Hierarchical Spatio-Temporal Representation Learning for Gait Recognition

Lei Wang[1,2], Bo Liu[1,2]*, Fangfang Liang[1,2], Bincheng Wang[1,2]
[1] College of Information Science and Technology, Hebei Agricultural University, China
[2] Hebei Key Laboratory of Agricultural Big Data, China
{20212060107, 20212060108}@pgs.hebau.edu.cn, {boliu, liangfangfang}@hebau.edu.cn

Abstract

Gait recognition is a biometric technique that identifies individuals by their unique walking styles, which is suitable for unconstrained environments and has a wide range of applications. While current methods focus on exploiting body part-based representations, they often neglect the hierarchical dependencies between local motion patterns. In this paper, we propose a hierarchical spatio-temporal representation learning (HSTL) framework for extracting gait features from coarse to fine. Our framework starts with a hierarchical clustering analysis to recover multi-level body structures from the whole body to local details. Next, an adaptive region-based motion extractor (ARME) is designed to learn region-independent motion features. The proposed HSTL then stacks multiple ARMEs in a top-down manner, with each ARME corresponding to a specific partition level of the hierarchy. An adaptive spatio-temporal pooling (ASTP) module is used to capture gait features at different levels of detail to perform hierarchical feature mapping. Finally, a frame-level temporal aggregation (FTA) module is employed to reduce redundant information in gait sequences through multi-scale temporal downsampling. Extensive experiments on CASIA-B, OUMVLP, GREW, and Gait3D datasets demonstrate that our method outperforms the state-of-the-art while maintaining a reasonable balance between model accuracy and complexity. Code is available at: https://github.com/gudaochangsheng/HSTL.

1. Introduction

Unlike other biometric technologies such as fingerprint, iris, and face, human gait can be captured at a distance without subject cooperation [34]. By evaluating individual-specific walking patterns, gait recognition has been applied in a variety of fields, including criminal investigations [31, 29], sports science [17, 6], and smart transportation [47]. However, the recognition can be challenging due to large variations in viewpoint [28, 18], occlusion [33, 43], and wearing [50, 48].

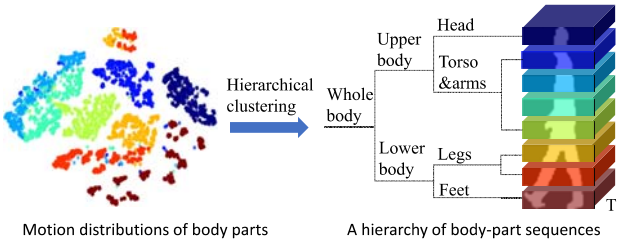


Figure 1: The motivation for our approach. Left: the motion distributions of body parts. The same color indicates the same spatial location across multiple gait sequences in the CASIA-B dataset. Right: an example hierarchy of body-part sequences where T denotes the temporal dimension of the sequence.

To address these issues, various approaches have been proposed for extracting gait features from silhouette sequences [4, 25, 16, 15, 55], 3D human structures [1, 22, 42, 59, 20], or gait templates [10, 35, 51]. Silhouette-based gait recognition methods have gained increasing attention due to the ease of obtaining silhouettes from raw videos while preserving essential temporal information. The alignment of the input silhouette makes it possible for some methods to extract local body features by horizontally slicing the silhouette image [56] or intermediate-layer features [8, 27]. This partitioning strategy, first introduced in person re-identification (ReID) [38], has been proven to be effective for gait recognition [4, 8, 12, 3].

However, the main limitation of the above part-based approaches is that they do not consider the hierarchical nature of local body movements [2]. For instance, within a gait cycle, the feet and lower body have distinct motion characteristics. Therefore, it is important to treat these body regions separately and investigate their part-whole relationships. Our motivation stems from the examination of body-part-specific motion clues. Specifically, each raw gait se-

*Corresponding Author

quence in the CASIA-B [52] dataset is uniformly divided into eight part sequences along the body axis, so that each division roughly match a particular body part [1]. The distributions of all body parts are shown in the left part of Fig. 1. Observably, some parts, e.g., the head and feet, are easily separated owing to their large changes in walking kinematics. Whereas other parts, such as the thighs and calves, overlap due to the strong motion correlations between them. Further, to identify the relational structure among the part sequences, a hierarchical clustering analysis [7] is performed. The results are shown in the right part of Fig. 1, indicating that the semantic body regions can be captured in the higher clustering levels without precise localization of the body parts.

Following the above findings, we propose a novel hierarchical spatio-temporal representation learning (HSTL) framework for gait representation. The HSTL framework consists of multiple adaptive region-based motion extractor (ARME) modules, which are stacked to learn hierarchical motion patterns implied in a gait sequence (as shown in Fig.1). In the ARME module, to account for inter-regional differences, non-shared 3D convolutions are used in correspondence with individual body regions. There regions are pre-identified by a hierarchical clustering process performed on fixed horizontal partitions, allowing each body region to cover one or more body parts. Consequently, the deeper the ARME is, the more local features it tends to extract. Moreover, an adaptive spatio-temporal pooling (ASTP) module is proposed, which couples with an ARME module on the corresponding level to obtain hierarchical gait embeddings.

In addition, changes in gait speed or sampling frequency may result in several redundant frames in a gait sequence. Although several temporal fusion strategies have been proposed, they lose spatial information [8, 15] or lack adaptability [27, 25]. To address this issue, we propose a frame-level temporal aggregation strategy (FTA). FTA fuses temporal features at multiple time steps, preserving significant motion information while compressing the sequence length. The main contributions of this paper are summarized as follows.

1. We propose a hierarchical spatio-temporal representation learning (HSTL) framework for gait recognition. HSTL takes into account the dependencies of body regions in gait motions, ensuring simplicity and scalability of the architectural design.

2. We introduce an adaptive region-based motion extractor (ARME) module to learn region-independent spatio-temporal representation for gait sequences, an adaptive spatio-temporal pooling (ASTP) module to perform hierarchical feature mapping, and a frame-level temporal aggregation (FTA) strategy to compress a gait sequence by removing redundant frames.

3. Extensive experiments on the widely used gait datasets CASIA-B [52], including OUMVLP [39], GREW [60] and Gait3D [59], demonstrate that our method achieves state-of-the-art performance while offering a suitable trade-off between model accuracy and complexity.

[1] Each part is pooled into a vector for visualization using t-SNE [44].

2. Related Work

2.1. Gait Recognition

Deep learning-based gait recognition methods can be broadly categorized into two categories: model-based and appearance-based. Model-based approaches extract structure and motion information from gait videos with the aid of pose estimation [1, 22, 42, 20, 41, 59]. Although these methods are robust to changes in viewpoint and appearance, they are sensitive to the accuracy of the pose parameters, making them incapable of handling low-resolution data. On the other hand, appearance-based approaches learn the feature representation from raw videos [57, 37, 24], or binary silhouette sequences [10, 35, 51, 4, 12, 8, 13, 27, 16, 14, 3, 5], which offer greater flexibility than model-based approaches. Our proposed method belongs to the family of appearance-based methods and uses silhouette sequences as inputs.

2.2. Hierarchical Model

Hierarchical feature representation has been successfully applied to a wide range of vision tasks. Here, we provide a brief review of the hierarchical object ReID approaches related to gait recognition.

In person ReID, some approaches [30, 54, 45, 40, 53] hierarchically learn local descriptions and aggregate appearance features at different levels. For example, Matsukawa *et al*. [30] described an image patch via hierarchical Gaussian distribution. Zhang *et al*. [53] proposed a framework to learn coarse-grained and fine-grained features according to body structure. To solve the occlusion problem, Tan *et al*. [40] devised a hierarchical mask generator to learn from both occluded and holistic joint images. In vehicle ReID, some approaches [46, 36, 19] extract features from vehicle images in a hierarchical manner. For instance, Wei *et al*. [46] proposed an RNN-based module for extracting latent cues from the model level to the vehicle level. Shyam *et al*. [36] developed an attention-based hierarchical feature extractor. In addition, Li *et al*. [19] proposed a global structural embedding module for investigating hierarchical relationships between vehicle characteristics by incorporating attribute and state information.

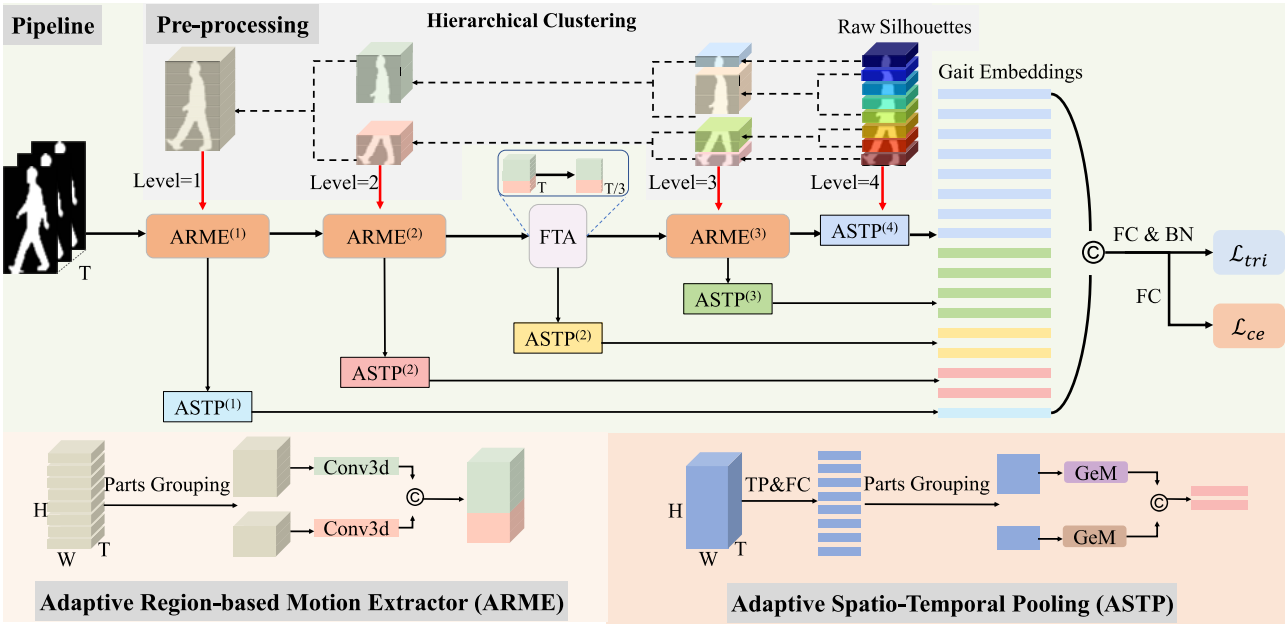


Figure 2: The framework of HFSL. It mainly consists of three modules: ARME (adaptive region-based motion extractor), ASTP (adaptive spatio-temporal pooling), and FTA (frame-level temporal aggregation). During pre-processing, a hierarchy of walking is obtained to guide the architectural design of HFSL. The framework uses multiple ARMEs to extract gait features from the entire body to individual regions. The ASTP module performs hierarchical feature mapping for the output of each level of ARME. The FTA module compresses local clips of each gait sequence to reduce the number of redundant frames. T, H and W denote dimensions of the feature maps. © represents the concatenation operation.

For gait recognition, fusing features of multiple granularities can improve performance [56, 8, 27, 25, 16, 3]. In particular, CSTL [15] proposed a temporal modeling network that integrates multi-scale temporal features adaptively. By combining part-level and sequence-level features, GaitPart [8] obtained a part-independent spatio-temporal expression. Additionally, GaitGL [27] considered both full body-based and part-based information to achieve discriminative feature learning. Instead of equally dividing the feature maps, in 3D Local [16], a localization operation was developed to find 3D volumes of body parts in a sequence. However, most existing gait recognition methods do not sufficiently exploit the hierarchical dependencies among body parts during walking. In this paper, the proposed HSTL performs a coarse-to-fine hierarchical strategy that integrates multi-level motion patterns from gait sequences.

2.3. Temporal Model

Temporal cues play a crucial role in gait recognition due to the periodic changes in body shape. Previous methods treat a gait sequence as an unordered set, either compressing it into a single gait template during preprocessing [35, 21] or learning order-independent gait representations from silhouette sets [4, 12, 13]. These methods assume that different subjects share similar global gait patterns, making ordering inputs unnecessary for gait assessment. However, ignoring the temporal nature of the gait sequence can result in missing discriminative local motion information. Recently, some approaches have achieved significant performance gains by explicitly modeling temporal information using LSTM [56], 1D convolution [15], and 3D convolution [26, 16]. Nevertheless, these spatio-temporal operators also significantly increase computational costs. Although some methods have been proposed to reduce video length by aggregating local clips [27, 25, 3], they lack adaptability to variations in pace. The main difference between our approach and others [23, 11] is that we employ multi-scale temporal pooling at the frame level while considering variations in motion across body regions, leading to a more adaptable reduction of the gait sequence length.

3. Proposed Method

In this section, we present the detailed description of HSTL, including the adaptive region-based motion extractor (ARME), the adaptive spatio-temporal pooling (ASTP), and the frame-level temporal aggregation (FTA).

3.1. Framework Pipeline

The overview of our HSTL is presented in Fig. 2. Given a gait dataset $\mathcal{D} = \{S_i\}_{i=1}^{N}$ with N gait sequences, where each sequence $S_i \in \mathbb{R}^{C\times T\times H\times W}$ is represented as a 4D tensor with C channels, T frames, and $H \times W$ pixels. During the preprocessing stage, each gait sequence S_i is divided horizontally and uniformly into k part sequences, indexed from 1 to k. Then, a hierarchical clustering algorithm [7] is applied to these part sequences to obtain a generic hierarchy of gait motions, which is denoted as $\mathcal{P} = \{\mathcal{P}^{(l)}\}_{l=1}^{L}$. Here, L is the number of levels in the hierarchy and $\mathcal{P}^{(l)}$ is the set of partitions at level l. The partitions at level l are defined as $\mathcal{P}^{(l)} = \{P_1^{(l)}, P_2^{(l)}, \ldots, P_{K_l}^{(l)}\}$, where $P_j^{(l)}$ is the j-th subset of part indices and K_l is the number of groups at level l. For instance, the top level $\mathcal{P}^{(1)} = \{\{1, 2, \ldots, k-1, k\}\}$ means all the k parts can be considered as a whole for the gait analysis. This hierarchy provides a structured property of the gait motion patterns and can be utilized to guide gait feature extraction. To achieve this, our proposed HSTL employs three modules: ARME for extracting independent multi-granularity motion features, ASTP for generating vectorized gait embeddings, and FTA for reducing redundant information at the frame level. The HSTL stacks these three modules according to the division in $\mathcal{P}$, and the main branch of the HSTL for the input sequence S_{in} can be formalized as:

$$Y^M = \Gamma^{(L)} \circ \Psi^{(L-1)} \circ \cdots \circ \Omega^{(2)} \circ \Psi^{(2)} \circ \Psi^{(1)}(S_{in}), \quad (1)$$

where $\Psi^{(l)}$, $\Gamma^{(l)}$, and $\Omega^{(l)}$ represent the ARME, ASTP and FTA modules at the l-th level of $\mathcal{P}$, respectively. Since FTA uses inter-frame compression to reduce redundant information, it is employed only once at the l_Ω-th level in $\mathcal{P}$ (e.g., $l_\Omega = 2$ in Eq. (1)) to prevent excessive loss of information.

To obtain the hierarchical gait embeddings, the output $Y^{(l)}$ of each $\Psi^{(l)}$ at levels $l \in \{1, 2, \ldots, L-1\}$ and the output of $\Omega^{(l_\Omega)}$ at level l_Ω, denoted as $Y_\Omega^{(l_\Omega)}$, are fed into the corresponding $\Gamma^{(l)}$. The resulting outputs from these L auxiliary branches are concatenated with the output of the main branch defined in Eq. (1), forming the final result, denoted as Y, which is given by:

$$Y = \Big[Y^M, \Gamma^{(L-1)}\left(Y^{(L-1)}\right), \ldots, \Gamma^{(l_\Omega)}\left(Y_\Omega^{(l_\Omega)}\right), \Gamma^{(2)}\left(Y^{(2)}\right), \Gamma^{(1)}\left(Y^{(1)}\right)\Big], \quad (2)$$

where $[,]$ denotes the concatenation operation.

Finally, Y undergoes feature mapping through separate fully connected layers. The model is then trained using a combination of triplet loss $\mathcal{L}_{tri}$ and cross-entropy loss $\mathcal{L}_{ce}$, which is a commonly adopted practice in gait recognition [12, 27, 15, 16, 5]. Further details regarding the relevant modules are described in the following subsections.

3.2. Adaptive Region-based Motion Extractor

The adaptive region-based motion extractor (ARME) aims to extract independent spatio-temporal patterns that are associated with different human body parts in a gait sequence. Unlike existing methods that uniformly slice gait images or sequences along the height axis [56, 8, 27, 25], ARME considers the inherent hierarchical relationships among different part sequences that are consistent with walking patterns. This allows ARME to effectively capture the unique walking kinematics of each part.

Given the hierarchical relation $\mathcal{P}$ introduced in Section 3.1, ARME first divides the input sequence X into K_l regions based on the partition of the l-th level $\mathcal{P}^{(l)}$, resulting in the set of regions $\{X_j^{(l)}\}_{j=1}^{K_l}$, where $X_j^{(l)} \in \mathbb{R}^{C\times T\times H_j^{(l)}\times W}$. $H_j^{(l)} = \frac{|P_j^{(l)}|}{k}H$ is the height of the j-th region of the l-th level. Then the l-th level of ARME, $\Psi^{(l)}$, can be defined as follows:

$$Y_\Psi^{(l)} = \Psi^{(l)}(X^{(l)}) = \left[f_1(X_1^{(l)}), f_2(X_2^{(l)}), \ldots, f_{K_l}(X_{K_l}^{(l)})\right], \quad (3)$$

where $f_j(\cdot)$ represents the independent 3D convolution operation applied to the j-th region. The output feature map $Y_\Psi^{(l)} \in \mathbb{R}^{C^{(l)}\times T\times H\times W}$ of level l has $C^{(l)}$ channels. This module only modifies the number of channels, preserving the spatial and temporal resolutions of the input feature map.

3.3. Adaptive Spatio-Temporal Pooling

It is a common procedure in gait recognition to obtain a compact and fixed-length feature representation by performing horizontal and uniform slicing of feature maps and strip-based pooling [4, 9, 26, 27, 15]. However, a non-uniform division of the feature maps is better at capturing the gait motion characteristics. Thus, the adaptive spatio-temporal pooling (ASTP) is devised to construct hierarchical feature mapping (as shown in Fig. 2). Similar to the ARME module described in Section 3.2, the hierarchy $\mathcal{P}$ enables us to obtain the j-th region of the l-th level, denoted as $X_j^{(l)} \in \mathbb{R}^{C\times T\times H_j^{(l)}\times W}$. The corresponding ASTP, denoted as $\Gamma_j^{(l)}$, can be expressed as follow:

$$Y_{\Gamma,j}^{(l)} = \Gamma_j^{(l)}(X_j^{(l)}) = \mathrm{GeM}_j \circ \mathrm{FC} \circ \mathrm{Max}(X_j^{(l)}), \quad (4)$$

where $\mathrm{Max}(\cdot)$ represents a max pooling operation along the temporal dimension, $\mathrm{FC}(\cdot)$ represents a fully connected layer, $\mathrm{GeM}_j(\cdot)$ represents a generalized mean pooling (GeM) operation [32] for j-th region, and $\Gamma_j^{(l)}$: $\mathbb{R}^{C\times T\times H_j^{(l)}\times W} \mapsto \mathbb{R}^{C\times 1\times H_j^{(l)}\times W} \mapsto \mathbb{R}^{C^{(l)}\times 1\times H_j^{(l)}\times W} \mapsto \mathbb{R}^{C^{(l)}\times 1\times 1\times 1}$, in other words, $Y_{\Gamma,j}^{(l)} \in \mathbb{R}^{C^{(l)}\times 1\times 1\times 1}$. Therefore, by concatenating outputs of K_l regions, we can

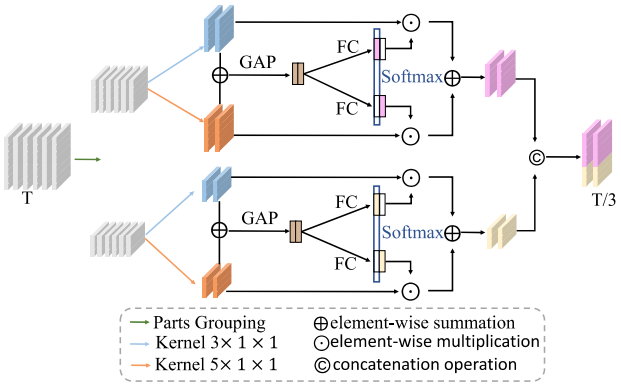


Figure 3: The detailed structure of the frame-level temporal aggregation (FTA). For simplicity, we omit the channel dimension C.

obtain $Y_{\Gamma}^{(l)} = \left[Y_{\Gamma,1}^{(l)}, Y_{\Gamma,2}^{(l)}, \ldots, Y_{\Gamma,K_l}^{(l)}\right]$, where $Y_{\Gamma}^{(l)} \in \mathbb{R}^{C^{(l)} \times 1 \times K_l \times 1}$ is the output of ASTP at level l.

3.4. Frame-level Temporal Aggregation

A gait sequence may contain several redundant frames due to factors such as the acquisition frame rate and pace frequency. To reduce computational costs, some methods compress a gait sequence by aggregating its local clips [27, 25]. In the proposed frame-level temporal aggregation (FTA) strategy, we consider both the gait structure and the multiscale temporal information. Given the j-th gait region at the l-th level, $X_j^{(l)} \in \mathbb{R}^{C \times T \times H_j^{(l)} \times W}$, we first fuse the features of the two temporal scales using the following formula:

$$\begin{aligned} \hat{U}_j^{(l)} &= U_{j,1}^{(l)} + U_{j,2}^{(l)} \\ &= \mathrm{Max}_{3\times1\times1}^{3\times1\times1}\left(X_j^{(l)}\right) + \mathrm{Max}_{5\times1\times1}^{3\times1\times1}\left(X_j^{(l)}\right), \end{aligned} \tag{5}$$

where $\mathrm{Max}_{3\times1\times1}^{3\times1\times1}(\cdot)$ and $\mathrm{Max}_{5\times1\times1}^{3\times1\times1}(\cdot)$ denote max pooling operations with kernel sizes of $3 \times 1 \times 1$ and $5 \times 1 \times 1$ respectively, both with stride of $3 \times 1 \times 1$. $\hat{U}_j^{(l)}$, $U_{j,1}^{(l)}$ and $U_{j,2}^{(l)}$ have the same size of $(C, \frac{T}{3}, H_j^{(l)}, W)$. The output of Eq.(5), $\hat{U}_j^{(l)}$, is the element-wise summation of the aggregation results of the two scales, $U_{j,1}^{(l)}$ and $U_{j,2}^{(l)}$, which reduces the temporal dimension of the input from T to $\frac{T}{3}$.

Then, the FTA model produces frame-level weights, which can be expressed as:

$$\begin{aligned} Z_{j,1}^{(l)} &= \mathrm{FC}_{j,1}^{(l)}\left(\mathrm{GAP}\left(\hat{U}_j^{(l)}\right)\right), \\ Z_{j,2}^{(l)} &= \mathrm{FC}_{j,2}^{(l)}\left(\mathrm{GAP}\left(\hat{U}_j^{(l)}\right)\right), \end{aligned} \tag{6}$$

where $\mathrm{GAP}(\cdot)$ represents the global mean pooling along spatial dimension. $\mathrm{FC}_{j,1}(\cdot)$ and $\mathrm{FC}_{j,2}(\cdot)$ are two independent fully connected layers that generate the frame selection weighting tensors, $Z_{j,1}^{(l)}$ and $Z_{j,2}^{(l)} \in \mathbb{R}^{C \times \frac{T}{3} \times 1 \times 1}$, for $U_{j,1}^{(l)}$ and $U_{j,2}^{(l)}$, respectively. The weights are further normalized across the two scales, which can be written as follows:

$$\mathcal{W}_{j,s,c,t}^{(l)} = \frac{e^{Z_{j,s,c,t}^{(l)}}}{e^{Z_{j,1,c,t}^{(l)}} + e^{Z_{j,2,c,t}^{(l)}}} \quad s \in \{1, 2\}, \tag{7}$$

where $\mathcal{W}_{j,s,c,t}^{(l)} \in \mathbb{R}^{1\times1\times1\times1}$ is the weight value of the c-th channel of the t-th frame. Combining Eq. (5) and Eq. (7), the j-th output region feature $Y_{\Omega,j}^{(l)} \in \mathbb{R}^{C^{(l)} \times \frac{T}{3} \times H_j^{(l)} \times W}$ for the l-th level of FTA can be obtained as follows:

$$Y_{\Omega,j}^{(l)} = \mathcal{W}_{j,1}^{(l)} \odot U_{j,1}^{(l)} + \mathcal{W}_{j,2}^{(l)} \odot U_{j,2}^{(l)}, \tag{8}$$

where $\mathcal{W}_{j,1}^{(l)}, \mathcal{W}_{j,2}^{(l)} \in \mathbb{R}^{C \times \frac{T}{3} \times 1 \times 1}$ are two weight tensors calculated using Eq. (5), and $\odot$ represents element-wise multiplication operation. The FTA module outputs $Y_{\Omega}^{(l)} \in \mathbb{R}^{C \times \frac{T}{3} \times H \times W}$ by concatenating the K_l gait regions of level l, where $Y_{\Omega}^{(l)} = \left[Y_{\Omega,1}^{(l)}, Y_{\Omega,2}^{(l)}, \ldots, Y_{\Omega,K_l}^{(l)}\right]$.

4. Experiments

4.1. Datasets and Evaluation Protocols

CASIA-B. The CASIA-B [52] dataset is a widely used benchmark for gait recognition. It contains video sequences of 124 subjects with 11 different views and three walking conditions (normal walking (NM), walking with a bag (BG), and walking with a coat (CL)). Our study follows the protocol outlined in previous works [4, 8, 27, 25, 15, 16]. The first 74 subjects are used for training, and the remaining 50 subjects are used for testing. During testing, the first four sequences under NM (NM#01-04) are regarded as the gallery set, and the rest (NM#05-06, BG#01-02, CL#01-02) are regarded as the probe set.

OUMVLP. The OUMVLP [39] is one of the largest gait datasets, containing silhouette sequences of 10,307 subjects. Each subject has a single normal walking condition (NM) with 14 views. According to the protocol provided by the dataset, the first 5,153 subjects are used for training while the remaining 5,154 subjects are used for testing. During the testing phase, the sequences of NM#01 are assigned to the gallery set, and the sequences of NM#02 are considered as the probe set.

GREW. GREW [60] is the first large-scale dataset for gait recognition in the wild, consisting of 128,671 sequences from 26,345 individuals captured by 882 cameras. It includes four data types: four data types: silhouettes, optical flow, 2D pose, and 3D pose. The dataset is divided into a training set with 20,000 subjects and 102,887 sequences, and a testing set with 6,000 subjects and 24,000 sequences. In the testing phase, each subject has two sequences for the

Table 1: Rank-1 accuracy (%) on CASIA-B under all views and different conditions, excluding identical-view cases. Std denotes the performance sample standard deviation across 11 views.

Gallery NM #1-4		0° – 180°											Mean	Std
Probe		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°		
NM #5-6	GaitSet [4]	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0	3.5
	GaitPart [8]	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2	3.1
	3D Local [16]	96.0	99.0	99.5	98.9	97.1	94.2	96.3	99.0	98.8	98.5	95.2	97.5	1.8
	CSTL [15]	97.2	99.0	99.2	98.1	96.2	95.5	97.7	98.7	99.2	98.9	96.5	97.8	1.3
	GaitGL [27]	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	99.3	98.8	94.0	97.4	1.7
	LagrangeGait [3]	95.7	98.1	99.1	98.3	96.4	95.2	97.5	99.0	99.3	98.9	94.9	97.5	1.6
	MetaGait [5]	97.3	**99.2**	99.5	**99.1**	97.2	95.5	97.6	99.1	99.3	**99.1**	96.7	**98.1**	1.3
	Ours	**97.6**	98.0	**99.6**	98.2	**97.4**	**96.5**	**97.9**	**99.3**	**99.4**	98.4	**97.0**	**98.1**	**1.0**
BG #1-2	GaitSet [4]	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2	4.9
	GaitPart [8]	89.1	94.8	96.7	95.1	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.5	4.2
	3D Local [16]	92.9	95.9	**97.8**	96.2	93.0	87.8	92.7	96.3	97.9	98.0	88.5	94.3	3.5
	CSTL [15]	91.7	96.5	97.0	95.4	90.9	88.0	91.5	95.8	97.0	95.5	90.3	93.6	3.0
	GaitGL [27]	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	98.2	96.9	91.5	94.5	2.8
	LagrangeGait [3]	94.2	96.2	96.8	95.8	94.3	89.5	91.7	96.8	98.0	97.0	90.9	94.6	2.7
	MetaGait [5]	92.9	**96.7**	97.1	96.4	94.7	90.4	92.9	**97.2**	98.5	**98.1**	92.3	95.2	2.6
	Ours	**95.0**	96.5	97.3	**96.6**	**95.3**	**93.3**	**94.6**	96.8	**98.6**	97.7	**92.9**	**95.9**	**1.7**
CL #1-2	GaitSet [4]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4	8.0
	GaitPart [8]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7	6.6
	3D Local [16]	78.2	90.2	92.0	87.1	83.0	76.8	83.1	86.6	86.8	84.1	70.9	83.7	6.2
	CSTL [15]	78.1	89.4	91.6	86.6	82.1	79.9	81.8	86.3	88.7	86.6	75.3	84.2	4.9
	GaitGL [27]	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6	6.3
	LagrangeGait [3]	77.4	90.6	93.2	90.2	84.7	80.3	85.2	87.7	89.3	86.6	71.0	85.1	6.3
	MetaGait [5]	80.0	91.8	93.0	87.8	86.5	82.9	85.2	90.0	90.8	89.3	78.4	86.9	**4.6**
	Ours	**82.4**	**94.2**	**95.0**	**91.7**	**88.2**	**83.3**	**88.0**	**92.3**	**93.1**	**91.0**	**78.5**	**88.9**	5.1

gallery set and two for the probe set. The GREW dataset also includes a distractor set, which contains 233,857 unlabeled sequences.

Gait3D. The Gait3D dataset [59] is a newly proposed dataset for gait recognition in uncontrolled indoor environments, particularly in large supermarkets. It contains 25,309 sequences of 4,000 subjects extracted from 39 cameras, with 18,940 sequences from 3,000 subjects for training and 6,369 sequences from 1,000 subjects for testing. The dataset mainly includes four data types: silhouettes, 2D pose, 3D pose, and 3D mesh. During testing, one sequence per subject is used as the probe set, while the remaining sequences are used as the gallery set. To evaluate the model's performance, we use accuracy as well as mean average precision (mAP) and mean inverse negative penalty (mINP) [49], which consider multiple instances and hard sample recall.

4.2. Implementation Details

Training details. 1) In our implementation, the margin m of the triplet loss is set to 0.2, and the parameter p of the GeM function used in the ASTP module is initialized to 6.5. In the hierarchy of gait motion, the number of partitions k at the bottom level is set to 8. 2) The batch size is set to (8,8) for CASIA-B, (32,8) for OUMVLP, (32,4) for GREW, and (32,4) for Gait3D. 3) We use gait silhouettes as the input

Table 2: The detailed architecture of the proposed HSTL on CASIA-B. The first column denotes the levels of the gait hierarchy and K_l is the number of groups at level l. C_{in} and C_{out} represent the input channel and output channel of each layer respectively. The body parts are indexed in spatial order from top to bottom, numbered 1 to 8.

Level	Block	Layer	C_{in}	C_{out}	Kernel	K_l	Parts Grouping
1	ARME	Conv3d	1	32	(3,3,3)	1	$\{\{1,2,3,4,5,6,7,8\}\}$
	ASTP						
2	ARME	Conv3d	32	32	(3,3,3)	2	$\{\{1,2,3,4,5\},\{6,7,8\}\}$
		Conv3d	32	64	(3,3,3)		
	ASTP						
2	FTA	MaxPool	64	64	(3,1,1)	2	$\{\{1,2,3,4,5\},\{6,7,8\}\}$
					(5,1,1)		
	ASTP						
3	ARME	Conv3d	64	128	(3,3,3)	4	$\{\{1\},\{2,3,4,5\},\{6,7\},\{8\}\}$
		Conv3d	128	128	(3,3,3)		
	ASTP						
4	ASTP					8	$\{\{1\},\{2\},\{3\},\{4\},$ $\{5\},\{6\},\{7\},\{8\}\}$

Table 3: Rank-1 accuracy (%) on OUMVLP under all views, excluding identical-view cases. Std denotes the performance sample standard deviation across 14 views.

Method	Probe View														Mean	Std
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°		
GaitSet [4]	79.3	87.9	90.0	90.1	88.0	88.7	87.7	81.8	86.5	89.0	89.2	87.2	87.6	86.2	87.1	4.0
GaitPart [8]	82.6	88.9	90.8	91.0	89.7	89.9	89.5	85.2	88.1	90.0	90.1	89.0	89.1	88.2	88.7	2.3
GLN [12]	83.8	90.0	91.0	91.2	90.3	90.0	89.4	85.3	89.1	90.5	90.6	89.6	89.3	88.5	89.2	2.1
CSTL [15]	87.1	91.0	91.5	91.8	90.6	90.8	90.6	89.4	90.2	90.5	90.7	89.8	90.0	89.4	90.2	1.1
GaitGL [27]	84.9	90.2	91.1	91.5	91.1	90.8	90.3	88.5	88.6	90.3	90.4	89.6	89.5	88.8	89.7	1.7
3D Local [16]	86.1	91.2	92.6	92.9	92.2	91.3	91.1	86.9	90.8	**92.2**	92.3	91.3	91.1	90.2	90.9	2.0
LagrangeGait [3]	85.9	90.6	91.3	91.5	91.2	91.0	90.6	88.9	89.2	90.5	90.6	89.9	89.8	89.2	90.0	1.4
MetaGait [5]	88.2	92.3	**93.0**	**93.5**	**93.1**	**92.7**	**92.6**	89.3	91.2	92.0	**92.6**	**92.3**	**91.9**	91.1	91.9	1.4
Ours	**91.4**	**92.9**	92.7	93.0	92.9	92.5	92.5	**92.7**	**92.3**	92.1	92.3	92.2	91.8	**91.8**	**92.4**	**0.5**

modality. During the training phase, we sample 30 frames following the strategy proposed in [8], while during testing, all frames are fed into the model. Moreover, we align each frame following the strategy presented in [39], and the input image size is cropped to 64×44 for all datasets. 4) For CASIA-B, the optimizer used is Adam with a weight decay of $5e$-4. The model is trained for 100K iterations with an initial learning rate (LR) of $1e$-5, and the LR is multiplied by 0.1 at 70K iterations. For OUMVLP, GREW, and Gait3D, the model is trained for 250K, 250K, and 210K iterations, respectively, with an initial LR of 0.1. The LR is multiplied by 0.1 at 150K and 200K iterations. The optimizer used for these datasets is SGD with a weight decay of $5e$-4.

Architecture details. Table 2 presents the detailed architecture of the model used for the CASIA-B dataset. To handle datasets with more subjects, such as OUMVLP, GREW, and Gait3D, we incorporate the label smoothing operation into the cross-entropy loss function, and deepen the network by adding an extra ARME module to the third level of the hierarchy. The output channels for the four ARMEs are set to 64, 64, 128 and 256, respectively. Additionally, we include a layer of spatial downsampling after the first ARME to improve the training efficiency.

Table 4: Rank-1 accuracy (%), Rank-5 accuracy (%), Rank-10 accuracy (%), Rank-20 accuracy (%) on GREW.

Methods	Rank-1	Rank-5	Rank-10	Rank-20
GaitSet [4]	46.28	63.58	70.26	76.82
GaitPart [8]	44.01	60.68	67.25	73.47
GaitGL [27]	47.28	63.56	69.32	74.18
MTSGait [58]	55.32	71.28	76.85	81.55
Ours	**62.72**	**76.57**	**81.32**	**85.24**

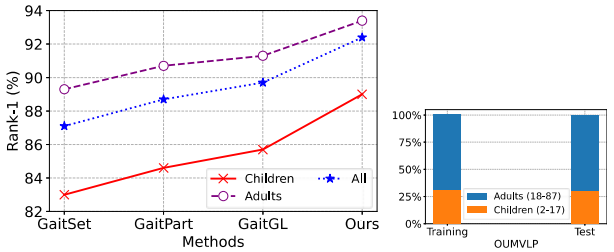
4.3. Comparison with State-of-the-Art Methods

Evaluation on CASIA-B. Table 1 compares the performance of the proposed HSTL with seven state-of-the-art

Table 5: Rank-1 accuracy (%), Rank-5 accuracy (%), mAP (%) and mINP on Gait3D.

Methods	Rank-1	Rank-5	mAP	mINP
GaitSet [4]	36.70	58.30	30.01	17.30
GaitPart [8]	28.20	47.60	21.58	12.36
GLN [12]	31.40	52.90	24.74	13.58
GaitGL [27]	29.70	48.50	22.29	13.26
CSTL [15]	11.70	19.20	5.59	2.59
SMPLGait [59]	46.30	64.50	37.16	22.23
MTSGait [58]	48.70	67.10	37.63	21.92
Ours	**61.30**	**76.30**	**55.48**	**34.77**

methods on the CASIA-B dataset. It can be seen that the proposed method obtains the best results in all three walking conditions while maintaining considerable stability across different views. The experimental results reveal that 1) the mean accuracy of the proposed method for the BG and CL walking conditions is 95.9% and 88.9%, respectively, which are higher by 0.7% and 2.0% than the second-best method (MetaGait [5]), demonstrating the advantage of our method in cross-view gait recognition. 2) The decrease in accuracy from NM to CL is 9.2% for our method, compared to 12.4% for 3D Local [16]. This indicates that the ARME module is more adaptable to various walking conditions. In addition, our method achieves an accuracy of 88.9% in the challenging CL condition, which is 5.2% higher than 3D Local. This may be due to the complex clothing conditions that affect the part localization accuracy in 3D Local. 3) Both CSTL [15] and our method extracts multi-scale temporal features but in different ways. CSTL first extracts spatial information and then fuses three scales of motion features. In contrast, we propose an FTA module to aggregate spatio-temporal information from multiple body regions. Therefore, the average rank-1 accuracy of our method is higher or equivalent to that of CSTL in all views. This suggests that FTA can be more adaptive to spatio-temporal changes.



(a) Average rank-1 accuracy (b) Age distribution

Figure 4: Cross-age comparison results for the OUMVLP dataset. (a) Average rank-1 accuracy in adults vs children. (b) Distribution of two age groups.

Evaluation on OUMVLP. To verify the model's generalizability, we conduct experiments on the large-scale OUMVLP dataset. As shown in Table 3, our approach achieves competitive results in most views. In particular, the proposed method outperforms the second-best method (MetaGait) by an average of 2.5% under three extreme views, i.e., 0°, 180°, and 270°, resulting in the best mean performance and cross-view stability. In addition, the OUMVLP dataset also provides annotations for the ages of the subjects. To evaluate the impact of age differences on recognition performance, we divided all subjects into two groups: adults (18-87 years old) and children (2-17 years old), as shown in Fig. 4(b). Fig. 4(a) presents the recognition accuracy based on age. It can be observed that, as adult sequences make up about 70% of the total sequences, all the compared methods show a bias toward the recognition results of adults. However, compared to other methods, our model effectively improves the accuracy of gait recognition for children, demonstrating the effectiveness of our hierarchical gait representation across ages.

Evaluation on GREW and Gait3D. Gait3D and GREW are two recently introduced datasets that contain challenging conditions, such as misalignment of the human body and partial occlusion. Tables 4 and 5 show the results of the comparison between the proposed method and the state-of-the-art methods. Our method shows superior performance in all metrics, indicating its ability to effectively model gait characteristics in realistic scenarios.

4.4. Ablation Study

Effectiveness of hierarchical feature extraction. To verify the effectiveness of our hierarchical gait partitioning, we conduct experiment with various grouping strategies. As shown in Table 2, this comparison only considers the different settings of the first three layers since a uniform partition is used at layer 4. Specifically, 1-1-1 indicates that the first three levels have the same number of groupings, and they are divided uniformly. 1-2-4* refers to the non-uniform division used in the proposed model. As shown in Fig. 5, the hierarchical feature extraction setting, e.g., 1-2-4, outperforms the other non-hierarchical approaches. A further mean performance improvement of 1.0% is achieved when the motion relationships between different body regions, i.e., 1-2-4*, are considered.

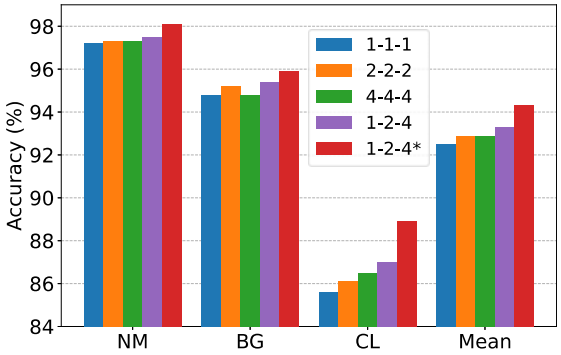


Figure 5: Ablation study on the effectiveness of hierarchical feature extraction on the CASIA-B dataset (best viewed in color).

Table 6: Ablation study on the effectiveness of ARME, ASTP, and FTA modules in terms of average rank-1 accuracy on the CASIA-B dataset.

Setting	ARME	ASTP	FTA		NM	BG	CL	Mean
			3	5				
a					97.0	94.5	84.2	91.9
b	✓				97.8	95.3	85.9	93.0
c	✓	✓			97.8	95.4	87.0	93.4
d	✓	✓	✓		97.7	95.3	87.5	93.5
e	✓	✓		✓	97.7	95.2	87.2	93.4
f	✓		✓	✓	97.8	95.4	87.5	93.6
g	✓	✓	✓	✓	**98.1**	**95.9**	**88.9**	**94.3**

Effectiveness of ARME, ASTP, and FTA. The results of the ablation experiments for ARME, ASTP, and FTA are shown in Tab. 6 The results indicate that the ARME module significantly contributes to the improvement of the recognition accuracy, with an average improvement of 1.1% compared to the baseline model (non-grouping version of ARME). The mean accuracy is further improved by 1.3% when the ASTP module is integrated and multi-scale fusion is utilized in FTA. These results demonstrate the effectiveness and complementarity of the three modules in the proposed gait recognition framework.

4.5. Trade-off between accuracy and efficiency

In this subsection, we evaluate the relationship between accuracy and efficiency for each compared method. As shown in Fig. 6, the 3D convolution-based approaches, such as GaitGL [27], 3D Local [16], LagrangeGait [3] and MetaGait [5], outperform the 2D convolution-based methods, like GaitSet [4] and GaitPart [8], in terms of accuracy, but

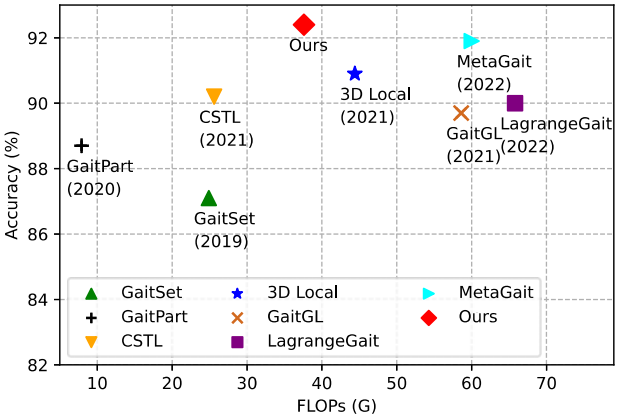


Figure 6: The trade-off between accuracy and FLOPs of our method and other comparison methods on the OUMVLP dataset.

at the cost of a significant increase in FLOPs (floating point operations). Our approach has a better trade-off between accuracy and efficiency. The main reason is that the proposed hierarchical learning architecture can extract multi-level motion features while reducing the number of 3D convolutions. More experimental results and ablation analysis are provided in the *supplementary material*.

5. Conclusion

This paper presents a hierarchical spatio-temporal representation learning (HSTL) framework for gait recognition. HSTL stacks multiple adaptive region-based motion extractors (ARMEs) and learns walking patterns in a coarse-to-fine manner. An adaptive spatio-temporal pooling (ASTP) module is proposed to perform hierarchical feature mapping for the output of each level of ARME. Additionally, a frame-level temporal aggregation module (FTA) is designed to compress local clips by fusing temporal information. The effectiveness of the proposed HSTL framework is demonstrated through extensive experiments conducted on four public datasets (CASIA-B, OUMVLP, GREW, and Gait3D).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61972132, 62106065) and the Research Project for Self-cultivating Talents of Hebei Agricultural University (Grant No. PY201810).

References

[1] Weizhi An, Shiqi Yu, Yasushi Makihara, Xinhui Wu, Chi Xu, Yang Yu, Rijun Liao, and Yasushi Yagi. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE TBIOM*, 2(4):421–430, 2020. 1, 2

[2] Pia Bideau, Aruni RoyChowdhury, Rakesh R Menon, and Erik Learned-Miller. The best of both worlds: Combining cnns and geometric constraints for hierarchical motion segmentation. In *CVPR*, pages 508–517, 2018. 1

[3] Tianrui Chai, Annan Li, Shaoxiong Zhang, Zilong Li, and Yunhong Wang. Lagrange motion analysis and view embeddings for improved gait recognition. In *CVPR*, pages 20249–20258, 2022. 1, 2, 3, 6, 7, 8

[4] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, pages 8126–8133, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[5] Huanzhang Dou, Pengyi Zhang, Wei Su, Yunlong Yu, and Xi Li. Metagait: Learning to learn an omni sample adaptive representation for gait recognition. In *ECCV*, pages 357–374. Springer, 2022. 2, 4, 6, 7, 8

[6] Jessica Maria Echterhoff, Juan Haladjian, and Bernd Brügge. Gait and jump classification in modern equestrian sports. In *ISWC*, pages 88–91, 2018. 1

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996. 2, 4

[8] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *CVPR*, pages 14225–14233, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[9] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, pages 8295–8302, 2019. 4

[10] Jinguang Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE TPAMI*, 28(2):316–322, 2005. 1, 2

[11] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *CVPR*, pages 2014–2023, 2021. 3

[12] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *ECCV*, pages 382–398, 2020. 1, 2, 3, 4, 7

[13] Saihui Hou, Xu Liu, Chunshui Cao, and Yongzhen Huang. Set residual network for silhouette-based gait recognition. *IEEE TBIOM*, 3(3):384–393, 2021. 2, 3

[14] Saihui Hou, Xu Liu, Chunshui Cao, and Yongzhen Huang. Gait quality aware network: Toward the interpretability of silhouette-based gait recognition. *IEEE TNNLS*, 2022. 2

[15] Xiaohu Huang, Duowang Zhu, Hao Wang, Xinggang Wang, Bo Yang, Botao He, Wenyu Liu, and Bin Feng. Context-sensitive temporal feature learning for gait recognition. In *ICCV*, pages 12909–12918, 2021. 1, 2, 3, 4, 5, 6, 7

[16] Zhen Huang, Dixiu Xue, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. 3d local convolutional neural networks for gait recognition. In *ICCV*, pages 14920–14929, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[17] Munish Kumar, Navdeep Singh, Ravinder Kumar, Shubham Goel, and Krishan Kumar. Gait recognition based on vision systems: A systematic survey. *Journal of Visual Communication and Image Representation*, 75:103052, 2021. 1

[18] Worapan Kusakunniran. Review of gait recognition approaches and their challenges on view changes. *IET Biometrics*, 9(6):238–250, 2020. 1

[19] Hongchao Li, Chenglong Li, Aihua Zheng, Jin Tang, and Bin Luo. Attribute and state guided structural embedding network for vehicle re-identification. *IEEE TIP*, 31:5949–5962, 2022. 2

[20] Xiang Li, Yasushi Makihara, Chi Xu, and Yasushi Yagi. End-to-end model-based gait recognition using synchronized multi-view pose constraint. In *ICCVW*, pages 4106–4115, 2021. 1, 2

[21] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. In *CVPR*, pages 13309–13319, 2020. 3

[22] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end model-based gait recognition. In *ACCV*, pages 3–20, 2020. 1, 2

[23] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *CVPR*, pages 510–519, 2019. 3

[24] Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In *ECCV*, pages 375–390. Springer, 2022. 2

[25] Beibei Lin, Yu Liu, and Shunli Zhang. Gaitmask: Mask-based model for gait recognition. In *BMVC*, page 363, 2021. 1, 2, 3, 4, 5

[26] Beibei Lin, Shunli Zhang, and Feng Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *ACM MM*, pages 3054–3062, 2020. 3, 4

[27] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *ICCV*, pages 14648–14656, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[28] Nini Liu and Yap-Peng Tan. View invariant gait recognition. In *ICASSP*, pages 1410–1413, 2010. 1

[29] Yasushi Makihara, Mark S Nixon, and Yasushi Yagi. Gait recognition: Databases, representations, and applications. *Computer Vision: A Reference Guide*, pages 1–13, 2020. 1

[30] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, pages 1363–1372, 2016. 2

[31] Daigo Muramatsu, Yasushi Makihara, Haruyuki Iwama, Takuya Tanoue, and Yasushi Yagi. Gait verification system for supporting criminal investigation. In *IAPR*, pages 747–748, 2013. 1

[32] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE TPAMI*, 41(7):1655–1668, 2018. 4

[33] Imad Rida, Noor Almaadeed, and Somaya Almaadeed. Robust gait recognition: a comprehensive survey. *IET Biometrics*, 8(1):14–28, 2019. 1

[34] Alireza Sepas-Moghaddam and Ali Etemad. Deep gait recognition: A survey. *IEEE TPAMI*, 2022. 1

[35] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *ICB*, pages 1–8, 2016. 1, 2, 3

[36] Pranjay Shyam, Kuk-Jin Yoon, and Kyung-Soo Kim. Adversarially-trained hierarchical feature extractor for vehicle re-identification. In *ICRA*, pages 13400–13407, 2021. 2

[37] Chunfeng Song, Yongzhen Huang, Yan Huang, Ning Jia, and Liang Wang. Gaitnet: An end-to-end network for gait based human identification. *PR*, 96:106988, 2019. 2

[38] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. 1

[39] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ TCVA*, 10:1–14, 2018. 2, 5, 7

[40] Lei Tan, Pingyang Dai, Rongrong Ji, and Yongjian Wu. Dynamic prototype mask for occluded person re-identification. In *ACM MM*, pages 531–540, 2022. 2

[41] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Towards a deeper understanding of skeleton-based gait recognition. In *CVPRW*, pages 1569–1577, 2022. 2

[42] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *ICIP*, pages 2314–2318, 2021. 1, 2

[43] Md Uddin, Daigo Muramatsu, Noriko Takemura, Md Ahad, Atiqur Rahman, Yasushi Yagi, et al. Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion. *IPSJ TCVA*, 11(1):1–18, 2019. 1

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 2

[45] Zhikang Wang, Lihuo He, Xiaoguang Tu, Jian Zhao, Xinbo Gao, Shengmei Shen, and Jiashi Feng. Robust video-based person re-identification by hierarchical mining. *IEEE TCSVT*, 2021. 2

[46] Xiu-Shen Wei, Chen-Lin Zhang, Lingqiao Liu, Chunhua Shen, and Jianxin Wu. Coarse-to-fine: A rnn-based hierarchical attention model for vehicle re-identification. In *ACCV*, pages 575–591, 2018. 2

[47] Jiachen Yang, Jianxiong Zhou, Dayong Fan, and Haibin Lv. Design of intelligent recognition system based on gait recognition technology in smart transportation. *Multimedia Tools and Applications*, 75(24):17501–17514, 2016. 1

[48] Lingxiang Yao, Worapan Kusakunniran, Qiang Wu, Jingsong Xu, and Jian Zhang. Collaborative feature learning for gait recognition under cloth changes. *IEEE TCSVT*, 2021. 1

[49] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE TPAMI*, 44(6):2872–2893, 2021. 6

[50] TzeWei Yeoh, Hernán E Aguirre, and Kiyoshi Tanaka. Clothing-invariant gait recognition using convolutional neural network. In *ISPACS*, pages 1–5, 2016. 1

[51] Shiqi Yu, Haifeng Chen, Qing Wang, Linlin Shen, and Yongzhen Huang. Invariant feature extraction for gait recognition using only one uniform model. *Neurocomputing*, 239:81–93, 2017. 1, 2

[52] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, pages 441–444, 2006. 2, 5

[53] Anguo Zhang, Yueming Gao, Yuzhen Niu, Wenxi Liu, and Yongcheng Zhou. Coarse-to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck. In *CVPR*, pages 598–607, 2021. 2

[54] Mingyang Zhang, Yang Xiao, Fu Xiong, Shuai Li, Zhiguo Cao, Zhiwen Fang, and Joey Tianyi Zhou. Person re-identification with hierarchical discriminative spatial aggregation. *IEEE TIFS*, 17:516–530, 2022. 2

[55] Shaoxiong Zhang, Yunhong Wang, and Annan Li. Cross-view gait recognition with deep universal linear embeddings. In *CVPR*, pages 9095–9104, 2021. 1

[56] Yuqi Zhang, Yongzhen Huang, Shiqi Yu, and Liang Wang. Cross-view gait recognition by discriminative feature learning. *IEEE TIP*, 29:1001–1015, 2019. 1, 3, 4

[57] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *CVPR*, pages 4710–4719, 2019. 2

[58] Jinkai Zheng, Xinchen Liu, Xiaoyan Gu, Yaoqi Sun, Chuang Gan, Jiyong Zhang, Wu Liu, and Chenggang Yan. Gait recognition in the wild with multi-hop temporal switch. In *ACM MM*, pages 6136–6145, 2022. 7

[59] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *CVPR*, pages 20228–20237, 2022. 1, 2, 6, 7

[60] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. Gait recognition in the wild: A benchmark. In *ICCV*, pages 14789–14799, 2021. 2, 5