2022 International Conference on Virtual Reality, Human-Computer Interaction and Artificial Intelligence (VRHCIAI) | 978-1-6654-9182-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/VRHCIAI57205.2022.00021

Multi-Level Fine-Tuned Transformer for Gait Recognition

1st Huimin Wu
College of Computer Science and Technology
Qingdao University
Qingdao, China
2021020693@qdu.edu.cn

2nd Aite Zhao*
College of Computer Science and Technology
Qingdao University
Qingdao, China
zhaoaite@qdu.edu.cn

***Abstract*—Gait Recognition aims to identify individuals by their walking patterns. Abundant spatial and temporal characteristics are contained in human walking for indicating unique personal identity. There have been a large number of existing gait recognition approaches achieving good performance via the analysis of spatial and temporal characteristics. However, most of them directly concatenate the separately extracted temporal features and spatial ones for gait recognition, rarely considering the overall spatiotemporal structure and the inner connection between them. Therefore, we establish a Multi-Level Fine-tuned Transformer trained on sequential gait data, to capture and analyze the overall spatiotemporal information for gait recognition. Specifically, we first design a point-level spatial feature extractor for spatial feature extraction of frames of the gait sequence, based on typical Transformer architecture with positional encoding and multi-head self-attention mechanism. Subsequently, a frame-level temporal feature extractor is employed for the overall spatiotemporal features by learning dynamic information of the extracted spatial features from frames. Moreover, fine-tuned strategy is followed in the process of model training, to profoundly explore intra-frame spatial information, as well as global spatiotemporal features of the whole walking sequence. Extensive experiments conducted on three public datasets (normal and abnormal gait) show promising classification results, compared with several state-of-the-art techniques.**

Keywords-gait recognition; gait sequences; spatiotemporal; Transformer network; positional encoding

I. INTRODUCTION

Gait recognition is a biometric technology that aims to recognize an individual's identity based on walking styles. In contrast to other popular objects of biometric recognition (e.g., facial recognition [1], fingerprint recognition [2] and palm recognition [3]), gait data can be captured over longer distances and does not necessarily require the individual's cooperation, which is an important advantage in many identification application fields (e.g., crime-prediction and intelligent traffic surveillance). Besides, human gait is not easy to camouflage as private physiological habits for a long time, which means that identification tasks based on gait features are more reliable. Thus, gait recognition has attracted large attention from the research community.

In recent years, a significant amount of research effort has been dedicated to the study of human identification based on gait data collected from different types of sensors, and they can be included by two main categories: Wearable Sensors and Ambit Sensors [4]. Wearable sensors, like wearable inertial, force and pressure sensors, having additional requirements to been placed on certain parts of the subject's body, can directly measure human motion data of the body segments. Although this specific measurement of moving parts facilitates capturing pure walking dynamics [5],it still limits their application in some scenarios. By comparison, Ambit Sensors are placed in the environment with non-contact and no subjects' active cooperation. These non-wearable gait recognition systems primarily rely on gait data from vision-based sensors, such as RGB images/videos, depth images and skeleton data. Among them, depth images collected by depth-based RGB-D cameras take the depth information of 3D space into consideration, as well as 2D color information included in those from single RGB cameras. It means that they can provide human richer characteristics for the identification task. What's more, skeleton data, extracted from RGB-D silhouette sequence or directly captured from newly developed depth-based RGB-D cameras such as Intel RealSense or Microsoft Kinect [6] , preserves 4D human motion data $(xyz\text{-}t)$ in depth images, while trying to minimize the interference from color and environmental background.

With the advantages of imaging sensors, amount of gait recognition methods based on skeletons or silhouettes has been developed. On the one hand, as the primary backbone architectures for computer vision, various deep learning models based on 2D convolution operation (CNNs) have been implemented with promising results. While these methods merely utilize static spatial features at the frame-level, with no consideration on gait changes of walking. Moreover, some researchers employ 3D convolution mapping multiple contiguous consecutive frames to capture motion information [7]. However, 3D CNN frameworks require a fix-length gait sequence as input, which means a lack of flexibility.

On the other hand, although some specialized deep learning algorithms, like recurrent neural network (RNN) architecture, are designed for temporal feature extraction, they still have some significant limitations when applied to gait recognition. Despite this architecture can flexibly utilize the relationship between variable time series data, it lacks detailed analysis of spatial information in each frame. In addition, RNNs require to get the input data in chronological sequence, making it difficult to learn dependencies between distant positions when addressing larger time series data. In particular, there are several state-of-the-art methods corresponding to sequential information, achieving great improvement in this case, such as bi-directional RNN, LSTM and GRU. However, these methods still follow the principle of sequential input, and the fundamental constraint of modeling long-range dependencies has not been fully addressed.

Furthermore, in recent years, another deep neural network, Transformer, has begun to capture the attention of many researchers. Unlike the structure of CNN and RNN, which designed for one kind of structured data specifically, Transformer can serve as a general-purpose backbone for both image data and time series data, and is very suitable for

978-1-6654-9182-2/22/$31.00 ©2022 IEEE
DOI 10.1109/VRHCIAI57205.2022.00021

the processing gait sequence data. Transformer was the first model developed for textual data using pure attention mechanism [8]. Attention mechanism in Transformer can flexibly learn global representations among embedding vectors in a sequence of data, without using sequence-aligned RNNs or convolution. Transformer and its variants have been shown to perform well in natural language processing (NLP) and other relevant sequence transduction tasks [9]. Therefore, many researchers studied its adaptability to computer vision, and recently showed promising results in various tasks related to vision [10],[11]. This inherent characteristic of attention mechanism in Transformers indicates it can be generalized to gait recognition tasks for learning both spatial and temporal feature in 4D skeleton sequence data.

Inspired by recent progress in deep learning, a multi-level fine-tuned spatiotemporal Transformer is designed to analyze gait patterns as well as dynamic changes in detail. It includes two components: a multi-level feature extraction module and a simple classification module. The multi-level module for extracting features consists of a point-level spatial feature extractor and a frame-level temporal feature extractor based on Transformer encoder architecture. Firstly, the point-level spatial feature extractor learns spatial features among all positions in the interior of each frame input. Then, the frame-level feature extractor mainly calculates the global dependencies of each frame to learn temporal features. Finally, a softmax classifier is employed to realize the recognition of extracted gait features. Besides the proposed model's training adopts fine-tuning strategy, where spatial structure and information are analyzed in as much detail as possible through pre-training the point-level spatial feature extractor using label monitoring.

The main contributions of this paper can be summarized as follows:

- A multi-level fine-tuned Transformer network is proposed for gait spatiotemporal feature extraction, fusion and recognition.
- A fine-tuned training strategy is employed to analyze interior spatial structure of each frame data sufficiently, while exploring the dynamic characteristics of gait sequence.
- The proposed method has been evaluated on two public normal gait datasets, SDUgait and UNITO datasets, and is shown to promising performance. Specifically, it can be extended to abnormal gait recognition tasks.

The structure of this paper is as follows. Section II discusses related work. The proposed method and details of each module is introduced in Section III. Section IV presents the results of the system evaluation. The conclusions are shown in Section V.

II. Related Work

Considerable approaches have been investigated in order to advance state-of-the-art in gait recognition. We here explore these methods from two perspectives of 1) data-processing based methods and 2) model-based ones.

A. Data-Processing Based Methods

There are several approaches to address the input of gait silhouette sequences and generate different feature representations, and they could be summarized as template-based and sequence-based. The template-based methods compress all frame data of a gait sequence into one image or gait template for identification. Gait Energy Image (GEI), as one of the most prevalent gait templates, compresses a 3D-video sequence onto a 2D-image template by averaging all frames in it [12]. Wu *et al.* acquired a GEI by aligning and averaging the silhouettes along the timeline, and then computed the similarity between probe and gallery GEIs for gait recognition [13]. Li *et al.* utilized the approach of reconstructing a gait template from a partially masked gait pattern for gait recognition invariant [14]. This compression method is simple and easy to implement. However, this sacrifice of the dynamic variation of silhouette sequences inevitably results in a loss of temporal features.

Different from gait templates, sequence-based methods reserve temporal relationships of silhouette sequence. Chao *et al.* [15] regarded a gait as a set of independent frames from which they learned identity information. Fan *et al.* designed Gaitpart to generate the spatiotemporal representations as a part-dependent approach [16]. Although they have learnt the input data in dimensions of time and space, directly learning contour images makes it more sensitive to the changes of environment and the subject's clothing. Extracting skeleton structure from raw silhouettes can enhance the models' anti-interference capability. Li and Zhao proposed a deep network to learn spatiotemporal characteristics from a sequence of human body skeleton coordinates [17]. This kind of data processing locates and tracks 4D skeleton coordinates from raw silhouette sequences as model input, and exhibits a significant level of robustness while maintaining temporal features of gait sequences.

B. Model-Based Methods

With remarkable advantages of the developments of deep learning, various model architectures are designed for different types of data, and plays an active role in many fields [18], [19]. Directed at visual related tasks, CNN-based methods have also been attempted to describe the spatial structure of human body during walking. In [20], gait features of a person were extracted by a deep 2D CNN from Gait Energy Image. Chao *et al.* added a max function after 2D convolution which extracting spatial features to generate a gait template [15]. To learn both spatial and temporal features, 3D CNN models were introduced into gait recognition [7]. Huang *et al.* proposed to integrate the decoupling process into a 3D convolution framework to enhance spatial-temporal salience [21]. In addition, integration of CNNs and RNNs for learning the temporal relationships following spatial encoding has also been used for gait recognition in many literatures. Zhang *et al.* proposed a method to divide gait silhouettes into 4 horizontal parts fed into an individual CNN and then use an attention-based LSTM to output frame-level attention scores for them [22]. Despite the fact that these methods adequately explain gait patterns, the structure and position relationships between body parts are not considered.

The position encoding in Transformer architectures can inject some information about the relative or absolute of the input tokens flexibly, in both time and space domain. It can be employed for spatial structure and temporal changes in gait sequences to improve recognition performance. Even

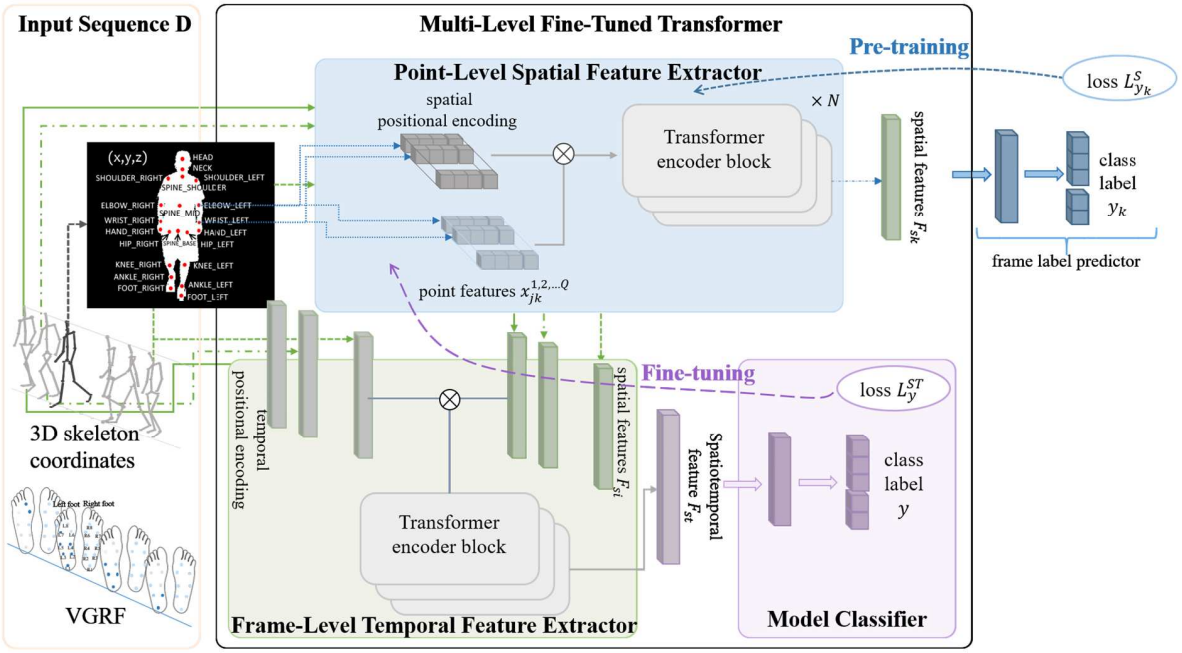


Figure 1. Framework of the proposed method. It includes two modules: a point-level spatial feature extractor and a frame-level temporal feature extractor based on Transformer network for learning spatiotemporal characteristics effectively. The training proceeds following fine-tuned strategy. The frame inputs are fed into the point-level feature extractor and the parameters are pre-rained by minimizing the frame label prediction loss. The spatial features of each frame of the gait sequence extracted from pre-trained point-level feature extractor are inputted into the frame-level temporal feature extractor for fine-tuning parameters of the whole model.

though partial mechanisms in Transformer have been integrated into various models for normal and abnormal gait recognition, like attention mechanism [23], [24], an entire Transformer architecture is rarely involved.

In this paper, we propose a multi-level fine-tuned Transformer network to process the input data and learn effective spatial structure representation from each frame data while grasping the streaming information. The proposed point- and frame- level feature extractor based on Transformer mechanism enhances the capability for gait recognition in both two dimensions. In addtion, we evaluate this method on three datasets, involving two skeleton datasets, SDUgait dataset and UNITO dataset for normal human identification, and PD dataset for severity rating of the abnormal gait.

III. METHOD

In order to obtain better results with full consideration of spatial features and dynamic changes in gait recognition, we propose a multi-level fine-tuning transformation method for spatiotemporal feature extraction. In this section, we will first outline the proposed approach. We then describe the point level and frame level spatiotemporal extractors based on transformer architecture. Finally, we introduced the details of the fine-tuned training process.

A. Overview

Our proposed method is displayed in Fig. 1. In this system, input sequences are fed into the model consisting of a point-level spatial feature extractor and a frame-level temporal one for gait recognition. The point-level spatial feature extractor utilizes an encoder based on Transformer to handle each frame of the input sequence for spatial relations among different sampling points, in which a positional encoding and self-attention mechanism are applied to store location information of points, as well as calculate the structural dependencies among them. Afterwards, the frame-level temporal feature extractor, sharing a similar architecture with the point-level one, extracts temporal features of different frames in the whole gait sequence, using learnt point-level spatial information in the previous stage.

First, we formulate the gait recognition problem. The gait dataset have N labelled walking sequences $D_S = \{D_1, D_2, \dots, D_N\}$ with corresponding C-class label set $Y_S = \{y_1, y_2, \dots, y_N\}$, $\forall y_j \in [C] = \{1, 2, \dots, C\}$. The sequence of a walking is defined as $D_j = \{x_i \in \mathbb{R}^{P\times Q}, i = 1, 2, \dots, T\}$, and the sequence $L_j = y_j \times 1_T$ indicates the corresponding truth labels of D_j. T represents the temporal length of the data sequence, while P and Q denote the number of sampling points recorded at each instant of time and the feature dimension of each point respectively. The model training follows fine-tuned strategy where two point-level and frame-level feature extractors are divided into two stages of training for spatiotemporal feature extraction tasks. For the stage of spatial feature extraction, we randomly sample K frames $\{x_{jk} \in D_j, k = 1, 2, \dots, K\}$ as input from the sequence D_j, and the goal is to make the submodule output spatial features $F_{sk} \in \mathbb{R}^{d_s}$ from inputs successfully, under the supervision of label y_j. For the stage of extracting temporal features, the frame-level feature extractor gets extracted T spatial features $\{F_{si} \in \mathbb{R}^{d_s}, i = 1, 2, \dots, T\}$ from pre-trained point-level spatial feature extractor as input, as well as aims to obtain a temporal feature vector $F_{st} \in \mathbb{R}^{d_t}$ with the shape of $d_t \times 1$, corresponding to the gait sequence D_j. For the classification's task, a simple connection layer with the softmax function is utilized to learn the output feature F_{st} and then produce the final classification result y'_j.

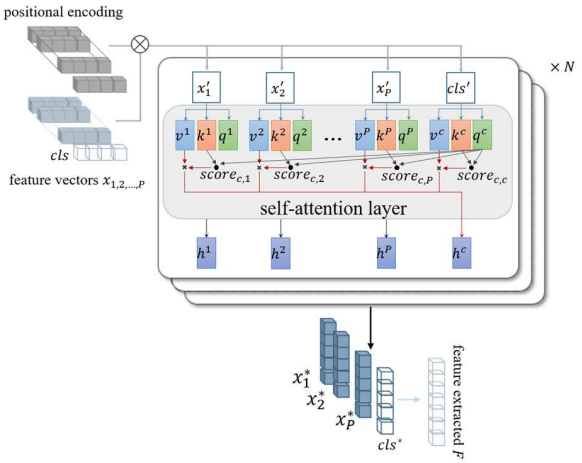


Figure 2. The feature extractor based on Transformer architecture. The complete flow of data is shown in the figure, and an additional CLS token

B. Point-level Spatial Feature Extractor

This section is intended to design an effective spatial feature extractor to analyze various points' structural characteristics in a frame of gait data. We reshape multiple frames in advance as a training sample, and input each frame consisting of P points into the point-level spatial feature extraction for effective structural spatial information. Since the relative position and structure relation of P points in different frames are consistent, this point-level spatial feature extraction could deal with all frames of the input sequence, which means smaller memory and computation cost during training and testing.

The point-level spatial feature extractor follows the architecture of Transformer encoder shown in Fig. 2, where multiple Transformer blocks are stacked. Each individual block has the same structure, the kernels of which are the multi-head self-attention mechanism and positional encoding. Here, we describe the process of extracting spatial information with the help of a frame of data.

1) Multi-head self-attention mechanism: In exploring the structure relationships of the points, we follow the typical Transformer model and computes similarity of points' features using scaled dot-product attention:

$$\text{Attention}(X) = \text{softmax}\left(\frac{(XW_Q)(XW_K)^T}{\sqrt{d_k}}\right)(XW_V) \quad (1)$$

where X is the feature metric of P' tokens, $W_Q, W_K, W_V \in \mathbb{R}^{P' \times d}$ are weight metrics of the queries, keys and values, and d_k is the key dimension. In this module, a token represents a point's feature vector, and P' tokens contain the features of P points in a frame as well as an additional CLS token, which is outputted as extracted spatial feature. Multi-head self-attention allows the module to concatenate different representation subspaces to enrich the information learnt.

2) Positional encoding: Since attention mechanism lack the relative positional information, embedded in convolution and recurrent architecture inherently, positional encodings is added to embedded characteristics of the inputs at the bottoms of the feature extractor based on Transformer network. There have been multiple existing positional encoding schemes exploited, mainly classified into absolute and relative positional encoding [25]. Here, we use the absolute one, with the purpose of less computational work. For each position index p, encoding vector is given by

$$PE(p,i) = \begin{cases} \sin(\omega_i p), & i\%2 = 1 \\ \cos(\omega_i p), & i\%2 = 0 \end{cases} \quad (2)$$

where ω_i is the hand-crafted frequency for each dimension i. The encoding vectors for different dimensions and positions can be computed directly, so it is flexible in researching both spatial structure and dynamic information. Proposed by the point-level spatial feature extractor, the relationships between points are fully explored and a feature vector $F_{si} \in \mathbb{R}^{d_s}$ representing structural information of points is extracted for a frame of input data $x_i \in \mathbb{R}^{P \times Q}$. About the sequence of a walking with T frames, the raw gait features $D_j = \{x_i, i = 1, 2, \dots, T\}$ are fed into this point-level feature extractor simultaneously, and then spatial features $\{F_{si}, i = 1, 2, \dots, T\}$ are outputted keeping the same timing as the input sequence.

C. Frame-level Temporal Feature Extractor

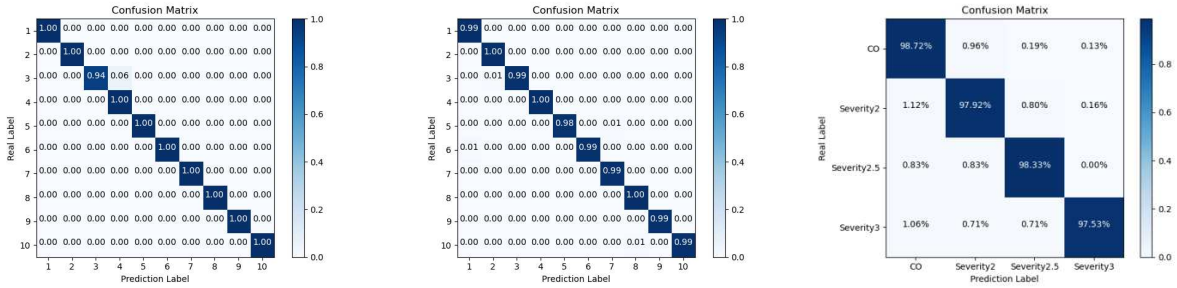
In temporal feature extraction, a network architecture is designed based on the similar architecture of Transformer encoder. Since spatial structure of each gait frame has been retrieved through the previous submodule, the role of this frame-level temporal feature extractor is to find dynamic changes in the temporal dimension.

Spatial features $\{F_{si}, i = 1, 2, \dots, T\}$ extracted by the point-level spatial feature extractor, with a timing correspondence to raw gait sequence $D_j = \{x_i\}$ of length T, are inputted into the frame-level module, and another positional encoding based on (2) are added to them for embedding timing information. In this frame-level temporal feature extractor, multi-layered self-attention blocks compute dynamic dependencies of a sequence of frames. Hence, spatiotemporal features $F_{st} \in \mathbb{R}^{d_t}$ of the sequence D_j are extracted successfully. In the final step, spatiotemporal features F_{st} are sent to a simple classifier, containing a connection layer and the softmax function, for the prediction result of the system.

D. Fine-tuned Training Strategy

For training, the model is optimized following fine-tuned training strategy. Two feature extractors are updated in two phases: 1) pre-training point-level spatial feature extractor by minimizing the frame-level prediction loss and 2) fine-tuning the whole model consisting of pre-trained point-level feature extractor, frame-level temporal feature extractor and final classifier for predicting gait sequence correctly.

For prediction, the sequence data of frames are input into the trained model in parallel, for effective spatiotemporal feature extraction as well as gait recognition directly. The recognition is an end-to-end process. Pre-training the point-level spatial feature extractor with the supervision of the frame labels can make spatial structure information inside frame data be excavated as sufficiently as possible, and it is also helpful for capturing temporal



(a) Confusion matrix on SDUgait Dataset (b) Confusion matrix on UNTIO Dataset (c) Confusion matrix on PDgait Dataset

Figure 3. The confusion matrix of the classification results on two normal and an abnormal gait dataset: (a) SDUgait Dataset of 52 subjects, (b) UNITO Skeleton Dataset of 20 subjects and (c) PDgait Dataset for 4 categories of PD severity diagnosis. Due to the page size, we select the subjects labeled 1-10 in (a) and (b) to show the recognition results which are 99.83% and 99.26% respectively, and the accuracy of PD severity diagnosis is 98.29% in (c).

features among frames of a sequence later. Furthermore, the end-to-end prediction process makes gait recognition efficient as well.

IV. EXPERIMENTS

In the experiments, our proposed multi-level fine-tuned Transformer network is compared with against several state-of-the-art methods and two submodules of it, on three normal and abnormal gait datasets. And the details of data and experiments are presented below.

A. Experimental Settings

1) Datasets: In this section, we give a brief description of three datasets in our experiment, including two normal skeleton-based datasets, SDUgait Dataset [26] and UNTIO Dataset [27], and an extensive abnormal gait dataset, PDgait Dataset [28].

a) SDUgait Dataset: This is a novel dataset based on the second generation Kinect (Kinect V2). Two Kinect V2 were used for simultaneously capturing silhouette images and corresponding 3D positions of 21 joins. The database includes 52 subjects, 28 males and 24 females. For each subject, 20 sequences with 8 different directions were recorded, with totally 1040 sequences in this datasets.

b) UNTIO Dataset: This includes 4D skeleton data of 20 joints coordinates acquired with the Microsoft Kinect sensor, which can track gait parameters in real-time. This dataset contains 20 subjects, where each of them was required for natural walking along a corridor 10 times, captured with 2 cameras (front view and rear view), for a total of 200 gait samples.

c) PDgait Dataset: This is an abnormal gait dataset about severity of Parkinson's disease (PD), containing the sequence data of vertical ground reaction force (VGRF) in Newton recorded by the 8 force sensors located under each foot. The subjects include 93 patients with idiopathic PD and 73 healthy controls, from three researchers (Ju, Ga, Si). The VGRF data of the subject are collected at a frequency of100 samples per second. And the collected gait data have been graded with severity label of PD according to two diagnostic criteria (H&Y and UPDRS).

2) Implementation details: the experiments are conducted with Pytorch and Python libraries. When training the module, different settings are employed on three datasets. For SDUgait dataset, we only select the skeleton data as inputs. The temporal length of the gait sequence T is set to 7, while the number of sampling points P in each frame and the feature dimension of each point Q are 21 and 3 respectively (21 joints, 3D coordinates). And the dropout in Transformer encoders is set to 0.3 to avoid over-fitting. For UNITO dataset, the dimension of the input in point-level spatial feature extractor is 21×3 (3D positions of 21 points), and the time step T in frame-level temporal feature extractor is set to 7. For the abnormal dataset PDgait, the shape of a training sample is $19 \times 1 \times 100$, indicating 19 points in a frame gait data and the time step of 100, which follows the sensor's frequency of 100 samples per second.

The other hyper-parameters of this model are set according to different datasets and the training process follows the fine-tuned strategy for good performance. The model is implemented with the system of GTX 3080 GPU, i7-11700 CPU and 12-GB RAM

B. Results on Normal Gait Datasets

In this subsection, we evaluate the proposed multi-level fine-tuned Transformer network on two normal gait datasets for identification tasks: SDUgait Dataset and UNTIO Dataset.

1) Experiments on the SDUgait Skeleton Dataset: This experiment aims to verify that our proposed model can perform well in 4D human motion dataset composed with sequences of 3D skeleton joints, which also shows certain advantages over the other methods applied to dataset.

In this dataset, we have 52 subjects for classification, and use 60% of the data for model training, 20% as validation dataset and 20% for final evaluation. The performance of trained network is shown in Fig 3. For lack of space, a partial confusion matrix is displayed with 10 categories. We can see that the model can realize effective recognition of gait data with considerable accuracy.

2) Experiments on the UNTIO Skeleton Dataset: This is an experiment on 3D skeletal sequence dataset with the purposes of verifying that multi-level fine-tuned Transformer outperforms other approaches for normal gait recognition.

In this dataset, data from 20 subjects were collected for identification. We randomly split the dataset following the

TABLE I. PERFORMANCE OF GAIT RECOGNITION COMPARED WITH ADVANCED MODELS ON THREE DATASETS

Methods		SDUgait			UNTIO			PD		
Deep Methods	*Others*	*acc*	*precision*	*recall*	*acc*	*precision*	*recall*	*acc*	*precision*	*recall*
CNN		92.85%	92.81%	93.24%	95.79%	95.73%	95.85%	89.82%	88.90%	90.88%
3D-CNN		91.06%	90.99%	91.42%	94.34%	94.24%	94.37%	-	-	-
GRU		95.30%	95.23%	95.47%	97.95%	97.19%	97.24%	92.81%	92.02%	92.50%
BiLSTM		96.01%	95.97%	96.14%	96.44%	96.37%	96.48%	91.70%	91.02%	91.57%
CNN+LSTM		94.51%	94.49%	94.80%	95.24%	95.16%	95.27%	97.26%	97.81%	97.08%
Transformer (Point-level)		95.34%	95.27%	95.56%	96.65%	96.72%	96.64%	97.82%	97.77%	97.14%
Multi-level Transformer		95.43%	95.47%	95.63%	96.87%	96.83%	96.84%	97.40%	97.49%	97.08%
Multi-level Fine-tuned Transformer		**99.83%**	**99.87%**	**99.89%**	**99.26%**	**99.23%**	**99.31%**	**98.29%**	**98.23%**	**98.12%**

same scheme. 20 percent of data are used for performance evaluation, and partial experiment results are shown in Fig. 3. The accuracy of test set is shown as 99.26% finally.

C. Results on Abnormal Gait Data

Furthermore, we evaluate the proposed method on the PD dataset for abnormal gait diagnosis [29].

In the PD Dataset, we use the whole dataset collected by all three institutions to classify PD severity. The severity levels of subjects were scored according to multiple scales in advance, and we use the criteria of H&Y scale as supervision information, containing three categories of PD severities and a healthy control group. The testing performance of trained method is shown in Fig 3. It is shown that the model is 98.72% for healthy controls and 97.53% for severely patients. It presents that the proposed method can be expanded to abnormal gait data for classifying gait pattern other than the subject identity.

D. Compare with Deep Methods

Furthermore, we compare the results on these normal and abnormal datasets with some deep methods for learning spatiotemporal information from them, i.e., CNN, 3D-CNN, modified recurrent networks and CNN+LSTM, for spatial or temporal feature extraction. To be fair, we experimented with the same data split scheme for the same dataset.

The quantitative results are summarized in Table I. For identification tasks based on two human normal skeleton datasets, the proposed method yielded a higher accuracy than other deep methods. For the abnormal PDgait Dataset, the performance of the model's classifying abnormal patterns is not inferior to them, as well. These promising results illustrated Transformer architecture's capacity of representing both spatial and temporal information, as well as the potential to be applied for exploiting the global spatiotemporal features from complicated 4D gait sequence.

E. Performance of the Whole Model

To further verify the validity of the proposed model, we evaluate the performance of the whole model and its internal components in this subsection.

The successful construction of the proposed multi-level fine-tuned Transformer explained in Section III, highly relies on multi-level Transformer architecture and fine-tuned training strategy. So experiments have been conducted to investigate the individual contributions of two subsections: 1) Transformer (point-level) and 2) Multi-level Transformer (point- and frame- level) without following fine-tuned strategy. Table I shows the experimental results on three datasets. The results indicate that both two components are necessary for the proposed method.

It can be seen that the average output of the Transformer for point-level feature extraction is not inferior to those excellent deep learning method, which demonstrates that the points' structure information related to gait features of a frame have been explored fully. Besides, the results of second subsection shows a slight over Spatial Transformer, due to finitude of information for gait representations contained in a certain moment. Finally, the performance of our proposed method has relatively significant improvements in all three datasets, with the only discrepancy of training strategy, indicating the important role it plays in integrating the characteristics of temporal and spatial dimension, as well as grasping the proper training direction for the model optimization. What's more, this kind of training strategy acting on the model performance can be seen as a combination of several smaller models that already have good results, to avoid the interference of individual outliers and enhance the model's robustness to some extent.

V. CONCLUSION

In this paper, we reported our investigation of gait recognition in time series data, involving with two normal gait recognition tasks and a severity rating of the abnormal gait sequence dataset. We proposed a multi-level fine-tuned Transformer network for exploring spatiotemporal representation characteristics of gait sequences. The proposed method yields promising performance on three datasets, and the constituting components in it play an individual role. Transformer networks have the ability to pass through spatial and temporal feature extraction. Multi-level feature extractor helps us to connect point-level spatial and frame-level temporal characteristics from sequence data of frames, and integrate into the global spatiotemporal representations. Meanwhile, fine-tuning strategy use frame labels as auxiliary supervision to ensure efficient spatial structure exploration within the frame. Moreover, extensive experiments were conducted on several submodules of the proposed model. Experiment results demonstrated crucial

contributions of multi-level feature extraction and fine-tuned strategy for model effect.

Since the subtle differences in abnormal gait patterns are less significant than subject-specific differences, the performance of our proposed method on abnormal gait dataset is relatively weak. This model will be extended to address the multi-modal data for the improvement of the accuracy of abnormal feature recognition.

ACKNOWLEDGMENT

This research was supported in part by National Natural Science Foundation of China under Grant No. 62106117, and Natural Science Foundation of Shandong Province under Grant No.ZR2021QF084.

REFERENCES

[1] M. Luo, J. Cao, X. Ma, X. Zhang, and R. He, "Face Recognition _FA-GAN: Face Augmentation GAN for Deformation-Invariant Face Recognition," *IEEE Transactions on Information Forensics and Security,* vol. 16, 2021, pp. 2341-2355, doi: 10.1109/tifs.2021.3053460.

[2] X. Yin, Y. Zhu, and J. Hu, "Fingerprint Recognition_3D Fingerprint Recognition based on Ridge-Valley-Guided 3D Reconstruction and 3D Topology Polymer Feature Extraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 43, no. 3, 2021, pp. 1085-1091, doi: 10.1109/tpami.2019.2949299.

[3] L. Fei, B. Zhang, L. Zhang, W. Jia, J. Wen, and J. Wu, "Palmprint Recognition_Learning Compact Multifeature Codes for Palmprint Recognition From a Single Training Image per Palm," *IEEE Transactions on Multimedia,* vol. 23, 2021, pp. 2930-2942, doi: 10.1109/tmm.2020.3019701.

[4] A. Sepas-Moghaddam and A. Etemad, "Deep Gait Recognition: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2022, pp. 1-1, doi: 10.1109/tpami.2022.3151865.

[5] M. D. Marsico and A. Mecca, "A Survey on Gait Recognition via Wearable Sensors," *ACM Computing Surveys,* vol. 52, no. 4, 2020, pp. 1-39, doi: 10.1145/3340293.

[6] M. Tölgyessy, M. Dekan, Ľ. Chovanec, and P. Hubinský, "Evaluation of the Azure Kinect and Its Comparison to Kinect V1 and Kinect V2," *Sensors,* vol. 21, no. 2, 2021, p. 413, doi: 10.3390/s21020413.

[7] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3D convolutional neural networks," in *2016 IEEE Inte*rnational Conference on Image Processing (ICIP), 2016: IEEE, doi: 10.1109/icip.2016.7533144.

[8] A. Vaswani *et al*., "Attention is all you need," in Advances in neural information processing systems, 2017, vol. 30, pp. 5998-6008, doi: https://doi.org/10.48550/arXiv.1706.03762.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2019, doi: https://doi.org/10.48550/arXiv.1810.04805.

[10] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929,* 2020, doi: https://doi.org/10.48550/arXiv.2010.11929.

[11] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012-10022.

[12] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, no. 2, 2006, pp. 316-322, doi: 10.1109/tpami.2006.38.

[13] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, no. 2, 2017, pp. 209-226, doi: 10.1109/tpami.2016.2545669.

[14] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait recognition invariant to carried objects using alpha blending generative adversarial networks," *Pattern Recognition,* vol. 105, 2020, p. 107376, doi: https://doi.org/10.1016/j.patcog.2020.107376.

[15] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition," *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 33, 2019, pp. 8126-8133, doi: 10.1609/aaai.v33i01.33018126.

[16] C. Fan *et al.*, "GaitPart: Temporal Part-Based Model for Gait Recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020: IEEE, doi: 10.1109/cvpr42600.2020.01423.

[17] N. Li and X. Zhao, "A Strong and Robust Skeleton-based Gait Recognition Method with Gait Periodicity Priors," *IEEE Transactions on Multimedia,* 2022, pp. 1-1, doi: 10.1109/tmm.2022.3154609.

[18] A. Zhao, Y. Wang, and J. Li, "Transferable Self-Supervised Instance Learning for Sleep Recognition," *IEEE Transactions on Multimedia,* 2022, pp. 1-1, doi: 10.1109/tmm.2022.3176751.

[19] Y. Wang, Z. Lv, Z. Sheng, H. Sun, and A. Zhao, "A deep spatio-temporal meta-learning model for urban traffic revitalization index prediction in the COVID-19 pandemic," *Advanced Engineering Inf*ormatics, vol. 53, 2022, p. 101678, doi: https://doi.org/10.1016/j.aei.2022.101678.

[20] P. Nithyakani, A. Shanthini, and G. Ponsam, "Human Gait Recognition using Deep Convolutional Neural Network," in *2019 3rd International Conference on Computing and Communications Technologies (ICCCT)*, 2019: IEEE, doi: 10.1109/iccct2.2019.8824836.

[21] T. Huang, X. Ben, C. Gong, B. Zhang, R. Yan, and Q. Wu, "Enhanced Spatial-Temporal Salience for Cross-view Gait Recognition," *IEEE Transactions on Circuits and Systems for Video Technology,* 2022, pp. 1-1, doi: 10.1109/tcsvt.2022.3175959.

[22] Y. Zhang, Y. Huang, S. Yu, and L. Wang, "Cross-View Gait Recognition by Discriminative Feature Learning," *IEEE Transactions on Image Processing,* vol. 29, 2020, pp. 1001-1015, doi: 10.1109/tip.2019.2926208.

[23] Y. Xu, W. Yang, M. Chen, S. Chen, and L. Huang, "Attention-Based Gait Recognition and Walking Direction Estimation in Wi-Fi Networks," *IEEE Transactions on Mobile Computing,* vol. 21, no. 2, 2022, pp. 465-479, doi: 10.1109/tmc.2020.3012784.

[24] Y. Xia, Z. Yao, Q. Ye, and N. Cheng, "A Dual-Modal Attention-Enhanced Deep Learning Network for Quantification of Parkinson's Disease Characteristics," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 28, no. 1, 2020, pp. 42-51, doi: 10.1109/tnsre.2019.2946194.

[25] Y. Xu *et al.*, "Transformers in computational visual media: A survey," *Computational Visual Media,* vol. 8, no. 1, 2022, pp. 33-62, doi: 10.1007/s41095-021-0247-3.

[26] Q. Li *et al.*, "Classification of gait anomalies from kinect," *The Visual Computer,* vol. 34, no. 2, 2018, pp. 229-241, doi: 10.1007/s00371-016-1330-0.

[27] E. Gianaria, M. Grangetto, M. Lucenteforte, and N. Balossino, "UNITO data," in *Biometric Authentication*: Springer International Publishing, 2014, pp. 16-27.

[28] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet," *Circulation,* vol. 101, no. 23, 2000, pp. e215-e220, doi: 10.1161/01.cir.101.23.e215.

[29] A. Zhao *et al.*, "Multimodal Gait Recognition for Neurodegenerative Diseases," *IEEE Transactions on Cybernetics,* 2021, pp. 1-15, doi: 10.1109/tcyb.2021.3056104.