\boldsymbol{Z_e}$ contains more high-dimensional information than initial pose.

Once all the intermediate variables $\boldsymbol{Z_e}$ of the training data are put into the memory bank, the greedy algorithm [27] is applied for downsampling, reducing the size of the memory bank and retaining the most representative part.

Abnormality Detection Strategy. In the test stage, when abnormal data $\boldsymbol{x_i}$ comes, we first obtain $\boldsymbol{Z_e}$ of it, and then conduct a nearest neighbor look up strategy, as shown in Eq. (2) and (3).

$$f(e_i) = \frac{1}{6\times6}\sum_{k=1}^{6}\sum_{m=1}^{6}(z_e^{(j,k,m)} - e_i^{(j,k,m)})^2 \tag{2}$$

$$Z_q = argmin(f(e_i)) \tag{3}$$

Replace $\boldsymbol{Z_e}$ with the searched $\boldsymbol{Z_q}$, and input $\boldsymbol{Z_q}$ into decoder. Then we obtain a normal pose $\widehat{\boldsymbol{x}}_{\boldsymbol{i}}$ that is most similar to the original abnormal pose.

The determination of frame-level and joint-level abnormality can be achieved by comparing $\boldsymbol{x_i}$ and $\widehat{\boldsymbol{x}}_{\boldsymbol{i}}$. If $\boldsymbol{x_i}$ is normal, a high degree of similarity between the two poses is expected, resulting in a minimal distance between them. Conversely, if $\boldsymbol{x_i}$ is abnormal and memory bank solely comprises normal poses, the distance between the two poses is significantly large, as shown in Fig. 1.

3.3 Gait Prediction Module

Temporal anomaly detection is targeted at abnormalities that may not be spatially apparent but are temporally significant, such as FOG. For such anomalies, we believe humans judge them by predicting the next movement based on experience, and they consider it abnormal if the actual movement deviates from the prediction. So we aim to train a prediction network to capture temporal features across a large number of normal human gaits, without overly emphasizing the unique features of individual walks. The attention mechanism aptly fulfills our requirements. We employ the self-attention mechanism to enable the model to learn more salient features in the temporal domain, which is to learn the periodicity of the walking sequence and the variations in limb movements.

Gait Prediction Network. The network is shown in Fig. 2 (c) referring to [10,34]. During the training phase, all data utilized are normal gait data. Given that too long sequences can lead to poor prediction accuracy, we slide T frames at a time in the temporal to form β sliding windows of length $t + T$ for sliding

detection. β can be calculated by (4). It is only necessary to judge whether each T frames in sliding windows is abnormal, allowing us to identify abnormalities within an extended sequence. Most of the time, the appearance of gait abnormalities is a gradual process, so the input of the initial sliding window is not abnormal. We default that there is no obvious abnormality (*i.e.*, $\mu(\widetilde{\boldsymbol{x}}_{\boldsymbol{i}})_{i=1}^{t} = 0$), which is indeed the case in datasets.

$$\beta = \begin{cases} [\frac{N-t}{T}] & (N-t)\%T = 0 \\ [\frac{N-t}{T}] + 1 & (N-t)\%T \neq 0 \end{cases} \tag{4}$$

Encoder. As shown in Fig. 2 (c), $\boldsymbol{x}_{\boldsymbol{i}}$ represents the pose data of frame i, while $\phi(\boldsymbol{x}_{\boldsymbol{i}})$ represents the data after the pose embedding, which increase the shape of the input data from $(t, S \times 3)$ to $(t, 512)$ through linear layer. In order to enhance the temporal information of the data, we use Positional Encoding (PE) method to encode the position of each frame. The encoding vector used is P in Fig. 2 (c), and the encoding method is shown in Eqs. (5) and (6), where pos indicates the moment that data are located in temporal dimension, and l means the position of features encoded at each moment. d_{model} denotes the size of the feature-space dimension. Positional Encoding guarantees the uniqueness of each location code value, the consistency of the distance between adjacent location code values and the adaptability to the length of the data [26].

$$PE(pos, 2l) = sin(\frac{pos}{10000^{(2l/d_{model})}}) \tag{5}$$

$$PE(pos, 2l+1) = cos(\frac{pos}{10000^{(2l/d_{model})}}) \tag{6}$$

We denote the sequence after embedding as I, and denote the positional encoding sequence as Q, input it into transformer's architecture and do *Multi-Head Self-Attention* (MHSA) computation [34]. Here, we assign higher weight to the features that are conducive to future gait prediction, that is, we get more general human gait features. Following this, we go through the *Add&Layer Norm* (LN) and the *Feed-Forward Networks* (FFN) with residual connection, and then through the LN again. Finally output the intermediate variable memory when the encoder is finished. Refer to [10] for the calculation method of these layers. The whole process can be seen as Eqs. (7) to (9). Eqs. (8) and (9) implement the *FFN* step. d_k denotes the dimension of queries and keys.

$$Q_1 = LN(I + MHSA(Q, Q, Q)),\ Q = I + P \tag{7}$$

$$Q_2 = LN(Linear(LN(Linear(Q_1)))) \tag{8}$$

$$Q_3 = LN(Q_1 + Q_2) \tag{9}$$

Decoder. In decoder, *Masked Multi-Head Attention* (MMHA) and *Multi-Head Cross Attention* (MHCA) [34] are applied. Using mask operation is to

Algorithm 1 Adaptive Score.

Input: $\mu(\widehat{\boldsymbol{X}})$, $\rho(\widetilde{\boldsymbol{X}})$, θ
Output: $\sigma(\boldsymbol{X})$
if $\overline{\mu(\widehat{\boldsymbol{x}}_{\boldsymbol{i}})} < \overline{\rho(\widetilde{\boldsymbol{x}}_{\boldsymbol{i}})}$ **then**
  $V_1 = \rho(\widetilde{\boldsymbol{x}}_{\boldsymbol{i}})$, $V_2 = \mu(\widehat{\boldsymbol{x}}_{\boldsymbol{i}})$
else
  $V_1 = \mu(\widehat{\boldsymbol{x}}_{\boldsymbol{i}})$, $V_2 = \rho(\widetilde{\boldsymbol{x}}_{\boldsymbol{i}})$
end if
if $\overline{V_2} < 0.6\theta$ (normal model deviation) **then**
  $\sigma(\boldsymbol{X}) = V_1$
else
  $\sigma(\boldsymbol{x}_{\boldsymbol{i}}) = \frac{\mu(\widehat{\boldsymbol{x}}_{\boldsymbol{i}}) \cdot \rho(\widetilde{\boldsymbol{x}}_{\boldsymbol{i}})}{\theta}$
end if
return Output

focus only on the data before the current moment and exclude the influence of the data after the current moment when doing self-attention. The specific implementation of mask operation is referred to [34]. The closer the distance from the predicted frame in time sequence, the more similar it is to the predicted frame. So the initial input of our decoder selects T frames $\boldsymbol{x_t}$ pose for auxiliary model training, then the decoder's input is its last output. When we input $T \times \boldsymbol{x_t}$, we apply the same embedding and positional encoding as the encoder, and then we call the resulting sequence O. Next, we perform the calculation of Eq. (10) - Eq. (13) to get the predicted sequence as shown in Fig. 2 (c). The function of the self-attention mechanism here is the same as described above. In addition, we also use the cross-attention mechanism, whose purpose is to utilize O_1 as a query for similar features in Q_3, so as to find the temporal features in Q_3.

$$O_1 = LN(O + MMHA(O, O, O)) \tag{10}$$

$$O_2 = LN(O_1 + MHCA(O_1, Q_3, Q_3)) \tag{11}$$

$$O_3 = LN(Linear(LN(Linear(O_2)))) \tag{12}$$

$$\widetilde{x}_{t+i} = \varphi(O_4^i) = Linear(O_4^i),\ O_4 = LN(O_3 + O_2) \tag{13}$$

The predicted frame $\widetilde{\boldsymbol{x}}_{t+i}$ can be obtained by this way. The training $MPJPE$ between $\boldsymbol{x_i}$ and $\widetilde{\boldsymbol{x}}_{t+i}$ is almost 28 mm.

3.4 Adaptive Score Module

We design an adaptive scoring module to integrate GRM and GPM outputs. To deal with multiple abnormalities in a gait, we divide a gait sequence into clips of length T according to the previous sliding windows, and then we introduce how

to perform operations in each clip. We first input a clip of GRM output $\mu(\widehat{\boldsymbol{X}}) = \{\mu(\widehat{\boldsymbol{x}}_{\boldsymbol{i}})\}_{i=k}^{k+T}$, $\mu(\widehat{\boldsymbol{x}}_{\boldsymbol{i}}) \in R$, a clip of GPM output $\rho(\widetilde{\boldsymbol{X}}) = \{\rho(\widetilde{\boldsymbol{x}}_{\boldsymbol{i}})\}_{i=k}^{k+T}$, $\rho(\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}) \in R$ and abnormal threshold θ. As illustrated in Algorithm 1, the fused vector $\sigma(\boldsymbol{X}) = \{\sigma(\boldsymbol{x}_{\boldsymbol{i}})\}_{i=t+1}^{N}$, $\sigma(\boldsymbol{x}_{\boldsymbol{i}}) \in R$ of GRM and GPM is obtained.

The output is then compared with θ to determine an abnormal score per frame, as shown in Eq. (14). For each joint, we use the pose generated by the module containing V_1. The Euclidean distance between each joint of $\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}/\widehat{\boldsymbol{x}}_{\boldsymbol{i}}$ and the corresponding joint of $\boldsymbol{x}_{\boldsymbol{i}}$ is calculated to determine the joint anomaly weight.

$$Score(x_i) = \begin{cases} 1 & \sigma(\boldsymbol{x}_{\boldsymbol{i}}) > \theta \\ 0 & \sigma(\boldsymbol{x}_{\boldsymbol{i}}) < \theta \end{cases} \tag{14}$$

4 EXPERIMENTS

4.1 Datasets

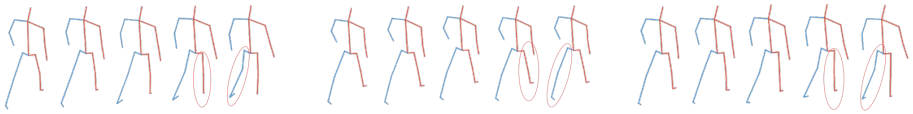
We utilize the pathological gait dataset to perform an ablation study and evaluate final impact on Abnormal Gait Dataset from CMU (AGD-CMU). Datasets are available at https://github.com/BingzhiDuan/FSGait.

AGD-CMU. The CMU MOCAP dataset comprises extensive human motion data, from which we specifically select 89 normal gaits for fine-tuning our model. Additionally, 49 abnormal gaits which contains 23 abnormal types (stiff walk, limp, zombie walk, etc.) and 20 normal gaits are selected to form a test set. Overall, AGD-CMU contains 158 gait sequences. This allows us to assess the performance of our method and compare it against the baseline. Detailed information regarding AGD-CMU is provided in the Supplementary Materials. In contrast to the pathological dataset, the motions in AGD-CMU consist of more frames and represent complete movements rather than partial segments.

Pathological gait dataset. The second dataset is a pathological gait dataset for gait analysis, including 10 subjects, each of which has 1 normal gait sequence and 5 abnormal gait sequence respectively [14]. The 3D pose data of 25 joints is captured using 6 Kinect cameras, with coordinates relative to each camera. This dataset offers both normal and pathological gaits, encompassing stiff-legged gait, lurching gait, steppage gait, antalgic gait, and Trendelenburg gait. Visualization of the dataset is available in the Supplementary Materials.

FOG data. Based on the normal gait data from pathological gait dataset, we also generate some FOG data to to evaluate the effectiveness of GPM, as depicted in Fig. 3. This process entails freezing a frame within a normal gait sequence and appending subsequent frames from that point onwards to simulate the freezing gait sequence.

Data label. Since pathological gait dataset only gives sequence-level labels and cannot evaluate frame-level anomalies, we visualize part of the data and label them at the frame level. We mark 111 abnormal fragments, including stiff-legged gait, lurching gait, steppage gait, and trendelenburg gait. We also add 59 normal gait sequences, so the test dataset has a total of 170 gait sequences.



(a) Normal Gait **(b)** Freezing Gait **(c)** Predicted Gait

Fig. 3: Distinction among three gaits. (a) is 5 successive frames. (b) is made by freezing the first frame of Normal Gait. (c) is the output of prediction network. The parts circled in red are where the Freezing Gait is obviously different from the other two gaits.

Antalgic gait is not marked, because this gait is difficult to recognize for non-medical professionals. Abnormal gait data of AGD-CMU exclude normal gait segments, so it does not require annotation.

Joints selection. The datasets have 25 and 31 joints, but our baseline method [22] use 17 joints. To ensure fair comparison, we also use 17 joints for testing. According to [18], gait recognition accuracy can potentially improve with reduced joint counts, emphasizing the highlighting of gait characteristics. The selected joints are detailed in the Supplementary Materials.

4.2 Training Settings

In our study, we set the sliding window parameters $T = 10$ and $t = 25$. In Gait Reconstruction Network settings, we set the *learning rate* to 3×10^{-5}, and batch size to 4. As for Gait Prediction Network, it is trained by Adam algorithm with batch size 4 and learning rate 10^{-4}. And the way of layer initialization is Xavier initialization. The number of encoder layer or decoder layer is 4. *Dropout* equals to 0.3. *Number of heads* is 8. The loss functions of two networks are illustrated in equations Eqs. (15) and (16), where $train\_var$ denotes the var of total training data, and MSE denotes the Mean Squared Error.

$$L_{recon} = \frac{(\widehat{x}_i - x_i)^2}{S \times 3 \times train\_var} \tag{15}$$

$$L_{pre} = MSE(\{\widetilde{x}_i\}_{i=t+1}^{t+T}, \{x_i\}_{i=t+1}^{t+T}) \tag{16}$$

4.3 Feasibility Test

We validate the effectiveness of our two modules by two specific examples (lurching gait and FOG).

Reconstruction module test. Fig. 4 presents the curve of reconstruction detection results. We select two segments of walking patterns, one normal and one abnormal (lurching), each consisting of 35 frames. The green curve represents the $MPJPE$ between the reconstructed pose and the input pose for each frame of the normal walk. Similarly, the blue curve depicts the $MPJPE$ for each frame of the abnormal walk sequence. The section with the red background indicates the position of the abnormal frames. As can be inferred from the figure, there is

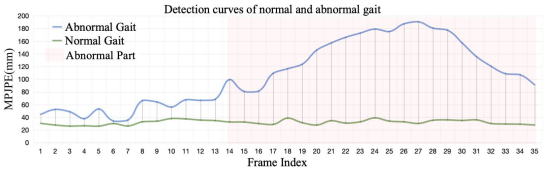


Fig. 4: Reconstruction detection curve of normal and abnormal gait.

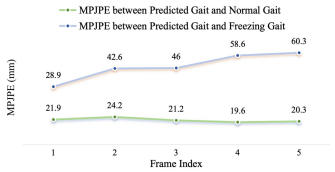


Fig. 5: Prediction difference between normal and freezing gait.

a high degree of coincidence between the actual position of the abnormal part and the prominent position on the blue curve, demonstrating the effectiveness of our algorithm in detecting such anomalies.

Prediction module test. As shown in Fig. 3 and Fig. 5, the detection of FOG is demonstrated in an example. We freeze 5 frames after t frame. The predicted gait closely resemble the original (normal) gait without freezing, but significantly differ from the gait with freezing, realizing FOG detection. The anomaly caused by freezing leads to an increasing MPJPE between the predicted gait and the input gait, whereas normally our MPJPE curve is flat.

4.4 Ablation Study

In this section, we introduce a comparative experiment for model selection, demonstrate the respective functions of the two modules, and discuss the selection of the anomaly threshold.

Reconstruction Model Selection. The task of reconstructing a single pose is simple. We attempt to employ the relatively complex Transformer model but found that it does not converge. Subsequently, we evaluate CNNs, Graph Convolutional Networks (GCN), and Multilayer Perceptrons (MLPs) for this task, comparing their performance over 10 epochs as detailed in Tab. 1. Among these models, CNN demonstrates superior reconstruction effect with the fewest parameters, thus it is selected.

Effect of GRM. Firstly, we present the ablation study results of frame-level anomaly detection. We take the currently found frame-level anomaly detection method [22] as our baseline. In Tab. 2, we use the labeled pathological gait dataset to test the effectiveness of memory bank and GRM. As shown in Tab. 2, our methods significantly improve the metrics compared to baseline. The highest

Table 1: Comparison of effects of different models in reconstruction tasks. RM denotes mean reconstruction MPJPE and Param denotes parameters number in models.

Model	Param	RM (mm)
MLP	1919751	21.7
CNN	63533	5.8
GCN	68101	26.2

Table 2: Frame-level Postural Anomaly Detection Result. Results are highlighted as first , second and third for each evaluation index.

Method	Memory Bank	AUC	Accuracy	Precision	Sensitivity	Specificity	F1-Score
baseline [22]	–	0.737	0.640	0.379	0.681	0.626	0.487
GRM(ours)	×	0.937	0.872	**0.823**	0.667	**0.947**	0.737
GRM(ours)	✓	**0.953**	**0.886**	0.743	**0.877**	0.889	**0.805**
GPM(ours)	–	0.890	0.818	0.630	0.780	0.832	0.697
FSGait(ours)	✓	0.943	0.879	0.740	0.852	0.890	0.792

Table 3: FOG Detection Result. Results are highlighted as first , second for each evaluation index.

Method	GPM	AUC	Accuracy	Precision	Sensitivity	Specificity	F1-Score
baseline [22]	–	0.457	0.471	0.569	0.491	0.441	0.527
GRM(ours)	–	0.503	0.402	0.702	0.254	0.748	0.373
FSGait(ours)	✓	**0.859**	**0.717**	**0.943**	**0.634**	**0.909**	**0.758**

values of AUC, Accuracy, Sensitivity and F1-Score appear in GRM with Memory Bank which indicates that the ability to detect abnormal frames is improved compared with GRM without Memory Bank.

Nevertheless, the highest Precision and Specificity values are observed in the absence of a memory bank, suggesting stricter anomaly identification criteria. As long as a frame is considered an anomaly, there is a high probability that it is actually an anomaly. However, such strict criteria can misclassify many abnormal samples as normal. After adding memory bank, some of these problems are avoided leading to overall improved accuracy.

Comparing GPM and FSGait reveals the effectiveness of adding GRM. Furthermore, the small gap in AUC and Accuracy between GRM and FSGait suggests that incorporating GPM has minimal impact on GRM itself, while adding GPM will greatly improve the accuracy in temporal anomaly detection.

Effect of GPM. In Tab. 3, the effectiveness of GPM is evaluated using self-generated FOG data. It is evident that incorporating GPM significantly enhances GRM's capability to detect temporal anomalies. As shown in Tab. 3, we can see that the addition of GPM greatly improves GRM's ability to recognize temporal anomalies. As can be seen from Tab. 3, both the baseline method and GRM alone struggle to capture temporal information, which is crucial for recognizing phenomena like FOG. In contrast, our approach (FSGait) achieves substantial improvements across all metrics compared to the baseline.

Abnormal threshold selection. As for the selection of abnormal thresholds, we test different thresholds, as shown in Fig. 6. When the threshold is about 50mm, the model has the best detection effect.

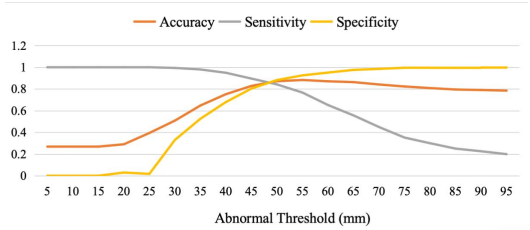


Fig. 6: Comparison of the results of different thresholds.

Table 4: MOCAP Dataset Result. Results are highlighted as first for each evaluation index.

Method	AUC	Accuracy	Precision	Sensitivity	Specificity	F1-Score
baseline [22]	0.668	0.742	0.821	**0.869**	0.212	**0.845**
FSGait(ours)	**0.873**	**0.768**	**0.937**	0.763	**0.787**	0.841

4.5 AGD-CMU Result.

We continue to verify the frame-level anomaly detection effect on the AGD-CMU, as shown in Tab. 4. Our method shows a 0.205 increase in AUC compared to the baseline, and a significant rise in Specificity from 0.212 to 0.787. These results demonstrate substantial enhancements in our ability to differentiate between normal and abnormal instances, significantly reducing the likelihood of misclassifying normal frames as abnormal.

To compare with more baseline and prove the generalization capability of FSGait, we select part of AGD-CMU and conduct comparative experiments, which are presented in Supplementary Materials.

Table 5: Sequence-level Detection Result.

Threshold	Sequence Number	TP+TN	FP+FN	Accuracy
1 frame	2728	2536	192	92.96%
2 frame	2728	2545	183	93.29%
3 frame	2728	2639	189	93.07%

4.6 Sequence-level Result.

Based on frame-level GAD, we assess the effect of sequence-level GAD, as depicted in Tab. 5. The pathological gait dataset have sequence-level labels, enabling us to conduct testing on a separate test set comprising a total of 2728 sequences. Here, a sequence is deemed abnormal if it contains at least n abnormal frames. Notably, when n increases, sequence-level accuracy increases until 93.29% and then decline. We also tested sequence-level detection on AGD-

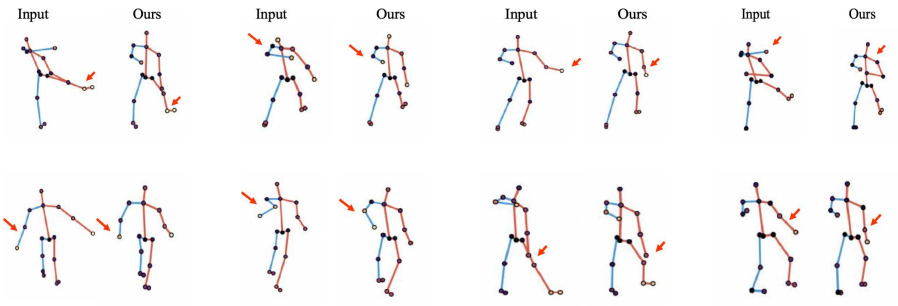


Fig. 7: Joint level detection result visualization. We compare the abnormal input pose and reconstructed (predicted) pose of ours to detect joint-level abnormality.

CMU, which contains 89 test samples, achieving an accuracy of 100% ($n = 1$). This means that we achieve fine-grained detection while still maintaining a high sequence-level detection accuracy.

4.7 Joint-level Result.

For joint-level GAD, the joints anomaly is measured by the Euclidean distance of each joint pair between $\boldsymbol{x_i}$ and $\widetilde{\boldsymbol{x}}_i/\widehat{\boldsymbol{x}}_i$. Due to the lack of joint-level labels, visual results are used to verify the effect of FSGait. As shown in Fig. 7, for different types of abnormalities, specific abnormal joints are detected. For example, in the first line of Fig. 7, the first two images are lurching gaits and its closest (reconstructed/predicted) normal gaits. The lighter colored joints on the foot are detected as the most abnormal parts, which is consistent with reality.

5 CONCLUSION

In this paper, we present a fine-grained self-supervised model for abnormal gait detection. Our key innovation is dividing gait abnormality to postural anomaly and temporal anomaly and devising two modules for them. With the reconstruction and prediction networks trained in self-supervised way, we can refine the detection level from sequence to frame and joint which enhances generalization ability simultaneously. Our method achieves state-of-the-art performance in frame-level and joint-level GAD. Additionally, we annotate a pathological gait dataset at the frame level and form an AGD-CMU dataset, which serves as a valuable resource for future abnormal gait detection methods. Moving forward, several challenges persist in the field of GAD, prompting the following considerations for future research:

– In terms of the fusion of spatial and temporal information, we hope to integrate them in the network training, so that they can reference each other.
– In the application field of abnormal gait, we hope to expand our works not only in the field of diagnosis, but also in rehabilitation or elderly care.

References

1. CMU Graphics Lab: Carnegie-Mellon Motion Capture (MoCap) Database, (2003), http://mocap.cs.cmu.edu 3
2. Bauckhage, C., Tsotsos, J.K., Bunn, F.E.: Automatic detection of abnormal gait. Image and Vision Computing **27**(1-2), 108–115 (2009) 3
3. Bonetto, R., Soldan, M., Lanaro, A., Milani, S., Rossi, M.: Seq2seq rnn based gait anomaly detection from smartphone acquired multimodal motion data. arXiv preprint arXiv:1911.08608 (2019) 3
4. Chaaraoui, A.A., Padilla-López, J.R., Flórez-Revuelta, F.: Abnormal gait detection with rgb-d devices using joint motion history features. In: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG). vol. 7, pp. 1–6. IEEE (2015) 3
5. Chai, T., Li, A., Zhang, S., Li, Z., Wang, Y.: Lagrange motion analysis and view embeddings for improved gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20249–20258 (2022) 1
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 5
7. Di Biase, L., Di Santo, A., Caminiti, M.L., De Liso, A., Shah, S.A., Ricci, L., Di Lazzaro, V.: Gait analysis in parkinson's disease: An overview of the most accurate markers for diagnosis and symptoms monitoring. Sensors **20**(12), 3529 (2020) 1
8. Elkholy, A., Makihara, Y., Gomaa, W., Ahad, M.A.R., Yagi, Y.: Unsupervised gei-based gait disorders detection from different views. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 5423–5426. IEEE (2019) 3
9. Endo, M., Poston, K.L., Sullivan, E.V., Fei-Fei, L., Pohl, K.M., Adeli, E.: Gaitforemer: Self-supervised pre-training of transformers via human motion forecasting for few-shot gait impairment severity estimation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 130–139. Springer (2022) 1, 3
10. Gong, D., Lee, J., Kim, M., Ha, S.J., Cho, M.: Future transformer for long-term action anticipation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3052–3061 (2022) 6, 7
11. Guo, Y., Deligianni, F., Gu, X., Yang, G.Z.: 3-d canonical pose estimation and abnormal gait recognition with a single rgb-d camera. IEEE Robotics and Automation letters **4**(4), 3617–3624 (2019) 3
12. Jani, D., Varadarajan, V., Parmar, R., Bohara, M.H., Garg, D., Ganatra, A., Kotecha, K.: An efficient gait abnormality detection method based on classification. Journal of Sensor and Actuator Networks **11**(3), 31 (2022) 3
13. Jun, K., Lee, D.W., Lee, K., Lee, S., Kim, M.S.: Feature extraction using an rnn autoencoder for skeleton-based abnormal gait recognition. IEEE Access **8**, 19196–19207 (2020) 1, 3
14. Jun, K., Lee, Y., Lee, S., Lee, D.W., Kim, M.S.: Pathological gait classification using kinect v2 and gated recurrent neural networks. IEEE Access **8**, 139881–139891 (2020) 3, 9
15. Khokhlova, M., Migniot, C., Morozov, A., Sushkova, O., Dipanda, A.: Normal and pathological gait classification lstm model. Artificial intelligence in medicine **94**, 54–66 (2019) 1, 3

16. Kour, N., Gupta, S., Arora, S.: Sensor technology with gait as a diagnostic tool for assessment of parkinson's disease: a survey. Multimedia Tools and Applications **82**(7), 10211–10247 (2023) 1
17. Li, X., Makihara, Y., Xu, C., Yagi, Y., Ren, M.: Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13309–13319 (2020) 1
18. Liao, R., Cao, C., Garcia, E.B., Yu, S., Huang, Y.: Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In: Biometric Recognition: 12th Chinese Conference, CCBR 2017, Shenzhen, China, October 28-29, 2017, Proceedings 12. pp. 474–483. Springer (2017) 10
19. Liao, R., Yu, S., An, W., Huang, Y.: A model-based gait recognition method with body pose and human prior knowledge. Pattern Recognition **98**, 107069 (2020) 4
20. Malik, O.A.: Deep autoencoder for identification of abnormal gait patterns based on multimodal biosignals. International Journal of Computing and Digital Systems **10**(1), 1–8 (2021) 1
21. Nguyen, T.N., Huynh, H.H., Meunier, J.: Skeleton-based abnormal gait detection. Sensors **16**(11), 1792 (2016) 3
22. Nguyen, T.N., Huynh, H.H., Meunier, J.: Estimating skeleton-based gait abnormality index by sparse deep auto-encoder. In: 2018 IEEE Seventh International Conference on Communications and Electronics (ICCE). pp. 311–315. IEEE (2018) 2, 3, 10, 11, 12, 13
23. Nguyen, T.N., Meunier, J.: Estimation of gait normality index based on point clouds through deep auto-encoder. EURASIP Journal on Image and Video Processing **2019**(1), 65 (2019) 3
24. Nieto-Hidalgo, M., Ferrández-Pastor, F.J., Valdivieso-Sarabia, R.J., Mora-Pascual, J., García-Chamizo, J.M.: A vision based proposal for classification of normal and abnormal gait using rgb camera. Journal of biomedical informatics **63**, 82–89 (2016) 3
25. Pachón-Suescún, C.G., Pinzón-Arenas, J.O., Jiménez-Moreno, R.: Abnormal gait detection by means of lstm. International Journal of Electrical & Computer Engineering (2088-8708) **10**(2) (2020) 1, 3
26. Ren, J., Wang, A., Li, H., Yue, X., Meng, L.: A transformer-based neural network for gait prediction in lower limb exoskeleton robots using plantar force. Sensors **23**(14), 6547 (2023) 7
27. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2022) 6
28. Sethi, D., Bharti, S., Prakash, C.: A comprehensive survey on gait analysis: History, parameters, approaches, pose estimation, and future work. Artificial Intelligence in Medicine **129**, 102314 (2022) 1
29. Shan, W., Liu, Z., Zhang, X., Wang, S., Ma, S., Gao, W.: P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In: European Conference on Computer Vision. pp. 461–478. Springer (2022) 5
30. Shen, C., Yu, S., Wang, J., Huang, G.Q., Wang, L.: A comprehensive survey on deep gait recognition: algorithms, datasets and challenges. arXiv preprint arXiv:2206.13732 (2022) 1
31. Sigcha, L., Borzi, L., Pavon, I., Costa, N., Costa, S., Arezes, P., López, J.M., De Arcas, G.: Improvement of performance in freezing of gait detection in parkinson's

disease using transformer networks and a single waist-worn triaxial accelerometer. Engineering Applications of Artificial Intelligence **116**, 105482 (2022) 1, 3
32. Sugiyama, Y., Uno, K., Matsui, Y.: Types of anomalies in two-dimensional video-based gait analysis in uncontrolled environments. PLOS Computational Biology **19**(1), e1009989 (2023) 3
33. Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G.: Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 2314–2318. IEEE (2021) 4
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 6, 7, 8
35. Zhang, A., Yang, S., Zhang, X., Zhang, J., Zhang, W.: Abnormal gait detection in surveillance videos with fft-based analysis on walking rhythm. In: Image and Graphics: 9th International Conference, ICIG 2017, Shanghai, China, September 13-15, 2017, Revised Selected Papers, Part I 9. pp. 108–117. Springer (2017) 3
36. 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. pp. 20228–20237 (2022) 1