Applied Intelligence (2023) 53:1535–1547
https://doi.org/10.1007/s10489-022-03543-y

# TransGait: Multimodal-based gait recognition with set transformer

**Guodong Li[1] · Lijun Guo[1] · Rong Zhang[1] · Jiangbo Qian[1] · Shangce Gao[2]**

Accepted: 22 March 2022 / Published online: 29 April 2022


**Abstract**

As a biological feature that can be recognized from a distance, gait has a wide range of applications such as crime prevention, judicial identification, and social security. However, gait recognition is still a challenging task with two problems in the typical gait recognition methods. First, the existing gait recognition methods have weak robustness to the pedestrians' clothing and carryings. Second, the existing temporal modeling methods for gait recognition fail to fully exploit the temporal relationships of the sequence and require that the gait sequence maintain unnecessary sequential constraints. In this paper, we propose a new multi-modal gait recognition framework based on silhouette and pose features to overcome these problems. Joint features of silhouettes and poses provide high discriminability and robustness to the pedestrians' clothing and carryings. Furthermore, we propose a set transformer model with a temporal aggregation operation for obtaining set-level spatio-temporal features. The temporal modeling approach is unaffected by frame permutations and can seamlessly integrate frames from different videos acquired in different scenarios, such as diverse viewing angles. Experiments on two public datasets, CASIA-B and GREW, demonstrate that the proposed method provides state-of-the-art performance. Under the most challenging condition of walking in different clothes on CASIA-B, the proposed method achieves a rank-1 accuracy of 85.8%, outperforming other methods by a significant margin (>4%).

**Keywords** Gait recognition · Multi-modal · Transformer

## 1 Introduction

Gait recognition is a human recognition technology based on a walking pattern of a person. Compared with other human biometric information used in human recognition methods, such as fingerprint, iris, and face, gait information is easy to obtain, hard to fake, and suitable for long-distance human recognition. Due to such advantages, it has been an active research topic in the fields of biometrics and computer vision, targeting a wide application perspective in public security and crime investigation. Most of the existing gait recognition methods extracted gait features from a human silhouette. Especially with the development of the deep convolutional network, the silhouette sequence-based method has been widely studied and used. Silhouette sequence has low computation cost but can effectively describe the gait of a person. However, the recognition accuracy is significantly affected by various external factors such as clothing and carrying conditions [1, 2]. For example, a recent state-of-the-art method, MT3D [3], achieved accuracies of 96.7% under different viewpoints with a normal walking condition on the CASIA-B gait dataset [4]. However, the accuracy was dropped to 81.5% for a clothing changes condition.

In order to reduce the influence of clothing and carrying conditions on gait recognition, we propose a multi-modal gait recognition method combining silhouettes and pose heatmaps. The silhouettes and pose heatmaps describe the pedestrian from different perspectives. The

✉ Lijun Guo
guolijun@nbu.edu.cn

Guodong Li
liguodong0520@foxmail.com

Rong Zhang
zhangrong@nbu.edu.cn

Jiangbo Qian
qianjiangbo@nbu.edu.cn

Shangce Gao
gaosc@eng.u-toyama.ac.jp

[1] Faculty of Electrical Engineering and Computer Science, NingBo University, Ningbo, China

[2] University of Toyama, Toyama, Japan

silhouette sequence describes the changes in pedestrian appearance during the gait cycle and contains rich pedestrian information. Hence, the silhouette gait features have strong discriminability. However, the silhouettes are susceptible to interference from the pedestrians' clothing and belongings, significantly affecting gait recognition accuracy. In contrast, the pose sequence describes the changes of the pedestrian's internal joints in the gait cycle. Thus, it does not contain the interference information of the pedestrian's clothing and carrying and is robust to the cloth change and carryings [5]. As shown in Fig. 1a, the silhouettes of the same pedestrian in different clothing conditions are considerably different because of the clothing information, but the pose heatmaps are similar across different clothing conditions. However, the pose heatmaps contain less information and are insufficient to distinguish different pedestrians. As shown in Fig. 1b, the pose heatmaps are very similar for different pedestrians under the same walking condition, but the silhouettes are of significant difference. It indicates that the silhouette and pose information are supplementary and can be combined to describe pedestrian gait accurately. The silhouettes have rich appearance information that is useful to distinguish different pedestrians, thus increasing inter-class discrimination. The pose heatmaps are robust to the changes of clothing and carryings. Accordingly, the influence of interference information is reduced in gait recognition, reducing the intra-class difference. Experiments on the CASIA-B dataset and the GREW demonstrate that the combination of silhouette and pose heatmap can improve the accuracy of gait recognition, and the multi-modal method is required.

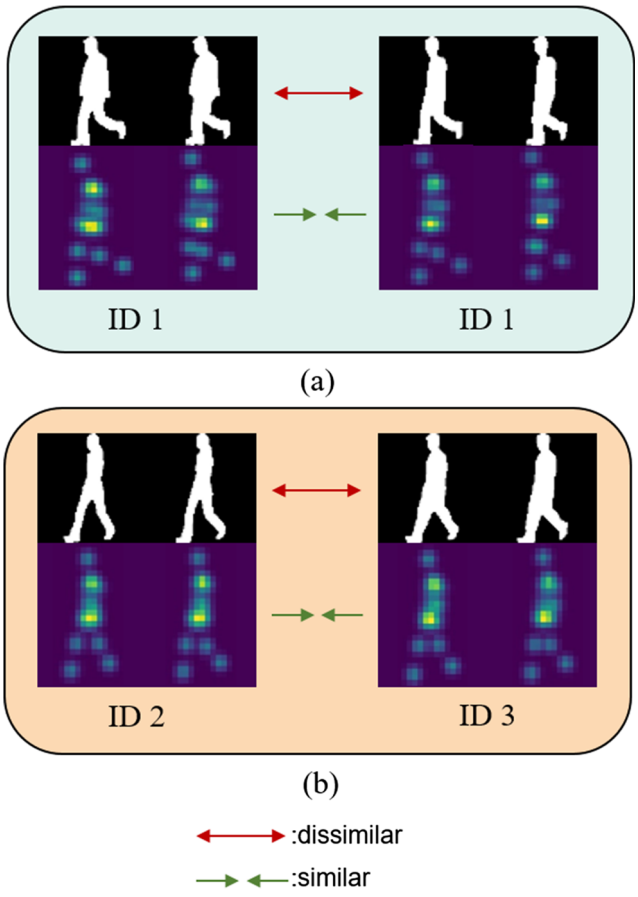


**Fig. 1** Examples of silhouettes (top) and poses (bottom).(a) The same pedestrian in different clothing conditions: the silhouettes are distinct, but the pose heatmaps are similar. (b) The different pedestrians: the pose heatmaps are similar across different pedestrians, but the silhouettes are dissimilar

Temporal modeling is one of the key tasks in gait recognition since the gait is inherently of motion. The gait was commonly temporally modeled using LSTM and 3DCNN in the existing methods. The LSTM can model the long-term temporal feature in the gait cycle. However, the LSTM cannot be trained in parallel. On the other hand, 3DCNN often requires a large number of parameters. Fan et al. [6] selected short-term temporal features as the most discriminative features to model human gait. However, only short-term temporal information is insufficient to extract discriminative characteristics of human gait. Although above methods preserve more temporal information, a significant degradation could be induced from discontinuous input frames and a different frame rate. This is because these methods retain unnecessary order constraints. Thus, we introduce the set transformer module (STM) into the gait recognition framework to model motion patterns on various time scales. First, STM imposes no constraints on the order of elements of the gait sequence to enable modeling interactions among gait frames under different viewpoints. Second, STM adaptively learns different motion patterns contained in the gait sequence, including short-, medium-, and long-term temporal information of the gait cycle. Each multi-head attention operator in the transformer focuses on a different movement pattern. Our main contributions are summarized as follows:

- We combine silhouettes and pose heatmaps to mine the robust and discriminative gait feature of a pedestrian. We construct part-based multi-modal features generated by assembling split deep features derived from silhouettes and pose heatmaps. Those multi-modal features corresponding to a specific part describe the part-level motion characteristics in a walking period.
- We propose the STM that is a novel temporal modeling module for gait recognition. The multi-modal feature sequence corresponding to a part is input into an STM to extract multiple motion features for gait recognition. The proposed STM network fuses multi-modal visual information, part-based fine-grained features, and temporal relativity of gait sequences.

Unlike the transformer model used in other vision tasks, STM is flexible due to its robustness to frame permutations.
- The proposed method outperforms the state-of-the-art gait recognition methods for the CASIA-B and GREW datasets.

## 2 Related works

### 2.1 Body representation in gait recognition

In terms of human body representation, gait recognition can be divided into silhouette-based methods [3, 6–13] and pose-based methods [14–19]. The silhouette has been the most commonly used body representation in literature. The silhouettes can effectively describe the changes in pedestrians' appearance during the gait period due to the advantage that unrelated information, such as color, to gait recognition is not described in the silhouettes [2]. However, it is very sensitive to the change of clothing and carryings because it describes pedestrians' appearance. Pose-based gait recognition methods commonly adopted 3D skeleton as human body representation since 3D skeleton is not easily affected by the clothing and carryings. However, there are two problems in the 3D pose estimation method: i) The 3D skeleton-based method heavily relies on accurate detection of body joints and is more sensitive to occlusion. ii) The 3D skeleton only describes the changes of the body joints in the gait period, which cannot fully reflect the pedestrians' gait.

In recent years, 2D pose estimation has made a great progress with the development of deep learning. Since pose information is of great importance in human gait recognition, 2D pose is a more feasible and lower cost technical solution than 3D pose. Feng et al. [20] used the human body joint heatmap extracted from a RGB image to extract temporal feature. However, the recognition rates of using only poses are not satisfactory when the silhouettes are completely ignored. Li et al. [21] integrated 3D joints, 2D joints and silhouettes of the human body. This method achieved state-of-the-art results but is relatively complex. Zhao et al. [22] extracted the unimodal gait features of silhouettes and poses respectively, rather than concatenating silhouettes and poses as a multimodal body representation to extract multimodal gait features. In this work, we are aiming at addressing the robustness of gait recognition to clothing and carryings. We propose a multi-modal gait recognition method using silhouette-pose body representation. The silhouette-pose body representation is more comprehensive to describe the change of pedestrian's gait. It is also robust to the change of pedestrian's clothing and carryings. In this paper, we choose 2D pose heatmaps to describe the change of pedestrian joints. Since the 2D pose heatmap is a probability map of human body joints, it is more robust to pose estimation error than the 3D skeleton.

### 2.2 Temporal representation in gait recognition

Temporal representation in gait recognition can be divided into template-based methods and sequence-based methods. The template-based approaches aggregated gait information into a single image using statistical functions, which can be divided into two sub-categories: temporal template and convolutional template. Temporal template aggregated gait information before inputting to the network, such as gait energy image (GEI) [23] and gait entropy image (GENI) [24]. The convolutional template aggregated the gait information after several layers of convolution and pooling operations, including set pooling [9] and gait convolution energy map (GCEM) [25]. Sequence-based methods learned the temporal relationship in gait sequences instead of aggregating them. Sequence-based methods can be divided into three sub-categories: LSTM-based methods [7, 25, 26], 3DCNN-based methods [3, 27] and micro-motion based methods [6]. Zhang et al. [7] divided the human body into several parts, where each part extracted spatio-temporal features of gait using the LSTM temporal attention model. Lin et al. [3] proposed a multi-time scale 3DCNN (MT3D) model, which improved the 3D pooling layer to aggregate the time information of each local temporal fragment. Fan et al. [6] proposed a micro-motion capture module (MCM), which consisted of a micro-motion template builder and a temporal pooling module. The micro-motion template generator uses the attention mechanism and statistical function to aggregate local adjacent frames and obtains several local micro-motion templates. Then, those micro-motion templates were aggregated to obtain gait features via a temporal pooling module. This method proves that micro-motion is effective for gait recognition. However, the micro-motion patterns are only considered without consideration of other movement patterns in this method. For example, the relationship between the starting motion and the future landing motion is beneficial to gait recognition. Therefore, we use the set transformer module to model interactions among elements in the input set, where each head in Multi-head attention in the transformer learns different motion patterns in gait sequences and then aggregate these motion pattern features for gait recognition.

### 2.3 Transformer

Transformer showed outstanding performance for sequence-based tasks, especially for natural language processing (NLP) tasks [28, 29]. It was originally designed to solve the problem that RNN cannot be trained in parallel [30]. The transformer consisted of a self-attention module

and a feed-forward neural network. The self-attention module learned the relationship between any two frames in an attention mechanism, providing better parallelism. Multi-head attention was composed of multiple self-attention. Each head extracted sequence features of different patterns, which helps capture richer sequence information. The transformer has been used in many computer vision tasks, such as action recognition [31, 32], and frame synthesis [33]. In recent years, the transformer has also been used for image spatial feature extraction [34, 35]. Dosovitskiy et al. [34] introduced a transformer instead of CNN for image space modeling for the first time. Liu et al. [35] proposed a hierarchical transformer structure based on the shifted windowing scheme, which had the flexibility to model at various scales with linear computational complexity with respect to image size. Yao et al. [36] used the transformer to model the spatial relationship of pedestrian joints in gait recognition. In this paper, we use the transformer for temporal modeling in gait recognition. As a permutation invariant attention-based neural network module, the STM is proposed to learn and aggregate different motion patterns in the gait cycle.

## 3 Proposed method

The overall structure of the proposed gait recognition model is depicted in Fig. 2. First, the silhouettes and pose heatmaps are obtained from the input gait sequence. Then, they are fed into the corresponding feature extraction modules, denoted as $E_s$ and $E_p$, to extract frame-level features. Then, the silhouette and pose feature maps are concatenated to get the silhouette-pose multi-modal frame-level body features. The multi-modal frame-level body features are horizontally split into part-level features by a Horizontal Pooling (HP) module. For each part, we use an STM to extract movement patterns on different time scales of the gait sequence and obtain spatio-temporal fine-grained features through temporal aggregation. Finally, the extracted set-level part motion features are used to recognize human gait.

### 3.1 Pipeline

Let denote the RGB image sequence of the subjects in the data set as $\{I_i \mid i = 1, \ldots, t\}$ where $t$ is the number of frames in the sequence. The background subtraction method and the pre-trained pose estimation network (CPM) [37] are used to extract the corresponding silhouette sequence and 2D pose heatmap sequence from the RGB image sequence, respectively, denoted as $\{S_i \mid i = 1, \ldots, t\}$ and $\{P_i \mid i = 1, \ldots, t\}$. Then, we extract the spatial features of the silhouette and 2D pose heatmap sequences by $E_s$ and $E_p$.

$$s_i = E_s(S_i) \tag{1}$$

$$p_i = E_p(P_i) \tag{2}$$

The silhouette feature map $s_i$ and the pose feature map $p_i$ are concatenated to obtain the silhouette-pose multi-modal feature map $m_i$, as follows:

$$m_i = [s_i, p_i] \tag{3}$$

where $[\cdot]$ represents the concatenate operation. The multi-modal feature is taken as the body representation feature, which is more robust to the pedestrians' cloth and carryings and stronger discriminability, respectively, than the silhouette feature and the pose feature.

Recent person re-identification methods generated deep representation from local parts for fine-grained discriminative features of a person [38–40]. Inspired by these works, we use the Horizontal Pooling (HP) module to extract the

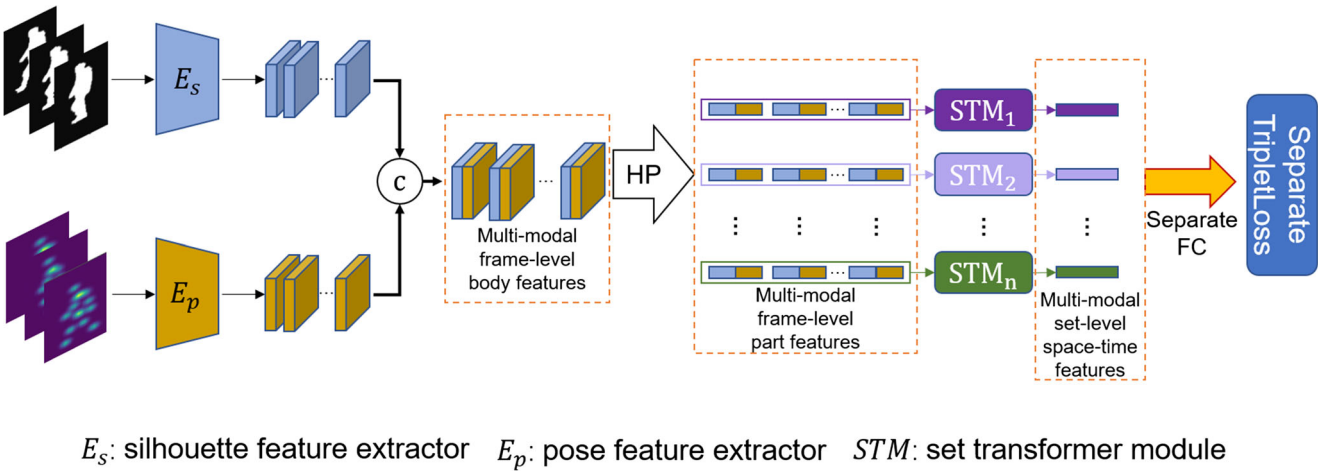


**Fig. 2** The overall framework of TransGait. The $E_s$ and $E_p$ represent silhouette feature extractor and pose feature extractor, respectively. c denotes concatenate operation. The HP represents Horizontal Pooling and STM represents set transformer module

discriminative part-informed features of the partial human body. As shown in Fig. 3, the HP module horizontally splits the multi-modal feature map $m_i$ into $n$ parts (we choose $n = 16$ in the experiment). Then, the HP module downsamples each part of $m_i$ by a global average and max pooling to generate column feature vector, $mp_{j,i}$.

$$mp_{j,i} = \text{Avgpool2d}\left(m_{j,i}\right) + \text{Maxpool2d}\left(m_{j,i}\right) \tag{4}$$

where $j \in 1, 2, \ldots, n$. We transform the multi-modal feature sequence into n part-level feature vectors and get the multi-modal part representation matrix $MP = \left(mp_{j,i}\right)_{n\times t}$. The corresponding row vector of the multi-modal part representation matrix is denoted as $MP_{j,\cdot} = \left\{mp_{j,i} \mid i = 1, \ldots, t\right\}$. Then, for part $j$ of MP, the STM extracts the set-level spatio-temporal features $v_j$. Note that STM does not require strictly sequential inputs, and the same output can be obtained even with mess-up inputs.

$$v_j = \text{STM}_j\left(MP_{j,\cdot}\right) \tag{5}$$

Finally, we use several separate FC Layers to map the feature vectors extracted from the STM to the metric space for gait recognition.

### 3.2 Multi-head attention

As an integral component of the transformers, the self-attention mechanism explicitly models the interactions between all entities of a sequence. The self attention is defined on receiving the tuple input (query, key, value) and performs the scaled dot-product as

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_q}}\right)V \tag{6}$$

where $\mathbf{Q} = \mathbf{X}\mathbf{W}^Q$ ($\mathbf{W}^Q \in \mathbb{R}^{n\times d_q}$), $\mathbf{K} = \mathbf{X}\mathbf{W}^K$ ($\mathbf{W}^K \in \mathbb{R}^{n\times d_k}$), $\mathbf{V} = \mathbf{X}\mathbf{W}^V$ ($\mathbf{W}^V \in \mathbb{R}^{n\times d_v}$) and $\mathbf{X}$ represents input sequence embeddings.

The multi-head attention comprises multiple self-attention blocks, where each self-attention head seeks different relationships among the sequence elements. The multi-head attention module is formulated as follows:

$$\text{MultiHead}\left(Q, K, V\right) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right)W^0 \tag{7}$$

$$\text{where head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{8}$$

### 3.3 Set transformer module (STM)

In this paper, we propose the STM, an attention-based module that extends the standard transformer network [28] to learn interactions between elements in the gait set. Note that, we use the temporal pooling (TP) suitable for gait recognition tasks for feature aggregation instead of [cls] tokens used by other transformers. And unlike the original transformers, the positional embedding is not added to the input. According to Gaitset [9], the silhouettes and pose heatmaps of each position in the gait sequence have a unique appearance and therefore contain their position information themselves.

As shown in Fig. 4, the STM consists of three sub-modules: the multi-head attention block (MAB), feed-forward module, and temporal pooling module. The MAB utilizes the multi-head attention mechanism to find different motion patterns of gait sequences on the time scales, which is formulated as follows:

$$\text{MAB}\left(X\right) = \text{MultiHead}\left(XW_{Q_T}, XW_{K_T}, XW_{V_T}\right) \tag{9}$$

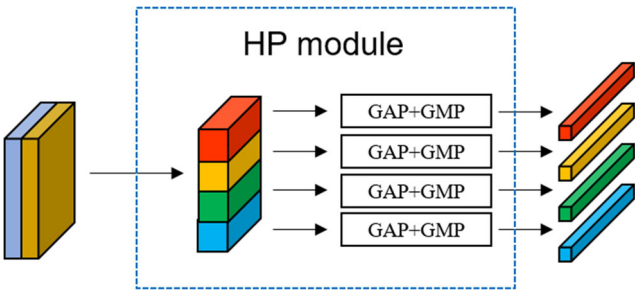


**Fig. 3** The structure of the HP ($n = 4$ as an example)

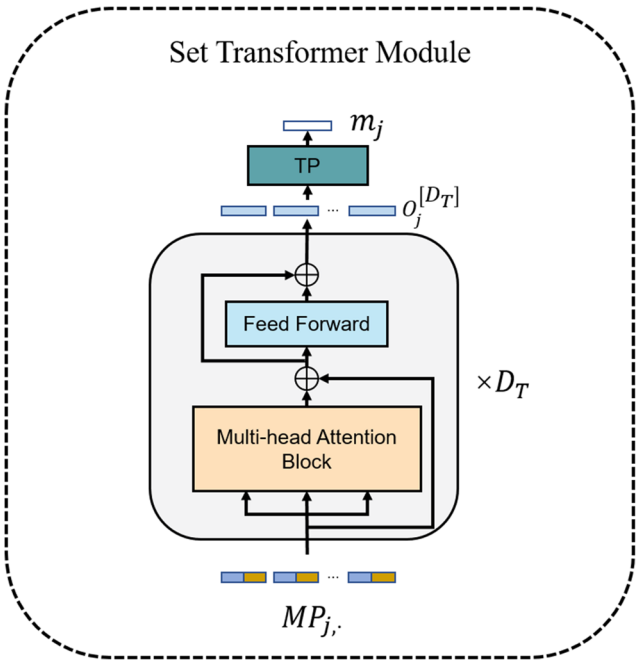


**Fig. 4** The block diagram of the STM

The feed-forward module contains a layer of $MLP$ and $ReLU$ activation functions. The temporal pooling module extracts the most discriminative motion feature vectors in the sequence, where $\max(\cdot)$ is used as the instance function of temporal pooling. The set transformer is formulated as follows:

$$O_{j,\cdot}^{[0]} = MP_j, \tag{10}$$

$$\hat{O}_{j,\cdot}^{[i]} = MAB_j(O_{j,\cdot}^{[i-1]}) + O_{j,\cdot}^{[i-1]} \tag{11}$$

$$O_{j,\cdot}^{[i]} = \hat{O}_{j,\cdot}^{[i]} + \mathrm{ReLU}\left(f_j^{\theta}\left(\hat{O}_{j,\cdot}^{[i]}\right)\right) \tag{12}$$

$$v_j = \mathrm{TP}\left(O_{j,\cdot}^{[D_T]}\right) = \max\left(O_{j,\cdot}^{[D_T]}\right) \tag{13}$$

where $f_j^{\theta}$ represents the feed-forward module corresponding to part $j$, and $\theta$ is the parameter. $D_T$ is the number of layers in the set transformer.

### 3.4 Implementation details

**Network hyper-parameters.** The $E_s$ and $E_p$ have the same structure but different parameters, which are composed of three convolution modules. Each convolution module comprises two 3×3 convolutional layers, a max-pooling layer [41], and a Leaky ReLU activation. The part number $n$ of the HP module is set to 16. The number of layers in STM is set to 2 and the number of heads is set to 8. STM can extract discriminative temporal features without deep stacking due to the advantage of the set transformer that can observe the whole sequence at the low-level layer. The ablation study for setting the hyper-parameter $D_T$ is discussed in Section 4.4.

**Loss and Sampler.** The separate batch all (BA+) triplet loss [42] is adopted to train the network. The corresponding column feature vectors among different samples are used to compute the loss. The batch size is set to (p,k), where p indicates the number of persons and k indicates the number of samples for each person in a batch.

**Testing.** At the test phase, the distance between gallery and probe is defined as the average euclidean distance of the corresponding feature vectors.

## 4 Experiments

### 4.1 Datasets and evaluation protocol

**CASIA-B** [4] is the most widely used gait data set, including RGB images and silhouettes of 124 subjects. An example of subjects in CASIA-B is shown in Fig. 5. Each subject contains 11 views, and each view contains ten sequences. The ten sequences are obtained under three different walking conditions; the first six sequences are obtained under normal conditions (NM), the second two sequences contain subjects carrying a bag (BG), and the last two sequences contain subjects wearing a coat or jacket (CL).

In our experiments, we use average Rank-1 accuracies to evaluate the performance of the gait recognition methods, excluding identical-view cases. Rank-1 accuracy indicates the probability of the correct matches in the first trial within galleries.

**GREW** [43] is the latest and most complex gait dataset. GREW is a more challenging dataset due to being constructed in the wild. It consists of 26,345 subjects and 128,671 sequences, which come from 882 cameras in wild environments. Moreover, GREW provides silhouettes and human poses data. This dataset is selected to verify the effectiveness of our method in gait recognition in the wild.

### 4.2 Training details

**1)** Common configuration: The input size of the silhouettes was 64×44. The joints with confidence greater than 0.1 synthesize the 2D pose heatmaps. We then cropped and resized the 2D pose heatmaps to match the size of the silhouettes. We randomly took 30 frames of silhouettes and their corresponding pose heatmaps from the sequence during each training epoch. Adam optimizer was used with

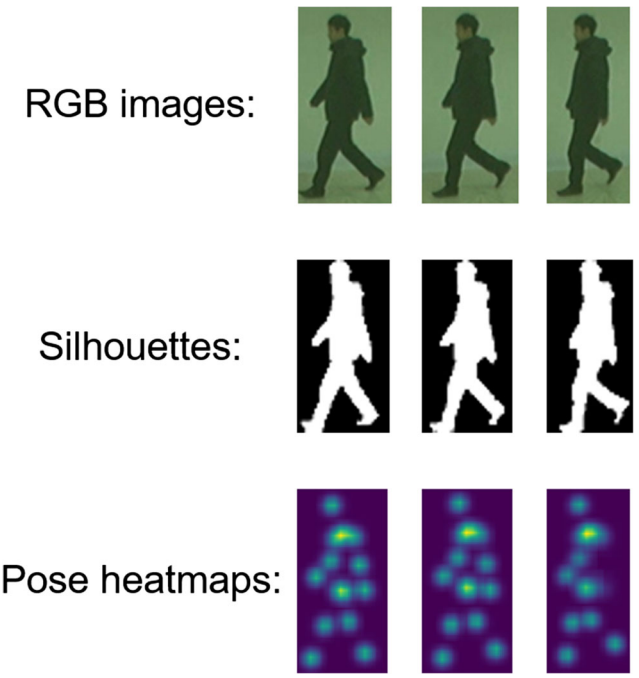


**Fig. 5** An example of RGB images, silhouettes and pose heatmaps from **CASIA-B**

**Table 1** Averaged Rank-1 accuracies on CASIA-B, excluding identical-view cases

| Gallery NM#1-4 | | 0°-180° | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Probe | | 0° | 18° | 36° | 54° | 72° | 90° | 108° | 126° | 144° | 162° | 180° | Mean |
| NM#5-6 | GaitSet [9] | 90.8 | 97.9 | 99.4 | 96.9 | 93.6 | 91.7 | 95.0 | 97.8 | 98.9 | 96.8 | 85.8 | 95.0 |
| | GaitPart [6] | 94.1 | 98.6 | 99.3 | 98.5 | 94.0 | 92.3 | 95.9 | 98.4 | 99.2 | 97.8 | 90.4 | 96.2 |
| | GLN [8] | 93.2 | 99.3 | 99.5 | 98.7 | 96.1 | **95.6** | 97.2 | 98.1 | 99.3 | 98.6 | 90.1 | 96.9 |
| | MT3D [3] | 95.7 | 98.2 | 99.0 | 97.5 | 95.1 | 93.9 | 96.1 | 98.6 | 99.2 | 98.2 | 92.0 | 96.7 |
| | TransGait(ours) | **97.3** | **99.6** | **99.7** | **99.0** | **97.1** | 95.4 | **97.4** | **99.1** | **99.6** | **98.9** | **95.8** | **98.1** |
| BG#1-2 | GaitSet [9] | 83.8 | 91.2 | 91.8 | 88.8 | 83.3 | 81.0 | 84.1 | 90.0 | 92.2 | 94.4 | 79.0 | 87.2 |
| | GaitPart [6] | 89.1 | 94.8 | 96.7 | 95.1 | 88.3 | 84.9 | 89.0 | 93.5 | 96.1 | 93.8 | 85.8 | 91.5 |
| | GLN [8] | 91.1 | 97.7 | **97.8** | 95.2 | 92.5 | 91.2 | 92.4 | **96.0** | **97.5** | 94.9 | 88.1 | 94.0 |
| | MT3D [3] | 91.0 | 95.4 | 97.5 | 94.2 | 92.3 | 86.9 | 91.2 | 95.6 | 97.3 | 96.4 | 86.6 | 93.1 |
| | TransGait(ours) | **94.0** | **97.1** | 96.5 | **96.0** | **93.5** | **91.5** | **93.6** | 95.9 | 97.2 | **97.1** | **91.6** | **94.9** |
| CL#1-2 | GaitSet [9] | 61.4 | 75.4 | 80.7 | 77.3 | 72.1 | 70.1 | 71.5 | 73.5 | 73.5 | 68.4 | 50.0 | 70.4 |
| | GaitPart [6] | 70.7 | 85.5 | 86.9 | 83.3 | 77.1 | 72.5 | 76.9 | 82.2 | 83.8 | 80.2 | 66.5 | 78.7 |
| | GLN [8] | 70.6 | 82.4 | 85.2 | 82.7 | 79.2 | 76.4 | 76.2 | 78.9 | 77.9 | 78.7 | 64.3 | 77.5 |
| | MT3D [3] | 76.0 | 87.6 | 89.8 | 85.0 | 81.2 | 75.7 | 81.0 | 84.5 | 85.4 | 82.2 | 68.1 | 81.5 |
| | TransGait(ours) | **80.1** | **89.3** | **91.0** | **89.1** | **84.7** | **83.3** | **85.6** | **87.5** | **88.2** | **88.8** | **76.6** | **85.8** |

Bold data the results with the highest recognition rate among the results of different methods

the learning rate of 1e-4. The margin of the triplet loss was set to 0.2. **2)** In CASIA-B: the batch size was set as (4, 16) following the manner introduced in Section 3.4, and the number of training epochs was 80K. **3)** In GREW: the batch size was set to (64, 4), the iterations was set to 200K, and the learning rate would be reduced to 1e-5 at 150k iterations.

### 4.3 Experimental result

**CASIA-B.** The proposed method is evaluated in comparison with several state-of-the-art gait recognition methods on the CASIA-B dataset, including GaitSet [9], GaitPart [6], GLN [8], and MT3D [3]. In order to make a systematic and comprehensive comparison, all the cross-view and cross-walking-condition cases were included in the comparison scope. As shown in Table 1, the proposed method outperforms the compared methods. For NM, the proposed method provides a 1.3% higher average accuracy than the best among compared methods, MT3D, whose accuracy is 96.7%. For BG, the proposed method achieves a 1% higher average accuracy than the GLN, whose accuracy is 94.04%. For CL, the average accuracy of the proposed method is 4% higher than the MT3D, whose accuracy is 81.5%. The most improvement is achieved for CL due to the introduction of pose information, which makes the model more robust to changes in pedestrian appearance.

**GREW.** In order to verify the generalization and robustness in complex scenarios, the TransGait is evaluated on the GREW dataset. As shown in Table 2, TransGait meets a new state-of-the-art in gait recognition in the wild. TransGait scores 56.27% in terms of Rank-1 metric, which exceeds GaitSet and GaitPart by about 10%. It is worth noting that GaitPart outperforms GaitSet on CASIA-B, but it performs worse than GaitSet on GREW. We believe that this is due to the lack of some frames in the gait cycle caused by imperfect detection and segmentation in the wild. GaitSet, which is not sensitive to the input order, is more suitable for this complex scenario than GaitPart, which is sensitive to the input order. TransGait is also insensitive to the input sequence and uses STM to mine the time relationship of multi-modal features, thereby achieving high recognition accuracy in complex natural environments.

**Table 2** Rank-1, Rank-5, Rank-10, Rank-20 accuracies on GREW

| Method | Rank-1 | Rank-5 | Rank-10 | Rank-20 |
|---|---|---|---|---|
| GEINet [44] | 6.82 | 13.42 | 16.97 | 21.01 |
| GaitSet [9] | 46.28 | 63.58 | 70.26 | 76.82 |
| GaitPart [6] | 44.01 | 60.68 | 67.25 | 73.47 |
| TransGait(ours) | **56.27** | **72.72** | **78.12** | **82.51** |

Bold data the results with the highest recognition rate among the results of different methods

**Table 3** Ablation Study. Control Condition: w/ and w/o applying silhouette, w/ and w/o pose, w/ and w/o applying STM. Rank-1 accuracies averaged on 11 views, excluding identical-view cases, are compared

| Group | Silhouette | Pose | STM | NM | BG | CL |
|---|---|---|---|---|---|---|
| A | ✓ | | ✓ | 97.3 | 92.8 | 80.6 |
| B | | ✓ | ✓ | 84.5 | 71.2 | 54.4 |
| C | ✓ | ✓ | | 96.9 | 92.5 | 80.3 |
| D | ✓ | ✓ | ✓ | 98.1 | 94.9 | 85.8 |

## 4.4 Ablation study

Several ablation studies with various settings are conducted on the CASIA-B dataset to verify the effectiveness of the silhouette-pose multi-modal fusion and the STM. We set up four groups of controlled experiments (denoted as A, B, C, and D, respectively). The experimental results are shown in Table 3, and the analysis is given in the following sections.

**Analysis of silhouette-pose multi-modal fusion.** In this paper, we propose a multi-modal gait recognition based on silhouette-pose fusion. The importance of silhouette-pose multi-modal fusion is verified by comparing the results for only silhouettes (group a), only pose heat maps (group b), and integrating silhouettes and pose heatmaps (group d). As summarized in Table 3, the ablation study shows that the concatenating of silhouette and pose features achieves better performance than a single feature, validating the potential of multi-modal fusion.

**Effectiveness of STM.** In order to validate the effectiveness of STM, we compare group c (without STM) and group d (with STM). Table 3 shows that STM significantly improves the accuracy of gait recognition, especially for CL. Moreover, in order to demonstrate the importance of STM, which generates temporal set features by fusing frame features containing a variety of different motion patterns on time scales for gait recognition, we compare our method with GaitPart, which is only based on short-term gait information. For a fair comparison, only the silhouettes and the STM (group a in Table 3) are used. The results are summarized in Table 4. As shown in Table 4, our model outperforms the Gaitpart, which proves that in addition to micro-motion, other motion patterns are also critical for gait recognition.

**Analysis of the layer number in STM.** The ablation experiment is designed to demonstrate the ability of the STM that models multi-motion patterns without deeply stacked structure. As shown in Table 5, the accuracy of STM is better with the layer numbers: 1 and 2 than the layer numbers: 3 and 4. This is because the multi-head attention block in STM can extract multi-motion (including short-, medium- and long-term motion) features by using global information at the low-level layer. The high-level layer of STM will model the relationship between the short- or medium-term motion features, which reduces the diversity of the motion information of the final features.

**Visualization of STM.** STM extracts multi-motion features by multi-headed attention. In order to observe the inter-frame relationships found by the different heads of STM, we further visualize the relationship between frame 0 and other frames with attention weights. As shown in Fig. 6, different heads find different inter-frame relationships. For example, head 1 focuses on the relationship between adjacent frames, while head 3 focuses on the relationship between distant frames. Due to the space limitation, we only show the visualization results of frame 0 and other frames in the three heads (24 frames and 8 heads). Note that the results for other frames are consistent with these results obtained by selecting frame 0.

## 4.5 Practicality

TransGait has great potential in more complicated practical conditions due to the invariance of the STM to frame permutations. This section investigates the practicality of TransGait through two novel scenarios. 1) Limited frames of input, and 2) different viewpoints of input frames. It is worth noting that our model was not retrained, and considering the comparison with GaitSet, we only use the silhouettes and the STM. All the experiments containing random selection were repeated ten times.

**Table 4** The effectiveness of multi-motion modeling. §denotes that only silhouettes and the STM are used

| method | NM | BG | CL |
|---|---|---|---|
| GaitPart [6] | 96.2 | 91.5 | 78.7 |
| TransGait§ | 97.3 | 92.8 | 80.6 |

**Table 5** Accuracy comparison (%) of different layer numbers of STM

| The number of layers | NM | BG | CL |
|---|---|---|---|
| 1 | 98.2 | 94.2 | 85.4 |
| 2 | 98.1 | 94.9 | 85.8 |
| 3 | 97.8 | 94.3 | 84.3 |
| 4 | 97.6 | 93.8 | 83.8 |

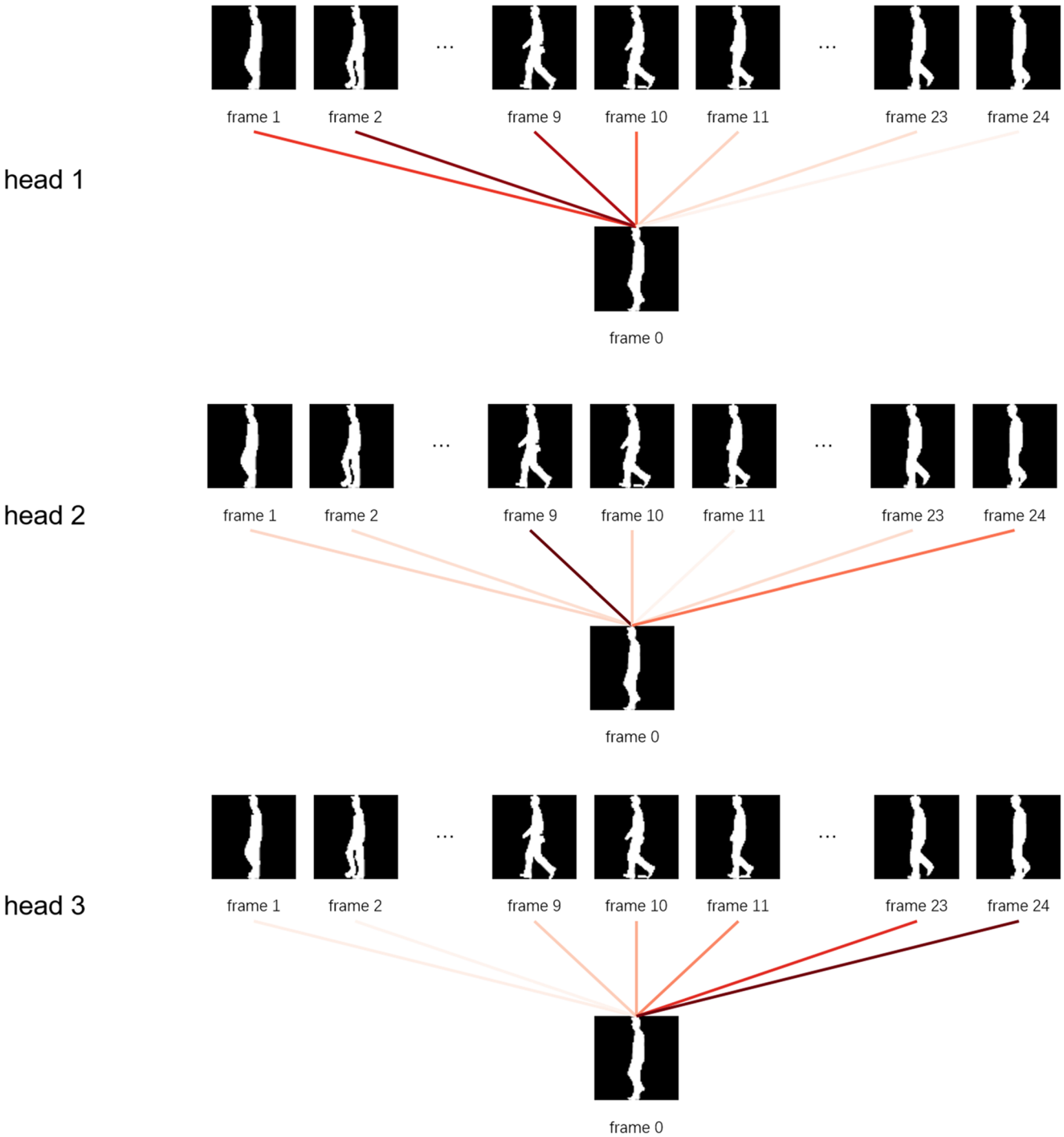


**Fig. 6** Visualization of the attention weight of the three heads of STM. Darker color means bigger weight

**Limited Frames.** In practical forensic identification scenarios, the gait information is often limited, where only some fitful and sporadic frames are available. This scenario was simulated by randomly selecting a certain number of frames from sequences to compose each sample in both gallery and probe. The proposed method is compared with GaitSet that treats the input as a set. As shown in Fig. 7, our model outperforms the GatSet on

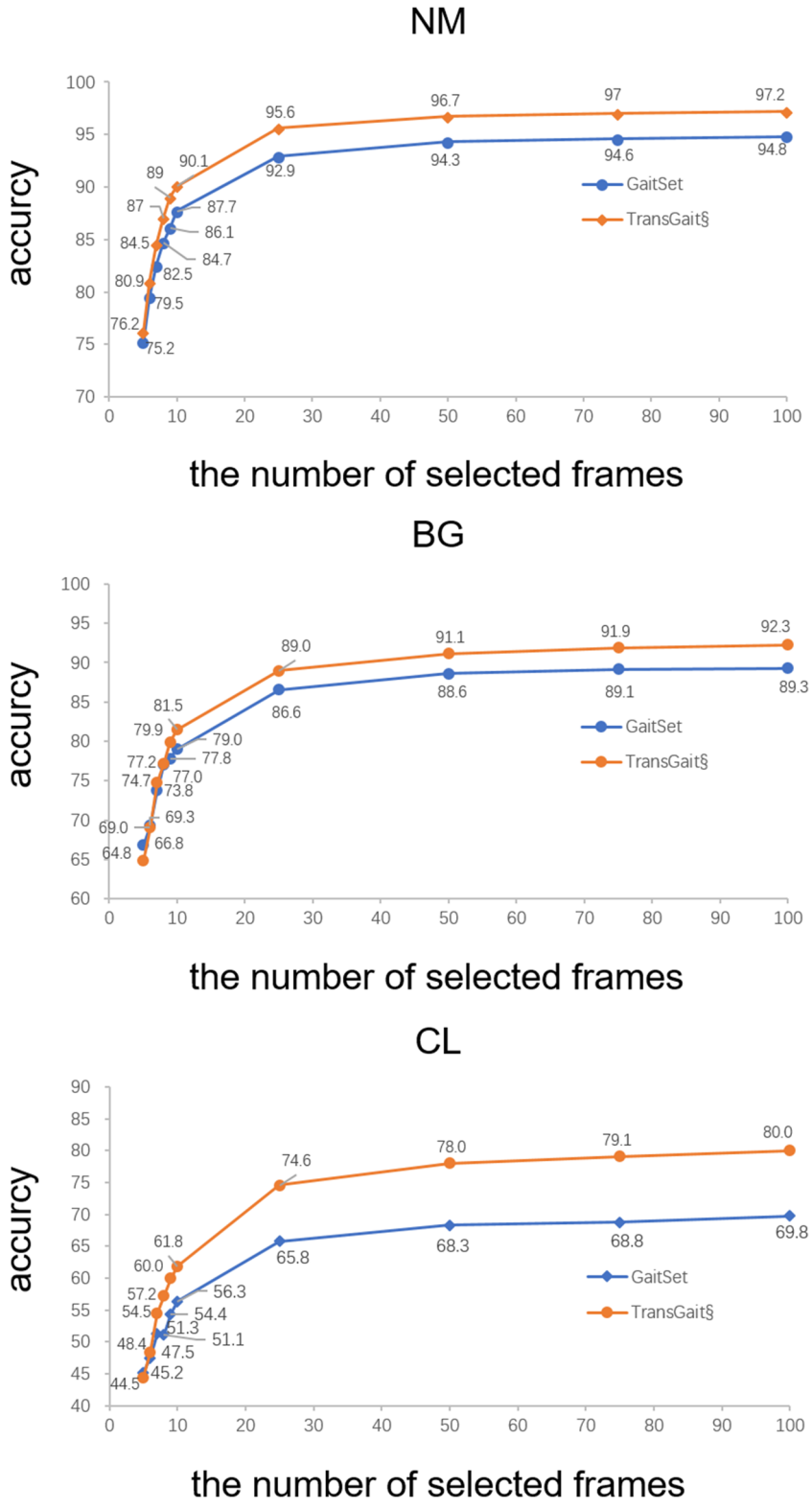


**Fig. 7** Average Rank-1 accuracies with constraints of silhouette volume on the CASIA-B dataset. The accuracy values are averaged on all 11 views excluding identical-view cases, and the final reported results are averaged across ten experimental repetitions. §denotes that only silhouettes and the STM are used

**Table 6** Multiview experiments on CASIA-B (%). §denotes that only silhouettes and the STM are used

| View difference | | 18°/162° | 36°/144° | 54°/126° | 72°/108° | 90° | Single view |
|---|---|---|---|---|---|---|---|
| NM | GaitSet [9] | 97.0 | 97.9 | 98.7 | 99.1 | 99.0 | 95.0 |
| | TransGait§ | 98.9 | 99.3 | 99.4 | 99.3 | 99.5 | 97.3 |
| BG | GaitSet [9] | 92.1 | 94.3 | 95.2 | 95.4 | 96.0 | 87.2 |
| | TransGait§ | 95.1 | 96.7 | 97.1 | 97.5 | 97.8 | 92.8 |
| CL | GaitSet [9] | 74.4 | 77.6 | 79.0 | 77.8 | 78.4 | 70.4 |
| | TransGait§ | 85.2 | 87.5 | 87.7 | 87.0 | 86.9 | 80.6 |

most numbers of selected frame in all walking conditions. This shows that STM is capable of efficiently exploiting the temporal relationships of sequences even with a limited number of frames.

**Multiple Views.** The experimental results in GaitSet showed that not only does the amount of input probe data improve the gait recognition effect, but also the data containing more information about view angle help to improve the accuracy of gait recognition. Similar to GaitSet, in this section, we also study the scenario where the gait is collected from different sequences with different views but the same walking conditions. We conducted the experiments on CASIA-B using the same test settings as GaitSet, whose results are given in Table 6.

The accuracy of each possible view difference is averaged to summarize the results for many view pairs. For example, the result of a 90° difference is averaged by the accuracies of 6 view pairs (0°&90°,18°&108°, ...,90°&180°). Furthermore, the nine view differences are folded at 90° and those larger than 90° are averaged with the corresponding view differences of less than 90°. For example, the results of 18° view differences are averaged with those of 162° view differences. As indicated in Table 6, regardless of the lack of different views in the training set, STM can effectively model interactions among gait frames from different views. And our model outperforms GaitSet in all walking conditions and view difference. Including multiple views in the input-set provides more gait information, and STM can use this information effectively to achieve better results.

## 5 Conclusion

In this work, we proposed a new gait recognition network, TransGait, where multi-modal gait features can be extracted with different movement patterns. Specifically, we combine the features of two different modes, silhouette and pose. The proposed multi-modal features are strongly discriminative and robust to the pedestrians' clothing and carryings. Also, STM is used to extract various motion patterns from the gait sequence. Unlike other existing temporal modeling approaches, STM can adaptively learn different motion patterns contained in gait sequences and is insensitive to input order. The experimental results on the CASIA-B dataset and the GREW dataset show that the proposed method achieves the outperforming accuracy over the state-of-the-art methods.

**Acknowledgments** This research work is supported by the Zhejiang Provincial Public Welfare Technology Research Project (No. LGF21F020008) and the Zhejiang Provincial Natural Science Foundation(No. LZ20F020001)

## References

1. Connor P, Ross A (2018) Biometric recognition by gait: A survey of modalities and features. Comput Vis Image Underst 167:1–27
2. Sepas-Moghaddam A, Etemad A (2021) Deep gait recognition:, A survey. arXiv preprint arXiv:2102.09546
3. Lin B, Zhang S, Bao F (2020) Gait recognition with multiple-temporal-scale 3d convolutional neural network. In: Proceedings of the 28th ACM international conference on multimedia, pp 3054–3062
4. Yu S, Tan D, Tan T (2006) A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: 18Th international conference on pattern recognition (ICPR'06), vol 4, pp 441–444
5. Verlekar T (2019) Gait analysis in unconstrained environments. PhD thesis, Ph. D. dissertation, Electrical and Computer Engineering, Instituto Superior
6. Fan C, Peng Y, Cao C, Liu X, Hou S, Chi J, Huang Y, Li Q, He Z (2020) Gaitpart: Temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14225–14233
7. Zhang Y, Huang Y, Yu S, Wang L (2019) Cross-view gait recognition by discriminative feature learning. IEEE Trans Image Process 29:1001–1015
8. Hou S, Cao C, Liu X, Huang Y (2020) Gait lateral network: Learning discriminative and compact representations for gait recognition. In: European conference on computer vision, pp 382–398
9. Chao H, He Y, Zhang J, Feng J (2019) Gaitset: regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 8126–8133
10. Xu C, Makihara Y, Li X, Yagi Y, Lu J (2020) Cross-view gait recognition using pairwise spatial transformer networks. IEEE Trans Circuits Syst Video Technol 31(1):260–274

11. Qin H, Chen Z, Guo Q, Wu QJ, Lu M (2021) Rpnet: Gait recognition with relationships between each body-parts. IEEE Transactions on Circuits and Systems for Video Technology
12. Ben X, Gong C, Zhang P, Yan R, Wu Q, Meng W (2019) Coupled bilinear discriminant projection for cross-view gait recognition. IEEE Trans Circuits Syst Video Technol 30(3):734–747
13. Xu W (2021) Graph-optimized coupled discriminant projections for cross-view gait recognition. Appl Intell, 1–13
14. Li N, Zhao X, Ma C (2020) A model-based gait recognition method based on gait graph convolutional networks and joints relationship pyramid mapping. arXiv preprint arXiv:2005.08625
15. Liao R, Yu S, An W, Huang Y (2020) A model-based gait recognition method with body pose and human prior knowledge. Pattern Recogn 98:107069
16. An W, Yu S, Makihara Y, Wu X, Xu C, Yu Y, Liao R, Yagi Y (2020) Performance evaluation of model-based gait on multi-view very large population database with pose sequences. IEEE Transactions on Biometrics, Behavior, and Identity Science 2(4):421–430
17. Jun K, Lee D-W, Lee K, Lee S, Kim MS (2020) Feature extraction using an rnn autoencoder for skeleton-based abnormal gait recognition. IEEE Access 8:19196–19207
18. Stenum J, Rossi C, Roemmich RT (2021) Two-dimensional video-based analysis of human gait using pose estimation. PLoS Computational Biology 17(4):1008935
19. Rao H, Wang S, Hu X, Tan M, Guo Y, Cheng J, Liu X, Hu B (2021) A self-supervised gait encoding approach with locality-awareness for 3d skeleton based person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence
20. Feng Y, Li Y, Luo J (2016) Learning effective gait features using lstm. In: 2016 23Rd international conference on pattern recognition (ICPR), pp 325–330
21. Li X, Makihara Y, Xu C, Yagi Y, Yu S, Ren M (2020) End-to-end model-based gait recognition. In: Proceedings of the Asian conference on computer vision
22. Zhao L, Guo L, Zhang R, Xie X, Ye X (2021) mmgaitset: multimodal based gait recognition for countering carrying and clothing changes. Appl Intell, pp 1–14
23. Han J, Bhanu B (2005) Individual recognition using gait energy image. IEEE Trans Pattern Anal Mach Intell 28(2):316–322
24. Bashir K, Xiang T, Gong S (2009) Gait recognition using gait entropy image
25. Sepas-Moghaddam A, Etemad A (2020) View-invariant gait recognition with attentive recurrent learning of partial representations. IEEE Transactions on Biometrics, Behavior, and Identity Science
26. Wang X, Yan WQ (2020) Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory. International Journal of Neural Systems 30(01):1950027
27. Wolf T, Babaee M, Rigoll G (2016) Multi-view gait recognition using 3d convolutional neural networks. In: 2016 IEEE international conference on image processing (ICIP), pp 4165–4169
28. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. arXiv preprint arXiv:1706.03762
29. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
30. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M (2021) Transformers in vision: A survey. arXiv preprint arXiv:2101.01169
31. Girdhar R, Carreira J, Doersch C, Zisserman A (2019) Video action transformer network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 244–253
32. Plizzari C, Cannici M, Matteucci M (2020) Spatial temporal transformer network for skeleton-based action recognition. arXiv preprint arXiv:2008.07404
33. Liu Z, Luo S, Li W, Lu J, Wu Y, Li C, Yang L (2020) Convtransformer:, A convolutional transformer network for video frame synthesis. arXiv preprint arXiv:2011.10185
34. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
35. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030
36. Yao L, Kusakunniran W, Wu Q, Xu J, Zhang J (2021) Collaborative feature learning for gait recognition under cloth changes. IEEE Transactions on Circuits and Systems for Video Technology
37. Wei S.-E., Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional pose machines. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4724–4732
38. Fu Y, Wei Y, Zhou Y, Shi H, Huang G, Wang X, Yao Z, Huang T (2019) Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 8295–8302
39. Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV), pp 480–496
40. Wang G, Yuan Y, Chen X, Li J, Zhou X (2018) Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM international conference on multimedia, pp 274–282
41. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25:1097–1105
42. Hermans A, Beyer L, Leibe B (2017) In defense of the triplet loss for person re-identification. arXiv:1703.07737
43. Zhu Z, Guo X, Yang T, Huang J, Deng J, Huang G, Du D, Lu J, Zhou J (2021) Gait recognition in the wild: A benchmark. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 14789–14799
44. Shiraga K, Makihara Y, Muramatsu D, Echigo T, Yagi Y (2016) Geinet: view-invariant gait recognition using a convolutional neural network. In: IEEE International Conference on Biometrics (ICB), pp 1–8




**Guodong Li** received the B.E. degree form China Jiliang University, China, in 2019. He is currently a master student in the Faculty of Electrical Engineering and Computer Science, Ningbo University, China. His research interests include person re-identification and gait recognition.

**Lijun Guo** received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011. He is currently a full professor with Ningbo University, Ningbo, China. His current research interests include computer vision and pattern recognition, intelligence science, medical imaging analysis.

**Jiangbo Qian** received the Ph.D. degree in computer-science from Southeast University, China,in 2006. He was a Visiting Scholar with the Department of Computer and Information Science, The University of Michigan-Dearborn, USA. He is currently a Professor with the Faculty of Electrical Engineering and Computer Science, Ningbo University, China. His research interests include database management, streaming data processing, deep learning, computer vision, and hardware/software co-design.

**Rong Zhang** received her M.S. degree in pattern recognition and intelligent systems from Harbin Institute of Technology, China, in 2001, and her Ph.D. degree in communication and information systems from Ningbo University, in 2015. She is currently an associate professor at Ningbo University, Ningbo, China. Her research interests include digital image forensics, computer vision and medical imaging analysis.

**Shangce Gao** received his B.S. degree from Southeast University, Nanjing, China in 2005, and M.E. and D.E. degrees from University of Toyama, Toyama, Japan in 2008 and 2011, respectively. He is currently an Associate Professor with the Faculty of Engineering, University of Toyama, Japan. His current research interests include nature-inspired technologies, mobile computing, machine learning, and neural networks for real-world applications.